

Geotagging matters?

The role of space and place in
politicised social media discourse

Adrian Paul Clive Tear

The thesis is submitted in partial fulfilment of the requirements

for the award of the degree of Doctor of Philosophy

of the University of Portsmouth.

September 2018

ABSTRACT

Voting results in marginal constituencies often determine wider political outcomes. Always a focus for campaigners, it is now apparent that voters in these areas have been individually and geo-behaviourally targeted by elaborate ‘psychological warfare’ operations designed to manipulate viewpoints and electoral results through advertising, (mis)information and/or ‘fake news’ disseminated online via popular social network sites. During the 2016 US Presidential Election, Russian operatives placed highly-politicised content supporting Donald Trump’s candidacy on Facebook and Twitter, segmenting audiences on these platforms by age, interests and location. In 2018, political marketing consultancy Cambridge Analytica was revealed to have earlier ‘hijacked’ data from Facebook, fusing it with other Big Data sources to promote Trump, in the US, and Vote Leave, in the 2016 UK European Union (‘Brexit’) Membership Referendum. Attempts to track the geographical diffusion of online politicking are hindered by incomplete geospatial referencing in available social media (meta)data; just ~1-2% of publicly-posted Twitter tweets, and even fewer Facebook posts, are typically ‘geotagged’ with Latitude and Longitude coordinates. Used successfully to monitor disaster situations or human mobility patterns this research examines ~8m interactions, created by ~2.4m users during the 2012 US Presidential Election and the 2014 Scottish Independence Referendum, to assess the role of space and place in politicised social media communications. Results of text, data-mining and statistical analyses demonstrate that coordinate-geotagging users of Twitter and Facebook, a) make fewer references to place in their message text, b) link to articles making fewer mentions of place in their content and; c) make far fewer links to external content than their non-coordinate-geotagging peers. Despite offering some valuable geospatial information, coordinate-geotagged interactions form only an inadequate and unrepresentative proxy for tracking the spread of all places, news, views, opinion, linked content or (mis)information shared online. Tackling the ‘crisis’ in deliberative democracy highlighted by recent data misuse and targeting scandals will, therefore, most likely require new political, regulatory and technical responses. One approach, suggested here, would store lower-resolution spatial information, e.g., identifiers uniquely referencing 1x1km grid squares or degraded Latitude and Longitude coordinates, alongside all social media interactions; enabling electoral officials, platform operators and others to more easily identify potentially nefarious content targeting *specific areas* as well as *specific individuals*.

TABLE OF CONTENTS

Abstract	ii
Table of Contents	iii
Declaration	xii
List of Tables.....	xiii
List of Figures	xvi
Abbreviations	xxiii
Acknowledgements.....	xxvii
Dedication	xxx
1 Introduction	1
1.1 Background.....	1
1.2 Social media data	11
1.3 Do social media data have any geographical value?.....	14
1.4 Defining key terms.....	17
1.4.1 Contentious meaning(s) of ‘space’ and ‘place’	17
1.4.2 Definition of key terms used in this thesis.....	18
1.5 Rationale for the research.....	22
1.5.1 Personal motivation	22
1.5.2 Wisdom of the Crowds?.....	24
1.5.3 Academic context.....	26
1.6 Relevance	30
1.7 Research hypothesis aim and objectives	34
1.7.1 Aim	35
1.7.2 Objectives.....	35

1.8	Introduction to the research process.....	37
1.8.1	Literature review.....	37
1.8.2	Case study-based data collection.....	39
1.8.3	Data storage, analysis, and interpretation	40
1.9	Originality and contribution to knowledge	40
1.9.1	Originality.....	40
1.9.2	Contribution to knowledge	41
1.10	Impact and engagement	43
1.10.1	Academic impact.....	44
1.10.2	Economic and societal impact.....	45
1.10.3	Engagement	47
1.11	Limitations of the research.....	47
1.12	Thesis structure	49
2	Literature and Context.....	51
2.1	Introduction.....	51
2.2	Text and data-mining in the literature review process.....	52
2.2.1	Methods	54
2.2.2	Results	57
2.3	Contextual synopsis.....	61
2.4	Political literature	64
2.4.1	Key publications	64
2.4.2	Key terms	65
2.5	Communications literature	72
2.5.1	Key publications	72

2.5.2	Key terms	72
2.6	Geographical literature	77
2.6.1	Key publications	77
2.6.2	Key terms	77
2.7	Technical literature.....	83
2.7.1	Key publications	83
2.7.2	Key terms	83
2.8	Gaps in knowledge	88
2.8.1	Geotagging rates	88
2.8.2	Representativeness.....	89
2.8.3	Toponymic usage	90
2.8.4	Localness	91
2.8.5	Testing the Geographality Assumption	91
2.9	Summary.....	92
3	Research Design.....	94
3.1	Introduction.....	94
3.2	Epistemology	96
3.3	Methodology	102
3.3.1	Case study methodology.....	105
3.3.2	Exploratory methodology	107
3.3.3	Hybrid case study/exploratory methodology	108
3.4	Ethics	111
3.5	Summary.....	116
4	Research Methods	118

4.1	Introduction.....	118
4.2	Data subjects	119
4.2.1	Characteristics of social media data	119
4.2.2	Social Network Analysis (SNA)	121
4.2.3	Technical proof of concept	123
4.2.4	Choice of case studies	126
4.2.5	Data acquisition.....	134
4.3	Data preparations.....	135
4.3.1	Data storage	135
4.4	Data procedures	147
4.4.1	Data augmentation	147
4.5	Data analysis.....	158
4.5.1	Data query, tabulation and analysis	159
4.5.2	Data visualisation	161
4.5.3	Statistical tests	163
4.6	Data measurements	164
4.6.1	Measuring and scoring ‘geographicality’ in OSN data	164
4.7	Summary.....	183
5	NLP/Geoparsing Results	186
5.1	Introduction.....	186
5.2	Research questions.....	188
5.2.1	RQ1 – How can baseline ‘geographicality’ be assessed and categorised in OSN data?	188
5.2.2	RQ2 – Does NLP-detectable ‘geographicality’ in message text increase in line with ‘spatiality’?	190

5.2.3	RQ3 – Does NLP-detectable ‘geographicality’ in linked/shared 3rd party content increase in line with ‘spatiality’?	205
5.3	Statistical tests.....	219
5.4	Software evaluation	221
5.4.1	Comparative evaluation.....	222
5.5	Summary.....	225
6	Discussion and Additional Findings	227
6.1	Introduction.....	227
6.2	Implications	229
6.2.1	Key implication.....	229
6.2.2	Other implications.....	237
6.3	Policy recommendations	238
6.3.1	Background	238
6.3.2	Regulatory responses.....	240
6.3.3	Technical responses	243
6.4	Additional findings.....	246
6.4.1	Spatiotemporality	246
6.4.2	Geo-retweeting.....	251
6.4.3	Data sparsity.....	255
6.4.4	Data fusion	262
6.4.5	Graph analysis	276
6.4.6	Data skewness.....	279
6.5	Summary.....	283
7	Conclusion.....	286

7.1	Introduction.....	286
7.2	New opportunities.....	289
7.3	[Geo]tagging, politics, prediction and tracking.....	292
7.4	Reflections and criticisms.....	297
7.5	Further (and future) research	299
7.6	Contributions to knowledge.....	305
7.6.1	Technological contributions.....	305
7.6.2	Substantive contributions	306
7.6.3	Policy contributions.....	308
7.7	Contributions to debate	310
7.8	Summary.....	317
	References.....	319
	Appendix 1 MaxMind GeoIP coding.....	407
A1.1	Introduction	407
A1.2	ColdFusion code.....	407
A1.2.1	application.cfm.....	407
A1.2.2	process_ips.cfm.....	407
A1.2.3	_process_one_ip_address.cfm	408
	Appendix 2 Word Cloud generation.....	411
A2.1	Introduction	411
A2.2	BibTeX file-based processing	411
A2.3	SQLite database-based processing	412
	Appendix 3 Academic literature text-mining	414
A3.1	Introduction	414

A3.2	Preparation	414
A3.3	Corpus creation.....	415
A3.4	Corpus analysis	416
A3.5	Word frequency export	418
Appendix 4	Ethical review correspondence.....	419
A4.1	Introduction	419
A4.2	Initial response.....	419
A4.3	Supplementary question	421
A4.4	Supplementary response.....	421
A4.5	Form UPR16	423
Appendix 5	DataSift Twitter licence	424
A5.1	Introduction	424
A5.2	Licence	424
Appendix 6	2012 French Presidential Election.....	427
A6.1	Technical proof of concept	427
A6.2	Outputs and analyses	427
Appendix 7	DataSift Stream Definitions.....	432
A7.1	Introduction	432
A7.2	2012 United States Presidential Election	432
A7.2.1	US2012_GEO	433
A7.2.2	US2012_NON_GEO	433
A7.2.3	US2012_NON_GEO_HISPANIC.....	434
A7.3	2014 Scottish Independence Referendum	435
A7.3.1	SCOT2014	435

Appendix 8	Computing environment.....	436
A8.1	Background	436
A8.2	Physical and Virtual computing environment	436
A8.3	System architecture	441
Appendix 9	Sparsity in the INTERACTIONS table	443
A9.1	Introduction	443
A9.2	INTERACTIONS table definition and metadata.....	443
A9.3	PL/SQL programme to count and store Zero Length Strings or NULLs into table ZERO_LEN_SPARSITY_INT_COLS	447
A9.4	SQL query to merge/update records into table ZERO_LEN_SPARSITY_INTERACTIONS.....	449
Appendix 10	AlchemyAPI processing	451
A10.1	Introduction	451
A10.2	Job controllers	451
A10.2.1	run_job.sh.....	451
A10.2.2	run_url_job.sh	451
A10.3	INTERACTION_CONTENT processing.....	451
A10.3.1	process_recs.rb	452
A10.3.2	mytest04.rb	454
A10.3.3	Included files.....	458
A10.4	LI_LINKS_URLS_DISTINCT processing	461
A10.4.1	process_url_recs.rb	461
A10.4.2	urltest04.rb.....	463
A10.5	Sample output	466
Appendix 11	SQL statements.....	479

Appendix 12	Statistical analysis in R.....	493
A12.1	R scripts.....	493
A12.2	Detailed statistical results and commentary.....	502

DECLARATION

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word count: 69,014

Signed:

A handwritten signature in black ink, appearing to read 'Adrian P. L. Teo'. The signature is fluid and cursive, with the first name 'Adrian' being more prominent.

Date: September 2018

LIST OF TABLES

Table 1-1 – Top 20 journal titles by number of stored articles	38
Table 3-1 – What case study and exploratory methodologies can tell us (after Labaree, 2017; quoted in italics).....	109
Table 3-2 – What case study and exploratory methodologies cannot tell us (after Labaree, 2017; quoted in italics).....	110
Table 4-1 – US2012: Summary of recorded OSN interactions (n=1,718,667)	128
Table 4-2 – SCOT2014: Summary of recorded OSN interactions (n=6,477,713).....	132
Table 4-3 – File listings, record counts and file sizes for CSV and JSON formatted data downloaded from DataSift.....	136
Table 4-4 – Transposition of nested, arrayed JSON into three CSV fields containing delimited string literals	146
Table 4-5 – Count of Interactions by Stream.....	160
Table 4-6 – Number and percentages of OSN Interactions by source, subtype and event	165
Table 4-7 – Descriptive statistics for interactions/user in the research data corpus (ALL), by event (US2012 and SCOT2014) and by source (FB=Facebook posts, TW=Twitter tweets, RT=Twitter retweets) whether all interactions (-ALL) or only coordinate-geotagged (-GEO)	167
Table 4-8 – Coordinate-geotagged Facebook posts, Twitter tweets and retweets by Stream and across the entire research data corpus.....	170
Table 4-9 – Numbers and percentages of original (Facebook posts and Twitter tweets) and reposted (Twitter retweet) coordinate-geotagged interactions in the research data corpus.....	171
Table 4-10 – Coding scores for 35 Potential Geographic Information (PGI) metadata fields	174
Table 4-11 – Example of TW_PLACE_ID co-populated metadata fields.....	176
Table 5-1 – US2012/SCOT2014: GATEcloud processing	194

Table 5-2 – US2012/SCOT2014: Number of resolved locations detected by GATEcloud in Facebook (FB), Twitter tweet (TW) and Twitter retweet (RT) interactions	197
Table 5-3 – US2012/SCOT2014: Number of records by tranche processed by AlchemyAPI	198
Table 5-4 – US2012/SCOT2014: Number of geographical entities detected in Twitter tweets by AlchemyAPI showing the rate (entities/tweet) for each sampled tranche	201
Table 5-5 – US2012/SCOT2014: Number of resolved locations detected by CLAVIN-rest in Facebook (FB), Tweet (TW) and Retweet (RT) interactions	204
Table 5-6 – US2012/SCOT2014: Number and percentage of linked URLs by Stream, OSN source and subtype (FB=Facebook, TW=Tweet, RT=Retweet) created by non-coordinate-geotagging and coordinate-geotagging users.....	207
Table 5-7 – US2012: Top 20 Domains and number of links for those interacting without and with coordinate-geotags (including retweets).....	209
Table 5-8 – SCOT2014: Top 20 Domains and number of links for those interacting without and with coordinate-geotags (including retweets).....	210
Table 5-9 – Entities, presented in tabular form, detected by AlchemyAPI against CNN’s Scottish Independence Referendum results page	212
Table 5-10 – US2012: Summary statistics of Welch’s paired T-tests for numbers of detected toponyms in geotagged/non-geotagged message text and linked/shared URLs at interaction and user levels by OSN type/subtype and parser	220
Table 5-11 – SCOT2014: Summary statistics of Welch’s paired T-tests for numbers of detected toponyms in geotagged/non-geotagged message text and linked/shared URLs at interaction and user levels by OSN type/subtype and parser	220
Table 6-1 – US2012/SCOT2014: Number of geo-retweet coordinate pairs, median, average and maximum straight-line geo-retweet distances	252
Table 6-2 – Abbreviations used in Column Names of the INTERACTIONS table	256
Table 6-3 – Top 10 fully populated fields/column names, commentary and utility	260

Table 6-4 – Next 7 highly populated fields/column names, commentary and utility	261
Table 6-5 – US2012: Top 10 users posting on Twitter by number of followers	280
Table 6-6 – US2012: Top 10 users posting with coordinate-geotagged messages on Twitter by number of followers	281
Table 6-7 – SCOT2014: Top 10 users posting without/with coordinate geotagged messages on Twitter by number of followers	282
Table 6-8 – File output sizes and database table sizes (in GB) for data augmentations used in this research	284

In Appendices

Table A7-1 – Summary of the US2012_GEO Stream	433
Table A7-2 – Summary of the US2012_NON_GEO Stream	433
Table A7-3 – Summary of the US2012_NON_GEO_HISPANIC Stream	434
Table A7-4 – Summary of the SCOT2014 Stream	435
Table A8-1 – Matrix of Host/Guest Operating System and Software Installations.	437
Table A9-1 – Columns in the INTERACTIONS table	443
Table A12-1 – US2012 geoparsing descriptive statistics (interaction level)	504
Table A12-2 – US2012 geoparsing descriptive statistics (user level)	505
Table A12-3 – SCOT2014 geoparsing descriptive statistics (interaction level)	506
Table A12-4 – SCOT2014 geoparsing descriptive statistics (user level)	507
Table A12-5 – US2012 geoparsing T-test results (x=NOTGEO, y=ISGEO) at interaction and user levels	508
Table A12-6 – SCOT2014 geoparsing T-test results (x=NOTGEO, y=ISGEO) at interaction and user levels	509

LIST OF FIGURES

Figure 1-1 – Facebook for Business location targeting options (Facebook, 2018d) ...	5
Figure 1-2 – 1997 UK General Election: World map showing GeoIP-derived geographical access to the website (~10% of page views, recovered from an archival fragment on Digital Audio Tape, DAT)	23
Figure 1-3 – US2012: Number of ‘mentions’ of the Candidates’ surnames in sampled Twitter tweets, retweets and Facebook posts	25
Figure 1-4 – US2012: World map showing sampled geotagged Twitter tweets and retweets (n=168,970).....	28
Figure 1-5 – CA (Cambridge Analytica) Political website, outlining company capabilities (<i>Cambridge Analytica, 2018</i>)	31
Figure 1-6 – US2012: Percentage population weighted mentions of State abbreviations in interaction message text.....	32
Figure 1-7 – US2012: Battleground States (BBC News, 2012)	33
Figure 1-8 – Top 100 terms from >1,250 stored article titles rendered as a Word Cloud	37
Figure 2-1 – Number of references by publication type selected for inclusion	56
Figure 2-2 – Number of references by year by type selected for inclusion	57
Figure 2-3 – Key terms (TF-IDF frequency >4,000) identified in the literature corpus	59
Figure 2-4 – Percentage publication titles by class.....	60
Figure 2-5 – Classification of papers according to authors’ academic research disciplines (after Figure 3, Steiger, de Albuquerque, et al., 2015, p815)	62
Figure 2-6 – Specific application domain of reviewed papers (after Figure 5, Steiger, de Albuquerque, et al., 2015, p817)	63
Figure 2-7 – Key political terms in the literature corpus identified by TF-IDF analysis	65
Figure 2-8 – Key communications terms in the literature corpus identified by TF-IDF analysis	72

Figure 2-9 – Key geographical terms in the literature corpus identified by TF-IDF analysis	78
Figure 2-10 – Key technical terms in the literature corpus identified by TF-IDF analysis	84
Figure 3-1 – Monthly Average User (MAU) counts, in millions, for major social media sites (Statista, 2018a)	95
Figure 4-1 – Simple representation of a Social Graph: 6 users A-F ('nodes') are connected to one another (by 'links') with node-size proportional to 'out-degree' (number of outbound links)	120
Figure 4-2 – Comparative size of Twitter's Streaming API (1% sample), Decahose (10% sample) and the full Firehose (100% of tweets)	124
Figure 4-3 – US2012: Search terms used, numeric and percentage contribution to OSN interactions sampled (n=2,676,331 total mentions of search terms in text) ..	127
Figure 4-4 – US2012: Worldwide distribution of coordinate-geotagged Twitter tweets (dark blue) and retweets (lighter blue)	129
Figure 4-5 – US2012: Spatialised network graph of 'Twitter mentions' relationships	130
Figure 4-6 – SCOT2014: Search terms used, numeric and percentage contribution to OSN interactions sampled (n=7,174,270 total mentions of search terms in text) ..	131
Figure 4-7 – SCOT2014: Worldwide distribution of coordinate-geotagged Facebook posts (orange), Twitter tweets (dark blue) and retweets (lighter blue)	133
Figure 4-8 – On-screen graphical representation of Presidential Candidate Barack Obama's tweet created on 05/11/2012 (from the raw JSON shown in Figure 4-9) ..	137
Figure 4-9 – A JSON formatted Twitter tweet sent from the account of Presidential Candidate Barack Obama and created on 05/11/2012	139
Figure 4-10 – Subjective scoring of the SQL and NoSQL data management systems used in this research (0=worst; 5=best)	141
Figure 4-11 – Scottish First Minister Alex Salmond's Twitter tweets processed using TwitIE on GATE Desktop	150

Figure 4-12 – Scottish First Minister Alex Salmond’s Twitter tweets processed using CLAVIN-rest running on a CentOS 7 virtual machine.....	155
Figure 4-13 – Shell script written to call CLAVIN-rest from the command line.....	156
Figure 4-14 – Log number of interactions/user by number of users in the research data corpus.....	165
Figure 4-15 – Log number of interactions/user by Log number of users: overall, by event and by OSN source (all records; all coordinate-geotagged records)	166
Figure 4-16 – Histogram of Log number of interactions/user against Log number of users	168
Figure 4-17 – Number of interactions by type created by prolific social network users ranked on number of interactions/user ($\geq 1,000$ interactions/user).....	169
Figure 4-18 – US2012/SCOT2014: Distribution of Geographicality Scores at interaction level across both events	178
Figure 4-19 – US2012/SCOT2014: Modal distribution of Geographicality Scores at user level across both events.....	178
Figure 4-20 – Percentage distribution of Modal Geographicality Scores for users making ≤ 5 interactions, ≥ 6 interactions and any number of interactions in the research data corpus.....	179
Figure 4-21 – Contribution of PGI metadata fields to each Geographicality Score class	180
Figure 4-22 – US2012: Interactions mapped by World Time Zones (yellow=low, red=high)	181
Figure 4-23 – SCOT2014: Interactions mapped by World Time Zones (yellow=low, red=high)	181
Figure 5-1 – US2012/SCOT2014: Average number of locations detected / interaction by Event, Geoparser and Geographicality Score	191
Figure 5-2 – US2012/SCOT2014: Average number of locations detected / user by Event, Geoparser and Modal Geographicality Score	191
Figure 5-3 – GATEcloud TwitIE output for Presidential Candidate Barack Obama’s Twitter tweet (Figure 4-9, p139).....	196

Figure 5-4 – US2012/SCOT2014: Number of records processed per day by AlchemyAPI	199
Figure 5-5 – US2012/SCOT2014: Number of distinct entities by type identified in message text by AlchemyAPI across all sampled tranches processed (n=311,575)	200
Figure 5-6 – CLAVIN-rest output showing resolved (i.e., successfully geoparsed) locations	203
Figure 5-7 – US2012/SCOT2014: Disambiguated geographic coordinates identified by AlchemyAPI in 641,472 distinct link URLs colour-coded by entity type	214
Figure 5-8 – US2012/SCOT2014: Number of distinct entities by type identified by AlchemyAPI in 641,472 distinct link URLs processed by the service	215
Figure 5-9 – US2012/SCOT2014: Top 10 entities for the two most detected entity types ('Person' and 'Organization') and three geographical entity types ('Country', 'City', 'StateOrCounty') identified by AlchemyAPI in 641,472 distinct link URLs	216
Figure 5-10 – Average number of toponymic or coordinate mentions identified by AlchemyAPI in linked URLs by grouped Geographicality Score at interaction level	218
Figure 5-11 – Average number of toponymic or coordinate mentions identified by AlchemyAPI in linked URLs by grouped Modal Geographicality Score at user level	218
Figure 5-12 – US2012: World map showing coordinate geotagged interactions (orange markers) and geoparsed locations (blue markers) identified by CLAVIN-rest	224
Figure 5-13 – SCOT2014: World map showing coordinate-geotagged interactions (orange markers) and geoparsed locations (blue markers) identified by CLAVIN-rest	224
Figure 6-1 – Detailed comments and snippet of 'geo-data "enrichment" code' created by Cambridge Analytica employee Michael Phillips (Sources: Albright, 2017; archive.today, 2017)	233
Figure 6-2 – Facebook campaign targeting parameters, and advertisement, for one of ~3,500 campaigns/advertisements set up by Russian state-sponsored actors during the 2016 US Presidential Election.....	235

Figure 6-3 – US2012: Spatiotemporal patterns of activity in 160,934 coordinate-geotagged interactions by day through to election night	251
Figure 6-4 – US2012/SCOT2014: Geographical dispersal of Twitter retweets in the UK and Ireland (both electoral events).....	253
Figure 6-5 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (first 50 columns).....	257
Figure 6-6 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (next 50 columns)	258
Figure 6-7 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (last 49 columns)	259
Figure 6-8 – 2010 Contiguous US population density at Census Tract level (light=low; dark=high)	267
Figure 6-9 – 2010 Contiguous US population density at Census Tract level and US2012 coordinate-geotagged OSN interactions	267
Figure 6-10 – US2012: Total population by age (OSN Tracts Indexed against US)..	268
Figure 6-11 – US2012: Male population by age (OSN Tracts indexed against US)..	269
Figure 6-12 – US2012: Female population by age (OSN Tracts Indexed against US)	269
Figure 6-13 – US2012: Population by race (OSN Tracts indexed against US).....	270
Figure 6-14 – US2012: Population in households and group quarters (OSN Tracts indexed against US).....	271
Figure 6-15 – SCOT2014: Population by age (OSN OAs indexed against UK).....	272
Figure 6-16 – SCOT2014: UK Output Areas (England=green, Wales=red, Scotland=blue) intersecting coordinate-geotagged OSN interactions	273
Figure 6-17 – SCOT2014: Male population by economic activity (OSN OAs indexed against UK)	274
Figure 6-18 – SCOT2014: Female population by economic activity (OSN OAs indexed against UK)	274
Figure 6-19 – SCOT2014: Population by country of birth (OSN OAs indexed against UK)	275

Figure 6-20 – Euler’s drawing of the bridges of Königsberg in 1736 and a graphical representation of the bridges of Königsberg in 1736 after Gribkovskaia et al. (2007, p200)	276
Figure 6-21 – SCOT2014: Gephi visualisation of ‘Twitter mentions’ (397,083 Nodes; 908,054 Edges) showing YouTube as the nexus between campaign-related interactions (in purple and green) and discussion of girl-band Fifth Harmony’s single ‘Better Together’ (in blue)	277
Figure 6-22 – American girl-band Fifth Harmony’s album, featuring title track Better Together, also the campaign slogan of the Unionist ‘Vote No’ coalition, was released on 18 October 2013 during data collection for the 2014 Scottish Independence Referendum	278
Figure 7-1 – US2012/SCOT2014: Number of toponymic mentions/user identified in message text (FB=Facebook, TW=Tweet, RT=Retweet) and linked/shared URL content	289
Figure 7-2 – SCOT2014: Number of social media posts per hour by times of day (light=day time; dark=night time) and day of week (smoothed by excluding counts from election day Thursday 18 September into Friday 19 September 2014)	294
Figure 7-3 – SCOT2014: Number of coordinate-geotagged social media posts per hour by times of day (light=day time; dark=night time) and day of week (smoothed by excluding counts from election day Thursday 18 September into Friday 19 September 2014).....	294

In Appendices

Figure A2-1 – Snippet of a BibTeX file	412
Figure A2-2 – Querying the Mendeley SQLite database to return lowercase article titles.....	413
Figure A3-1 – Mapping the working directory in VirtualBox to a Shared Folder on the Ubuntu virtual machine	414
Figure A4-1 – Dr Malcom Bray’s Ethical review response: ‘Favourable opinion with conditions’	421

Figure A4-2 – Email of 22/02/2016 (Adrian Tear - Ethical Review - Further Question)	421
Figure A4-3 – Email of 25/02/2016 in response to the supplementary question...	422
Figure A4-4 – Email of 25/02/2016 confirming acceptance of the opinion regarding the supplementary question.....	422
Figure A6-1 – Minute by minute Count and Cumulative total of OSN interactions recorded 4pm and 5pm on 6 May 2012 during the French Presidential Election second-round runoff	427
Figure A6-2 – % Null rows within columns in the French Presidential Election technical proof of concept data set; few columns/fields have fully populated rows	429
Figure A6-3 – Gender in the French Presidential Election technical proof of concept data set.....	429
Figure A6-4 – European distribution of explicitly geotagged interactions (n=736, 1.39% of the total) in the French Presidential Election technical proof of concept data set.....	430
Figure A6-5 – Saliency score and count of interactions mentioning ‘Hollande’ sent by Twitter users (with over 10,000 followers) in the French Presidential Election technical proof of concept data set	431
Figure A8-1 – ‘Shed-hosted’ personal Private Cloud (1*2U Dell PowerEdge 2950, 2*4U IBM System x3850 M2).....	439
Figure A8-2 – Maintenance activities on the University of Portsmouth’s SCIAMA supercomputer.....	440
Figure A8-3 – Schematic representation of containers (i.e., physical host, virtual machines, Cloud), data flows, data transfer mechanisms and software applications employed in this research	441
Figure A12-1 – Hierarchical levels for statistical analysis of toponymic place name detection in case study social media data	502

ABBREVIATIONS

ACORN	A Classification of Residential Neighbourhoods
AGI	Ambient Geographic Information
AL	AlchemyAPI NLP (message text)
ANNIE	A Nearly-New Information Extraction system
ANSI	American National Standards Institute
AOL	America Online
API	Application Programming Interface
AWS	Amazon Web Services
BCP	Bulk Copy
BLOB	Binary Large Object
CASA	Centre for Advanced Spatial Analysis
CIA	Central Intelligence Agency
CL	CLAVIN-rest NLP (message text)
CLI	Command Line Interface
CLOB	Character Large Object
CNN	Cable News Network
COTS	Commercial Off the Shelf
CS	Computer Science
CSDL	Curated Stream Definition Language
CSS	Computational Social Science
CSV	Comma Separated Values
DAT	Digital Audio Tape
DAU	Daily Active Users
DMA	Designated Market Area
DoD	US Department of Defense
DTM	Document Term Matrix
DTS	Data Transformation Services
ESRC	Economic and Social Research Council
ESRI	Environmental Systems Research Institute, Incorporated

ETL	Extract, Transform, Load
EU	European Union
EXIF	Exchangeable Image File
FB	Facebook post
FOAF	Friend of a Friend
GATE	General Architecture for Text Engineering
GB	Gigabyte (1024 ³ bytes)
GBP	GB Pound (£)
GIR	Geographical Information Retrieval
GIS	Geographical Information System
GMT	Greenwich Mean Time
GOP	Grand Old Party; Republicans
GPS	Global Positioning System
GT	GATEcloud TwitIE NLP
GUI	Graphical User Interface
HDFS	Hadoop Distributed File System
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
I/O	Input/Output
IBM	International Business Machines
ICG	Institute of Cosmology and Gravitation (UoP)
IJGIS	International Journal of Geographical Information Science
IMR	Internet Mediated Research
IP	Internet Protocol
IPv4/IPv6	Internet Protocol version 4/version 6
IS	Information Systems
IT	Information Technology
JAR	Java ARchive
JISC	Joint Information Systems Committee
JSON	JavaScript Object Notation
KML	Keyhole Markup Language

LAT/LON	Latitude/Longitude
LBS	Location Based Services
LI	AlchemyAPI NLP (linked/shared URL content)
MAU	Monthly Active Users
NER	Named Entity Recognition
NLP	Natural Language Processing
NoSQL	Not-only Structured Query Language
OCI8	Oracle Call Interface 8
ODBC	Open Database Connectivity
OSN	Online Social Network
PDF	Portable Document File
PGI	Potential Geographic Information
POS	Part of Speech
RAID	Redundant Array of Inexpensive Disks
RAM	Random Access Memory
RDBMS	Relational Database Management System
REGEXP	REGular EXPression
REST	Representational State Transfer
RGS-IBG	Royal Geographical Society (with the Institute of British Geographers)
RT	Twitter retweet
SaaS	Software as a Service
SCOT2014	2014 Scottish Independence Referendum OSN data set (A7.3)
SNA	Social Network Analysis
SNP	Scottish Nationalist Party
SNS	Social Network Site
SQL	Structured Query Language
SQLldr	Oracle SQL*Loader
SSD	Solid State Disk
SSDT	SQL Server Data Transformation (tools)
SSIS	SQL Server Integration Services

TCP/IP	Transmission Control Protocol/Internet Protocol
TDM	Term Document Matrix
TF-IDF	Term Frequency – Inverse Document Frequency
TOR	The Onion Ring
TW	Twitter tweet
UGC	User Generated Content
UK	United Kingdom
UoP	University of Portsmouth
URL	Uniform Resource Locator
US	United States
US2012	2012 US Presidential Election OSN data set (A7.2)
USD	US Dollar (\$)
UTC	Universal Time Coordinated
UTF-8	Unicode Transformation Format – 8 bit
UUID	Universally Unique IDentifier
VGI	Volunteered Geographic Information
VM	Virtual Machine
VPN	Virtual Private Network
WMD	Weapons of Mass Destruction
WWW	World Wide Web
XML	eXtensible Markup Language

ACKNOWLEDGEMENTS

This research was inspired by work conducted in 1997 with my friends and business partners Tony Sellen and Mark Watson together with Robert Waller, Byron Criddle, Vanessa Lawrence, Ed Parsons and others to produce a website covering the UK General Election held that year. As well as providing an extensive range of maps, voting data, constituency and candidate profiles this project, which won the Association for Geographic Information's 1997 annual award, sparked an interest in politics and the online consumption of political information, which is further developed in this thesis. I am still in touch with most of my collaborators from those days and must thank Mark Watson, particularly, for his assistance with geodemographic profiling and various statistical techniques.

The UK General Election website of 1997 was a largely one-directional 'Web 1.0' publication, adapted in 2000 to disseminate facts about the London Mayoral Election and later fully updated to cover the UK General Election held in 2001. Conversely, the current research would not have been possible without the development and growth of more participatory 'Web 2.0' sites and mobile applications enabling the multi-directional spread of User Generated Content (UGC). The opinions, expressed in the public domain, of ~2.4 million users of Online Social Network (OSN) platforms, predominantly of the popular micro-blogging site Twitter, but also of Facebook, have provided the raw material examined in this research. In 1997, 2000 and 2001 Internet users consumed political information from publishers' websites; today they consume, link to, share and comment on publishers' material as well as creating original content of their own.

The ~8 million records of social media message text, metadata and linked/shared content accessed during the 2012 US Presidential Election and the 2014 Scottish Independence Referendum campaigns have provided an excellent, if at times hard to interpret, politically discursive corpus that has been text and data-mined in

various ways. I thank these users, the social media platforms which have made this possible and data aggregator DataSift, without whom it would have been impossible to access so many individual thoughts and opinions.

Naturally, I must also thank my first supervisor Professor Richard Healey for supporting me in this endeavour. Richard and I have known each other for over twenty-five years since I studied under him at the Department (now the Institute) of Geography at the University of Edinburgh. When in 2010 I found that Richard had moved to the University of Portsmouth, and was now somewhat closer to my own home location, he was fully supportive of my research proposal to investigate the relevance of geography in modern-day social media communications. Various other individuals have assisted during this research. My second supervisor, Professor Humphrey Southall, is an expert in historical geography and well-versed in the vagaries of gazetteer search. Our work together on a 'social media' chapter for the forthcoming book, *Data in Society*, has been particularly instructive, helping to validate and confirm several opinions also expressed in this thesis. Academic and support staff at the University of Portsmouth and my friend and colleague Bruce Gittings at the University of Edinburgh, where I deliver a lecture series to students of the M.Sc. Geographical Information Science (GIS) degree in connection with my Honorary Fellowship in the School of GeoSciences, have also provided useful comments on drafts and research outputs.

Large parts of the data storage and analysis work conducted during this research have been made possible by continual advances in database technology and 'Big Data' processing systems. I must thank Richard Pitts and Alastair Fraser at Oracle and Adam Fowler at MarkLogic for their assistance in getting the most out of their respective database systems. Michael Hausenblas and Leon Clayton have assisted with the installation and use of MapR's Hadoop-based ecosystem tools, particularly Apache Drill. Professor Kalina Bontcheva, Ian Roberts, and colleagues in the Natural Language Processing Group in the Department of Computer Science at the

University of Sheffield have been particularly helpful, providing access to Sheffield's GATEcloud natural language processing system and extending its functionality in several directions to meet requirements identified in this research.

Gary Burton, High Performance Computing Support Officer in the Institute of Cosmology and Gravitation, has provided invaluable assistance in the setup of a five-node MapR Hadoop cluster on the University of Portsmouth's SCIAMA supercomputer. David Marshall, Principal Database Administrator, University of Portsmouth, has provided equally valuable assistance in the setup of Oracle 11g and Oracle 12c RDBMS instances on two other SCIAMA supercomputer nodes. Rich and Josh (surnames unknown) at IBM Bluemix Watson Services Support have helped by providing academic access to the Cloud-hosted AlchemyAPI semantic text analysis suite. Much other software used in this research, e.g., the R Project for Statistical Computing, Gephi graph and Tableau data visualisation packages, are either open-sourced or available under favourable academic licensing terms.



Finally, I must thank my wife Elspeth McVey, and my children Patrick, Roslyn and Charlie for their support, encouragement and patience during this period. Our dog, Lena, has also spent many hours by my side in the office; patiently waiting for me to throw her the ball. Hopefully, now this thesis has been completed, there will be a bit more time for doing so!

DEDICATION

I dedicate this work to my late father, Group Captain R.C. Tear.

1 INTRODUCTION

1.1 Background

Online social networks, politics and Big Data, are in the news. Explosive revelations surrounding Cambridge Analytica's long-standing misuse of Facebook data for political marketing purposes have prompted a US Congressional Committee inquiry to investigate data usage, sharing and privacy policies at Facebook (McKinnon & Seetharaman, 2018; U.S. House of Representatives, 2018b) 'leaving the internet giant scrambling to contain a growing scandal over how it treats its users' (Dinan, 2018). Political campaigning using advanced behavioural and psychographic targeting, alongside geographical micro-marketing (Albright, 2017) designed to bring out or win over key voters, may even have affected the outcome of the 2016 US Presidential Election, a contest which Cambridge Analytica claimed to have 'won' for Donald Trump (P. Lewis & Hilder, 2018). The misuse of large amounts of personal data, together with Russian state-sponsored interference in electoral processes through the promotion of frequently inflammatory material on popular social networks, including Facebook and Twitter (BBC News, 2018c), has been widely reported in the mainstream media (Cadwalladr & Graham-Harrison, 2018; New York Times, 2018).

In the UK, connections between political strategists from Cambridge Analytica (2018) and pro-Brexit Vote Leave campaigners contesting the 2016 UK European Union Membership Referendum have led to claims that 'data-analytics' provided by CA political (Figure 1-5, p31) and AggregateIQ (Ram, 2018) have effectively 'hijacked our democracy' (Cadwalladr, 2017). These assertions will soon be tested by a group of British expatriates (Bowcott, 2018) taking a case to the High Court challenging the legality of the 'Brexit' result (S. Wilson, 2018) following Vote Leave's £449,079 over-spend during the campaign; recently detected, and penalised with several fines, by The Electoral Commission (2018b). Campaign spending on 'digital'

has risen steadily in the UK (Sabbagh, 2018) as ‘trends captured by the terms professionalization, marketization and mediatization explain dramatic shifts in the way parties execute their election campaigns’ (Lilleker, Tenscher, & Štětka, 2014, p747). These developments, and the related issues surrounding the promulgation of ‘fake news’, which researchers have found travels particularly rapidly across online social networks (Vosoughi, Roy, & Aral, 2018), have prompted a House of Commons Select Committee Inquiry (Digital Culture Media and Sport Committee, 2018) to call on noted academic experts in media and communications studies for evidence (Fuchs, 2017c). Amongst Ministers and Members of Parliament there is widespread concern that targeted digital communications and/or fake news, disseminated over the Internet or on social media networks, may be ‘crowding out’ real news, creating a ‘crisis’ for British democracy (BBC News, 2018h).

In the US, the newly-elected 45th President of the United States, Donald Trump, faces a smouldering investigation by Special Counsel Robert Mueller into his links, and his campaign team’s links, with Russian operatives seeking to influence the outcome of the 2016 election which brought him to power. Since his appointment as head of the ‘independent federal investigative and prosecutorial agency’ (U.S. Office of Special Counsel, 2018) on 17 May 2017, Mueller has ‘indicted 22 criminal defendants and garnered five guilty pleas, including from Michael Flynn, the former national security adviser; Richard Gates III, a former deputy to campaign chairman Paul Manafort; and George Papadopoulos, a former Trump foreign policy aide’ (McCarthy, 2018). Trump’s campaign ally, Roger Stone, may be next in Mueller’s firing-line (Swaine, 2018) as ‘special sections’ in many newspapers (e.g., The New York Times, 2018) suggest that coverage of the Special Counsel’s investigations into Trump’s electoral campaign will continue to run and run.

In academia, scholarly articles now ask ‘Can Democracy survive the Internet?’ (Persily, 2017) where once there had been ‘a relatively brief period of euphoria about the possibility that social media might usher in a golden age of global

democratization’ (Tucker et al., 2018, p3). Social networks, the data they hold and the extensive and targetable reach they enable have become newsworthy – and an increasingly important subject area for academic research – as the widespread realisation has dawned that several of the phenomenally successful Internet businesses established in the last decade or so (Facebook, Google, Twitter et al.) have grown rich and perhaps, at times, been somewhat careless in navigating the Faustian bargain made by so many people in exchanging ever more personal information for free, expertly-developed, and frequently pervasive software applications designed to touch so many aspects of our ‘connected’ lives (Andersson, 2018).

Several other revelations regarding the (mis)use of social media platforms and Big Data repositories have also emerged. For example, the development of a quiz ‘app’, or application, created by a University of Cambridge academic (Etter & Frier, 2018) and installed by just 305,000 people led directly to the ‘harvesting’ of ~87 million Facebook user profiles in 2015 including, ironically, details from Facebook-founder Mark Zuckerberg’s own personal account (The Independent, 2018). Data collected by the ‘ThisIsYourDigitalLife’ app, and shared with Cambridge Analytica, exploited the ‘social graph’, or set of inter-linkages between Facebook users, and the platform’s API, or Application Programming Interface (BBC News, 2018e). These data were later used to ‘build psychological profiles of voters in the United States’ and elsewhere (Frenkel, Rosenberg, & Confessore, 2018).

Governments, regulatory agencies and lawyers in the US, UK, EU and elsewhere are currently examining the multiple breaches of online trust engendered by these developments, while the social science and scientific community to some extent play ‘catch-up’; even though issues surrounding ‘homophily’ (Abbasi, Zafarani, Tang, & Liu, 2014), ‘echo chambers’ (Gilbert, Bergstrom, & Karahalios, 2009), ‘polarization’ (De Nooy & Kleinnijenhuis, 2013), ‘participation’ (de Zúñiga, Veenstra, Vraga, & Shah, 2010) and ‘misinformation’ (Budak, Agrawal, & El Abbadi, 2011) in


politicised online social media discourse have been researched for quite some time (Chapter 2, p51).

The present study, drawing on a part-time research programme commenced in 2011, is situated within both of these evolving and contemporary contexts. The research examines how ‘place’ and ‘space’ are differentially used, referenced and/or shared in politically discursive messages by individuals interacting on two popular Online Social Network (OSN) platforms; Facebook and Twitter. Two relatively recent political events provide the backdrop to the collection of case study material examined in this work; the 2012 United States Presidential Election (US2012) and the 2014 Scottish Independence Referendum (SCOT2014). These case studies have been selected as elections offer a data-rich environment with contemporaneous opinion polling, known post-electoral outcomes (Bond & State, 2015) and clearly-defined, and geographically ‘bound’, voting districts or constituencies (Cox, 1969; Elden, 2005; Giddens, 1985).

Electorates in these areas are now increasingly targetable, and are being actively targeted online, by political parties and marketing professionals seeking to influence turnout or voting behaviour in the small number of marginal ‘swing’ states or constituencies that often determine wider political outcomes in established Western democracies (Lilleker et al., 2015; Moore, 2016; The Electoral Commission, 2018a). Bespoke geographical targeting campaigns, as developed by Cambridge Analytica, may exploit (Section 6.2, p229) toponymic references found in users’ self-reported ‘Location’ fields (Hecht, Hong, Suh, & Chi, 2011), toponymic references found in users’ publicly-posted message text (Stock, 2018) and/or Latitude and Longitude coordinates deposited in OSN metadata when users optionally choose to ‘geotag’ their social media posts (Kumar, Morstatter, & Liu, 2014). Platform operators, such as Facebook (2018d), also provide many facilities (Figure 1-1, p5) for locational targeting; on County or Region, City, Designated Market Area (DMA, in the US), Postcode or Business address.

facebook business

Reach people in the areas where you do business



Location targeting helps you to find people where you do business, helping you to create adverts that are relevant to people according to their location.

You can already choose from areas near you, including country and postcode, but we have now expanded features that will give you even more ways to reach people in specific areas.

[Create Advert](#)

Selecting your location

Location targeting lets you select your audience within a custom radius from the following locations:

- Country
- County or region
- City
- DMA®
- Postcode
- Business address

*DMA® (Designated Market Area) regions are the geographic areas in the US in which local television viewing is measured by Nielsen.

Audiences
Choose a Custom Audience
Browse
Create New Custom Audience...

Locations
United States, California
San Francisco + 50 mi
Berkeley + 25 mi
1501 Willow Rd, Menlo Park, CA, USA + 10 mi
Add a country, state/province, city, ZIP or address
Everyone in this location
Age
18 - 65+
Gender
All Men Women
Languages
Enter a language...
More Demographics

Refining your audience

Audiences can be refined based on which audience would be most interested in your business. The choices for audiences within a location are:

- (Default) **Everyone in this location.** People whose current city on their Facebook profile is that location, as well as anyone determined to be in that location via mobile device.
- People who live in this location.** People whose current city from their Facebook profile is within that location. This is also validated by IP address and their Facebook friends' stated locations.
- Recently in this location.** People whose most recent location is the selected area, as determined only via mobile device. This includes people who live there or who may be travelling there.
- People travelling in this location.** People whose most recent location is the selected area, as determined via mobile device, and are greater than 100 miles from their stated home location from their Facebook profiles.

Audiences
Choose a Custom Audience
Create New Custom Audience...

Locations
United States, California
San Francisco + 50 mi
Berkeley + 25 mi
Add a country, state/province, city, ZIP or address
Everyone in this location
People who live in this location
People recently in this location
People travelling in this location
Age
Gender
Languages
Enter a language...
More Demographics

'Targeting audiences by location and age has truly increased my ROI on advertising.'

Sandra Iheuwa, Owner, Fashion Bee Hair Store

Get started with location targeting today

[Create Advert](#)

Figure 1-1 – Facebook for Business location targeting options (Facebook, 2018d)

Advertisements can be displayed to all Facebook users in (or within a radius of) selected locations, users who live in those locations (also 'validated' by Internet Protocol, IP, address), users currently in those locations ('as determined only by

mobile device’) or people just passing through (‘as determined by mobile device [when it is] greater than 100 miles from their stated home location from their Facebook profile’). Other major social media or ‘destination’ websites operated by Google (2018c), Instagram (2018a), Snapchat (2018a), Twitter (2018a) and YouTube (2018) offer broadly similar facilities to advertisers – or political campaigners – using their services. All also offer targeting on age, basic demographics (e.g., gender) and, in several cases, on more advanced behavioural or similarity traits (e.g., interests and ‘Lookalike Audiences’ in Facebook’s case). It is currently unclear whether recent attempts to distort the outcome of democratic elections through geo-behavioural targeting, a type of online gerrymandering, have shown clear ‘monolithic effects [but] the impact of social media in political campaigning around the world is undeniable’ (Dimitrova & Matthes, 2018).

Steiger, de Albuquerque, & Zipf (2015, p816) have noted that coordinate-geotagged OSN interactions, sourced primarily from Twitter, have demonstrated high degrees of utility in ‘research on event detection [particularly in the] investigation of abnormal spatial, temporal and semantic tweet frequencies [surrounding] disaster and emergency [situations].’ The current research uses a mixture of coordinate-geotagged and non-coordinate-geotagged social media data from Facebook and Twitter to determine whether a similarly high level of utility may be observed in political contexts. Understanding how different classes of social media users imprint their message text with place or, less frequently, space – or consume, link to and share 3rd party Uniform Resource Locator (URL) content imprinted with place – is essential when attempting to accurately track the downstream diffusion of deliberately geo-targeted political advertising.

Computational techniques have been used in this study to sample, collect, store, query, map and measure any ‘geographicality’ (Relph, 1985) expressed in publicly-posted OSN interactions, the message text and metadata bundles downloadable from several social network platforms. As Relph (1985, p16) has noted, ‘The

experiences of places, spaces and landscapes in which academic geography originates are a fundamental part of everyone's experience.' Different forms of geographical expression, including toponymic references to place in message text and linked/shared content and the spatial coordinates deposited alongside messages in OSN metadata by geotagging users, are widely-made in social media discourse. This is especially true during elections, which are both political *and* geographical events; defined in modern democracies as much by the concepts of 'territorial, representative' constituencies (Rehfield, 2005) and 'place' (Johnston & Pattie, 2006) as by the exigencies of political candidacy and/or entrenched local allegiances.

Over 8 million OSN interactions created by ~2.4 million users, ~90% sourced from Twitter and ~10% from Facebook, have been recorded and analysed (Chapter 4, p118) in this research. Typically, only a small proportion (~1-2%) of Twitter interactions are geotagged with Latitude and Longitude coordinates (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013) and similarly low coordinate-geotagging rates are reported here (Chapter 5, p186). Large absolute numbers of these small percentages of coordinate-geotagged OSN interactions have, however, often been used somewhat uncritically (Bertrand, Bialik, Virdee, Gros, & Bar-Yam, 2013; Compton, Jurgens, & Allen, 2014), as Leszczynski & Crampton (2016) have argued, to track or map the spread of message text, sentiment or links shared online.

Many such studies rely on the *Localness Assumption* recently identified and tested by I. L. Johnson, Sengupta, Schöning, & Hecht (2016) in which it is 'implicitly assumed' that geo-references in text are proximal to coordinate-geotagging users' recorded locations. This is found to be true in only around 75% of cases, in turn raising a more fundamental question; who references place in message text or linked/shared content most? Is it the small percentage of coordinate-geotagging users on OSN platforms or the majority comprised of their non-coordinate-geotagging peers? Answering this question helps determine whether messages sent

by coordinate-geotagging users may be reliably used as proxies for the flow of all geographically-referenced information exchanged on these sites. This knowledge may be used to inform tactics subsequently adopted to track the spread of political (mis)information disseminated over social media channels since unless, or until, legislation forces a change in policy (BBC News, 2018f) political parties, candidates, campaign managers and OSN platform operators are not currently compelled to reveal online geo-targeting strategies (The Electoral Commission, 2018a).

Using modern relational and document store databases and several proprietary and open-source software systems and APIs the case study data sets have been extensively 'mined' to provide quantitative measures of expressed geographicality, using coordinates and toponymic mentions in message text and linked/shared content as an online substitute for 'understanding people's relationship with places and geographical environments' (Seamon & Lundberg, 2017). As the storage and analysis of large numbers of social media interactions present some peculiar challenges, especially when handling terse free-form text (Derczynski, Maynard, Aswani, & Bontcheva, 2013; Tear & Healey, 2017), the computing environment and associated technologies used to produce these outputs are described in detail. The work has been conducted within an exploratory spatiotemporal research methodology, proposed by N. Andrienko, Andrienko, & Gatal'sky (2003), and developed here to provide useful exemplars of methods and techniques (Chapter 4, p118) which can be used to derive meaning and find locations in a massive research data corpus containing over 230 million space-tokenised words.

Key outputs from the research, results of which are presented in Chapter 5 (p186) with a discussion and additional observations given in Chapter 6 (p227), include the findings that:

- Messages from coordinate-geotagging users on Facebook are, on average, found to be 'liked' slightly less (1.33 vs. 1.50) than those from the universe of users. On Twitter, the source of 89.72% of all sampled messages,

coordinate-geotaggers have fewer 'friends' (median 325 vs. 345) and fewer 'followers' (median 275 vs. 348) than non-coordinate-geotagging users.

These differences, while apparently small, are highly significant when calculated for 2,117,577 Twitter users across the research data corpus, a much larger number than that used in many other political studies using social media data and a number far larger than any doctoral research presented in the era preceding the availability of OSN Big Data.

- Coordinate-geotagging/non-coordinate-geotagging users' median Klout scores, a measure of 'influence scoring' within and across several different social networks (Rao, Spasojevic, Li, & Dsouza, 2015), are near identical at 40 and 41 respectively. Median values, rather than averages, are used here as social networks are highly skewed by major celebrities or political figures, such as gossip columnist Perez Hilton or former President Barack Obama, having many millions of Facebook 'friends' or Twitter 'followers' (Section 6.4.6, p279).
- Analysis of the case study interactions, using three Natural Language Processing (NLP) systems and data-mining via Structured Query Language (SQL) constructs, shows that coordinate-geotagging users mention identifiable locations less in their messages, link to less 3rd party content and link to 3rd party content containing fewer identifiable geographical 'entities' (cities, towns, states etc.) than non-coordinate-geotagging users during the two case study events (Sections 5.2.2, p190 and 5.2.3, p205).

Contrary to expectation, and the research hypothesis set out below (Section 1.7, p34), it appears that the small, spatially valuable and (apparently) *most geographical* class of coordinate-geotagging OSN users – whose Latitude and Longitude locations allow straightforward and potentially accurate mapping of online information consumption and sharing patterns – are somewhat *less geographically expressive* and *link to less 3rd party content* than OSN users in general. Consequently, and significantly, this finding implies that tracking or

mapping the spread of places, news, views or opinion by searching for phrases or toponyms in message text created by coordinate-geotagging users alone, or by searching for specific URL links shared alongside these messages, does not provide an adequate proxy for tracking the geographical spread of all politically discursive material created and shared online over social media networks. This conclusion is reported at exactly the time when electoral officials and other analysts wishing to trace the diffusion of micro-marketed, geo-targeted political communications disseminated by companies such as Cambridge Analytica or external agents, such as Russian ‘trolls’ (BBC News, 2017c) seeking to influence or interfere in democratic processes by promoting particular content to particular people in particular places, might turn to coordinate-geotagged social media data for precisely this purpose.

If governments, regulators, researchers or citizens want to know *where* social media content are being consumed or shared, in order to detect potential interference in electoral processes through geo-targeted advertising aimed at specific marginal ‘swing’ states or constituencies, more transparency in social media (meta)data and reporting are required. Policy recommendations in this area are outlined in Chapter 6 (Section 6.3, p238) and would involve recording lower-resolution geographical coordinates in metadata alongside all social media interactions; safeguarding users’ high-resolution locational privacy (unless full coordinate-geotagging were opted-in to, as now) while providing useful geographical oversight. Such a change would, undoubtedly, require legislation to regulate the operations of large technology companies; an idea which commentators, politicians and mainstream media organisations across the political spectrum have suggested now appears increasingly likely (Fildes, 2018; Sunstein, 2018; Taylor, 2018).

The following sections of this introductory chapter describe social media data in more detail before defining key terms used throughout this thesis and explaining the rationale for conducting this research. The hypothesis, aim and objectives of

the research are stated and a brief introduction to methodology is given before the contribution to knowledge, impact and engagement are set out. The final section details the overall structure of this thesis.

1.2 Social media data

Social media data are typically, although not exclusively (Marechal, 2016), created by individuals as they produce, share and comment on content online. These data sets are widely used in government, corporate and academic environments. Applications include the surveillance and monitoring of citizens (Fuchs, 2017b), business brand and reputation management (Grabher & König, 2017) and wide-ranging investigations in the Computational Social Science (CSS) and Information System (IS) domains (see Kapoor et al., 2017 for a useful summary of major research topics). As the ‘participatory’ Web 2.0 model (O’Reilly, 2005) has superseded ‘publication’ on the World Wide Web (WWW) several rapidly-evolving websites and applications, e.g., Facebook, Flickr, Twitter, Wikipedia and YouTube have promoted the creation and enabled the storage and, to varying extents, retrieval of increasingly large volumes of User Generated Content (UGC). Some of these human-made digital artefacts – consisting of text, shared URL links, audio, image or video files – are ‘publicly posted’ online (Hough, 2009) allowing widespread, although seldom free (Zelenkauskaitė & Bucy, 2016), access to potentially huge volumes of material.

Social media data are generally time-stamped in Universal Time Coordinated (UTC) allowing sequencing by creation date and time. Individual records are often packaged for downloadable access, with metadata, in JavaScript Object Notation (JSON) format (ECMA International, 2013, 2017). Some data, e.g., Flickr images or Twitter tweets, may be geotagged with Latitude and Longitude coordinates allowing straightforward mapping of social data phenomena (Miller & Goodchild, 2015). Key demographic or address information, e.g., age, sex, street, town or postcode are not, for privacy reasons, available in downloadable social media data,

although some, e.g., gender, may be imputed with varying levels of success (Diaz, Gamon, Hofman, Kiciman, & Rothschild, 2016). In some cases, however, and especially when users have given sufficient 'read access' to 3rd party social media applications, these variables may be visible. Aleksandr Kogan's ThisIsYourDigitalLife app collected data in exactly this way, while also exploiting Facebook's 'friend' relationships to 'harvest' data from inter-connected user accounts, explaining Cambridge Analytica's interest in his work (W. Davies, 2018).

The two prominent OSNs studied most frequently in social science research are Facebook, founded in 2004, and Twitter, founded in 2006. These sites, or 'platforms' (Barreneche & Wilken, 2015), are now accessed by >2 billion and >300 million Monthly Active Users (MAU) respectively (Facebook, 2018b; Statista, 2018b). Facebook (2018b) claim that 66% (1.37 billion) of its user base are Daily Active Users (DAU), thought to upload >350 million images (Macagba, 2017) and share 4.75 billion content items every day (Fu, Wu, & Cho, 2017). Twitter, which does not report DAU statistics, is thought to publish ~500 million tweets per day (Worldometers, 2018). Massive usage of Facebook has necessitated development of complex, large-scale storage infrastructures designed to handle daily multi-petabyte (10^{15} bytes) uploads of text, image and video content (Wiener & Bronson, 2014). Twitter data scientists, likewise, describe 'plumbing' together multiple 'Big Data' (Magoulas & Lorica, 2009) software systems to run 'jobs [accomplishing] everything from data cleaning to simple aggregations and report generation to building data-powered products to training machine-learned models for promoted products, spam detection, follower recommendation, and much, much more' (Lin & Ryaboy, 2013, p6).

End users experience fast, responsive and highly-personalised websites and mobile applications which promote networked content sourced from 'friends' (Facebook) or other user accounts being 'followed' (Twitter). Increasingly, as the two major operators have developed successful 'advertising monetisation' models (N.

Newman, Fletcher, Levy, & Nielsen, 2016), Facebook ‘walls’ and Twitter ‘timelines’, the personalised home pages users navigate around on these sites, have also featured targeted interjections from advertisers, which may include both ‘real’ and ‘fake’ news (Hogan, 2018) and ‘clickbait’ (Kirkby, 2016). OSN users are encouraged to constantly engage in continual exploration of content through user interfaces featuring ‘infinite’ or ‘never ending’ page scrolling (J. Kim, Zhang, Kim, Miller, & Gajos, 2014) and are psychologically rewarded for ‘liking’ content, ‘retweeting’ posts or building large ‘friend’ or ‘follower’ networks through addictive triggers which have been deliberately designed-in to most social network applications (Andersson, 2018; Andreassen, Pallesen, & Griffiths, 2017).

Social media data consumers in government, commerce or academia may download or ‘stream’ publicly-posted data either directly through a number of individual publishers’ APIs (Facebook, 2018a; Lane, 2017; Twitter, 2017) or by using 3rd party social data aggregators, such as DataSift, which ‘[manages] upstream API integration and [provides] a single-point-of-access to upwards of twenty individual social media data sources’ (Tear, 2014, p223). A second major aggregator, GNIP, recently acquired by Twitter, offers access to the full ‘Firehose’ of current and historic tweets and is now the only such source (Hern, 2014). Twitter is the most widely researched OSN platform, ‘despite being only 11th in global rankings by number of users’ (Stock, 2018, p227). Twitter’s Streaming API, through which it is possible to freely capture a 1% sample of all social media interactions posted on the site in real time (Stone, 2006), is chiefly responsible for this bias (Tufekci, 2014).

Many other sources of social media data do, however, exist. DataSift (2018), for example, offers ‘free real-time monitoring’ of Wikipedia edits, alongside paid-for access to feeds aggregated from news, social network, video and blogging sites including Blogger, DailyMotion, IMDb, LexisNexis, NewsCred, Reddit, Topix, Tumblr, WordPress, YouTube and many other smaller sites. Rising stars of the OSN world, such as Instagram (2018b) and Snapchat (2018b), have also enabled API

functionality. Instagram's images, for example, have recently been 'fused' with Twitter text in a 'social sensing' experiment (Giridhar, Wang, Abdelzaher, Amin, & Kaplan, 2017). The photographic image storage and sharing site Flickr (2018) offers an API which has been used, e.g., to detect and geolocate images of endangered wildlife species in protected areas for criminological research (Lemieux, 2015). Google+, the OSN developed by the Web search giant, also offers 'read-only access to public data' through its own API (Google, 2018a). In use, all of these social media data sources may be queried to subset manageable numbers of records for further analysis. The potential for deriving geographical value from sampled social media data is discussed below.

1.3 Do social media data have any geographical value?

The availability and use of geotagging functionality on OSN websites has provided Geographical Information Scientists ('GIScientists'), geographers and others accessing data from Twitter, since 2009 (Sarver, 2009), and Facebook, since 2010 (Parr, 2010), with significant amounts of 'Volunteered' (Goodchild, 2007) or 'Ambient' (Stefanidis, Crooks, & Radzikowski, 2013) Geographic Information (VGI/AGI). Coordinate-geotagging involves the practice of sharing a posting location alongside message text or other content, available in social media interaction metadata as a Latitude and Longitude pair. Many articles and maps, both in the academic literature and more widely published in print or on the Web, have exploited coordinate-geotagged OSN data to produce a variety of geographical outputs. Different types of coordinate-geotagged or 'geosocial' (Bahir & Peled, 2013) Big Data have been used to monitor road traffic congestion (Work, Blandin, Tossavainen, Piccoli, & Bayen, 2010), manage crisis events (Goodchild & Glennon, 2010) or plot multiple attributes of London life (O'Brien & Cheshire, 2014). Steiger, de Albuquerque, et al. (2015) provide a useful summary of the many application domains using coordinate-geotagged OSN data.

Typically, and somewhat unfortunately for geographers, only small percentages of social media interactions are geotagged with Latitude and Longitude coordinates. Leetaru et al. (2013) report that just 1.6% of ~1.5 billion Twitter interactions analysed in their study contained 'Exact locations'. Slightly higher rates have been reported elsewhere (Croitoru, Crooks, Radzikowski, & Stefanidis, 2013) with variability attributed to event type (e.g., an elevated 16% following the Fukushima nuclear disaster in Japan), cultural practice (e.g., some nations use OSNs more frequently than others) and differing technological factors (e.g., smartphone adoption rates). While widely-used in many countries there are also notable 'black holes' in worldwide OSN coordinate-geotagged space, particularly in North Korea and China where neither Facebook or Twitter are allowed to operate (Graham, Stephens, & Hale, 2013; Leetaru et al., 2013; M. S. Weber & Monge, 2011).

'Place' is more commonly used than 'space' in OSN communications. Frequent toponymic mentions of place are found in both message text and associated interaction metadata (Gelernter & Mushegian, 2011; Pavalanathan & Eisenstein, 2015; Stefanidis, Cotnoir, et al., 2013). Coordinate-geotags, when present, are most commonly appended to OSN interactions by smartphone 'apps' (S. Li et al., 2016; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; Wei, 2013) installed on mobile devices equipped with Global Positioning System (GPS) chipsets designed to capture and/or share locational data (Kumar et al., 2014; L. Li, Goodchild, & Xu, 2013). While most users' mobile devices are perfectly capable of imprinting coordinates alongside their Web or OSN posts geotagging is an 'opt-in' feature which users must explicitly enable in their software application (Sui, 2017), although reports have surfaced which suggest that Google collected Android phone users' locations 'even when location services [were] disabled' (Collins, 2017). Few users choose to deliberately activate geotagging facilities (Tasse, Liu, Sciuto, & Hong, 2017) and, consequently, most mapping and geographical analyses of social media interactions are enabled not by the majority of OSN users, but by *a distinct minority* who choose to post with coordinates. It is thought that low rates of

coordinate-geotagging amongst social media users reflect a strong desire to protect 'locational privacy' online (Cottrill, 2011; de Souza e Silva, 2013; Tsou & Leitner, 2013) although Egelman, Felt, & Wagner (2013, p21) have found that some users, in a more generic context, may 'view the location permission [request on Android smartphones] as an indicator of desirable functionality rather than an indicator of privacy risk.' C. W. Chang & Chen (2014, p36) have found that study participants 'were more likely to disclose their location on Facebook if their friends did so, a concept called subjective norm' while others (Cottrill, 2011) have suggested that press coverage of tongue-in-cheek websites such as PleaseRobMe.com (Van Grove, 2010) have alerted users to the dangers of posting coordinates online, where differences in 'home' and 'away' locations can easily be used to infer presence or absence with potentially dire consequences. Overall it appears that most OSN users have no interest in coordinate-geotagging and do not turn the feature on, a) unless their friends do so, or; b) they feel the usefulness of the feature, even if used temporarily (Tasse et al., 2017), outweighs their more general predisposition to safeguard locational privacy.

As a ~1-2% function of the 'vast' volume of social media interactions made online (S. C. Lewis, Zamith, & Hermida, 2013), absolute numbers of coordinate-geotagged posts used in research may be high, particularly if a long-running OSN recording filtered on the *presence* of coordinates has been used. Several long-term studies designed to capture and map very large numbers of coordinate-geotagged posts have been reported, analysing geotagging rates (Leetaru et al., 2013), socio-economic phenomena (L. Li et al., 2013), global patterns of human synchronization (Morales, Vavilala, Benito, & Bar-Yam, 2017) and the demographics of coordinate-geotagging users (Sloan & Morgan, 2015). Physical devices, hardware and software amalgams such as the 'Tweet-o-meter' developed by University College London's Centre for Advanced Spatial Analysis (CASA), have also been produced; designed to continually update with coordinate-geotagged Twitter tweets made around 16 major world cities (S. Gray, Milton, & Hudson-Smith, 2015).

To varying degrees these works all rely on the assumption that ‘this one percent [of spatialised interactions] is already large enough’ for meaningful geographical analyses (Jiang, Ma, Yin, & Sandberg, 2016, p349). Many studies also implicitly assume, incorrectly in around 25% of cases as I. L. Johnson et al. (2016) have demonstrated, that a high degree of ‘localness’ is exhibited in coordinate-geotagged OSN interactions; i.e., that locations mentioned in spatialised OSN message text are proximal to the coordinates of the post.

The current research questions whether there is an over-reliance on ‘geosocial’ data deposited by just ~1-2% of all social media users and whether expressions of ‘place’ in message text and linked/shared content are highly correlated with ‘space’ in coordinate-geotagged OSN interactions. The work does not repeat I. L. Johnson et al.'s (2016) study, by comparing coordinate geotags to locations derived from toponymic references in adjacent message text, but addresses a more fundamental question: who makes, or links to external content containing, the most place-based references on social media networks; coordinate-geotagging or non-coordinate-geotagging users of these sites? Answering this question, to determine the toponymical representativeness of coordinate-geotagging users, helps determine whether or not this minority group may be used as ‘markers’ to accurately and spatially, through their Latitude and Longitude coordinates, trace the geographical diffusion of politically discursive material.

1.4 Defining key terms

1.4.1 Contentious meaning(s) of ‘space’ and ‘place’

The preceding sections have detailed the background to this study (Section 1.1, p1), introduced social media data (Section 1.2, p11) and discussed whether, and how, social media data might offer any geographical value to researchers (Section 1.3, p14). Concepts surrounding the identification of ‘space’ and ‘place’ in social media (meta)data have been outlined, yet these two terms – whose nuanced and

multifarious meanings are amongst the most heavily contested in academic geography and related social science disciplines (Hubbard & Kitchin, 2011) – require further definition as they apply to the current research. Definitions of several key terms used throughout this thesis are given in the following section.

1.4.2 Definition of key terms used in this thesis

Agnew (2011) has suggested that ‘place becomes a particular or lived space’ as humans associate a ‘location somewhere’, or their occupation of that location, with its spatial (or locational) position. Place, therefore, may take on multiple meanings or refer to spaces – the living room, the home, the home town or the country – which have very different geographical extents and which may vary, conceptually, from one person to another. Social media interactions holding discursive text, or text-based metadata, are not immune from such semantic or conceptual dichotomies; a Twitter tweet containing the word ‘Kansas’ may refer either to the US State of that name or to Kansas City, or to both. Equally, depending upon context, the mention might be shorthand referring to a popular baseball or football team located in the State or City in question, or to some other location or logical entity altogether (e.g., the Wildcats football team of Kansas State University). Place has social meaning, but determining the context within which a place is mentioned in social media message text, and the exact meaning implicit in that mention of place, is not necessarily a straightforward task.

Senses of known-place(s), affirmed-place(s) and space(s), some of which may be accompanied by apparently ‘accurate’ Latitude and Longitude coordinates, are often highly conflated in social media data. Users of Twitter, for example, when registering, are asked ‘Where in the world are you?’ (Hecht et al., 2011) and may just as reasonably answer ‘BRICK city bitch’ or ‘Somewhere, Overthere’ as ‘Concord, NC’ or ‘iPhone: 40.699490,-73.891556’. Difficulties inherent in identifying and parsing Potential Geographic Information (PGI) in free-form social media message text and associated (meta)data are amplified considerably when, as in this research,

place-based geographical references must be detected computationally (Section 4.4, p147). Consequently, and as huge data volumes preclude individual human examination of over ~8 million social media interactions, necessarily focused definitions of ‘space’ and ‘place’ are adopted in this research:

- **Space** – Refers in this thesis to geographically and explicitly *locational* data, i.e., to a point (most often) or a geographical extent (much less frequently) defined by one or more pairs of Latitude and Longitude coordinates. Almost all space-based data emanates directly from the small subset of coordinate-geotagged social media interactions in the research data corpus. However, additional spatial data (or ‘spatialities’) may be inferred from non-coordinate-geotagged interactions either by post-processing (meta)data or by ‘geoparsing’ message text to append coordinates, where possible, to detected references of place.
- **Place** – Refers in this thesis to computationally-identifiable geographical references in text, i.e., to *toponymic* place names, e.g., of towns, cities, counties, states or countries etc. Information Extraction (IE) and Named Entity Recognition (NER) techniques from Natural Language Processing (NLP) are used to detect such geographical references in social media data and linked/shared content. The software systems used either rely upon large, open-source gazetteers of toponymic place names (e.g., the ~11 million records available from GeoNames, 2016) or use smaller gazetteers supplemented by logical ‘rules’ (Tear & Maynard, personal communication, 2018) to boost place identification based upon the co-occurrence of certain terms (e.g., ‘Isle of...’, ‘Mount...’, ‘Cape...’) associated with place names.

Space, where it exists in social media interaction (meta)data, may *generally* be regarded unambiguously; the Latitude and Longitude coordinates of a user’s location have been recorded alongside their message text by a GPS-equipped mobile device just at the moment of message creation. Exceptions exist, of course,

such as the production of coordinate-geotagged messages by robotic networks ('botnets'), described by Marechal (2016) and exemplified by Echeverría & Zhou's (2017) detection of the 'Star Wars' botnet, responsible for creating 1.2 million coordinate-geotagged Twitter tweets in North America and Western Europe; or through the presence of unlikely, or nonsensical, spatial coordinates such as the 227 interactions with 0 Latitude and 0 Longitude in the research data corpus.

Place, in social media data, retains many of the elements of ambiguity identified by Tuan (2001) and other geographical theorists but is referenced, on the admittedly narrower grounds adopted here, much more widely in message text, metadata and linked/shared content than space (Section 5.2.2, p190 and 5.2.3, p205). In addition, there is both, a) some overlap between 'space' and 'place' in digital social media data, and; b) some potential to 'move' from space to place, or vice versa. For example:

- a) The metadata of some OSN interactions records users' time zone offsets relative to Greenwich Mean Time (GMT) in seconds. While these values (e.g., 14,400 or -18,000) are neither explicitly space- or place-based, converting seconds to hours and minutes allows data fusion with world time zone boundaries enabling the small-scale worldwide mapping of online activity (Figure 4-22 and Figure 4-23, p181).
- b) The Latitude and Longitude coordinate pairs deposited alongside individual's geotagged interactions may, likewise, be 'fused' to official areal units such as US or UK Census boundaries using GIScience techniques (Section 6.4.4, p262). This process enables other types of imputation and reporting which, by exploiting the geographical hierarchy implicit in US and UK Census data (e.g., US Census Tracts aggregate to Counties and States; UK Output Areas to Wards, Local Authorities and Counties etc.) may yield place-based results from purely spatial coordinate data. Conversely, and as an example of the movement from place to space, all toponymic place names detected in

message text and successfully geoparsed may be mapped to expand the locational scope of the case study data sets beyond just the small subset of spatially coordinate-geotagged social media interactions present in the research data corpus (Figure 5-12, p224 and Figure 5-13, p224).

A final term, used throughout this thesis, encapsulates the different space- and place-based meanings in social media data outlined above:

- **Geographicality** – Refers in this thesis to the multiplicity of geographical forms of expression evident in social media data, ranging considerably both, a) in *scale*, from world time zones covering parts of continents to point-based locations of message creation, and; b) in *nature*, e.g., incorporating mentions of place(s), again at many different scales, either in message text or linked/shared content, which may or may not be amenable to spatial augmentation through geoparsing. Measuring and scoring geographicality in social media (meta)data (Section 4.6.1, p164) enables cross-comparison of space- and place-based facets of geographical expression at several levels; by case study event; by OSN platform; by user and, most atomically; by interaction (i.e, individual message and metadata bundle). The results of this work are presented in Chapter 5 (p186) with additional findings presented in Chapter 6 (p227).

In their Introduction to *Key Thinkers on Space and Place*, a bibliographic compendium detailing theoretical contributions from 66 scholars of geography and related social science disciplines, Hubbard & Kitchin (2011, p7) state that ‘given the way space and place have been operationalised, they remain relatively diffuse, ill-defined and inchoate concepts.’ In this thesis, meanings of ‘space’ and ‘place’ are measured and operationalised much less diffusely, having been clearly defined above. While necessarily focused, the definitions adopted here enable machine-based classification of very large volumes of social media data; affording an opportunity to determine how space and place are used online, and whether

different user groups – especially the most-spatial, coordinate-geotagging, users of two popular Online Social Network platforms – make differential references to place. The rationale for conducting this research is set out below. The relevance of determining who makes the most mention of place in message text, or links to and shares content making the most mention of place via online social media channels, is detailed later in Section 1.6 (p30).

1.5 Rationale for the research

1.5.1 Personal motivation

This research project is informed by much earlier work (Tear, 1997), undertaken during the relative infancy of the World Wide Web, to develop a website covering the 1997 UK General Election. At that time, in what is now thought of as the ‘Web 1.0’ era (Helles, 2013), the project involved the collection and publication of facts (UK constituency boundary maps, turnout and voting statistics, candidate and constituency profiles) and the provision of search functionality (clickable maps, postcode to constituency lookups etc.) enabling public access to those facts.

The 1997 UK General Election website was widely-used, recording well over one million ‘hits’ in web server log files which were analysed to monitor site performance, understand load issues and identify popular pages (Chu, Wipfli, & Valente, 2013). Heatmaps generated using GeoIP mapping functionality (MaxMind, 2012; Appendix 1, p369) built in to Webtrends (2018) log file analysis software showed fascinating, and geographically uneven, patterns of public information access to the published material. The ‘cash for questions’ scandal, for example, which prompted BBC journalist Martin Bell (Independent) to stand against incumbent Neil Hamilton (Conservative) in the Tatton constituency (Farrell, McAllister, & Studlar, 1998) sparked nationwide interest in that page, while other pages were generally viewed by users more proximal to the constituency in question.

Accepting known imperfections in IP addressing, e.g., the fact that all America Online (AOL) traffic originated from just a few IP addresses, and the varying spatial accuracy of the GeoIP database it was reasonably straightforward, and has now become commonplace for webmasters (Google, 2018b), to map the spatial origin of requests for given pages; to know *where* people access pages from and *which* pages they access.

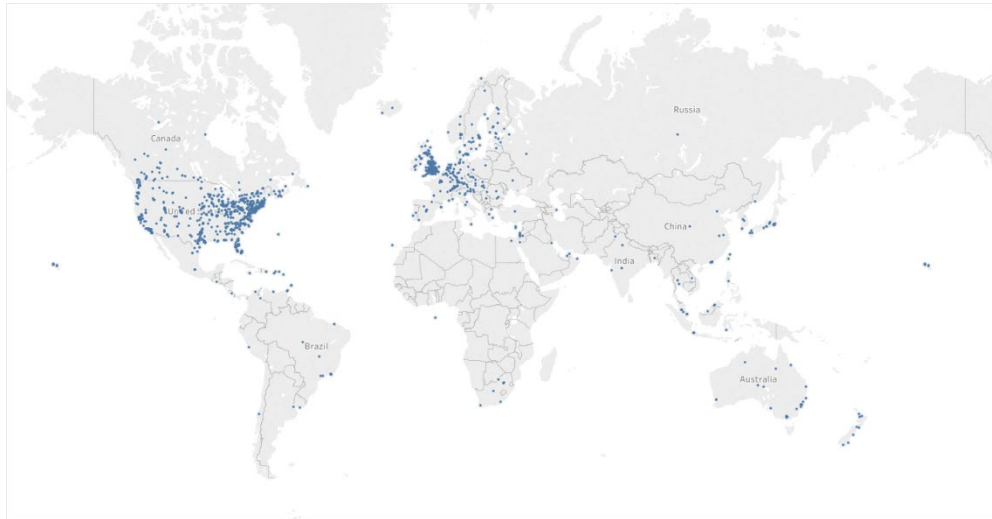


Figure 1-2 – 1997 UK General Election: World map showing GeolIP-derived geographical access to the website (~10% of page views, recovered from an archival fragment on Digital Audio Tape, DAT)

Early work to model distance decay effects in web server information flows (Murnion & Healey, 1998, p285) had suggested that latency, time-delays in network communications, might be used as a proxy ‘to determine the effect of Internet distance on the number of expected visits to a web server.’ Sharp distance decay curves were observed in worldwide access to UK academic web servers yet the 1997 UK General Election website was both widely accessed from overseas, particularly from the US and Europe (Figure 1-2, p23), and was also differentially accessed page-by-page as users consumed information about each of the UK’s, then, 641 parliamentary constituencies. Some of these differences could be attributed to elevated levels of interest prompted by events during the 1997 election campaign (e.g., Tatton), or large numbers of inbound links to certain

constituency pages from 3rd party websites, but others could not. This raised the tantalising prospect that further analysis of geographical patterns of politicised online information consumption might prove informative, or perhaps even predictive in terms of modelling turnout or voting intention, if enough representative political opinions were known. In 1997 there was no easy way of knowing *what* large numbers of people said or thought as they consumed political information online. The arrival of social media platforms in the mid-2000s dramatically reversed this deficiency.

1.5.2 Wisdom of the Crowds?

Only recently has it become possible, as ‘Web 2.0 desires to read, write, and share personal information’ have developed (Jung, 2015, p53), to know *what* large numbers of people are *saying* or *thinking* as they comment on, share and interact with content online. The rapid growth of OSNs such as Facebook and Twitter has brought billions of users and countless user messages into the public domain. UGC now abounds and the traditional communications model of Habermas’ (2011) *Public Sphere* incorporating governmental, judicial and media power-players appears to have moved towards a more pluralistic model involving overlap between public and newly digitally-enabled ‘private’ spheres (Papacharissi, 2010).

Political opinions expressed online are now widely-made, shared, and increasingly accessible for download from OSN platforms; some records are coordinate-geotagged and many more make frequent toponymic mention of place (Han, Cook, & Baldwin, 2014). For privacy reasons, other than to site operators, IP addresses are not made available in OSN data downloads; removing one of the easiest – although not necessarily accurate (Backstrom, Sun, & Marlow, 2010) – methods to geographically estimate the location of social media communications. If, as Surowiecki (2004) has suggested, the ‘Wisdom of the Crowds’ really can provide more accurate prediction, is it possible that ‘mining’ massive amounts of OSN data for spatiotemporally expressed political opinion could help to detect events, or

possibly even ‘call’ an election? After all, as some scholars have suggested, ‘there is a strong relationship between political information [consumption, news seeking] and political participation; the more we know about politics, the more, and more effectively, we participate in political activities’ (Feezell, 2016, p495).

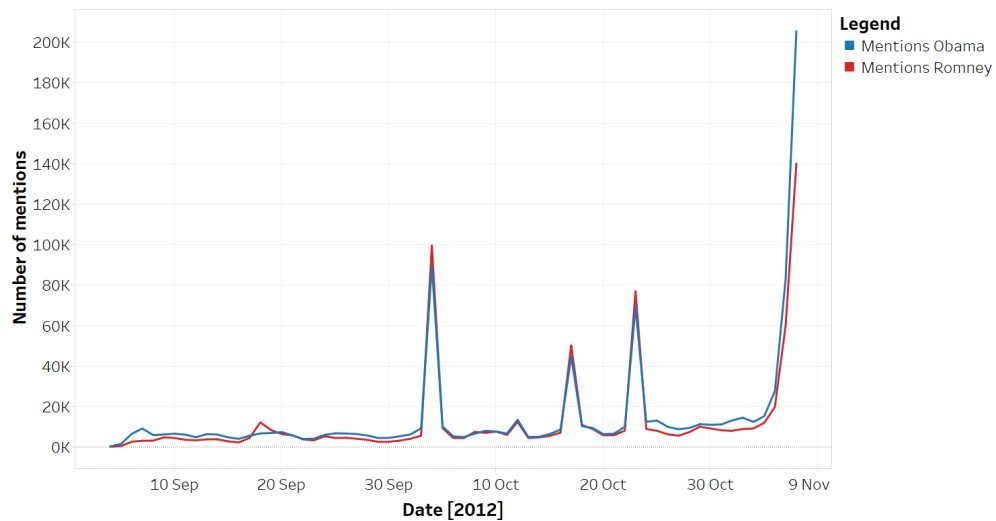


Figure 1-3 – US2012: Number of ‘mentions’ of the Candidates’ surnames in sampled Twitter tweets, retweets and Facebook posts

Figure 1-3 illustrates just one sort of analysis made possible using OSN data, a modern Relational Database Management System (Oracle 12c RDBMS) and a query written in Structured Query Language (SQL). Some 1,718,667 records collected from Facebook and Twitter during the 2012 US Presidential Election campaign (Section 4.2.4.1, p126) have been queried on Candidate’s surname and counted by day using SQL (Appendix 11 listing 1, p479). Daily counts have been plotted using Tableau (Felt, 2016; Tableau, 2017b) data visualisation software (Section 4.5.2, p161). The chart shows that Obama largely leads Romney in mentions throughout the campaign but that Romney is mentioned more often during the three large, and clearly visible, temporal peaks for both candidates coinciding with the three televised US 2012 Presidential Candidate Debates. Aside from these peaks, Romney only ‘beats’ Obama in mentions on one other occasion and mentions of Obama rise rapidly on election day, 6 November 2012. Obama, of course, goes on to win the 2012 US Presidency. Opinion polls at the time, however, frequently suggested that

Romney was level-pegging with Obama (Langer, 2012) or even forging ahead (Gallup, 2012). Can social media data collection, import, storage and query using one reasonably straightforward SQL statement prove a ‘Wisdom of the Crowds’ effect or does further academic enquiry into online information consumption behaviours and by-products, including many works which discount the ability of OSN data to provide predictive political power (Gayo-Avello, 2012b; Iacus, 2014; Murthy, 2015; Vergeer, 2013) suggest better alternative avenues for research?

1.5.3 Academic context

The detailed and extensive literature search and review presented in the following chapter, a) confirms the validity of examining politicised social media data, b) provides many reasons to question the usage of OSN data for electoral prediction, and; c) identifies several gaps in the literature addressed by this research. These strands of academic thought are introduced below and more fully explored in Chapter 2 (p51).

Prominent geographers such as Elwood, Goodchild, & Sui, (2012, p571) were quick to note that the ‘convergence of newly interactive Web-based technologies with growing practices of user-generated content disseminated on the Internet [are] generating a remarkable new form of geographic information.’ This ‘Volunteered Geographic Information’, the authors suggested (p571), ‘represent[s] a paradigmatic shift in how geographic information is created and shared.’ Kuhnian paradigmatic shifts surrounding social media data availability, usage and analytical possibilities have also been reported in the disciplines of Politics (Jenkins, Slomczynski, & Dubrow, 2016), Communications (Van Dijck, 2014) and Technology (Olshannikova, Olsson, Huhtamäki, & Kärkkäinen, 2017). Just under 250 papers (20%) from a research literature corpus comprised of over 1,250 articles (Section 2.2.2, p57) contain the word stem ‘paradigm’. Social media research, in political contexts or otherwise, is a rapidly growing area (Figure 2-2, p57). In the academic literature corpus curated here Facebook is mentioned in 600 papers (48%) and

Twitter in over 650 (52%). While some (e.g., Fuchs, 2017a) have criticised 'Internet Studies' for a lack of theory and others (Boyd & Crawford, 2012; Tufekci, 2014) have identified significant methodological problems inherent in OSN 'Big Data' research, the corpus of academic literature in this area is relatively recent (as are the networks themselves), is still growing and still contains gaps (Section 2.8, p88). Published research has shown that political prediction using OSN data is difficult (Phillips, Dowling, Shaffer, Hodas, & Volkova, 2017) and that a narrow-minded focus on coordinate-geotagged OSN interactions alone is inadvisable (Crampton et al., 2013; Leszczynski & Crampton, 2016).

Repeated attempts to predict political events from social media data in the scientific literature (Franch, 2013; Jain & Kumar, 2017; Tumasjan, Sprenger, Sandner, & Welp, 2010), some claiming success and others reported in the mainstream media (BBC News, 2016), have been comprehensively criticised by Gayo-Avello (2012a, 2012b, 2013) and more recently reviewed by Phillips et al. (2017). Gayo-Avello (2012a, p2) identified 'eight flaws in [...] research regarding electoral prediction' and suggested that the most significant of these is that 'it's not prediction at all! [as all papers reviewed] claim that a prediction could have been made; i.e. they are *post-hoc* analysis and, needless to say, negative results are rare to find.' Academic publishing schedules may be partly to blame, but challenges surrounding Twitter 'vote counting', the application of sentiment analysis to message text and demographic unrepresentativeness, amongst others (Section 7.3, p292), suggest that the challenge of electoral prediction using social media data is considerable, both in terms of reliable 'location inference' and the separation of 'buzz' from voting intent (Han et al., 2014; Jungherr, Schoen, Posegga, & Jürgens, 2017).

Phillips et al. (2017, p10), in their systematic review, have stated that 'Poor findings in the field of election prediction overall suggest that volume of [social media] posts alone, with or without sentiment analysis, is likely a poor method for predicting

election outcomes.’ Jungherr et al. (2016, p1) have also highlighted the intrinsic dangers of using Twitter-sourced OSN data for political prediction, noting that ‘All indicators tested [...] suggest caution in the attempt to infer public opinion or predict election results based on Twitter messages. In all tested metrics, indicators based on Twitter mentions of political parties differed strongly from parties’ results in elections or opinion polls. This leads us to question the power of Twitter to infer levels of political support of political actors. Instead, Twitter appears to promise insights into temporal dynamics of public attention toward politics.’

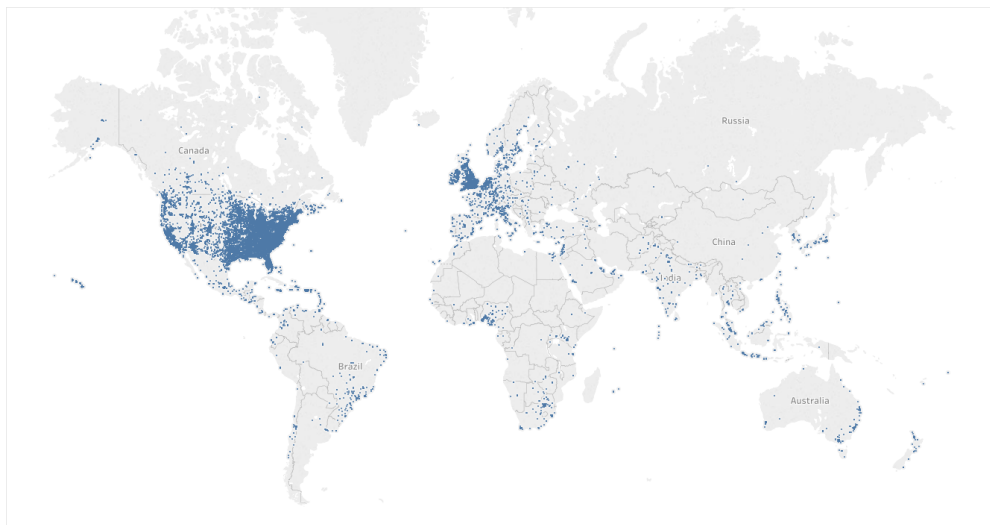


Figure 1-4 – US2012: World map showing sampled geotagged Twitter tweets and retweets (n=168,970)

As Figure 1-4 shows, using a data visualisation package (Tableau in this case) or Geographic Information System (GIS) and the Latitude and Longitude pair embedded within the metadata of some social media interactions, it is extremely straightforward to map coordinate-geotagged social media messages. It is possible to visualise coordinate-geotagged ‘mass sentiment’ (Edwards, Housley, Williams, Sloan, & Williams, 2013) and, by clicking on the map markers in the software or executing a query for a word or phrase, to know *who* is saying *what*, *when* and *where* (Yuan, Cong, Ma, Sun, & Thalmann, 2013). Mapping coordinate-geotagged interactions is far from difficult (Crampton et al., 2013) but, in any given study,

there are typically not enough spatialised records to model OSN sentiment at a local level, the precursor to any accurate electoral prediction (Giuliani, 2018), even if it were believed that text-mining of often terse OSN message text could accurately infer political preferences.

Leszczynski & Crampton (2016, p3), have warned against ‘a fixation on “the geotag” [which] engenders a fetishization of data that is mapped or mappable’ tending towards ‘an implicit commitment to a spatial ontology’ which is ‘divorced from social relations.’ As so few social media interactions are spatially tagged an over-reliance on coordinate-geotagged data is flawed, and a total reliance probably dangerous. OSN data does, however, hold a larger amount of ‘Ambient Geospatial Information’; many Facebook posts or Twitter tweets contain ‘geographic footprints’ in the form of toponymic mentions of place within message text or in associated metadata fields (Stefanidis et al., 2013, p319). A great deal of effort has been expended attempting to geocode or ‘geoparse’ locational references in text (Ajao, Hong, & Liu, 2015; Jurgens, 2013; Poulston, Stevenson, & Bontcheva, 2017; Purves, Clough, Jones, Hall, & Murdock, 2018) but geographers and social scientists have not answered the more fundamental question addressed by this research; given so many geotagged OSN interactions are used as indicators of localised opinion (I. L. Johnson et al., 2016) who makes *most* mention of detectable geographical references in message text and linked/shared URL content, coordinate-geotagging or non-coordinate-geotagging users of social media sites?

Rzeszewski & Beluch (2017, p3), summarising the current state of academic research into ‘Geosocial Media Production’, have suggested that insufficient ‘attention’ has been focused on the ‘subgroup’ of social media users who choose to coordinate-geotag their posts. The research presented here redresses this imbalance by contrasting the posting behaviour of this small but significant ‘subgroup’ of OSN users against that of the non-coordinate-geotagging majority, using data from two case studies, and two social media platforms, separated by a

period of two years. This is unusual for a social media research project (Gayo-Avello, 2012a; Stock, 2018; Tufekci, 2014), most of which rely on only one social media data source (Twitter) sampled during one event. Coordinate-geotaggers are the most *explicitly spatial* users of online social networks but are they also the *most geographically expressive*? By examining differences in expressions of geographicality in message text, interaction metadata and URL link sharing this research determines how 'space' and 'place' are used, referenced and shared in politicised OSN discourse, and most widely by whom.

1.6 Relevance

Social media users in general (Mellon & Prosser, 2017; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011), and coordinate-geotagging users of social media in particular (Sloan & Morgan, 2015), are not thought to be representative of the population at large. However, improved capabilities to target locations on (Figure 1-1, p5), or abstract geographical information from (Figure 6-1, p233), social media networks raise several defining questions for candidates and political parties (Moore, 2016), campaign teams (Buchanan, 2016) and the media (BBC News, 2016). The ability of political candidates, such as Barack Obama, to communicate online with over 29m Facebook Friends and over 20m Twitter Followers is a remarkable innovation in the 'personalisation' of political communication (Enli & Skogerbø, 2013). The potentially 'Orwellian' nature of this new form of interaction (Chamley, Scaglione, & Li, 2013) has been highlighted, somewhat disconcertingly, by an anonymous member of Obama's campaign team, who reportedly stated (McGregor, 2011) that '[online] we want to serve you with stuff that you are going to like' and that, in doing so, 'the information that is interesting to us is behavioural.' Subsequent events during the 2016 US Presidential Election and the 2016 UK European Union Membership Referendum, characterised by attempts to deliberately manipulate democratic outcomes through geo-psychographic targeting (Figure 1-5, p31) of social media users (Albright, 2017; Cadwalladr, 2017; Persily,

2017), suggests that candidates and/or political parties must increasingly endorse such strategies; why else would they spend money on this form of campaigning?

The screenshot displays the CA Political website. The header includes the company name 'CA Political' and navigation links for 'FLIGHTS' and 'NEWS'. A search bar is located in the top right corner. The main content area is divided into two columns. The left column features a 'CA Political' section with an introductory paragraph about the company's data-driven marketing solutions, followed by sections on 'Achievements', 'Efficiency and Transparency', 'Multidisciplinary Expertise', and 'Services'. The 'Services' section is further divided into four sub-sections: 'Market Research', 'Data Integration', 'Audience Segmentation', and 'Targeted Advertising'. The right column contains a 'Recent Posts' section with a list of recent news items, including awards, dataset announcements, election results, and committee publications.

CA Political

FLIGHTS / NEWS

CA Political

CA political has been provided the stronger relationship between data and marketing. Basically, it combines the predictive data analytics, behavioural sciences, and innovative ad tech into one award winning approach. With the use of CA political, you can engage the customers more effectively and efficiently. Recently, CA-political.com online service has received awards like Global Periodical Publisher and Employee Benefits Provider. CA-political is a data-driven marketing solution and it works together with the companies like SCL Group Limited, Cambridge Analytica Limited, SCL Elections Limited, SCL political Limited, and SCL Social Limited in order to deliver best solutions to its clients.

Essentially, CA political is providing an advantage of improving the brand's marketing effectiveness by influencing the customer behaviour with the use of behavioural sciences, predictive analytics, data driven solutions, and other technological benefits.

Achievements

CA-political has up to 5000 data points on over 230 million individual American consumers. By embedding with these data instruments, you can add to your own customer data and gain the benefit of establishing custom target audiences which lets to engage and motivate the individuals.

Efficiency and Transparency

CA political doesn't hide any margins on CPM rates. That means, the data driven insights allows to focus the investment on your best customers. It supplies real-time reporting dashboards which lets to view the true ROI of your campaigns.

Multidisciplinary Expertise

It has included a global team of data scientists, PhD researchers, psychologists, and digital marketing experts who always strive to deliver a holistic approach to exceeding the marketing goals.

CA political Data-driven insights provide you different tools that needed to cut through crowded advertising landscape and speak directly to the customers as well. Based on each unique behavioural profile, it optimizes the work specifically included visuals, copy, and positioning.

Services

CA political Data-driven services really helpful for you to better understand your audiences. You can run your marketing with end-to-end campaign by combining the insights of behavioural psychology with the precision of data analytics and individually addressable technology.

- **Market Research**

In order to get a complete view of customer behaviour, competition, and trends, CA-political collaborates the behavioural psychology with the statistically robust methodology.

Once your existing customer knowledge base has been assessed, it will conduct a custom research on projects according to your particular needs. It is also providing an in-depth picture of your audiences based on qualitative and quantitative techniques by embedding the unique behavioural science methodology. Available services under market research subsumed knowledge gap analysis, audience behavioural insights, and research design and execution.

- **Data Integration**

By making centralization on your first party data and enriching it with data from third party resources, CA-political implements a rich and holistic view of your customers' behaviour. To produce data deeper and richer insights, it collects the data from public sources and reputable data providers. The data integration service helps to find and influence the customers quickly and efficiently. Data Integration tools are data cleaning, data management consultation, and CRM integration.

- **Audience Segmentation**

For creating targetable groups of similar individuals, CA political provides audience segmentation from the collected profiles of customers. By grouping behavioural similarities and identifying population segments, it helps to reach your most persuadable customers and campaigns.

- **Targeted Advertising**

It always strives to promote the products and services to reach the customers in a unique way to which they are most likely to respond. As CA-political uses data science technology, behavioural science, and the latest advancements in advertising technology, it has been defined who should be targeted, how the messages should be constructed, and make sure the right message gets to the right person as well.

Recent Posts

- CA Receives Top Honor in the 2017 ARF David Ogilvy Awards
- CA responds to announcement that GSR dataset potentially contained 87 million records
- CA, the Data Gurus Who Anticipated the Election Result
- Digital, Culture, Media and Sport Committee publishes documents
- Data-driven services
- Cambridge Analytica responds to committee hearing
- Message from acting CEO Dr. Alexander Tayler
- Cambridge Analytica responds to Facebook announcement
- Google Flights Helps you to get Cheap Airfare This ThanksGiving 2018
- Cambridge Analytica responds to use of entrapment and mischaracterization by Channel 4 News

Figure 1-5 – CA (Cambridge Analytica) Political website, outlining company capabilities (Cambridge Analytica, 2018)

Early analysis of the 2012 US Presidential Election data set collected as part of this research suggested that many outputs could be produced, some perhaps reflecting contemporaneous online ‘politicking’ (Panagopoulos, Gueorguieva, Slotnick, Gulati, & Williams, 2009) as well as potentially partisan mainstream media coverage of events (Searles, Smith, & Sui, 2018). Figure 1-3 (p25), for example, has illustrated the timeline of counts of ‘mentions’ of the 2012 US Presidential Candidates’ surnames, and shows some temporal contradiction with several of the opinion polls published at the time. However, these daily counts do not offer anything like a predictive model of turnout or political outcomes or allow more detailed mapping of political sentiment or voting intention for local areas. Instead, the data tends to reveal the spatiotemporal ‘buzz’ (Roy & Zeng, 2015) surrounding events. Figure 1-6 shows the percentage of population weighted mentions of State abbreviations (OH=Ohio, VT=Vermont, etc.) from OSN posts in the US2012 data set.

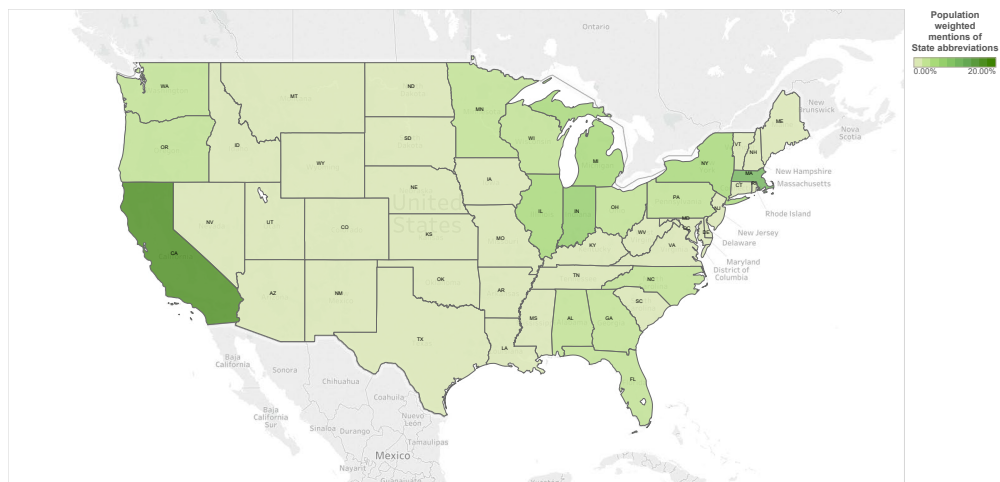


Figure 1-6 – US2012: Percentage population weighted mentions of State abbreviations in interaction message text

The map has been produced, not by counting the small number of explicitly coordinate-geotagged records in the data set, but by looping over a list of US State abbreviations and full State names stored in an Oracle 12c database table (Section 4.3.1.3, p145) and using SQL to count occurrences of these terms in message text (‘OH’, Ohio, is illustrated in Appendix 11 listing 2, p479). The number of text

matches for each State has been saved, and weighted by 2012 State level population estimates (U.S. Census Bureau, 2017). It is apparent that many, but not all, of the State abbreviations most mentioned in the OSN research data corpus coincide with those defined by contemporary media reports identifying key 'Battleground States' in the 2012 US Presidential Election contest (BBC News, 2012; Figure 1-7, p61).

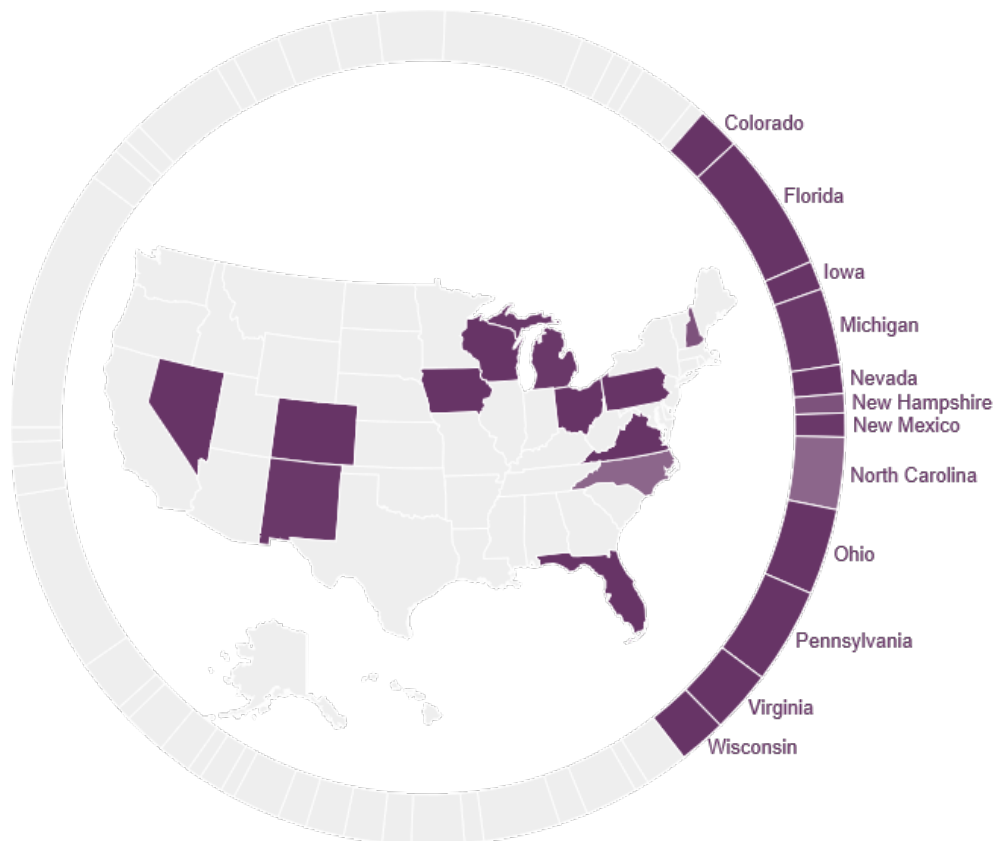


Figure 1-7 – US2012: Battleground States (BBC News, 2012)

As Figure 1-6 shows, identifying the online social media 'buzz' surrounding specific areas should be of interest to anyone who votes in geographically bound democratic elections. Candidates and political parties are now making increasing efforts to accurately target their campaign messages towards voters in the often small number of 'swing' states or constituencies which typically determine the outcome of electoral contests in many modern democracies (Henneberg & O'Shaughnessy, 2009; Lilleker et al., 2015; Moore, 2016). Gaining more

understanding of how ‘place’ and ‘space’ are used, differentially referenced and shared by social media users in politicised online social media discourse is, therefore, a highly relevant exercise.

1.7 Research hypothesis aim and objectives

Building upon the *Localness Assumption* identified by I. L. Johnson et al. (2016, p515, authors' italics), in which ‘*a unit of social media VGI [is implicitly assumed almost] always [to represent] the perspective or experience of a person who is local to the region of the corresponding geotag*’, this research is framed around an equally simple hypothesis.

The *Geographicality Assumption* tested here asserts that *coordinate-geotagging users are the most geographically expressive of all OSN users*.

- As I. L. Johnson et al. (2016) have noted, geographical ‘localness’ has been widely assumed in most research regarding coordinate-geotagged OSN data. This supposition, however, rests upon another equally implicit and, so far, untested assumption; that geotagging social media users make at least equal, and possibly especially frequent, mentions of place in message text when compared to their non-coordinate-geotagging counterparts.
- We know that only ~1-2% of Twitter interactions are typically coordinate-geotagged by their creators (Leetaru et al., 2013) and we know that toponymic mentions of place are widespread in Twitter and other OSN interactions (Gelernter & Mushegian, 2011; Pavalanathan & Eisenstein, 2015; Stefanidis, Crooks, et al., 2013).
- We do not know whether coordinate-geotagging or non-coordinate-geotagging users of Twitter or other OSNs exhibit *differential* affinity for place, in terms either of a) more mentions of place in message text or, b) more mentions of place in linked/shared content from URLs found alongside message text in OSN metadata.

This research tests the assumption that the most ‘spatialised’ class of coordinate-geotagging social media users are also the most geographically expressive users of these sites. Doing so, as previous sections have outlined, will indicate whether social media interactions sourced from this minority group can reliably be considered representative of all OSN interactions when using place to track the diffusion of political opinion, or potentially geo-targeted (mis)information, online.

1.7.1 Aim

This research tests the *Geographicality Assumption* through detailed analysis of coordinate and non-coordinate-geotagging users’ toponymic mentions of place in OSN message text and linked/shared URL content. Using a range of techniques including Information Extraction (IE) and Named Entity Recognition (NER) from Natural Language Processing (NLP), alongside more traditional gazetteer-based geoparsing approaches, the work answers three research questions:

1. How can baseline ‘geographicality’ be assessed and categorised in OSN data?
2. Does NLP-detectable ‘geographicality’ in message text increase in line with ‘spatiality’?
3. Does NLP-detectable ‘geographicality’ in linked/shared 3rd party content increase in line with ‘spatiality’?

Large volumes of OSN interaction data have been analysed to answer these questions. These new forms of online ‘geographicality’ echo the *géographicit  * of   ric Dardel (Dardel, 1952), in which the human fascination with the lived environment is evidenced through mentions of place.

1.7.2 Objectives

Potential Geographic Information (PGI) exists in OSN interaction message text and in some associated metadata fields. It is not clear to what extent the sharing of

spatial (coordinate) or toponymic (place) references matters; either to those reading, reposting, retweeting or otherwise consuming content on social networks, or to those analysing such content for an increasingly wide variety of purposes.

The objectives of this research are to:

- Sample, record and store sufficient volumes of OSN data to provide good case study material to address the research questions.
- Select and use appropriate 'Big Data' technologies to detect toponymic mentions of place in text and shared URLs from the OSN case study data.
- Extract information and mine social media interactions to determine how geographical expressions are made by all types of OSN users.
- Compare results of Information Extraction (IE) operations across political events, by system and non-/coordinate-geotagging user classes.

The research examines ~1.7m OSN interactions recorded during the two-month run-up to the 2012 US Presidential Election and ~6.5m recorded during the twelve-month run-up to the 2014 Scottish Independence Referendum. The political context within which the research is framed is now particularly apposite as both succeeding scholars, and regulators, embark on even larger-scale investigations into the probable (mis)use of geo-behaviourally targeted social media advertising during the 2016 US Presidential Election and the 2016 UK European Union Membership Referendum.

Results from this research are presented in Chapter 5 (p186) with a discussion and additional findings in Chapter 6 (p227). The thesis concludes, in Chapter 7 (p286), by asserting several contributions to knowledge, critically reflecting on the work conducted and setting out suggestions for future and further research.

1,250 article titles stored in Mendeley Desktop (Appendix 2, p411). McNaught & Lam (2010, p631) find that '[word clouds seem] to be particularly useful for studies that involve qualitative/thematic analyses of written or transcribed spoken text' specifically offering a 'tool for preliminary analysis, quickly highlighting main differences and possible points of interest [which can] provide an additional support for other analytic tools.' Key words from the (lower-cased) article titles shown in Figure 1-8 (p37) include 'social', 'political', 'media', 'twitter' and 'data'. The words 'geography', 'geographic' and 'spatiotemporal' appear in smaller type in the word cloud indicating lower word frequency counts for these terms in article titles.

Table 1-1 – Top 20 journal titles by number of stored articles

Position	Journal	References
1	<i>Computers in Human Behavior</i>	35
2	<i>New Media & Society</i>	30
3	<i>Social Science Computer Review</i>	23
4	<i>Information, Communication & Society</i>	19
5	<i>International Journal of Geographical Information Science</i>	17
6	<i>arXiv</i>	11
7	<i>Cartography and Geographic Information Science</i>	10
8	<i>Expert Systems with Applications</i>	9
9	<i>Journal of Broadcasting & Electronic Media</i>	9
10	<i>Journal of Communication</i>	9
11	<i>American Behavioral Scientist</i>	8
12	<i>Big Data & Society</i>	8
13	<i>Journal of Computer-Mediated Communication</i>	8
14	<i>Journal of Information Technology & Politics</i>	8
15	<i>Political Communication</i>	8
16	<i>Communications of the ACM</i>	7
17	<i>Electoral Studies</i>	7
18	<i>European Journal of Communication</i>	7
19	<i>First Monday</i>	7
20	<i>Social Networks</i>	7

While this thesis is concerned primarily with geography and geographicality these themes clearly exist within a much wider academic context. Table 1-1 shows the top 20 journal titles, by number of articles stored in Mendeley, from the research

literature corpus. An expressly geographical journal (*IJGIS*) appears only 5th by total number of references collected when references are summed by publication title. Leading sources of references include the journals *Computers in Human Behavior* (established 1985), *New Media & Society* (1999), *Social Science Computer Review* (1983) and *Information, Communication & Society* (1998). The study of geographicality in OSN posts must be contextualised within a wider scholarly framework considering political, communications and socio-technological aspects of social media usage. These themes are more fully explored in Chapter 2 (p51) which considers literature and context.

1.8.2 Case study-based data collection

Given massive growth in OSN usage, improved availability in the form of access to public posts and the fact that so much political activity now takes place over the Internet (Hong & Nadler, 2012; Quintelier & Theocharis, 2012; Shehata & Strömbäck, 2018), this research uses political case studies to assess the role and significance of geography, coordinate and non-coordinate-geotagging users and their interactions in these newly emerging contexts. As the research literature corpus exhibits considerable cross-disciplinary overlap between subjects the selection of political case study material is especially apt. Clark & Jones (2013, p305) have confirmed ‘the scope and potential for spatialising new institutionalist studies, by demonstrating how fluidities of political behaviours predicated by post-structural accounts of place and space come to be ‘fixed’ within certain ‘sticky’ institutional places. Consequently, we argue that a spatialised new institutionalism offers promising conceptual and methodological possibilities for developing research collaborations between political geography and political science on the placing and spacing of political behaviours.’ Chapter 3 (p94) and Chapter 4 (p118) detail the methodology adopted and methods used to capture, store and analyse >8 million OSN records sourced from Facebook and Twitter during the two case study events to address ‘the great implications of spatialization’ (Ethington &

McDaniel, 2007) for ‘political behaviour research’ (Clark & Jones, 2013) addressed in this investigation.

1.8.3 Data storage, analysis, and interpretation

Over 46GB (2^{10} bytes) of raw OSN data in JavaScript Object Notation (JSON) and Comma Separated Values (CSV) formats have been stored and queried using both conventional (SQL-based) RDBMSs, Microsoft SQL Server 2012 R2 (2013) and Oracle 12c (2014c), and a number of Not-only-SQL (NoSQL) alternatives, including the MarkLogic (2014) document data store and Apache Software Foundation's (2014) Drill running within the MapR (2014) ‘Hadoop ecosystem’ (Cutting, 2013). These systems offer competing benefits when importing, storing and querying JSON and CSV data files, as well as differences in access and infrastructural integration. Section 4.3, p135) describes the rationale behind the choice of database management system, Oracle 12c RDBMS, used most widely in this research programme following several comparative evaluation exercises. As the research has progressed, new software releases have provided functionality missing in earlier versions; the effective use of up-to-date technology is, therefore, a defining feature of the research, linking theoretical and practical aspects to create the relevant research outputs detailed in Chapter 5 (p186) and Chapter 6 (p227).

1.9 Originality and contribution to knowledge

1.9.1 Originality

This research fuses long-established approaches with very recent technological developments in computational systems, designed to operate at extremely large scale, to examine two unique case study data sets. The principal methods include:

- Data-mining using the latest generation of RDBMS software and SQL;
- Information Extraction (IE) using NLP/geoparsing software;
- Spatiotemporal visualisation using specialist software, and;

- Statistical analysis using R (The R Foundation, 2018).

While some of the approaches are well-established in theoretical terms – NLP, for example, stems in part from Turing's (1950) seminal paper *Computing Machinery and Intelligence* – only recently have systems been scaled to work effectively with very large data sets (Tablan, Roberts, Cunningham, & Bontcheva, 2012). The combinatorial use of these approaches itself contributes to knowledge. As the extensive code bases in Appendix 10 (p451), Appendix 11 (p479) and Appendix 12 (p493) demonstrate, and Lin & Ryaboy (2013) have argued, a great deal of work is required to successfully integrate different data sources and software systems in a Big Data project. In addition, trialling and participation in beta programmes, has resulted in several enhancements to existing software packages which are discussed in more detail, below, in Section 1.10 (p43).

1.9.2 Contribution to knowledge

The practice of coordinate-geotagging messages on OSNs is a relatively recent phenomenon, providing massive amounts of geographically-referenceable data produced (primarily) by human 'social sensors' (Rosi et al., 2011). Users of OSNs are a self-selecting subset of the population and the percentage coordinate-geotagging their posts appears to be a low (Leetaru et al., 2013) if somewhat variable one (Croitoru et al., 2013). The coordinate-geotagging rate observed here across US2012 and SCOT2014 data sets collected in several different 'Streams' (Appendix 7, p432) falls only in the range 0.86-1.45%. Consequently, OSN interactions have been data and text-mined to search for locational references in text and linked/shared URL content. In doing so, the work makes use of a 'software stack' which shares some similarities with the software capabilities detailed in Croitoru et al.'s (2013) 'Geosocial gauge' and moves 'beyond the geotag', as Crampton et al. (2013, p138) have advocated, by paying 'closer attention to the diversity of social and spatial processes, such as social networks and multi-scalar

events, at work in the production, dissemination, and consumption of [geoweb] content’.

In addition to the original methodological and technological contributions outlined in Section 1.9.1 (p40) key, and original, contributions to knowledge from this case study research include the findings that:

- Coordinate-geotagging users of Twitter and Facebook OSNs, a) make fewer references to place in their message text (Section 5.2.2, p190), b) link to external content making fewer mentions of place in text (Section 5.2.3, p205), and; c) make fewer links to external third-party content (Section 5.2.3, p205) than corresponding non-coordinate-geotagging users of these platforms.
- A disproportionate number of US2012 coordinate-geotagging users (Section 6.4.4.1, p265) are found in areas with a higher than expected percentage of ‘non-institutionalised’ population (halls of residence, nursing accommodation etc.); geodemographic profiling of SCOT2014 coordinate-geotaggers (Section 6.4.4.2, p272) also suggests these users are found in areas with relatively youthful age profiles when compared to UK averages.
- The median and average straight-line geo-retweet distances (Section 6.4.2, p251) across both electoral events are, respectively, 2.72km and 17.22km. These distances are significantly lower, in this politicised data, than the 1,698km median and 955km average distances reported by van Liere, (2010) using a much smaller, non-politicised, random sample of Twitter data (n=13,399 retweets). Geo-retweet distances recorded here are also significantly lower than the ‘749 statute miles’ reported by Leetaru et al. (2013) analysing a much larger 10% sample of all Twitter tweets and retweets made between 23 October and 30 November 2012 (n=1,535,929,521), a time period which overlapped with one of the data collection exercises undertaken in this study. The lower values reported

here may reflect a more ‘local pattern’ (van Liere, 2010) of opinion re-dispersal in politicised social media discourse and is an area suggested for future research (Section 7.5, p299). Further details and maps showing dyadic geo-tweet/retweet dispersal are presented in Section 4.2.4.1 (p126);

Chapter 5 (p186) presents the NLP/geoparsing results of this research which shows, through the application of a number of Big Data processing techniques, just how relevant space and place are to users of Facebook and Twitter interacting online during the 2012 US Presidential Election and 2014 Scottish Independence Referendum campaigns. These results are discussed, and additional findings detailed, in Chapter 6 (p227) of this thesis. From these observations it is possible to conclude (Chapter 7, p286) that tracking the spread of (mis)information using coordinate-geotagged interactions alone is likely to prove inaccurate. The comparative lack of coordinate-geotagged records, and the unrepresentative geographicality characteristics of coordinate-geotagging users, also makes any attempt at political prediction based around analysis of the sentiment of geotagging users’ messages, or the inclusion of given URL links shared alongside those messages, difficult to recommend. For governments, regulators, researchers and citizens this conclusion has profound implications. It is not possible to accurately track the geographical diffusion of political opinion or potentially nefarious geo-targeted social media advertisements using coordinate-geotagged interactions alone. These implications are more fully discussed in Section 6.2 (p229) while policy recommendations, which could be adopted to ameliorate this problem, are outlined in Section 6.3 (p238).

1.10 Impact and engagement

Research Councils UK define impact as ‘the demonstrable contribution that excellent research makes to society and the economy’ (RCUK, 2016) further identifying:

- **Academic impact** is the demonstrable contribution that excellent social and economic research makes to scientific advances, across and within disciplines, including significant advances in understanding, method, theory and application.
- **Economic and societal impact** is the demonstrable contribution that excellent social and economic research makes to society and the economy, of benefit to individuals, organisations and nations.

This research project has demonstrated impact under both RCUK headings, more fully described below.

1.10.1 Academic impact

The substantive results of this research are reported in Chapter 5 (p186) with a discussion and additional findings presented in Chapter 6 (p227). Key contributions arising from this investigation into differing patterns of geographical expression amongst coordinate and non-coordinate-geotagging users of social media platforms during electoral campaigns have been outlined above, the implications of which, including policy recommendations, are more fully detailed later in Section 6.2 (p229) and Section 6.3 (p238). It now appears certain that considerable scholarly and regulatory attention will focus on improving our understanding of geo-behaviourally targeted online social media advertising (mis)use during electoral campaigns, following the series of damaging revelations and scandals which have recently come to light, described earlier in the introductory pages of this chapter (p1). The methods, results, recommendations and suggestions for further work detailed in this thesis should provide successive researchers with much useful information on which to base their future investigations.

In addition, the project's use of Datasift's JSON data has already been of value to developers at the University of Sheffield (Bontcheva & Greenwood, personal communication, 2014) and the University of Cardiff (Morgan, personal

communication, 2015) who have used sample files to write parsers for the General Architecture for Text Engineering (GATE) and the Collaborative Online Social Media Observatory (COSMOS) software packages, respectively. The project's use of a clustered MapR Hadoop ecosystem incorporating Apache Software Foundation's Drill package has, likewise, already been of value to researchers at the University of Keele (Lam, personal communication, 2016) in verifying the successful installation of the University's own MapR system.

1.10.2 Economic and societal impact

Ministers, parliamentarians, regulators, scholars and many other commentators have expressed deep concern over the breaking revelations highlighting the misuse of geo-behaviourally targeted political advertising, and/or misinformation, disseminated online via social media channels. In the UK, at least – with tightly enforced regulations regarding 'above the line' political campaign spending and message attribution on television, radio, newspaper and outdoor advertising, as well as in door-dropped pamphlets – unregulated campaign (or external) spending on social media advertising is now under intense scrutiny (The Electoral Commission, 2018a). Even in the US, with its generally more relaxed regulations surrounding campaign spending limits or political messaging, Senator Mark Warner, Vice Chair of the Senate Intelligence Committee, has warned that 'the era of the Wild West in social media is coming to an end' (Charter, 2018).

Writing in the *Financial Times*, Gapper (2018) has suggested that 'Mark Zuckerberg cannot control his own creation', going on to point out that despite tightening data controls, especially surrounding access to the 'social graph' of user inter-relationships (Hogan, 2018), 'things cannot be fixed because they are beyond Mr Zuckerberg's control, lost in myriad encounters among Facebook's 2bn users. The technical term is emergence, the powerful and unpredictable outcome of millions of users interacting freely with others. Anything from joke videos to fake news can spread like a virus, changing how people feel and act.' Tracking the geographical

and social spread of these virus-like outbreaks is not straightforward. This thesis contributes to academic and societal understanding of these issues by demonstrating the difficulty involved in accurately geographically tracking the downstream diffusion of such material and providing suggestions for the future. It is currently much easier for political advertisers, whether *bona fide* or not, to set up and run a geo-targeted campaign using Facebook (Figure 1-1, p5), or several other major websites (Section 1.6, p30), than it is for government, regulators, researchers or citizens to track the geographical consumption and sharing patterns of such advertising. This situation arises largely from platform operators' privacy policies, some perfectly well-intentioned to safeguard user privacy, but also from a lack of 'transparency' in available (meta)data and reporting (The Electoral Commission, 2018a).

Early indications (e.g., BBC News, 2018d) suggest that the political will to regulate some of the world's largest 'tech giants' exists, but regulating global corporations will not be easy. The policy recommendations incorporated in this thesis (Section 6.3, p238) offer one possible solution to the problem, by encoding lower-resolution coordinates alongside all social media interactions. However, this technical response is only one strand of a much larger debate in which societies around the world must engage with what *The Atlantic's* correspondent and author, Franklin Foer (2017), in an interview with *The Guardian* (Taylor, 2018), has suggested is 'the real problem [...that we] have two or three companies that are the masters of the global public sphere.' Tackling the 'real problem' may take some time but, hopefully, the results and suggestions contained within this thesis will assist.

Meanwhile, on a more prosaic and technical level, the project's use of Datasift's JSON data, and particularly the 'ingestion' of 2014 Scottish Independence Referendum data with its larger number of longer (and/or multi-byte encoded) JSON key names, has already been of value to Oracle Corporation (Pitts & Venzl,

personal communication, 2015), whose developers are improving JSON document data storage in forthcoming versions of Oracle's RDBMS software.

1.10.3 Engagement

During the research, findings have been presented at:

- RGS-IBG Annual International Conference 2013; Big, Open Data and the Practice of GIScience (Tear, 2013)
- The 14th International Conference on Computational Science and Its Applications (ICCSA 2014) (Tear, 2014)
- The 19th AGILE International Conference on Geographic Information Science; Workshop "GIS with NoSQL" (Tear, 2016)
- The 25th GIS Research UK Conference presenting *Wading through the swamp: filter systems for geospatial data science* (Tear & Healey, 2017)

The paper presented at ICCSA 2014 was later published in Springer's *Lecture Notes in Computer Science* series.

1.11 Limitations of the research

Eisenhardt (1989, p532) states that a case study 'approach is especially appropriate in new topic areas', particularly as the 'the process [...] is highly iterative and tightly linked to data.' The technological infrastructure required to satisfy the aim and objectives of this study (Chapter 4, p118) comprises an important part of the research method, in turn framed within an exploratory and largely quantitative empirical research methodology (Chapter 3, p94) that is intimately bound with the politicised social media data under investigation, much of it discursive and hence inherently qualitative in nature.

The two case study data sets are large (>46GB raw), rich in some attributes (e.g., message text, date/time stamps) but poorly populated in others (e.g., exact

geographical locations, demographics). A wide range of computerised software systems have been used to collect, store, query and mine these data, producing the largely aggregated results presented in Chapter 5 (p186) and Chapter 6 (p227). This type of approach, sometimes termed 'Computational Social Science' (Lazer, Brewer, Christakis, Fowler, & King, 2009), has been criticised from a methodological standpoint by some scholars (Fuchs, 2017a; Tufekci, 2014). Arguments over the validity and extensibility of Big Data research, sometimes compounded by a lack of accompanying theory, in many ways reflect academic geography's earlier struggles with its own 'quantitative revolution' (I. Burton, 1963; Cresswell, 2014).

Wyly (2014, p35) has written an excellent and extremely thought-provoking article on this topic, eloquently explaining how 'the torrential acceleration of data flows [and] the circulation and partially autonomous replication of mobile data streams [...] seem to be reconstituting some of the most important [methodological] debates of geography from the 20th century.' The new quantitative revolution which Wyly (p35) describes has 'only recently come into pragmatist existence with the new abilities to observe and quantify online human attention in real time.' As this study examines data of this type – and ethical good practice and University recommendations (Section 3.4, p111; Appendix 4, p419) largely preclude the reporting of most, except the most politically prominent, of individual's social media message text or metadata – these factors are all limitations in this work. Consequently, an 'awareness' of the 'representativeness, validity and other methodological pitfalls' in social media Big Data research, which Tufekci (2014, p524) identifies, is acknowledged here. In deference to Tufekci's suggestion that a discussion of these issues should go 'beyond soliciting "limitations" sections' in research work the reader is invited to turn to Chapter 3 (p94) of this thesis, where epistemological, methodological and ethical limitations are discussed in much greater detail.

1.12 Thesis structure

This thesis is comprised of seven chapters:

1. **Introduction** – The current chapter has introduced the study, set out the hypothesis, aim and objectives of the research, the research context, contribution to knowledge and impact. The background to the research has been described together with an outline of the research process. Limitations of the study have also been outlined and a chapter plan given.
2. **Literature and Context** – Chapter 2 (p51) contextualises the study providing a theoretical framework for the research based on cross-disciplinary readings in Politics, Communications, Geography and Computer Science. The use of OSN data in other application domains is discussed, and the gaps in existing knowledge are identified.
3. **Research Design** – Chapter 3 (p94) outlines the exploratory case study methodology employed to harvest significant volumes of OSN interactions generated during two major political events. Epistemological and ethical considerations are discussed. Methodological difficulties inherent in ‘Big Data’ social media analysis are introduced.
4. **Research Methods** – Chapter 4 (p118) describes file data outputs and the Extract, Transform and Load (ETL) procedures developed to effectively store data in a number of conventional (SQL) and unstructured (NoSQL) database systems, together with the range of software packages and computer systems used to augment, mine and visualise these data.
5. **NLP/Geoparsing Results** – Chapter 5 (p186) presents results addressing three research questions using the data collected and the research methods outlined in the preceding chapter. A wide range of statistics, maps and tables are presented illustrating the role of geography in OSN interactions based on metadata, NLP, geoparsing and data-mining operations.

6. **Discussion and Additional Findings** – Chapter 6 (p227) discusses results presented in the preceding chapter and details several additional findings made possible by the flexible set of research methods and exploratory research methodology described in earlier chapters. UK and US Census data are ‘fused’ to OSN interactions and other analyses reported.
7. **Conclusion** – Chapter 7 (p286) concludes the thesis by drawing together all of the themes introduced in earlier chapters, confirming the validity of the approach and laying claim to an original contribution to knowledge. Ideas for further research are identified, and complementary areas of cross-disciplinary research are outlined.

Data licences, ethics correspondence and code listings are largely confined to Appendices (pp407-491). As the research is highly technical in nature, complete programmes, virtual machines (VMs), and archival database backups are too voluminous (>2TB) for inclusion. These digital artefacts may be obtained from the author upon request.

2 LITERATURE AND CONTEXT

2.1 Introduction

The review presented in this chapter is based upon an extensive search for literature relating to the usage of Online Social Networks, particularly in political contexts. The review process is complemented by several computerised bibliometric, analytical and text-mining methods designed to synthesise results from published scholarly works and related, largely news-based, material. Over 1,250 articles, and 1.22GB of associated Adobe Portable Document Format (PDF) files, have been saved to Mendeley's (2016) desktop bibliographic reference management software as a result of this literature search. The academic advisory body JISC (2012, p3) has noted that 'Vast amounts of new information and data are generated everyday through economic, academic and social activities', suggesting that techniques 'such as text and data mining and analytics are required to exploit this potential.' JISC go on to state that:

In systematic reviews of literature, text mining is used to automatically identify literature that should be reviewed by researchers wishing to establish the current state of knowledge in a particular field. The mining takes place across both traditional peer-reviewed academic journals and grey literature such as technical reports, policy documents and pre-prints. Researchers can use the information extracted to identify relevant documents from a much wider source pool, including from other disciplines and non-traditional sources.

(JISC, 2012, p15)

Text and data mining techniques are applied here to analyse the curated research literature corpus. Section 2.2, below, details the background to this new type of literature review process while Section 2.2.1 (p54) details the methods used to

interrogate the substantial body of literature collated during this research. The results of text and data-mining operations against the literature corpus are presented in Section 2.2.2 (p57); outputs from which have been used to systematically identify and rank key terms from text stored in article PDFs, helping to frame a thematic overview of the wide-ranging and cross-disciplinary research literature. Section 2.3 (p61) provides a contextual synopsis of these results before key themes in the political (Section 2.4, p64), communications (Section 2.5, p72), geographical (Section 2.6, p77) and technical (Section 2.7, p83) literatures are elucidated. Finally, Section 2.8 (p88) identifies gaps in knowledge and several influential papers, fundamental to the design and conduct of this research project.

2.2 Text and data-mining in the literature review process

As Webster & Watson (2002, pXIII) make clear ‘A review of prior, relevant literature is an essential feature of any academic project. An effective review creates a firm foundation for advancing knowledge. It facilitates theory development, closes areas where a plethora of research exists, and uncovers areas where research is needed.’ In an era in which ever-increasing numbers of journals and journal articles examine emergent phenomena, such as Online Social Networks, new techniques are required to search for, select and synthesise academic material (JISC, 2012). Ware & Mabe (2015, p6) have estimated that around 2.5 million peer-reviewed scientific, technical and medical articles were published in the English language in 2014. As the rate of production of scholarly output increases and the ease with which electronic documents can be stored and searched improves, nascent text-mining and knowledge discovery technologies – used elsewhere in this research to interrogate social media data (Chapter 5, p186) – now offer high degrees of utility when analysing large corpora of published academic work.

Fully ‘systematic’ literature reviews, particularly popular in the medical research community and increasingly being applied by ‘early adopters’ in the social sciences and humanities (JISC, 2012, p4), rely upon searches for literature, based on key

terms, executed against multiple online academic repositories or databases, e.g., JSTOR, Web of Science, PubMed etc. The approach adopted here, more fully detailed in Section 2.2.1 below, may best be described as ‘semi-systematic’; over 1,250 articles have been selected for inclusion in the research literature corpus over a period of more than 7 years, based upon searches executed on popular online repositories and publishers’ websites as well as alerts set up on, and emails received from, Google Scholar and learned societies such as the Political Studies Association. Articles selected for inclusion in the research literature corpus are stored in Mendeley Desktop bibliographic management software and have been read either in full or sectionally, by searching for key terms and the paragraphs or sections that contain them. The literature review which follows (Sections 2.4 to 2.7, pp64-88) therefore mixes a conventional scholarly approach to the task, with a synopsis of key themes given in Section 2.3 (p61), alongside several computerised methods outlined in the following paragraph and more fully described in Sections 2.2.1 and 2.2.2, below.

Usai, Pironti, Mital, & Aouina Mejri (2018) have suggested that a systematic review of literature may be conducted ‘by applying “text mining at the term level, in which knowledge discovery takes place on a more focused collection of words and phrases that are extracted from and label each document” (Feldman et al., 1998, p1). This approach involves extracting labels which correspond to keywords, which consequently represent the main topic of an article.’ Term labelling may be achieved *manually* (e.g., by the researcher marking up article text identifying key terms based upon their own domain expertise) for training in machine learning applications or *automatically*, as here, through the use of algorithmic processing, e.g., the creation of Term Frequency – Inverse Document Frequency (TF-IDF) matrices (Section 2.2.2.2, p58). Either approach is designed to ‘find nuggets in mountains of textual data’ (Dörre, Gerstl, & Seiffert, 1999), helping to identify key themes running through substantial bodies of literature and to organise the review process accordingly.

The following sections of this chapter describe the methods used to search for literature (Section 2.2.1) and present results of data and text-mining analysis (Section 2.2.2, p57). Through this work 136 key terms, each mentioned over 4,000 times within the research literature corpus, have been identified programmatically. Using acquired domain knowledge (Alexander, 1992), based upon a reading of these articles, identified terms have been assigned (Section 2.2.2.3, p59) to four disciplinary categories; political, communications, geographical and technical. The literature relating to Online Social Network usage, as it applies to each category, is discussed below, starting in Section 2.4 (p64). First, the methods used to search for and store literature are set out.

2.2.1 Methods

This section details the methods used in the literature review, considering the scope of the review, sources of literature and the potential for bias in study selection (Section 2.2.1.1). Section 2.2.1.2 (p56) outlines the specific methods used to interrogate the research literature corpus held in (Mendeley, 2016) bibliographic management software, further details of which are given in Appendix 3 (p414).

2.2.1.1 Scope of the review, sources of literature and potential for bias

Recognising that a literature review is inherently a ‘retrospective, observational’ task (A. F. Smith & Carlisle, 2015), and that the search process is ‘no more free from the impact of human subjectivity than other research’ (Okoli & Schabram, 2010, p2) the approach adopted here is *semi-systematic*; aiming to be as ‘explicit’, ‘comprehensive’ and ‘reproducible’ as possible, in line with Fink's (2005) recommendations. The literature search uses a range of Web-hosted databases and email-based alert tools, as advocated by Dunleavy (2003), Trafford & Leshem (2008) and M. Wallace & Wray (2011). Various systems have been used, including those developed by the University of Portsmouth Library, the Joint Information Systems Committee (JISC) and the British Library. Searches have also been conducted on the Web of Science, Google Scholar and on websites developed by academic publishers

including Sage Publishing, Taylor & Francis and John Wiley & Sons, amongst others. Regular email alerts from Google Scholar, each covering specific topic areas and typically returning around ten potentially relevant articles per email, have also been used.

In this study:

- Searches are conducted in the English language, although non-English texts have not been specifically excluded;
- Preference for inclusion is shown towards published works, particularly works published in journals in recent years;
- Search terms used in alert services have evolved iteratively with the longest running searches on Google Scholar (>1,200 emails since 2011) indexing:
 - [politics \"social network\"]
 - [intitle:\"geo tagging\"]
- Cross-referencing and article recommender systems have also been used to expand the 'pool' of available literature (Teppan & Zanker, 2015).

In 'screening for *inclusion* and *exclusion*' (Okoli & Schabram, 2010) consideration has been given to:

- The quality of academic writing including the use of English (grammar, spelling, punctuation) and the accuracy and extent of referencing;
- The quantitative measurement of 'relevance' as exhibited by citation and/or other bibliometric scores (e.g., journal 'impact factor').

Altogether, over 1,250 bibliographic references have been saved to Mendeley during the course of this research. The following section briefly describes how features in Mendeley Desktop, allied to third-party technologies, usefully enable bibliometric analysis of the research literature corpus.

2.2.1.2 Bibliometric analysis

Using the Help -> Create Backup... menu in Mendeley Desktop it is possible to create a backup of stored references. These are saved to an SQLite (2016) database file that may be opened, viewed and queried using open-source software (DB Browser for SQLite, 2016). Figure 2-1, based on analysis from this workflow, illustrates the number of references by publication type (journal article, book etc.) selected for inclusion in the literature search and stored in Mendeley Desktop.

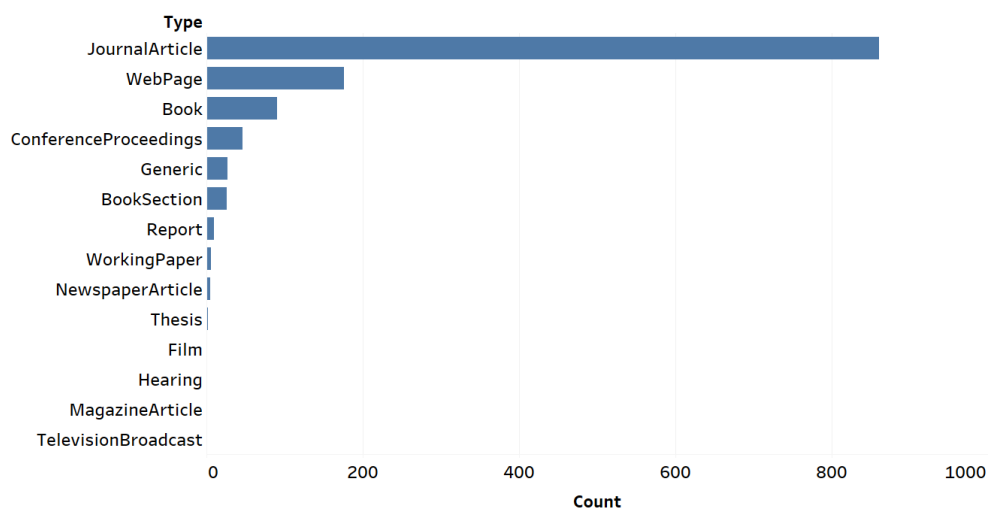


Figure 2-1 – Number of references by publication type selected for inclusion

Mallig (2010) has highlighted the benefits of using ‘relational databases [...] in the field of bibliometrics.’ RDBMSs such as SQLite, the underlying storage technology used by Mendeley Desktop, not only store data (e.g., year of publication etc.) in tables but enable queries to be executed against this stored data. Figure 2-1, which shows that journal articles comprise the majority (68.7%) of saved references in the research literature corpus and Figure 2-2 (p57), which shows the number of references by publication type by year, could not have been created within Mendeley Desktop itself, but can be graphed using Mendeley’s SQLite backup file and a SQL query run in DB Browser for SQLite (Appendix 11 listing 3, p479). Further details of this, and alternate, techniques for querying bibliographic data are given in Appendix 3 (p414) of this thesis. Pertinent results from this bibliometric analysis

exercise are detailed in the following section, alongside a report of the results of text-mining operations conducted in R (The R Foundation, 2018).

2.2.2 Results

2.2.2.1 Literature recency by publication type

A data-based approach provides useful information about the *shape* (Figure 2-1, p56 and Figure 2-2, below) and *composition* (Table 1-1, p38) of the research literature corpus. It is possible to draw two key conclusions from this analysis:

1. The literature search exhibits a strong degree of recency, and;
2. The literature search exhibits a strong degree of cross-disciplinarity.

Outwith academic 'Geography' many of the references collected as part of this search have been published in 'Political', 'Communications' or 'Computer Science' journals, some of which, e.g., *Mobile Media & Communication* (Volume 1, 2013), have only recently been established.

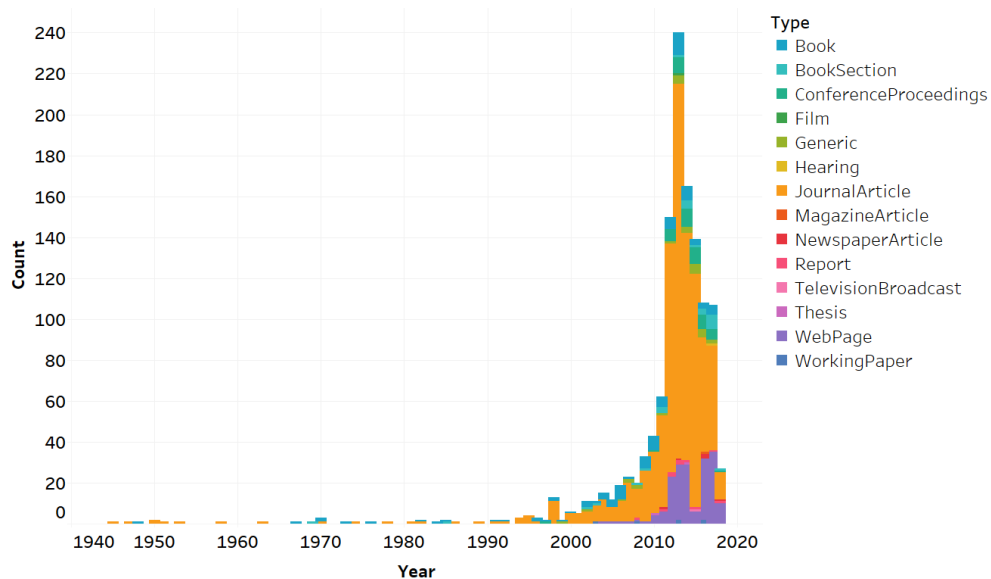


Figure 2-2 – Number of references by year by type selected for inclusion

These conclusions support the view a) that growth in OSNs and other forms of mobile communication are leading to new forms of scholarship (R. M. Chang,

Kauffman, & Kwon, 2014; Cresswell, 2014), and; b) that plenty of geographically relevant content can be found in journals outside the traditional publication bounds of the discipline of geography itself (Miller & Goodchild, 2015).

2.2.2.2 Literature mining for key terms

Around 90% of the ~1,250 references stored in Mendeley Desktop include a PDF file containing source literature content. Using computer programmes (Appendix 3, p414) developed in R (The R Foundation, 2018) the content of 1,111 PDF files has been text-mined using a Term Frequency – Inverse Document Frequency (TF-IDF) algorithm.

TF-IDF scores account for ‘the frequency of terms appearing in a document, the length of the document in which any particular term appears, and the overall uniqueness of the terms across documents in the entire corpus’ (Russell, 2011, p151). A large number of terms (185,577) from 1,111 PDFs containing 159.6MB written text (28.5 times more than the seminal, 5.6MB, *Complete Works of William Shakespeare* digitised by Project Gutenberg) have been identified using R’s Text Mining (TM) package (Feinerer, Hornik, & Artifex Software Inc, 2016).

In pseudo-code the steps involve:

- Mounting Mendeley’s PDF document repository as a ‘shared folder’ on a Linux Virtual Machine (VM) set up with the R and RStudio packages, and;
- Running scripts written in R to create a corpus from the PDF files, converting all text to lower case, removing punctuation, numbers and English-language stop words (‘and’, ‘the’ etc.) before performing statistical analysis.

Results may be tabulated or, as in Figure 2-3 (p59), visualised using a Word Cloud. Top ranked terms include, as expected, the words ‘political’, ‘tweets’, ‘twitter’, ‘social’ and ‘media’. Less prominent terms include ‘spatial’, ‘geography’, ‘analytics’ and more.

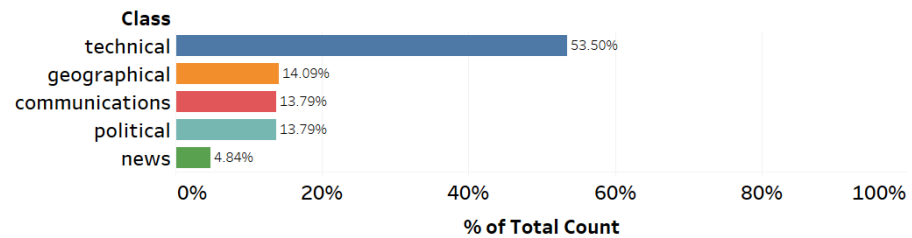


Figure 2-4 – Percentage publication titles by class

Figure 2-4 shows the percentage of 556 distinct publication titles (e.g., *New Media & Society*, *The Guardian*) allocated to each of the same four classes. As with key terms, technical publications comprise the majority (> 53%) of all references held. Geographical, communications and political classes together comprise ~42% of all references. A fifth class, news, of which there has been a great deal in the subject area during the research programme, accounts for just under 5% of all references selected for inclusion by the literature search.

2.2.2.4 Value and benefit of text-mining

The quantitative and thematic analyses detailed above, together with a great deal of reading, have helped to distill several key contextual *leitmotifs* from a large research literature corpus examining OSN usage across four cross-disciplinary boundaries, further illustrating the ‘value and benefit of text-mining’, identified by JISC (2012), in conducting literature reviews (Section 2.2, p52). A contextual synopsis and overview of the curated research literature corpus follows in Section 2.3, after which key terms, concepts and select papers from each of the four main thematic classes shown in Figure 2-4 are discussed consecutively in Sections 2.4 to 2.7 (pp64-88). Technical terms, and the technical literature, are covered lastly in this synthesis as politics, communications and geography do most to conceptually frame the current research project.

2.3 Contextual synopsis

There is widespread public and academic concern that increased political marketing or more overt interference in electoral processes using geographically and behaviourally targeted messaging on social networks may be altering the nature of democracy. The fast pace of communications and immediacy which the Internet, Web and social media allow appears to have weakened, rather than strengthened, the Public Sphere. Decreasing participation rates in political activity visible in falling levels of activism, voting attendance and party membership are also causes for anxiety. Declining readership of newspapers – particularly amongst the young, who increasingly ‘get’ their news from homophilous friend-network or recommender systems on social media – appears to have significantly changed the nature of many people’s exposure to, or consumption of, reliable information. This is especially dangerous as the spread of ‘fake news’ on Internet and social media channels is both more rapid, and more extensive, than the spread of ‘real news’.

More widely there is a perception in the literature, especially in the wake of the Facebook and Cambridge Analytica scandal, that the rise of the networked society elevates Big Data into the realms of an Orwellian Big Brother, constantly storing information about individuals’ daily actions and using these data for manipulation or control. Earlier revelations from ex-US spy Edward Snowden, published in *The Guardian* (2018), revealing that government surveillance agencies make extensive use of social media feeds, as well as any other Internet and telecommunications data they can access, had already done little to inspire public confidence in online digital privacy and, apparently, ‘a week after President Donald Trump’s inauguration’ on 20 January, 2017, following several post-electoral months in which Big Data driven campaigning techniques had come to light, ‘George Orwell’s “1984” [was] the best-selling book on Amazon.com’ (Broich, 2017).

Academic articles in the geographical and technical literature are not immune from these concerns, particularly where one or other crosses into the political or

communications space. However, the application of geographical approaches in social media Big Data analysis have been relatively limited and appear especially limited in politicised settings. With the exception of just one paper covering electoral events (H. Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012), nowhere else in their *Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data* do Steiger, de Albuquerque, et al. (2015) find any examples of GIScience-based research into geo-referenced OSN interactions situated in a political context.

Steiger, de Albuquerque, et al.'s (2015) review, graphically summarised in Figure 2-5 demonstrates that over 75% of papers came from schools of Computer Science or Information Science, according to a classification of their authors' academic research disciplines.

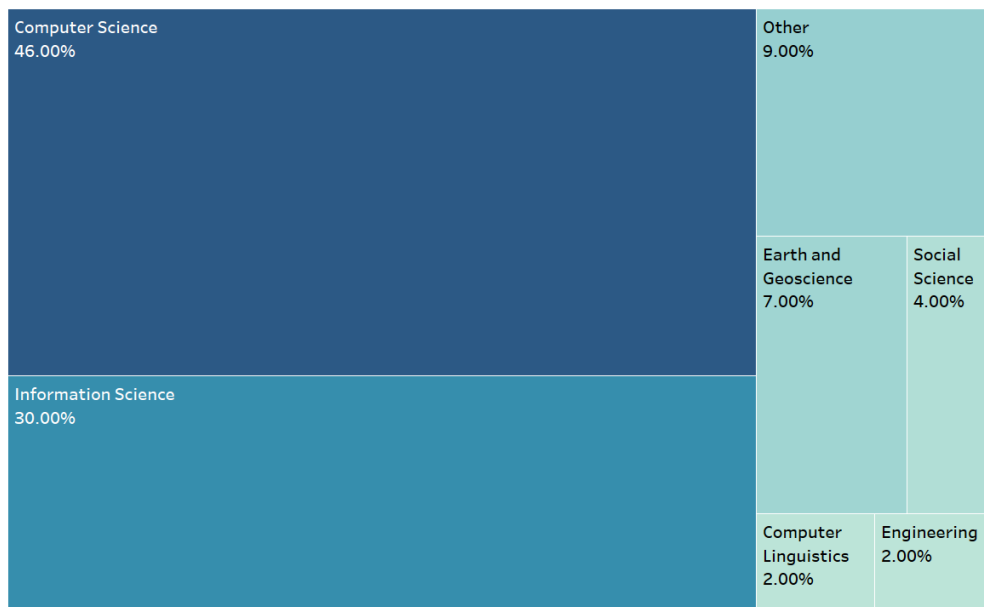


Figure 2-5 – Classification of papers according to authors' academic research disciplines (after Figure 3, Steiger, de Albuquerque, et al., 2015, p815)

Just 7% of reviewed papers emanated from 'Earth and Geoscience' departments and most of these (46%, Figure 2-6, p63) have focused on what the authors term 'Event Detection', e.g., disaster (27%), traffic (14%) or disease management (5%)

where ‘study outcomes have demonstrated a high spatiotemporal reliability and usefulness of tweets’, particularly in earthquake detection.

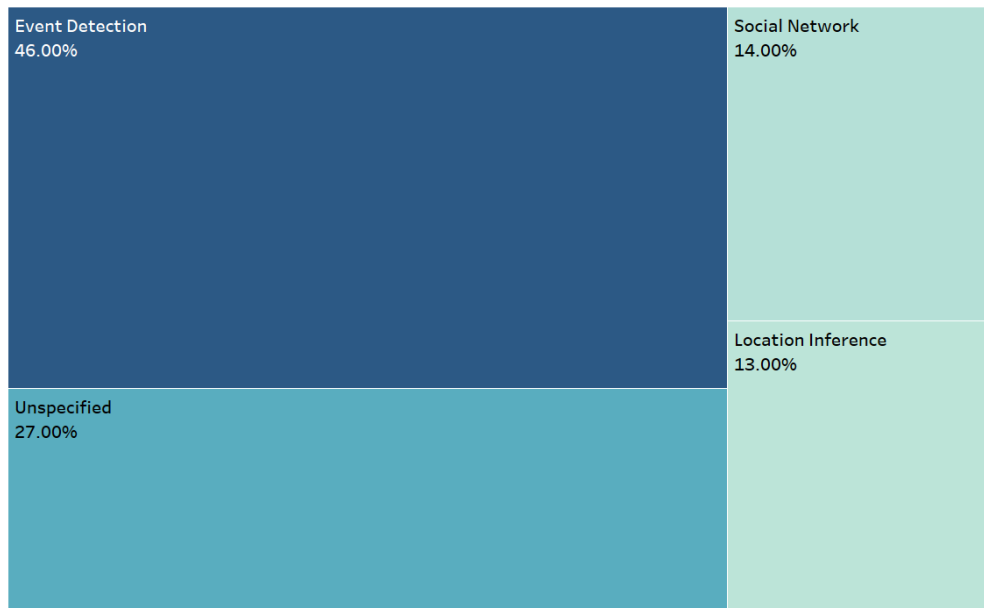


Figure 2-6 – Specific application domain of reviewed papers (after Figure 5, Steiger, de Albuquerque, et al., 2015, p817)

Social Network investigations, examining ‘individual user characteristics and their social relationships within a network’ comprise 14% of papers selected in Steiger, de Albuquerque, et al.'s (2015) review and Location Inference, focused ‘on retrieving direct or indirect geolocation information from Twitter’ comprise 13% of papers; the remaining 27% of reviewed papers could not be classified.

Online Social Networks, and Facebook in particular, clearly provide political marketers, or even state-sponsored agents such as the Russian ‘trolls’ (BBC News, 2017b) thought to have interfered in the 2016 US Presidential Election, with an unprecedented level of control over user geo-targeting and messaging. Tucker et al. (2018, p3) identify ‘widespread concern in many segments of society—including the media, scholars, the philanthropic community, civil society, and even politicians themselves—that social media may [...] be undermining democracy.’ While concerning, Agnew (2014) has suggested that geography may act as a ‘mediating’

force in politics as the boundaries of voting districts change infrequently while the socio-economic and political behaviour of many bounded electorates exhibit quite considerable degrees of stability over time; this is evidenced through the ‘geographical rootedness of political life’ and the ‘persistence of of place-specific and regional voting patterns’, some of which may also be observed in Online Social Networks.

As the fast-paced communications changes brought about by social media influence political life, geographical interpretations have increased significance. However, as the review above has shown, few published articles have considered this new media landscape from a distinctly geographical perspective. Recent events suggest that this is bound to change. The 2018 Facebook and Cambridge Analytica scandal (Cadwalladr & Graham-Harrison, 2018), and presumed Russian state-sponsored interference in the 2016 US Presidential Election campaign (U.S. House of Representatives, 2018a), demonstrate how the use of geo-behavioural targeting may be applied in attempts to perturb normal democratic processes.

There is a requirement to develop a greater understanding of the role of geography in politicised OSN discourse, and to determine whether coordinate-geotags or other forms of expressed geographicality may allow physical tracking of real or nefarious content promulgated in the virtual world. The research presented in this thesis addresses these essential questions, building upon literature in the political, communications, geographical and technical fields discussed, consecutively, below.

2.4 Political literature

2.4.1 Key publications

Political journals, articles and book excerpts provide around 14% of literature references with key publications including *Political Communication*, *Electoral Studies*, *The Annals of the American Academy of Political and Social Science*,

American Political Science Review and *Politics*. Adding more overtly sociological works to the mix incorporates leading journals including *American Behavioral Scientist*, *Social Science Quarterly*, *Sociology*, *American Sociological Review* and *Theory & Psychology*.

2.4.2 Key terms

The number of relevant politically-coded terms (Figure 2-7) in the research literature corpus is high, at 33, and the politically (and sociologically) themed literature provides an important theoretical backdrop to the technical work undertaken as part of this research.

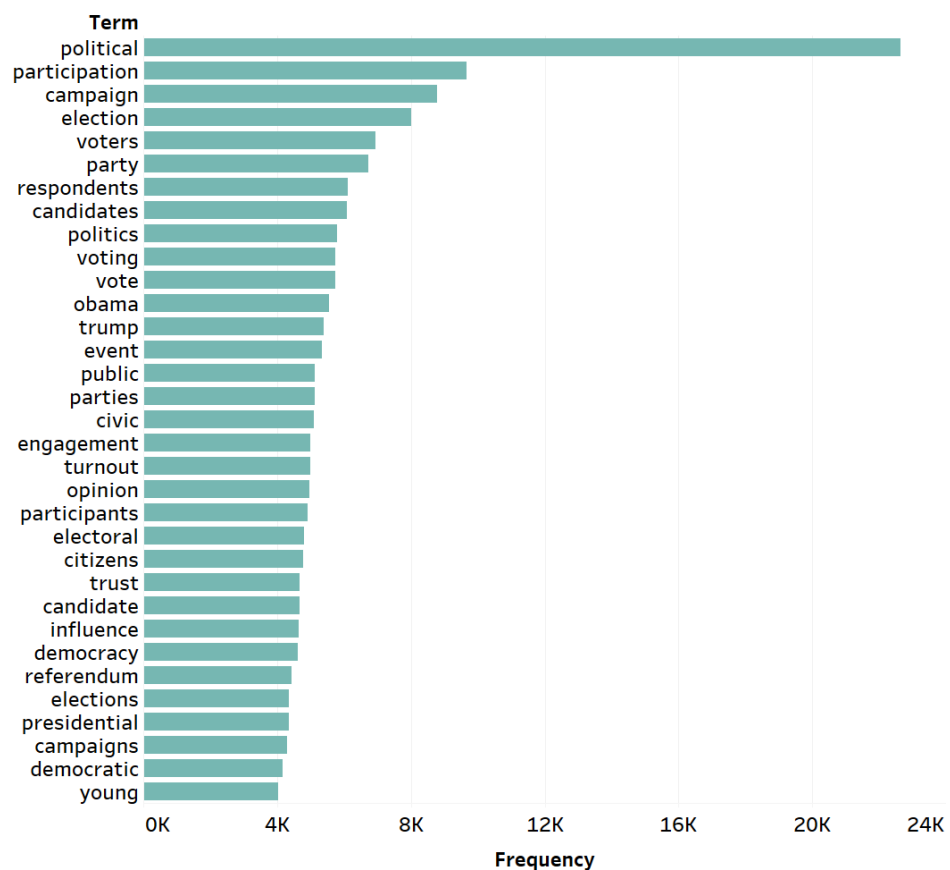


Figure 2-7 – Key political terms in the literature corpus identified by TF-IDF analysis

Key themes include matters of political participation, campaigning, engagement and trust. Political leaders including US Presidents Obama and, more latterly, Trump

both appear frequently in the literature, as do several other key terms relating to the mechanics of politics; voters and voting, parties, turnout and opinion. All of the identified terms and concepts are crucial in understanding the political impacts of newly-emergent communication systems, including online social media networks, discussed later in Section 2.5, (p72). As the introductory pages of this thesis (p1) have demonstrated, numerous politicians, regulators, scholars and commentators have suggested that there is a 'relentless threat to our democracy' (Charter, 2018) from online social networks. The perceived threat centres around the destabilising effect widespread and near-constant engagement with social media may have on individuals' political decision-making processes, which may have been purposefully manipulated by micro-targeted campaigns playing on deep-seated fears or emotions, or more simply distorted by the spread of 'fake news'.

Political deliberation, and the related concepts of deliberative democracy and participation, feature regularly in the politically themed references of the research literature corpus (Dahlgren, 2005; de Zúñiga, Copeland, & Bimber, 2014; Raphael & Karpowitz, 2013). Chambers (2003, p309) has defined deliberation as 'debate and discussion aimed at producing reasonable, well-informed opinions in which participants are willing to revise preferences in light of discussion, new information, and claims made by fellow participants.' A deliberative democracy in which citizens are well-informed and fully engaged with debate in order to make 'rational-critical' decisions represents something of an ideal in political theory (Polat, 2005), but now appears to be at risk in an age of 'weaponised' (Nissen, 2015) social media political propagandisation (Ward, 2018).

The notion of 'rational-critical' decision-making stems from the 'Habermasian' theory and concept of the 'Public Sphere'. First published in German in 1962, Habermas' (2011) work *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society* has proved an enduring and influential work (Borah, 2017). According to Habermas, the Public Sphere is 'a category that is

typical of an epoch' (Habermas, 2011, pXVII) which *transforms* over time. The Public Sphere is a product not only of systems of governance, law and economics but also of systems of communication. The Internet, and the growing reach of websites and applications enabled by the development of the World Wide Web, is the most recent communications development to have affected contemporary political discourse (L. M. Weber, Loumakis, & Bergman, 2003).

During a 'relatively brief period of euphoria', which Tucker et al. (2018) suggest is now well past, some scholars had suggested that technology and the Internet could 'save' democracy; allowing citizens to rebuild trust in their elected representatives through improved communications (Westen, 1998), or replacing 'outdated and inadequate voting technology' with Internet-based systems (Mercurio, 2004). These hopes arose from observations that democracy was 'in trouble', as evidenced by 'comparative turnout decline' (Gray & Caul, 2000, p1092), 'declining party membership' (B. Lee, 2014) and 'significant' declines in newspaper readership (N. Newman et al., 2016). These deteriorations in 'a strong and active civic society' (Putnam, 1995, p65) are contra-indicators, in Habermasian theory, of a healthy Public Sphere. Online communications, it was suggested, including enhanced 'political relationship marketing', could be used to arrest the 'rapid decline in direct participation in politics' which had been observed (Henneberg & O'Shaughnessy, 2009). Relationship marketing over social media channels has now been roundly criticised, leading some (e.g., Persily, 2017) to question whether 'democracy can survive the Internet?'

A defining concern of the political literature is its attempt to reconcile new forms of online engagement with the more traditionally 'direct' (and offline) political participation of earlier ages (Vissers & Stolle, 2014). Key contributors to the debate, including Peter Dahlgren and Zizi Papacharissi, have framed their contributions within Habermasian constructs (Dahlgren, 2005; Dahlgren & Sparks, 1991; Papacharissi, 2002, 2004, 2010). While the writings of 'Jürgen Habermas [are]

difficult to read and often [leave] students perplexed' (Thomassen, 2010, p1) the German philosopher's thinking has provided a useful model for analysing political discourse across several different epochs. Dahlgren (2005, p148) offers the following conceptual overview: 'In schematic terms, a functioning public sphere is understood as a constellation of communicative spaces in society that permit the circulation of information, ideas, debates – ideally in an unfettered manner – and also the formation of political will (i.e., public opinion).' Where these ideas or debates have been significantly manipulated, as recent events have suggested (Cadwalladr, 2017; U.S. House of Representatives, 2018a), the Public Sphere and 'rational-critical' decision-making are endangered.

In an influential article published in *New Media & Society*, Papacharissi (2002, p9) cautiously suggested that 'The internet and its surrounding technologies hold the promise of reviving the public sphere.' More recently, Papacharissi (2010, pVI) has questioned whether the Public Sphere may have 'expired'; to be supplemented (or supplanted) by a 'Private Sphere' characterised by a 'networked self and [...] remote connectivity', the 'new narcissism [of] blogging', the 'rebirth of satire and subversion', 'social media news aggregation' and the 'agonistic pluralism of online activism.' These changing political interpretations of the Web, online 'politicking' (Panagopoulos et al., 2009) and social media (Harris & Harrigan, 2015) are reflected in over 45 articles in the research literature corpus including the word 'participation' in their titles. Many allude to the possibility that increased online exposure to politics and political conversations may boost interest both in discourse (Hoffman, Jones, & Young, 2013; Yunhwan Kim, Russo, & Amnå, 2017; Ostman, 2012) and in 'measurables' such as turnout and voting numbers (Borthakur et al., 2011; Moeller, de Vreese, Esser, & Kunz, 2014; Vassil & Weber, 2011) but no decisive conclusions have been reached.

While earlier works (Mercurio, 2004; Westen, 1998) tended to view the possibilities offered by Internet-based Information and Communication Technologies (ICTs)

optimistically, more latterly – and particularly in the wake of the Facebook and Cambridge Analytica data harvesting ‘scandal’ (BBC News, 2018e) – a much more sceptical tone has set in (Persily, 2017; Tucker et al., 2018; Tucker, Theocharis, Roberts, & Barberá, 2017). Persily (2017, p64) has suggested that the ‘void’ left by declining mainstream media and political-party organisations ‘was filled by an unmediated populist nationalism tailor-made for the Internet age.’ This was particularly evident, Persily suggests, during the 2016 US Presidential Election, but has also been evidenced by the ‘rise of the Five Star Movement in Italy, the Pirate Party in Iceland, the “keyboard army” of President Rodrigo Duterte in the Philippines, and the use of social media by India’s Prime Minister Narendra Modi, who has 39 million followers on Facebook and 27 million on Twitter.’ Political use or misuse of the Internet, where (mis)information spreads quickly through social media networks (Vosoughi et al., 2018), is having a profound effect on democracies world-wide.

Sean Parker, an ex-Founding President of Facebook, has said (BBC News, 2017a) that the ‘little dopamine hits’ his social network rewards through site activity, and the strong growth of the platform, has had ‘unintended consequences’ in changing people’s relationships with society and with each other, and may be having yet more profound effects on childrens’ development. In the political sphere, Tucker et al. (2017, p46) neatly summarise this duality, asking ‘How can one technology – social media – simultaneously give rise to hopes for liberation in authoritarian regimes, be used for repression by these same regimes, and be harnessed by antisystem actors in democracy?’ The authors suggest that these contradictions can be explained by understanding ‘1) that social media give voice to those previously excluded from political discussion by traditional media, and 2) that although social media democratize access to information, the platforms themselves are neither inherently democratic nor nondemocratic, but represent a tool political actors can use for a variety of goals, including, paradoxically, illiberal goals.’

While contradictions and debate surrounding the effect of social media on politics continue, other incongruities regarding voting participation in the online age have emerged. The 2012 US Presidential Election was, according to the Pew Research Center (DeSilver, 2015), characterised by a turnout of '53.6%, based on 129.1 million votes cast and an estimated voting-age population of just under 241 million people.' In the US, this turnout figure (DeSilver, 2015) has 'been fairly consistent over the past several decades, despite some election-to-election variation. Since 1980, voting-age turnout has varied within a 9-percentage-point range – from 48% in 1996, when Bill Clinton was re-elected, to 57% in 2008, when Barack Obama won the White House.' Conversely, during the 2014 Scottish Independence Referendum, a turnout of 84.6% of the registered electorate was recorded which 'was the highest recorded at any Scotland-wide poll since the advent of universal suffrage' (The Electoral Commission, 2014, p1).

Voter participation, as measured by turnout, has shown both stability and growth in a period when interest in politics is thought to have increased as a result of participation in new forms of online communication, including social media (Lilleker et al., 2015). Two additional UK and US results do little to clarify the picture; the 2016 UK European Union Membership Referendum recorded a turnout of 72.2% (The Electoral Commission, 2016) while CNN reported 'Voter turnout at 20-year low in 2016' in the surprise victory for Donald Trump in the 2016 US Presidential Election (G. Wallace & Yoon, 2016). Habermas (2016) has stated, commenting on the unexpected 2016 'Brexit' referendum result in Great Britain, that 'It never entered my mind that populism would defeat capitalism in its country of origin.' He continues:

The relatively high turnout suggests that the populist camp succeeded in mobilising sections of previous non-voters [overwhelmingly] found among the marginalised groups who feel hung out to dry [and amongst the] poorer, socially disadvantaged and less educated strata [who]

voted more often than not for Leave. [Contrary] voting patterns in the country and in the cities [and] the geographical distribution of Leave votes, piling up in the [de-industrialised] Midlands and parts of Wales [...] point to the social and economic reasons for Brexit.

(Habermas, 2016)

The differing geographical distribution of turnout in the 2016 US Presidential Election, along with a populist surge felt in several countries (Agence France-Presse, 2016), may have helped Donald Trump in to office. Early results from the 2016 US Presidential Election (G. Wallace & Yoon, 2016) suggest that ‘some of the key states that propelled President-elect Donald Trump to his win [cast more] ballots this year than in 2012, even though overall turnout was down.’ The increasing sophistication, professionalisation and internationalisation of political marketing, leading to what Lees-Marshment & Lilleker (2012, p343) have termed the ‘marketization’ of politics, raises concerns that ‘Political marketing consultants who offer specialist skills and experience in political marketing – polling, strategy, voter profiling, segmentation, micro-targeting, voter-responsive product design – [now appear to] wield global power.’

As many of these efforts, as Moore (2016) has noted, use social media ‘to target specific voters in marginal constituencies with tailored messages’ it is clear that an understanding of politics, communications, geography and technology are required to identify these effects. Issues identified in the communications literature are discussed next, many of which overlap substantially with the concepts and debates outlined above.

2.5 Communications literature

2.5.1 Key publications

Communications journals, articles and book excerpts provide around 14% of literature references with key publications including *New Media & Society*, *Journal of Communication*, *Journal of Broadcasting & Electronic Media*, *Theory, Culture & Society* and *Journal of Computer-Mediated Communication*.

2.5.2 Key terms

The number of communications-coded terms (Figure 2-8) in the research literature corpus is high, at 24, and the communications literature – including articles appearing in newly published journals such as *Information, Communication & Society* (Volume 1, 1998), *New Media & Society* (Volume 1, 1999) and *Mobile Media & Communication* (Volume 1, 2013) – reflects the comparative recency of scholarly enquiry into emergent online communications spaces.

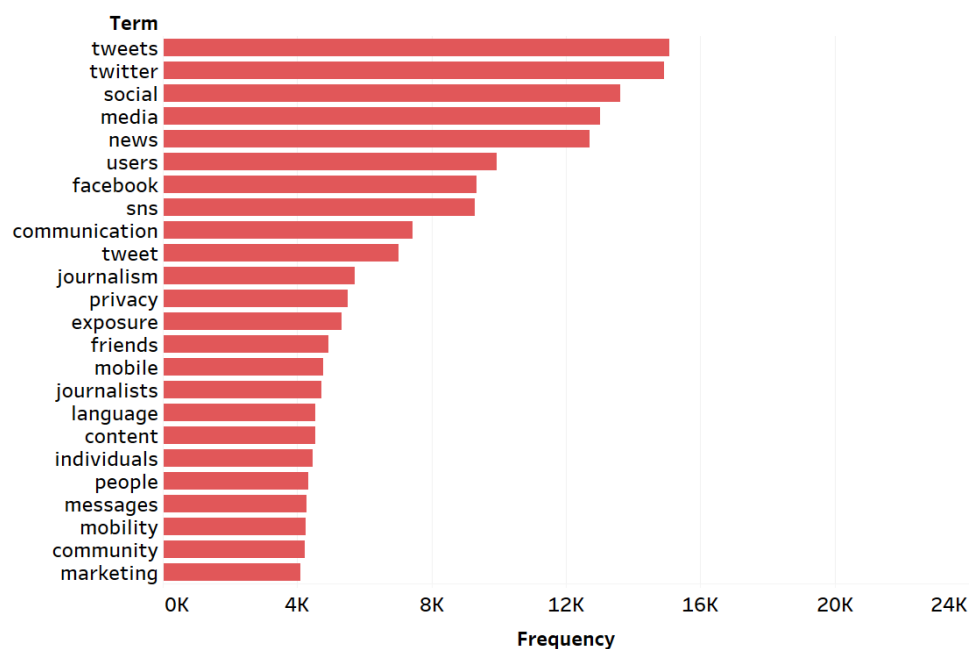


Figure 2-8 – Key communications terms in the literature corpus identified by TF-IDF analysis

Key themes include Twitter, Facebook, social media, networks and the Web. Within these are themes relating to matters of engagement and trust, sentiment, privacy and exposure. The interplay between journalistic output and communications created, consumed or shared by individuals (through tweets, posts or blogs) is also relevant in the literature.

Former British Prime Minister Harold Macmillan is supposed to have remarked that 'Events, [my] dear boy. Events' were most likely to derail the careers of senior politicians or the process of smooth governance (Ratcliffe, 2016). Whether or not Macmillan's oft-cited words are, in fact, a misquotation (Knowles, 2006) it is clear that news diffusion and, especially, the accelerated pace of communication surrounding events, represents one of the defining characteristics of politics in the Internet age. Wyly (2014, pp28-29) has suggested that 'the speedy exponential cascade' of events, now recorded in 'vast networks of proprietary corporate digital dossiers [enabling] truly revolutionary transformations in the nature of observation' poses a theoretical challenge for human geographers. Where once, he argues (Wyly, 2014, pp29-30), researchers found 'common ground' in the use of major public data sources, such as the population Census, new spatiotemporalities of Big Data and an algorithmic revolution have produced 'a Kantian temporal distortion [where] each day, Facebook is given more than 64,000 years of human expression in a digital form readily suited for advanced analyses that exceed the wildest mathematical hypotheticals of the quantitative revolutionaries of the 1950s.' Nowhere is this manifestation more apparent than in online political debate, which now far surpasses more traditional forms of political expression such as activism, public demonstration or party membership (M. Gray & Caul, 2000; Harris & Harrigan, 2015; T. J. Johnson & Kaye, 2014).

In a pre-Internet era, Calhoun (1992) summarised the contribution media channels make to a discursive Public Sphere. Referencing Habermas (2011), he suggested that the 'immediacy' with which we 'experience radio, film and television' leads to a

focus on 'personal attributes' which makes concentration on Habermas' rational-critical arguments more difficult to sustain. Calhoun (1992, p24) concluded by warning that 'A personalised politics revives representative publicity by making candidates into media stars at the same time the new public relations industry finds it easy to engineer consent among the consumers of mass culture.' The 'immediacy' in radio, film and television communications which Calhoun identified in 1992 has, of course, more recently been joined by new, even more immediate and participatory online media systems. Social media websites and mobile applications offer both a more 'immersive' experience than traditional media, and share their ability to transmit sound, still and moving images (Unwin, 2012).

Dimmick, Chen, & Li (2004, p19) have suggested that 'in light of the niche theory and the theory of uses and gratifications, a new medium survives, grows, competes, and prospers by providing utility or gratification to consumers.' Growth in online information access has resulted in 'changes in use of traditional media' and what Dimmick et al. (2004, p27) consider 'a competitive displacement effect [...] in the daily news domain with the largest displacements occurring for television and newspapers.' Earlier studies (Althaus & Tewksbury, 2000, p21) recorded patterns of Web browsing amongst university students, 'where Internet use is woven into the fabric of daily life, largely as 'a source of entertainment' and speculated that 'the World Wide Web as a news source seems unlikely to diminish substantially use of traditional news media.' More recently, the Reuters Institute for the Study of Journalism (N. Newman et al., 2016, p8), based on a survey of over 50,000 people in 26 countries, has reported that:

- 51% [of respondents] say they use social media as a source of news each week. Around one in ten (12%) say it is their main source. Facebook is by far the most important network for finding, reading/watching, and sharing news.

- Social media are significantly more important for women (who are also less likely to go directly to a news website or app) and for the young. More than a quarter of 18–24s say social media (28%) are their main source of news – more than television (24%) for the first time.

The Reuters study acknowledges that ‘television news still remains most important for older groups’ (Newman et al., 2016, p8) but notes an overall decline in usage ‘particularly for “appointment to view” bulletins and amongst younger groups’. It appears that ‘smartphone usage for news is sharply up, reaching half of [the] global sample (53%), while computer use is falling and tablet growth is flattening out.’ Many communications studies (Hargittai, Neuman, & Curry, 2012; Macafee, 2013; Pennington, Winfrey, Warner, & Kearney, 2015; Weeks, Ardèvol-Abreu, & de Zúñiga, 2015) are predicated on an examination of the changing patterns of online news consumption, with a general acceptance that rates of online news consumption are increasing. If Diehl et al. (2016, p2) are correct in their assertion that ‘[even] non-political discussion and social interaction on social media can serve as a catalyst for political expression and participation’ then the whole nature of political communications is changing.

Writing in the Editorial introduction to the first edition of *Information, Communication & Society*, Loader & Dutton (1998, pV) state that ‘A new social and economic paradigm is said to be restructuring the traditional dimensions of time and space within which we live, work and interact, which is based around information as the primary resource for social and economic development.’ Manuel Castell’s influential trilogy (Castells, 1996, 1997, 1998) proposed *The Rise of the Network Society*. Writing in the Preface to the second edition, Castells (2009, p1) recounts how, around the close of the second Millennium, ‘A technological revolution, centred around information technologies, began to reshape, at accelerated pace, the material basis of society.’ Castells (fore)saw this development arising not only through increasing globalisation and the collapse of ‘Soviet-statism’,

bringing with it an end to the threats of the Cold War era, but as an interdependency amongst world economies based on ‘a new form of relationship between economy, state, and society, in a system of variable geometry.’ Key to this development, which Castells (2009, p2) identified in financial markets, trade systems and even the criminal underworld are ‘Interactive computer networks [which] are growing exponentially, creating new forms and channels of communication, shaping life and being shaped by life at the same time.’

The communications literature is now centred comprehensively on these themes. Online ‘Web 1.0’ *publications* have, largely, been replaced by interactive ‘Web 2.0’ *applications* (O’Reilly, 2005) which promote highly personalised views of the world, frequently through advanced recommender systems (Bontcheva & Rout, 2014; Mittelstadt et al., 2016; Tao, Zhou, Lau, & Li, 2013) or reliance on content-sharing within friend networks (C. S. Lee & Ma, 2012; Ma, Lee, & Goh, 2012; Oeldorf-Hirsch & Sundar, 2015). These developments appear to have produced some profound effects, with homophilous sharing of ‘likeable’ content displacing wider or more rounded views of news and opinion (Hasell & Weeks, 2016; Himelboim, Smith, Rainie, Shneiderman, & Espina, 2017; Mousavi & Gu, 2014).

Mummery & Rodan (2013, p25) have suggested that ‘Because political blog networks tend to seek and reinforce existing political opinions they tend to be “homophilous”, falling “well short of the deliberative ideal”.’ Finding that ‘agreement out-numbers disagreement in blog comments by more than 3 to 1’ Gilbert et al. (2009, p1) suggest that blogs, including Online Social Networks such as Twitter, are ‘Echo Chambers’, where opinions are reinforced and criticism is discouraged. Outside academia, the impact of these changes in the nature of modern communications have prompted some to question whether or not we face the ‘End of Truth’ or a ‘Post Truth’ environment (Noble & Lockett, 2016), a time in which ‘Facts are now a quaint hangover from a time of rational discourse, little annoyances easily upended’ (Cohen, 2016).

The ‘traditional dimensions of time and space within which we live, work and interact’ (Loader & Dutton, 1998, pV) are undoubtedly altering as Web-based systems predominate. The role of social media in spreading ‘false news’, which research shows ‘cascades diffused to between 1000 and 100,000 people, whereas the truth rarely [diffuses] to more than 1000 people’ (Vosoughi et al., 2018), is yet another manifestation of the rapid dissemination of misinformation which networked systems of individuals, computers and robotic systems (‘bots’) allow (Bessi & Ferrara, 2016; Bessi, Scala, Rossi, Zhang, & Quattrociocchi, 2014; Metaxas & Mustafaraj, 2012; Silva, Silva, Pinto, & Salles, 2013).

One remarkable facet of this new media and communications landscape is the ability for users to share spatialised (coordinate-geotagged) locational information alongside posts (Batty, Hudson-Smith, Milton, & Crooks, 2010). The analysis of this phenomenon in a political context is the purpose of this research and the literature devoted to this development is discussed below.

2.6 Geographical literature

2.6.1 Key publications

Geographical journals, articles and book excerpts provide around 14% of literature references with key publications including *International Journal of Geographical Information Science*, *Cartography and Geographic Information Science*, *Political Geography*, *Environment and Planning A* and *Dialogues in Human Geography*.

2.6.2 Key terms

The number of high-ranking ‘geographical’ terms (Figure 2-9, p78) in the research literature corpus is lower (14) than the preceding two thematic classes yet, as so often in Geography, papers with a geographical theme frequently unite ideas expressed in greater isolation in the political or communications literature detailed above. Concepts of ‘space’ and ‘location’ feature prominently in the geographical

literature while the word stem ‘geo’ (for ‘geography’ and ‘geographic’) appears in 690, or 54.6%, of all saved references.

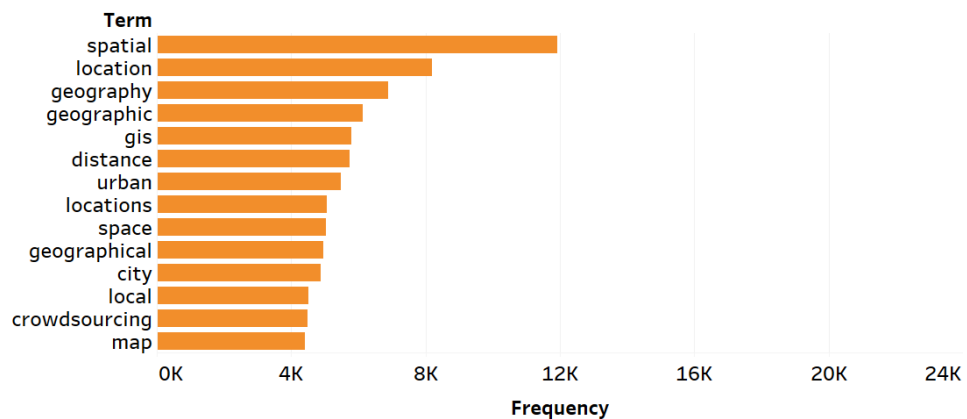


Figure 2-9 – Key geographical terms in the literature corpus identified by TF-IDF analysis

Political and communications themes frame this research as the case study OSN interactions under investigation are, by their nature, political communications. It is reasonably straightforward to view *who*, *what* and *when* someone posted OSN content publicly online, but ‘geographical biases and demographic confounds’ (Pavalanathan & Eisenstein, 2015) make it much more difficult to determine *where* a tweet or post originated from.

The challenge of the current research is to determine, using data sets in which IP addresses have been masked to protect personal locational privacy, how many of the ~8m OSN case study interactions captured in 2012 and 2013-2014, and how much of the 3rd party URL content linked to and shared alongside these messages, are imprinted with explicit coordinate or more implicit toponymic geographical references. Some authors (Z. Cheng, Caverlee, & Lee, 2010) have suggested that ‘You are where you Tweet’, but since so few users of OSNs post with coordinates this is not often the case (Leetaru et al., 2013). Understanding how geographical references are expressed, used and shared over online social media, and most widely by whom, will determine whether the coordinates posted alongside geotagging users’ interactions may be used to reliably track the spread of political

opinion and/or (mis)information cascading in message text or linked content through social networks.

Spatialities at different resolutions in OSN data are recorded in Latitude and Longitude pairs, in toponymic mentions of place or, at the lowest resolution, in time zone encodings from the temporal metadata of interaction creation date/time (Section 4.6.1, p164). Only a subset of records (Table 4-8, p170) store precise geographical information, while the wildly varying levels of spatial accuracy which may be derived from data-mining free-form text or metadata fields such as time zone encodings (Tear & Healey, 2017) make accurate assessment of posting locations particularly challenging (Section 5.4, p221).

Social networks, as Quercia, Capra, & Crowcroft (2012) have identified, are human-built. All OSNs are dependent upon a technological framework of software, servers, hosting and the Web, yet OSNs are not simply a product of technological design. Rather, they are populated and created by the self-organising sets of relationships established amongst individuals and, consequently, tend to mirror several key characteristics of the societies within which the individuals are situated. The introduction of geotagging functionality on Twitter and Facebook has, therefore, provided several fascinating opportunities to derive insights about society using Quercia et al.'s (2012) and others' (Arthur & Williams, 2017; L. Li et al., 2013; Longley & Adnan, 2016; Luo, Cao, Mulligan, & Li, 2016; Xu, Wong, & Yang, 2013; Yin, Soliman, Yin, & Wang, 2017) varied, and often quantitatively-biased, methods of computational Social Network Analysis (SNA).

Geospatial and spatiotemporal analysis of social media data in the research literature corpus springs from the growing 'etherization of geography' identified by Sui & Goodchild (2001). In a prescient Guest Editorial for the *International Journal of Geographical Information Science*, Sui & Goodchild (2001, p387) noted the 'dazzling development of GIS technology' which, they argued, rendered the 'traditional , mostly instrumental, views of GIS – as spatial database, mapping tool,

and spatial analytical tool – inadequate to capture the fundamental essence of this technology and its social implications.’ Instead, the authors proposed that ‘the complex relationship between GIS and society can be better understood if one conceives of GIS as new media.’

Evidence for this has included widespread deployment of the technology on the Web, improved data availability and data sharing over the Internet and use of GIS software in systems such as automotive navigation. The development, miniaturisation and ubiquitous deployment of Global Positioning System (GPS) micro-processor hardware in affordable mobile phone handsets, initially in the US, ‘to an accuracy of 100m—in the interests of [providing] accurate response to emergency calls’ (Sui & Goodchild, 2001, p387) has done most to cement mapping and locational awareness into people’s everyday experiences of geography.

Many applications and studies of what Turner (2006) has termed ‘neogeography’ focused on the use of geographical information in disaster response (Gelernter & Mushegian, 2011; Goodchild & Glennon, 2010; Granka, 2010). Latterly, there has been a more general shift into research surrounding the creation of ‘Volunteered Geographic Information’ (VGI; Goodchild, 2007) or ‘Ambient Geographic Information’ (AGI; Stefanidis et al., 2013). As Steiger, de Albuquerque, et al. (2015, p810) have noted, the ‘growing availability of mobile devices equipped with GPS sensors, high performing computers and broadband internet connections with advanced server and client-side key technologies, allows users to participate actively and create content through mobile applications and location-based services’, including OSN platforms.

The literature covering VGI is extensive, with several notable contributions (Hardy, Frew, & Goodchild, 2012; Warf & Sui, 2010; M. W. Wilson & Graham, 2013) and much discussion of key enabling technologies, such as OpenStreetMap (Birkin, Malleson, Hudson-Smith, Gray, & Milton, 2011; Elwood & Leszczynski, 2013; Perkins, 2014). The study of AGI phenomena, which include the OSNs Facebook and

Twitter, where the production of geographical data is a by-product of ‘sensor-based’ activity rather than the purpose of it is, as Steiger, de Albuquerque, et al. (2015, p809) have noted, ‘not clearly visible and not easy to locate’; an imbalance which this thesis aims to redress.

Tolbert & McNeal (2003, p175) have suggested that ‘new communications technology has changed the way many people gather news and participate in politics’ and that the Internet ‘permits users to exchange large amounts of information quickly regardless of geographical distance.’ Despite these advances, many Western democratic political systems are still characterised by geographical forms of political organisation originating ‘with the constitutional reforms of Cleisthenes’ in 509 BC (Tridimas, 2011). Democracy, ‘rule of the commoners’, requiring open debate in the Assembly of the Ancient Greeks and informing Habermas’ notion of the Public Sphere, has now been replaced by a number of ‘representative democratic’ systems (Dahlberg, 2013) where governance is determined by voting outcomes in multiple geographically ‘bound’ constituencies (Elden, 2005).

As Moore (2016) and The Electoral Commission (2018a) have noted, the ability to perturb the results of democratic elections through sophisticated electioneering and highly targeted political geo-marketing in ‘swing’ constituencies now represents a significant threat to existing democratic systems. The election of Donald Trump in the 2016 US Presidential Election has, likewise, been surrounded by concerns over geo-data-driven campaigning (Albright, 2017; Kohn, 2016) and the amazing possibility, identified by the US Central Intelligence Agency (CIA), that state-supported ‘Russian hackers’ may have intervened electronically in support of Trump’s candidacy (Sanger & Shane, 2016).

Earlier sections (2.4, p64; 2.5, p72) have detailed political and communications themes within the academic literature. It is clear that the study of geo-enabled social networks, where much political debate and marketization now takes place,

requires a cross-disciplinary approach. Clark & Jones (2013, p312) have suggested that ‘Ideally, “mixed-method” approaches should be used to tackle cross-disciplinary work on spatialization, bringing together for example the strengths of quantitative modelling in political science with the robust qualitative case history approaches of human geography.’

Geography is frequently thought of as a ‘holistic’ discipline (Archer, 1995) and the growth in OSNs and increasing production of geotagged content, leading to what S. W. Campbell & Kwak (2011), Lee (2012) and Bahir & Peled (2013) have termed ‘Geo-Social Networks’, is evident both in growing bodies of research output (Chen, Vasardani, & Winter, 2017; Hawelka et al., 2014; Qunying Huang & Wong, 2015; Humphreys, 2013; Koylu, 2018; S. Li et al., 2016; Licoppe, 2013; McKenzie & Janowicz, 2014; Pradeepa & Manjula, 2016; Purves et al., 2018; Steiger, de Albuquerque, et al., 2015; Van Diepen, Twigg, Ekinsmyth, & Moon, 2017; Yin et al., 2017) and novel journal publications such as *Mobile Media & Communication* investigating the characteristics of this new ‘locative media’ (Wilken, 2012).

Tsou (2015) has suggested that analysing ‘geo-social media’ geographically raises several ‘major research challenges for GIScientists’ including, in abridged form:

1. improving demographic information about users;
2. creating a multi-scale spatiotemporal analysis framework;
3. protecting user privacy and locational privacy;
4. using multi-disciplinary techniques;
5. linking ‘content’ with ‘context’;
6. reducing ‘noise’, and;
7. addressing problems regarding the ‘repeatability’ of research.

Subsequent sections of this research address these challenges, moving the study of geo-located OSN data ‘beyond the geotag’, as Crampton et al. (2013) have recommended. Doing so requires sophisticated computer systems and intelligent

analyses; the technical and technological themes which frame this research are discussed in the following section.

2.7 Technical literature

2.7.1 Key publications

Technical or technological journals, articles and book excerpts provide around 54% of literature references with key publications including *Computers in Human Behavior*, *Social Science Computer Review*, *Information, Communication & Society*, *arXiv* and *First Monday*.

2.7.2 Key terms

The list of relevant technical terms (Figure 2-10, p84) identified in the research literature corpus is extensive, at 65, although some terms (e.g., 'semantic', 'actors') undoubtedly cross thematic coding boundaries. Overall, the technical terms in the literature corpus are concerned with the Internet, social networks, big data, database systems for query and analysis, graph theory, and software development. These themes are explored below. Detailed technical articles, used to inform decisions regarding data collection and storage or data analysis and visualisation, are referenced more extensively in Chapter 4 (p118).

The key word 'online' is mentioned in 843 (66.7% of all) articles, either in title text or in body copy; a slightly lower number of references (727, or 57.5%) mention the key word 'Internet'. Clearly, a literature corpus which results from the collection of many recent references covering Online Social Network usage is bound to refer to 'the online' extensively. Ignoring this self-referential loop, the literature frames two much broader questions: 1) How should the online be viewed? 2) How should the online be studied?

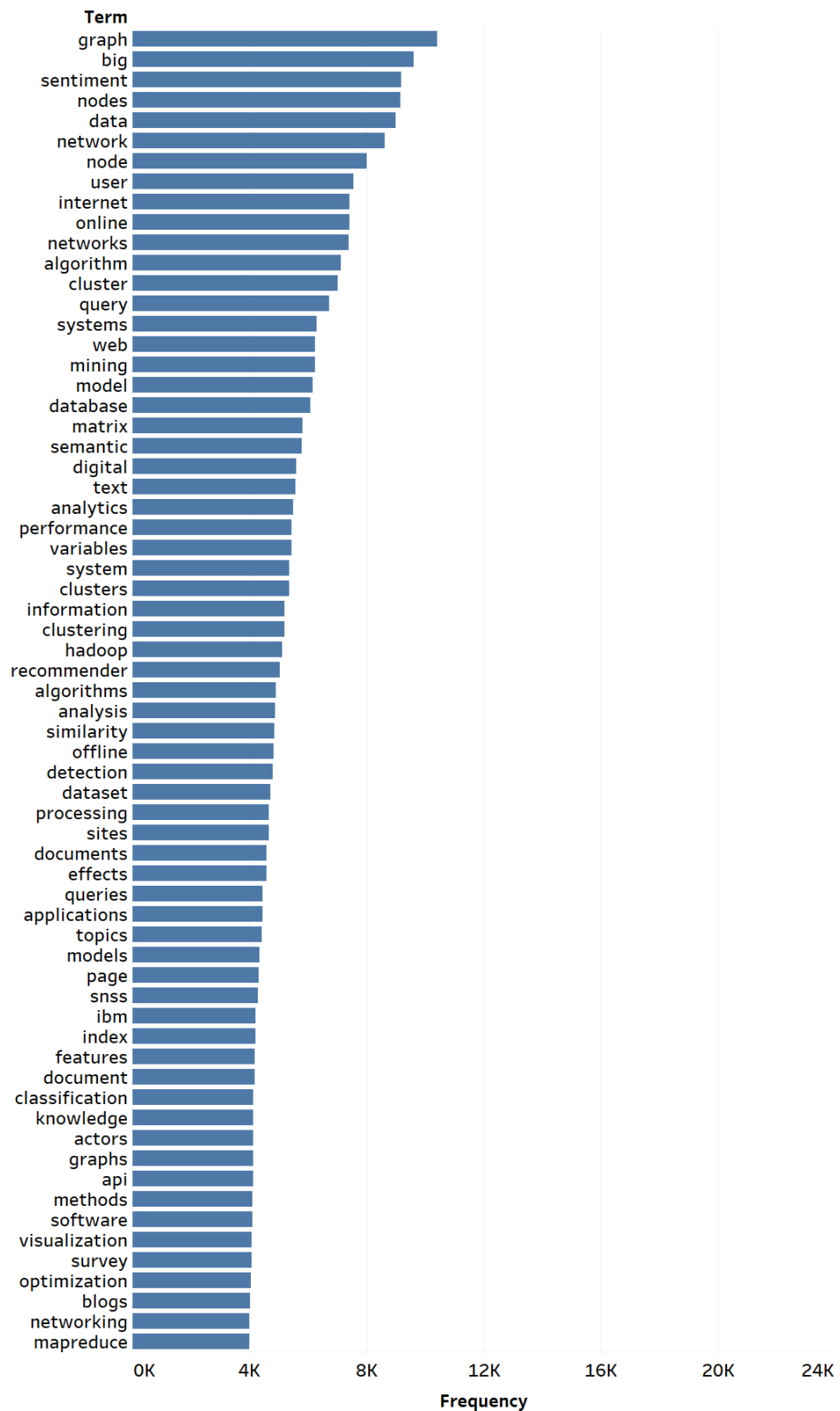


Figure 2-10 – Key technical terms in the literature corpus identified by TF-IDF analysis

In the context of this research, Papacharissi's (2010) description of *A Private Sphere* provides a useful techno-sociological answer to the first question. Here it is suggested that online activities 'may possess public and private essence, imperatives may be personal and political, communications may be intimate and mediated, and audiences may be individual or multiplied' (Papacharissi, 2010, p162). The private sphere may be seen as a 'technologically equipped bridge of overlapping and networked spheres [...whose continual change in form...] illuminates the function of online technologies in a democracy: to connect, to create new space...' (Papacharissi, 2010, p164). These contemporary developments are driven by several of the changes in politics, communications and geography outlined earlier, but also require an explicitly developed, socio-technologically-based and cross-disciplinary mode of study.

Berners-Lee et al. (2006, p771) state that 'the scale, topology, and power of decentralized information systems such as the Web [...] pose a unique set of social and public-policy challenges' as the Web is both a technically 'engineered space' with multiple languages and protocols but has been created, and is used, by humans whose 'interactions are, in turn, governed by social conventions and laws' (Berners-Lee et al., 2006, p769). To fully understand these complexities the authors propose *Creating a Science of the Web* which 'is about more than modeling the current Web. It is about engineering new infrastructure protocols and understanding the society that uses them [...] It uses powerful scientific and mathematical techniques from many disciplines to consider at once microscopic Web properties, macroscopic Web phenomena, and the relationships between them' (Berners-Lee et al., 2006, p771).

Web Science has been called *A Provocative Invitation to Computer Science* (Shneiderman, 2007, p25) as 'the social perspective [creates a] disruptive shift [which] involves moving away from studying the technology toward studying what users can do with the technology.' Berners-Lee and fellow collaborators have subsequently expanded considerably upon their original article in *Science*, setting

out a 130-page ‘agenda for the science of decentralised information systems’ (Berners-Lee, Hall, Hendler, O’Hara, et al., 2006) and, later (Hendler, Shadbolt, Hall, Berners-Lee, & Wietzner, 2008), a further paper describing the ‘systems approach, in the sense of “systems biology”’ required to understand the interplay between ‘Social Interactions, Application Needs and Infrastructure Requirements’ (Hendler et al., 2008, Figure 1, p62).

The idea of the Internet or, more specifically, the World Wide Web as an ‘organism’ (Heylighen, 2007) or even ‘emergent Global Brain’ (Mayer-Kress & Barczys, 1995) is a fascinating concept, but one that cannot exist without an understanding of some of the major drivers supporting this new Kuhnian (1970) paradigmatic shift in society’s relationship with technology. Chief amongst these is the emergence of Big Data (Halavais, 2015; Jenkins et al., 2016). While many types of human (e.g., social network, digital photographic) or sensor-based (e.g., server logs, remotely-sensed imagery) Big Data exist, the ‘computational social sciences’ (Lazer et al., 2009) most frequently analyse data extracted from online social media platforms in research.

As Zelenkauskaitė & Bucy (2016) have noted:

Recent decades have witnessed an increased growth in data generated by information, communication, and technological systems, giving birth to the ‘Big Data’ paradigm. Despite the profusion of raw data being captured by social media platforms, Big Data require specialized skills to parse and analyze — and even with the requisite skills, social media data are not readily available to download. Thus, the Big Data paradigm has not produced a coincidental explosion of research opportunities for the typical scholar.

(Zelenkauskaitė & Bucy, 2016, p1)

Several factors are at work, including ‘the cost and control of social media data’ and also the ‘order of magnitude increases in the complexity of data storage and retrieval and the need for sophisticated statistical tools to analyze it’, prompting

Zelenkauskaitė & Bucy (2016) to ‘illuminate a curious but growing “scholarly divide” between researchers with the technical know-how, funding, or institutional connections to extract big social data and the mass of researchers who merely hear big social data invoked as the latest, exciting trend in unattainable scholarship.’ The gatekeeping concept invoked, surrounding access to data and technology, originates from Lewin's (1947) ground-breaking research into ‘group dynamics’ and ‘action research’.

Lewin's theory has been widely applied in several mediated systems, including Information and Communications Technology where, for example, news editors or systems administrators decide which articles to publish, or which users should be granted access rights to data (Deluliis, 2015). OSN Big Data appear to elevate and conflate gatekeeping issues. Concerns regarding individual user privacy (Small, Kasianovitz, Blanford, & Celaya, 2012), overall data cost, volume, and storage – coupled with the specialist technological and statistical skills required of the analyst (Iacus, 2014) – all present potential barriers to research. Epistemological and ethical limitations in social media Big Data research are more fully discussed in Chapter 3 (p94) of this thesis. The costs of data acquisition, and practical matters regarding the storage and query of large data sets, are discussed in Chapter 4 (p118).

Outwith these broader concerns the technically themed articles in the research literature corpus discuss numerous technical matters; e.g., approaches to geoparsing free form text (Alonso-Lorenzo, Costa-Montenegro, & Fernandez-Gavilanes, 2016; Gelernter & Mushegian, 2011; K.-S. Kim, Kojima, & Ogawa, 2016) or methods for storing ‘unstructured’ data (Demchenko, de Laat, & Membrey, 2014; Kambatla, Kollias, Kumar, & Grama, 2014; Manyika et al., 2011; Puglisi, Montanari, Petrella, Picelli, & Rossetti, 2014). As a review of the many terms used in the technical literature (Figure 2-10, p84) is outside the scope of this thesis only the most salient works touching on political, communications and geographical themes have been discussed above. Elsewhere, throughout this thesis, highly relevant technical material is referenced extensively.

2.8 Gaps in knowledge

The rise of Castells' (1996) *Networked Society*, lays bare many elements of contemporary human life in digital form. It has been stated (Lazer et al., 2009, pp1-2) that 'We live life in the network', checking email, making phone calls, using digital 'mass transit cards' and buying food with credit cards; 'Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago.' OSN interactions – some coordinate-geotagged, many containing geographical references in text or linked/shared URL content – now offer a remarkable resource enabling researchers to examine the spatiotemporal characteristics of modern-day life in incredible detail, and at unprecedented scale. These developments hinge upon several known findings or, in some cases, implicit assumptions – discussed in turn below – which define the current gaps in our knowledge:

2.8.1 Geotagging rates

It is clear that coordinate-geotagged 'location data is incredibly valuable as it enables us to establish the geographic context in which the tweeter is immersed at the point of data creation' (Sloan & Morgan, 2015, p2). However, it has also become increasingly clear that only small percentages (~1-2%) of OSN data are typically geotagged (Leszczynski & Crampton, 2016; Paraskevopoulos & Palpanas, 2016) with Leetaru et al.'s (2013) trawl through over 1.5 billion tweets in search of coordinates oft-cited in the literature. Despite these findings, and some others which suggest slightly higher geotagging rates in response to particular events such as manmade or natural disasters (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013), coordinate-geotagged OSN data have been widely used in several application domains. There is an implicit assumption that, due to scaling laws and massive data volumes, 'this one percent is already large enough' (Jiang et al., 2016,

p349) for meaningful geographical analyses. It is appropriate to question, however, whether ~1% really is 'enough', or is representative enough, for coordinate-geotagged interactions to be used as accurate geographical proxies for all social media communications, particularly in political contexts.

2.8.2 Representativeness

Users of Facebook and Twitter are not thought to be representative of the general population (Mellon & Prosser, 2017) and coordinate-geotagging users of OSN platforms are not thought to be representative of all such users (Sloan & Morgan, 2015). There is evidence, during elections (Barberá & Rivero, 2015, p712), that 'Twitter users who write about politics tend to be male, to live in urban areas, and to have extreme ideological preferences.' There has been a 'perennial criticism [...] regarding the lack of demographic information' in social media data in its application to social science research (Sloan et al., 2013, p1). Mislove et al. (2011, p554) have found 'that the Twitter population is a highly non-uniform sample of the [US] population' and Tufekci (2014) has warned of the dangers of 'representativeness, validity and other methodological pitfalls' associated with 'social media big data research'. While some advances have been made in understanding the demographic composition of coordinate-geotagging (Sloan & Morgan, 2015) and non-coordinate-geotagging (Sloan, Morgan, Burnap, & Williams, 2015) users of OSN platforms, 'little is known [about] divides between Twitter users, based on the spatial and temporal distribution of the content they produce' (Rzeszewski & Beluch, 2017, p1). One, so far unexplored, aspect of this puzzle is the degree to which differential usage of toponymic mentions of place in OSN interactions created by coordinate-geotagging and non-coordinate-geotagging users may intersect with the 'localness assumption' (I. L. Johnson et al., 2016) often implicitly adopted when spatially-tagged social media messages are used in research.

2.8.3 Toponymic usage

Expressions of 'place' in OSN interactions and metadata are much more prevalent than coordinate-geotagged 'spatialities'. Ambient geospatial information in OSN messages 'often has geographic footprints, for example, in the form of locations from where the tweets originate, or references in their content to geographic entities' (Stefanidis, Crooks, et al., 2013, p319). It has been suggested (Goodchild, 2013, p280) that 'a large proportion of what is currently being envisioned as big data will be georeferenced, that is, will specify observations or facts about some location on or near the earth's surface.' In this study, around 25% of ~8 million OSN interactions (Section 5.2.2, p190) contain some form of toponymic information in message text, with more found in associated metadata. Toponymic mentions may be 'geo-parsed [using] location inference techniques' (Ajao et al., 2015) and an expanding number of 'geographic information retrieval' tools (Purves et al., 2018). A substantial body of research work is devoted to geoparsing both modern-day (Chen et al., 2017; Smart, Jones, & Twaroch, 2010; Zhang & Gelernter, 2014) and historical (Southall, Mostern, & Berman, 2011; Southall, von Lunen, & Aucott, 2009) toponyms using a range of gazetteer, natural language processing and other methods, e.g., neural networks or language modelling. Some of these systems have been open-sourced (Berico-Technologies, 2017; Bontcheva et al., 2013; Defence Science and Technology Laboratory, 2015; Language Technology Group, 2014) and others (IBM, 2017a) are proprietary. Three systems have been used here (Section 4.4.1, p147) to help determine how differential expressions of geographicality have been made in message text and linked/shared URL content by coordinate-geotagging and non-coordinate-geotagging users of OSN platforms. Doing so deepens our understanding of how geographical senses of 'space' and 'place' are expressed in online social media interactions, and most frequently by whom.

2.8.4 Localness

The ‘localness assumption’, identified by I. L. Johnson et al. (2016), has only recently been stated (Section 1.7, p34) but has been ‘implicitly assumed’ in most studies involving coordinate-geotagged social media data. In this type of research, the authors suggest (p515), ‘An important [underlying] assumption [...] is that social media VGI is “local”, or that its geotags correspond closely with the general home locations of its contributor [however, a]nalysis of ‘three separate social media communities (Twitter, Flickr, Swarm) [shows] that this localness assumption holds in only about 75% of cases.’ The remaining 25% of coordinate-geotagged content is, on average, ‘non-local to an area’ as smartphone-based online locational posting and changing patterns of human mobility have perturbed previously-observed, and generally more static, relationships between geotagged and home locations. I.L. Johnson et al.’s work on ‘localness’ raises another fundamental question, which is the primary topic addressed by this research. It is assumed that coordinate-geotagging is important, that *geotagging matters*, because geotagging users mention proximal locations as well as depositing their Latitude and Longitude coordinates when they post, but do they? Who makes most geographical references in their message text? Who links to and shares URLs making the most frequent toponymic mention of place; coordinate-geotagging or non-coordinate-geotagging users of OSN sites?

2.8.5 Testing the Geographicality Assumption

The research presented here tests this *Geographicality Assumption* using a research design and methodology described in the following chapter and a set of data collection, storage, augmentation and visualisation systems detailed in Chapter 4 (p118). Results from this research are presented in Chapter 5 (p186) with additional findings given in Chapter 6 (p227).

2.9 Summary

Sui & Goodchild (2011, p1746) have stated that ‘social media [has] become more locationally aware’, yet little attention has been paid by Geographers to the neogeography of electoral events (Steiger, de Albuquerque, et al., 2015; Section 2.5, p91). Why do some people choose to post with coordinates? Do they post differently? Are they atypical demographically? Do they link to different types of material? Do they use geography differently? Some work has been conducted in these areas (e.g., Hemsley & Eckert, 2014) but much scholarly research has exhibited a simplistic ‘fetishization of [the Latitude and Longitude] geotag’ itself (Leszczynski & Crampton, 2016).

The political and communications literature features considerable debate regarding recent growth in online ‘political consumerism’ (de Zúñiga et al., 2014), questioning whether this is increasing political participation or whether it is simply a form of ‘prosumption’ (Paltrinieri & Esposti, 2013) or ‘media fandom’ (Sandvoss, 2013, p252). Habermas’ (2011) description of a *Public Sphere*, an open and liberal arena in which public deliberation informs democratic opinion, is contrasted with what Papacharissi (2010, p131) has described as ‘a private sphere of interaction [...] located within the individual’s personal and private space.’ The sharing of news and opinion within this space has been found to exhibit a good deal of homophilous behaviour (Glynn, Huge, & Hoffman, 2012; Yonghwan Kim, Hsu, & de Zúñiga, 2013; Messing & Westwood, 2012) such that ‘if politics [...] is discussed, new information may not be acquired because people are only sharing like-minded views’ (Gerber, Huber, Doherty, & Dowling, 2012, p851). This is a problem which appears especially acute as researchers have now shown that ‘fake news’ travels further and faster in social networks than ‘real news’ (Vosoughi et al., 2018).

As so many political discussions touch on geography (e.g., ‘What will the vote be in Ohio?’, ‘Can the SNP win in Dundee?’) the purposeful distribution of geo-targeted political advertising, itself containing geographical text, may be highly effective.

Campaigning of this type, especially when behavioural triggers are also employed (Section 1.1, p1), is generally considered detrimental to ‘rational-critical’ political deliberation (Habermas, 2011) and is thought to pose considerable dangers to democracy. Writing in *The Times*, Charter (2018), reporting from Washington D.C. on the US Senate’s latest investigations into misinformation campaigns ahead of the November 2018 primaries, has stated that ‘Western democracies face a relentless threat from hackers determined to undermine elections.’ Charter’s report contains quotes from Sheryl Sandberg, Mark Zuckerberg’s deputy at Facebook, stating that the site ‘stopped millions of attempts to register bogus accounts every day’ while admitting that ‘3 to 4 per cent of current users were “inauthentic” and did not represent real people’, a number amounting to tens of millions of Facebook’s ~2 billion total user base. Jack Dorsey, founder and chief executive of Twitter, also present at the hearing, claimed to be blocking ‘500,000 phone log-in attempts a day and pledged “tectonic” changes [to the site] to prevent manipulation.’ Speaking for the US Department of Homeland Security, Kirstjen Nielsen recounted how her department was created to ‘prevent another 9/11’ but that she now believed ‘an attack of that magnitude is now more likely to reach us online than on an aeroplane. Cyberspace is now the most active battlefield, and the attack surface spreads into every single American home’ (Charter, 2018).

Attempts to track or monitor the spread of political opinion or (mis)information deliberately disseminated via social media are hampered by online anonymity, low levels of coordinate-geotagging amongst OSN users, platform privacy policies redacting potentially useful information in interaction metadata (e.g., IP addresses) and wide-ranging technical difficulties associated with storing and text-mining very large volumes of data (Chapter 4, p118). There are also several profound epistemological, methodological and ethical issues surrounding the investigation of message text and related data created by individuals’ who, although posting publicly online, have not been conventionally co-opted into the research study. These subjects are examined in the following chapter of this thesis.

3 RESEARCH DESIGN

3.1 Introduction

This research is designed to test the *Geographicality Assumption*, that *coordinate-geotagging users are the most geographically expressive of all OSN users*, by addressing three research questions:

1. How can baseline ‘geographicality’ be assessed and categorised in OSN data?
2. Does NLP-detectable ‘geographicality’ in message text increase in line with ‘spatiality’?
3. Does NLP-detectable ‘geographicality’ in linked/shared 3rd party content increase in line with ‘spatiality’?

To answer these questions, it is necessary to acquire, store, augment and query social media data and to tabulate, analyse and visualise results. It is also necessary to understand epistemological facets of social media production and to adopt and develop an appropriate methodology for the research. This chapter begins by addressing these latter topics before the following chapter on research methods (p118) describes ‘what was done to answer the research question[s], describe[s] how it was done, justif[ies] the experimental design, and explain[s] how the results were analyzed’ (Kallet, 2004, p1229). Finally, this chapter concludes by considering ethical issues in social media analysis, which are especially relevant as none of the ~2.4m users whose messages have been analysed here have been conventionally co-opted into this study.

It has been suggested that ‘The era of big data has created new opportunities for researchers to achieve high relevance and impact [as the] *scientific paradigm shift toward computational social science* [has enabled] fundamental changes [in] the research questions we can ask, and the research methods we can apply’ (R. M.

Chang et al., 2013, p67, author's italics). Billions of users (Figure 3-1) now post messages on social media platforms such as Facebook and Twitter; deliberately, or unwittingly in some cases (Bertino & Matei, 2015; Cresswell, 2014), depositing their thoughts and opinions in the public domain. The users' message text or link shares and the associated metadata bundle which contains the date of posting, the number of 'friends' and 'followers', indicators of importance within the network site, coordinate-geotags if available, and much else besides, provides a rich set of content and attributes for research.

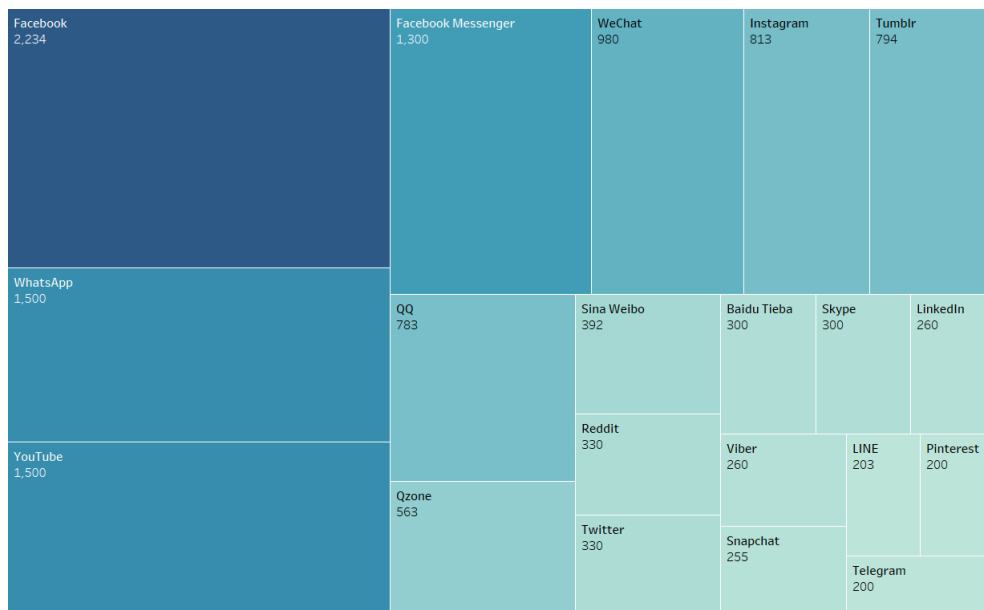


Figure 3-1 – Monthly Average User (MAU) counts, in millions, for major social media sites (Statista, 2018a)

There has been an explosion in social media usage and in Big Data research (Figure 2-2, p57), with a corresponding increase in the number of scholarly papers devoted to these subjects (Chapter 2, p51 and Figure 2-2, p57). These developments have not been universally welcomed. Longley has warned (Groom & Booth, 2016) of the 'delusional and at best misleading' nature of Big Data which, he says, 'can be described as the "exhaust" from millions of daily transactions, including social media, from which it is possible to gain some insight but to what populations?' Wyly (2014, p35) has warned of the dangers of a 'new quantitative revolution' in

Geography as 'Big data enables and encourages empiricist data-mining logics unhinged from the positivist framework of assumptions, hypotheses, and causal explanation.' Methodological problems with OSN Big Data research have been comprehensively detailed by Tufekci (2014) and several other authors have raised questionmarks over the utility of Big Data research, both generally (Fuchs, 2017a; Hargittai, 2015; Iacus, 2014) and in a Geographic Information Science (GIScience) context (Goodchild, 2013; Miller & Goodchild, 2015; Tsou, 2015).

The intersection of Big Data, location, sentiment, unfamiliar data structures, technologies and processes presents particular challenges to GIScience, forcing new requirements to efficiently handle and analyse huge scale, text heavy, un/semi-structured, spatiotemporal data, often derived from Web-based sources (S. Wang, 2013; Yousfi, Chiadmi, & Nafis, 2016). This research uses mapping, data analytic and text-mining approaches to examine the online 'spatialisation of political behaviours' (Clark & Jones, 2013, p313). The following sections set out the epistemological and methodological background to the study. Later in the chapter ethical matters are addressed.

3.2 Epistemology

United States Secretary of Defense Donald Rumsfeld memorably stated, when asked about the lack of evidence linking the Iraqi government to the supply of 'weapons of mass destruction' (WMD) to terrorist groups at a US Department of Defense (DoD) news briefing on 12 February 2002, that:

Reports that say that something hasn't happened are always interesting to me because as we know, there are known knowns: there are things we know we know. We also know there are known unknowns: that is to say we know there are some things [we know] we do not know. But there are also unknown unknowns – the ones we don't know we don't

know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult one.

(Rumsfeld, 2011, pXIII)

As Logan (2009, p712) explains, the idea of ‘Known knowns, known unknowns, [and] unknown unknowns’ is not a new one. Indeed, ‘the concept of the unknown unknown existed long before Donald Rumsfeld gave it a new audience.’

Epistemologically, however, the notion of different levels of knowledge, or uncertainty, fits well with OSN data analysis. Publicly available OSN data sets are characterised by uncertainty in multiple dimensions:

- **Demographically** – It is impossible to know, with certainty, whether individual users are male or female or, even, how old they are. The same holds true for ethnicity and a raft of other demographics (Qunying Huang & Wong, 2016; Mislove et al., 2011) commonly used as controls in population-based or survey research (Gittelman, Thomas, Lavrakas, & Lange, 2015). Automated demographic classification has been attempted (Pennacchiotti & Popescu, 2011) but, since we do not know whether users are who they claim to be (e.g., their avatar photograph may have been misappropriated and is not their true likeness or, even, ethnicity; they may use a female name when they are male) results may have only limited utility.
- **Geographically** – it is equally impossible to know, with certainty, whether individual users are located where they say they are. It is difficult to determine whether geo-references in user taglines (e.g., ‘I’m [forename], from Denver’), user posts (e.g., ‘Hi, visited [some location] today’) or Latitude and Longitude geotags are accurate. Also, not all users posting with coordinates are humans using GPS-enabled mobile smartphone devices; Echeverría & Zhou (2017) have detected substantial ‘botnets’ making spatialised posts on Twitter. Whether or not coordinates are fully trusted, it is difficult to determine whether individual OSN posts are made from home,

work, vacation or other locations (Aladwani, 2015; Warf & Sui, 2010).

Diurnal analyses of OSN feeds (Morales et al., 2017) goes some way to addressing these problems, but significant doubts about ‘inflated granularity in Spatial Big Data’ remain (Dalton & Thatcher, 2015)

- **Semantically** – Not unsurprisingly, given the above, it is difficult to know whether what we observe on OSNs is genuinely the opinion of the user (Which user? Where?), whether accounts or topics might have been hijacked (Ferrara et al., 2013) or whether content is being deposited robotically (Boyd & Crawford, 2012; Roy & Zeng, 2015). ‘Ghost accounts’ or ‘bots’ (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015) are thought to have distorted the outcome of the 2016 US Presidential Election (Bessi & Ferrara, 2016) and may have played a role in earlier events (Shin, Jian, Driscoll, & Bar, 2017). ‘Fake news’ stories are easily spread on OSNs (Kaplan & Haenlein, 2010) and it appears that growing numbers of individuals, e.g., in Macedonia, are involved in the production of fake, and highly profitable, online ‘clickbait’ masquerading as news (Kirkby, 2016).

Targeted ‘computational propaganda’ appears to be a growing problem in several established democracies. In the US, Howard, Kollanyi, Bradshaw, & Neudert (2017, p4) have found that ‘Many of the swing states getting highly concentrated doses of polarizing content [in 2016] were also among those with large numbers of votes in the Electoral College.’ In Kenya, Cambridge Analytica mined voters’ data ‘to help President Uhuru Kenyatta win disputed elections. Over two presidential election cycles, it presided over some of the darkest and most vicious campaigns Kenya has ever seen’, ‘poisoning’ democracy in that country (Madowo, 2018). It is currently unclear, and may eventually prove impossible to determine (W. Davies, 2018), whether these interventions by companies or state-sponsored actors have had any direct impact or ‘clear monolithic effects’ on electoral outcomes (Dimitrova & Matthes, 2018, p333). More research is clearly required to understand how (mis)information travels through social media channels. Tracking the geographical

spread of such communications is complicated by low rates of coordinate-geotagging (Section 2.6, p77) and imprecise locational references in other metadata fields (e.g., user place of registration) sometimes present in OSN data.

As Martin (2010, pIX) has stated, our questioning of where knowledge comes from and how we gain ‘belief’ in it is something that ‘Philosophers have been thinking about [...] and arguing with each other about [...] for at least two thousand years.’ Boyd & Crawford (2012, p662) have argued that ‘Given the rise of Big Data as a socio-technical phenomenon [...] it is necessary to critically interrogate its assumptions and biases.’ These arise from the ‘interplay of technology, analysis and mythology’ in which increased computational power and ‘algorithmic accuracy’ have led to a ‘widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy’ (Boyd & Crawford, 2012, p663). The authors suggest that these developments pose six ‘critical questions’ for Big Data research, reproduced in bold text below, alongside a precis of their arguments and select quotes:

1. **Big Data changes the definition of knowledge** – Just as ‘Fordism’ changed the nature of manufacturing and work the availability of Big Data has produced a ‘radical shift in how we think about research’ raising profound epistemological and ethical concerns, including ‘how we should engage with information, and the nature and the categorization of reality.’ Algorithmic certainty (also criticised by Mittelstadt et al., 2016) should be avoided, alongside any adoption of the ‘end of theory’ ideas professed by Anderson (2008), who has argued that if ‘we can track and measure [...] why people do what they do [...] with unprecedented fidelity [then with] enough data, the numbers [will] speak for themselves’ without the need for theory.
2. **Claims to objectivity and accuracy are misleading** – Web-sourced data sets, which are ‘often unreliable, prone to outages and losses’ should not be

treated inviolably. The 'mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts' leads to inflated claims of objectivity and accuracy which are implausible as Big Data analysis 'is still subjective' and human behaviours expressed in social media messages cannot be reduced to quantitative certainties. Cresswell (2014) has also stated that this 'culture of numbers' is dangerous, affecting institutional decisions to fund numerical research in the humanities as 'soft' subjects attempt to become 'hard'.

3. **Bigger data are not always better data** – Giardullo (2015) has published a paper with a similar title, arguing (p529) that 'claims for the methodological power of bigger and bigger datasets, as well as increasing speed in analysis and data collection, are creating a real hype in social research.' This hype raises concerns, he argues, when the validity of research may be called into question if Big Data sources are the primary 'or (even worse) [the] unique source of information' used in such studies. Representativeness, 'wholeness' (Boyd & Crawford, 2012, p669) and sampling strategies may skew research findings and, even if Big Data are bigger, these problems do not necessarily make them better than more traditional, and tightly-controlled, fieldwork.
4. **Taken out of context, Big Data loses its meaning** – Here Boyd & Crawford (2012, p671) argue that 'Context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model.' A fixation on mathematically modelling traits of human behaviour, particularly 'social graph' elements recorded in Twitter retweet or Facebook friend data that broadly appear, but imprecisely equate, to real sociological network relationships (e.g., tie strength) fails to acknowledge that 'Not every connection is equivalent to every other connection, and neither does frequency of contact indicate strength of relationship.' Social networks do not provide a fully accurate reflection of societies or of all social processes, interactions and relationships which exist in real life.

5. **Just because it is accessible does not make it ethical** – The recent Cambridge Analytica scandal (Albright, 2017; Cadwalladr & Graham-Harrison, 2018; Persily, 2017) has added fuel to the fire identified by Boyd & Crawford's (2012) earlier reporting of the somewhat surreptitious monitoring of 1,700 college students by K. Lewis, Kaufman, Gonzalez, Wimmer, & Christakis (2008), itself later eclipsed by the 'Facebook experiment' (BBC News, 2014), in which researchers sought to model 'massive-scale emotional contagion' by actively manipulating 'the extent to which people (N = 689,003) were exposed to emotional expressions in their News Feed' (Kramer, Guillory, & Hancock, 2014). Issues of privacy, trust and consent are involved and are discussed in more detail in Section 3.4 (p111).
6. **Limited access to Big Data creates new digital divides** – The sixth 'critical question' for Big Data research involves access and divides, as Boyd & Crawford (2012, p673) note that 'Much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data' before asking 'But who gets access? For what purposes? In what contexts? And with what constraints?' These issues have more recently been covered by Zelenkauskaitė & Bucy (2016) who identify a dangerous trend for 'unattainable scholarship' as only those with the funds, and Big Data skill sets, are able to purchase, access, store and analyse 'digital traces of human behavior that are available online.'

Boyd & Crawford's (2012) work has been influential, with over 1,100 CrossRef citations, 44 of which are recorded in the research literature corpus. These epistemological and ethical concerns, which overlap significantly with various methodological issues identified in the following section, frame this research. Digital data deposited on, and sampled from, OSN platforms clearly exhibit several profound questions of believability, and the use of these data sets in research raises yet more important questions regarding contextualisation, over-quantification, privacy, consent, ethics and the 'problematic ontological and epistemological claims

[of] datafication [in which] (meta)data have become a regular *currency* for citizens to pay for their communication services and security' (Van Dijck, 2014, pp197-198). These issues form an area of active and ongoing research (Barassi, 2016; Fuchs, 2017a), and are discussed in more detail below and in Section 3.4 (p111) of this thesis.

3.3 Methodology

This research adopts a methodology which Andrienko, Andrienko, & Gatalsky (2003, p503) describe as an 'exploratory analysis' approach to visualising and understanding spatiotemporal data, combined with 'the robust qualitative case history approaches of human geography' (Clark & Jones, 2013, p312). Although all OSN messages are inherently discursive, and hence qualitative in nature, associated metadata and the large number of records in the case study data sets demand a largely quantitative, computerised, set of research methods. Consequently, analytical processes (Chapter 4, p118) and subsequent results (Chapter 5, p186 and Chapter 6, p227) may be criticised through choice of methodology, method or both. The 'intellectual weakness' stemming from 'the wave of super-positivism and the mania for quantification which swept all the social sciences in the nineteen-sixties' was identified by Massey in her Introduction to *Geography matters!* (Massey & Allen, 1984, p2). More recent reviews (Ceron, Curini, Iacus, & Porro, 2014; Ceron & Memoli, 2016; Darmon, Omodei, & Garland, 2014; Iacus, 2014; Tufekci, 2014) identify similar concerns in the current 'Gold Rush' (Felt, 2016; Tsou, 2015) towards a Big Data-driven, quantitatively-based, analytical future.

Tufekci (2014), in an expansive article, has outlined several *Big questions for social media big data*, which include *Representativeness, validity and other methodological pitfalls*. She concludes by stating (Tufekci, 2014, p513) that 'Social media big data is a powerful addition to the scientific toolkit. However, this emergent field needs to be placed on [a] firmer methodological and conceptual footing. Meaning of social media imprints, context of human communications, and

nature of socio-cultural interactions are multi-faceted and complex. People's behavior differs in significant dimensions from other objects of network analyses.' Tufekci's (2014) methodological pitfalls in 'social media big data' research cover five areas which are discussed, with some quotations from her paper under the original headings, reproduced in bold type, below:

1. **Model Organisms and Research: Twitter as the Field's *Drosophila Melanogaster*** – Twitter, in social media analysis, has become analogous to biologists' experimentation using fast-breeding (and cheap-to-use) *Drosophila melanogaster* fruit flies. Twitter is chosen 'mostly due to availability of data, tools and ease of analysis' but also because most data from the other major English-language OSN, Facebook, is not available publicly owing to tighter default or user-optional privacy settings.
2. **Hashtag Analyses, Selecting on the Dependent Variable, Selection Effects and User Choices** – Filtering on established Twitter hashtags (e.g., #Obama) 'select[s] on a dependent variable, and hence display[s] the concomitant features and weaknesses of this methodological path [...as...] inclusion of a case in a sample depends on the very variable being examined.'
3. **The Missing Denominator: We Know Who Clicked But We Don't Know Who Saw Or Could** – 'One of the biggest methodological dangers of big data analyses is insufficient understanding of the denominator. It's not enough to know how many people have "liked" a Facebook status update, clicked on a link, or "retweeted" a message without knowing how many people saw the item and chose not to take any action.'
4. **Missing the Ecology for the Platform** – 'Most existing big data analyses of social media are confined to a single platform (often Twitter, as discussed.) However, most of the topics of interest in such studies, such as influence or information flow, can rarely be confined to the Internet, let alone to a single platform [as information] in human affairs flows through all available channels.'

5. **Inferences and Interpretations** – ‘The question of inference from analyses of social media big data remains underconceptualized and underexamined. What’s a click? What does a retweet mean? In what context? By whom? How do different communities interpret these interactions? As with all human activities, interpreting online imprints engages layers of complexity.’

An awareness of these methodological problems, Tufekci (2014) suggests, ‘should be incorporated into the review process and go beyond soliciting “limitations” sections’ in research work; hence, these issues are covered here in some depth. In response to Tufekci’s observations, and following the same numbering system, it should be stressed that:

1. **The current study examines both Twitter *and* Facebook data** – Twitter provides the bulk (~90%) of the OSN data under examination. Facebook data does, however, offer a useful counterpoint. Facebook message text is much longer, and metadata differs substantially. Differences between the two OSN data sources are examined in Chapter 4 (p118) and Chapter 5 (p186).
2. **Hashtag filtering has *not* been used** – A wide range of search terms, rather than one or two hashtags, have been used as filters (Appendix 7, p432) in both case studies (Section 4.2.4, p126). Also Twitter’s Firehose, accessed through DataSift, has been used for data collection rather than the free and much more commonly used 1% Streaming API. Even though the selection of terms is much wider than ‘hashtag filtering’, there is still a danger of ‘selecting on dependent variables’. In practice, as the data purchase and technical cost of consuming *all* OSN data surrounding any given event would be enormous, filtering of some sort is an inevitable feature of this methodology.
3. **‘The missing denominator’ is missing**– It is undeniably true that there remain great difficulties in knowing how many people have looked at Twitter tweets or Facebook posts and chosen not to retweet, or like, the

content. Currently, re-weighting retweets or likes relative to impressions in this way is impossible as the pageview data which would enable this analysis is not released by any OSN platform operator.

4. **Examining the ecology surrounding the platform** – The geographicality of interactions made during electoral events are considered here. It is apparent that the surrounding media or ‘information ecology’ significantly affects OSN traffic, as evidenced, e.g., by the three spikes in OSN posts accompanying the televised 2012 US Presidential Candidate Debates (Figure 1-3, p25). Conventional media both influences OSN traffic and is influenced by it; this research considers the wider ecologies surrounding these bi-directional effects.
5. **Suggesting inferences and interpretations** – Tufekci points out that ‘As with all human activities, interpreting online imprints engages layers of complexity.’ This research, drawing on a wide range of literature with interdisciplinary political, communications, geographical and technical themes (Chapter 2, p51), suggests several inferences and interpretations. As Eisenhardt (1989, p532) has noted, the case study ‘research approach is especially appropriate in new topic areas’ where ‘the tie to actual data [may permit] the development of testable, relevant, and valid theory.’

The data under investigation are, at once, both qualitative (text messages, links, linked text) and quantitative, featuring many associated metadata fields of varying importance. As a product of Big Data, the ~8 million OSN messages collected are simply too voluminous to analyse individually. Consequently, a hybridised case study/exploratory analysis methodology, making full use of modern computerised techniques, has been adopted. This is described in the following sections.

3.3.1 Case study methodology

Labaree (2017) states that ‘A case study is an in-depth study of a particular research problem rather than a sweeping statistical survey or comprehensive comparative

inquiry. It is often used to narrow down a very broad field of research into one or a few easily researchable examples. The case study research design is also useful for testing whether a specific theory and model actually applies to phenomena in the real world. It is a useful design when not much is known about an issue or phenomenon.' Currently, 'not much is known about [the] phenomenon' (Labaree, 2017) under investigation; the differential usage of 'space' and 'place' in politicised social media discourse (Chapter 2, p51).

Gerring (2006) highlights the difference between 'cross-case', and 'within case' (or 'case study'), methodologies. In order to learn how to build a house, Gerring (2006, p1) explains, one could 'study the construction of many houses – perhaps a large subdivision or even hundreds of thousands of houses [or] one might study the construction of a particular house.' When applied to the social sciences, the latter, case study-based approach, enables researchers to study 'a few cases more intensively' rather than 'observ[ing] lots of cases superficially.' Social science research work of this type 'rests implicitly on the existence of a micro-macro link in social behaviour', which may take several forms (Gerring, 2006), e.g., 'For anthropologists and sociologists, the key unit is often the social group (family, ethnic group, village, religious group, etc.). For psychologists, it is usually the individual. For economists, it may be the individual, the firm, or some larger agglomeration. For political scientists, the topic is often nation-states, regions, organizations, statutes, or elections.'

In this research, the two case studies chosen (Section 4.2.4, p126) focus on elections taking place in nation states; the US and Scotland. As political communications now transcend national boundaries (Agnew, 2013; Ó Tuathail, 1998) no geographical boundaries have been enforced during case study data collection, although filtering on language (English, almost exclusively) has been applied. The unit of study is the 'interaction', in the form of a Twitter tweet or Facebook post, and the 'micro-macro link' proceeds from interactions, to users, to

places (e.g., identifiable towns or cities), to place ‘agglomerations’ (e.g., states, constituencies or, even, time zones) and ultimately to countries and the entire world.

3.3.2 Exploratory methodology

Labaree (2017), states that an exploratory design is especially useful when ‘there are few or no earlier studies to refer to or rely upon to predict an outcome’ and that this approach may be used to ‘establish an understanding of how best to proceed in studying an issue or what methodology would effectively apply to gathering information about the issue.’ Exploratory research methodologies can provide a ‘well grounded picture of the situation being developed’ leading to the ‘generation of new ideas and assumptions’ which can be used to develop ‘tentative theories or hypotheses’ (Labaree, 2017) in a process of refinement which may help determine the feasibility of future studies and directions for future research.

In a spatiotemporal GIScience context, the work of Gennady and Natala Andrienko and collaborators has proven particularly instructive (G. Andrienko et al., 2013, 2010; G. Andrienko, Andrienko, Fuchs, & Wood, 2017; G. Andrienko, Andrienko, & Wrobel, 2007; N. Andrienko, Andrienko, Fuchs, Rinzivillo, & Betz, 2015; N. Andrienko et al., 2003; Keim et al., 2008). These authors have focused on ‘exploratory spatio-temporal’ visualisation, or analytical, approaches, noting (N. Andrienko et al., 2003, p504) that ‘Modern computer technologies provide better than ever before opportunities for storage, management, visualization, and analysis of dynamic, i.e. temporally variable, data, including dynamic spatial data (further referred to as spatio-temporal data).’ Building on the work of Peuquet (1994) in identifying ‘three components [of spatio-temporal] data: space (*where*), time (*when*) and objects (*what*)’ N. Andrienko et al. (2003, p509) propose a classification scheme, ‘taking time as [its] focus’ in which:

1. Time is [a] given while other types of information (objects, locations, properties, relationships) need to be discovered and described. We shall schematically designate this type of task as *when* → *where* + *what*.
2. Time needs to be discovered for given information of other types. This type of task will be further designated as *where* + *what* → *when*.

This framework has proven particularly useful in exploring spatiotemporal OSN data, which encodes locations in various ways (geotags, toponyms etc.) and features many types of ‘objects’ (interactions, users etc.) together with many potential spatial and/or temporal aggregations ranging from locations at different scales to the passing of time measured in minutes, hours or weeks; all of which may be accompanied by shorter or longer-phased political events. As the amount of text in the OSN corpus is vast, at over 230 million space-tokenised words, a further exploratory technique involves the use of NLP software (Section 4.4.1, p147), and associated database work, to make sense of a very large volume of free-form text.

3.3.3 Hybrid case study/exploratory methodology

The methodology used in this research is a hybrid of the case study and exploratory approaches discussed above. Two case study events (Section 4.2.4, p126) provide data. ‘Exploratory spatio-temporal visualisation’ and ‘visual analytics’ methods (G. Andrienko et al., 2010) provide several valuable computer-driven and, in some cases, GIScience-specific research techniques. These have been supplemented by Natural Language Processing pipelines (GATEcloud, AlchemyAPI and CLAVIN-rest, discussed in Chapter 4, p118), newly available ‘at Web scale’ (Berners-Lee, Hall, Hendler, O’Hara, et al., 2006), and able to operate on, text-mine and extract meaningful information from millions or even billions of records (Tablan et al.,

2012). Labaree (2017) usefully summarises what case study and exploratory methodologies can (Table 3-1) and cannot (Table 3-2, p110) tell us.

Table 3-1 – What case study and exploratory methodologies can tell us (after Labaree, 2017; quoted in italics)

Case study methodology	Exploratory methodology
<i>Approach excels at bringing us to an understanding of a complex issue through detailed contextual analysis of a limited number of events or conditions and their relationships.</i>	<i>A useful approach for gaining background information on a particular topic.</i>
A researcher using a case study design can <i>apply a variety of methodologies and rely on a variety of sources</i> to investigate a research problem.	<i>Exploratory research is flexible</i> and can address research questions of all types (what, why, how).
<i>Can extend experience or add strength</i> to what is already known through previous research.	<i>Provides an opportunity to define new terms and clarify existing concepts.</i>
Social scientists make wide use of this methodology to <i>examine contemporary real-life situations</i> and provide the basis for the application of concepts and theories.	<i>Exploratory research is often used to generate formal hypotheses and develop more precise research problems</i> by identifying patterns or irregularities in data hitherto unsuspected.
<i>Can provide detailed descriptions of specific and rare cases.</i>	In the policy arena or applied to practice, <i>exploratory studies help establish research priorities</i> and where resources should be allocated.

These summations are tabulated above and below, although it must be noted that Labaree did not juxtapose his conclusions in a comparative row/column-based table of this type, instead preferring bulleted lists of points. Strengths and weaknesses inherent in the two approaches from his original text have been italicised for emphasis.

Table 3-2 – What case study and exploratory methodologies cannot tell us (after Labaree, 2017; quoted in italics)

Case study methodology	Exploratory methodology
<i>A single or small number of cases offers little basis for establishing reliability or to generalize the findings to a wider population of people, places, or things.</i>	Exploratory research <i>generally utilizes small sample sizes and, thus, findings are typically not generalizable to the population at large.</i> The availability of Big Data may well have changed this assumption.
<i>Intense exposure to the study of a case may bias a researcher's interpretation of the findings.</i>	The exploratory nature of the research <i>inhibits an ability to make definitive conclusions about the findings.</i> They provide insight but not definitive conclusions.
<i>Design does not facilitate assessment of cause and effect relationships.</i>	The research process underpinning exploratory studies is <i>flexible but often unstructured, leading to only tentative results that have limited value to decision-makers.</i>
<i>Vital information may be missing in any given case; this may make the case hard to interpret and add to difficulties in making generalisable conclusions.</i>	<i>Design lacks rigorous standards applied to methods of data gathering and analysis</i> because one of the areas for exploration could be to determine what method or methodologies could best fit the research problem. More systematic approaches to exploration help mitigate this problem.
<i>The case may not be representative or typical of the larger problem being investigated and, unless more cases are studied, it may be hard to know this.</i>	
<i>If the criteria for selecting a case is because it represents a very unusual or unique phenomenon or problem for study, then interpretation of the findings can only apply to that particular case.</i>	

Table 3-1 (p109) indicates that a hybrid case study/exploratory approach has many advantages, including the ability to understand complex issues in a flexible way, helping to clarify existing concepts and frame new directions for research. Where Labaree (2017) identifies weaknesses, these tend to be in terms of case study selection, unrepresentativeness (a definite possibility) and 'small sample sizes' (less relevant here as many millions of records have been sampled), all of which may limit the methodology's ability to offer 'definitive findings'. These observations match several of Tufekci's (2014) points above (Section 3.3, p102), as applied to the study of 'social media big data'. An awareness of these potential shortcomings in the research methodology is acknowledged. As any deficiencies in the methodological approach may also stem from ethical concerns or practice these issues are discussed in the following section.

3.4 Ethics

None of the 2,436,167 social media users whose 8,196,380 messages have been analysed here gave explicit consent to take part in this research and contacting all of them to request their co-operation would be practically impossible. What ethical issues arise from the collection, download, storage and analysis of so many messages publicly posted by so many users of two leading OSN sites? Has a new paradigm (Kuhn, 1970) of implicit co-option and tacit co-operation in research been reached?

Users who have 'opted-in' to, or failed to 'opt-out' of, public posting on OSNs are sometimes surprised by the amount of data deposited in the public domain and how these data are used. Tear & Southall (2019, in press) identify how 'modern social media enables [sophisticated communications] at no upfront cost to its users, who have, thus far, made a [...] bargain by imparting their own personal information for access to these platforms, accepting increasingly targeted advertising in exchange.' This process of 'datafication', through which Web users accept and acknowledge that their data (and its metadata) have value, and use this

as a 'currency' in exchange for services, has also been noted by Van Dijck (2014). In the wake of Edward Snowden's revelations on the extent to which government surveillance agencies make use of social media, telephony and other 'digital traces' of modern-day life (The Guardian, 2018) issues surrounding 'datafication', 'dataism' (a belief in objective, quantitative measurement and prediction from data) and perpetual state-organised 'dataveillance' have come to the fore.

Van Dijck (2014, p206) has stated that 'The popularization of datafication as a neutral paradigm, carried by a belief in dataism and supported by institutional guardians of trust' raises several 'precarious matters' which are still to be addressed. These centre around society's relationship with democracy and 'dataveillance' by state actors but also require that '[academics take] responsibility for maintaining credibility of the [data] ecosystem as a whole.' In exhibiting sometimes 'unbridled enthusiasm' for using the by-products of datafication in research, based on a generally misplaced epistemological belief in 'objective quantified understanding', Van Dijck (2014, p204) recommends that 'To keep and maintain trust, Big Data researchers need to identify the partial perspectives from which data are analyzed; rather than maintaining claims to neutrality, they ought to account for the context in which data sets are generated and pair off quantitative methodologies with qualitative questions.' This is the approach that has been adopted in this study which, against the complex multi-media backdrop of two major political campaigns, merges quantitative and numerical reporting with examination and identification of more qualitative expressions of place evident in social media messages and link shares. The research has, of course, been ethically reviewed (Appendix 4, p419) and a summary of this process, and issues highlighted by it, follows.

In his email of 29/05/2015, reporting a 'Favourable opinion with conditions', the University of Portsmouth's Geography Department Ethics Co-ordinator (Bray, personal communication, 2015), noted that this research project 'is an unusual

study for which relatively little clear research ethics exists.’ This observation has been echoed in the literature. The University of Cardiff’s, Economic and Social Research Council (ESRC) funded, Social Data Science Laboratory (2016) has stated that ethics have become a particularly ‘salient’ feature of research using social media data as ‘The digital revolution has outpaced parallel developments in research governance and agreed good practice. Codes of ethical conduct that were written in the mid twentieth century are being relied upon to guide the collection, analysis and representation of digital data in the twenty-first century.’

Examples of ethical challenges abound. Swirsky, Hoop, & Labott (2014, p60) give examples in which ‘the investigator wonders whether the requirement for informed consent can be waived because viewing publicly accessible Facebook pages is akin to observing public behavior’ and conclude that this ‘may not be the case.’ Users, it is noted, ‘may feel that their Facebook page, even if publicly accessible, is still somewhat private.’ Moreno, Goniou, Moreno, & Diekema (2013, p708) consider several ethical dilemmas in ‘observational research, interactive research, and survey/interview research’, highlighting problems with ‘issues regarding privacy, consent, and confidentiality.’

From a geographical perspective, there also appears to be a particularly strong desire amongst OSN users, identified by both academic (Barreneche & Wilken, 2015; Cottrill, 2011) and non-academic contributors (Schwartz, 2013), to protect personal locational privacy. Utter dismay has accompanied several real or perceived breaches of locational trust. Cottrill (2011, p49), for example, describes the “Locationgate” scandal when, in ‘April of 2011, at Where 2.0 (a conference focusing in part on location-aware technology) [it was disclosed] that the Apple iPhone and 3G iPad were recording the locations of the devices, unencrypted, to a hidden file.’ As a result, hardware manufacturers, Facebook, Twitter and other Web platforms such as eBay – which may once inadvertently have disseminated geospatial information recorded in the Exchangeable Image File (EXIF) metadata of publicly-

posted digital images from users listing items for sale (Schwartz, 2013) – now place significant importance on locational ‘safeguarding’ (Lazer et al., 2009; Min & Kim, 2015).

Twitter’s *Geo Guidelines for Developers* (Twitter, 2014), for example, require that:

- Users must opt-in to use the Tweeting with Location feature (turn location “on”).
- Users must give explicit permission for their exact location to be displayed with their Tweets.
- It must be clear to users what level of location information, if any, will be displayed in association with their Tweet.
- Users should be able to turn on and off their location each time they compose a Tweet.

Taking the above into account, none of the maps or charts presented in this thesis are at large enough scale to identify individual users’ locations or addresses. In software, however, it is perfectly possible to zoom in to specific coordinate-geotagged Twitter tweets or Facebook posts, or to identify and map all of the locations from which any given user has deposited social media messages. While it is possible, it is unethical to report on the data in this way. Consequently, and in line with both the University of Portsmouth’s ethical review and general ‘good practice’ recommended in this research area (Williams, 2015), in this study:

- Licensing restrictions imposed by Twitter, Facebook and DataSift (Appendix 5, p424) are respected.
- Research outputs (Chapter 5, p186) report at aggregate levels.
- No quotes from identifiable social media users are used, unless those users are public figures (e.g., Obama, Salmond).

- No data gathered in the research are shared with others, except for project partners (e.g., by transmission to secure servers operated by the University of Sheffield's GATEcloud.net service).
- All data are always stored in secure, password-protected systems.

The ethical approach adopted here *does limit* the presentation of results. Research outputs are based upon aggregated analyses of atomic data, and no individual messages (other than those posted by public figures such as electoral candidates or major celebrities) are reported. These recommendations, concepts and limitations surrounding Internet Mediated Research (IMR) were discussed at the University of Portsmouth's Ethics and Governance conference (Sugira, Carpenter, Evans, & Parry, 2016). At this event, the outgoing chair of the University's Ethics Committee, David Carpenter, self-deprecatingly asked whether IMR prompts 'the end of research ethics as we know it?' Answering his own question, Carpenter's assertion that a 'paradigm shift' has occurred in the face of new-found abilities to access and interrogate social media Big Data has been affirmed extensively elsewhere in the literature (Borah, 2017; R. M. Chang et al., 2014; Deluliis, 2015; Fuchs, 2017a; S. Li et al., 2016; Rowe, 2015; Wei, 2013; Zelenkauskaitė & Bucy, 2016).

Ethical considerations inevitably influence the conduct of research and the presentation of results. Fuchs (2017, p43) has warned that analytical positivism in Big Data research frequently 'results in often very superficial analyses that highlight major topics, users or social relations in large amounts of data gathered from Twitter, Facebook and other social media platforms.' Fuchs (2017a) has stated that 'the trouble' with Big Data analytics is 'that it often does not connect statistical and computational research results to a broader analysis of human meanings, interpretations, experiences, attitudes, moral values, ethical dilemmas, uses, contradictions and macro-sociological implications of social media.' This, he suggests, requires a rethink 'about theoretical (ontological), methodological (epistemological) and ethical dimensions of an alternative paradigm.'

The critical approach which Fuchs (2017a) advocates is designed not just to understand ‘what people do on the Internet but also why they do it’, ideally by incorporating ‘traditional sociological methods’ including interviews, observation, surveys, statistical analysis of secondary data and so forth. Doing so, while admirable, requires larger budgets, more time and the sort of co-option onto survey panels which has not been attempted here, and remains an area for future research (Section 7.5, p299). Fusion of OSN data with other secondary sources of statistical data, including US and UK Census data has, however, been conducted and is reported upon as an additional finding in Section 6.4.4 (p262). Wherever possible a digital positivist methodology is avoided in this research which considers content, and context, in preference to (meta)data alone. The profound and limiting epistemological, methodological and ethical issues which frame the research, and have been outlined above, are returned to in the concluding chapter (Section 7.4, p297) of this thesis.

3.5 Summary

In the Internet era politics, and to a degree life itself, is arguably becoming increasingly ‘deterritorialized’ (after Deleuze & Guattari, 1972 reprinted in translation 2004; cf. Ó Tuathail, 1998), while at elections political outcomes are still based around geographically bound constituencies (Agnew, 2013; Elden, 2005). As Johnston has noted in a review (2009, p511) of Rehfield's (2005) work *The Concept of Constituency* ‘For electoral and political geographers in almost all of the English-speaking world, the role of territorially-defined constituencies in legislative elections is virtually taken-for-granted.’ This territorial definition, or ‘boundedness’ of space, has held a long-running fascination for geographers (Cox, 1969; Giddens, 1985) and has continuing relevance during a period when candidates, political parties, companies and state-sponsored actors have attempted to influence opinion online, particularly in the small and (now easily) targetable number of

constituencies or voting districts which often determine political outcomes in Western democracies.

The recent exponential growth in OSNs, coupled with improved access to OSN Big Data and developments in computing technology, have enabled new forms of social science research, including the analysis of geographicality in politicised social media discourse set out in this thesis. Vergeer (2012, p12) has suggested that where theory-driven, smaller-scale socio-political and larger-scale information science approaches intersect is 'where scientific innovations will most likely surface.' The following chapter details methods used in this study to innovate in this way. Results from this work are presented in Chapter 5 (p186) with additional findings from the research given in Chapter 6 (p227).

4 RESEARCH METHODS

4.1 Introduction

This chapter describes the research methods employed to acquire (using several randomised 1-in- n samples), store, augment, query, tabulate, analyse and visualise ~8 million OSN interactions recorded during the 2012 US Presidential Election and the 2014 Scottish Independence Referendum campaigns. The chapter, after Kallet (2004), describes, a) what was done, b) how it was done, c) justifies the experimental design, and; d) explains how results were analysed.

Five main sections detail the methods adopted in this research, describing:

1. **Data subjects** – Section 4.2 (119) describes the material used in this research, consisting of data subjects in the form of digital social media messages (or ‘interactions’) and accompanying metadata created by ~2.4m users of Twitter and Facebook during two case study events; the 2012 US Presidential Election and the 2014 Scottish Independence Referendum.
2. **Data preparations** – Section 4.3 (p135) describes the preparations required to render the files of collected social media interaction data usable, through storage (and the testing of different storage technologies) allowing for the efficient querying and interrogation of the data. The several file formats and database management systems used are introduced.
3. **Data procedures** – Section 4.4 (p147) describes the procedures adopted to augment the collected data using three Natural Language Processing (NLP) systems, each of which was used to text-mine interaction message text and, in one case, linked/shared URL content deposited alongside message text for place-based (toponymic) geographical references.
4. **Data analysis** – Section 4.5 (p158) describes the set of methods used to analyse the data subjects, including data query, tabulation and analysis, data visualisation and statistical tests. The software systems used to perform

these tasks are introduced and, through reference to Appendices, the computing environment used to perform the analysis is described.

- 5. Data measurements** – Section 4.6 (p164) describes the development of a measurement and scoring system used here to categorise baseline levels of ‘geographicality’ in social media metadata. Geographicality Scores, based on the presence of identified Potential Geographic Information (PGI) in interaction metadata, aid cross-comparison when presenting results.

The various, largely technical, methods detailed in this chapter are used to provide answers to the three research questions outlined in Section 1.7 (p34); these results are presented in the following Chapter 5 (p186) with additional relevant findings detailed and discussed in Chapter 6 (p227).

4.2 Data subjects

The subjects of this research are politically discursive social media messages; 8,196,380 OSN interactions created by 2,436,167 individual users of Twitter and Facebook in a roughly 90:10 ratio during the campaigns leading up to the 2012 US Presidential Election (US2012) and the 2014 Scottish Independence Referendum (SCOT2014). This section outlines the characteristics of social media data (Section 4.2.1, p119) and several of the methods and technologies used in Social Network Analysis (SNA; Section 4.2.2, p121) before detailing the technical ‘proof of concept’ exercise undertaken first in this research programme (Section 4.2.3, p123). The conduct of this test, along with outputs and analyses derived from it (given in Appendix 6, p427), led to the selection of case studies (Section 4.2.4, p126) and the data acquisition phase (Section 4.2.5, p134) of this research.

4.2.1 Characteristics of social media data

It is generally difficult to consume entire streams of social media data; the characteristic 3 Vs of ‘volume, velocity and variety’ (Laney, 2001) make ingestion

and storage of real-time feeds problematic without high-end IT infrastructures. Consequently, and because many OSN sites are global in scope, subsetting is common prior to analysis. The subsetting of social media data may be achieved temporally or by filtering against available text, location, language or metadata fields (e.g., image or video descriptions). Filtering may also exploit characteristics of the ‘social graph’ at the heart of many OSN sites.

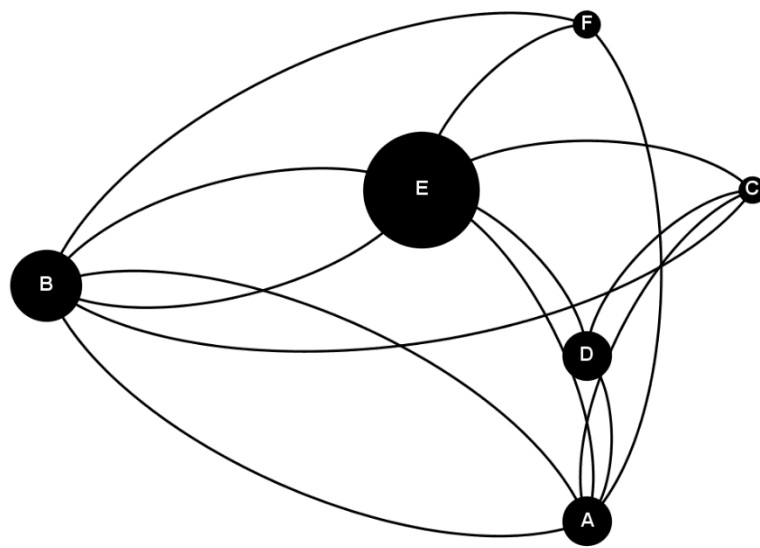


Figure 4-1 – Simple representation of a Social Graph: 6 users A-F (‘nodes’) are connected to one another (by ‘links’) with node-size proportional to ‘out-degree’ (number of outbound links)

Figure 4-1 shows a conceptualised set of inter-relationships between imaginary users (e.g., A knows B and C; B:A,E,F; C:B; D:A,C; E:A,B,C,D,F; F:A) rendered as a spatialised graph in the Gephi software package (Bastian, Heymann, & Jacomy, 2009). This type of visualisation is based upon Euler’s puzzle regarding possible walking routes over the seven bridges of Königsberg (Gribkovskaia, Halskau, & Laporte, 2007; J. R. Newman, 1953; Shields, 2012) which led to the foundation of

graph theory in mathematics (I. Robinson, Webber, & Eifrem, 2015). Interactions between users form the mainstay of the workings of all OSN sites; e.g., on Facebook you find friends, on LinkedIn you connect to fellow professionals, on Twitter you follow or mention other users and may retweet their messages.

In the current research, subsetting has been achieved through temporal cut-offs and text-matching against key terms, with some additional filtering on language. In one case (US2012; Section 4.2.4.1, p126) the scale of the event and the sampling strategy adopted precluded the collection of an entire social graph. In the other (SCOT2014; Section 4.2.4.2, p129), a much larger data set and complete graph was collected for the selected key terms. These data acquisition events are described later, following a brief description of how social media data from Twitter and Facebook may be stored and analysed. The introductory overview given in Section 4.2.2, below, is expanded upon in subsequent sections of this chapter.

4.2.2 Social Network Analysis (SNA)

Social Network Analysis (SNA) primarily focuses on understanding interaction relationships in social media data. Computerised systems are required, commonly including:

- **NoSQL and/or SQL relational databases** – Social media data are typically ‘semi-structured’ comprising variable length sets of key/value pairs frequently with nesting and arrays held in JSON format (ECMA International, 2013, 2017). NoSQL (not only Structured Query Language) databases (e.g., Apache Drill running on Hadoop, MongoDB) are well-suited to ‘ingesting’ such data, although conventional relational databases (e.g., MySQL, Oracle, PostgreSQL) now offer significantly-improved JSON storage and query facilities, using familiar SQL syntax, and are perfectly well-suited to querying the generally more-structured elements, e.g., username, user home page, creation date/time etc. held in OSN metadata. File formats encountered and

database management systems used in this research are detailed in Section 4.3 (p135).

- **Natural Language Processing (NLP) systems** – Most non-audio/visual social media data is comprised of free-form ‘unstructured’ text. As data volumes preclude individual examination of text, computerised systems must be used. Social media data, and Twitter tweets especially, often feature terse and ungrammatical language (Batista & Figueira, 2017) making Information Extraction (IE) and Named Entity Recognition (NER) difficult. The University of Sheffield’s open-source GATE and GATEcloud instances of TwitIE (Maynard, Roberts, Greenwood, Rout, & Bontcheva, 2017) perform particularly well against Twitter data. Other software or systems (e.g., AlchemyAPI, Lexalytics, R’s Quanteda or TextMining packages) may be better-suited to longer text or to specific tasks (e.g., NER of shared links). The NLP systems used in this research are detailed in Section 4.4, (p147).
- **Graph databases and visualisation systems** – The ‘social graph’ may be stored and analysed in graph databases (e.g., Neo4j, Oracle Spatial & Graph) and visualised using specialist software (e.g., Gephi, Pajek). Social media users, in these notations, are termed ‘nodes’ and their inter-relationships ‘links’ or ‘vertices’. Attributes, e.g., liked, following, followed, mentioned, retweeted are attached to the vertices allowing extended traversal through graph networks. LinkedIn, the leading OSN aimed at professionals, runs on a custom-built graph database (Clemm, 2015) and is particularly good at finding Friend of a Friend (FOAF) and FOAF-FOAF-FOAF relationships out to several degrees. Likes on Facebook, together with retweets and mentions on Twitter, are commonly studied in the SNA literature (Scott, 2017). The range of systems and specialist software (below) used in this research are detailed in Section 4.5 (p158).
- **Specialist software** – An extensive array of open-source and commercial products are available, running either locally on desktop or server class

hardware or in the Cloud, to perform sentiment analysis, machine learning, image content classification, geographical visualisation and countless other analytical tasks. A comprehensive listing of the many methods and technologies employed in SNA is outside the scope of this thesis, but is covered in considerable depth by Scott (2017) and Russell (2011).

Massive usage growth in highly participatory online platforms and mobile-enabled applications is the defining characteristic of the modern-day Web as the ‘underlying dynamics of code and networking [...have enabled...] computer corporations [to take] over the media’s natural field or, at least, [diverge] into new forms of corporate consumer business’ (Allen, 2017, p177). Many of us now voluntarily create content online, either personally or professionally, and accept that it is stored digitally in corporate databases some of which allow access to publicly-posted material (Van Dijck, 2014). While content platforms have increasingly ‘locked down’ access to social graph relationships (Hogan, 2018) and more sophisticated privacy settings remove much data from public view, social media data (from Twitter, especially) are now readily accessible, e.g., using Social Feed Manager (George Washington University Libraries, 2016), and provide a fascinating source of research material for social and information scientists. The technical ‘proof of concept’ exercise undertaken first in this programme, designed to trial social media data collection and provide material for test analysis, is described below. Results from this exercise helped to inform subsequent choice of case studies (Section 4.2.4, p126) and methods used in this research.

4.2.3 Technical proof of concept

On 6 May 2012, a technical proof of concept exercise was undertaken to record OSN interactions made during the final stages of the 2012 French Presidential Election. France uses a ‘two-round runoff’ voting system to elect the President of the Republic and members of the National Assembly. In 2012, ten Presidential candidates, including the incumbent Nicolas Sarkozy (Union for a Popular

Movement), stood in the first round. Sarkozy and his Socialist opponent, François Hollande, collected the most votes (9,753,629 and 10,272,705 respectively) in the first-round of voting held on 22 April 2012, and proceeded to the second-round runoff contest set for 6 May 2012. The outcome of this election was hotly anticipated as the result in the first-round had been so close, with Sarkozy winning 27.18% of total votes cast and Hollande just edging him with 28.63%.

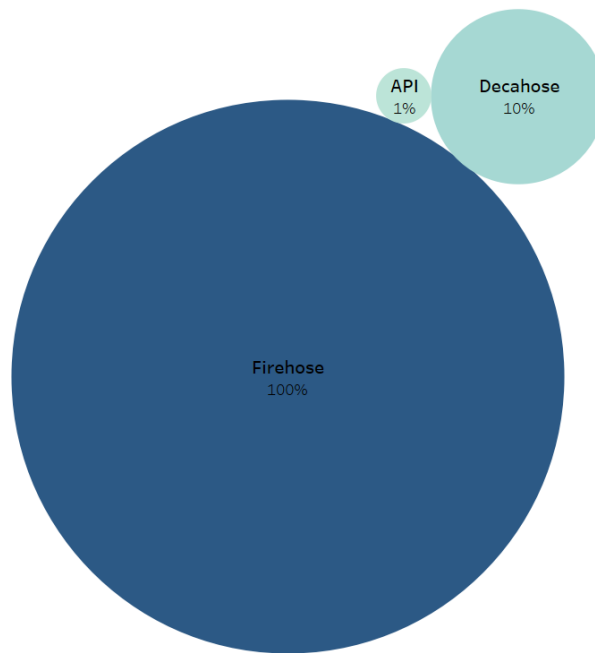


Figure 4-2 – Comparative size of Twitter’s Streaming API (1% sample), Decahose (10% sample) and the full Firehose (100% of tweets)

Around three hours before the result was due to be announced, test collection of social media data commenced using the DataSift platform, a content aggregation system capable of accessing Twitter’s ‘Firehose’ (Figure 4-2). Most studies in the published academic literature use Twitter’s ‘Streaming API’ to download tweets, a facility provided by Twitter since 2006 (Stone, 2006). This Application Programming Interface (API) allows developers to authenticate against Twitter’s servers (Twitter, 2013b, 2017) and stream, in real time, a 1% sample of the full ‘Firehose’ of tweets, nowadays equating to ~500 million tweets/day (Worldometers, 2018). As it is free to use and data volumes are manageable the Streaming API has been used

extensively in academic research (Deitrick & Hu, 2013; Elisa Omodei, Manlio De Domenico, & Alex Arenas, 2015; Kwon, Wang, Raymond, & Xu, 2015; Wachowicz & Liu, 2016). While the 1% sample provided by the Streaming API is thought to be representative of the full Firehose of tweets (Morstatter, Pfeffer, Liu, & Carley, 2013) access to the 10% Decahose or 100% Firehose is typically 'very hard to come by and potentially very expensive to realistically consume' (Twitter, 2013a). Consequently, many studies using the Streaming API have collected data over substantial time periods measured in months or, in some cases, years (S. Gray et al., 2015). This is especially the case where researchers filter on rarely used keyword terms or on the presence of certain rarely-populated interaction metadata, e.g., 'geotagged' coordinates.

In order to test the collection of a sufficient number of social media interactions quickly DataSift's (2013c) content aggregation service was used. In 2012 DataSift offered new users a limited free trial of its services, able to access both the full Firehose of Twitter tweets and messages posted on several other OSNs, including Facebook. DataSift managed upstream technical integration with platform operators, a useful feature as APIs change frequently (Claburn, 2018), and allowed ongoing 'pay as you go' access to social media data through its own easily understandable and programmable Curated Stream Definition Language, CSDL (DataSift, 2013a). As a test, a real-time recording was created on the DataSift platform using the CSDL statement below:

```
interaction.content CONTAINS_ANY "french election,  
presidential election, sarkozy, hollande"
```

The CSDL was designed to record OSN interactions with message text containing any of the case-insensitive phrases (e.g., 'french election') shown within double quotes above. The recording started on Sunday, 6 May 2012 at 16:17:47 and was stopped at 17:31:03 on the same day, some 1 hour, 13 minutes and 16 seconds later. This trial exercise, outputs and analyses of which are presented in Appendix 6 (p427), proved – through the collection of ~50,000 records in under an hour and a

quarter during the final stages of the 2012 French Presidential Election – that OSN interactions could be filtered, recorded, saved and downloaded using the DataSift platform at speed, in large volume and with controllable costs. Hollande was declared winner of the contest to become President of the French Republic, with 51.64% of the vote, at 8pm on 6 May 2012, some 2 hours 30 minutes after test data collection ceased.

4.2.4 Choice of case studies

Much larger and longer-running OSN recordings were required to test the *Geographicality Assumption* against the research questions set out earlier in this thesis (Section 1.7, p34). Two, then forthcoming, electoral events were selected as case studies for further investigation; these are detailed below.

4.2.4.1 Case study #1: 2012 US Presidential Election

During the two-month run up to the US Presidential Election of 6 November 2012, 1,661,402 Twitter tweets and 57,265 Facebook posts were sampled from contemporaneous OSN communications. Three sample sets ('Streams' in DataSift terminology) from Twitter and Facebook were recorded using a range of identical text search terms and controlling for explicit presence/absence of geographical coordinates, extent (country) and/or language. The interactions were filtered (Appendix A7.2, p432) on any case-insensitive words or phrases matching those illustrated in Figure 4-3 (p127), which shows the contribution of each search term to the sample, noting that some terms (e.g., 'US President' and 'Obama') may have appeared multiple times within message text. When sampled, using a 1-in- n strategy to record only a percentage of all Twitter tweets or retweets from the Firehose, DataSift's CSDL construct, `interaction.sample`, was used. This generates a 'floating-point random number between 0 and 100' for each interaction enabling control of sample size (DataSift, 2018a).

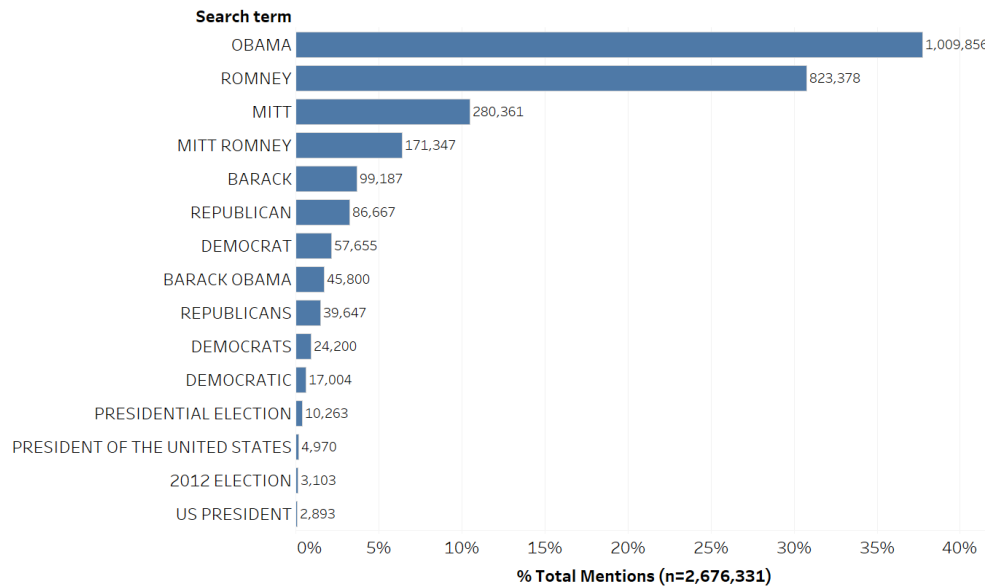


Figure 4-3 – US2012: Search terms used, numeric and percentage contribution to OSN interactions sampled (n=2,676,331 total mentions of search terms in text)

Despite filtering on 15 search terms the top 3 terms account for 78.97% of interactions sampled in the data set. Filtering has selected for inclusion mainly on candidate surname, forename or a combination of the two. In both political events (see also Figure 4-6, p131), the top 2 terms usefully, and reasonably evenly, select interactions for the major protagonists in both contests although, in retrospect, the addition of another two single-word terms ('President' for US2012 and 'Referendum' for SCOT2014) may have yielded somewhat differently-shaped data sets. As Tufekci (2014) has noted, choice of filter terms is clearly critical when abstracting OSN content (Section 3.2, p96). In this research, however, the two data sets are considered to demonstrate both a reasonable split between OSN sources and subtypes (Facebook, Twitter tweets, Twitter retweets) and a balance between opposing sides of the political debate in each election under investigation.

To collect many spatially referenced Twitter tweets or Facebook posts in 2012, and not knowing at the outset how many records (at what cost) would be captured, a 1 in 5 sample of explicitly geotagged content was recorded, along with a 1 in 50 sample unconstrained by explicit presence of geographical coordinates. This second

Stream was required to determine the overall proportion of OSN interactions containing coordinate-geotags and/or toponymic references. A third 1 in 50 sample considered posts in Spanish as the Hispanic population was thought at the time to comprise a key voting bloc in the election (Choy, Cheong, Laik, & Shung, 2012; Klofstad & Bishin, 2014).

Table 4-1 – US2012: Summary of recorded OSN interactions (n=1,718,667)

Details	US2012_GEO	US2012_NON_GEO	US2012_NON_GEO_HISP
Start Date	04/09/2012	06/09/2012	05/10/2012
End Date	06/11/2012	06/11/2012	06/11/2012
Duration	63 days	61 days	32 days
Sample	1 in 5	1 in 50	1 in 50
Coverage	Worldwide	Worldwide	United States
Language	English	English	Spanish
Geo Required?	Yes	No	No
Geotagged	146,424	22,424	122
Non-geotagged	0	1,538,543	11,154
Total	146,424	1,560,967	11,276
% Geotagged	100%	1.44%	1.08%

The three US2012 Streams (Table 4-1) were recorded, stored and downloaded from DataSift’s servers. The data set consists of 1,718,667 rows across three files each with up to 146 fields (Table 4-3, p136). The source files in both Comma Separated Values (CSV) and JavaScript Object Notation (JSON) formats were loaded into the Oracle 12c RDBMS (Section 4.3.1.3, p145). Altogether across the US2012 data set 168,970 Twitter tweet and retweet interactions, and no Facebook posts, were coordinate-geotagged; these geotagged records could easily be mapped (Figure 4-4, p129) using Tableau software (Section 4.5.2, p161). Most coordinate-geotagged interactions came, not unsurprisingly, from the US2012_GEO Stream which explicitly *required* coordinates to be present in metadata (Appendix A7.2.1, p433).

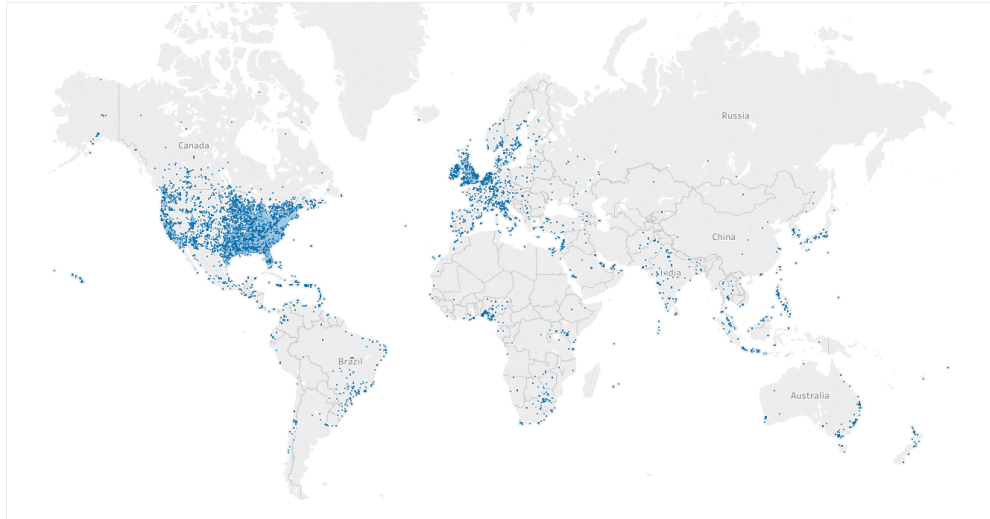


Figure 4-4 – US2012: Worldwide distribution of coordinate-geotagged Twitter tweets (dark blue) and retweets (lighter blue)

In this Stream 99.94% of records held a valid Latitude and Longitude coordinate pair, with 85 recording useless 0/0 coordinates. In those Streams without an explicit geographical filter only ~1% of records were geotagged. If the sampled data were grossed up ~750,000 records ($5 * \sim 150,000$) out of ~75 million ($50 * \sim 1.5$ million) filtered on the search terms used, i.e. ~1% (again in line with Leetaru et al., 2013), could easily be mapped using coordinate-geotags.

4.2.4.2 Case study #2: 2014 Scottish Independence Referendum

The necessity for a second case study was prompted by early analysis of data from the first. The US2012 data set consisted of three sampled Streams in 1:5 (one Stream) and 1:50 ratios (two Streams). Sampling was used to restrict data volume and control costs (Section 4.2.5, p134) but also resulted in an ‘incomplete’ data set where the full network graph of tweeting, mentioning and retweeting could not be examined. This can be illustrated visually (Figure 4-5, p130) and by running modularity class calculations (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) using the open-source Gephi graph analysis package (Bastian et al., 2009).

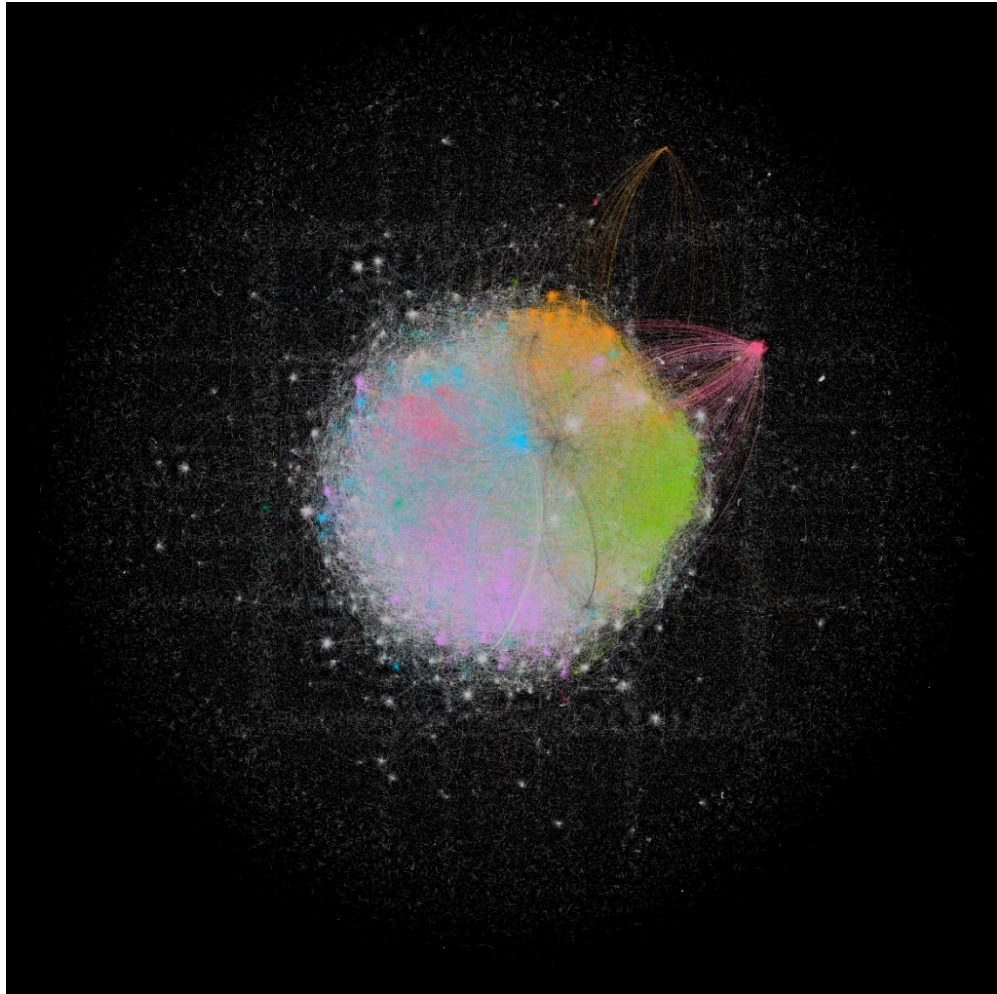


Figure 4-5 – US2012: Spatialised network graph of ‘Twitter mentions’ relationships

Figure 4-5 shows the ‘Twitter mentions’ network in the US2012 data set, visualising 314,427 ‘nodes’ (Twitter users) mentioning other Twitter users in their tweets to form 330,164 connections, or ‘edges’, linking users. The chart has been produced using Gephi’s OpenOrd plugin (S. Martin, Brown, Klavans, & Boyack, 2011); nodes have been coloured by modularity class, a calculation which ‘measures how well a network decomposes into modular communities’ (Gephi, 2018b), a higher score indicating a more sophisticated community structure. While the software does successfully identify 73,799 distinct communities in the data set, the visualisation (and modularity class score of 0.792) also shows that many nodes (outlying pale dots in the chart) are isolated, since not all Twitter users mentioned were sampled. For this reason, and to provide further opportunities for analysis and

cross-comparison between events, a second real-time recording of OSN interactions during the much longer run up to the 2014 Scottish Independence Referendum was started on 18 September 2013, exactly one year before the vote was due to take place.

Scotland, with a much smaller population (~5 million) than the US (~320 million), was thought unlikely to generate the sorts of OSN data volumes a 1:1 sampled US recording would have created.

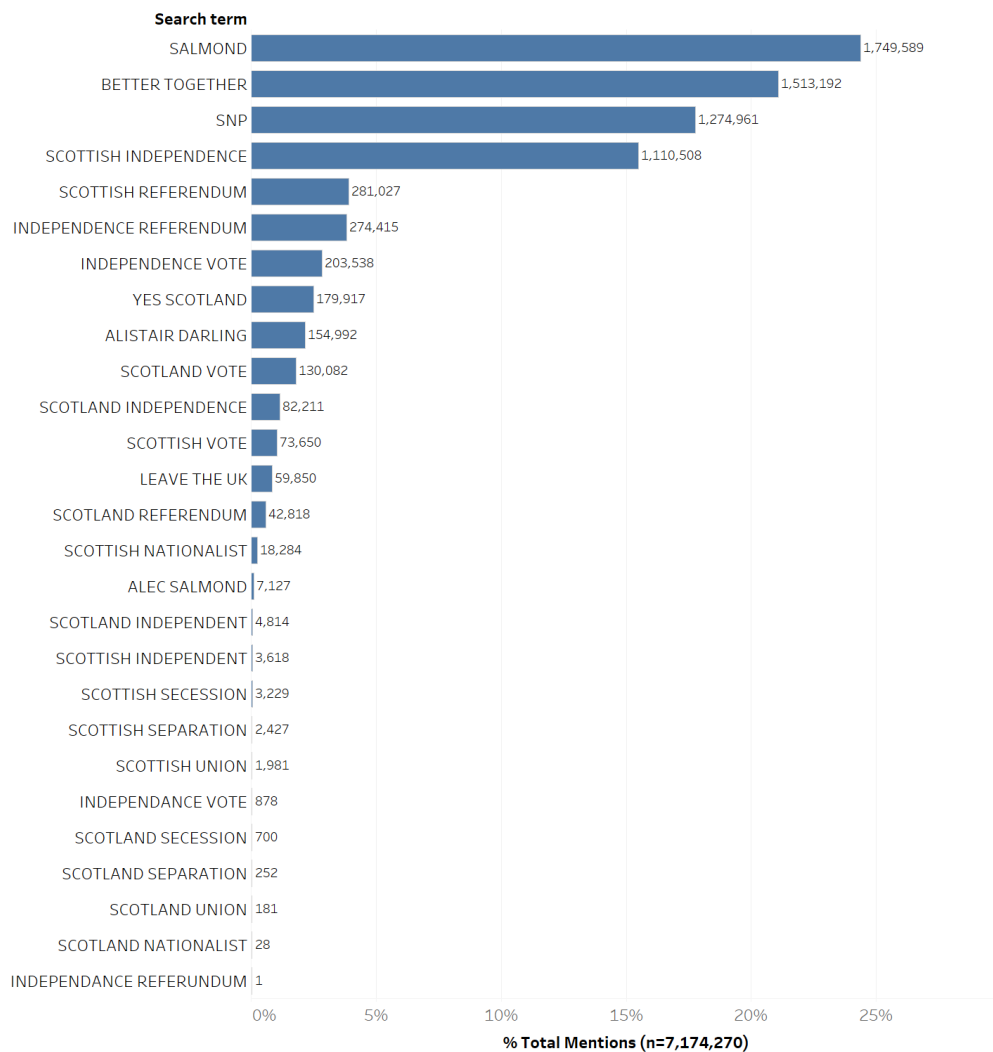


Figure 4-6 – SCOT2014: Search terms used, numeric and percentage contribution to OSN interactions sampled (n=7,174,270 total mentions of search terms in text)

Interactions were filtered (Appendix A7.3, p435) on any case-insensitive words or phrases matching those illustrated in Figure 4-6 (p131), which also shows the contribution of each search term to the sample; again noting that some terms will have appeared multiple times within message text. Deliberate misspellings ('independance' etc.) were incorporated in the CSDL design as misspellings are a common feature of OSN communications (Deitrick & Hu, 2013; Maruyama, Robertson, & Douglas, 2014; Russell, 2011). Despite Scotland's small population size worldwide interest in the outcome of the Referendum, coupled with the longer-running nature of the recording, eventually resulted in the collection of ~6.5 million OSN interactions. The top 3 of 27 search terms account for 63.25% of interactions sampled in the data set. Filtering has selected for inclusion on a mix of First Minister (and Vote Yes leader) Alex Salmond's surname, the campaign slogan ('Better Together') of the Vote No (remain united) coalition, where no one political figure spearheaded the campaign, and the abbreviation 'SNP' (Scottish Nationalist Party), the name of the pro-independence party in Scotland. As noted earlier (Section 4.2.4.1, p126), the top 2 terms usefully, and reasonably evenly, select interactions for the major protagonists in the 2014 Scottish Independence Referendum and are thought to offer a balance of messages for inclusion in the sample for both opposing sides of the political debate.

Table 4-2 – SCOT2014: Summary of recorded OSN interactions (n=6,477,713)

Details	SCOT2014
Start Date	18/09/2013
End Date	30/09/2014
Duration	378 days
Sample	All
Coverage	Worldwide
Language	English
Geo Required?	No
Geotagged	187,975
Non-geotagged	6,289,738
Total	6,477,713
% Geotagged	2.90%

The data set is summarised in Table 4-2 (p132) and consists of 6,477,713 records with 411 fields, many more than had been recorded in the US2012 data set two years earlier.

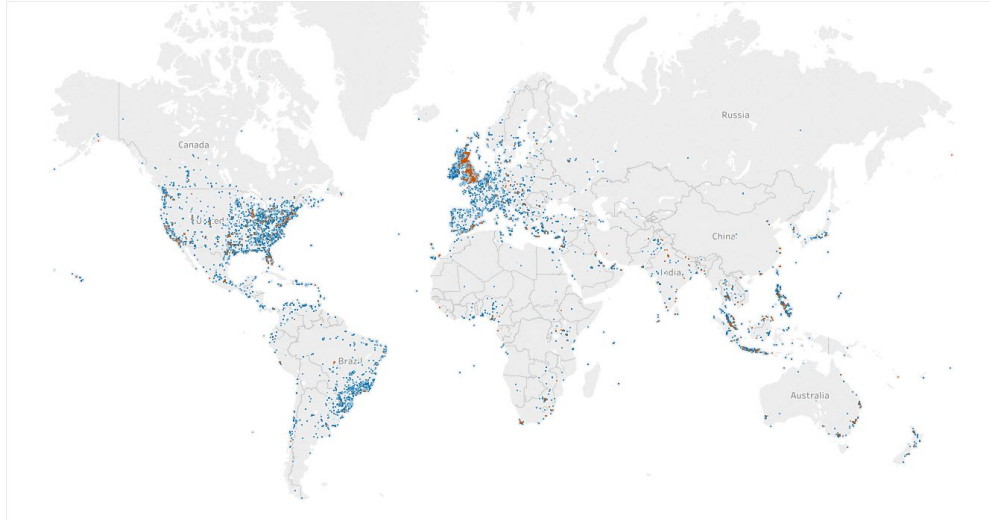


Figure 4-7 – SCOT2014: Worldwide distribution of coordinate-geotagged Facebook posts (orange), Twitter tweets (dark blue) and retweets (lighter blue)

Downloadable from DataSift, again in both CSV and JSON formats, output files (Table 4-3, p136) were again loaded into the Oracle 12c RDBMS (Section 4.3.1.3, p145). Altogether across the SCOT2014 data set 187,975 records, a mixture of 1,231 Facebook posts, 92,437 Twitter tweets and 94,307 retweets were coordinate-geotagged; these could easily be mapped (Figure 4-7) using Tableau data visualisation software (Section 4.5.2, p161) or a GIS. At 2.90% the SCOT2014 coordinate-geotagging rate (Table 4-2, p132) was somewhat higher than the overall percentage (0.92%) found in the US2012 data set (Table 4-1, p128), primarily due to a large number of coordinate-geotagged retweets. However, the figure is still broadly in line with earlier findings (Section 4.2.3, p123; Section 4.2.4.1, p126; Leetaru et al., 2013) reporting low geotagging rates. A more detailed description of geotagging rates by OSN source and subtype is given in Table 4-8 (p170) later in this chapter.

4.2.5 Data acquisition

Two distinct phases of data acquisition in 2012 and 2013-2014 yielded a total of 8,196,437 politically discursive OSN interactions. Data costs totalled \$4,968 or around £4,000, tending to echo Zelenkauskaitė & Bucy's (2016) assertion that 'both the cost and control of social media data limit opportunities for research.' Costs associated with the 2012 US Presidential Election may be considered reasonable value, at \$1,728 (around £1,400) for ~1.7m records, or 0.1 US cents per record. The longer-running 2014 Scottish Independence Referendum data cost around twice that amount but yielded nearly four times as many records, at 0.05 US cents per record, including a full set of social network relationships for the phrases used in the sample design. Scottish interactions were cheaper to acquire as DataSift charges are largely based on the complexity of the CSDL used, this being more complex during the US2012 data collection exercise (Appendix 7, p432). By the standards of much conventional academic or commercial market research, however, these costs are remarkably low for the number of 'responses' received.

Overall, the two case study data sets are unique:

- Twitter's full 'Firehose' was used for data acquisition, filtered and sampled using DataSift technologies, rather than the free and more commonly-used 1% Streaming API.
- Around 10% of the OSN records sampled came from Facebook, most of which were recorded during the 2014 Scottish Independence Referendum. Many studies examine data from only one OSN source; having two enables comparison between posting characteristics observed on both.

As the large OSN operators tighten their data access policies in response to user privacy concerns (Hogan, 2018), commercial considerations (Hern, 2014) and mounting regulatory pressures (McKinnon & Seetharaman, 2018) it is possible that accessing such data sets in the future may become increasingly difficult.

4.3 Data preparations

The previous section has detailed the nature and acquisition of the data subjects; a large set of ~8 million social media interactions in digital form. This section details the procedures adopted and systems used to store files of interaction data acquired from DataSift. Without data storage (Section 4.3.1) in a data management system offering advanced querying and software integration capabilities no further processing or interrogation of data would have been possible.

4.3.1 Data storage

In 1970, Codd (p377) stated that 'Future users of large data banks must be protected from having to know how the data is organized in the machine' describing a relational framework designed to provide 'the independence of application programs and terminal activities from growth in data types and changes in data representation.' Although well aware of the potential for 'natural growth in the types of stored information' Codd and other early designers of RDBMS software would probably not have anticipated the recent step-change in volumes of stored data and diversity of data types. Data in two commonly-used text interchange formats, CSV and JSON (ECMA International, 2013, 2017), downloaded in files from DataSift's servers (Section 4.2.4, p126) or transmitted in streams by Web-based APIs (Section 4.4, p147), have been accessed and stored in this research.

Surprisingly, given the long-standing ubiquity of CSV files in Extract, Transform and Load (ETL) operations, early attempts to verifiably load large numbers of OSN interactions in this format into several data management systems proved difficult. Importing JSON files or 'ingesting' streaming JSON data into these systems also proved challenging in many instances and impossible in one (Section 4.3.1.2, p140). Table 4-3 (p136) shows the file names, each available in CSV and JSON variants, record counts and file sizes of data downloaded from DataSift during the data acquisition phase of this project (Section 4.2.5, 134).

Table 4-3 – File listings, record counts and file sizes for CSV and JSON formatted data downloaded from DataSift

Event	Files (.csv and .json)	Records	CSV Size	JSON Size
US2012	us_2012_geo	146,424	178.69MB	264.41MB
	us_2012_non_geo	1,560,967	2,385.72MB	2,879.62MB
	us_2012_non_geo_hisp	11,276	17.03MB	20.77MB
	Subtotal	1,718,667	2,581.44MB	3,164.80MB
SCOT2014	part-r-00000	2,159,069	7,058.96MB	6,661.80MB
	part-r-00001	2,158,686	7,056.14MB	6,660.25MB
	part-r-00002	2,160,015	7,058.37MB	6,662.54MB
	Subtotal	6,477,770	21,173.47MB	19,984.59MB
TOTAL		8,196,437	23,754.91MB	23,149.39MB

Altogether, 8,196,437 records had to be stored in a database management system capable of importing 23.75GB of file input held in CSV and/or 23.15GB held in JSON format. Neither file format would load straightforwardly and the storage of JSON formatted data, discussed below, proved particularly taxing.

4.3.1.1 JSON

Most OSN interactions and much other data returned by Representational State Transfer (RESTful) APIs such as the GATEcloud.net, AlchemyAPI and CLAVIN-rest systems used for data augmentation in this research (Section 4.4, p147) are stored or exchanged in one of the rapidly evolving ‘interchange languages’ of the Web (Decker et al., 2000); either Extensible Markup Language (XML) or, more recently, JavaScript Object Notation (JSON). The latter is ‘a lightweight, text-based, language-independent data interchange format’ that ‘facilitates structured data interchange through a syntax of braces, brackets, colons, and commas’ (ECMA International, 2013, ppl-II). JSON can represent ‘objects and arrays [which] nest [allowing] trees and other complex data structures [to] be represented’ (ECMA International, 2013, ppl).

The JSON data interchange format has been widely adopted by major OSN platform operators including Twitter (2013b) and Facebook (2013) and may be illustrated (Figure 4-8, p137 and Figure 4-9, p139) using the on-screen and raw JSON data

representations of a tweet downloaded from DataSift sent from the Twitter account of Presidential Candidate Barack Obama on 5 November 2012 by his campaign staff, recorded in the US2012 data set.



Figure 4-8 – On-screen graphical representation of Presidential Candidate Barack Obama’s tweet created on 05/11/2012 (from the raw JSON shown in Figure 4-9)

On Twitter, users see Obama’s tweet as it appears in Figure 4-8, graphically rendered and simply laid out. In JSON, used to record and share Twitter tweets and many other OSN interactions, the corresponding raw data (Figure 4-9, below) are represented by a sequence of Unicode code points, certain characters (e.g., the solidus character or forward slash ‘/’) are escaped and file encoding is in UTF-8 (Yergeau, 2003). The number of ‘key/value pairs’, broadly analogous to ‘fields’ or ‘columns’ in tabular database systems, degree of nesting and the length of arrays varies from record to record.

```
{
  "interaction": {
    "author": {
      "avatar":
      "http://a0.twimg.com/profile_images/2764236884/901
      02995f6e328d7f90c43c8b337a0c7_normal.png",
      "id": 813286,
```

```

    "link": "http:\\\\twitter.com\\BarackObama",
    "name": "Barack Obama",
    "username": "BarackObama"
  },
  "content": "Happening now: President Obama speaks in
Ohio about the choice in this election. RT so your
friends can watch, too. http:\\\\t.co\\d42qgdn8",
  "created_at": "Mon, 05 Nov 2012 21:39:41 +0000",
  "id": "1e227914e2f4ac80e0740cf699462aae",
  "link":
"http:\\\\twitter.com\\BarackObama\\statuses\\265569098
132516864",
  "schema": {
    "version": 3
  },
  "source": "web",
  "tags": ["Democratic Party", "Neutral", "Barack
Obama"],
  "type": "twitter"
},
"klout": {
  "score": 98
},
"language": {
  "confidence": 100,
  "tag": "en"
},
"links": {
  "created_at": ["Mon, 05 Nov 2012 03:10:59 +0000"],
  "retweet_count": [0],
  "title": ["Watch live: Barack Obama on the campaign
trail \\u2014 Barack Obama"],
  "url": ["http:\\\\www.barackobama.com\\live"]
},
"salience": {
  "content": {
    "sentiment": 0
  }
},
"trends": {
  "content": ["ohio", "can"],
  "source": ["twitter"],
  "type": ["Canada", "daily"]
},
"twitter": {
  "created_at": "Mon, 05 Nov 2012 21:39:41 +0000",
  "domains": ["OFA.BO"],

```

```

    "id": "265569098132516864",
    "links": ["http:\\\\OFA.BO\\PfvCKP"],
    "source": "web",
    "text": "Happening now: President Obama speaks in
    Ohio about the choice in this election. RT so your
    friends can watch, too. http:\\\\t.co\\d42qgdn8",
    "user": {
      "created_at": "Mon, 05 Mar 2007 22:08:25 +0000",
      "description": "This account is run by #Obama2012
      campaign staff. Tweets from the President are signed -
      bo.",
      "followers_count": 21753954,
      "friends_count": 670840,
      "id": 813286,
      "id_str": "813286",
      "lang": "en",
      "listed_count": 179596,
      "location": "Washington, DC",
      "name": "Barack Obama",
      "screen_name": "BarackObama",
      "statuses_count": 7779,
      "time_zone": "Eastern Time (US & Canada)",
      "url": "http:\\\\www.barackobama.com",
      "utc_offset": -18000,
      "verified": true
    }
  }
}

```

Figure 4-9 – A JSON formatted Twitter tweet sent from the account of Presidential Candidate Barack Obama and created on 05/11/2012

Around 150,000 records containing Latitude and Longitude coordinates resulted from sampling during the run-up to the 2012 US Presidential Election. These records, and other geotagged interactions from the various Streams recorded in 2012 and 2013-2014 (Appendix 7, p432), hold an additional `geo` key/value pair nested within the `interaction` key which, in JSON, takes the form:

```
"geo":{"latitude":40.8183573,"longitude":-73.965401}
```

The ability of JSON to systematically describe arbitrarily structured and/or volatile data makes it both extremely powerful for programmers, who like flexibility, and Web platform operators, who like fast development. It also makes JSON potentially

challenging for database administrators and data analysts, both of whom typically like well-defined structure. Many commonly-used RDBMS and GIS software packages rely upon tabular row and columnar data storage accessing somewhat inflexible ‘designed-in-advance’ schemas (Tear, 2014) whereas the structure of Twitter tweets, and many other OSN or Web data interchange formats, is more likely to be unpredictable and also changes frequently (Faber, Matthes, & Michel, 2016; Twitter, 2018b). Between 2012 and 2014 the maximum number of fields in the CSV files acquired and downloaded from DataSift and the number of key/value pairs in the corresponding JSON file variants of these data sets more than doubled, from 163 to 411. Most of the additional keys were found in Facebook data, not much of which had been recorded in 2012, and many of the key/value pairs held values in different language alternatives, e.g., several variants of Arabic, probably reflecting a change in Facebook’s multilingual architecture around this time.

An additional 12.66GB of augmented data, derived from three NLP/geoparsing systems (Section 4.4.1, p147), all output in JSON format with wildly varying record structures, necessitated the selection of a database system capable of efficiently storing and querying (Section 4.5.1, p159) data held in this popular, but extremely variable, ‘semi-structured’ interchange language. As this proved much more difficult than anticipated, several different data management systems were evaluated. This process of evaluation is described in the following section.

4.3.1.2 Data management systems

As the research has progressed the ability to store, query and analyse unstructured text has proven a key requirement. The `US2012` data set consists of ~1.7m date and time-stamped OSN messages, many of which may be geo-referenced directly or indirectly through text matching, entity extraction or other Natural Language Processing (NLP) techniques run against 30,125,821 space-tokenized words. The larger `SCOT2014` data set consists of ~6.5m OSN messages and a massive 205,089,540 space-tokenized words. It is clear that ‘the huge amount of free-form

unstructured text in the blogosphere, its increasing rate of production, and its shrinking window of relevance, present serious challenges to the [...] analyst who seeks to take public opinion into account' (Till, Longo, Dobell, & Driessen, 2014, p71).

	Type	Installation	Import CSV	Import JSON	Query data	Serve data	Overall score
SQL Server 2012 R2	SQL	5	1	0	5	5	16
Oracle 12c R1	SQL	5	5	5	5	5	25
MarkLogic 7	NoSQL	4	1	4	2	3	14
Oracle Endeca	NoSQL	4	2	2	3	2	13
MongoDB	NoSQL	5	2	5	1	3	16
MapR Hive	SQL	2	5	1	4	4	16
MapR Drill	NoSQL	2	5	5	4	5	21

Figure 4-10 – Subjective scoring of the SQL and NoSQL data management systems used in this research (0=worst; 5=best)

Figure 4-10 shows the matrix of data management systems used in this research, subjectively colour-coded from experience of installing software, importing CSV and JSON files or streams, querying and serving data to external 3rd party applications, e.g., Tableau visualisation software or Web-based APIs. It is apparent that all of the software products used had different strengths and weaknesses and that Oracle 12c Release 1 (Section 4.3.1.3, p145) offered the strongest feature set overall. The competing systems are briefly evaluated below:

- **Installation** – The software used and evaluated consists of a number of primarily Commercial Off the Shelf (COTS) products with one open-source NoSQL database (MongoDB) and a commercialised distribution (MapR) of Apache's open-source Hive and Drill projects. The two conventional SQL databases (Microsoft SQL Server 2012 Release 2 and Oracle 12c Release 1) installed straightforwardly on physical hardware. MarkLogic 7, Oracle Endeca and MongoDB could be installed reasonably easily on Linux virtual machines (Appendix 8, p436). Installation of MapR in a clustered environment proved much more difficult. Experiments using a cluster installed as a set of Hyper-V Ubuntu Linux virtual machines on a Windows 2012 Server host (Figure A8-1, p439) failed before professional assistance

from staff in the Institute of Cosmology and Gravitation (ICG) resulted in the successful installation of a five-node cluster on SCIAMA, the University of Portsmouth's supercomputer (G. Burton, 2017).

- **Import CSV** – Despite its ubiquity, CSV formatted data proved harder to import than expected. Initial imports using the 'Import Data Wizard' of Microsoft SQL Server 2012 R2's Data Transformation Services (DTS) resulted in field truncation and data loss; two import attempts failed to correctly handle UTF-8 encoded strings resulting in data loss both of international characters (e.g., Spanish diacritics) and emoticons and these problems could not be overcome using either the more advanced SQL Server Data Transformation Tools (SSDT; Microsoft, 2013) or changed working practice (Murray, 2013). In response, Oracle 12c Release 1 was adopted and, after much coding of SQLLDR control files (Oracle, 2018d), successfully and verifiably imported all US2012 CSV data, and almost all SCOT2014 CSV data, rejecting just 57 badly formatted records out of 6,477,770. MapR Hive and Drill could both read UTF-8 encoded CSV data easily while the other data management systems, more attuned to document or semi-structured data storage, were not primarily designed to handle CSV data.
- **Import JSON** – Aside from Microsoft SQL Server 2012 Release 2, which had no capabilities to import semi-structured JSON data, most of the software systems evaluated could import or 'ingest' this format. MapR Drill and MongoDB were the most elegant, the former simply requiring JSON files (still GZIP compressed) to be placed in the correct directory of the Hadoop Distributed File System (HDFS) cluster disk and the latter rapidly 'ingesting' flat files of multi-row UTF-8 encoded JSON into its internal database. The process in MarkLogic 7 was somewhat less straightforward as several large files containing millions of JSON records had to be split atomically into millions of separate files, each containing one JSON record. Oracle Endeca was able to read JSON straightforwardly but, as a developer edition, could

only read a limited number of records or a very limited number of key/value pairs for a larger number of records. MapR Hive was not designed for JSON data storage, although this can now be achieved (MapR, 2018). Oracle 12c Release 1, while requiring a more complicated setup routine (Section 4.3.1.3, p145), successfully and verifiably imported all JSON data including the 57 records that had been rejected during import of the CSV variant files.

- **Query data** – The most important consideration in adopting a data management system to store data is the ability of that system to query data (Codd, 1972). In this respect the seven systems evaluated differed markedly, although some of these differences undoubtedly arose from a combination of operator familiarity, expediency and inertia; each of which are common factors in decisions surrounding IT systems adoption (Agarwal & Prasad, 2000; Venkatesh, Davis, & Morris, 2007). Considerable effort has been expended, drawing upon over thirty years of IT experience using multiple software systems, to read, store and query OSN data imported from CSV and JSON formats. Microsoft SQL Server 2012 Release 2, first used in this research, proved through SQL querying that the ETL steps undertaken to load CSV data into the database had not worked properly. SQL queries revealed the problems of field truncation and character set mishandling outlined above so that, although Microsoft SQL Server 2012 R2 was subsequently abandoned, its querying facilities (and the value of comprehensively checking data in this way upon loading) may be scored highly. Oracle 12c Release 1, with a broadly comparable and comprehensive querying tool, SQL Developer (Oracle, 2017a), scored similarly highly. Of the other systems MapR Hive and MapR Drill both offered the familiarity of SQL, although the software interface to execute queries was much less polished than the two mainstream SQL RDBMSs. The three document-store NoSQL databases provided extremely unfamiliar querying environments; MarkLogic 7 required the development of an application, Oracle Endeca presented a pre-built Web browser-based application aspects of which, such as ‘faceted’

search, were well-executed, and MongoDB presented a black-box; only addressable through JavaScript programming.

- **Serve data** – The final component of this evaluation considers the ability of the data management systems to serve data. All of the software systems used are database servers and all of them are highly scaleable. Tests conducted on the 5-node MapR Hive cluster installed on the SCIAMA supercomputer, for example, showed that the software could store a massive ~3TB file of generated sample data which could be remotely accessed and queried using Hive (Tear, 2017). Other demonstrations showed how JSON data stored in MapR Drill on SCIAMA could be accessed over the Internet, live from a conference in Helsinki (Tear, 2016), and mapped and plotted using Tableau visualisation software on a laptop (Section 4.5.2, p161). Hence, the scores for MapR and for the two mainstream SQL RDBMSs, both of which are well-established and known to integrate well with a wide-range of 3rd party tools and applications, are high. The three document-store NoSQL databases did not score so highly in this evaluation as, despite some extremely impressive demonstrations at conferences, it proved difficult to access data stored in these systems using programming standards such as Open Database Connectivity (ODBC) or 3rd party tools including, until much more recently, the visualisation software Tableau.

The files imported or ingested during this research are not ‘Big’ enough to be fully representative of some of the immense storage, manipulation or analysis problems occurring in significantly larger data sets. However, the experiences set out above usefully highlight technical, workflow and integration issues of relevance to individual researchers or small research teams. Zelenkauskaitė & Bucy (2016) have stated that ‘even if physical access to data is available to scholars, computational skills and technical expertise become a limiting factor, thereby introducing a schism

in scholarly opportunities that conventional social science and humanities traditions are not entirely prepared to deal with.’

As data get ever bigger this problem may be amplified, necessitating better training in Big Data ‘spatial science and quantitative analysis’ for geographers (Johnston et al., 2014) in case geography, which in its spatial component – if not its history, theory and tenets – may be reduced to ‘just another data type’, and comes to be studied by computer scientists alone. Other researchers might have chosen other systems, however, Oracle’s 12c RDBMS was selected as the data management system used in this research, and is further detailed in the following section.

4.3.1.3 Oracle 12c Release 1

Only recently have commercial (Oracle 12c Release 1, July 2014) and open-source (e.g., PostgreSQL 9.3, September 2013) RDBMSs offered native storage and query support for JSON data, the preferred format for OSN and much other Web-based data interchange, which has moved ‘from being an underground secret, known and used by very few, to becoming the clear choice for mainstream data applications’ (Severance, 2012). Early difficulties in reading and storing JSON files focused attention on the need to import parallel CSV versions of these data in which, e.g., nested JSON key/value pairs containing arrays (such as the trends listed in Barack Obama’s earlier tweet, Figure 4-9, p139) were transposed by DataSift into three CSV fields containing delimited string literals (Table 4-4, p146).

New JSON-handling features built into the Oracle 12c Release 1 (version 12.1.0.2.0) RDBMS were then used to successfully import JSON formatted variants of the same sampled OSN interactions into the `OSNDATA` database. Data are stored as Character Large Objects (CLOBs) with JSON constraint (Oracle, 2014a) in a permanent database table populated from an external staging table used to read files downloaded from DataSift (Table 4-3, p136). The SQL statement used to access one of these files is shown in Appendix 11 (listing 5, p479).

Table 4-4 – Transposition of nested, arrayed JSON into three CSV fields containing delimited string literals

Type	Content
JSON	<pre>"trends": { "content": ["ohio", "can"], "source": ["twitter"], "type": ["Canada", "daily"] },</pre>
CSV	Content
TRENDS_CONTENT	["ohio", "can"]
TRENDS_SOURCE	["twitter"]
TRENDS_TYPE	["Canada", "daily"]

The SQL code opens and reads a JSON file from a directory on the machine, using newlines as the delimiter, in the UTF8 character set with an extended `READSIZE` of 1,048,576 bytes so each line of file input fits into the RDBMS memory buffer. From this external staging table data (effectively, the file) may be `INSERTED` into a permanent database table (defined in Appendix 11 listing 6, p480) using another SQL statement (Appendix 11 listing 7, p480).

Successful querying, tabulation and analysis (Section 4.5.1, p159) of this stored JSON data led to the adoption of Oracle 12c Release 1 as the *de facto* data management system for this project. Oracle's database software sits at the centre (Figure A8-3, p441) of a complex, multi-tenant computing environment described in Appendix 8 (p436). The software has been called upon to store ~8m OSN interactions recorded in 23.75GB of raw CSV and 23.15GB in raw JSON, supplemented by another 12.66GB of raw JSON output from three NLP/geoparsing systems used to detect toponymic mentions and other entities (e.g., persons, organisations) in OSN message text, linked/shared URLs and metadata. Procedures relating to data augmentation are discussed in the following section.

4.4 Data procedures

The previous section has described how social media interaction data, downloaded from DataSift in CSV and JSON files, were loaded into the Oracle 12c RDBMS.

Although Oracle's (2012) Text features enabled increasingly sophisticated indexing and querying of free-form text, additional procedures – using even more advanced Natural Language Processing (NLP) software – were adopted to search for and find mentions of place in OSN interaction message text and linked/shared content.

These text-mining ‘augmentations, and the systems used to perform them, are described in the following section.

4.4.1 Data augmentation

Massive recent growth in the amount of unstructured electronic text available for analysis (JISC, 2012; Manyika et al., 2011) has spurred the development of many commercial and open-source software systems designed to ‘mine’, ‘augment’ or ‘enhance’ textual data. As JISC (2012, p13) state, the ‘availability [of large amounts of text data] does not equate to being able to analyse easily the content to find sought after information or to develop new insights.’ There is simply too much text for individual researchers to read; e.g., JISC note that upwards of ‘1.5 million [journal] articles are added [by 11,500 journals] per year’ and specialist domain knowledge is required to make sense of certain text terms (e.g., ‘tree’, ‘branch’, ‘leaf’ in JISC’s example) that may have very different meanings in different disciplines.

JISC (2012, p13) propose that ‘Text mining offers a solution to these problems, drawing on techniques from information retrieval, natural language processing, information extraction and data mining/knowledge discovery’ in four stages:

1. Enhanced information retrieval
2. Linguistic analysis and entity recognition

3. Information extraction
4. Data mining/Knowledge discovery

Enhanced information retrieval has been used in this research programme, a) to search for relevant academic literature to contextualise the study (Chapter 2, p51), and; b) to search for relevant OSN interactions to provide case study material for the research (Chapter 3, p94 and Chapter 4, p118). As the ~8 million social media interactions recorded here could not possibly be examined individually, three Natural Language Processing (NLP) systems have been used to address JISC's suggested stages 2 and 3. Two of the three systems, GATEcloud and CLAVIN-rest, offer somewhat similar geoparsing information extraction functionality allowing cross comparison, while the third, AlchemyAPI, is particularly well-suited to information extraction and knowledge discovery operations against Web-hosted URLs, which are widely-shared in OSN interactions. Several data mining and knowledge discovery processes address the fourth stage of JISC's suggestions, using a mixture (Sections 4.5.1 and 4.5.2, pp159-161) of relational, non-relational and graph databases, queries, visualisation and statistical analyses (Section 4.5.3, p163).

Stock (2018, p209) has noted that 'During the last ten years, a large body of research extracting and analysing geographic data from social media has developed.' Reviewing 690 papers accessing 20 social media platforms she states that 'a wide array of [...] approaches have been developed, with methods that extract place names from message text providing the highest accuracy.' The three NLP packages successfully used for geographical entity recognition and extraction from message text and linked/shared content in this research are discussed below. A fourth subsection (4.4.1.4, p157) briefly describes two others since, as Gritta, Pilehvar, Limsopatham, & Collier (2018) have noted, there is a 'substantial disparity' between working or workable NLP entity extraction systems and those that are difficult to install, use or simply will not run.

4.4.1.1 General Architecture for Text Engineering (GATE)

GATE, released as open-source software by a team at the University of Sheffield, 'is over 15 years old and is in active use for all types of computational task[s] involving human language' (GATE, 2017). Originally a desktop application, or a set of Java Archives (JARs) for use in 'embedded' applications, two more recent developments prompted adoption of the software for use in this research. First, the team released TwitIE (Bontcheva et al., 2013), a Twitter Information Extraction engine and 'open-source NLP pipeline customised to microblog text at every stage.' Around 90% of the ~8m records in the case study research data corpus originate from Twitter, and tweets are notoriously 'difficult [to process]: the genre is noisy, documents have little context, and utterances are very short' (Bontcheva et al., 2013, p1).

TwitIE is designed to process terse and frequently ungrammatical tweets using the 'sentence splitter' and 'name gazetteer' functions of ANNIE (A Nearly-New Information Extraction system and another GATE component), supplemented by specially developed functions for language identification, tokenisation, normalisation, Part of Speech (POS) tagging and Named Entity Recognition (NER). GATE software operates on a Corpus, a set of documents or, in this case, a set of tweets. A Corpus can be constructed by searching for records within the Oracle 12c database, e.g., to find the 19 Twitter tweets recorded in the SCOT2014 data set made by Scottish First Minister Alex Salmond during the 2014 Scottish Independence Referendum campaign (Appendix 11 listing 8, p480).

After loading the necessary plugins (File -> Manage CREOLE Plugins... to activate plugins `Format_Twitter` and `Twitter`) the TwitIE Ready Made Application may be launched. The appropriate Corpus is selected, and TwitIE run against it, yielding the sort of output shown in Figure 4-11 (p150). In this example checkboxes on the rightmost panel of GATE Desktop have been used to highlight Locations (pink), Persons (purple) and Organizations (green) recognised by the software in the text of Salmond's 19 Twitter tweets. No human intervention has been necessary.

The software works well on the desktop, but is constrained by memory limits, and cannot deal with the millions of tweets in the research data corpus. Recognising the need to analyse increasingly massive text corpora, the GATE team developed GATEcloud.net, ‘a platform for large-scale, open-source text processing on the cloud’ (Tablan et al., 2012). NLP pipelines developed on GATE Desktop software can be exported to GATEcloud.net and run on dedicated machines, ‘harnessing the vast, on-demand compute power of the Amazon cloud’ (Tablan et al., 2012, p1).

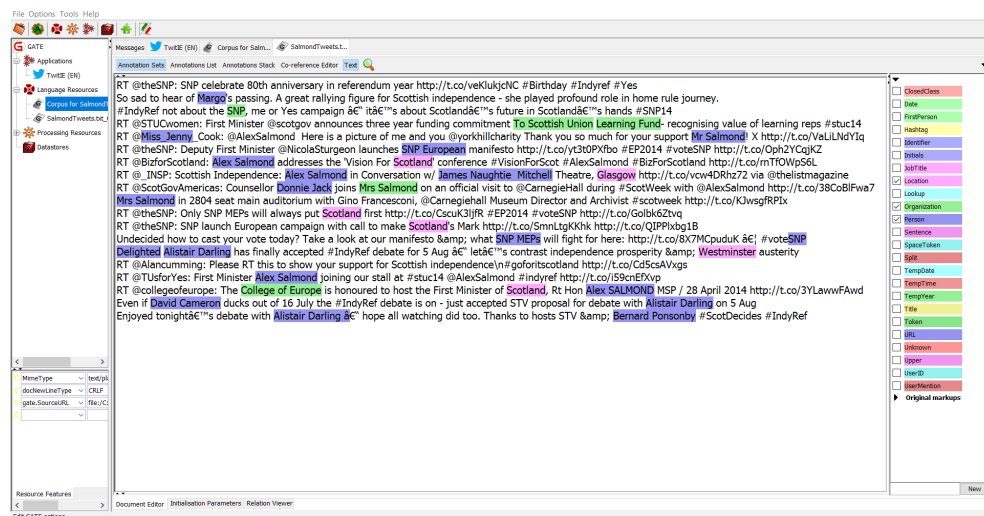


Figure 4-11 – Scottish First Minister Alex Salmond’s Twitter tweets processed using TwitIE on GATE Desktop

Sharding, load-balancing and other ‘important infrastructural issues’ of the process are handled by the GATEcloud.net application which, the authors’ suggest, helps enable the ‘democratization of science’ by providing individual researchers or small research groups with ‘cutting-edge, data-driven, text-processing’ systems that are otherwise extremely difficult to set up (Tablan et al., 2012, p2). Having used GATE Desktop experimentally for some time, the entire research data corpus of ~8 million records was processed using GATEcloud.net (Figure A8-3, p441). The run was designed to perform Information Extraction and Named Entity Recognition, particularly of locations and, coincidentally, helped in beta testing of a new deployment of GATEcloud software (Roberts, personal communication, 2016). Input

files in DataSift's JSON format, used earlier to help develop GATE's DataSift reader (Bontcheva & Greenwood, personal communication, 2014), were processed using TwitIE with output to a set of 86 (US2012=16, SCOT2014=70) JSON files subsequently re-imported to the Oracle 12c database.

Crucially, rather than outputting only input text (i.e., the message) and TwitIE's augmentations, co-development work with Roberts (2017) ensured that 'tweet IDs [were] passed through into the output'; making it much easier to join input and output for analysis using SQL in the database (Section 5.2.2.1, p193). The ability to join two tables of data together on a common key (e.g., an ID or identifier field) held in each is a central concept in data management (Codd, 1970). While tables can be joined on text fields (e.g., the message text of OSN interactions) there are many duplicated rows of message text (n=4,739,827), mainly Twitter retweets, in the OSNDATA database; each retweeted by an individual user with different characteristics (e.g., coordinate-geotagging or not). Using message text as the key, in this or similar instances, would not enable correct joining of metadata to the respective message input to and output from GATEcloud. The co-developed functionality in GATEcloud newly resulting from this research should, therefore, prove extremely useful for subsequent researchers.

GZIP-compressed in Linux, GATEcloud.net output totalled 170MB (US2012) and 790MB (SCOT2014) in size. Uncompressed, the US data set required 901MB of file storage and the Scottish data set 4.19GB. The tables storing this data in Oracle 12c are 2.34GB and 11.19GB in size respectively.

4.4.1.2 AlchemyAPI

AlchemyAPI, now re-branded Watson Natural Language Understanding and part of IBM's Watson Developer Cloud service (IBM, 2017b), is Cloud-hosted, commercial software, available on a rate-throttled basis upon request to academic researchers. This RESTful API service has been used by several scholars (Cios & Kurgan, 2006; Gelernter & Mushegian, 2011; Kulshrestha, Zafar, Espin-Noboa, Gummadi, &

Ghosh, 2017; Quercia et al., 2012; Saif, He, Fernandez, & Alani, 2016) particularly for Twitter sentiment analysis, where it may help to ‘alleviate data sparsity [and] performs better than [other Web-hosted systems including Zemanta or OpenCalais] in terms of the quality and the quantity of the extracted entities [returned]’ (Saif, He, & Alani, 2012, p4).

IBM (2017a) documentation states that AlchemyAPI offers:

- Entity Extraction
- Sentiment Analysis
- Emotion Analysis
- Keyword Extraction
- Concept Tagging
- Relation Extraction
- Taxonomy Classification
- Author Extraction
- Language Detection
- Text Extraction
- Microformats Parsing
- Feed Detection
- Linked Data Support

As a RESTful web service, calls to AlchemyAPI are made over Hyper Text Transfer Protocol (HTTP) using an API Key for authentication. Academic usage is restricted to 30,000 ‘daily transactions’, compared to 2 million/day or more for commercial users, and the number of ‘transactions’ used to process each piece of text (e.g., OSN message text or text found at a linked/shared URL) will vary according to which calls, from the list above, are made to the service.

Bespoke software was developed using Ruby (2017) scripts running on a CentOS 7 virtual machine (Appendix 8, p436) to select data from Oracle 12c, pass it to the

AlchemyAPI service and store the returned JSON directly in the database. As the service is rate-limited, the XML-parsing Nokogiri (2017) plugin for Ruby was used to decode responses from the AlchemyAPI management URL to determine how many ‘daily transactions’ remained to be consumed (Appendix 10, p451). Two applications were developed:

1. **PROCESS_RECS** – A Ruby script, executed through a shell script called from `cron`, running every 10 minutes to process up to 150 records per run (Appendix A10.3, p451) selected from a ‘queueing’ table (`ALCHEMY_API`) created in Oracle 12c on the VM host, a Dell Latitude E7440 laptop running Windows 10. The queueing table was populated, with five SQL `INSERT` statements, to store five tranches of OSN interaction messages for AlchemyAPI processing (`US2012_GEO Stream=146,424`, `SCOT2014 geo tagged=1,074`, `US2012_NON_GEO 1% sample tweets=92,304`, and `SCOT2014 1% sample tweets=56,622` records). Message text was processed using AlchemyAPI calls for Entity Extraction, Keyword Extraction, Concept Tagging, Sentiment Analysis, Relation Extraction, Text Extraction and Taxonomy Classification. The data processing allows for comparison, according to these various augmentations, for all coordinate geotagged records from each Stream against a random sample of non-coordinate-geotagged records from both `US2012` and `SCOT2014` data sets. Results are presented in Chapter 5 (p186).
2. **PROCESS_URL_RECS** – A Ruby script, executed through a shell script called from `cron`, running every 15 minutes to process up to 250 records per run (Appendix A10.4, p461) selected from the `LI_LINKS_URLS_DISTINCT` ‘queueing’ table created in Oracle 12c on the Dell laptop VM host, as above. The queueing table was populated, using a SQL `INSERT` statement, with 641,472 distinct link URLs (pointers to linked URLs made by Twitter or Facebook users in an approximate 80:20 ratio) derived from 3,485,840 URL links to online media (e.g., newspaper websites, blogs, YouTube videos etc.)

recorded in the `LINKS_URL` field of the main `INTERACTIONS` table.

Linked URLs were processed using AlchemyAPI calls for Entity Extraction.

The data processing allows for comparison of many detected entity types, particularly location, in linked URLs. Numbers of locations referenced in linked URLs may then be compared by user class, e.g., coordinate-geotagging or not. Results are presented in Chapter 5 (p186).

While GATEcloud.net could be set up and used within days to process ~8 million OSN interactions, rate-throttling of the AlchemyAPI service required the development of queuing tables, populated by numbers of records likely to be processed within the timescales available. All coordinate-geotagged `US2012` and `SCOT2014` OSN interactions were processed, and all distinct linked URLs, but only a 1% sample of all other OSN interactions from each of the two case study events.

4.4.1.3 Cartographic Location and Vicinity INDEXER (CLAVIN)

CLAVIN (the Cartographic Location and Vicinity Indexer) is, according to its developers Berico-Technologies (2017), ‘an award-winning open source software package for document geotagging and geoparsing that employs context-based geographic entity resolution. It extracts location names from unstructured text and resolves them against a gazetteer to produce data-rich geographic entities.’ It is one of several gazetteer-based geoparsing solutions evaluated in this research (see Section 4.4.1.4, p157) but the only one that would compile and run reliably. The version of CLAVIN used, the CLAVIN-rest variant, is a ‘DropWizard RESTful micro-service demonstration of CLAVIN, GeoNames, and OpenNLP or CLAVIN-NERD’ (Berico-Technologies, 2017). The software uses the Stanford CoreNLP toolkit (Manning et al., 2014; Stanford University, 2017) which ‘can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations

between entity mentions, get quotes people said, etc.’ (Stanford University, 2017). Stanford CoreNLP is widely used and can also be used in GATE. Like CLAVIN-rest, Stanford’s CoreNLP code is written in Java.

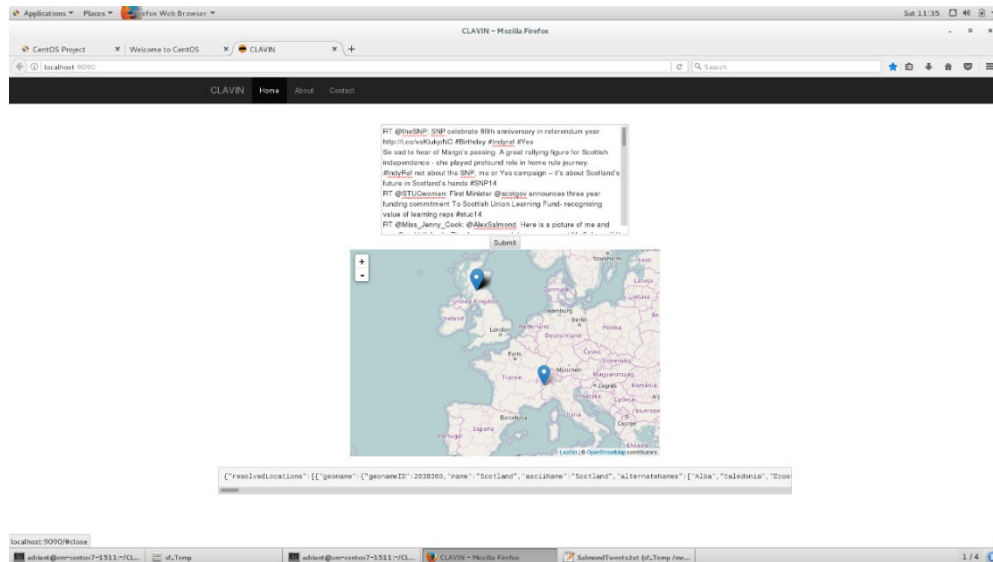


Figure 4-12 – Scottish First Minister Alex Salmond’s Twitter tweets processed using CLAVIN-rest running on a CentOS 7 virtual machine

The Apache Maven ‘software project management and comprehension tool’ (Apache Software Foundation, 2017) was used to build and compile the code (`mvn package` was run against the repository downloaded from GitHub) on a CentOS 7 virtual machine (Appendix 8, p436). The GeoNames (2016) gazetteer database `allCountries.zip`, listing 11,370,639 place names and locations with many language-specific spelling alternatives (e.g., Londres for London), was downloaded on 4 April 2017 and used as the location master file. When started, CLAVIN-rest presents a Web browser-based interface on localhost:9090 as shown in Figure 4-12.

Text data (in this case First Minister Alex Salmond’s 19 sampled Twitter tweets from the 2014 Scottish Independence Referendum) may be copied into the `TEXTAREA` at the top of the browser page and, once submitted, will be geoparsed by CLAVIN-rest. The mappable locations found in this example include ‘Scotland’ and ‘Europe’. Others, including ‘Westminster’, identified as a location by GATE Desktop (Figure

4-11, p150) in another of Alex Salmond's Twitter tweets, has not been found. Geoparsers have different success rates (Gritta et al., 2018) and GATEcloud.net, AlchemyAPI and CLAVIN-rest could all be fooled by sentence structure, a problem returned to in Section 5.4 (p221).

As a RESTful web service, CLAVIN-rest could also be called using `curl`, the Linux command to call URLs from the terminal. A shell script (Figure 4-13, p156) was developed to pass Universally Unique Identifiers and message content from OSN data (concatenating `UUID` and `INTERACTION_CONTENT` fields with the characters `'|~|'`, which did not appear anywhere else in message text) to CLAVIN-rest.

```
#!/bin/bash
OLDIFS=$IFS
IFS="|~|"
while read f1 f2
do
    #echo "UUID is      : $f1"
    #echo "CONTENT is   : $f2"
    curl -s --data "$f2" --header "Content-Type:
text/plain"
http://localhost:9090/api/v0/geotagmin -w
"|~|$f1\n"
done <
~/Desktop/vw_int_content_compdelim_utf8.txt |
grep -v '{"resolvedLocationsMinimum":\[ \]}'
IFS=$OLDIFS
```

Figure 4-13 – Shell script written to call CLAVIN-rest from the command line

The standard output of CLAVIN-rest appends all GeoNames data to the input text, including multiple language-alternative gazetteer spellings (n=185 in the case of London, UK), creating extremely verbose and excessively large files (1.7GB in; ~103GB out). This can be controlled through the use of the `geotagmin` URL argument (Figure 4-13), which prevents output of multiple language-alternative spellings. File sizes were further minimised by piping output through `grep` to store only the `UUID`s, and resolved locations in JSON, of text that could be geoparsed.

This resulted in a much smaller output file size of 487.13MB. The script could be run in a Linux terminal on the Centos7 virtual machine using the command:

```
./test_curl_line_at_a_time_minjson.sh > out.txt
```

All 8,196,380 OSN records were passed through CLAVIN-rest and 1,978,404 records (24.14%) containing `resolvedLocationsMinimum`, an array of `GeoNames` locations with Latitude and Longitude coordinates in JSON, and `UUID`, to join back to the input text and associated metadata, were imported into the Oracle 12c database. Results from this exercise, and a comparison of CLAVIN-rest and the other NLP-based NERs used in this research, are presented in Chapter 5 (p186).

4.4.1.4 Others

Gazetteer search has played a sometimes confounding role in the development of GIS technology on the modern-day Web (G. Cheng & Du, 2008; Pradeepa & Manjula, 2016; Shi & Barker, 2011) and in historical applications (Southall et al., 2011, 2009). The spelling of place names may change over time, many alternate spellings may be used, or places (e.g., Kaliningrad) may change their name altogether. Software may, or may not, be able to pick up on these subtleties, and few geoparsers come close to human levels of accuracy when identifying probable place names within text (Gritta et al., 2018).

In addition to the GATEcloud.net, AlchemyAPI and CLAVIN-rest NLP-based NERs described above, several other geoparsers were assessed. Unfortunately, while showing promise, these systems failed to deliver either due to setup, coding or software compilation problems.

- **BALEEN** – from the UK’s Defence Science and Technology Laboratory (2015) is another RESTful entity extraction framework designed to ‘extract information from unstructured and semi-structured text.’ The software uses Ordnance Survey-derived gazetteers which might have improved geoparsing

results against the SCOT2014 dataset. Unfortunately, the available downloadable version of Baleen would neither compile or run.

- **Edinburgh Geoparser** – from the Language Technology Group (2014) at the University of Edinburgh (Alex, Byrne, Grover, & Tobin, 2014) has been widely used, and scored particularly highly in Gritta et al.'s (2018) review of five geoparsing systems. Version 1.1 (16/03/2016) was downloaded and installed on a Scientific Linux virtual machine. Packaged tests ran but the software would not run against the OSN data corpus examined here.

It is probable that more time spent with either, or both, of these software packages would eventually have yielded results. However, both distributions are open-source projects and, as such, the onus is on the user to attempt to solve installation or setup problems. Neither system offered dedicated support and one of them (the Edinburgh Geoparser) is now a 'retired' project. Results presented in Chapter 5 (p186) therefore rely upon NLP-based data augmentations produced using GATEcloud.net, AlchemyAPI and CLAVIN software.

4.5 Data analysis

All of the data feeding into the Oracle 12c RDBMS used in this research (Section 4.3.1.3, p145), from case study data acquisition (Section 4.2.5, p134) and data augmentation (Section 4.4.1, p147), must be queried, tabulated and analysed to produce results. A database management system has no practical utility if it can only store data; files and file systems do that. This section details the range of data query, tabulation and analysis (Section 4.5.1, p159), data visualisation (Section 4.5.2, p161) and statistical software (Section 4.5.3, p163) used to manipulate stored data. Measurements of geographicality resulting from these analytical methods are described in Section 4.6 (p164).

4.5.1 Data query, tabulation and analysis

In 1972 Codd set out to 'define a collection of operations on relations [or a] relational algebra' which could be used to build a 'query language'. He stated that 'In a practical environment [this query language] would need to be augmented by a counting and summing capability, together with the capability of invoking any one of a finite set of library functions tailored to that environment' (Codd, 1972, p1). This work led directly to the creation of Structured Query Language (SQL) which has been used extensively in database management systems, whether relational or not, ever since.

The two case study data sets, with added DataSift source indicator (`STREAM`), sequential numeric identifier (`STREAMID`) and Universally Unique Identifier (`UUID`) fields held in the main Oracle 12c table `INTERACTIONS`, is comprised of 1,196,671,480 data points. Clearly it is impossible to analyse a >1 billion cell matrix without using sophisticated computer systems. Relational Database Management Systems (RDBMS) and Structured Query Language (SQL) have been developed to address large, complex, tasks of this type. As Wolfram (2006, p301) explains, 'DBMSs are primarily used to house structured textual and numeric data that have been compartmentalized into records and fields. This compartmentalization facilitates access and retrieval of data contents. Through the use of the SQL data manipulation language, one is able to readily summarize and process content. The power associated with these capabilities makes relational DBMSs and, specifically, SQL well suited for the storage and manipulation of informetric data.'

SQL is particularly 'well suited' to generating descriptive counts and aggregations, using the `COUNT... GROUP BY...` syntax. In practice, many data investigations start by counting records, since '*aggregation* [offers] the ability to summarize information' (van Renesse, 2003, p87, author's italics). Initial data investigation in this research followed the 'manageability' practice described by Knobbe, Siebes, & Marseille (2002), using standard SQL constructs such as `COUNT`, `COUNT`

`DISTINCT`, `MIN`, `MAX`, `SUM`, `AVG` and, significantly in the context of heavily-skewed OSN data (Section 6.4.3, p255), `MEDIAN`.

Upon successful completion of ETL data import processes (Section 4.3.1.3, p145), which themselves relied upon many SQL queries to check data consistency along the way, the first query run against the newly-created database counted the number of records in the `INTERACTIONS` table (Appendix 11 listing 9, p480).

Table 4-5 – Count of Interactions by Stream

STREAM	COUNT
US2012_GEO	146,424
US2012_NON_GEO	1,560,967
US2012_NON_GEO_HISP	11,276
SCOT2014	6,477,713

This simple statement was swiftly followed by a slightly more complicated SQL query designed to count the number of records by Stream (Appendix 11 listing 10, p480), yielding the result set shown in Table 4-5. On a laptop equipped with Solid State Disks (SSD)s, using a function-based bitmap index (Oracle, 2016b) on the `STREAM` field, the query ran in 2.104 seconds over 8,196,380 records. Many SQL queries of this type, some taking much longer to compute, or running within looping PL/SQL programmes (Feuerstein & Pribyl, 2005), have been issued as part of the investigatory process and are referenced throughout this thesis (Appendix 11, p479). Queries have been designed to assess sparsity, coordinate and toponymic geographicality, temporality, spatiotemporality and skewness in the research data corpus of OSN interactions. Output has been visualised mainly through the use of Tableau and various GIS packages, together with Gephi graph analysis software. These software systems are described in the following section.

4.5.2 Data visualisation

Most of the maps and charts in this thesis, excluding the maps in Section 6.4.4 (p262) produced with QGIS (2018), have been generated using Tableau (2017b) Desktop Professional Edition versions 8.2 through 10.5 running under 64-bit Microsoft Windows Server 2012 R2 or Windows 10 Professional. Tableau (2017a) is made freely available to academic users having grown out of a data-based visualisation project ('Polaris') at Stanford University (Stolte, Tang, & Hanrahan, 2002). Presenting Polaris, the authors stated that:

In the last several years, large multidimensional databases have become common in a variety of applications, such as data warehousing and scientific computing. Analysis and exploration tasks place significant demands on the interfaces to these databases. Because of the size of the data sets, dense graphical representations are more effective for exploration than spreadsheets and charts. Furthermore, because of the exploratory nature of the analysis, it must be possible for the analysts to change visualizations rapidly as they pursue a cycle involving first hypothesis and then experimentation.

(Stolte et al., 2002, p52)

Polaris was designed 'to discover structure, find patterns, and derive causal relationships' in large databases. The software featured tight integration between visual design and data query, achieved through the use of Pivot Tables and 'n-dimensional data cubes [where] each dimension in [the] cube corresponds to one dimension in the relational schema' (Stolte et al., 2002, p52). The design of Polaris exploited many of the graphing techniques (use of size, shape, colour etc.) first developed in Bertin's (1967) *Semiology of Graphics*, later codified in Wilkinson's (1999) computerised *Grammar for Graphics*. Now commercialised as Tableau, this system (VizQL, or Visual Query Language) allows the user to '[make] interactive

data visualization an integral part of understanding data' (Tableau, 2017c) through the use of a 'drag and drop' interface which translates visual requirements ('draw a map', 'plot a line graph') into standard SQL queries, which can be executed against a large number of databases, including the Oracle 12c RDBMS used here.

Tableau is not a conventional Geographic Information System (GIS), but does have mapping capabilities. Point plotting of Latitude/Longitude data and geocoding, with display against an OpenStreetMap (2017) backdrop, has been available since version 8. Version 10, released in 2017, supports inclusion of third-party maps in Keyhole Markup Language (KML), ESRI Shapefile, MapInfo (Tables and MapInfo Interchange Format) and GeoJSON formats (Marten, 2017). The software also features 'Paging', the computerised animation of time-series data – including spatiotemporal data – which has been used to temporally visualise coordinate-geotagged OSN interactions. Even on a commodity SSD-equipped laptop, also running the Oracle 12c RDBMS, Tableau can rapidly produce graphical output from large (~8 million row) database tables or views. As the software has evolved, a wide range of connections to other database servers have been added, including major NoSQL software releases such as MapR Hadoop Hive and MongoDB mentioned earlier (Section 4.3.1.2, p140). This combination of features has made Tableau a particularly useful component of the exploratory spatiotemporal data analysis and visualisation methodology adopted in this research (Section 3.3, p102).

Aside from Tableau, the usual range of desktop computing applications (e.g., Microsoft Excel) have been used together with QGIS (2018) and MapInfo (Pitney Bowes, 2018) GISs. One other specialist scientific computing package, Gephi (2018a), has also been used. This application, which describes itself as 'the leading visualization and exploration software for all kinds of graphs and networks' has proven particularly useful when analysing or visualising social network graphs, e.g., Twitter mentions relationships (Figure 4-5, p130; Figure 6-21, p277). The software stems from academic research (Bastian et al., 2009) and encodes several algorithms

commonly used in graph network analysis (Jacomy, Venturini, Heymann, & Bastian, 2014; S. Martin et al., 2011) in a flexible and reasonably easy-to-use package.

4.5.3 Statistical tests

A series of 80 Oracle 12c database views, created in SQL (e.g., Appendix 11 listing 37, p491), output numbers of NLP-detectable toponymic mentions in OSN interaction message text and linked/shared URL content for further analysis in R statistical computing software (The R Foundation, 2018). Paired (Welch's) T-tests were used to measure statistical significance (Spector, 2018).

Welch's T-test is a variant of Student's T-test and is more reliable when the two samples for comparison exhibit unequal variances and unequal sample sizes. This is the case in the current research where numbers of NLP-detectable toponymic mentions vary substantially, according to interaction or user categorisation (e.g., coordinate-geotagged/geotagging or not) and by OSN source/subtype (i.e., Facebook post, Twitter tweet or retweet) within event. For example, at least 1 and at most 706 toponymic mentions per interaction were detected in non-coordinate-geotagged Facebook interactions by GATEcloud in the SCOT2014 data set (Table A12-3, p506). Within events there are also sizeable discrepancies in 'like-for-like' sample sizes, e.g., in the US2012 data set GATEcloud detected toponymic mentions in 125,758 Twitter tweets; 21,455 of which were coordinate-geotagged and 104,303 were not (Table A12-1, p504).

R scripts computed descriptive statistics (Appendix A12.1 listing 2, p494), T and P scores (Appendix A12.1 listing 3, p494) for numbers of NLP-detectable toponymic mentions in Facebook (FB), Twitter tweet (TW) and retweet (RT) message text or linked/shared URL content, whether coordinate-geotagged (GEO=Y) or not (GEO=N), for each of the three NLP/geoparser systems described in Section 4.4 (p147); TwitIE on GATEcloud (GT), AlchemyAPI (AL against message text; LI against links) and CLAVIN-rest (CL). These computations are repeated for the two

case study electoral events, US2012 and SCOT2014. Summary statistics are presented in Section 5.3 (p219) while detailed statistical results and commentary are presented in Appendix A12.2 (p502).

4.6 Data measurements

The data subjects (Section 4.2, p119), preparations (Section 4.3, p135), procedures (Section 4.4, p147) and analysis methods (Section 4.5, p158) detailed in the preceding sections of this chapter were used to measure and score ‘geographicality’ in case study OSN interactions. The following section describes the measurement and scoring process in detail. Geographicality scores, at interaction and user levels, are reported upon further in Chapter 5 (p186) which presents the results of this investigation; answering the three research questions formulated earlier in this thesis (Section 1.7, p34) to determine whether coordinate-geotagged social media interactions and the users that create them, the most *spatially* expressive user class, are also the most *geographically* expressive in terms of toponymic mentions of place.

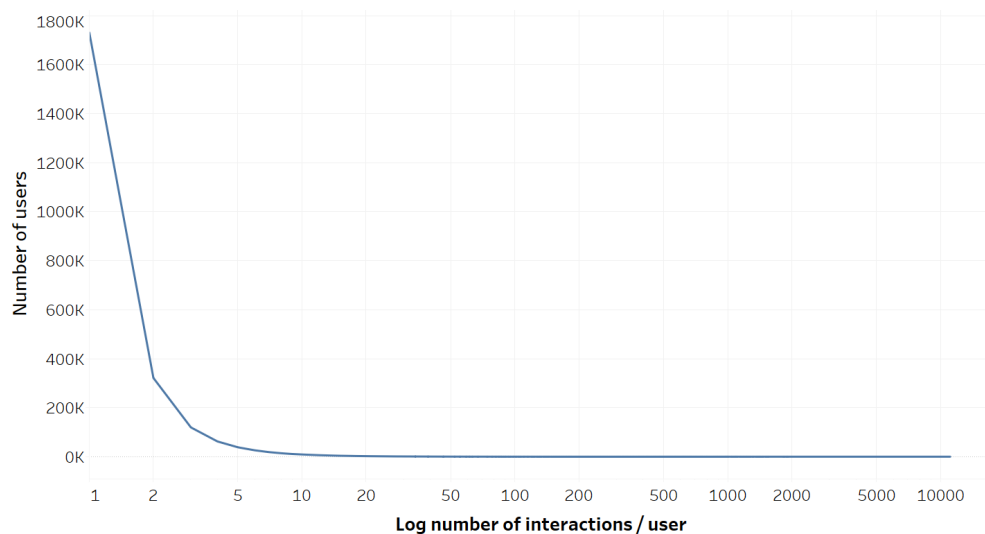
4.6.1 Measuring and scoring ‘geographicality’ in OSN data

The ~8 million record research data corpus examined in this study is comprised of a mixture of Facebook posts and Twitter tweets and retweets sampled in 2012 (US2012) and 2013-14 (SCOT2014). Table 4-6 (p165) shows the number and percentage of OSN interactions by source, subtype and event. In the US2012 data set Facebook posts comprise just 3.33% (n=57,265) of all interactions sampled (n=1,718,667). In the SCOT2014 data set 12.12% (n=785,237) of all interactions sampled (n=6,477,713) are Facebook posts. Interactions sourced from Facebook comprise 10.28% (n=842,502) of the research data corpus overall (n=8,196,380) while interactions sourced from Twitter comprise the remaining 89.72% (n=7,353,878), split reasonably evenly between Twitter tweet (n=3,712,847) and retweet (n=3,641,031) subtypes.

Table 4-6 – Number and percentages of OSN Interactions by source, subtype and event

Source	Facebook Post	Twitter Tweet	Twitter Retweet	Total
n US2012	57,265	866,160	795,242	1,718,667
% US2012	3.33%	50.40%	46.27%	100.00%
n SCOT2012	785,237	2,846,687	2,845,789	6,477,713
% SCOT2014	12.12%	43.95%	43.93%	100.00%
n TOTAL	842,502	3,712,847	3,641,031	8,196,380
% TOTAL	10.28%	45.30%	44.42%	100.00%

Data-mining in SQL (Appendix 11 listing 11, p480) shows that most users ($n=1,730,748$; 71.04%) make only one interaction, 89.14% three or fewer (cumulative $n=2,171,589$), and 93.26% five or fewer interactions (cumulative $n=2,271,917$) in the research data corpus (Figure 4-14). Users making ≤ 5 interactions created 3,171,447 or 38.69% of all interactions (Appendix 11 listing 12, p480).

**Figure 4-14** – Log number of interactions/user by number of users in the research data corpus

The remaining 164,250 users (6.74% of $n=2,436,167$ total users) created 61.31% ($n=5,024,933$) of all interactions. Figure 4-14 shows that across the research data corpus the distribution of numbers of interactions/user is heavily skewed. However,

skewness by OSN source, event and subtype differs substantially (Figure 4-15) and is somewhat lower amongst coordinate-geotagging users.

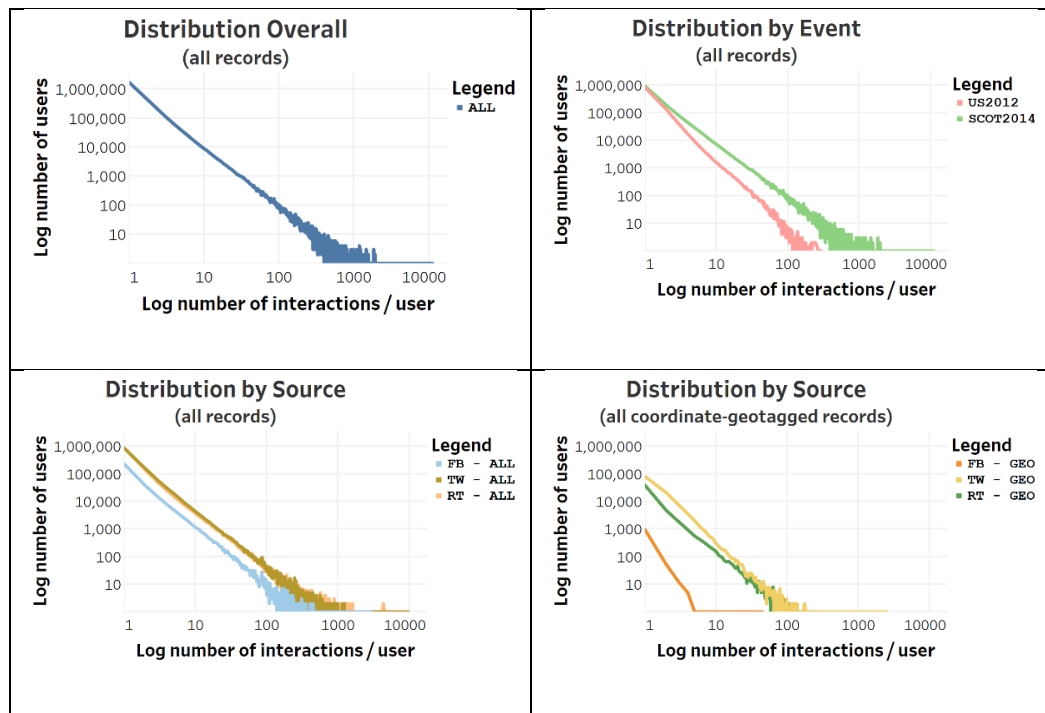


Figure 4-15 – Log number of interactions/user by Log number of users: overall, by event and by OSN source (all records; all coordinate-geotagged records)

Table 4-7 (p167) shows descriptive statistics for interactions/user in the entire research data corpus (ALL) further broken down by event (US2012 and SCOT2014) and OSN source/subtype, both for all user interactions by OSN source (e.g., Facebook posts=FB-ALL, Twitter tweets=TW-ALL, Twitter retweets=RT-ALL) and for all user coordinate-geotagged interactions by source/subtype (FB-GEO, TW-GEO and RT-GEO). Statistics were created using the `psych` package (Revelle, 2018) in R (The R Foundation, 2018) from a script (Appendix A12.1 listing 1, p493) reading counts of numbers of interactions/user derived from the main INTERACTIONS table in the Oracle 12c database using SQL (Appendix 11 listing 13, p480). Table 4-7 (p167) shows that the median number of interactions/user across both events (ALL) is 1, with a maximum number of 11,046 interactions/user recorded during the SCOT2014 event. Skewness and kurtosis are discussed in more depth overleaf.

Table 4-7 – Descriptive statistics for interactions/user in the research data corpus (ALL), by event (US2012 and SCOT2014) and by source (FB=Facebook posts, TW=Twitter tweets, RT=Twitter retweets) whether all interactions (-ALL) or only coordinate-geotagged (-GEO)

	n	mean	sd	median	min	max	range	skew	kurtosis	se
ALL	2,436,167	3.4	32.7	1	1	11,046	11,045	112	22,315	0.02
US2012	1,060,163	1.6	3.3	1	1	295	294	21	761	0.00
SCOT2014	1,424,087	4.6	42.7	1	1	11,046	11,045	87	13,230	0.04
FB-ALL	318,688	2.6	13.4	1	1	2,881	2,880	70	10,213	0.02
TW-ALL	1,216,471	3.0	29.2	1	1	9,623	9,622	142	31,565	0.03
RT-ALL	1,170,795	3.1	29.0	1	1	6,482	6,481	78	10,296	0.03
FB-GEO	1,052	1.1	1.5	1	1	45	44	24	647	0.05
TW-GEO	121,045	2.1	9.4	1	1	2,503	2,502	170	42,271	0.03
RT-GEO	50,120	2.0	5.5	1	1	300	299	17	507	0.02

Skewness ranges from 17 (RT-GEO, coordinate-geotagged Twitter retweets) to 170 (TW-GEO, coordinate-geotagged Twitter tweets) reflecting substantial differences in the distribution of numbers of interactions/user in each case (Table 4-7). High kurtosis values (up to 42,271 for TW-GEO and 31,565 for TW-ALL) show, as does (Figure 4-14, p165), a ‘heavy’ rightwards or ‘long-tailed’ distribution of interactions/user.

Further analysis in SQL (Appendix 11 listing 14, p484) showed that 333 users made $\geq 1,000$ interactions in the research data corpus, accounting for 653,782 interactions comprised of 10,892 Facebook posts, 292,440 Twitter tweets and 350,450 retweets. One Twitter user created the maximum number of 11,046 interactions/user; 4,564 tweets and 6,482 retweets. Figure 4-16 (p168), a binned log-log histogram, shows the tail-off in prolifically interacting users occurs at around 1,000-2,000 interactions per user. Although behaviour of this type is often indicative of roboticised posting (Blank & Lutz, 2017; Howard et al., 2018; Silva et al., 2013) the account in question (Mulder1981) is both a real user and a particularly prolific Twitter tweeter,

especially active during the 2014 Scottish Independence Referendum campaign (The Herald, 2017). The same user is also responsible for the maximum number of 2,503 coordinate-geotagged Twitter tweets (TW-GEO in Table 4-7, p167) recorded in the SCOT2014 data set, many of which were coordinate-retweeted by other Twitter users throughout Scotland allowing measurement of dispersal effects (see, Section 6.4.2, p251 and Figure 6-4, p253).

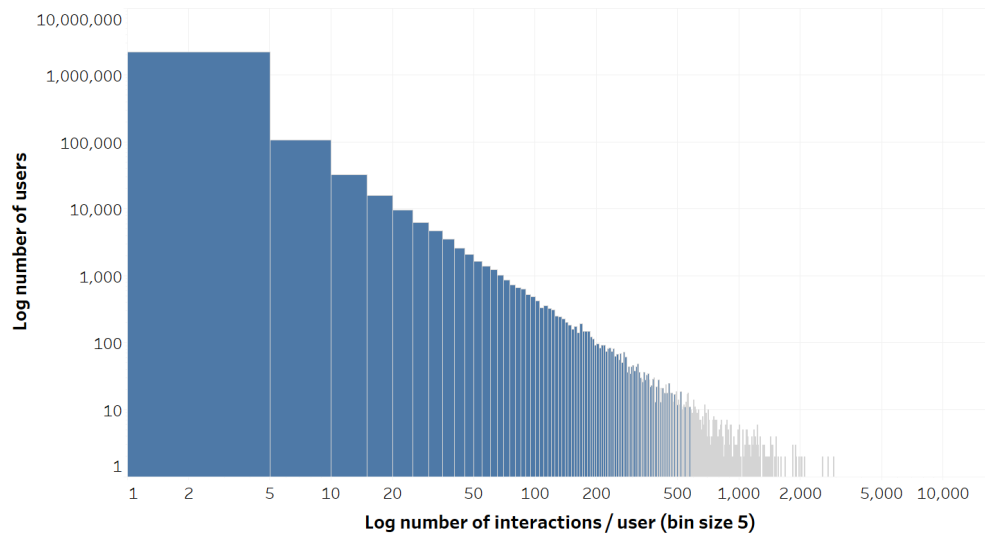


Figure 4-16 – Histogram of Log number of interactions/user against Log number of users

Amongst the 333 most prolific users, making $\geq 1,000$ interactions each (Figure 4-17), there was a mixture of those who tweeted ($n=52$, 15.62%) or retweeted exclusively ($n=4$, 1.20%) and those, in the majority, who both tweeted and retweeted ($n=270$, 81.08%). The second most prolific Twitter user made 9,623 interactions in the SCOT2014 event, all of them tweets comprised of just 91 distinct messages, and was probably posting robotically. Just 7 users (2.10% of prolific users) on Facebook account for $\geq 1,000$ interactions/user and many of their messages contained duplicate text. Skewness is common in OSN data and can pose problems for analysis (Lerman et al., 2018; A. Smith & Gaur, 2018). Various tests in SQL and R address this issue and the subject of skewness is returned to again in Section 6.4.6 (p279).

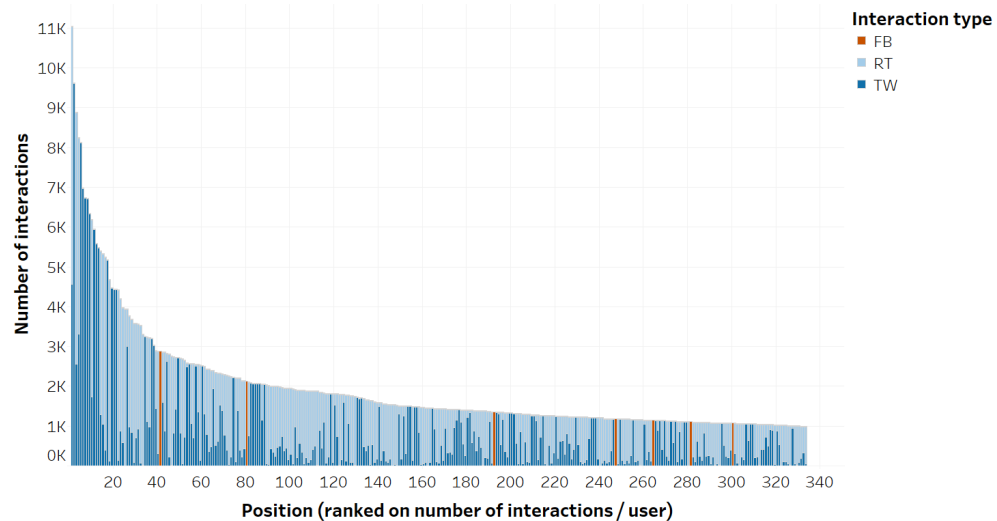


Figure 4-17 – Number of interactions by type created by prolific social network users ranked on number of interactions/user ($\geq 1,000$ interactions/user)

Only 1,231 Facebook interactions are coordinate-geotagged, all of which were sampled during the SCOT2014 event (Table 4-8, p170). The other 355,714 coordinate-geotagged interactions were comprised of Twitter tweets or retweets sampled by the various Streams shown in Table 4-8 and detailed in Appendix 7 (p432). Overall, 356,945 (or 4.35% of) OSN interactions held coordinates but many of these ($n=146,424$ or 41.02%) were recorded by the US2012_GEO Stream which was deliberately designed to sample *only* coordinate-geotagged interactions (Appendix A7.2.1, p433). This inflates the percentage of coordinate-geotagged Twitter tweets to 6.82% across the research data corpus, where it would otherwise have been 3.00%, and the overall rate to 4.35% where it would otherwise have been 2.62%.

The percentage of coordinate-geotagged retweets (2.81%) or Facebook posts (0.15%) is unchanged as neither source or subtype was sampled by this Stream. Geotagging rates by Stream are slightly lower, ranging 1.08-2.90% for any type of coordinate-geotagged interaction, and are shown in Table 4-1 (p128, US2012) and Table 4-2 (p132, SCOT2014) earlier in this chapter. All interactions in the

US2012_GEO Stream were, of course, coordinate-geotagged, although 85 records exhibited useless 0 Latitude, 0 Longitude coordinates.

Table 4-8 – Coordinate-geotagged Facebook posts, Twitter tweets and retweets by Stream and across the entire research data corpus

US2012	<i>n</i> Posts	<i>n</i> Tweets	<i>n</i> Retweets	TOTAL
US2012_GEO ⁽¹⁾	0	146,424	0	146,424
US2012_NON_GEO	0	14,411	8,013	22,424
US2012_NON_GEO_HISP	0	99	23	122
Geotagged total	0	160,934	8,036	168,970
Geotagged total excl. ⁽¹⁾	0	14,510	8,036	22,546
US2012_GEO ⁽¹⁾	0	0	0	0
US2012_NON_GEO	57,210	697,871	783,462	1,538,543
US2012_NON_GEO_HISP	55	7,355	3,744	11,154
Non-geotagged total	57,265	705,226	787,206	1,549,697
TOTAL	57,265	866,160	795,242	1,718,667
TOTAL excl. ⁽¹⁾	57,265	719,736	795,242	1,572,243
% GEOTAGGED	0.00%	18.58%	1.01%	9.83%
% GEOTAGGED excl. ⁽¹⁾	0.00%	2.02%	1.01%	1.43%
SCOT2014	<i>n</i> Posts	<i>n</i> Tweets	<i>n</i> Retweets	TOTAL
Geotagged total	1,231	92,437	94,307	187,975
Non-geotagged total	784,006	2,754,250	2,751,482	6,289,738
TOTAL	785,237	2,846,687	2,845,789	6,477,713
% GEOTAGGED	0.16%	3.25%	3.31%	2.90%
RESEARCH DATA CORPUS	<i>n</i> Posts	<i>n</i> Tweets	<i>n</i> Retweets	TOTAL
Geotagged total	1,231	253,371	102,343	356,945
Geotagged total excl. ⁽¹⁾	1,231	106,947	102,343	210,521
TOTAL	842,502	3,712,847	3,641,031	8,196,380
TOTAL excl. ⁽¹⁾	842,502	3,566,423	3,641,031	8,049,956
% GEOTAGGED	0.15%	6.82%	2.81%	4.35%
% GEOTAGGED excl. ⁽¹⁾	0.15%	3.00%	2.81%	2.62%

The coordinate-geotagging rates reported here are slightly higher than findings reported in the 2012 French Presidential Election technical proof of concept exercise (1.39%; Section 4.2.3, p123) but are broadly in line with, and usefully corroborate, results from Leetaru et al. (2013, Table 4) who report, in their research into over 1.5 billion Twitter interactions, ‘1.6 percent having Exact Location’ and coordinate-geotagged retweet rates as high as 4.57% in New York and 2.98% in London.

The choice of base used to calculate geotagging rates – research data corpus including or excluding the US2012_GEO Stream or geotagged Twitter tweets against total Twitter tweets etc. (Table 4-8, p170) – alters these percentages just as the choice of base, e.g., UK, England, Scotland etc. does when calculating Census-based population profiles (Section 6.4.4, p262) or other data of this type.

Table 4-9 – Numbers and percentages of original (Facebook posts and Twitter tweets) and reposted (Twitter retweet) coordinate-geotagged interactions in the research data corpus

Type	Posts + Tweets	Retweets	Total
n Geotagged	254,602	102,343	356,945
n Not geotagged	4,300,747	3,538,688	7,839,435
TOTAL	4,555,349	3,641,031	8,196,380
OVERALL % Geotagged	3.11%	1.25%	4.35%

To avoid recalculation for different bases the summary totals and percentages shown in Table 4-9, above, ('posts+tweets' and 'retweets' over 'all records') are adopted in the following pages.

Rates of coordinate-geotagging are low in the case study data sets. SQL queries, counts and further examination of the stored data show that:

1. **Spatiality in OSN interactions is binary** – users either do, or do not, choose to record coordinates alongside their Facebook posts or Twitter tweets.
2. **Spatiality in Twitter retweets forms an unusual case** – Sloan & Morgan (2015) note that 'Retweets generated by invoking the retweet command in the Twitter user interface are not classed by Twitter as original content and are never geotagged. However, retweets generated by copying and pasting the content of a tweet into the tweet-composition box are classed as original content and can be geocoded (if the user chooses).' There are many retweets (n=102,343) of this type in the research data corpus (Table 4-8; Table 4-9) and these are addressed separately in Section 6.4.2 (p251).

3. **Geographicality in OSN interactions is scalar** – Varying amounts of geographically referenceable material are created by all users, whether coordinate-geotagging or not. Message text and metadata, particularly in the more numerous fields available in OSN records sourced from Twitter, hold information with potential geographical value, including several of the ambient ‘geographic footprints’ mentioned in the literature by Q. Huang, Cao, & Wang (2014) and Stefanidis, Crooks, et al. (2013).

To determine how ‘geographicality’ can more accurately be assessed, aside from simply calculating counts of explicitly spatially-referenced interactions, it is necessary to examine OSN metadata. This exhibited substantial disparity by OSN source; interactions sourced from Twitter were accompanied by many metadata fields while Facebook posts were not. Hence, it is also necessary to ‘mine’ OSN message text (Stock, 2018) and linked/shared URL content from both sources to detect toponymic mentions of place. Data augmentation of this type, using the NLP/geoparsing tools introduced earlier (Section 4.4, p147), can determine whether coordinate-geotagging or non-coordinate-geotagging users create or link to content making the most mention of NLP-detectable geographical entities. These results are presented in the following Sections 5.2.2 (p190) and 5.2.3 (p205) once the framework for calculating ‘baseline’ Geographicality Scores has been set out.

Only 21 fields holding Facebook metadata and 50 holding Twitter metadata were apparent in the 146 fields imported from DataSift CSV files into the main `INTERACTIONS` table stored in Oracle 12c (Appendix A9.2, p443). The remaining 75 fields held message text, creation date/time or other types of data (e.g., linked/shared URLs) common to either OSN platform, together with a number of DataSift augmentations including gender and salience discussed earlier (Section 4.2.3, p123). Many of the OSN metadata fields record `NULL`, no data, values in rows. Just 33.01% of the 1,196,671,480 cells in the `INTERACTIONS` table contained values; the remaining 801,595,515 cells (66.99%) were null. Sparsity

analysis, more fully detailed later in Section 6.4.3 (p255), was used to determine which fields in the data set warranted examination as some fields were so sparsely populated (>99.99% null) that any analysis based upon their content would have been almost entirely worthless. Several metadata fields contained potentially useful geo-information (e.g., toponyms), but only for a subset of records, while row-based sparsity levels for all metadata fields varied substantially across Streams.

Metadata fields storing Potential Geographic Information (PGI) were identified and scored using four integer values (0=None, 1=Low, 100=High, 200=High) through SQL analysis and exploratory data-mining. High value scores were given to original (100) or re-posted (200) Latitude and Longitude coordinate pairs, which allow straightforward mapping, and low value scores to any of the other PGI metadata fields requiring significant post-processing to extract geographical information from stored data, at highly variable levels of spatial resolution. The interaction message text itself, stored in the `INTERACTION_CONTENT` field, was not coded in this exercise as all message text may contain PGI and the search for this is considered separately, alongside RQ2, in Section 5.2.2 (p190).

The scoring scheme was designed to produce a scalar ‘Geographicality Score’ for non-coordinate-geotagged interactions by summing integer scores for non-null fields across rows for each interaction in the database. Original and re-posted coordinate-geotagged interactions were given higher scores, of 100 and 200 respectively, alongside any other PGI metadata fields so that a score for a coordinate-geotagged tweet could, e.g., end up at 107, or a retweet at 209. Higher integer scores for these two types of coordinate-geotagged interactions helped separate them into distinct classes for later analysis. PGI metadata fields and scores were recorded in a Microsoft Excel spreadsheet containing column/field names from the `INTERACTIONS` table and various other calculations derived from SQL queries, e.g., % non-null rows across the entire data set for each field. The total score for any given interaction was designed to replicate programmatic strategies

frequently adopted when developing a geoparser (F. Liu, Vasardani, & Baldwin, 2014; Shi & Barker, 2011; Zhang & Gelernter, 2014), e.g., if Field₁ exists then check Field₂; if Field₂ and Field₃ are not null check for coordinates in Field₄ etc.

The integer scores assigned to each of the 35 identified PGI metadata fields are shown in Table 4-10 alongside a commentary detailing potential utility.

Table 4-10 – Coding scores for 35 Potential Geographic Information (PGI) metadata fields

PGI metadata field	Comment	Score
TW_RT_USER_GEO_ENABLED	LOW VALUE	1
TW_RTED_USER_GEO_ENABLED	LOW VALUE	1
TW_RTED_USER_TIME_ZONE	LOW VALUE	1
TW_RTED_USER_UTC_OFFSET	LOW VALUE	1
TW_RTED_USER_LOCATION	LOW VALUE (Mars etc.), NEEDS NLP	1
TW_RT_USER_LOCATION	LOW VALUE (Mars etc.), NEEDS NLP	1
TW_USER_LOCATION	LOW VALUE (Mars etc.), NEEDS NLP	1
TW_USER_TIME_ZONE	LOW VALUE	1
TW_USER_UTC_OFFSET	LOW VALUE	1
TW_RT_USER_TIME_ZONE	LOW VALUE	1
TW_RT_USER_UTC_OFFSET	LOW VALUE	1
INTERACTION_GEO_LATITUDE	IF SO GREAT, HALF THE HIGH SCORE	50
INTERACTION_GEO_LONGITUDE	IF SO GREAT, HALF THE HIGH SCORE	50
TW_GEO_LATITUDE	IF SO GREAT, BUT DO NOT DOUBLE COUNT	0
TW_GEO_LONGITUDE	WILL BE IN INTERACTION_GEO_LATITUDE	0
TW_PLACE_FULL_NAME	CLEAN DATA, NEEDS NLP	1
TW_PLACE_ID	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_PLACE_NAME	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_PLACE_PLACE_TYPE	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_PLACE_URL	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0

TW_PLACE_COUNTRY	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_PLACE_COUNTRY_CODE	RELATED TO TW_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_COUNTRY	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_COUNTRY_CODE	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_FULL_NAME	CLEAN DATA, NEEDS NLP	1
TW_RTED_PLACE_ID	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_NAME	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_PLACE_TYPE	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_PLACE_URL	RELATED TO TW_RTED_PLACE_FULL_NAME, DO NOT DOUBLE COUNT	0
TW_RTED_GEO_LATITUDE	VERY HIGH VALUE (GEO DISPERSAL)	100
TW_RTED_GEO_LONGITUDE	IF SO GREAT HALF THE VERY HIGH SCORE	100
TW_PLACE_ATT_ST_ADDRESS	CLEAN DATA, NEEDS NLP	1
TW_RTED_PLACE_ATT_ST_ADDRESS	CLEAN DATA, NEEDS NLP	1
TW_PLACE_ATT_LOCALITY	TOPONYMIC, MAY NEED LOOKUP	1
TW_PLACE_ATT_REGION	TOPONYMIC, MAY NEED LOOKUP	1

As Table 4-10 shows, the scoring system was only applicable to OSN interactions sourced from Twitter. Facebook posts sampled during the 2012 US Presidential Election, and in 2013-2014 during the 2014 Scottish Independence Referendum campaign, held no PGI metadata capable of generating such a scoring scheme, except for 1,231 coordinate-geotagged interactions (Table 4-8, p170), scored at 100, sampled by the SCOT2014 Stream.

In designing a Geographicality Score for interactions, or aggregating interaction level scores to calculate modal scores for each user, the presence of Potential Geographic Information in metadata fields enables the calculation of summary distributions of geographicality via SQL constructs without the need to develop a geoparser, several of which already exist (Gritta et al., 2018) and have been used here to identify toponymic mentions of place in users' message text and linked/shared URL content. Results of this work are presented in Section 5.2.2 (RQ2, p190) and Section 5.2.3 (RQ3, p205) later in this chapter.

As the data in the `INTERACTIONS` table are derived from DataSift's conversion of Twitter (2018) JSON formatted data to CSV format, care must be taken not to 'double-count' RDBMS fields which were originally 'children' of another JSON key 'parent' (see Figure 4-9, p139 and Table 4-4, p146). For example, the field `TW_PLACE_ID` is associated with several related fields and a query run against it to return non-null values (Appendix 11 listing 15, p484) also returns data from these co-populated fields (Table 4-11, p176). Consequently, only one member of a 'grouped' field-set is scored to avoid double-counting. This explains several 0 scores for PGI metadata shown in Table 4-10 (p174), where comments denote how one field is related to another and allocates a score to the lead variable only or, in some cases (e.g., coordinate pairs), splits scores between two metadata fields.

Table 4-11 – Example of `TW_PLACE_ID` co-populated metadata fields

Country	Code	Full Name	Place ID	Place Name	Type
United States	US	New York, NY	27485069891a7938	New York	city
United States	US	West Deptford, NJ	3b524e5ca68b7923	West Deptford	city
United Kingdom	GB	Belfast, Belfast	a5d1791165a6517e	Belfast	city

The same Microsoft Excel spreadsheet used to code metadata field scorings was also used to dynamically construct SQL to create a persistent database view (Appendix 11 listing 16, p484) for the 21 non-zero-coded and scored fields shown in

Table 4-10 (p174). The logic exploits the SQL CASE statement (Oracle, 2018b) using the following syntax for any given metadata [FIELD] and [SCORE]:

```
SELECT UUID,
(case
  when
    (FIELDn is not null) then SCOREn
    else 0
  end) as FIELD_SCOREn
FROM INTERACTIONS
```

The sum of the coding scores computed across each row, testing for nullness by case for each scored PGI metadata field, created an overall Geographicality Score for all interactions in the research data corpus using SQL (Appendix 11 listing 17, p479) the distribution of which is shown in Figure 4-18 (p178). Most interactions (n=8,051,729 or 82.39%) held one or more low-value PGI metadata fields. High value coordinate metadata was evident in 3.11% of records comprised of a mixture of original Facebook posts and Twitter tweets and a further 1.25% was comprised of re-posts in the form of coordinate-geotagged Twitter retweets. The majority of non-coordinate-geotagged interactions (n=6,395,784 or 78.03% of the research data corpus) held 1-9 non-null PGI metadata fields. Just 4 records (0.000049%) held 9 items of PGI metadata; these were grouped with the preceding class (8). Class 100+ is a grouping of original coordinate-geotagged interactions scored 100-105 and 107. Class 200+ is a grouping of re-posted (retweeted) coordinate-geotagged interactions scored 202-210 inclusive. Those interactions with a 0 Geographicality Score (n=1,443,651) were comprised mainly (58.27%) of Facebook posts (n=841,271, i.e. almost all Facebook posts; Table 4-8, p170) which could not be assigned to other classes due to a lack of available PGI metadata. A further 598,544 Twitter tweets (41.46% of the 0-scored interactions) and 3,836 Twitter retweets (0.27%) held no PGI metadata.

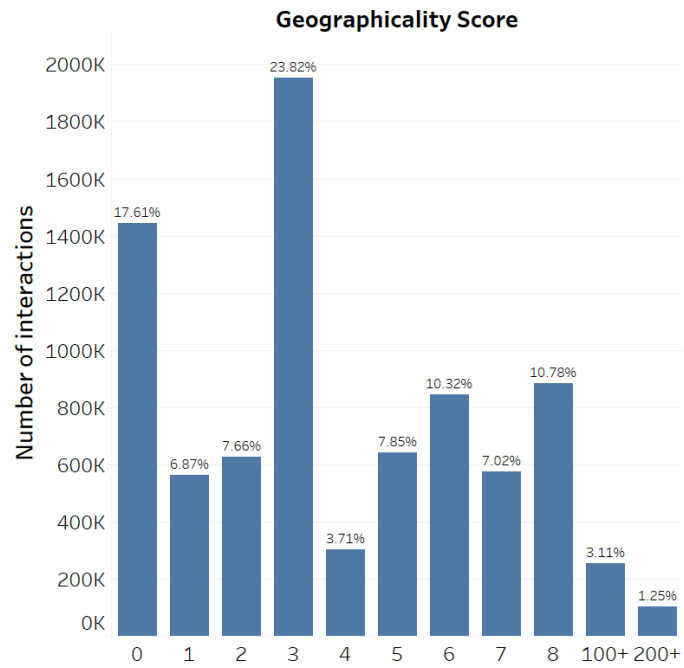


Figure 4-18 – US2012/SCOT2014: Distribution of Geographicality Scores at interaction level across both events

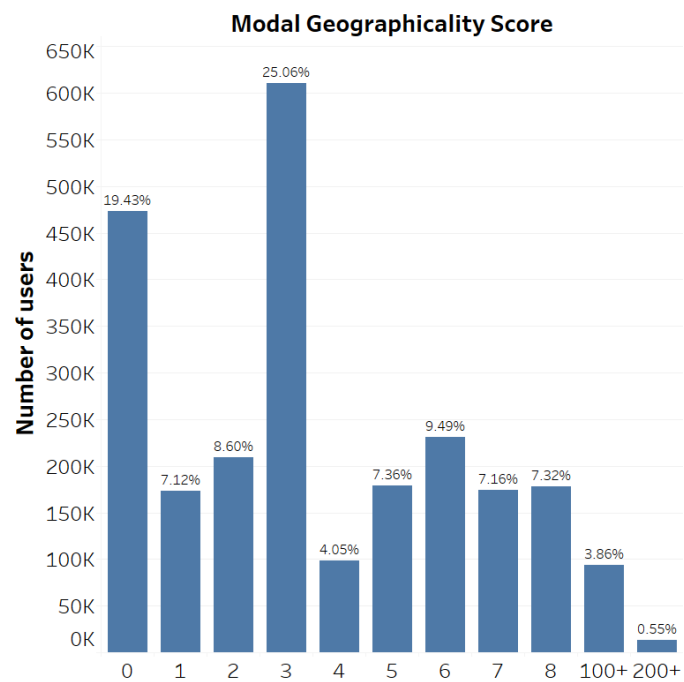


Figure 4-19 – US2012/SCOT2014: Modal distribution of Geographicality Scores at user level across both events

Figure 4-19 (p178) shows the distribution of the Geographicality Score at user level, computed using Oracle's (2017b) `STATS_MODE` function in SQL to score all 2,436,167 users by their modal (most frequent) Geographicality Score at interaction level (Appendix 11 listing 18, p486). The similarity in the distribution of scores at interaction and user levels is explained by OSN posting behaviour observed in the research data corpus; 71.04% of users made just one interaction and 93.26% five or fewer so the distribution of Geographicality Scores at interaction and user levels is bound to be similar.

Despite significant skewness in numbers of interactions/user (Table 4-7, p167) the distribution of modal Geographicality Scores by posting frequency (Figure 4-20, p179) shows strong similarity (Pearson Correlation coefficient 0.96) between infrequent (users making ≤ 5 interactions) and frequent contributors (users making ≥ 6 interactions) to the research data corpus.

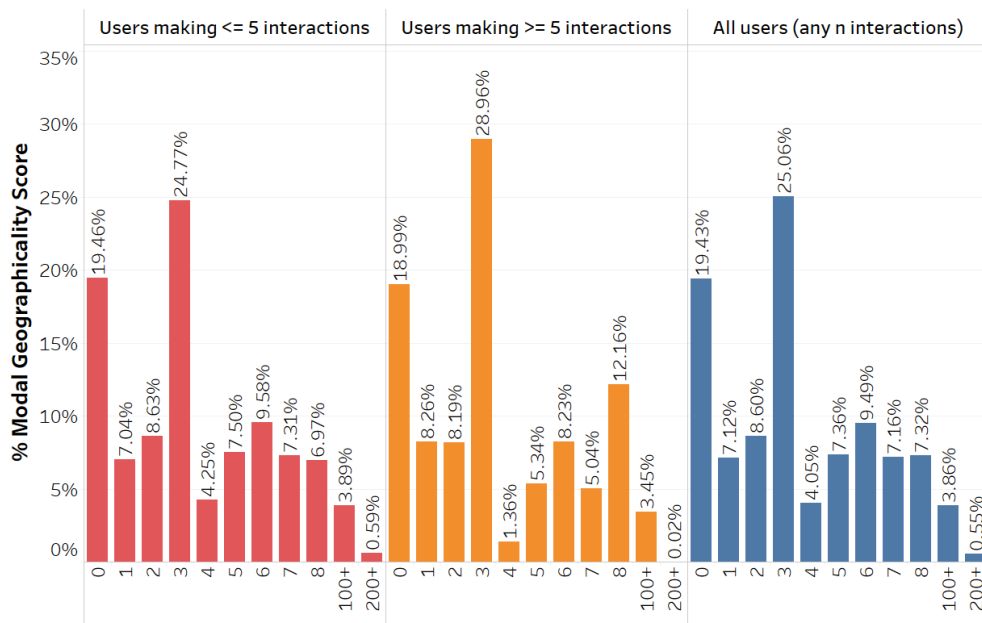


Figure 4-20 – Percentage distribution of Modal Geographicality Scores for users making ≤ 5 interactions, ≥ 6 interactions and any number of interactions in the research data corpus

The high correlation in modal Geographicality Scores between users who made ≤ 5 interactions and all users (Pearson's $r=0.99$) and between those who made ≥ 6

interactions and all users ($r=0.96$) suggests that the scoring scheme is not badly affected by skewness in numbers of interactions/user. The relative contribution of PGI metadata fields to each class of the computed Geographicality Score is shown in Figure 4-21.

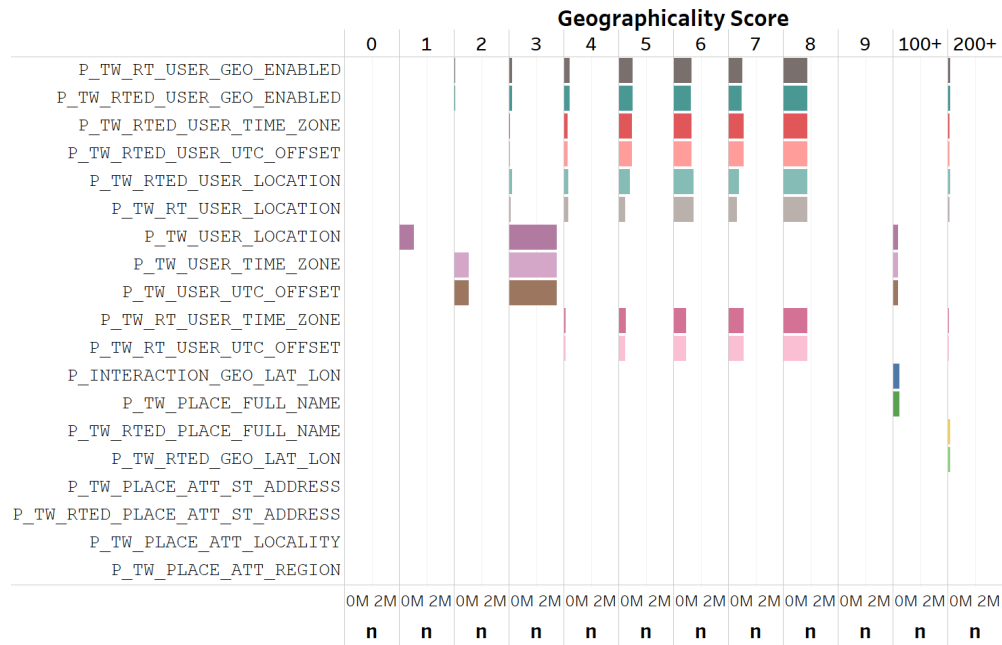


Figure 4-21 – Contribution of PGI metadata fields to each Geographicality Score class

The largest contributor to scores, TW_USER_LOCATION, present in 1,752,373 (21.38% of) interactions and derived from the free-form self-reported user location field on Twitter, was generally ‘assumed [to be] strongly typed geographic information with little noise and good precision’ until Hecht, Hong, Suh, & Chi's (2011, p237) report ‘found that 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools.’ Similar results were evident here, where the TW_USER_LOCATION field, completed in response to a Twitter registration form asking ‘Where in the world are you?’ (Hecht et al., 2011) was found to contain many non-geographical references (e.g., ‘World Wide Web’ or ‘In my skin’) as well as mentions of cities, states, countries and, in some cases, Latitude and Longitude coordinates.

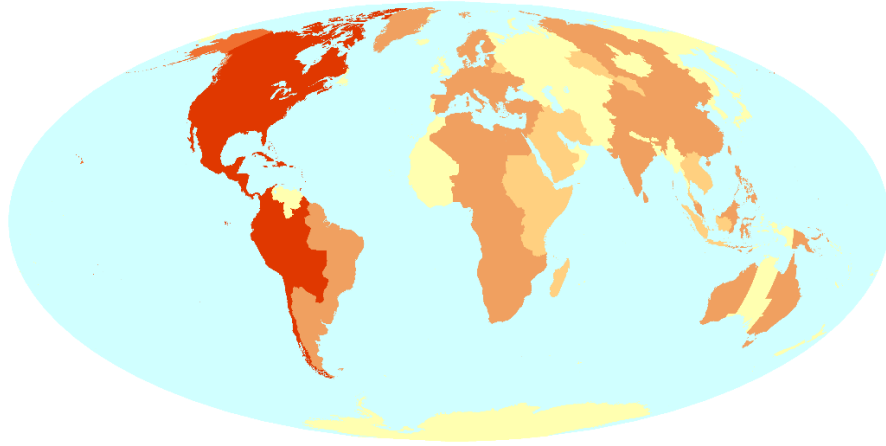


Figure 4-22 – US2012: Interactions mapped by World Time Zones (yellow=low, red=high)

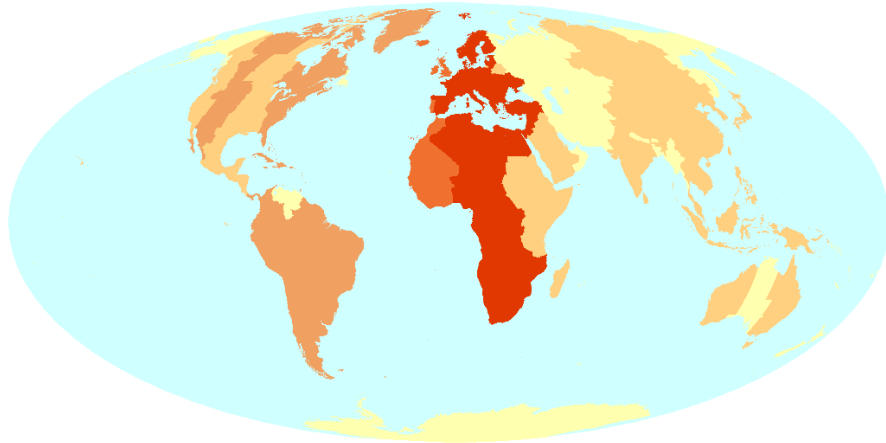


Figure 4-23 – SCOT2014: Interactions mapped by World Time Zones (yellow=low, red=high)

Other highly contributory PGI metadata fields shown in Figure 4-21 (p180) included `TW_USER_TIME_ZONE` ($n=1,752,379$, 21.38% of interactions) and another time-related field, `TW_USER_UTC_OFFSET`, with just two fewer populated records. As Twitter does not adhere to the Olson standards for time zone naming (Internet Assigned Numbers Authority, 2017) analysing or mapping Twitter's named time zones may require significant post-processing (Mahmud, Nichols, & Drews, 2012, 2014) and has not been attempted here. It is possible, however, to aggregate (Appendix 11 listing 19, p486) and map (Figure 4-22 and Figure 4-23, p181)

numbers or percentages of interactions grouped by the `TW_USER_UTC_OFFSET` field. Recorded by Twitter in seconds these values (e.g., 14,400 or -10,800) can be converted into hours and minutes (i.e., +04:00, -03:00), joined to World Time Zone boundaries and mapped. These maps usefully show levels of activity at time zone level for the two case study electoral events but otherwise provide very little additional 'spatial granularity' for further analysis (Dalton & Thatcher, 2015). A similar set of metadata (Figure 4-21, p180) relating to Twitter retweet or retweeted and potentially also erroneous self-reported location fields, alongside further time zone data, furnishes many of the other frequently encountered, but low value, PGI metadata fields found in the `INTERACTIONS` table.

The difficulty inherent in deriving geo-information from Twitter and other OSNs has spurred much research effort in Geography and related disciplines (Section 2.6, p77; References, p319) but is not the focus of this study. Stock (2018) has found that of the 42 methods used to geographically augment OSN data, applied predominantly in the literature to Twitter interactions (the top source of OSN data analysed in 54.2% of 821 surveyed papers) and much less-widely to Facebook posts (in 6th place with just 2.1% of surveyed papers; Stock, 2018, Table 1, p213), techniques which '[extract] place names from messages' have given the best accuracy. These techniques are applied here (Section 5.2.2, p190) not to test the accuracy of any given geoparsing approach, although three systems have been evaluated (Section 5.4, p221), but to test the *Geographicality Assumption* (Section 1.7, p34) often implicitly adopted in the analysis of geotagged OSN messages; that *coordinate-geotagging users are the most geographically expressive of all OSN users*.

There is a general expectation that coordinate-geotagging users of social media platforms will mention place in their messages, and an equally strong expectation that the places they mention will be proximal to the geotagged coordinates of their post. I. L. Johnson et al. (2016) have demonstrated that 'localness' of this type is

exhibited in only ~75% of coordinate-geotagged OSN messages (from Twitter, Flickr and Swarm). Nobody, until now, has asked the more fundamental question; who mentions place *most*; coordinate-geotagging or non-coordinate-geotagging users of OSN sites? The measurement and scoring typology outlined in this section has been used to categorise Twitter tweet and retweet interactions according to the type and value of any Potential Geographic Information held in key metadata fields. The technique does not extend well to Facebook posts, which have no analogous PGI metadata fields and only 1,231 of which are coordinate-geotagged. All Twitter tweets, retweets and Facebook posts may, however, be augmented by detecting geographicality in message text and linked/shared content using NLP/geoparsing techniques. Chapter 5 (p186) presents the results of these text-mining investigations which are reliant upon the research methods detailed above.

4.7 Summary

The role of the Data Scientist is apparently ‘The Sexiest Job of the 21st Century’ (Davenport & Patil, 2012) as Castells’ (1996, 2009) *Rise of the network society* creates a data-driven ‘information age economy’. Popular online ‘Cheat Sheets’ (Ohri, 2014) for budding Data Scientists encourage their readers to ‘write code, understand statistics [and] derive insights from data’ detailing a number of skills or technologies which these titans of the new information age should command, including ‘R, Python, Java, SQL, Hadoop (Pig, Hive Query Language, MapReduce) etc.’ Data Scientists have been described as ‘Engineers of the Future’ (van der Aalst, 2014); analytical experts (Agarwal & Dhar, 2014) offering improved ‘data-driven’ decision-making (Baesens, 2014) and predictive insight into human or machine behaviour (Dhar, 2013).

The availability of large data sets appears to be revealing, as Zelenkauskaitė & Bucy (2016) have argued, an ‘emerging [Kuhnian] research paradigm [in the growth of] computational social science’ which is exposing a ‘scholarly divide’ between those with the resources and skills needed to access, handle and store Big Data and those

without. Geographic data have always been ‘big’ (S. Li et al., 2016) but are usually highly structured (Graham & Shelton, 2013) whereas the types of social media interactions used in this research can be a lot bigger, measured in terms of the ‘petabytes, exabytes, zettabytes and, yottabytes’ of Foley’s (2013) *Extreme Big Data*, and are often ‘messy’, characterised by a mixture of un-/semi-/structured elements such as text, images or toponymic geographical references.

These characteristics present ‘challenges’ for many established or conventional forms of computational analysis (Kambatla et al., 2014; Tsou, 2015). The ‘Data Lake’ may, perhaps, be more accurately be thought of as a ‘Data Swamp’; requiring significant skills in Virtual Machine (VM), Operating System (OS) and software setup to create ‘filtering systems’ capable of revealing meaningful information hidden within ‘otherwise turbid depths’ (Tear & Healey, 2017).

The methods used in this research have been selected to meet the requirement to quantitatively ‘mine’ large amounts of qualitative data, the ~230 million words of free-form text deposited by ~2.4 million users in ~8 million social media messages which, in turn, link to yet more content in ~3.5 million (~650,000 distinct) shared URLs. Consequently, this research adopts technologies centred around a well-established Relational Database Management System, Oracle 12c (Appendix 8, p436; Figure A8-3, p441), capable of storing and querying large amounts of structured (CSV) and semi-structured (JSON) data collected online from Facebook and Twitter, all of which has been augmented by three NLP/geoparsing systems. Unless, or until, a high performance, large scale and potentially all-encompassing ‘CyberGIS’ is developed (S. Wang, 2013) it seems likely that those managing, analysing and visualizing text heavy spatiotemporal OSN data will continue to integrate a number of products or technologies to fulfil their individual operational or research requirements.

Zelenkauskaitė & Bucy (2016) have correctly stated that ‘There is no such thing as a small scale, qualitative analysis of big social data.’ The results of this research,

presented in the following chapter, depend upon careful collection, preparation, augmentation, analysis and measurement of social media data. Answers to the three research questions set out in the introductory chapter of this thesis (Section 1.7, p34) are presented overleaf.

5 NLP/GEOPARSING RESULTS

5.1 Introduction

Earlier chapters have introduced (Chapter 1, p1) and contextualised this study (Chapter 2, p51) detailing the methodology adopted (Chapter 3, p94) and methods used (Chapter 4, p118) in this investigation. The current chapter presents results from this research, challenging the prevailing *Geographicality Assumption* (Section 1.7, p34) that *coordinate-geotagging users are the most geographically expressive of all OSN users*.

One main section, with three subsections, addresses the three research questions:

1. **How can baseline ‘geographicality’ be assessed and categorised in OSN data?** Section 5.2.1 (p188) lays the foundation for the following analyses by examining the utility of the classification scheme designed in Section 4.6.1 (p164) to assess and categorise ‘baseline’ geographicality in OSN data. The research data corpus consists of ~8m records sourced overwhelmingly (~90%) from Twitter, in a roughly even mix of tweets and retweets, with another ~10% of interactions (i.e., individual messages and accompanying metadata) comprised of Facebook posts. Few of these OSN interactions are coordinate-geotagged but many contain items of Potential Geographic Information (PGI) such as toponymic mentions of place in self-reported user location fields, which may be text-mined, or time zone offsets in seconds, which may be mapped. Data analysis and mining, in SQL, were used to score each of the identified PGI metadata fields to create an overall ‘Geographicality Score’ at the interaction, i.e., message, level. As multiple interactions were created by some users in the research data corpus the Geographicality Score was also computed at user level. The classification scheme provides a baseline for subsequent NLP and geoparsing operations which can be used to show, a) how much these processes can augment

geographicality and, b) how evenly any such uplifts in NLP-detectable geographicality are distributed between coordinate-geotagging and non-coordinate-geotagging users of the two case study OSNs.

2. **Does NLP-detectable 'geographicality' in message text increase in line with 'spatiality'?** Section 5.2.2 (p190) builds upon the preceding section by applying NLP and geoparsing techniques, reported against Geographicality Scores, to detect toponymic mentions of place in interaction message text, the field which offers the greatest source of potential geographical uplift available in OSN data (Stock, 2018). Message text has been widely studied and much research, e.g., in geographic information retrieval (Purves et al., 2018), is devoted to extracting geo-information from social media posts. Less attention has been devoted to determining *who* tweets, retweets or posts with the *most* geographical information. Answering this question, using three NLP/geoparsing systems, addresses the second research question. The methods adopted all produce output which can be data-mined in SQL to determine, a) whether uplifts in geographicality are evenly distributed amongst all users of OSNs by Geographicality Score, or b) whether uplifts are unevenly distributed, particularly amongst coordinate-geotagging and non-coordinate-geotagging users of OSN sites.
3. **Does NLP-detectable 'geographicality' in linked/shared 3rd party content increase in line with 'spatiality'?** Section 5.2.3 (p205) addresses the third research question by considering how coordinate-geotagged and non-coordinate-geotagged interactions and the users that create them link to and share 3rd party content. The posting of linked and shared material in the form of URLs to internal (OSN) or external (Web) content is an important characteristic of social media communications (Bartlett & Miller, 2013). While message text has been widely studied, much less attention has been paid to linked/shared content, partly as consuming and processing it can be difficult and partly as a result of the prevailing focus on analysing message text. This section details the results of text-mining

~650,000 distinct URLs in the search for references to locations and various other detectable 'entities' including, e.g., persons and organisations. SQL data-mining is used to determine which Geographicality Score classes, particularly amongst coordinate-geotagging and non-coordinate-geotagging users, are most likely to link to 3rd party material making most frequent mention of place.

Results detailing statistical tests of significance are presented in Section 5.3 (p219) and a comparative evaluation of the effectiveness of the three NLP/geoparsing methods employed in this research is given in Section 5.4 (p221), ahead of the chapter summary (p225). Further discussion of results follows in Chapter 6 (p227), which also presents additional findings arising from the exploratory spatiotemporal analytical methodology adopted (Chapter 3, p94) and the range of methods employed in this research (Chapter 4, p118).

5.2 Research questions

This chapter is organised around the three research questions defined earlier in this thesis (Section 1.7, p34), outlined above, and addressed in sequence below.

5.2.1 RQ1 – How can baseline 'geographicality' be assessed and categorised in OSN data?

Section 4.6.1 (p164) describes the development of the Geographicality Score used in this research, a measurement system based on the presence of different types of Potential Geographic Information (PGI) in OSN interaction metadata. Analysis and coding in SQL (Appendix 11 listing 16, p484) derived integer Geographicality Scores for each interaction in the research data corpus, based upon a coding scheme for 21 out of 35 identified PGI metadata fields shown in Table 4-10 (p174). Eleven distinct classes were observed in the two case study data sets; zero geographicality, some geographicality (1-9) and spatialised geographicality (class 100+ for original

coordinate-geotagged Twitter tweets and class 200+ for coordinate-geotagged Twitter retweets). Geographicality Scores, calculated at the interaction level (Figure 4-18, p178), were also computed at user level (Figure 4-19, p178) by determining the modal (most common) score for all interactions made by each user (Appendix 11 listing 18, p486). Geographicality Scores provide a baseline for assessing and categorising geographicality – at a range of spatial scales, e.g., from small-scale world time zones through to (apparently accurate) coordinate-geotagged Latitude and Longitude point locations – in OSN interaction metadata. The system has two principal advantages, Geographicality Scores are:

1. Easy to understand and straightforward to code.
2. Flexible; allowing, e.g., cross-tabulation of results in later analyses.

The scoring system does, however, have one distinct disadvantage:

1. Interactions cannot be scored where they lack available metadata.

Almost all Facebook posts (n=841,271 ; 58.27% of zero-scored interactions), a large number of Twitter tweets (n=598,544, 41.46%) and 3,836 Twitter retweets (0.27%) completely lacked PGI metadata. While these interactions could not be scored using the methods developed in Section 4.6.1 (p164) they can, along with all other positively-scored interactions, be geographically ‘augmented’ by searching for toponymic mentions of place in message text. Stock (2018), Gritta et al. (2018) and Purves et al. (2018) all concur that searching for toponymic place names in OSN interaction message text is most likely to prove successful in enhancing geographical referencing in social media data. This method, based upon searches conducted using the three NLP/geoparsing systems detailed earlier in Section 4.4.1 (p147), has been employed here to detect and count numbers of toponymic mentions in the message text (Section 5.2.2, p190) and linked/shared URL content (Section 5.2.3, p205) of all OSN interactions in the research data corpus. Cross-tabulation of these counts against Geographicality Scores, the results of which are

presented below, helps to determine which classes of interactions and users are most likely to make most mention of place in message text or linked/shared content. The following two sections answer RQ2 and RQ3 set out in the introductory chapter of this thesis (Section 1.7, p34) using these methods.

5.2.2 RQ2 – Does NLP-detectable ‘geographicality’ in message text increase in line with ‘spatiality’?

Three NLP/geoparsers (Section 4.6.1, p164) have been used to search for and augment any ‘geographicality’ detected in OSN message text. Two graphs (Figure 5-1 and Figure 5-2, p191) help to provide a summary of the findings reported below. Despite adopting technically different solutions to the ‘challenge’ of location-detection in free form text (S. Li et al., 2016) both GATEcloud and CLAVIN-rest, the two systems successfully used to text-mine all ~8m social media messages in the research data corpus, produced highly comparable results. Processing rate restrictions in the third NLP/geoparser used, AlchemyAPI, meant that only subsets of the research data corpus could be processed and these results are presented separately in Section 5.2.2.2 (p198).

In the `US2012` data set (Figure 5-1, p191) message text contains a similar average number of NLP-detectable locations/interaction regardless of geoparser (CLAVIN-rest in light blue, GATEcloud in darker blue). Although original coordinate-geotagged messages (Twitter tweets and Facebook posts, Geographicality Score 100+) contain as many as 1.449 average locations/interaction (according to GATEcloud) the average number of locations detected in zero-scored interactions is slightly higher at 1.586. In the `SCOT2014` data set (Figure 5-1, p191; CLAVIN-rest in light grey, GATEcloud in darker grey) original coordinate-geotagged messages contain on average 1.829 locations/interaction (GATEcloud) against a higher average of up to 4.213 locations/interaction for interactions with a 0 Geographicality Score. This class is comprised of many records (n=841,271; 58.27%

of all 1,443,651 zero-scored interactions) sourced from Facebook (Section 4.6.1, p164).

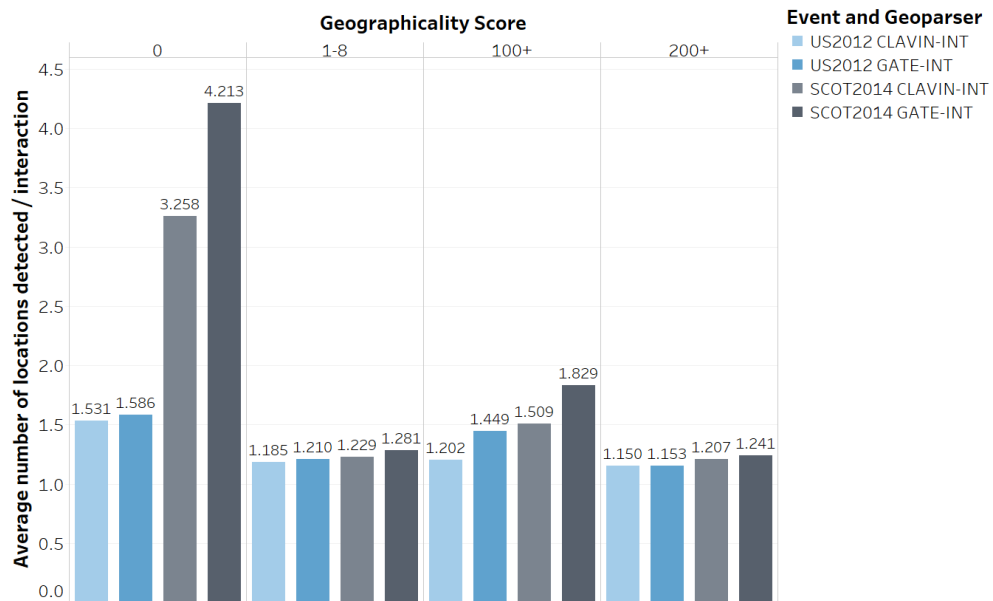


Figure 5-1 – US2012/SCOT2014: Average number of locations detected / interaction by Event, Geoparser and Geographicality Score

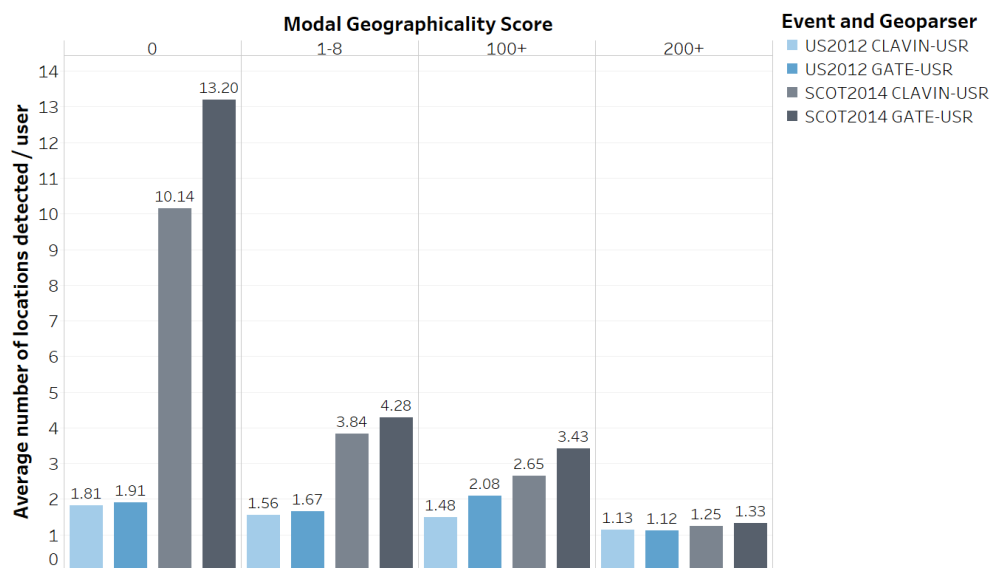


Figure 5-2 – US2012/SCOT2014: Average number of locations detected / user by Event, Geoparser and Modal Geographicality Score

When results are cross-tabulated against modal Geographicality Scores at user level (Appendix 11 listing 18, p486) the distribution of average numbers of locations

detected/user differs markedly (Figure 5-2, p191) from the interaction level distribution shown in Figure 5-1. In the `US2012` data set, which has very few Facebook interactions, CLAVIN-rest (light blue) and GATEcloud (darker blue) again detect similar average numbers of locations/user, with the highest average (2.08) found amongst coordinate-geotagged, and original, Twitter tweets and Facebook posts (Geographicality Score 100+). In the `SCOT2014` data set, however, CLAVIN-rest (light grey) and GATEcloud (darker grey) find, on average, more locations/user in Geographicality Score classes 1-8, recording 4.28/user against 3.43/user for coordinate-geotagging Twitter or Facebook users. The average number of NLP-detectable locations/user in the zero-scored class ranges 10.14 (CLAVIN-rest) to 13.20 (GATEcloud) depending upon geoparser. As noted above, this class is comprised of large numbers of Facebook posts lacking in PGI metadata fields (Section 4.6.1, p164). Facebook's lack of geographic metadata does not prevent interactions sourced from this OSN platform having potential, and potentially greater, geographical value than Twitter-sourced interactions. On the contrary, Facebook posts appear to have significantly more geographical value, following NLP-based location entity detection, than Twitter tweets or retweets, including coordinate-geotagged variants of either.

Stock (2018, p227) has noted that 'It is currently very difficult for researchers who wish to apply social media data to a specific research question [...] to determine the best approach to use to extract geographic data, to evaluate the limitations of alternative approaches, and then to use the methods for their own research.' Summarising the state of location mining in social media, Stock (2018, p227) also points out that 'Twitter is by far the most frequently used social media platform for geospatial research, despite being only 11th in global rankings by number of users, and research on more popular platforms (e.g. Facebook) is much more limited. There is a need for research into some of these less frequently used platforms, including the analysis of the location of content of particular kinds across and within the platforms.' The remainder of this section presents detailed results from the

three alternative NLP/geoparsing systems used to text-mine Twitter *and* Facebook message text in this research, valuably contributing to knowledge in the two main areas identified by Stock (2018). Augmenting geographicality in OSN interactions remains technically challenging but is possible using any of the three systems detailed below, each of which offers distinctive approaches to the problem.

5.2.2.1 GATEcloud

Effective collaboration with researchers at the University of Sheffield's Department of Computer Science (Bontcheva & Greenwood, personal communication, 2014; Roberts, personal communication, 2016) and participation in beta programmes has helped develop enhanced functionality for GATEcloud.net, the Cloud-hosted instance of Sheffield's General Architecture for Text Engineering (GATE) software (Section 4.4.1.1, p149). GATE provides a 'family of open source text analysis tools and processes [and] is one of the most widely used systems of its type with yearly download rates of tens of thousands and many active users in both academic and industrial contexts' (Cunningham, Tablan, Roberts, & Bontcheva, 2013). GATEcloud works at much larger scale on Amazon's Cloud computing infrastructure, enabling researchers to run specially developed 'NLP algorithms [which] tend to be complex, [making] their parallelization and deployment on cloud platforms a non-trivial task' (Tablan et al., 2012). As Bontcheva et al. (2013, p1) note, 'Processing microblog text is difficult: the genre is noisy, documents have little context, and utterances are very short.' GATE's TwitIE pipeline 'has been specifically adapted to microblog content' and runs at scale on the GATEcloud service.

All US2012 and 98.14% of SCOT2014 social media messages have been successfully processed using TwitIE on GATEcloud (Table 5-1, p194). To avoid a `Java OutOfMemoryError` on GATEcloud servers, 120,727 longer Facebook posts from the SCOT2014 data set were discarded as TwitIE is geared towards the analysis of ~140-character Twitter-type text. The software has, nonetheless, successfully processed 721,775 Facebook posts.

Table 5-1 – US2012/SCOT2014: GATEcloud processing

US2012	N Records	N Processed	% Processed
Facebook	57,265	57,265	100.00%
Twitter Tweet	866,160	866,160	100.00%
Twitter Retweet	795,242	795,242	100.00%
Totals	1,718,667	1,718,667	100.00%
SCOT2014	N Records	N Processed	% Processed
Facebook	785,237	664,510	84.83%
Twitter Tweet	2,846,687	2,846,687	100.00%
Twitter Retweet	2,845,789	2,845,789	100.00%
Totals	6,477,713	6,356,986	98.14%

Named Entity Recognition and Information Extraction using TwitIE on GATEcloud can detect `Token` (i.e., individual words), `Emoticon`, `Hashtag`, `URL`, `Address`, `Date`, `Location`, `Organization`, `Person`, `Money`, `Percent`, `SpaceToken` (spaces) and `Sentence` types in Twitter-type OSN message text. The detection of all possible types, particularly `Tokens`, creates extremely large JSON output files (US2012=3.02GB in, 4.74GB out; SCOT2014=19.0GB in, 54.9GB out) which, initially, could only be joined to input records stored in the Oracle 12c database using the (potentially duplicated) message text of each interaction as the key.

Co-development work with University of Sheffield staff (Roberts & Tear, personal communication, 2017) updated TwitIE's data import routine to preserve unique identifiers (DataSift's `INTERACTION_ID`) for output during processing. Running TwitIE on GATEcloud *without* `Token`, `Money`, `Percent`, `SpaceToken` or `Sentence` detection results in significantly smaller JSON output files for re-import to the Oracle 12c database (US2012=901MB, SCOT2014=4.19GB). Once imported the JSON output from TwitIE on GATEcloud could easily be joined to source tables in Oracle 12c using the retained `INTERACTION_ID` field as the key. This development, It is hoped, will be of significant future value to researchers processing large social media corpora held in DataSift JSON, standard Twitter JSON or other OSN JSON formats.

The main entity type of interest returned in GATEcloud JSON following TwitIE processing is `Location`, in the `locType` key, which references indexed characters (i.e., the n^{th} characters in the message text) containing the detected location; these are coded `region`, `province`, `post`, `unknown`, `country`, `country_abbrev`, `city`, `airport`, `racecourse` or `pre`. These codings are mainly self-evident, except for `pre` and `post` which refer to first/last matches to ‘bit[s] of a location [covering] things like “Mount”, “East”, “Cape”, “Isle of” etc. [which co-occur with] a proper noun’ (Tear & Maynard, personal communication, 2018). TwitIE on GATEcloud detects `locType` in the message text of 263,296 US2012 interactions, identifying a further 2,088,788 messages containing locations in the SCOT2014 data set.

Overall, 2,352,084 interactions, or 28.69% of all 8,196,380 interactions, are found to contain locations in message text. As each interaction may mention multiple locations, the JSON array of locations returned by TwitIE must be ‘unpacked’ into a relational view using Oracle 12c’s `NESTED PATH` syntax for further analysis (Hammerschmidt, 2015; Appendix 11 listing 20, p451). TwitIE’s output may be illustrated by joining the input `INTERACTION_ID` and output `ID_STR` fields to the `INTERACTIONS` table in Oracle 12c (Appendix 11 listing 21, p486), selecting Presidential Candidate Barack Obama’s Twitter tweet shown earlier in this thesis (Figure 4-9, p139) and the output of GATEcloud processing against this message text (Figure 5-3, p196).

In this example, TwitIE finds entities of `Person`, `URL` and `Location` type. The `Location.locType` is “province” and the indices `[41, 45]` point to characters 41-45 of the message text, containing the geographically-relevant word ‘Ohio’. SQL queries designed to count against this view (e.g., Appendix 11 listing 22, p486) showed that TwitIE on GATEcloud detected 330,389 locations in the US2012 data set with a further 4,315,548 in the larger SCOT2014 data set (Table 5-2, p197).

```
{
  "text": "Happening now: President Obama speaks
in Ohio about the choice in this election. RT so
your friends can watch, too.
http://t.co/d42qgdn8\n\n",
  "entities": {
    "Person": [{
      "indices": [25, 30],
      "surname": "Obama",
      "kind": "fullName",
      "rule": "GazPerson",
      "gender": "male",
      "ruleFinal": "PersonFinal"
    }
  ],
  "URL": [{
    "indices": [116, 136],
    "rule": "URL",
    "temp_category": "NN",
    "kind": "URL",
    "length": 20,
    "string": "http://t.co/d42qgdn8",
    "replaced": 12,
    "category": "URL"
  }
  ],
  "Location": [{
    "indices": [41, 45],
    "kind": "locName",
    "rule": "InLoc1",
    "locType": "province",
    "ruleFinal": "LocFinal"
  }
  ],
  "id_str": "1e227914e2f4ac80e0740cf699462aae"
}
```

Figure 5-3 – GATEcloud TwitIE output for Presidential Candidate Barack Obama’s Twitter tweet (Figure 4-9, p139)

TwitIE is not a dedicated geoparser and does not return coordinates alongside detected `Location` entities. Although gazetteers are used to resolve location entities in text these are much smaller than the ~11m row GeoNames (2016) gazetteer used by CLAVIN-rest (Section 5.2.2.3, p202). Instead the software is

designed to apply ‘constructive [NLP-based] rules that look for common words that are often part of a location name like "River", "Hill", "Island", etc.’ (Roberts & Tear, personal communication, 2017) as well as considering sentence structure, e.g., the words ‘to’ and ‘from’ are often followed by location mentions. The number of locations detected in message text by TwitIE on GATEcloud compares favourably with the 24.14% of location-bearing interactions found by CLAVIN-rest (yielding an overall total of 3,524,958 resolved locations against GATEcloud’s 4,645,937) using the more extensive GeoNames gazetteer

The ratio between numbers of interactions and numbers of locations resolved by TwitIE on GATEcloud shows that Facebook posts, with their longer text content, produce more location mentions per message (1.9 against an average of 1.22 for Twitter tweets or retweets in the US2012 data set) rising significantly to 5.66 location mentions per message in the SCOT2014 data set, where Facebook-based interactions make up a higher proportion of the data set (12.12% of total, vs. 3.33% US2012, see Table 4-8, p170). The relevance of this finding is expanded upon in the summary (p225) of this Chapter and in the discussion (p227) which follows.

Table 5-2 – US2012/SCOT2014: Number of resolved locations detected by GATEcloud in Facebook (FB), Twitter tweet (TW) and Twitter retweet (RT) interactions

	US2012	RATIO	SCOT2014	RATIO	TOTAL
FB resolved locations	25,405	1.90	2,097,506	5.66	2,122,911
FB <i>n</i> interactions	13,341		370,774		384,115
TW resolved locations	153,085	1.23	1,087,698	1.31	1,240,783
TW <i>n</i> interactions	123,960		833,235		957,195
RT resolved locations	151,899	1.21	1,130,344	1.28	1,282,243
RT <i>n</i> interactions	125,995		884,779		1,010,774
Total Resolved	330,389	1.25	4,315,548	2.07	4,645,937
Total Interactions	263,296		2,088,788		2,352,084

It possible to produce numerous analyses by joining GATEcloud output on `INTERACTION_ID` to the main `INTERACTIONS` table in Oracle 12c and, particularly, to determine whether interactions or users with different

Geographicality Scores tweet, retweet or post with differing propensity to mention locations in their message text. Results of this analysis are shown in Figure 5-1 (p191) and Figure 5-2 (p191) and are statistically tested, alongside results from the other NLP/geoparser systems discussed below, in Section 5.3 (p219).

5.2.2.2 AlchemyAPI

A bespoke program written in the Ruby language (Appendix A10.3, p451) was used to pass interaction message text to the Cloud-hosted AlchemyAPI service, recently rebranded by IBM as Watson Natural Language Understanding (IBM, 2017c).

AlchemyAPI (for convenience), is another NLP system able to detect various geographical and other entity types found in interaction message text. In contrast to TwitIE on GATEcloud or CLAVIN-rest, the first of which has very low costs for academic users and the second of which is free open-source software, AlchemyAPI is a commercial offering, available to academic users on a rate-limited basis. To pass all ~8m social media messages through AlchemyAPI at the permitted academic use rate of 30,000 API transactions per day would require ~1,866 processing days if the number of transaction ‘credits’ used to process each snippet of interaction message text matched the 7 credits/interaction used here.

Table 5-3 – US2012/SCOT2014: Number of records by tranche processed by AlchemyAPI

Tranche	N Records
US2012_GEO Stream	146,424
US2012_NON_GEO 1% sample tweets	15,151
SCOT2014 geo-tagged tweets	93,378
SCOT2014 1% sample tweets	56,622

As this timescale was considered unrealistic, several smaller tranches of data that could be processed within more realistic time frames were inserted into a queuing table (`ALCHEMY_API`) used to control daily AlchemyAPI processing. The number of records in each tranche is shown in Table 5-3 (p198).

Focusing on detecting toponymical differences in the message text of coordinate-geotagged and non-coordinate-geotagged interactions, all records from the US2012_GEO Stream (n=146,424), comprised wholly of coordinate-geotagged Twitter tweets, were inserted into the queueing table using SQL (Appendix 11 listing 23, p487). Similar statements populated the other tranches shown above using the SQL in Appendix 11 (listings 24-26) and, in the case of the 1% samples, using Oracle's `SAMPLE` clause (Hornick, 2010) to insert numbers of records processable within a reasonable ~3-4 months time frame. The Ruby program, running on a CentOS 7 virtual machine, communicates with the Oracle 12c database on the laptop host over a TCP/IP network using Oracle Call Interface 8 (OCI8) middleware (Figure A8-3, p441).

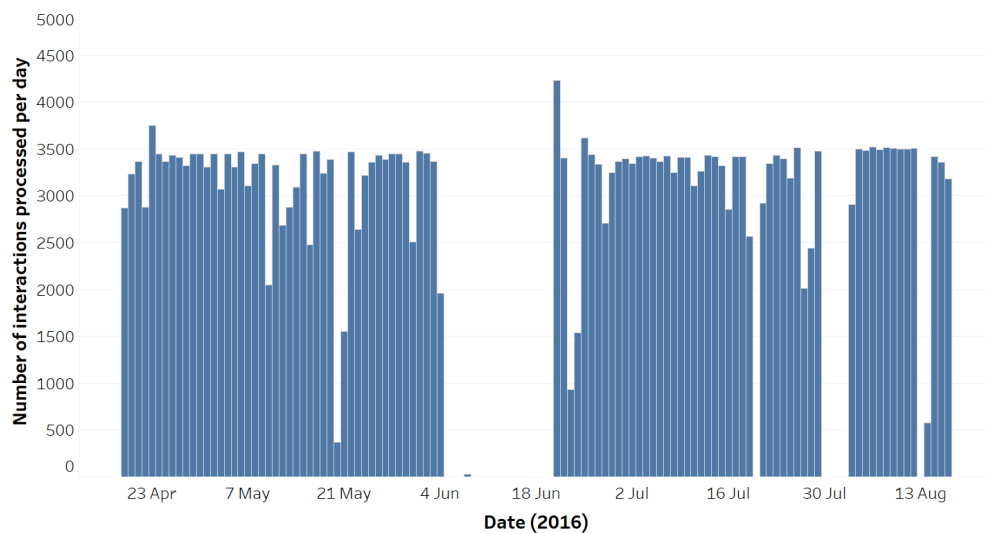


Figure 5-4 – US2012/SCOT2014: Number of records processed per day by AlchemyAPI

It proved impossible to fetch CLOB data storing message text from the database using this interface and, consequently, as the Ruby programme (Appendix A10.3, p451) processed each queued item of message text this was `CAST` (Oracle, 2018c) to a `VARCHAR` data type of length 140 (the maximum length of a Twitter tweet at the time) and processed, at the rate of ~3,000 interactions/day, over a period of 120 days (Figure 5-4, p199). Gaps in Figure 5-4 show periods when the CentOS 7 virtual machine running on laptop hardware, or the laptop itself, were switched off.

Running the code from an always-on computer, or without rate throttling, would yield results much more quickly, as would a higher daily rate limit. Paying customers accessing Watson Natural Language Understanding may process 5 million records or more per month for ~\$1,800/month (IBM, 2017c). Overall (Figure 5-5), the number of entity types detected in message text shares some similarities, but important differences, with the number of entity types found in linked/shared URL content (Section 5.2.3, p205; Figure 5-8, p215).

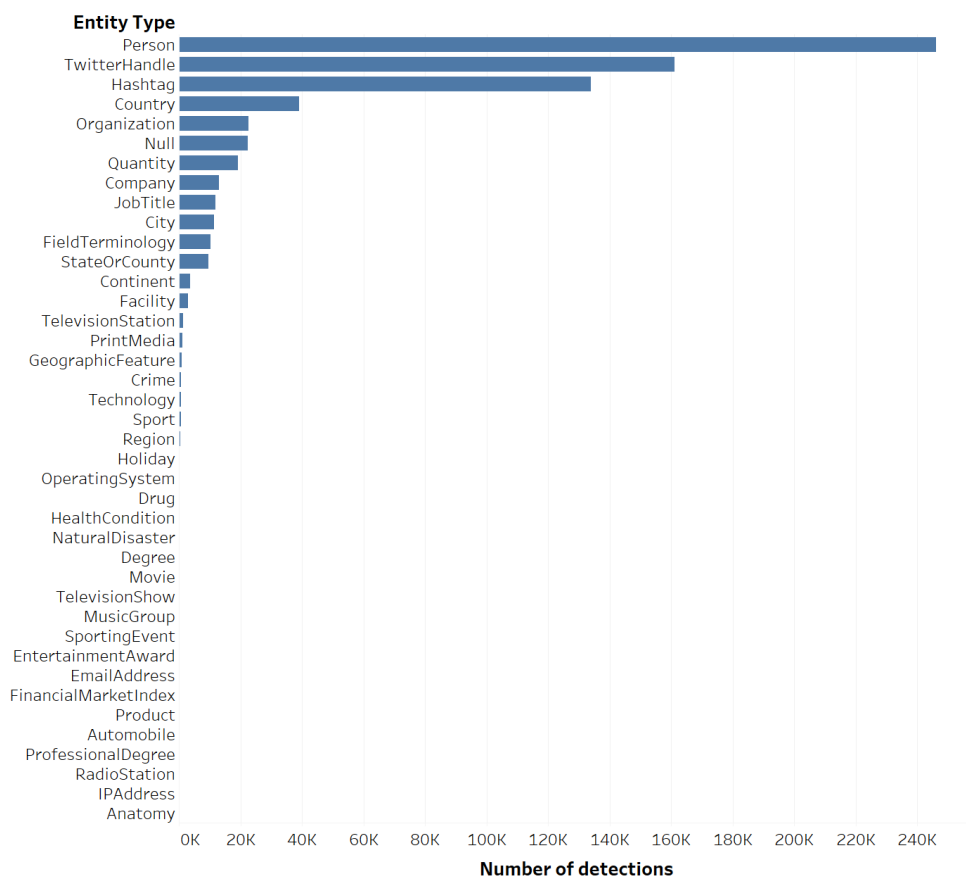


Figure 5-5 – US2012/SCOT2014: Number of distinct entities by type identified in message text by AlchemyAPI across all sampled tranches processed (n=311,575)

In both interaction message text (Figure 5-5, p200) and linked/shared URL content (Figure 5-8, p215), AlchemyAPI detects most mentions of the `Person` entity type. This reflects frequent mentions of political figures (Barack Obama, Mitt Romney, Alex Salmond and others) across both corpora. Subsequently, linked/shared URL

content most frequently mentions `Organization`, `Country`, `Quantity` and `Company` types. Conversely, in interaction message text, `AlchemyAPI` next identifies `TwitterHandle`, `Hashtag`, `Country` and `Organization` entity types. This accurately reflects the different nature of the two corpora, Twitter tweets and linked articles, and the inclusion of many `TwitterHandles` (e.g., `@BarackObama`) and `Hashtags` (e.g., `#indyref`) in users' Twitter tweets.

Taking the clearly geographical entity types (`Country`, `City`, `StateOrCounty`, `Continent`, `GeographicFeature` and `Region`), as well as testing for any coordinates returned in JSON by `AlchemyAPI`, it is possible to count numbers of geographical entities and determine the rate of geographical entity detection by tranche using SQL (Appendix 11 listing 27, p487). Table 5-4 shows these results for each tranche. In both `US2012` and `SCOT2014` events coordinate-geotagged Twitter tweets have slightly lower rates of geographical entity detection than corresponding 1% samples drawn from the same electoral event.

Table 5-4 – `US2012/SCOT2014`: Number of geographical entities detected in Twitter tweets by `AlchemyAPI` showing the rate (entities/tweet) for each sampled tranche

Tranche (Twitter tweets)	Entities	Tweets	Rate
<code>US2012_GEO geotagged</code>	21,522	146,424	0.15
<code>US2012_NON_GEO 1% sample</code>	2,597	15,151	0.17
<code>SCOT2014 geotagged</code>	23,974	93,378	0.26
<code>SCOT2014 1% sample</code>	17,407	56,622	0.31

Differences in the rate of geographical entity detection between coordinate-geotagging users' Twitter tweets and corresponding 1% samples are slight within electoral events. Perhaps counter-intuitively, however, and as with results from `AlchemyAPI`-based linked/shared URL content analysis (Section 5.2.3, p205), explicitly coordinate-geotagged interactions have slightly fewer detectable mentions of geographical entities in their message text than the corresponding 1% sample of non-coordinate-geotagged Twitter tweets in both events. Between the two political events, it appears that the 2014 Scottish Independence Referendum

generated a rate (0.31 geographical entities per tweet) roughly twice as high as that shown (0.17) in the 2012 US Presidential Election. There is no obvious explanation for this difference in geographical entity detection rates, suggesting that further research is necessary to determine how such rates might vary between events of various types in the future.

As with TwitIE on GATEcloud (Section 5.2.2.1, p193) geoparsing using AlchemyAPI suggests that those who are most geographic in depositing coordinate information alongside their Twitter tweets are slightly less geographically expressive than all users as measured by numbers of mentions of NLP-detectable geographical entities in message text. Statistical tests demonstrating the significance of this previously unreported result, alongside results from the other NLP/geoparser systems discussed in this chapter, are presented in Section 5.3 (p219).

5.2.2.3 CLAVIN-rest

CLAVIN-rest, Berico Technologies' open-source Cartographic Location and Vicinity INdexer (Berico-Technologies, 2017), is a geoparser employing Stanford CoreNLP (Stanford University, 2017) and the extensive ~11 million record GeoNames (2016) gazetteer to detect references to locations in free-form text, returning coordinates from GeoNames in JSON when detecting a match. If TwitIE on GATEcloud and AlchemyAPI, detailed above, may be considered specialist NLP toolkits capable of detecting locations (amongst many other entities) in text, CLAVIN-rest may be considered a specialist geoparser, capable of augmenting detected locations with coordinates but incapable of detecting other non-geographical entity types in text.

Using a version of CLAVIN-rest compiled on a CentOS 7 virtual machine (Appendix 8, p436) and the Linux shell script detailed earlier (Section 4.4.1.3, p154; Figure 4-13, p156) all 8,196,380 items of interaction message text were geoparsed by CLAVIN-rest using a file output from the Oracle 12c database. This UTF-8 encoded file held `UUID` (for uniqueness, and subsequent SQL joins) and OSN message text exported in a delimited format using a compound delimiter (Section 4.4.1.3, p154) which did

not otherwise appear (as commas, pipes and most other common CSV file delimiters did) within interaction message text. From the input file, CLAVIN-rest wrote 1,978,404 lines in 487.13MB consisting of UUID and geoparser output in JSON. As with GATEcloud (Section 5.2.2.1, p193; Figure 5-3, p196) and AlchemyAPI (Section 5.2.2.2, p198) this output consists of an array of detected locations, in the JSON key `resolvedLocationsMinimum`, as shown in Figure 5-6.

```
{
  "resolvedLocationsMinimum": [{
    "geonameID": 5174035,
    "name": "Toledo",
    "countryCode": "US",
    "latitude": 41.66394,
    "longitude": -83.55521
  }, {
    "geonameID": 4990729,
    "name": "Detroit",
    "countryCode": "US",
    "latitude": 42.33143,
    "longitude": -83.04575
  }, {
    "geonameID": 4839335,
    "name": "New Center Cemetery",
    "countryCode": "US",
    "latitude": 41.60704,
    "longitude": -72.65815
  }
]}
}
```

Figure 5-6 – CLAVIN-rest output showing resolved (i.e., successfully geoparsed) locations

Once re-imported into Oracle 12c, a database view was created around this stored output to ‘unpack’ array data into rows using SQL (Appendix 11 listing 28, p488). Counts against this view showed 1,978,404 OSN interactions, 24.14% of 8,196,380 total interactions, contained one or more resolved locations in message text. As any one OSN interaction may mention several locations, the view can be used to count the number of resolved locations detected in the entire research data corpus. Altogether, 3,524,958 spatial locations were identified in the ~2m records

successfully geoparsed by CLAVIN-rest. Numbers of resolved locations by OSN type are shown in Table 5-5.

Table 5-5 – US2012/SCOT2014: Number of resolved locations detected by CLAVIN-rest in Facebook (FB), Tweet (TW) and Retweet (RT) interactions

	US2012	RATIO	SCOT2014	RATIO	TOTAL
FB resolved locations	22,019	1.80	1,573,145	3.98	1,595,166
FB <i>n</i> interactions	12,199		395,112		407,311
TW resolved locations	120,491	1.19	832,941	1.25	953,433
TW <i>n</i> interactions	101,664		667,263		768,927
RT resolved locations	119,979	1.18	856,383	1.22	976,363
RT <i>n</i> interactions	101,599		700,567		802,166
Total Resolved	262,489	1.22	3,262,469	1.85	3,524,959
Total Interactions	215,462		1,762,942		1,978,404

It is apparent, as with TwitIE on GATEcloud (Section 5.2.2.1, Table 5-2, p197), that 45.25% of resolved locations ($n=1,595,166$) detected by CLAVIN-rest stem from 407,311 successfully geoparsed OSN interactions sourced from Facebook (i.e., 20.59% of all interactions processed), most of which ($n=395,112$) were collected during the 2014 Scottish Independence Referendum. Larger numbers of Twitter tweet ($n=768,927$) and retweet ($n=802,166$) interactions (total $n=1,571,093$) were geoparsed by CLAVIN-rest but yielded a total of only 1,929,794 ($n=953,433$ and $n=976,363$, respectively) resolved locations, 87.54% of which ($n=1,689,324$) were found in the SCOT2014 data set, which features a higher proportion of Twitter retweets. Most locations resolved in retweets will, of course, be duplicates of locations found in the originating tweet.

The ratio of resolved locations to interactions is, on average, higher (1:1.85 vs 1:1.22) for the more recent SCOT2014 data set and is substantially higher (1:3.98) for Facebook posts. This is a significant finding as most academic studies examining OSN data source publicly-posted interactions more readily, cheaply or freely available from Twitter (Giardullo, 2016; Stock, 2018; Tufekci, 2014). Facebook posts, which make up a higher proportion of the SCOT2014 data set (12.12% of

total, vs. 3.33% US2012; Table 4-6, p165), offer significantly greater possibilities for text geoparsing than Twitter-sourced interactions. The longer message format of Facebook allows more text per record and, on average, ~4 locations are resolved for every interaction by CLAVIN-rest as against just ~1.2 for Twitter-based message text. Broadly similar results from GATEcloud text-mining corroborate these findings (Section 5.2.2.1, p193).

Facebook contributes 842,502 interactions (Table 4-8, p170) to the combined US2012/SCOT2014 research data corpus, of which just 1,231 records (0.15%) are coordinate-geotagged. Twitter-sourced interactions are more numerous and contain higher percentages of coordinate-geotagged interactions (Table 4-8, p170) yet offer comparatively fewer detectable locations/message when geoparsed. Much less frequently used in most social media studies than Twitter data, with its lower cost, easy availability (Stock, 2018) and more obviously coordinate-geotagged nature, it appears that Facebook posts offer a richer seam for text-mining and geoparsing operations than Twitter tweets or retweets, whether these are coordinate-geotagged or not. Statistical tests demonstrating the significance of these findings, alongside results from the other NLP/geoparser systems discussed in this chapter, are presented in Section 5.3 (p219).

5.2.3 RQ3 – Does NLP-detectable ‘geographicality’ in linked/shared 3rd party content increase in line with ‘spatiality’?

The sharing of media, in the form of URL links to content posted elsewhere on OSN sites, or on 3rd party websites, is a key component of social media usage (Bartlett & Miller, 2013; Hermida, Fletcher, Korell, & Logan, 2012; Kamath, Caverlee, Cheng, & Sui, 2012). According to Bartlett & Miller (2013, p42), understanding media sharing ‘helps identify influencers’, whose highly active sharing activities tend to drive ‘subsequent traffic’ on OSN web sites. Identifying influencers has become especially relevant as traditional models of news publication and dissemination (e.g., buying a newspaper) are replaced by often free, frequently social, online ‘Internet news

market[s]' (M. S. Weber & Monge, 2011) which 'are changing the way individuals consume and share news' (C. S. Lee & Ma, 2012). While research, especially in the communications literature (Section 2.5, p72), has addressed some of these changing behaviours, little is known about differential patterns of link sharing amongst differing types of spatially or geographically expressive OSN users. Do users posting with coordinate-geotags, for example, link to content containing more identifiable locational entities than other users? The following pages address this research question using NLP-based text-mining and SQL-based data-mining techniques. Results are reported for URL link sharing against the Geographicality Scores developed earlier in Section 4.6.1 (p164).

The `INTERACTIONS` table holds several links columns; `FB_LINK` stores link URLs shared using Facebook; `TW_LINKS` and `TW_RT_LINKS` store Twitter tweet and retweet link URLs, respectively. Whenever one of these mutually exclusive fields is not null, a 'flattened' version of the original JSON array (as per earlier examples given in Table 4-4, p146) is stored in the `LINKS_URL` field of DataSift's CSV file, taking the form, e.g., for an array of 3 links:

```
[ "URL1", "URL2", "URL3" ]
```

Altogether, 3,281,150 records (40.03% of all interactions) have a non-null `LINKS_URL` field. Because each non-null value may contain an array of length n , each array must be transposed into n rows, using regular expressions (`REGEXP`) in SQL (Oracle, 2016a) to create a new table for AlchemyAPI processing (Appendix 11 listing 29, p488). This table holds 3,485,840 records, i.e. some interactions mention multiple URLs. Table 5-6 (p207) shows how these linked URLs are split amongst coordinate and non-coordinate-geotagged interactions sampled from the two OSN sources (by subtype) for the four Streams recorded in 2012 and 2013-2014.

Table 5-6 – US2012/SCOT2014: Number and percentage of linked URLs by Stream, OSN source and subtype (FB=Facebook, TW=Tweet, RT=Retweet) created by non-coordinate-geotagging and coordinate-geotagging users

NOT GEOTAGGED	FB	% FB	TW	%TW	RT	%RT	TOTAL	% TOTAL
US2012_GEO	0	0.00%	0	0.00%	0	0.00%	0	0.00%
US2012_NON_GEO	20,602	2.63%	182,456	11.89%	161,107	13.80%	364,165	10.45%
US2012_NON_GEO_HISP	13	0.00%	3,694	0.24%	1,475	0.13%	5,182	0.15%
SCOT2014	762,552	97.31%	1,317,646	85.87%	990,803	84.85%	3,071,001	88.10%
SUBTOTAL	783,167	99.94%	1,503,796	98.00%	1,153,385	98.77%	3,440,348	98.69%
GEOTAGGED	FB	% FB	TW	%TW	RT	%RT	TOTAL	% TOTAL
US2012_GEO	0	0.00%	13,145	0.86%	0	0.00%	13,145	0.38%
US2012_NON_GEO	0	0.00%	1,273	0.08%	795	0.07%	2,068	0.06%
US2012_NON_GEO_HISP	0	0.00%	15	0.00%	0	0.00%	15	0.00%
SCOT2014	464	0.06%	16,224	1.06%	13,576	1.16%	30,264	0.87%
SUBTOTAL	464	0.06%	30,657	2.00%	14,371	1.23%	45,492	1.31%
TOTAL	783,631	22.48%	1,534,453	44.02%	1,167,756	33.50%	3,485,840	100.00%

Subtotals in Table 5-6 show that the vast majority of links (98.69%) are created by non-coordinate-geotagging users, responsible for all but 45,492 of the total number of 3,485,840 linked/shared URLs in the research data corpus. Furthermore:

- Links sourced from Facebook interactions account for 22.48% (n=783,631) of all links but are sourced overwhelmingly (97.37%) from the SCOT2014 Stream, with just 2.63% of all Facebook links recorded during the 2012 US Presidential Election data collection exercise. Of these, none were recorded in the exclusively geographical Stream US2012_GEO (which consists of Twitter tweets only) and just 13 came from the geographically agnostically sampled Stream US2012_NON_GEO_HISP. Only 464 links from coordinate-geotagging Facebook users are present in the research data corpus, all sampled during the SCOT2014 event.
- Links sourced from Twitter interactions account for 77.52% of all links (n=2,702,209) comprised 44.02% of tweets (n=1,534,453) and 33.50% (n=1,167,756) of retweets:

- The SCOT2014 Stream again contributes most linked/shared URL records (n=2,353,477; 87.10%) and just 29,800 of these links were created by coordinate-geotagging users.
- In the US2012 event most linked/shared URLs (n=364,165; 10.45% of all links) were recorded in the US2012_NON_GEO Stream, set up without the requirement to sample exclusively coordinate-geotagged interactions (Appendix A7.2.2, p433). Numbers and percentages of URLs linked and shared in other Streams, whether coordinate-geotagged or not, are low.
- Altogether 45,028 URLs have been linked or shared in coordinate-geotagged Twitter tweets (n=30,657) or retweets (n=14,371). Added to the 464 links shared in coordinate-geotagged Facebook posts just 45,492 links (or 1.31% of all 3,485,840 links) have been created by coordinate-geotagging users.

It is apparent that coordinate-geotagging users, according to the counts and percentages in Table 5-6 (p207), make far fewer URL link shares than non-coordinate-geotagging OSN users. This finding has not been reported elsewhere.

The top 20 World Wide Web domain names referenced, by number of links for all interactions in the US2012 data set, are shown in Table 5-7 (p209). Counts have been generated using SQL (Appendix 11 listing 30, p488) and results are ranked by count, at domain-level, of links from interactions created without, and with, Latitude and Longitude coordinates. A similar ranking, for the SCOT2014 data set, is shown in Table 5-8 (p210). In each electoral event, there is considerable overlap between the set of popular linked domains recorded in OSN interactions made without, or with, coordinates. However, disparities in the order of the rankings also exist, along with higher order preferences amongst coordinate-geotaggers to link to some domains (e.g., yatown.com in the US2012 case; dk.pairsonnalites.org in the SCOT2014 case) that do not appear at all in the top 20 rankings for non-coordinate-geotagged interactions.

Table 5-7 – US2012: Top 20 Domains and number of links for those interacting without and with coordinate-geotags (including retweets)

US2012 links without coordinates			US2012 links with (any) coordinates	
Position	Domain	Number	Domain	Number
1	www.huffingtonpost.com	14,395	instagram.com	1,976
2	www.youtube.com	14,109	www.youtube.com	861
3	instagram.com	10,651	www.huffingtonpost.com	694
4	www.breitbart.com	10,192	www.twitlonger.com	241
5	www.washingtonpost.com	7,552	www.politico.com	240
6	thinkprogress.org	6,317	www.washingtonpost.com	236
7	twitpic.com	5,148	thinkprogress.org	223
8	www.facebook.com	4,659	www.breitbart.com	210
9	www.barackobama.com	4,350	twitpic.com	180
10	www.politico.com	4,303	twitter.com	146
11	www.foxnews.com	3,982	www.foxnews.com	137
12	www.dailykos.com	3,624	www.buzzfeed.com	131
13	edition.cnn.com	3,527	www.barackobama.com	126
14	www.motherjones.com	3,447	yatown.com	119
15	news.yahoo.com	3,446	www.motherjones.com	116
16	dailycaller.com	3,393	myloc.me	114
17	www.theblaze.com	3,012	www.argojournal.com	95
18	www.reuters.com	2,854	politicalticker.blogs.cnn.com	92
19	twitchy.com	2,792	www.dailykos.com	87
20	www.twitlonger.com	2,448	www.nytimes.com	83

Yatown.com was a ‘neighborhood social network that connects individuals with their neighbors, and allows them to share information’ (Crunchbase, 2018). Now closed, the website might well have been of interest to coordinate-geotagging users linking to locally-relevant content during the 2012 US Presidential Election. The Pairsonnalites.org website, regularly linked to by coordinate-geotagging users during the 2014 Scottish Independence Referendum, is a more puzzling inclusion. This self-proclaimed ‘Nordic’ but US-registered (ICANN, 2018) website, which is still online, describes itself as a blog ‘Keeping up-to-date on social exclusion worldwide’ and featured many articles in its ‘Nordic | Scotland’ edition covering Scottish Independence during 2014.

Table 5-8 – SCOT2014: Top 20 Domains and number of links for those interacting without and with coordinate-geotags (including retweets)

SCOT2014 links without coordinates			SCOT2014 links with (any) coordinates	
Position	Domain	Number	Domain	Number
1	www.facebook.com	410,204	www.youtube.com	2,826
2	www.bbc.co.uk	203,250	dk.pairsonnalites.org	2,050
3	www.youtube.com	196,966	www.theguardian.com	2,004
4	www.theguardian.com	149,192	instagram.com	1,536
5	www.blackfarce.com	91,531	www.scotsman.com	1,238
6	www.telegraph.co.uk	75,836	www.bbc.co.uk	1,196
7	www.scotsman.com	62,779	www.heraldsotland.com	1,141
8	www.heraldsotland.com	53,293	www.telegraph.co.uk	904
9	www.independent.co.uk	41,486	www.independent.co.uk	571
10	www.dailyrecord.co.uk	38,292	path.com	538
11	fw.to	33,732	fw.to	537
12	news.google.com	32,311	www.dailyrecord.co.uk	447
13	itunes.apple.com	31,034	www.buzzfeed.com	410
14	www.huffingtonpost.co.uk	27,910	www.huffingtonpost.co.uk	374
15	www.snp.org	25,075	www.facebook.com	372
16	twibbon.com	23,403	twitter.com	365
17	www.dailymail.co.uk	22,245	www.dailymail.co.uk	322
18	news.stv.tv	20,899	www.trendinalia.com	277
19	www.twitlonger.com	20,597	wingsoverscotland.com	277
20	www.nytimes.com	20,534	blogs.wsj.com	276

While this analysis shows some differences in terms of domain, or website level, destinations from link shares it cannot reveal anything about the content found at individual linked and shared URLs. AlchemyAPI, the Cloud-hosted NLP service (IBM, 2017a, 2017b) used to detect locations in four tranches of interaction message text (Section 5.2.2.2, p198), has also been used to determine how many mentions of geographical entities can be detected in linked/shared URL content during the US2012 and SCOT2014 campaigns. This work builds on the ‘system prototype for knowledge discovery from social media’ presented by Croitoru et al. (2013) to create new types of geographically-relevant results that have not previously been reported. While Croitoru et al.'s (2013) *Geosocial gauge* could interrogate social

media messages from a number of platforms, ‘starting with Twitter and Flickr’, it did not branch out to consider linked/shared URL content.

As the 3,485,840 links in the database include many duplicates, mainly due to retweeting, a table listing distinct links was created using SQL (Appendix 11 listing 31, p489) to be used as a ‘queueing’ table for AlchemyAPI processing. The column `ENTITY_JSON`, a JSON-constrained Character Large Object (CLOB) field to store AlchemyAPI responses, together with fields to uniquely identify rows and record processing date, enabled programmatic control of queue processing. The table stores 641,472 distinct linked/shared URLs, posted to AlchemyAPI servers using custom Ruby scripts (Appendix A10.4, p461) accessing data on the laptop host Oracle 12c database over TCP/IP and OCI8 middleware and executed periodically using the Linux `cron` job controller on a CentOS 7 virtual machine (Appendix 8, p436).

Appendix A10.5 (p466) shows the JSON returned by AlchemyAPI NLP software when run against a CNN (Cable News Network) URL reporting the results of the 2014 Scottish Independence Referendum. This is the 34th most shared link in the research data corpus, with 4,167 shares, and is still available online. The amount of augmented data returned by AlchemyAPI is substantial; 47 entities are detected in the HTML content at CNN’s URL, these entity types include `Country`, `City` and `Person`. In some cases, AlchemyAPI returns coordinates for identified entities in the JSON path `entities.disambiguated.geo`. In other cases, links to authoritative sources which store geographical information (e.g., GeoNames) are returned in the JSON path `entities.disambiguated.geonames`. Table 5-9 (p212) shows the breakdown of entity `type`, `relevance`, `text`, `count` and `geo` for the JSON returned by AlchemyAPI against CNN’s URL.

Using query capabilities built into Oracle 12c (Oracle, 2014a) a relational view was built over the path elements of the JSON returned from AlchemyAPI output, including the `NESTED PATH` represented by the array `entities` shown in

Appendix A10.5 (p466). The view, defined in SQL (Appendix 11 listing 32, p489), had to incorporate the `IS JSON STRICT` clause (Maram, 2017) in order to enforce strict validation of JSON. If not, SQL queries run against the view (counts, attempts to export to CSV etc.) would fail if any one record contained malformed JSON.

The view exposes 7,159,609 entities from the 641,472 distinct URLs passed through AlchemyAPI. Querying the view in SQL it is possible to select key entities for a given URL (Table 5-9 lists key features from the CNN URL whose AlchemyAPI JSON is shown in Appendix A10.5, p466), the entire data set (Figure 5-8, p215), or a subset of it. In Table 5-9 only 2 of the 47 detected entities have `disambiguated.geo` paths and corresponding Latitude and Longitude coordinates in JSON, the ‘Scottish Parliament’ and ‘Strichen’. This `City` entity type is, indeed, mentioned in the CNN article (McKirdy, Smith-Spark, & Robertson, 2014), where ‘Scotland's First Minister Alex Salmond, who has led the pro-independence "Yes Scotland" campaign, cast his ballot Thursday morning in the village of Strichen, Aberdeenshire.’

Table 5-9 – Entities, presented in tabular form, detected by AlchemyAPI against CNN’s Scottish Independence Referendum results page

type	relevance	count	text	geo
Country	0.813336	22	Scotland	(null)
City	0.436839	6	Glasgow	(null)
City	0.408424	5	Edinburgh	(null)
Company	0.321778	4	CNN	(null)
Person	0.321343	3	Alex Salmond	(null)
Country	0.288505	3	United Kingdom	(null)
Person	0.284499	2	Prime Minister David Cameron	(null)
City	0.260409	2	Edinburgh	(null)
StateOrCounty	0.233592	2	Aberdeenshire	(null)
Region	0.227693	2	Northern Ireland	(null)
Country	0.222848	2	Wales	(null)
Organization	0.222649	1	Scottish Parliament	55.95194 -3.17513
Country	0.220393	2	England	(null)
JobTitle	0.219891	1	Prime minister	(null)
Person	0.218625	2	Sue Bruce	(null)
JobTitle	0.21754	2	officer	(null)
Person	0.212998	1	Prime Minister Gordon Brown	(null)

City	0.208436	1	Dundee	(null)
FieldTerminology	0.207396	1	oil-rich city	(null)
Organization	0.206967	1	Glasgow City Council	(null)
City	0.196357	1	Aberdeen	(null)
Person	0.191401	1	Phil MacHugh	(null)
Person	0.188891	1	Alistair Darling	(null)
Person	0.187902	1	Nic Robertson	(null)
Organization	0.18525	1	EU	(null)
Person	0.184526	1	Mary Pitcaithly	(null)
City	0.183566	1	Hong Kong	(null)
City	0.183279	1	Dumfries	(null)
Person	0.181844	1	Angus	(null)
City	0.179558	1	London	(null)
Region	0.177794	1	East Dunbartonshire	(null)
Crime	0.177649	1	fraud	(null)
Person	0.174821	1	Euan McKirdy	(null)
City	0.174539	1	Kirkcaldy	(null)
City	0.173872	1	Galloway	(null)
City	0.172553	1	Strichen	57.5865 -2.0904
Person	0.170637	1	Laura Smith-Spark	(null)
Person	0.163373	1	Richard Allen Greene	(null)
Person	0.153864	1	Greg Botelho	(null)
Person	0.144819	1	Lindsay Isaac	(null)
Quantity	0.144819	1	17-year	(null)
Quantity	0.144819	1	46%	(null)
Quantity	0.144819	1	54%	(null)
Quantity	0.144819	1	75%	(null)
Quantity	0.144819	1	80%	(null)
Quantity	0.144819	1	86%	(null)
Quantity	0.144819	1	8%	(null)

While some other obviously geographical features (e.g., the `City` entities ‘Glasgow’ and ‘Edinburgh’) were correctly identified by AlchemyAPI, and disambiguated, the service does not always return coordinates in the `disambiguated.geo` JSON path. Altogether, AlchemyAPI returned coordinates for 182,619 records; 2.55% of all 7,159,609 entities identified in the set of 641,472 distinct linked URLs. These geotagged records are plotted in Figure 5-7 (p214), and represent the readily-mappable output of AlchemyAPI-based text-mining of all

linked/shared URL content deposited by users during the US2012 and SCOT2014 electoral events.

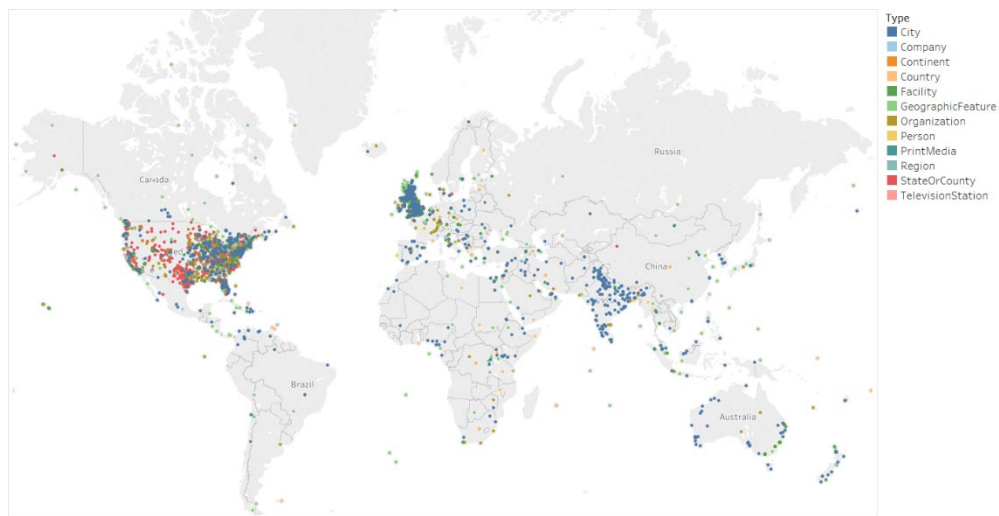


Figure 5-7 – US2012/SCOT2014: Disambiguated geographic coordinates identified by AlchemyAPI in 641,472 distinct link URLs colour-coded by entity type

The number of mappable entities detected by AlchemyAPI could be increased, by up to 433,846 records, if `disambiguated.geonames` were post-processed against the GeoNames (2016) gazetteer. However, to test RQ3, it is more useful to understand the distribution of entity types by Geographicality Score; i.e., do the most geographic coordinate-geotagging users link to the most geographically expressive online content such as, for example, news articles containing multiple toponymic references? Across all linked/shared URLs the most-detected entities (Figure 5-8, p215) are of type `Person` followed by `Organization`, `Countries`, `Quantities` and `Companies`.

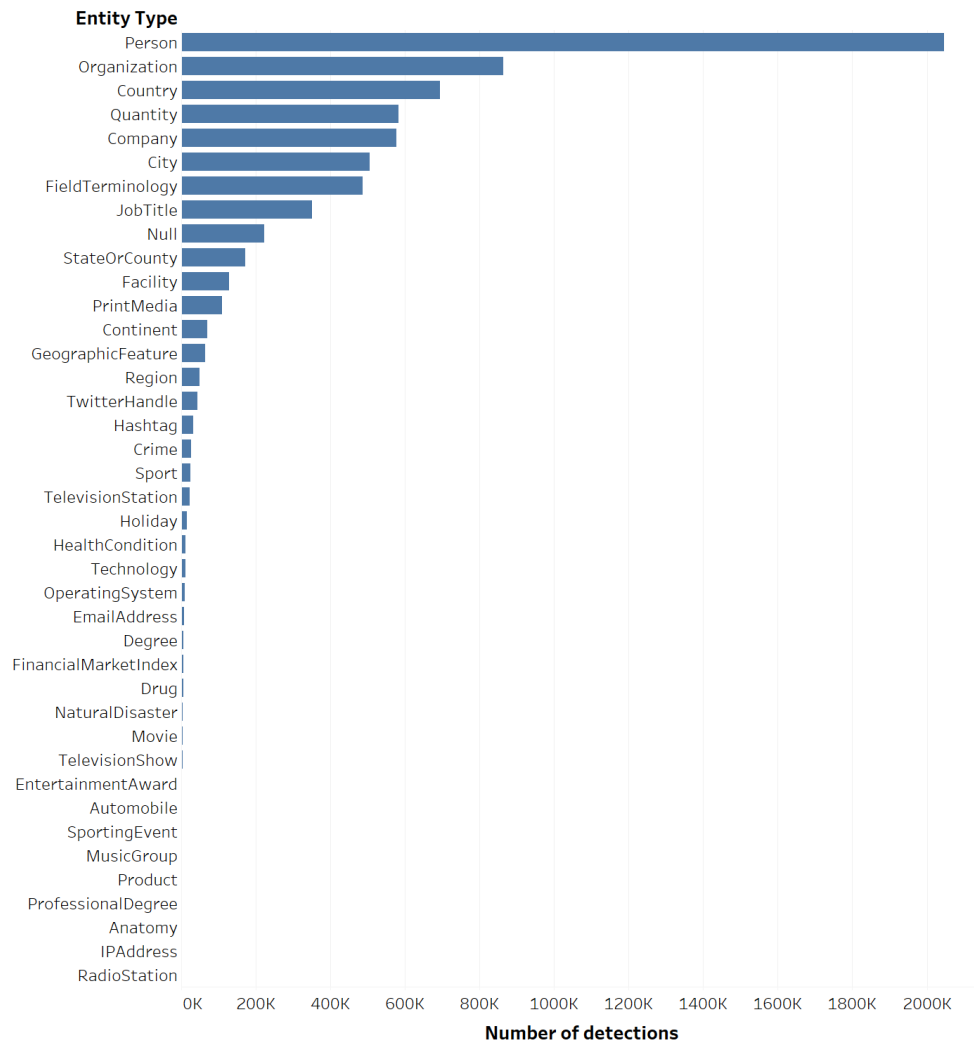


Figure 5-8 – US2012/SCOT2014: Number of distinct entities by type identified by AlchemyAPI in 641,472 distinct link URLs processed by the service

The top 10 most frequently identified entity types *Person* and *Organization*, together with top 10s for the three most frequently detected geographical entity types (*Country*, *City* and *StateOrCounty*) are shown in Figure 5-9 (p216). Key political figures, such as Alex Salmond, Barack Obama and Alistair Darling feature prominently, as do organizations including the SNP, EU (European Union) and GOP (Grand Old Party; Republicans). These named entities were, of course, frequently mentioned during online coverage of the two electoral contests chosen as case studies in this research (Section 4.2.4, p126).

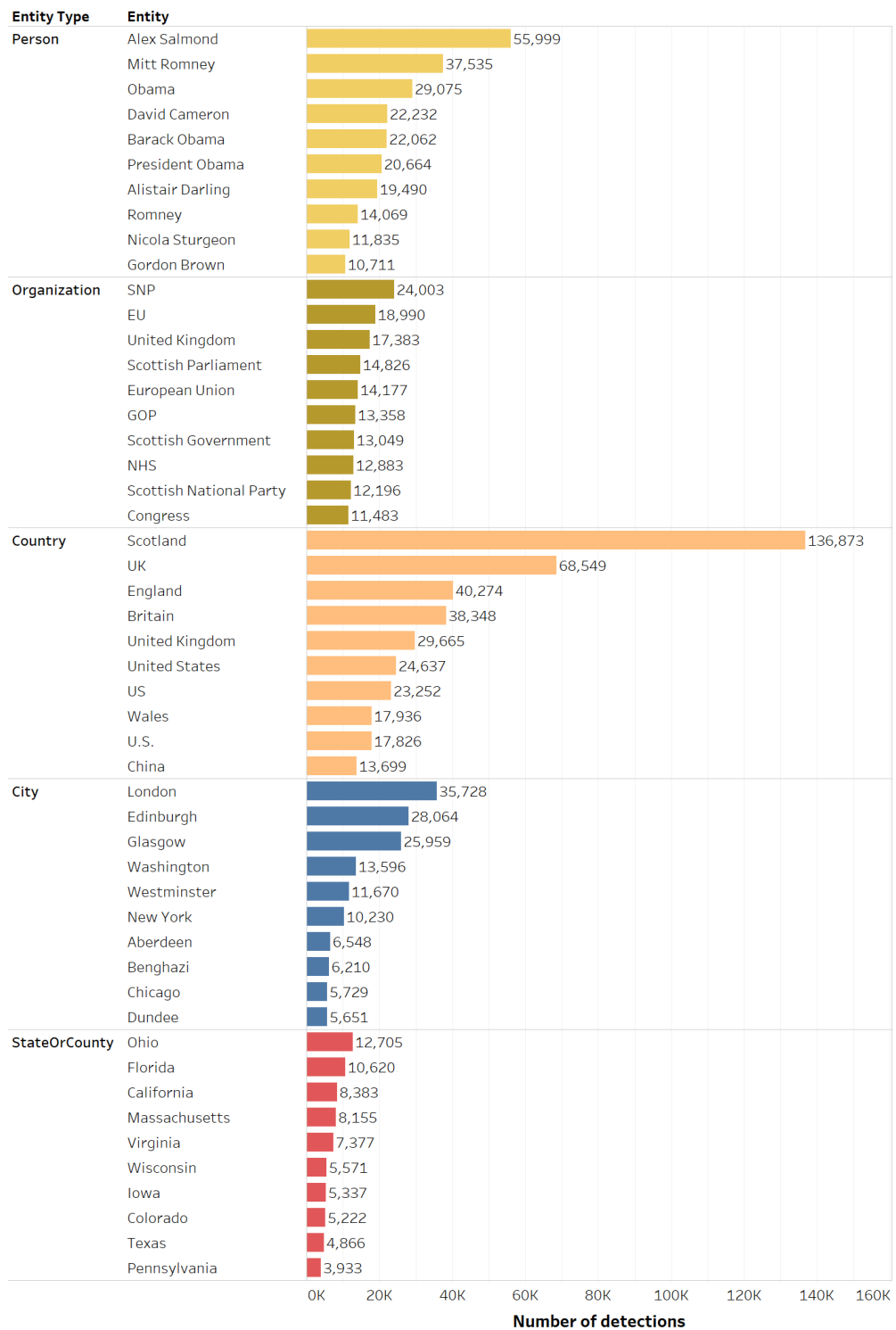


Figure 5-9 – US2012/SCOT2014: Top 10 entities for the two most detected entity types ('Person' and 'Organization') and three geographical entity types ('Country', 'City', 'StateOrCounty') identified by AlchemyAPI in 641,472 distinct link URLs

Key toponymic mentions, as expected, include Scotland, UK, United States, US and several major cities either side of the Atlantic including London, Edinburgh, Glasgow, Washington, New York and Chicago. States or Counties identified by AlchemyAPI are exclusively US-based and include Ohio, Florida, California and Massachusetts; either key swing or highly-populous states or, in the case of Massachusetts, the home state of Presidential Candidate Mitt Romney. Entities identified by AlchemyAPI NLP software in linked URL content clearly reflect contemporary commentary found, and shared, online at the time of the US2012 and SCOT2014 campaigns. Even the initially puzzling inclusion of Benghazi, as the eighth most-frequently identified city entity in Figure 5-9 (p216), can be explained by an attack against the US Consulate in that Libyan city which took place in 2012, fallout from which was widely viewed as a potentially ‘game-changing’ political moment during the course of the 2012 US Presidential Election (McGreal, 2012).

Data from AlchemyAPI text-mining of ~650,000 linked/shared URLs offers an accurate and highly searchable reflection of contemporaneous online text-based content surrounding US2012 and SCOT2014 events. By joining the AlchemyAPI JSON tables and relational views stored in Oracle 12c to views of Geographicality Scores at interaction and user levels (Section 4.6.1, p164) it is possible to determine whether patterns of link sharing by numbers of detected geographical entities in URL content differ according to Geographicality Score, answering RQ3. This metric is calculated as the average number of mentions of toponymic or coordinate geography (‘geo-entities’) detected in linked/shared URL content by AlchemyAPI NLP software at interaction (Appendix 11 listing 33 and 35, p489) and modal user levels (Appendix 11 listing 34 and 36, p490). Figure 5-10 (p218) shows the average number of geo-entities detected per interaction by AlchemyAPI using this logic. This count is cross-tabulated against grouped Geographicality Scores calculated earlier for each interaction.

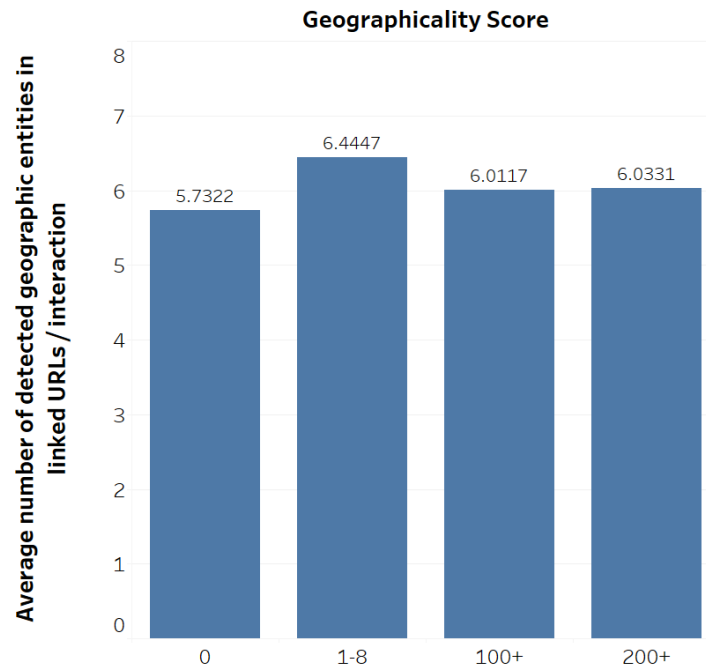


Figure 5-10 – Average number of toponymic or coordinate mentions identified by AlchemyAPI in linked URLs by grouped Geographicality Score at interaction level

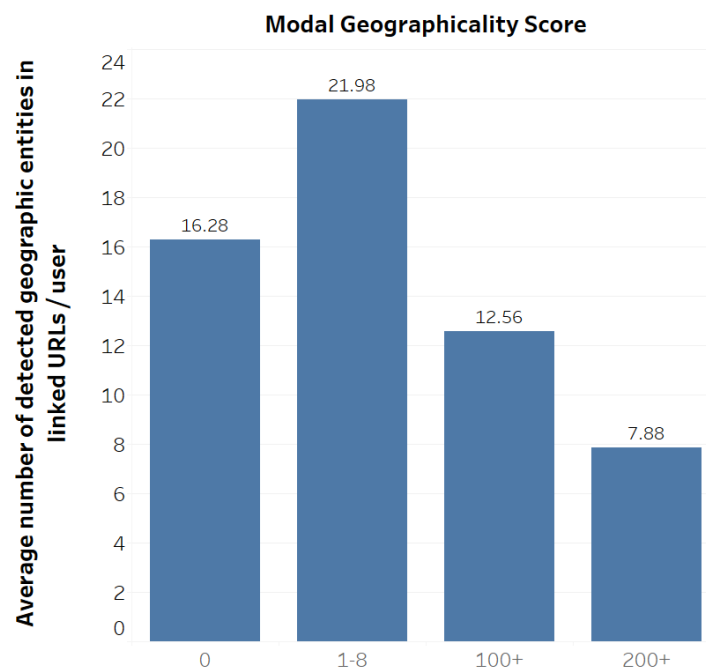


Figure 5-11 – Average number of toponymic or coordinate mentions identified by AlchemyAPI in linked URLs by grouped Modal Geographicality Score at user level

All classes at interaction level exhibit similar average numbers of geo-entities detected in linked/shared URL content. At user level (Figure 5-11, p218), averaging numbers of geo-entities detected byAlchemyAPI in linked/shared URLs for all links made by each user, cross-tabulated against the modal Geographicality Score for each user, the distribution differs. Users with a zero Geographicality Score (no PGI metadata; mainly Facebook users) link to content containing, on average, 16.28 AlchemyAPI-detectable geo-entities in their linked/shared URLs. Users with some PGI metadata (classes 1-8 in Figure 5-11) link to content containing, on average, 21.98 geo-entities. Amongst coordinate-geotagging users who tweet (100+) or retweet (200+) the number of AlchemyAPI-detectable geo-entities in linked/shared URL content is lower, at 12.56 and 7.88 detected geo-entities per user respectively.

This analysis shows that NLP-detectable ‘geographicality’ in linked/shared URL content does not increase in line with ‘spatiality’; the most spatially expressive coordinate-geotagging users do not link to the most toponymically expressive content. Users tweeting, retweeting or posting with coordinate-geotags are not only much less likely to link to external content altogether (Table 5-6, p207) but the content they link to contains fewer mentions of NLP-detectable geo-entities than that shared by non-coordinate-geotagging users. This is an original result which has not been reported elsewhere in the scientific literature.

5.3 Statistical tests

Summary results from Welch’s paired T-tests (Section 4.5.3, p163) comparing the distribution of numbers of toponymic mentions detected by the geoparsers used to examine the message text and linked/shared URL content of coordinate-geotagged and non-coordinate-geotagged interactions, and coordinate-geotagging and non-coordinate-geotagging users, are presented in Table 5-10 (US2012, p220) and Table 5-11 (SCOT2014, p220). The tables summarise the statistical significance of differences in numbers of toponyms detected in geotagged and non-geotagged

Facebook posts, Twitter tweets and retweets, at both interaction and user levels, by each NLP/geoparser used.

Table 5-10 – US2012: Summary statistics of Welch’s paired T-tests for numbers of detected toponyms in geotagged/non-geotagged message text and linked/shared URLs at interaction and user levels by OSN type/subtype and parser

US2012 - INTERACTION LEVEL						
LEVEL						
OSN TYPE	Facebook post		Twitter tweet		Twitter retweet	
PARSER	t > ±2	p < .05	t > ±2	p < .05	t > ±2	p < .05
GATEcloud	NA	NA	✓	✓	✗	✗
AlchemyAPI			✗	✗		
CLAVIN-rest	NA	NA	✓	✓	✓	✓
AlchemyAPI (URLs)	NA	NA	✗	✗	✓	✓
US2012 - USER LEVEL						
LEVEL						
OSN TYPE	Facebook post		Twitter tweet		Twitter retweet	
PARSER	t > ±2	p < .05	t > ±2	p < .05	t > ±2	p < .05
GATEcloud	NA	NA	✓	✓	✓	✓
AlchemyAPI			✓	✓		
CLAVIN-rest	NA	NA	✓	✓	✗	✗
AlchemyAPI (URLs)	NA	NA	✓	✓	✗	✗

Table 5-11 – SCOT2014: Summary statistics of Welch’s paired T-tests for numbers of detected toponyms in geotagged/non-geotagged message text and linked/shared URLs at interaction and user levels by OSN type/subtype and parser

SCOT2014 - INTERACTION LEVEL						
LEVEL						
OSN TYPE	Facebook post		Twitter tweet		Twitter retweet	
PARSER	t > ±2	p < .05	t > ±2	p < .05	t > ±2	p < .05
GATEcloud	✗	✗	✓	✓	✓	✓
AlchemyAPI			✓	✓		
CLAVIN-rest	✓	✓	✓	✓	✓	✓
AlchemyAPI (URLs)	✗	✗	✓	✓	✓	✓
SCOT2014 - USER LEVEL						
LEVEL						
OSN TYPE	Facebook post		Twitter tweet		Twitter retweet	
PARSER	t > ±2	p < .05	t > ±2	p < .05	t > ±2	p < .05
GATEcloud	✓	✓	✓	✓	✓	✓
AlchemyAPI			✓	✓		
CLAVIN-rest	✓	✓	✓	✓	✓	✓
AlchemyAPI (URLs)	✓	✓	✓	✓	✓	✓

In Table 5-10 and Table 5-11 a green tick indicates that the null hypothesis can be rejected, i.e. that differences do exist in the number of toponymic mentions detected in geotagged/non-geotagged interaction message text or linked/shared URL content. In 27 out of 40 cases, like-for-like comparisons are statistically significant ($t > \pm 2$) with >95% confidence. In 6 other cases statistics could not be calculated owing to a complete lack of coordinate-geotagged Facebook posts in the US2012 data set (marked NA in Table 5-10, p220). This finding is itself significant, and must reflect DataSift's changing access to Facebook-sourced OSN interactions over time, as both coordinate-geotagged and non-coordinate-geotagged Facebook interactions are present in the SCOT2014 data set (Table 5-11, p220). During the 2014 Scottish Independence Referendum event, sampled by one consistent and long-running 1:1 DataSift Stream, most like-for-like comparisons of numbers of toponymic detections (e.g., in Facebook posts, Twitter tweets or retweets by GATEcloud, AlchemyAPI or CLAVIN-rest) are statistically significant.

Detailed results show that coordinate-geotagged interactions hold, and coordinate-geotagging users make, fewer toponymic mentions than their non-coordinate-geotagged/tagging peers (Appendix A12.2, p502).

5.4 Software evaluation

Results obtained from three, quite different, computerised NLP/geoparsing systems are presented above. Several other geoparsers, such as Baleen (Defence Science and Technology Laboratory, 2015) and the Edinburgh Geoparser (Language Technology Group, 2014), were also tested but either failed to compile or could not read the input files and, hence, could not be properly evaluated (Section 4.4.1.4, p157). The difficulties encountered here in installing and running open-sourced geoparsing software have also been encountered by others. Gritta et al. (2018, p619), for example, comment on the 'prohibitively cumbersome [software] set up' involved in their tests of five geoparsing systems, reporting 'substantial disparity' in

terms of software availability, installability and support. The following section gives a comparative evaluation of the three systems used in this research.

5.4.1 Comparative evaluation

All NLP/NER-based entity extraction engines to differing degrees, can easily be fooled by input sentence structure; a significant problem when much of the text in the research data corpus is terse and ungrammatical. Consider the following (lucid) sentences mentioning locations:

- I used to live in London
- After visiting Newcastle I went to Durham
- I went past Stonehenge the other day
- I'd love to know what's happening in Newquay tonight

These locations, with many country alternatives (e.g., London, Texas), all appear in the GeoNames gazetteer. The CLAVIN online demonstrator (Berico-Technologies, 2018) successfully resolves and provides coordinates for the locations ‘City of London’ and ‘Newquay’; a success-rate of 40%. Changing the second sentence to read ‘After visiting Newcastle I went *over* to Durham’ adds ‘County Durham’ to the list of resolved locations, boosting the success rate (allowing for semantic differences between County Durham and the City of Durham) to 60%. Changing the third sentence to ‘I went to Stonehenge the other day *near Amesbury*’ resolves Amesbury but continues to miss Stonehenge. Newcastle is missed every time, although if the sentence is changed to ‘*I lived in Newcastle and often* went to Durham’ Newcastle is resolved but Durham is not. CLAVIN-rest, compiled on a Centos 7 virtual machine, resolves London, Durham and Newquay in the initial sentence and returns coordinates in JSON, finds Amesbury (but not Stonehenge) when the sentence is modified and finds both Newcastle and Durham in the final iteration of sentence structure.

The AlchemyAPI demonstrator (IBM, 2018) does somewhat better, finding all four location entities except Stonehenge in the original sentence. It does not return coordinates for any of these locations in JSON and, while it misses Stonehenge as a 'location entity', it does recognise it as a 'keyword'. GATE Developer desktop software also finds all location entities, except Stonehenge, when presented with the initial sentence structure and the online GATEcloud demonstrator (GATE, 2018) produces the same result. Neither desktop or Cloud variants of GATE return coordinate data. Like AlchemyAPI, TwitIE on GATE and GATEcloud is a specialist NLP/NER system tuned to extract multiple entity types from text. While GATE/GATEcloud do not return coordinates by default alongside identified location entities, University of Sheffield developers are working on this functionality in future releases by creating bespoke software processing pipelines (Roberts & Tear, personal communication, 2017). As an open-source product, GATE Developer desktop software is free to use. Running jobs on GATEcloud costs, but costs substantially less than it would on AlchemyAPI when processing large data sets, particularly for academic users, and has no daily rate restrictions (or 'throttling') for research use.

In terms of speed, with the computing resources available (Appendix 8, p436), CLAVIN-rest offers the fastest solution for geoparsing, processing large files within minutes on a CentOS 7 virtual machine using a local copy of the GeoNames (2016) gazetteer database. However, CLAVIN-rest *only* geoparses, and does not resolve any other entities in text. Resolved locations can be used to add some value to maps, by showing additional locations found within text (e.g., Figure 5-12 and Figure 5-13, p224), but the software cannot identify people, URLs, quantities, companies, key words or sentiment etc. or provide any additional contextual information. Knowing that a place *is* mentioned, even with the addition of coordinates, is not as valuable as knowing *why* it is mentioned. CLAVIN-rest therefore adds less inferential value to analyses than the other, more general-purpose, NLP/NER systems evaluated here.

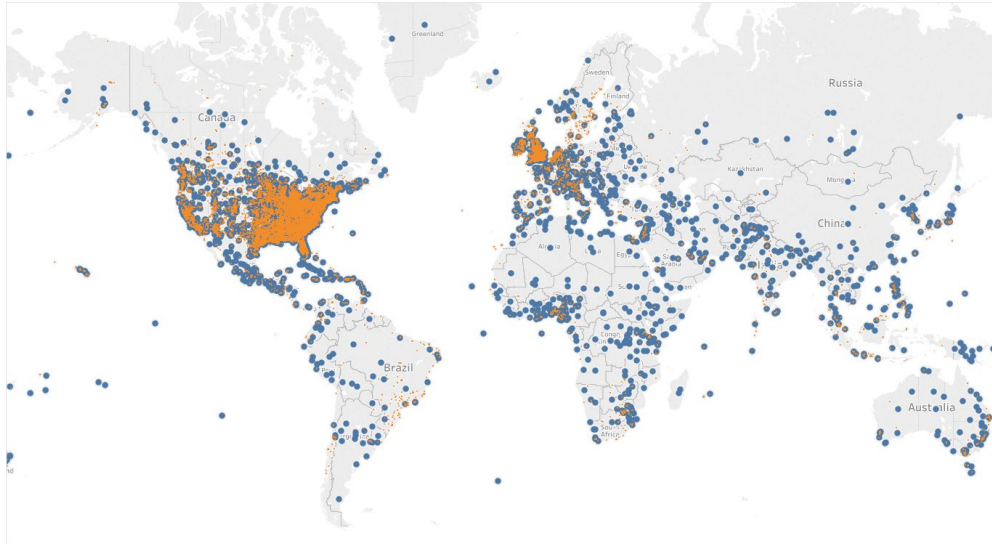


Figure 5-12 – US2012: World map showing coordinate geotagged interactions (orange markers) and geoparsed locations (blue markers) identified by CLAVIN-rest

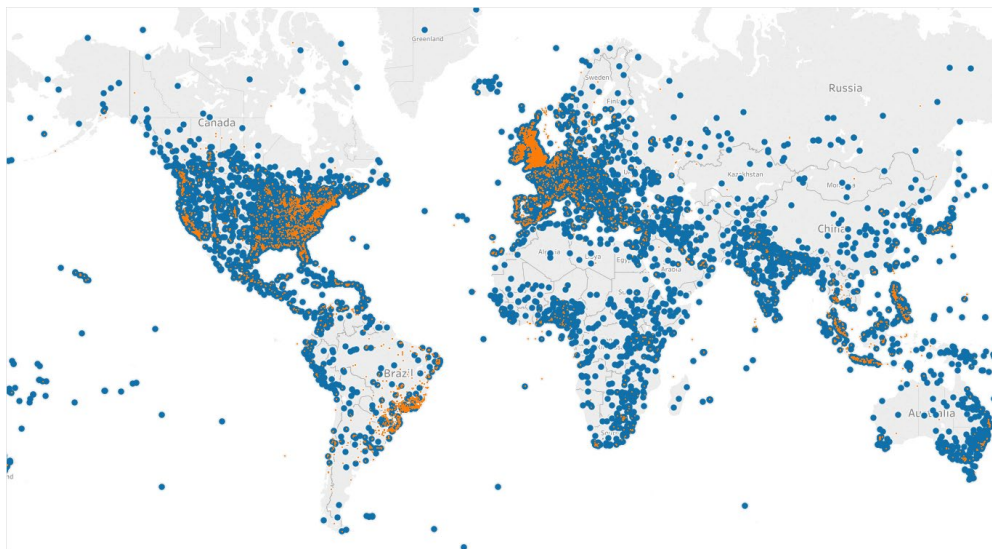


Figure 5-13 – SCOT2014: World map showing coordinate-geotagged interactions (orange markers) and geoparsed locations (blue markers) identified by CLAVIN-rest

Given the ≤ 140 -character message length limit in $\sim 90\%$ of the research data corpus, comprised largely of Twitter tweets and retweets (Table 4-8, p170), it is sensible to adopt an NLP solution aimed specifically at microblog text. As GATE's development team have noted (Derczynski et al., 2013, p21), 'Using semantic technologies for mining and intelligent information access to microblogs is a challenging, emerging research area. Unlike carefully authored news text and other longer content,

tweets pose a number of new challenges, due to their short, noisy, context-dependent, and dynamic nature.’ The median length of all user interactions in the research data corpus is 127 characters, falling precisely into the ‘short, noisy’ category identified by Derczynski et al. (2013).

While the different NLP-based NERs evaluated here have different strengths, weaknesses and costs when used as geoparsers they all point to a similar result: *the most geographic coordinate-geotagging OSN users are somewhat less geographically expressive* when it comes to mentioning NLP/NER-detectable toponymic locations in their online message text or linked/shared URL content. This conclusion has not been reached elsewhere.

5.5 Summary

In a geographical context, coordinate spatiality in Twitter and Facebook communications is characterised by a ‘lack of user adoption of geo-based features [which suggests that] the promise of [OSN data] as a location-based sensing system may have only limited reach and impact’ (Z. Cheng et al., 2010, p1). Just as massive data sets open up the possibility for straightforward mapping of many hundreds of thousands, or millions, of human interactions it is apparent that most of these social media interactions are not imprinted with spatial coordinates. Furthermore, the results presented in this chapter show that spatialised social media message text, or linked/shared URL content, deposited on Facebook and Twitter by coordinate-geotagging users is not toponymically representative of the majority of interactions created on these platforms by non-coordinate-geotagging users. In summary:

1. Fewer toponymic mentions are found in the message text of coordinate-geotagging users’ interactions, whether on Facebook or Twitter.
2. Coordinate-geotagging users make fewer links to 3rd party content in their interactions than others on both Facebook and Twitter.

3. The linked/shared content of coordinate-geotagging users contains fewer toponymic mentions than that shared by others on both platforms.
4. Coordinate-geotagging users link to and share content from largely overlapping, but different, sets of Web domains to others.

The results of this research:

- Demonstrate the validity of the methodological approach adopted, and methods used, to examine politically discursive online social media interactions collected during two separate electoral events, using two social media data sources and three alternative NLP/geoparsing systems.
- Refute the *Geographicality Assumption* tested here that *coordinate-geotagging users are the most geographically expressive of all OSN users*.

These findings suggest that tracking or mapping the spread of political opinion or (mis)information by searching for toponymically-infused message text or linked/shared URL content, deposited solely by coordinate-geotagging OSN users of two major social media platforms during electoral campaigns, will be inaccurate. The implications of this conclusion are discussed in more depth in the following chapter, alongside several additional findings resulting from the exploratory spatiotemporal research methodology adopted throughout this work.

6 DISCUSSION AND ADDITIONAL FINDINGS

6.1 Introduction

The World Wide Web is a vast decentralised content repository, built upon an interconnected network of computing infrastructure originally designed to withstand full-blown nuclear conflict (Salus, 1995). Much of the Web's content consists of unstructured text, predominantly created by human hands. Scharl (2007, p6), citing Delboni, Borges, & Laender (2005), has suggested that 'At least 20 percent of Web pages contain easily recognizable and unambiguous geographic identifiers.' A broadly similar percentage is evident in social media message text, where around 25% of the ~8m interactions in the research data corpus are found to contain toponymic references to place (Section 5.2.2, p190). Few records (~1-2%) are spatially imprinted with Latitude and Longitude pairs (Table 4-8, p170) while statistically significant results show that coordinate-geotagged interactions contain fewer toponymic mentions in message text and linked/shared URL content than that found in corresponding, non-coordinate-geotagged, social media data.

Place detection in OSN interactions may be attempted using the `LIKE` or `CONTAINS` functions in SQL and a full-text index (Oracle, 2012) on interaction content. Searching for Perth, for example, finds five mentions of the Fair City during the 2012 US Presidential Election, with many more occurrences (n=560) found during the 2014 Scottish Independence Referendum. The SQL query works (Appendix 11 listing 38, p491), but the approach does not work well for all places:

- Many identical place names appear both within-country (e.g., Newport, UK) and across-country (e.g., Perth, Scotland; Perth, Western Australia). Computerised searching for string literals, or matching to published gazetteers, may mismatch common place names.

- Some place abbreviations (e.g., 'IN' for Indiana, 'OH' for Ohio), however cased, are commonly used in OSN message text. A search against message text in the US2012 data set using the SQL phrase `LIKE '% IN%'` returns 11,320 OSN interactions, but not all of these refer to Indiana.
- Computerised text-matching against very large gazetteers, e.g., the popular, open-source GeoNames (2016) database of over 11 million place names, may be attempted using looping PL/SQL programmes or pattern-matching database indices but is best attempted using specialist NLP/geoparsing software capable of breaking down terse or ungrammatical social media message text into parts of speech (POS; nouns, verbs, adverbs etc.), using these structures to determine likely location-bearing constructs in text.

As Gritta et al. (2018, p603) have observed, 'The ability to geo-locate events in textual reports represents a valuable source of information in many real-world applications such as emergency responses, real-time social media geographical event analysis, understanding location instructions in auto-response systems and more. However, geoparsing is still widely regarded as a challenge because of domain language diversity, place name ambiguity, metonymic language and limited leveraging of context.' In this research, Natural Language Processing 'pipelines' including TwitIE (Bontcheva et al., 2013), AlchemyAPI (IBM, 2017a) and CLAVIN-rest (Berico-Technologies, 2014) have been used to search for geo-references within Twitter and Facebook message text and linked/shared URL content. The research demonstrates that coordinate-geotagging users are *not* always the most geographically expressive of all OSN users (Section 5.3, p219). The implications of this finding are discussed in the following section while Section 6.4 (p246), later in this chapter, presents the results of several further investigations conducted during this research, each of which shed additional light on the sometimes perplexing 'confounds and consequences' (Pavalanathan & Eisenstein, 2015) of coordinate-geotagged and non-coordinate-geotagged social media data.

6.2 Implications

6.2.1 Key implication

The key implication of this study, in the contemporary political context within which it is situated, centres around the efficacy of using coordinate-geotagged OSN interactions to geographically track, map or monitor the dispersal of opinion or (mis)information produced and shared online over social media during electoral campaigns. This may be illustrated through the logical progression below:

- Various authors, as noted in the introduction to this thesis (p1), have suggested there are grounds for believing that geo-behavioural targeting may have impacted the outcome of the 2016 US Presidential Election and the 2016 UK European Union Membership Referendum.
 - Albright (2017), at the Tow Center for Digital Journalism, has found computer code revealing how geographical and behavioural targeting techniques were used by an employee of Cambridge Analytica to geolocate and mine Twitter data in 2016, which ‘shows the inner workings of client voter file geo-data “enrichment” and presumably automated voter database processing for clients by Cambridge Analytica.’ A snippet of this code with the programmer’s detailed comments is shown in Figure 6-1 (p233).
 - Documents laid before Congress by Facebook during the U.S. House of Representatives (2018c) Permanent Select Committee on Intelligence inquiry into Social Media Advertisements detail how ~3,500 geo-behavioural campaigns were set up by Russian state-sponsored actors using Facebook’s own campaign management and targeting facilities (Figure 6-2, p235). *The New York Times* (Collins, 2018) has built a useful interactive web page around these data, allowing users to see what sorts of advertisements were being

shown on social media in 2015-17 to different age groups, in different regions of the US, with different interests.

- This evidence builds a picture of how social media have been (mis)used to target and direct political, and often populist or inflammatory, material at voters in given areas. However, to date, there remains no compulsion for political parties, marketing companies or campaign teams in the US, UK or elsewhere to release this information.
- Subsequently, electoral officials and researchers can only view the end results of marketing or (mis)information campaigns by searching for the spread of messages, content or URLs publicly-posted and shared on OSN platforms, blogs or websites.
 - While it is relatively straightforward to track the spread of messages or URLs amongst users of social network sites or, at least, those users (predominantly on Twitter) who publicly-post their messages, it is much more difficult to know where these users are located and, hence, whether geo-targeted messages are reaching, or possibly influencing, audiences in specific areas.
- Coordinate-geotagged interactions, and coordinate-geotagging users, offer one mechanism through which accurate spatial tracking could be achieved. However, this research has shown that:
 - Few politically discursive OSN interactions are coordinate-geotagged (Section 5.2.1, p188)
 - Coordinate-geotagging users do not refer to locations in message text, or link to URLs referencing locations (which might have been targeted or promoted), as often as non-coordinate-geotagging users (Section 5.2.2, p190).
 - Coordinate-geotagging users do not share as much content, or quite the same sorts of content, as non-coordinate-geotagging users of the Twitter and Facebook platforms studied (Section 5.2.3, p205).

- Both of these conclusions, along with other results indicating probable youthful skewness and urban living amongst these users (Section 6.4.4, p262), suggest that coordinate-geotagged social media interactions from geotagging users, which comprise only a small 1-2% minority of all OSN messages, are not sufficiently representative to allow for accurate geographical tracking of the online dispersal of opinion or (mis)information disseminated during electoral campaigns.

Steiger, de Albuquerque, et al. (2015, p826) have stated that across all application domains they reviewed, and particularly in disaster management, ‘georeferenced tweets provided accurate location information [with] study outcomes [demonstrating] a high spatiotemporal reliability and usefulness of tweets.’ The authors have suggested (p826) that ‘Earthquake detection from Twitter is one successful example in a number of reviewed studies where disaster events have been localized in a real-time manner, showing a high correlation in comparison with official earthquake sensor data. A similar outcome can be stated within the application of disease and health management. Tweets indicating disease incidents have shown a similar spatiotemporal distribution in comparison with official reports.’ In natural disaster, emergency or terrorism situations, or in use cases which examine large-scale population trends and movements, OSN data – mainly sourced from Twitter – have demonstrated high degrees of utility. When tracking the geographical spread of opinion or (mis)information online, or attempting to use messages or shared content posted on Twitter or Facebook to poll or predict political outcomes, the results are much more ambiguous (Section 7.3, p292).

It is much easier to set up a geographically-targeted campaign, political or otherwise, on Twitter, Facebook, Instagram etc. than it is to use publicly-available OSN interaction (meta)data sourced from these platforms to accurately monitor and assess the reach of such a campaign. Figure 6-1 (p233) shows lines 191-239 of a 401-line programme written in Python by Michael Phillips, an intern working for

Cambridge Analytica, whose GitHub repository was accidentally left open and whose work has been investigated by ‘Professor and researcher in news, journalism, and #hashtags’, Jonathan Albright (2017).

```

"""
This function is really the purpose of this script.
Essentially what it does is:
For each address in the addresses file, try to get an
accurate lng/lat quickly (comparing available data
from Aristotle/IG to the zip code file data to determine
accuracy), but if we can't, we fetch it from ArcGIS.
addresses is an array of addresses each in the form
    , address_id, voter_id, AddressLine, ExtraAddressLine,
HouseNumber, PrefixDirection, StreetName, Designator,
SuffixDirection, ApartmentNum, Zip, ZipPlus4, City, County,
CongressionalDistrict, State, latitude, longitude,
ar_latitude, ar_longitude
    only lines 5, 7, 8, 13, 16 are used though, the rest
can be blank.
    lines 17,18,19,20 are optional, they are the data from
Aristotle and IG lat/lng data.

_zips is an array of zip codes in the form:
zip, city, state, latitude, longitude, timezone, dst
    where latitude and longitude correspond to the center
of the zip code.
    note that zip codes should be in the same format as
provided by the addresses file.  sometimes this means
trimming leading zeroes.

latLngFunc is the function you want to use to fetch
lat/lngs that are not supplied in the addresses array.
    getLatLngArcGIS is recommended due to accuracy and the
fact that there is no usage restriction.

The function returns an array which adds 2 extra columns to
the original addresses array.  The extra columns are the
accurate lat/lngs.
Output is a little confusing, but the important bits are
the fetch rate (basically dictates how quickly the function
goes),
    and the errors (which is going to be the number of
address lines which we couldn't get accurate data for from
any source.)
*****THINGS THAT CAN BE ADDED*****
Right now the exception clause basically just says that the
latLngFunc failed to find the address and give a valid
lat/lng.
    Instead, it could use googleMaps to fill in the field.
It would have to be limited to 2500 uses in a day, but at
the average hit rate for
    ArcGIS, it would be difficult for it to fail 2500 times
in a single day.  This would increase coverage to possibly
100%, but if we did

```

```

hit our googleMaps fetch limit, it could cause the
program to crash or take virtually forever.

"""
def completeAddresses(addresses, _zips, latlngFunc):
    completeAddr = np.hstack((addresses,
np.zeros([len(addresses), 2])))
    completeAddr = pd.DataFrame(completeAddr)
    #create dictionary for zip lookups
    zipList = _zips[:,0].tolist()
    latitudeList = _zips[:,3]
    longitudeList = _zips[:,4]
    latlngList = zip(latitudeList, longitudeList)
    zipDict = dict(zip(zipList, latlngList))

    radius = 15

    errorCount = 0
    igHitNumber = 0
    igMissNumber = 0
    arHitNumber = 0
    arMissNumber = 0
    numberOfFetches = 0

    for index, line in completeAddr.iterrows():

```

Figure 6-1 – Detailed comments and snippet of ‘geo-data “enrichment” code’ created by Cambridge Analytica employee Michael Phillips (Sources: Albright, 2017; archive.today, 2017)

This computer code trawls Twitter data to match addresses and expands a list of sentiment-specific keyword groups or concepts which Albright associates with the 2016 Trump campaign, e.g., ‘hilarySentiments’, ‘gunsSentiments’ etc. The code references Aristotle, an internal Cambridge Analytica system, and ‘IG’; an abbreviation for Instagram. ArcGIS, the Geographical Information System from ESRI (2018), is also referenced in the code along with comments referring to rate limits in the Google Maps API, also used to geocode social media data. The programme shows – as previous sections and programmatic listings in Appendices to this thesis also show – how computers, databases, systems and code may be used to interrogate and augment OSN data and how, in the case of Cambridge Analytica, this augmented database was used, to target political campaigns on `voter_id`

and geo-attributes including City, County, CongressionalDistrict and State.

In an interview with *The Guardian* (Cadwalladr, 2018a) Christopher Wylie, the whistle-blower who helped bring the Cambridge Analytica scandal to light, describes how he ‘ended up creating “Steve Bannon’s psychological warfare mindfuck tool”’. In the video which accompanies the piece, Wylie states that he and Cambridge Analytica were ‘playing with the psychology of an entire nation’, using data science and machine learning to ‘combine micro-targeting with new constructs from psychology’ to ‘build cultural weapons.’ Data harvested from Facebook, using Kogan’s online quiz app (Etter & Frier, 2018), had only ‘to touch a couple of hundred thousand people’, Wylie states in *The Guardian* (Cadwalladr, 2018a) interview, to fan out through Facebook’s friend network and ‘scale to the entire US’, collecting an estimated 50-60 million user profiles; a figure which later emerged as an under-estimate of the ~87 million profiles really ‘harvested’ by this operation (BBC News, 2018e). Personal data, including information about people’s Facebook Likes and Interests, were used to determine what ‘kinds of messaging’ individual voters were susceptible to and ‘where [they] were going to consume’ targeted messages. Teams of ‘creatives, designers and geographers’ were involved in making and targeting bespoke messages which, according to Wylie, ‘whispered into the ears’ of individual voters; pushing them towards specially created websites, blogs and content designed to reinforce messages and political standpoints. Tracing the source and tracking the consumption of such material is a necessity if free and fair democratic elections are to continue in an era in which, as a former chief of the UK’s General Communications Headquarters (GCHQ) has said, social media organisations ‘have huge power’ over governments (BBC News, 2018c).

The growing ‘weaponisation’ of social media (Nissen, 2015) has become most visible in the US Central Intelligence Agency’s (CIA) detection of Russian state-

sponsored involvement and interference in the 2016 US Presidential Election, in support of Presidential Candidate Donald Trump's campaign.

Ad ID 1254

Ad Text Secured borders are a national priority. America is at risk and we need to protect our...

Ad Landing Page <https://www.facebook.com/Secured.Borders/>

Ad Targeting Location - Living In: United States

Interests: Immigration to the United States, Conservatism, Deportation, Stop Illegal Immigration, Julian Assange, Laura Ingraham, Ron Paul, National identity, Welfare state, United States Department of Homeland Security, Donald Trump, Bill O'Reilly (political commentator), Illegal immigration, WikiLeaks, Law enforcement, Republican Party (United States), Patriotism, Old Glory, United States Constitution, Immigration law, Conservatism in the United States, Foreign policy, Christopher Hitchens, United States Bill of Rights, Sean Hannity, Michael Savage, Mike Huckabee, Racism in the United States or Politics and social issues

Excluded Connections: Exclude people who like Secured Borders

Age: 18 - 65+

Placements: News Feed on desktop computers, News Feed on mobile devices or Right column on desktop computers

Ad Impressions 184


Ad Clicks 16

Ad Spend 204.19 RUB

Ad Creation Date 07/16/17 03:51:18 PM PDT


Ad End Date 07/19/17 03:51:23 PM PDT

Suggested Page



Secured Borders
Sponsored

Secured borders are a national priority. America is at risk and we need to protect our...



Secured Borders
News & Media Website
135,301 people like this.

Like Page

Redactions Completed at the Direction of Ranking Member
of the US House Permanent Select Committee on Intelligence

P(1)0003117

Figure 6-2 – Facebook campaign targeting parameters, and advertisement, for one of ~3,500 campaigns/advertisements set up by Russian state-sponsored actors during the 2016 US Presidential Election

Figure 6-2 shows the targeting criteria used in one of the ~3,500 advertisements set up on Facebook and designed, according to the U.S. House of Representatives

(2018a) Permanent Select Committee on Intelligence, to ‘sow discord online’. In this operation ‘[Russian] Defendants, posing as U.S. persons and creating false U.S. personas, operated social media pages and groups designed to attract U.S. audiences. These groups and pages, which addressed divisive U.S. political and social issues, falsely claimed to be controlled by U.S. activists when, in fact, they were controlled by Defendants. Defendants also used the stolen identities of real U.S. persons to post on ORGANIZATION-controlled social media accounts [reaching] significant numbers of Americans for [the] purposes of interfering with the U.S. political system, including the presidential election of 2016’ (U.S. House of Representatives, 2018a, authors' capitalisation). Tracking this sort of interference in large social media data sets is not straightforward:

- The cost of purchasing and consuming entire data streams may be prohibitive outside governmental or law enforcement agencies.
- Platform operators’ advertisement placement and campaign management systems are unavailable to all, unless legally requested.
- Platform operators’ privacy policies, in many cases, preclude access to all or most user data, or to some data points that would be highly useful.
- Geographical tracking is hindered by redaction of IP addresses in metadata and low rates of coordinate-geotagging, typically just ~1-2%.
- Spatialised interactions and the users who create them are not entirely representative of all OSN users, as this research has shown.

Van Dijck (2014) has stated that Web users now accept that they ‘exchange’ their personal data for Web-hosted services including search, email and social media. Many users of these systems think that they are the customers of these services, even though they use them for free. However, the real customers of major Internet corporations including Facebook, Google and Twitter are the advertisers who pay to promote their content on these platforms, whether endorsing a product, a party, a candidate, an ideology or some form of misinformation. Much more transparency

in online advertising is desirable, particularly where it is used in support of political campaigning via highly-targetable social media networks. Policy recommendations in this area are discussed below (Section 6.3, p238) following a brief description of other, mainly technical, implications arising from this research.

6.2.2 Other implications

Several other implications, largely of a technical nature, flow from this research.

These centre around:

- **Data availability** – In 2012 and 2013-14 DataSift was used (Section 4.2.5, p134) to collect and store Twitter and Facebook data. DataSift no longer has access to any Twitter data (Lunden, 2015) but can still access data from Facebook. Platform operators' policies, and growing regulatory imperatives to protect user privacy, may restrict such data availability in the future.
- **Computing infrastructure** – Although challenging (Appendix 8, p436) it is possible for a single researcher to store and augment reasonably large volumes (~8m records) of social media 'Big Data'. The data volumes and systems used here are not, however, fully representative of the difficulties to be expected when storing or analysing much larger data sets.
- **Computer software** – This research presents results collated using three NLP/geoparsing systems (Section 4.4.1, p147) but many more are available. While there is considerable agreement between two of the systems used to geoparse interaction message text (TwitIE on GATEcloud and CLAVIN-rest) similar results might not be found using alternative software packages.

These, and other, implications and observations are expanded upon in the concluding chapter (p286) of this thesis. The remainder of this chapter outlines policy recommendations stemming from this research, below, and details several other findings (Section 6.4, p246) relevant to the discussion.

6.3 Policy recommendations

6.3.1 Background

Ever more personal data are being deposited in the ‘corporate digital dossiers’ (Wyly, 2014) of giant Internet and Web-based businesses including, amongst others, Facebook, Google, Microsoft and Twitter. Not only have these corporations become increasingly dominant ‘but their online domination increasingly allows them to dictate terms’ particularly surrounding the dissemination of news (Tear & Southall, 2019, in press). While Google, Facebook and others, as Tear & Southall have noted, ‘do of course pass much of their income on to content providers, including editorially curated sources [...] Facebook in particular is not simply directing users to newspaper sites but providing a “News Feed” where traditional media compete with lower-cost providers; and Google’s YouTube is somewhat similar. Where the content is simply entertainment, providing a platform for individual “creators” promotes diversity. However the lowest cost way to produce “news” is to invent it.’

Invented, fake or alternative news, which spreads particularly quickly online and via social media (Vosoughi et al., 2018), is thought to pose a real threat to both political processes and civil society. Persily (2017, p63) has stated that ‘Whereas the stories of the last two [US Presidential Election] campaigns focused on the use of new tools, most of the 2016 story revolves around the online explosion of campaign-relevant communication from all corners of cyberspace. Fake news, social-media bots (automated accounts that can exist on all types of platforms), and propaganda from inside and outside the United States – alongside revolutionary uses of new media by the winning campaign – combined to upset established paradigms of how to run for president.’ Tackling these sorts of problems through policy will not be straightforward, particularly when other research (Müller & Schwarz, 2017) has shown how social media can ‘fan the flames of hate’ with ‘right-wing anti-refugee sentiment on Facebook [accurately] predict[ing] violent crimes against refugees in

otherwise similar [German] municipalities with higher social media usage [levels, suggesting] that social media can act as a propagation mechanism between online hate speech and real-life violent crime.'

Politicians and the massive global corporations running some of the Web's most popular sites, platforms and applications must tackle these problems. Promisingly, there are now signs that policy is developing in these areas. In the UK, Labour Party Leader Jeremy Corbyn has suggested that an 'internet tax' on tech companies could help create a 'public interest media fund' dedicated to investigative journalism (BBC News, 2018b). In the US, President Donald Trump has 'warn[ed] Google, Facebook and Twitter' about perceived political bias which, somewhat typically for the current Presidential incumbent, appears to centre around his own sensitivities towards stories from the 'Fake News Media' or 'Negative Left-Wing Media' appearing first in listings on these sites when searching online for 'Trump news' (BBC News, 2018g). As Taylor (2018) writing in *The Guardian* has pointed out, however, 'political heat from Trump and the left may signal reckoning ahead' for the Big Tech companies.

From an electoral perspective it appears almost certain, in the UK at least, that online political marketing and spending will come under increasing scrutiny and control. The Electoral Commission (2018a) report on *Digital campaigning*, subtitled *Increasing transparency for voters*, has made nine key recommendations. In abridged form these include:

1. Imprinting online material to say 'who is behind [a political] campaign and who created it';
2. Improving spending regulations to 'give more information about the money spent on digital campaigns';
3. Requiring campaigners to 'provide more detailed and meaningful invoices from their digital suppliers to improve transparency';

4. Suggesting social media companies ‘should work with’ the Electoral Commission to ‘improve their policies on campaign material and advertising for elections and referendums’;
5. Clearly labelling election and referendum adverts on social media platforms to ‘make the source clear’;
6. Clarifying that ‘spending on election or referendum campaigns by foreign organisations or individuals is not allowed’;
7. Improving ‘rules and deadlines for reporting spending’ both during and after election and referendum campaigns, and;
8. Increasing ‘the maximum fine [the Electoral Commission] can sanction campaigners for breaking the rules, and [strengthening] powers to obtain information outside of an investigation.’
9. Preventing ‘spending on election or referendum campaigns by foreign organisations or individuals.’

All of these are sensible recommendations, but results from this research suggest additional regulatory and/or technical responses are desirable. Recommendations in these areas are discussed, in turn, below.

6.3.2 Regulatory responses

The operations of major social media networks, including Facebook and Twitter, and other ‘Internet giants’, including Google, Instagram, Microsoft etc., are not currently regulated – outside normal company legislation – despite these corporations wielding increasing influence in our ‘always-on’ societies (Reich, 2018). While external regulation is not, yet, enforced by national governments most large technology companies have ‘self-regulated’; historically in response to user-privacy concerns (Burkell, Fortier, Wong, & Simpson, 2014) and, more recently, following a series of significant ‘data misuse’ scandals (Digital Culture Media and Sport Committee, 2018). These self-regulatory effects have been most visible at Facebook, which has progressively introduced much more granular user control

over privacy settings protecting personal data (excepting various transgressions, discussed below), but also extend to several other social media platforms and popular Internet sites or applications, e.g., Google, Instagram and Snapchat (Iosifidis & Wheeler, 2016; Muhammad, Dey, & Weerakkody, 2018). Most enhancements to user privacy settings are designed to restrict what content is visible when publicly-posted online or shared, increasingly ‘locking-down’ access to material. Others have been designed to restrict access to social graph inter-relationships via platform operator’s APIs; preventing, again with notable exceptions, extended traversal through social networks (Hogan, 2018). While user control over geodata, e.g., place or space-based posting locations or coordinate-geotags embedded within the EXIF metadata of GPS-encoded photographic images, has been enabled or enhanced through these developments self-regulatory policies surrounding the use of geographical information are generally not so opaque.

Geographical knowledge, and Location Based Services (LBS; Küpper, 2005), have done much to cement the utility of the World Wide Web. Knowing that a user is sitting in New York or London, e.g., enables Amazon to serve pages from country-specific sites (priced in USD and GBP, respectively) or encourage its user to visit the appropriately localised site to make purchases. Early LBS systems, often based around the MaxMind (2012b) GeoIP database, were rudimentary and somewhat error-prone (Shavitt & Zilberman, 2010) but have, with time and new technology, improved significantly; most evidently in the mobile arena. It is now possible to locate users, or their devices, based upon satellite, cellular and even indoor positioning using GPS, cell mast or WiFi data (Küpper, 2005). The developers of most web sites and mobile applications do their utmost to know where users are situated. Using LBS positioning techniques or HTML5 Geolocation (World Wide Web Consortium, 2018) and Google's (2018c) Reverse Geocoding API it is now easy, in code, to move from Latitude and Longitude coordinate pairs to an address. Hence, systems from Facebook (e.g., Figure 1-1, p5), Google, Twitter and others almost always know – with increasing spatial granularity over time – where their users are

located. This knowledge, regrettably for geographers and others (e.g., electoral regulators) with an interest in these subjects, does not translate into the ready availability of locational information alongside digital traces of online activity; platform privacy policies generally restrict access to geodata unless users have explicitly 'opted-in' to geolocation sharing (e.g., Twitter, 2014) and corporations are unwilling to release any more information than they have to.

Geolocation tracking by major Internet businesses does, however, now present a somewhat pernicious intrusion in daily life. The European Consumer Organisation (BEUC; Bureau Européen des Unions de Consommateurs in French) has recently filed seven complaints with national data protection authorities regarding Google's location tracking, its director Monique Goyens reportedly stating (Keane, 2018) that 'Google's data hunger is notorious but the scale with which it deceives its users to track and monetise their every move is breathtaking.' The BEUC report, titled *EVERY STEP YOU TAKE: How deceptive design lets Google track users 24/7*, notes that location tracking is pervasive, and cannot be switched off, on smartphones running the market-leading Android mobile device operating system developed by Google (Forbrukerrådet, 2018). Similar accusations have been levelled against Facebook whose various APIs, lobby group Privacy International have detected, when used on partner sites including Skyscanner and Duolingo, '[track] Android users even if they don't have a Facebook account' (Cuthbertson, 2018).

Geographical and locational knowledge are of great value to Internet businesses but are also of great value to society, government, regulators and researchers. It seems increasingly likely that moves to stem the worst excesses of the 'Wild West' identified in OSN advertising and data (mis)use, highlighted by the Facebook and Cambridge Analytica scandal, will be forthcoming (Charter, 2018). Facebook, as a result of its lax attitudes to data protection, has already received a £500,000 fine from the Information Commissioner's Office in the UK (BBC News, 2018d) with a much larger, £8.9m fine, levied by Italian data protection authorities (Embury-

Dennis, 2018). Law makers in the UK, US and EU are reportedly working up legislative responses to Web and social media misuses just as others, e.g., Reich (2018) have suggested that Facebook, Google, Apple and Amazon should be 'broken up' in anti-trust actions, as the original corporate 'robber barons [of the first] Gilded Age' in America were; a suggestion returned to in Section 7.7 (p310). Geographical data collected and used, but not shared, by the Internet giants in support or pursuit of their operational and 'advertising monetisation' strategies (N. Newman et al., 2016) is itself of immense value in tracing how content is consumed and disseminated online. Legislators and regulators, in addition to their existing responses, should also consider how access to geographical data might be improved, even if they were anonymised, aggregated or degraded. The following section offers suggestions in this area which might better enable the geographical tracking of political campaign material disseminated over the Internet.

6.3.3 Technical responses

Opinion, information and misinformation travel through 'cyberspace' via multiple channels. Very few of these can be tracked geographically, particularly shared communications made on social media networks. IP addresses, which are available to platform operators but redacted from publicly-accessible OSN data to protect user privacy, allow reasonably accurate locational estimations to be built from GeoIP databases (MaxMind, 2012b, 2012a). Many larger websites or applications also collect GPS-coordinates, or spatially-referenceable WiFi or cellular telephony mast information (Stackexchange, 2016), from smartphone-based users as they interact online.

The solution proposed here would:

- Store, alongside each message in all OSN interaction metadata for all users, a low resolution Latitude and Longitude coordinate pair or, alternatively, a lookup to, e.g., a 1x1km grid square. This would enable low-resolution

geographical tracking of all OSN content consumed or shared online and has several other advantages:

- Platform operators will already know the geolocation of most pageviews using server-side GeoIP lookups or client-side GPS or WiFi/cell mast data, and could degrade known coordinates to a lower resolution or allocate coordinates to a grid square identifier or the top/left coordinate of a bounding box etc.
- Personal locational privacy would not be affected as high-resolution coordinates would not be saved or imparted unless users, as now, opted-in to full coordinate-geotagging.
- Access to this low-resolution information, the degraded Latitude and Longitude coordinates or grid square lookup, could be restricted to accredited researchers or government agencies as, or if, applicable.
- Even if content originators ‘cloaked’ their whereabouts by using Virtual Private Networks (VPNs) or The Onion Ring (TOR) enabled anonymous routing and Web browsers, geographical patterns of consumption and sharing of this content by the vast majority of OSN users would reveal whether the content was intentionally geo-targeted, prompting additional and more detailed investigations into sources of origination as required.

Internet Protocol, version 6 (IPv6) provides increased capabilities for ‘including geolocation information in the headers of IPv6 packets (IPv6 GEO)’, the standard proposes optional storage of Latitude and Longitude pairs as both 16-bit integer and 32-bit fractions together with the storage of altitude (Skeen, 2017). Increased usage of this TCP/IP transmission protocol as IPv4 runs out of namespace – and/or World Wide Web Consortium (W3C) standards-based mechanisms to provide lookups to degraded coordinates or anonymised geographical areas or grid squares in IPv4 or IPv6 packets, rather than to full-resolution Latitude and Longitude coordinate pairs – would enable rudimentary geographical tracking of all Web or

OSN-based content consumption. This idea might raise privacy concerns (Groat, Dunlop, Marchanyy, & Tront, 2011) but, if the data were sufficiently anonymised and geographically sufficiently degraded, the benefits to democracy could be considerable.

As Section 6.2 (p229) has shown, during electoral campaigns potentially nefarious or inflammatory content is targeted not just at *individuals* but at *areas*; key ‘swing’ States or constituencies (Figure 1-6, p32; Figure 1-7, p33) whose results often dictate wider democratic outcomes. In many ways this targeting highlights the vulnerability of the democratic, and particularly ‘first-past-the-post’, voting systems used in the UK and several other countries historically influenced by Great Britain, including Canada, India, New Zealand and the United States of America. Diamond & Morlino (2004) have stated that ‘At a minimum, democracy requires: 1) universal, adult suffrage; 2) recurring, free, competitive, and fair elections; 3) more than one serious political party; and 4) alternative sources of information.’ While politicians have sought the popular vote ever since representative democracies were established, and changing media environments (e.g., television in the 1950s) have provided sometimes contentious platforms for doing so (Calhoun, 1992), new-found possibilities to more subtly manipulate electorates through social media advertising provide a particularly concerning development. As the Cambridge Analytica revelations have revealed, the material may be so well micro-targeted – on interests, behaviours and geography – that its existence may be extremely difficult to trace.

Knowing what has been seen, *roughly where*, would significantly improve campaign oversight. It might even enable platform operators, citizens, researchers or authorities to arrest social-media promulgated hate crime which, worryingly, appears to be the next stage in ‘advanced’ geo-behavioural targeting (Müller & Schwarz, 2017).

6.4 Additional findings

Pavalanathan & Eisenstein (2015, p2146) have noted that coordinate-geotagged 'Twitter data offers an invaluable resource for studying the interaction of language and geography, and is helping to usher in a new generation of location-aware language technology. This makes critical investigation of the nature of this data source particularly important.' The authors acknowledge (p2145) that '[several] papers draw similar conclusions, showing that the the distribution of geotagged tweets over the US population is not random, and that higher usage is correlated with urban areas, high income, more ethnic minorities, and more young people.' They also suggest that these results may have arisen from demographic and linguistic 'confounds and consequences' in coordinate-geotagged Twitter tweets and, by extension, other geotagged OSN data, which are biased towards these age groups and geographies.

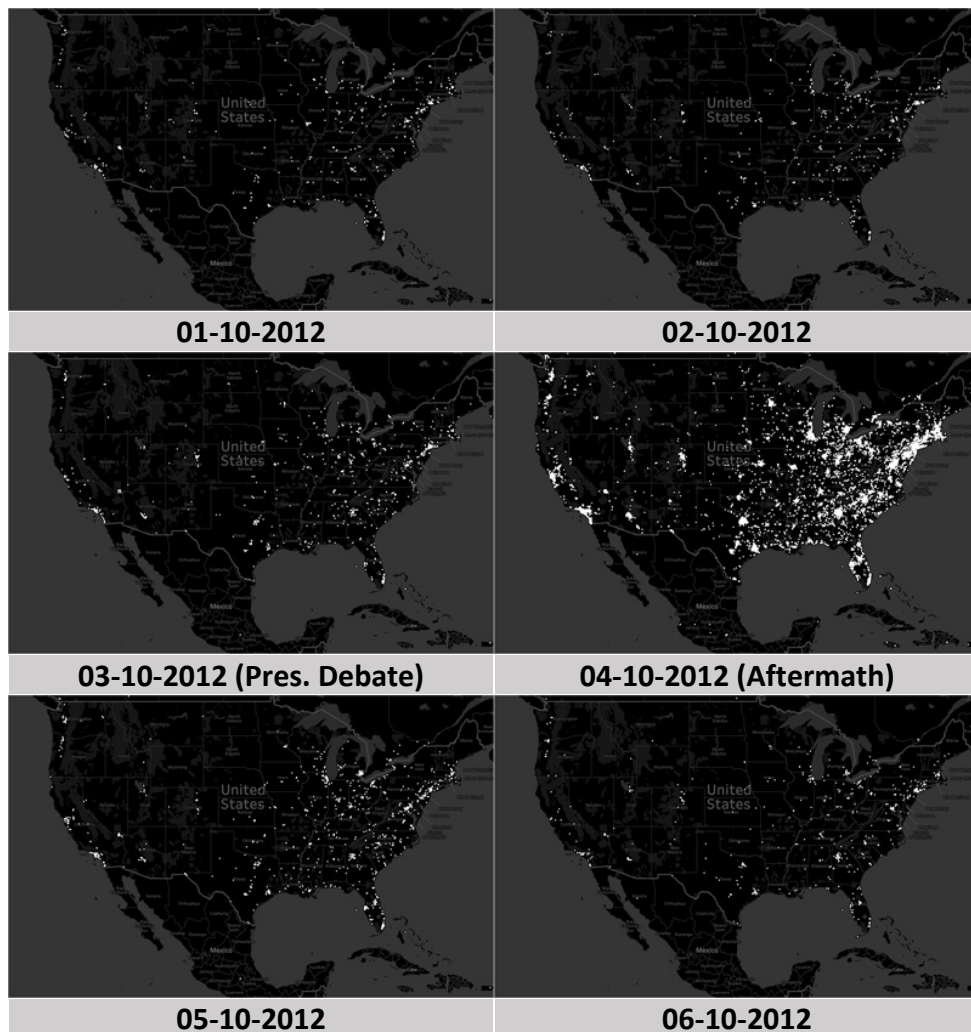
This section presents the results of several additional investigations conducted to identify these effects in `US2012` and `SCOT2014` case study data including spatiotemporality, geo-retweeting, data sparsity, OSN-Census data fusion, graph analysis and data skewness. As in the US, coordinate-geotagged OSN messages found within the boundaries of the UK exhibit several similar probable age and urban biases which suggests, as the results in Chapter 5 (p186) have already confirmed, that the representativeness of geotagged OSN data is limited and that this limitation (Section 6.2, p229) must be considered in future research.

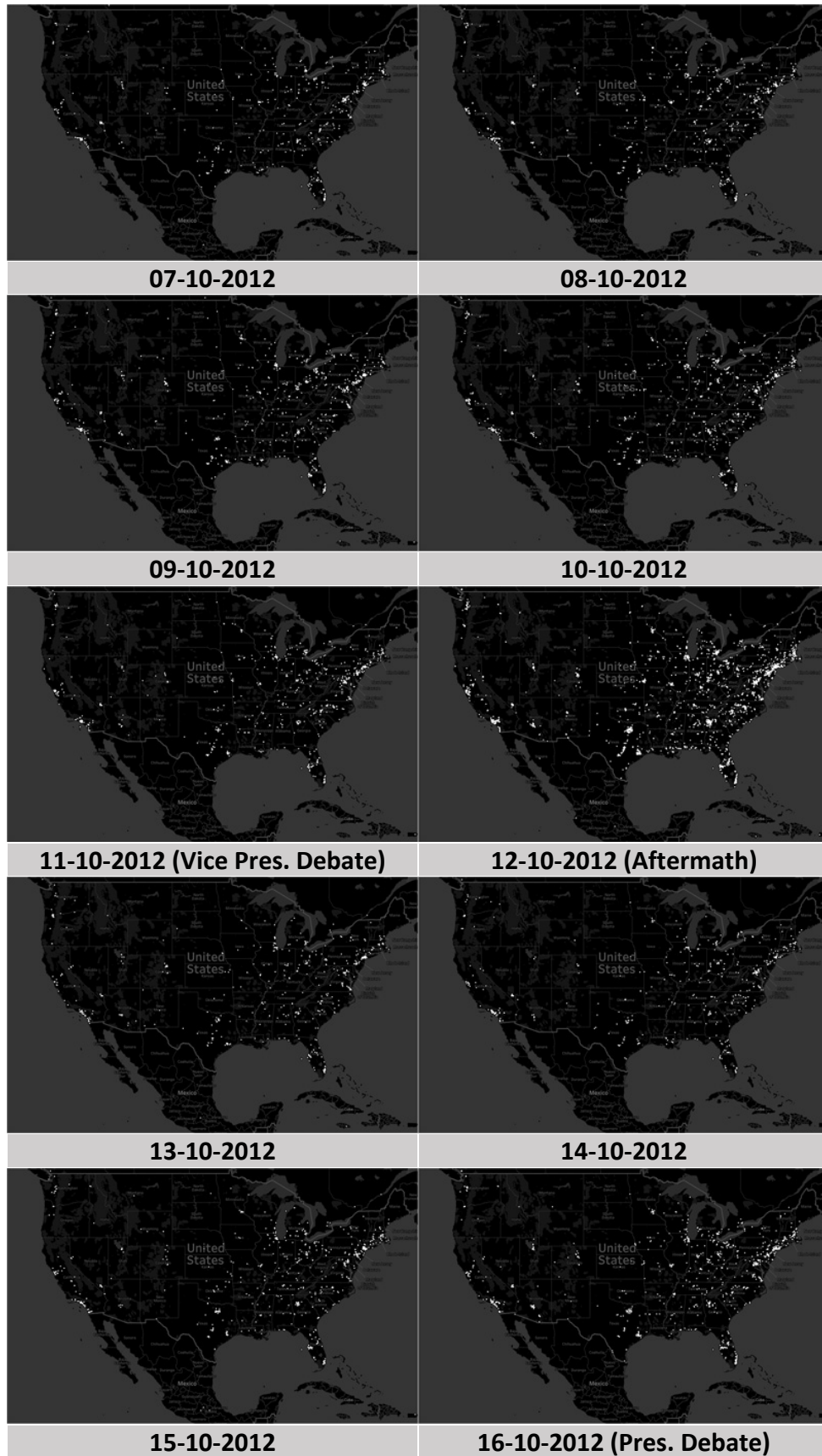
6.4.1 Spatiotemporality

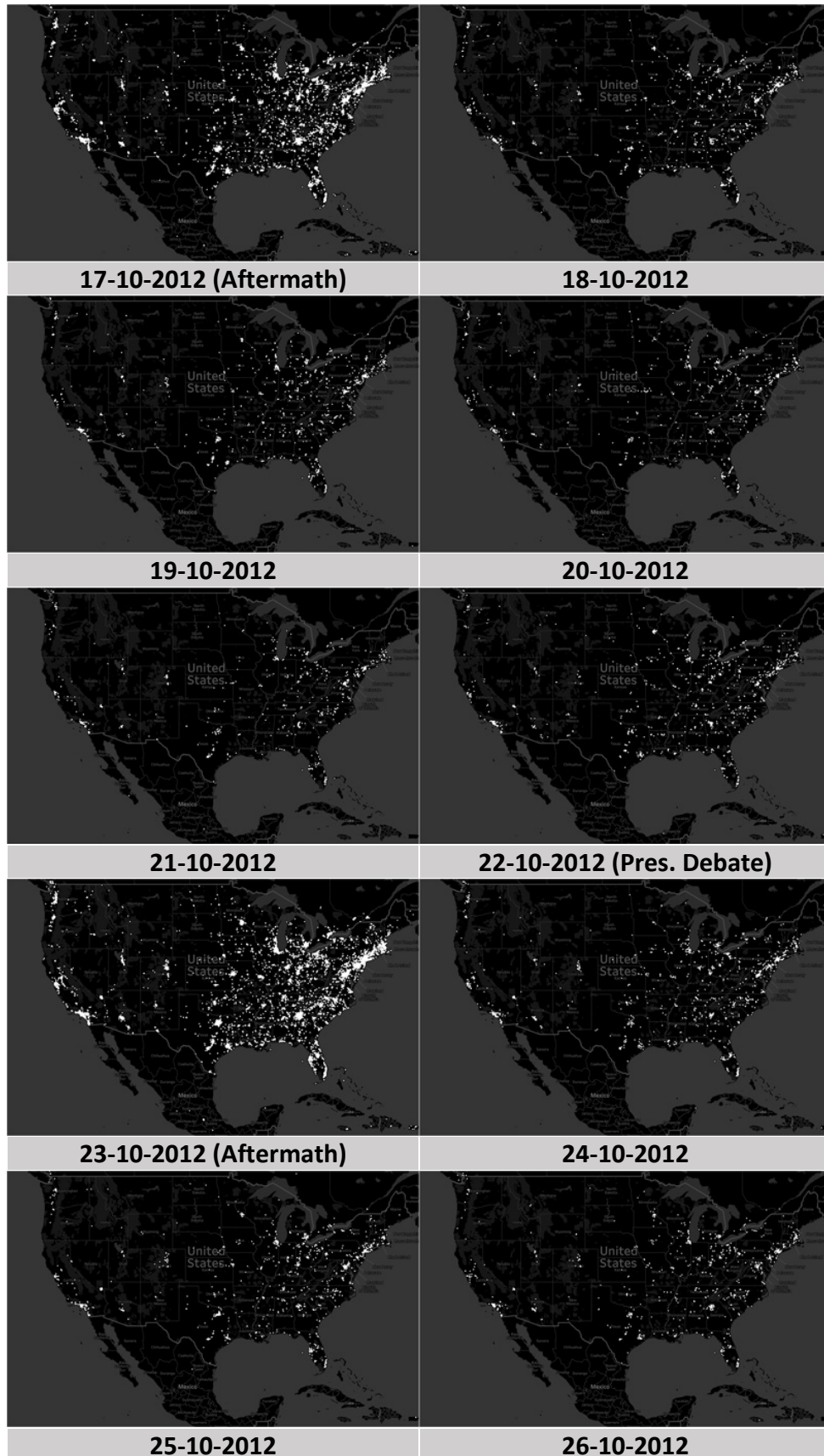
Coordinate-geotagged records may be mapped by day and visualised as an animation, using Tableau or other GIS software (Section 4.5.2, p161).

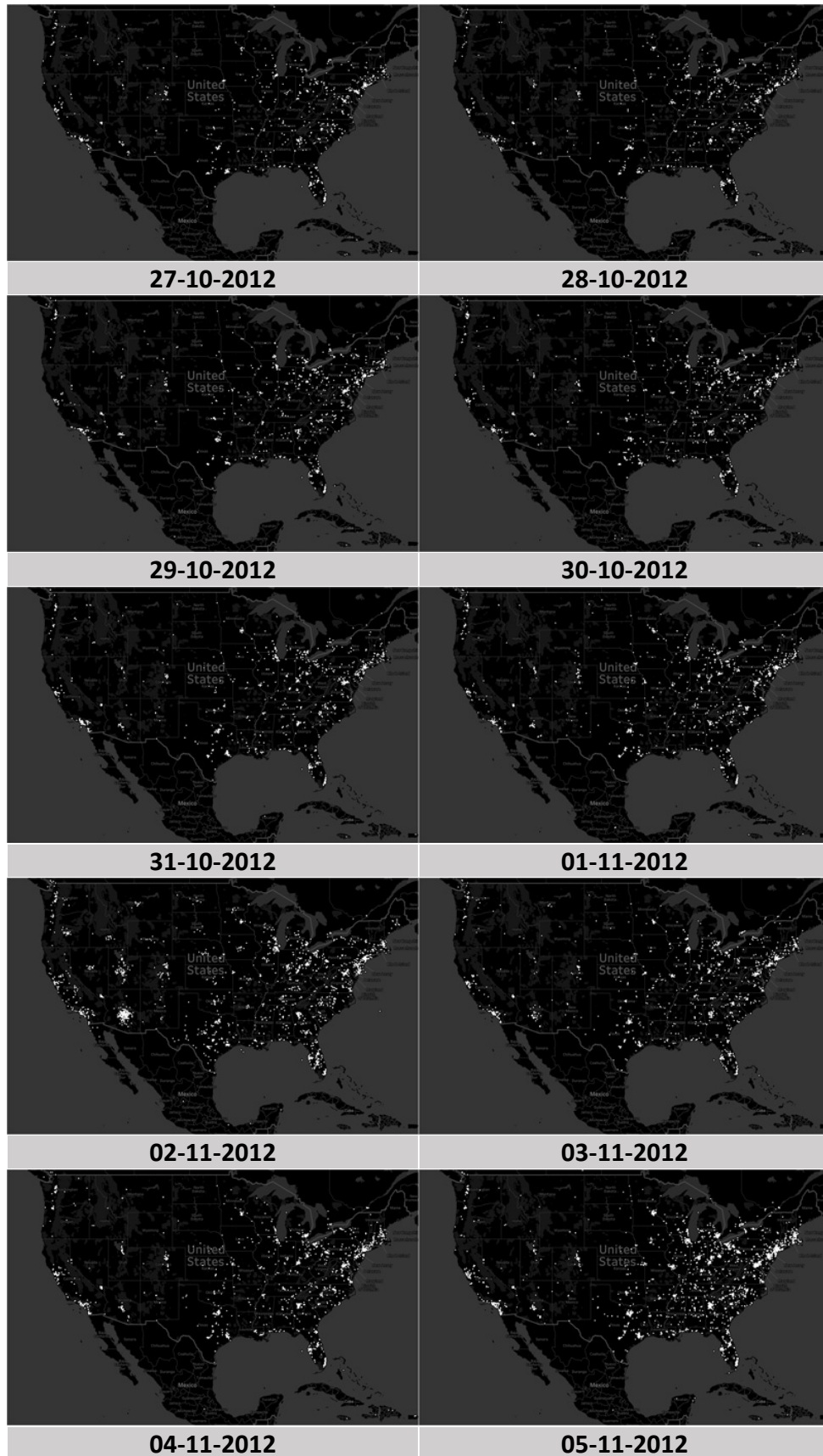
Spatiotemporal visualisations of this type work well on-screen (An et al., 2015) but are notoriously difficult to reproduce on paper; except, perhaps, as a 'flip book' or 'flick animation', as used in Scotese's (2004) 'cut-out-and-keep' paper-based

animation of Continental Drift. Figure 6-3 presents coordinate-geotagged OSN interactions from the US2012 data set in this way. Cross-referencing the graph of mentions of Presidential Candidate's surnames recorded by day (Figure 1-3, p25) with the map sequence shown in Figure 6-3 (below) it is clear that impacts from real-world events, such as the televised Presidential (or Vice Presidential) Candidate Debates of 3, (11), 16 and 22 October 2012 appear distinctly, both as peaks in the timeline and geographically on the map.









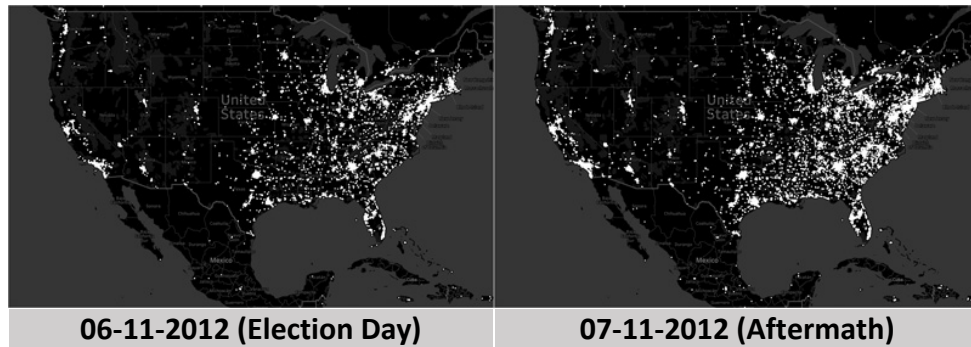


Figure 6-3 – US2012: Spatiotemporal patterns of activity in 160,934 coordinate-geotagged interactions by day through to election night

While the aftermath of each of the televised debates and the larger peak on and around election day are visible geographically as a ‘lighting up’ of the Eastern Seaboard of the United States, there are also many localised subtleties within the observed spatiotemporal record. Throughout the 2012 US Presidential Election campaign, and during the 2014 Scottish Independence Referendum, clusters of coordinate-geotagged OSN interactions sporadically appear, disappear or remain reasonably consistently in place; reflecting the observed spatiotemporal spread of coordinate-geotagging activity on OSNs in response to events such as political meetings or breaking local news stories occurring in given towns, cities or regions.

If all OSN interactions were coordinate-geotagged, or IP addresses or lower-resolution Latitude and Longitude coordinate pairs were made available alongside message text in OSN metadata (Section 6.3, p238), this type of data would provide a remarkable and comprehensive spatiotemporal resource. Unfortunately, as this thesis has demonstrated, only a small percentage of OSN interactions are coordinate-geotagged, and geotagging users are not representative of all users on either of the Twitter or Facebook social media platforms examined in this research.

6.4.2 Geo-retweeting

While the overall percentage of coordinate-geotagged Twitter retweets (at 1.25%, Table 4-8, p170) is in line with the low rates found in those OSN Streams sampled without the explicit need for geography (0.88%-1.45%, Table 4-1, p128 and Table

4-2, p132) many coordinate-geotagged retweets (n=102,343) reference, with different coordinates, Twitter tweets which were themselves originally coordinate-geotagged. These coordinate-geotagged retweets could only have been made by users ‘copying and pasting’ the body of the original Twitter interaction into a new tweet composition box before retweeting it with their own coordinates (Sloan & Morgan, 2015). Analysis of coordinate-geotagged retweets therefore provides an important strand of investigation, providing information about geographical dispersion of opinion, even if for only a small proportion of all retweeted interactions. There are 3,641,030 Twitter retweets (Table 4-6, p165) in the research data corpus. Of these, 466,043 (12.80%) can be ascribed to the originating tweet by linking Twitter ID columns using SQL (Appendix 11 listing 39, p491). Data-mining shows that 102,343 retweets have been coordinate-geotagged (Appendix 11 listing 40, p491) and 94,474 of these interactions can be linked back to the originating Twitter tweet (Appendix 11 listing 41, p491). Further SQL queries, designed to determine whether original and retweeted coordinate pairs differ (Appendix 11 listing 42, p491), detect 73,400 retweets of this type, originating from 14,546 distinct Twitter tweets (Appendix 11 listing 43, p492). The two sets of coordinates have been used to map dispersal effects, as shown in Figure 6-4 (p253). The median, average and maximum straight-line retweet distances across both events are, respectively, 2.72km, 17.22km and 2,223.26km. While the longer lines shown in Figure 6-4 emphasise long distance retweeting, the average retweet distance across the political two events is <18km and the median <4km. There are, however, some differences in these geographical straight-line measurements between the two events, as shown in Table 6-1.

Table 6-1 – US2012/SCOT2014: Number of geo-retweet coordinate pairs, median, average and maximum straight-line geo-retweet distances

Event	N pairs	Median	Average	Maximum
US2012	1,029	3.63km	26.30km	519.92km
SCOT2014	72,371	2.71km	17.10km	2,223.26km

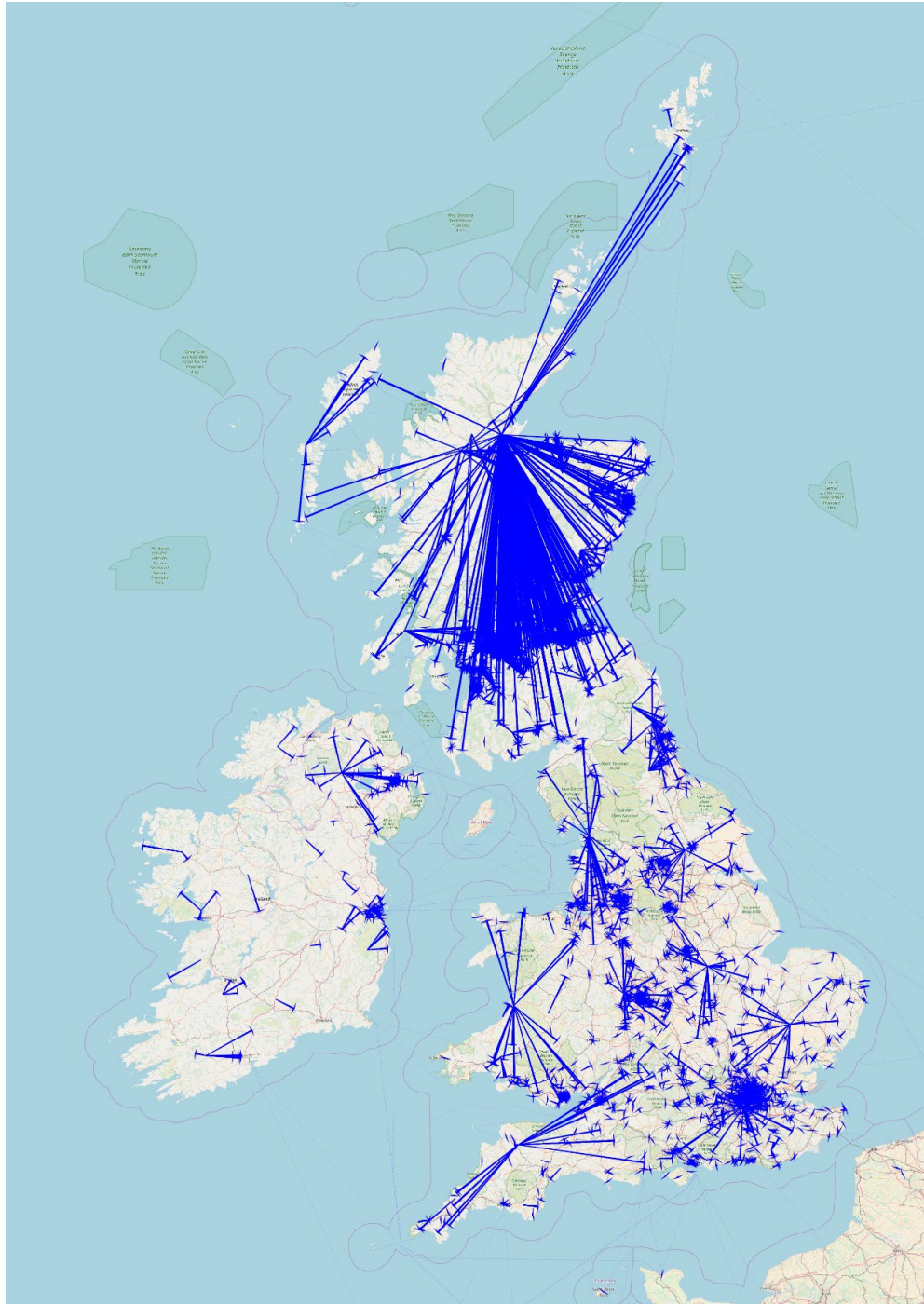


Figure 6-4 – US2012/SCOT2014: Geographical dispersal of Twitter retweets in the UK and Ireland (both electoral events)

These findings are at odds with the higher dispersion distances (1,698km median, 955km average) reported by van Liere (2010) using a much smaller sample of Twitter data (n=13,399 retweets) and the ‘749 statute miles’ reported by Leetaru et

al. (2013) analysing a much larger 10% sample of all Twitter tweets and retweets (n=1,535,929,521) made between 23 October and 30 November 2012, a time period which overlapped with US2012 data acquisition (Section 4.2.4.1, p126).

Many more dyads of original tweet and retweeted coordinates are observed in the SCOT2014 data set, which was captured using one continuous 1:1 sample (Appendix A7.3, p435) over a much longer interval than that used during the US2012 event, most of which were sampled in a 1:50 ratio leading to ‘misses’ between tweet origination and subsequent retweet. As well as demonstrating lower median and average straight-line geo-retweet distances, in line with Scotland’s smaller geographical extent (Table 6-1, p252), these results suggest that electoral events may foster a more ‘local’ pattern of communication in social media. When reviewing his findings, based on the 12-hour collection of all Twitter message text containing the ‘RT’ retweet identifier, van Liere (2010, p3) argues that ‘the [955km] average and [1,698km] median distance are too large to speak of local communication which suggest that the information broker pattern is the most appropriate pattern for this sample.’ This information sharing pattern, van Liere continues, ‘is based on following people with shared interests and not necessarily following friends.’ Results from the current research, analysing politically discursive material, suggests that the alternative ‘local communication’ pattern, which van Liere defines as ‘conversations [...] mainly between people who are friends in the off-line world’, is more prevalent during the two case study electoral campaigns.

The large number of tweets originating North of Invergordon in Highland Scotland, and coordinate-retweeted throughout Scotland, stem from one particularly prolific coordinate-geotagging social media user during the 2014 Scottish Independence Referendum, Mulder1981. This Twitter user, a Scottish Tory Councillor and ‘influential BritNat Twitter troll’ (The Herald, 2017), posted from 2,503 locations; largely in Scotland but from as far afield as Turkey and the West Coast of America. Coordinate-geotagged retweets are considered extremely valuable in analysing

geographical dispersal of content and have been coded with a ‘High’ score (200) when categorising PGI metadata fields (Table 4-10, p174).

6.4.3 Data sparsity

Data held in tabular rows and columns (or ‘fields’) in an RDBMS, such as Oracle 12c used here (Section 4.3.1.3, p145), must store empty cells, cells containing no data, as `NULL` values. In this respect, tabular data storage is less efficient than JSON-based data storage, in which only the values of non-null fields (or JSON ‘keys’) are recorded (ECMA International, 2013, 2017). While less efficient, the storage of `NULL`s in RDBMSs enables straightforward computation of row-level sparsity statistics across tables. The following charts (Figure 6-5, Figure 6-6 and Figure 6-7 starting on p257) show row level sparsity by field, or column name, across the research data corpus stored in the main `INTERACTIONS` table, further broken down by Stream (Appendix 7, p432). The charts are ordered by descending sparsity level (% null records), then by column name. Column names appearing at the top of Figure 6-5 are fully or well-populated with values in all rows, and become progressively more highly-populated by nulls in ensuing charts.

In order to fit within the 30-character limit for American National Standards Institute (ANSI) SQL column names, which Oracle (2018a) adheres to, long JSON key names or CSV header names have been shortened and consistently formatted. The abbreviations shown in Table 6-2 have been used. In the `INTERACTIONS` table, the dot notation of the original lower-case JSON key name `twitter.in.reply.to.screen.name` (also used in the header row of the DataSift CSV files) becomes `TW_IN_RE_TO_SCREEN_NAME` after dots are replaced with underscores and the abbreviations above are applied. Sparsity counts have been calculated using a PL/SQL programme detailed in Appendix 9 (p443). This program loops over 149 columns in the `INTERACTIONS` table for each of the 4 Streams recorded. Counts of `NULL` or zero length values are made using SQL queries and results stored, before these are transposed (Appendix A9.4, p449) and

used in Tableau (Section 4.5.2, p161) to create the graphical output shown in the three figures below.

Table 6-2 – Abbreviations used in Column Names of the INTERACTIONS table

Abbreviation	Meaning
ATT	Attributes
FB	Facebook
RE	Reply
RT	Retweet
RTED	Retweeted
ST	Street
TW	Twitter

Only 10 of the 146 fields, or column names, present in the `INTERACTIONS` table (6.85% of all fields) are fully populated, with no `NULL` values in any rows. The distribution of highly populated fields in each Stream varies. Some fields, e.g., `INTERACTION_GEO_LATITUDE` and `INTERACTION_GEO_LONGITUDE` in the `US2012_GEO` Stream (Figure 6-7, p259) are highly populated as a result of sample design; this Stream was designed to record *only* coordinate-geotagged interactions (Appendix A7.2.1, p433). Elsewhere there is general uniformity in sparsity, except where some fields (e.g., `INTERACTION_TAGS`, `DEMOGRAPHIC_GENDER`) appear highly populated in the three `US2012` Streams but are not well-populated in the `SCOT2014` Stream. There is no obvious explanation for these discrepancies, although they probably arise from changes in the operation of the DataSift platform or its access to underlying social media feeds during the two-year interim between recordings.

The `US2012_GEO` Stream, filtered on the presence of geographic coordinates (Appendix A7.2.1, p433), exhibits clear differences to the other Streams. The columns `FB_AUTHOR_AVATAR` and `FB_AUTHOR_ID`, present in 10.28% of records overall, are wholly absent and no OSN interactions from Facebook have been sampled.

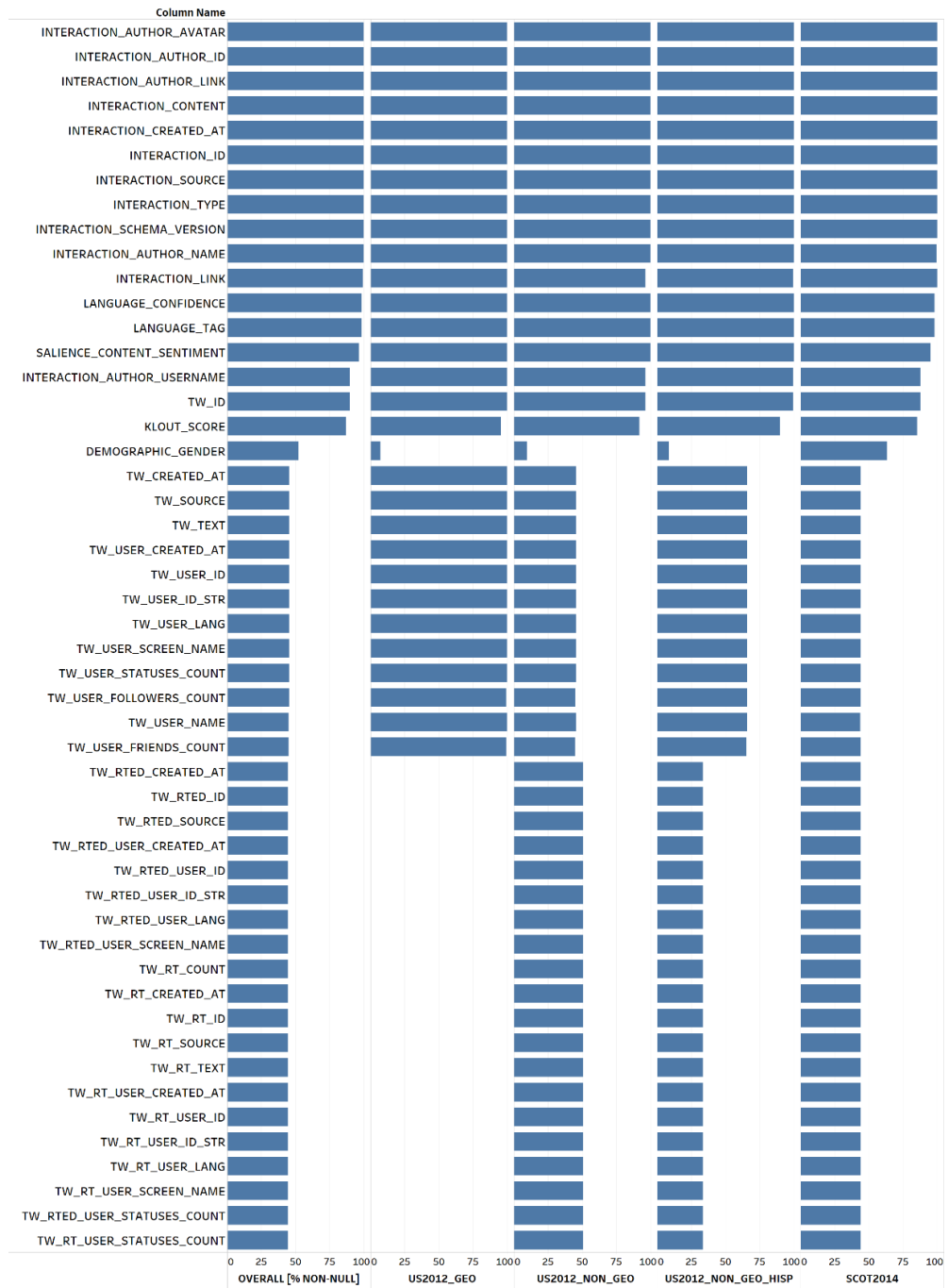


Figure 6-5 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (first 50 columns)

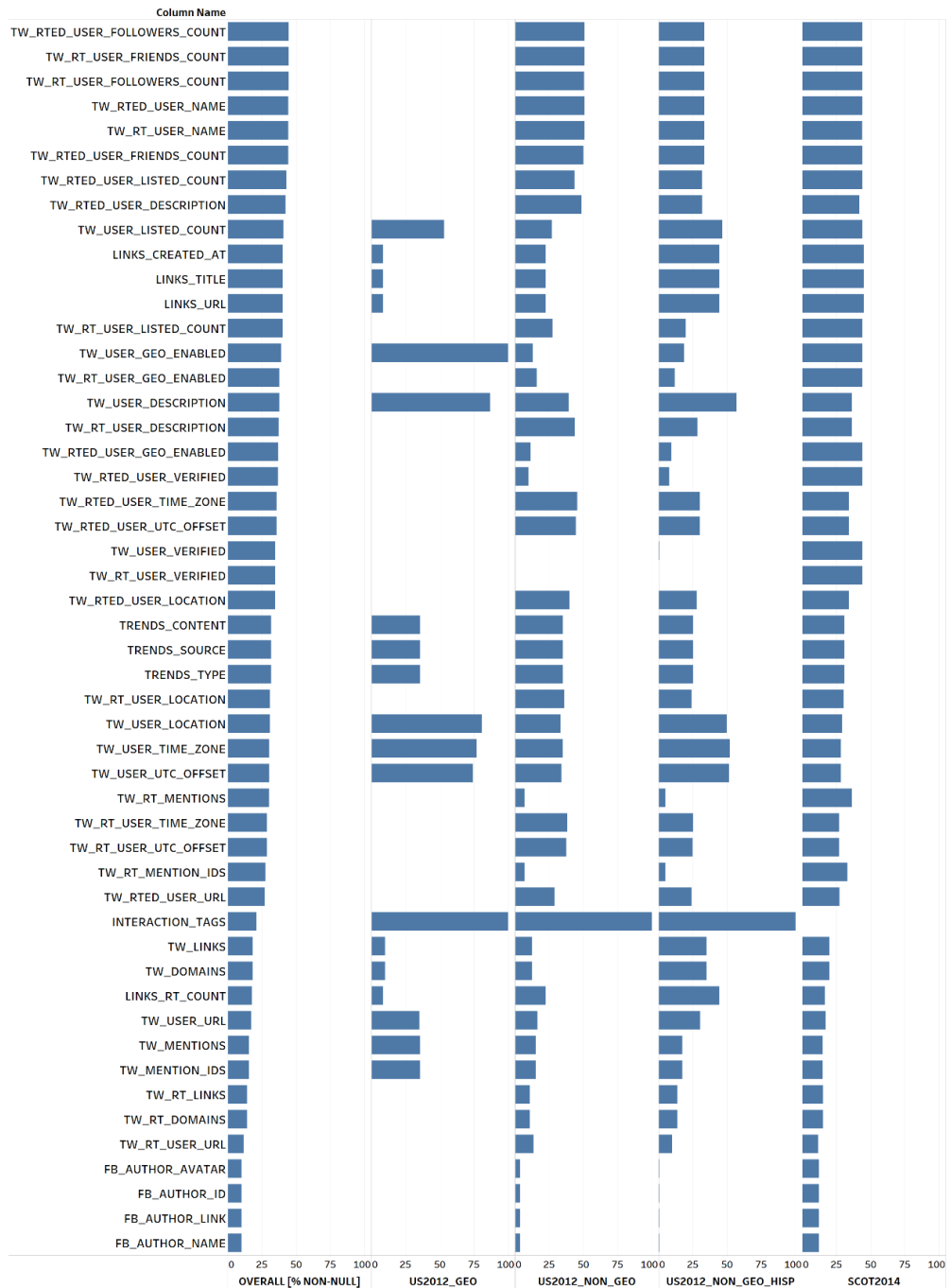


Figure 6-6 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (next 50 columns)

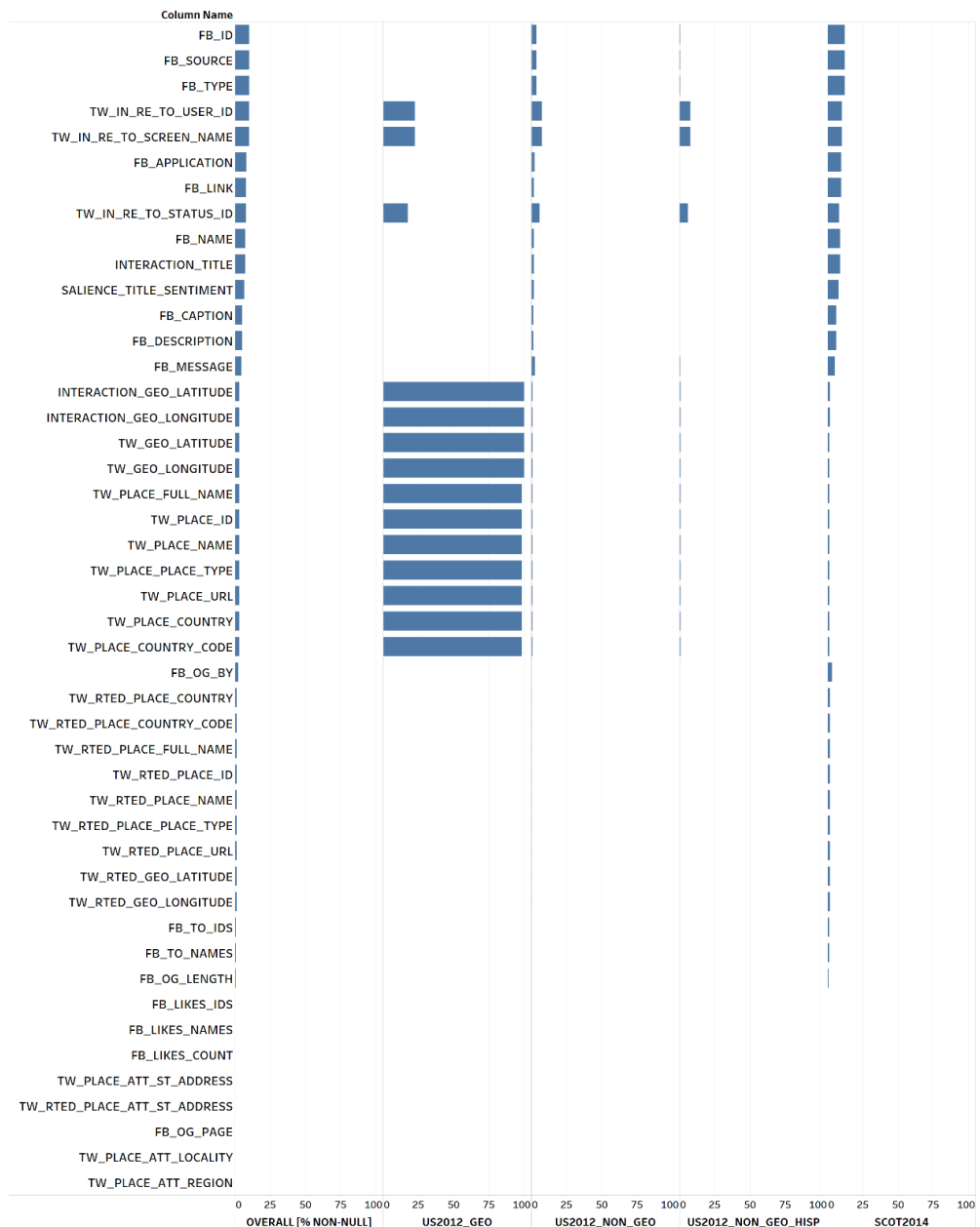


Figure 6-7 – Row-level sparsity (% non-null) by field/column names; overall and by Stream (last 49 columns)

Differences in the field sparsity of the `US2012_GEO` Stream are an unintended consequence of the restrictive CSDL condition `interaction.geo exists` used to record these interactions. No coordinate-geotagged Facebook posts were sampled during the `US2012` event. Later, 1,231 coordinate-geotagged Facebook posts were sampled (Table 4-8, p170) by the `SCOT2014` Stream without this

condition, suggesting that DataSift's access to Facebook data may have changed in the interim.

Table 6-3 – Top 10 fully populated fields/column names, commentary and utility

Column Name	Utility
INTERACTION_AUTHOR_AVATAR	Low
<i>URL to an image used as the author's avatar. Low utility as the image may not be a true likeness of the user, or even the same sex/ethnicity.</i>	
INTERACTION_AUTHOR_ID	Low
<i>DataSift's unique identifier for the author. Low utility except when counting or grouping messages by author.</i>	
INTERACTION_AUTHOR_LINK	Medium
<i>URL to the author's Twitter/Facebook home page. Medium utility in this study, potentially higher utility if the link is followed and social graphs are mined.</i>	
INTERACTION_CONTENT	High
<i>Text content of the Twitter Tweet or Facebook Post recorded. High utility variable length message text, ranging from 3 to 82,478 characters; median length 127.</i>	
INTERACTION_CREATED_AT	High
<i>UTC date/time stamp of interaction creation. High utility date/time stamp allowing temporal (and/or spatiotemporal) analysis of message flow.</i>	
INTERACTION_ID	Low
<i>DataSift's unique identifier for the interaction. Low utility as the identifier is unique to the message, which should itself be unique.</i>	
INTERACTION_SOURCE	Medium
<i>Source of the interaction. Medium utility as the top 20 sources account for 85.19% of all messages, and 65.66% of these are made using mobile phone applications.</i>	
INTERACTION_TYPE	Low
<i>Type of interaction, either Twitter (7,353,878 interactions, 89.72% Overall) or Facebook (842,502 interactions, 10.28% Overall). Low utility.</i>	
INTERACTION_SCHEMA_VERSION	Low
<i>DataSift's schema version identifier. Low utility as all interactions bar 10 (which are null) have Schema Version = 3.</i>	
INTERACTION_AUTHOR_NAME	Low
<i>Author's name (may be real, no way of knowing) as opposed to their username (@POTUS etc.). Low utility, usage identifies users, raising ethical issues.</i>	

The ten fields, or column names, fully-populated across all Streams are shown in Table 6-3, together with a commentary on field content and an assessment of analytical utility. Examination of the data shows that many of these fields have little practical analytical utility, with six of the top ten fully-populated fields containing

low utility data. Two of the fully-populated fields are of medium, and two of high, utility. A guide of this type has not been found elsewhere in the academic literature and should prove useful to subsequent researchers.

A further seven fields are generally well populated both overall (86.9 – 99.3% non-null) and across the four Streams. These fields, or column names, are shown in Table 6-4, again with commentary and an assessment of analytical utility.

Table 6-4 – Next 7 highly populated fields/column names, commentary and utility

Column Name	Utility
INTERACTION_LINK	Low
<i>URL to the interaction recorded. Low utility as all the data on the page is already recorded alongside the interaction in the database.</i>	
LANGUAGE_CONFIDENCE	Low
<i>DataSift's 0-100 confidence score for language detection. Low utility in this study, where the English language is used in 95.44% of messages.</i>	
LANGUAGE_TAG	Low
<i>DataSift's language tag (EN=English; ES=Spanish etc.). Low utility as only 3.12% of interactions are not in English. Spanish and Portuguese are 2nd and 3rd most used.</i>	
SALIENCE_CONTENT_SENTIMENT	Medium
<i>DataSift's sentiment score, ranging -38 to +45, median value 3 for non-zero records. Medium utility for sentiment analysis (using a 'black box' approach).</i>	
INTERACTION_AUTHOR_USERNAME	Low
<i>Author's username (Twitter or Facebook username, e.g., @POTUS). Low utility, usage identifies users, raising ethical issues.</i>	
TW_ID	Low
<i>Twitter's unique identifier for an interaction. Low utility as other columns provide links back to the same information. Can be shared with others to rebuild corpus.</i>	
KLOUT_SCORE	High
<i>Index of Social Media Impact, ranging in this data set 10 through 99, median 41. High utility in ranking users' impact. Media firms have the highest Klout scores.</i>	

The tail-off in row-level completeness (Figure 6-5, p257) amongst the remaining 136 fields stored across the Streams in the database is particularly noticeable. After the heavily populated fields described in Table 6-3 (top 10) and Table 6-4 (next 7), row-level sparsity increases markedly as the percentage of NULL values present in rows within each field increases. Values for some fields, such as the potentially

useful indicator `DEMOGRAPHIC_GENDER`, are present in 52.11% of records overall, most from the `SCOT2014` Stream with much lower percentages in each of the three `US2012` Streams. The least populated field, `TW_PLACE_ATT_REGION`, stores values in only 6 of 8,196,380 rows and therefore has next to no practical utility. The data are ‘Big’, but far from complete. This characteristic separates social media data from other types of population or administrative data often used in social science research and is not always explicitly stated in the literature. Fusion of coordinate-geotagged social media data to these more richly-populated data sets, to produce population profiles from 2010 US and 2011 UK Census data, in line with recommendations from Crampton et al. (2013) and Fuchs (2017a), has been conducted to partially redress this data sparsity problem and is reported upon below.

6.4.4 Data fusion

The previous section described the large, if sparse, nature of the OSN data sets examined in this research. There are no age or sex data for any users, although a DataSift-derived and text-content-inferred gender flag of questionable quality is present featuring the same ‘male’, ‘female’, ‘mostly male’, ‘mostly female’, ‘unisex’ etc. categories found in the 2012 French Presidential Election technical proof of concept exercise (Figure A6-3, p429). Ethnicity, marital status, educational attainment level and other demographic data commonly used as controls in social science research are wholly absent, prompting Mislove et al. (2011, p554) to observe that ‘despite the enormous potential presented by this remarkable data source, we still do not have an understanding of the Twitter population itself.’ Mellon & Prosser (2017) have highlighted similar problems in data publicly-available from Facebook, the other major OSN data source used in this research.

Following one of Crampton et al.’s (2013) recommendations to move ‘beyond the geotag’, by *fusing* OSN data to public (e.g., census) data sets, a suggestion also echoed by Fuchs (2017a), this section details the work undertaken to produce

population profiles of coordinate-geotagging OSN users' likely areas of origin. The analyses are a best estimate, dependent upon the assumption that all geotagged coordinates represent home locations, which I. L. Johnson et al. (2016) and this research partially discount, but does nevertheless provide some demographic information about the areas, both in the US and the UK, from which coordinate-geotagging OSN users may well have created their social media messages. Several of the results reported here also usefully corroborate findings from other studies examining social media demographics in these two countries (Blank & Lutz, 2017; Longley & Adnan, 2016; Longley, Adnan, & Lansley, 2015; Sloan et al., 2015). In the US, Mislove et al. (2011, p555) have reported that 'Twitter users significantly overrepresent the densely [populated] regions of the U.S., are predominantly male, and represent a highly non-random sample of the overall race/ethnicity distribution.' In the UK, Mellon & Prosser (2017, p1) have found that 'On average social media users are [found to be] younger and better educated than non-users, and they are more liberal and pay more attention to politics', mainly as a result of their demographic composition. Social media users in the UK also appear to share similar age and education characteristics with their US counterparts, Mellon & Prosser (2017, p1) citing several surveys which show that US 'Facebook and Twitter users tend to be younger and more educated than the general population, with Twitter having a more skewed distribution.'

Geodemographic techniques have been widely-used, particularly in marketing and market analysis, to segment customer groups (Voas & Williamson, 2001) ever since Richard Webber's work led to the creation of ACORN (A Classification of Residential Neighbourhoods) in the 1970s (McElhatton, 2004). Leventhal (2016), citing Sleight (2004), defines geodemographics 'as the analysis of people by where they live' proceeding to identify the two principles that underpin the methodology, '1) [that] two people living in the same neighbourhood are more likely to have similar characteristics than two people chosen at random' and '2) [that] neighbourhoods can be categorized according to the characteristics of their residents; two

neighbourhoods belonging to the same category are likely to contain similar types of people, even though they may be geographically far apart.’ ACORN, and later geodemographic classifications such as MOSAIC (Experian, 2018; Webber, 2004), categorise small areas (US Census Blocks, UK Census Output Areas or sets of zip codes or postcodes) into typologies (e.g., MOSAIC’s ‘A01 World-Class Wealth’ and ‘B08 Bank of Mum and Dad’ types) based upon public Census counts (e.g., age composition, number of cars in households) and private data sources, including electoral roll, credit checking and/or consumer spending data which together are often used as proxies for wealth. These geodemographic typologies, or ‘discriminators’, are commercial products and are not generally available to academic researchers. Typically, they also rely upon zip or postcode-level matching to ‘fuse’ records to classification types. Neither US zip codes nor UK postcodes are present in the Facebook and Twitter interactions which comprise the research data corpus. A GIScience-based approach provides the solution to this problem, using GIS software (MapInfo Professional 8.0) to allocate (‘point-in-polygon’) coordinate-geotagged OSN interactions to publicly-available 2010 US Census Tract and UK 2011 Census Output Area boundaries. Lerman et al (2017, p210) have used an identical approach and report that their findings ‘highlight the role of [...] demographic factors in online interactions and demonstrate the value of traditional social science sources, like US Census data, within social media studies.’

Barr (1996) has stated that ‘The US and UK censuses have many similarities’ as well as some ‘instructive’ differences, mainly based upon their ‘constitutional basis [and] the way [in which] they are administered.’ The US Census, for example, ‘is taken, and used, for electoral re-districting to a greater extent than [is the case in] the UK’ and has historically been freely accessible from government whereas the UK Census has not. Since the 2011 Census, however, the UK government has opened access to counts and accompanying digital mapping files so that US and UK data availability are now comparable, even if differences in age breaks etc. used in data collection or reporting remain. Results from the fusion exercise, linking US2012 and SCOT2014

coordinate-geotagged interactions to 2010 US Census and 2011 UK Census data, are presented below.

6.4.4.1 US Census / US2012 data fusion

In 2010, the US Census Bureau (2012a) defined 74,002 Census Tracts with an average population of 4,222 persons per Tract (ranging 1-37,452, median 3,993) built upon 11,155,486 much smaller Census Blocks. US Census Tracts are larger in terms of population and, in many cases, areal extent than the Output Areas (OAs) used to build the UK Census (Section 6.4.4.2, p272). US Census Tract boundaries were chosen over Census Blocks for point-in-polygon intersection as these are available nationally with 'Selected Demographic and Economic Data' (US Census Bureau, 2012b) whereas the Block-level data set is not, and is only available for download State by State, requiring significant post-processing. Of 168,873 non-0/0 Latitude/Longitude coordinate-geotagged interactions (n Facebook posts=0, n Twitter tweets= 160,837, n Twitter retweets= 8,036) in the US2012 data set (Table 4-8, p170) 151,567 records (89.75%) could be allocated to 38,645 distinct, or 52.22% of all, US Census Tracts. The remaining 17,306 coordinate-geotagged records fell outside US boundaries.

In pseudo-code the GIScience steps involve:

- Intersecting each of the OSN points with Census Tract polygons;
- Attaching Census counts for the underlying polygon to each point, and;
- Saving and exporting the result for further summation.

As each interaction picks up the underlying counts from US Census Tracts, the counts for 'OSN Tracts' containing n interactions appear n times in the 151,567 record output, i.e., a Tract containing 4 geocoded interactions will have its Census counts repeated 4 times. While the totals for OSN Tracts (Figure 6-10, p268 onwards) are therefore *higher in total* than totals for the US population the calculation of *percentages* for each Census count against this weighted figure allows

comparison with percentages also calculated for each count against the US overall. The ‘% OSN Tracts’ figure in each chart shown in Figure 6-10 (p268) to Figure 6-14 (p271) is compared to the percentage for each count (e.g., Males or Females aged 20-24 years old) for the US as a whole (‘% United States’) and indexed against the latter to show under/over representation.

Using individuals’ locations to create summary areal reports in this way does raise some dangers of stumbling into the ‘ecological fallacy’ identified many years ago by W. S. Robinson (1950) and Selvin (1958). W. S. Robinson (1950, p357) warned that ‘ecological correlations [cannot] validly be used as *substitutes* for individual correlations’ even when re-weighting takes place. However, in the absence of *any* demographic data from Facebook and Twitter the method used here enables comparison of areas containing coordinate-geotagging OSN users against national percentage bases. These profiles should not be used to suggest that all coordinate-geotagging OSN users in areas share a common demographic profile, or that coordinate-geotagging OSN users in any given area are generally representative of the population of that area. The analysis is only possible for those users recording Latitude and Longitude coordinates alongside their OSN interactions, and this percentage is a) a small one, and; b) is somewhat unrepresentative of OSN users overall, as this thesis has demonstrated (Chapter 5, p186). However, over one half of all US Census Tracts contain coordinate-geotagged interactions and, as Figure 6-8 (p267) and Figure 6-9 (p267) show, there is a significant overlap between areas of high population density and high OSN posting density, a result that has also been reported elsewhere (Mislove et al., 2011).

The maps in Figure 6-8 (p267) and Figure 6-9 (p267), produced using QGIS (2018), show a clear relationship between US population density and coordinate-geotagged OSN interaction locations. Most explicitly geotagged Twitter tweets or retweets, and the few geotagged Facebook posts in the research data corpus, are made in the most densely populated parts of the United States.

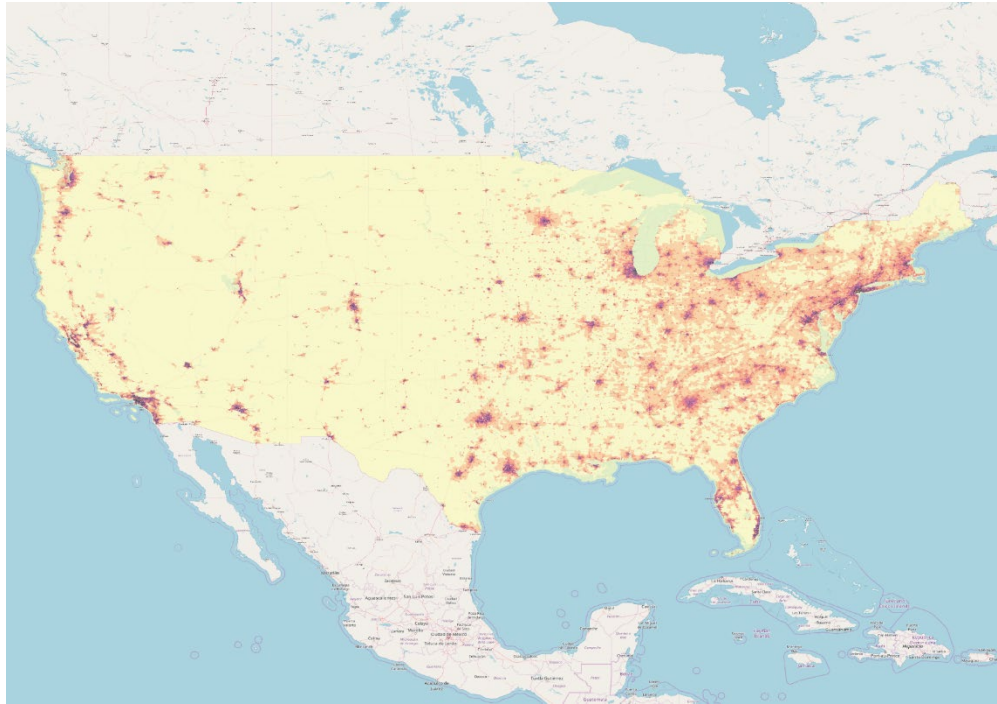


Figure 6-8 – 2010 Contiguous US population density at Census Tract level (light=low; dark=high)

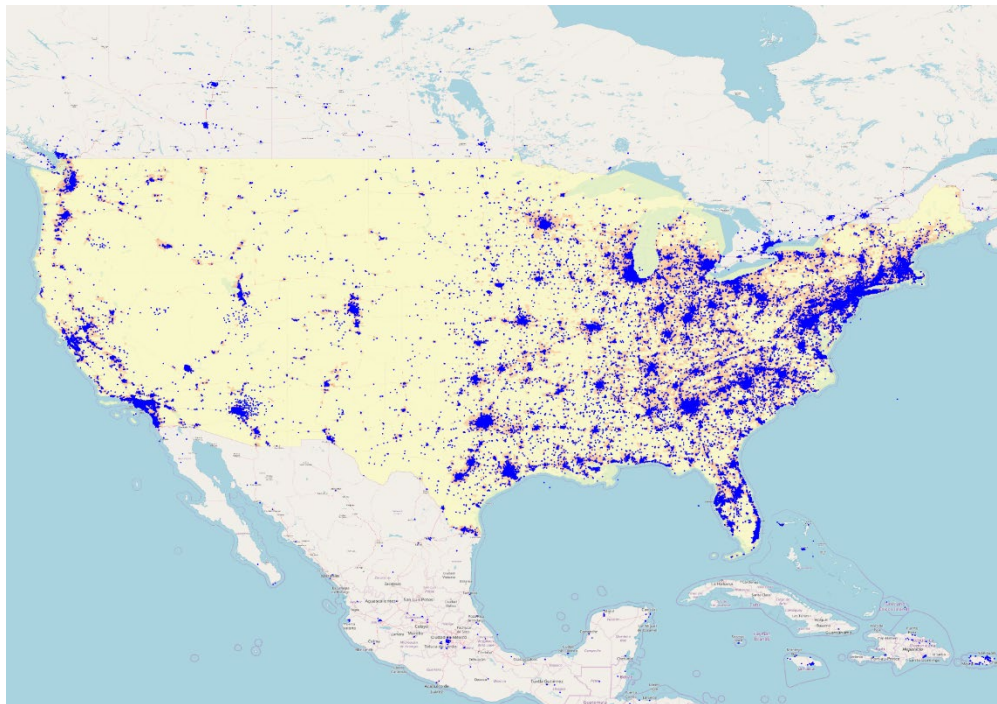


Figure 6-9 – 2010 Contiguous US population density at Census Tract level and US2012 coordinate-geotagged OSN interactions

Similar relationships between population density and OSN coordinate-geotagged locations have been reported in Germany (Hahmann, Purves, & Burghardt, 2014), the UK (Steiger, Westerholt, Resch, & Zipf, 2015), and several cities worldwide, leading Jiang, Ma, Yin, & Sandberg (2016, p349) to ‘conjecture that the spatial distributions of tweets [...] quite accurately reflect those of urban populations.’ As Y. Liu et al. (2015, p517) have suggested, these geographical distributions suggest that in ‘urban fringe areas or rural areas’ communities appear under-represented on social media platforms.

When comparing ‘% OSN Tracts’ with ‘% United States’ (Figure 6-10), populations in younger age groups (20-24 years old) are quite significantly (index=137, i.e. 37%) over-represented. Other age groups, such as those ‘Under 5 years’ old will, of course, hopefully not be posting online at all using either Twitter or Facebook, whatever the profile suggests about the areal demographic composition of Census Tracts containing OSN users’ coordinate-geotagged interactions.

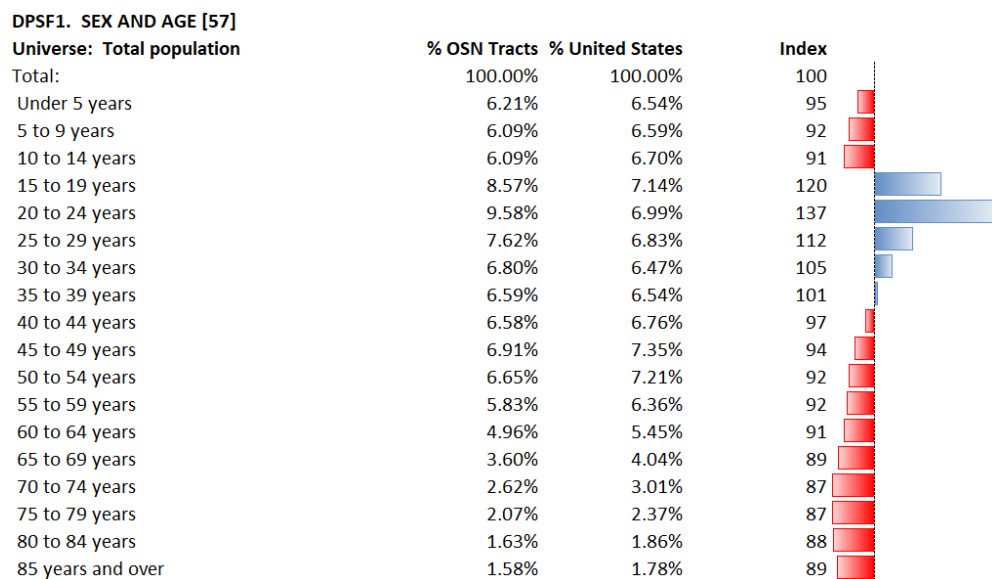


Figure 6-10 – US2012: Total population by age (OSN Tracts Indexed against US)

Breaking down the geodemographic analysis further, by gender, OSN Tracts are somewhat over-represented by younger, male age groups when compared to the US base, again in the age range 20-24 (index=136, Figure 6-11, p269).

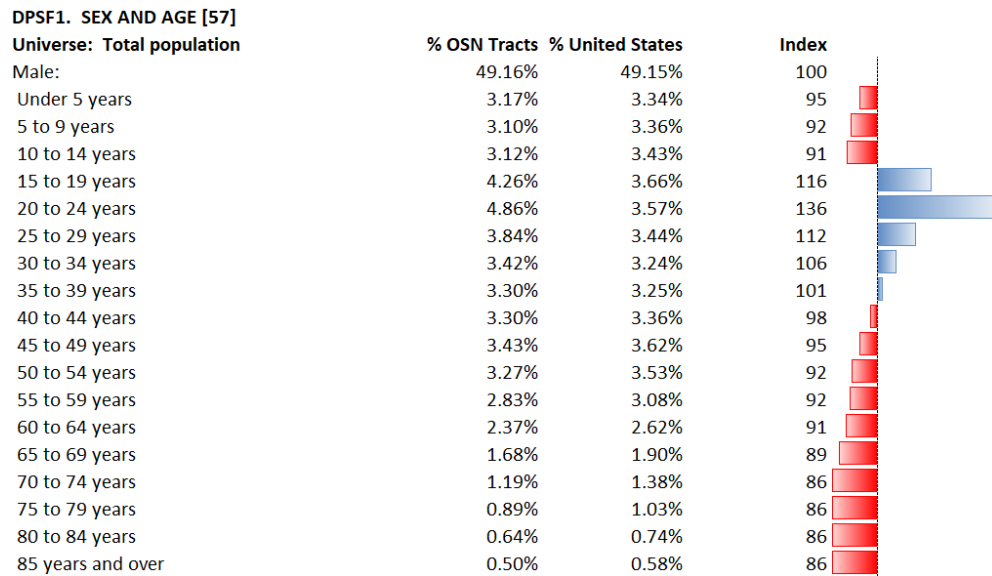


Figure 6-11 – US2012: Male population by age (OSN Tracts indexed against US)

Likewise, OSN Tracts are also somewhat over-represented by younger, female age groups when compared to the US base, again particularly in the age range 20-24 (index=138, Figure 6-12).

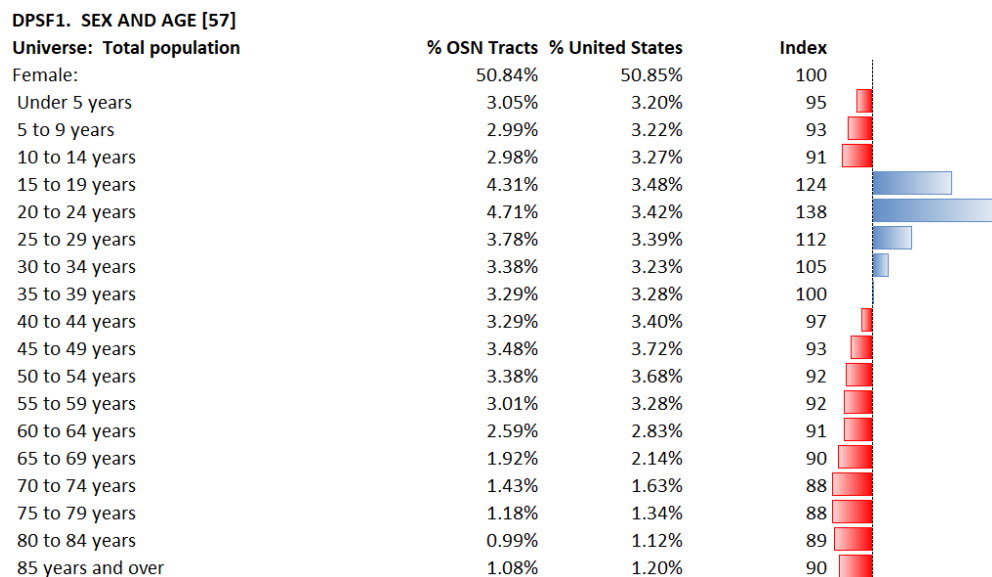


Figure 6-12 – US2012: Female population by age (OSN Tracts Indexed against US)

OSN Tracts are also over-represented by ethnically diverse non-White groups when compared to the US base (Figure 6-13, p270).

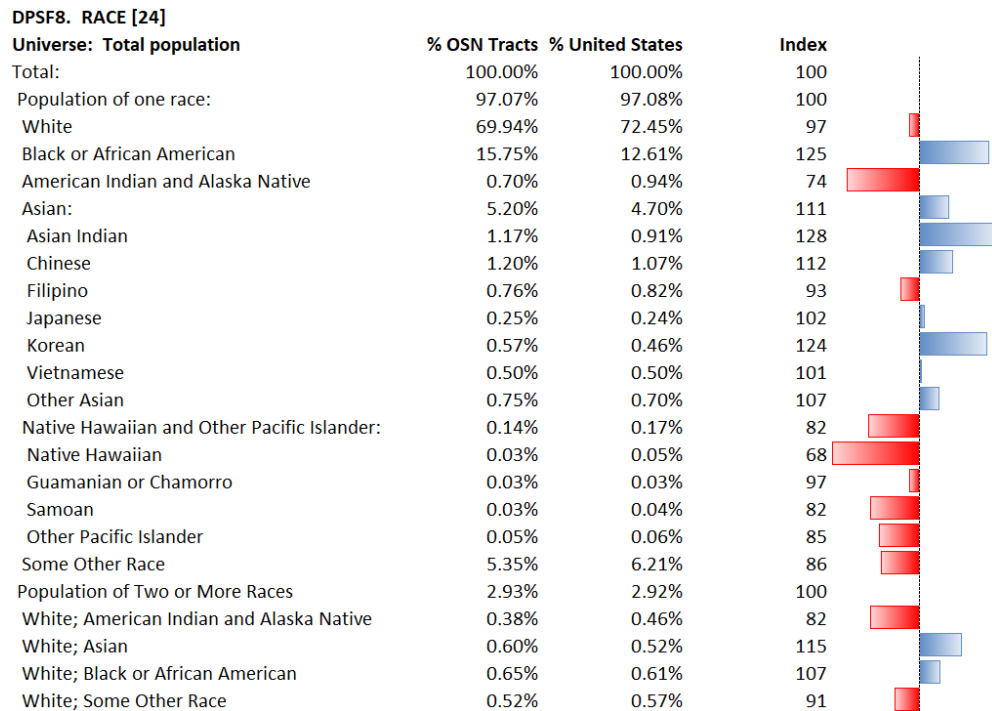


Figure 6-13 – US2012: Population by race (OSN Tracts indexed against US)

The profile shown in Figure 6-13 suggests that OSN Tracts are slightly below US norms for White population and above US norms, particularly for ‘Black or African American’, ‘Asian Indian’, and ‘Korean’ ethnic groups. While the latter two ethnic groups comprise only a very small percentage of total US population, the ‘Black or African American’ ethnic group comprises a larger proportion (12.61%) of US population and formed a key target group in both of Barack Obama’s ‘online’ 2008 (Kiyohara, 2009) and 2012 Presidential campaigns (Bimber, 2014).

Finally, OSN Tracts are significantly over-represented (index=239) by population ‘In group quarters’ when compared to the US as a whole (Figure 6-14, p271), particularly in terms of the US Census Bureau’s ‘Noninstitutionalized population’ class (index=381) living in ‘group quarters’. The US Census Bureau (2010, p1) defines Group Quarters as ‘a place where people live or stay, in a group living arrangement, that is owned or managed by an entity or organization providing housing and/or services for the residents’.

DPSF12. RELATIONSHIP [20]			
Universe: Total population	% OSN Tracts	% United States	Index
Total:	100.00%	100.00%	100
In households:	93.86%	97.43%	96
Householder	37.81%	37.79%	100
Spouse	16.74%	18.28%	92
Child	26.45%	28.83%	92
Own child under 18 years	19.50%	20.98%	93
Other relatives	5.97%	6.63%	90
Under 18 years	2.20%	2.53%	87
65 years and over	0.87%	0.96%	91
Nonrelatives	6.87%	5.90%	117
Under 18 years	0.37%	0.43%	86
65 years and over	0.23%	0.26%	88
Unmarried partner	2.53%	2.51%	101
In group quarters:	6.14%	2.57%	239
Institutionalized population:	1.26%	1.29%	98
Male	0.87%	0.88%	100
Female	0.39%	0.41%	95
Noninstitutionalized population:	4.88%	1.28%	381
Male	2.51%	0.69%	365
Female	2.37%	0.59%	398

Figure 6-14 – US2012: Population in households and group quarters (OSN Tracts indexed against US)

Group Quarters ‘include such places as college residence halls, residential treatment centers, skilled nursing facilities, group homes, military barracks, correctional facilities, and workers’ dormitories’ (US Census Bureau, 2010). The non-institutionalised population is comprised of people who are 16 years or older and are not inmates of penal, mental or elder-care institutions and who are not serving in the armed forces. The high index values (381 overall, 365 male and 398 female) of OSN Tracts in non-institutionalised population against the US base (Figure 6-14) are suggestive of a significant student population of coordinate-geotagging social media users, i.e. ~3.8 times more in OSN Tracts than would be expected generally in the US. Using a GIS to map coordinate-geotagged interactions and college and university sites, extracted from OpenStreetMap (2018), against non-institutionalised population broadly confirms this conclusion, as many educational institutions are proximal to large numbers of coordinate-geotagged OSN interactions and in or near Tracts, many of which are in urban areas, with high percentages of non-institutionalised population.

6.4.4.2 UK Census / SCOT2014 data fusion

At the 2011 Census 181,408 OAs in England and Wales contained, on average, a population of 309 (Office for National Statistics, 2012) while in Scotland the average for 46,351 OAs was 110 (National Records of Scotland, 2013). Using the same methods detailed above, running a point-in-polygon intersection of coordinate-geotagged OSN interactions (n Facebook posts= 1,227, n Twitter tweets= 92,311, Twitter retweets= 94,307) against UK Output Area boundaries, several similarities in population profiles to the US results (Section 6.4.4.1, p265) were revealed, albeit with different Census counts.

Age (Base: all usual residents)	% OSN OAs	% United Kingdom	Index	
Aged 0-4	4.7	6.2	76	
Aged 5-7	2.8	3.4	82	
Aged 8-9	1.7	2.1	81	
Aged 10-14	5	5.8	86	
Aged 15	1.1	1.2	92	
Aged 16-17	1.8	2.5	72	
Aged 18-19	3	2.6	115	
Aged 20-24	10.1	6.8	149	
Aged 25-29	9.2	6.8	135	
Aged 30-44	21.1	20.5	103	
Aged 45-59	19.1	19.6	97	
Aged 60-64	5.7	6.1	93	
Aged 65-74	8.1	8.7	93	
Aged 75-84	4.9	5.6	88	
Aged 85-89	1.3	1.5	87	
Aged 90+	0.5	0.8	63	

Figure 6-15 – SCOT2014: Population by age (OSN OAs indexed against UK)

Altogether, 140,673 Latitude and Longitude coordinate pairs from 187,845 coordinate-geotagged interactions (74.89%) could be allocated to 21,811 distinct 2011 UK Output Areas, 9.58% of all OAs in the UK. The remaining 47,172 coordinate-geotagged interactions in the SCOT2014 data set fell outside the territorial boundary of the UK. Of the 140,673 intersected coordinate-geotagged interactions, 67.10% fell to OAs within Scotland, 30.93% to OAs within England and 1.96% in Wales (Figure 6-16, p273). These OAs (Figure 6-15, above), as per the Census Tracts containing coordinate-geotagged US2012 interactions (Figure 6-10, p268), also exhibit a bias towards younger age groups, particularly those aged 20-24 (index=149).

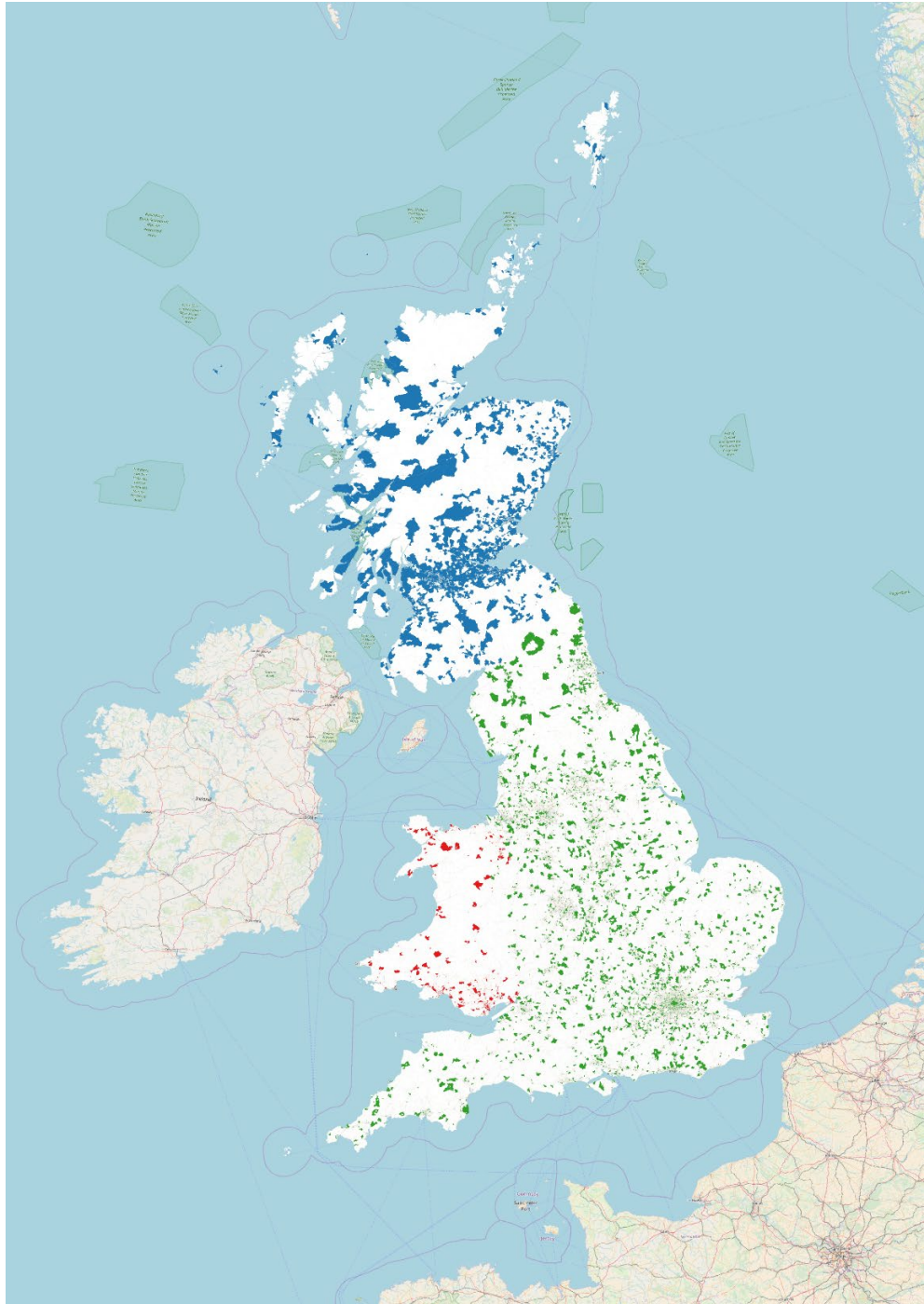


Figure 6-16 – SCOT2014: UK Output Areas (England=green, Wales=red, Scotland=blue) intersecting coordinate-geotagged OSN interactions

Looking at the composition of this population by gender and economic activity, OSN OAs (Figure 6-17, p274) have higher than normal proportions of ‘economically

active male full time students' (index=172) and 'economically inactive male students' (index=129).

Working Status (Base: all MALES aged 16-74)			
Economically Active Males			
	% OSN OAs	% United Kingdom	Index
Employee: Part time	4.9	6	82
Employee: Full time	47.6	46.7	102
Self-employed	12.9	13.5	96
Unemployed	4.6	5.3	87
Economically active FT student	5.5	3.2	172
Economically In-active Males			
	% OSN OAs	% United Kingdom	Index
Retired	9.6	12	80
Student	7.6	5.9	129
Looking after home or family	0.6	0.8	75
Long-term sick or disabled	4	4.5	89
Other	2.8	2	140

Figure 6-17 – SCOT2014: Male population by economic activity (OSN OAs indexed against UK)

The picture for economically active/inactive females (Figure 6-18) is similar, with above average numbers of female students in OSN OAs; economically active full time female students index at 143 and economically-inactive female students at 139.

Working Status (Base: all FEMALES aged 16-74)			
Economically Active Females			
	% OSN OAs	% United Kingdom	Index
Employee: Part time	17.6	21.2	83
Employee: Full time	33	30.5	108
Self-employed	6	5.6	107
Unemployed	3.5	3.5	100
Economically active FT student	5.3	3.7	143
Economically In-active Females			
	% OSN OAs	% United Kingdom	Index
Retired	14.3	15.8	91
Student	7.8	5.6	139
Looking after home or family	6.4	7.6	84
Long-term sick or disabled	3.9	4	98
Other	2	2.3	87

Figure 6-18 – SCOT2014: Female population by economic activity (OSN OAs indexed against UK)

Lastly, in case there were any doubt that the outcome of the 2014 Scottish Independence Referendum most affected and interested natives of Scotland, Figure 6-19 (p275) shows the population profile of OSN OAs by country of birth.

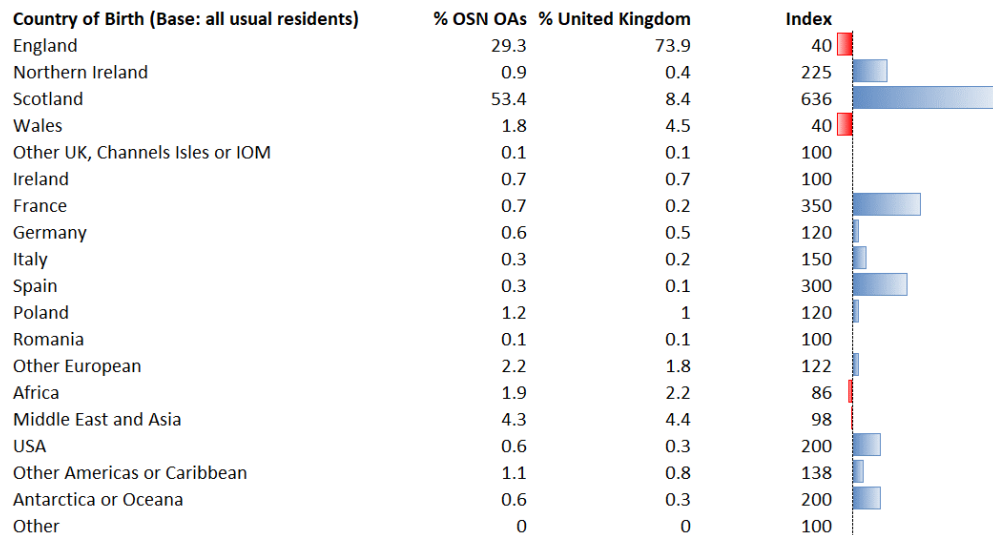


Figure 6-19 – SCOT2014: Population by country of birth (OSN OAs indexed against UK)

Output Areas with a high proportion of residents born in Scotland exhibit a high index value of 636, reflecting a great deal of Scottish interest in the Referendum.

6.4.4.3 Data fusion summary

The findings reported above, of a relatively youthful coordinate-geotagging user base, are broadly in line with several studies in the literature (Ajao et al., 2015; Arribas-Bel, 2014; Jiang et al., 2016; Longley et al., 2015; Murthy, Gross, & Pensavalle, 2016; Steiger, Westerholt, et al., 2015; Wachowicz & Liu, 2016), which use a variety of techniques to estimate age, sex, and the urban concentration of OSN users. The finding that US Census Tracts containing coordinate-geotagging OSN users have a much higher than expected percentage of non-institutionalised population against the US base has not been reported elsewhere. The picture is imperfect, with results dependent on geodemographic profiling of a small proportion of U2012 and SCOT2014 users whose geotagged coordinates are assumed to be home locations. However, in the total absence of any reliable demographic information from Facebook, Twitter or DataSift, data fusion provides an indication of the sorts of areas we might expect geotagging users to come from. If Jiang et al. (2016, p349) are correct in their assertion that ‘this one percent [of geotagging users] is already large enough’ to sustain these types of analyses, then

the US and UK Census-based profiling of 292,240 coordinate-geotagged OSN interactions provides another useful investigatory output from this research.

6.4.5 Graph analysis

Graph analysis, and a new range of emergent graph databases (I. Robinson et al., 2015), are particularly well-suited to the visualisation, network mapping and examination of OSN data sets, which are themselves inherently graph-based. Inspired by Euler's study of the *Seven Bridges of Königsberg* (Shields, 2012), graph theory focuses on connections formed between 'nodes' by 'edges'. In Euler's study, the nodes comprised four parts of the city of Königsberg (now Kaliningrad), which were linked by seven bridges crossing the River Pregel (Gribkovskaia et al., 2007). Challenged to determine, mathematically, whether it was possible to walk through each part of the city (Figure 6-20) crossing each bridge only once, Euler approached the problem topologically, laying down the foundations for graph theory.

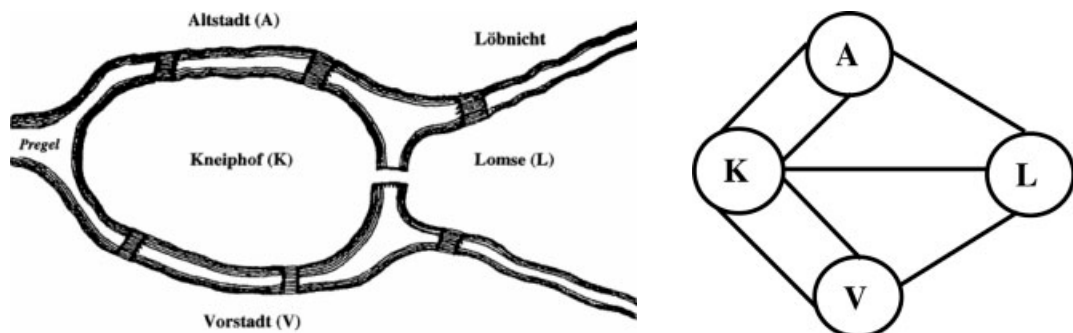


Figure 6-20 – Euler's drawing of the bridges of Königsberg in 1736 and a graphical representation of the bridges of Königsberg in 1736 after Gribkovskaia et al. (2007, p200)

Regardless of the geometrical layout of the city and its bridges, Euler determined that the solution rested on the 'connectedness' of the graph and the overall number of odd or even 'degrees' (or connections) at the nodes. A connected graph has a link between every pair of nodes. In Königsberg each of the four nodes (or parts of the city) had an odd number of degrees (the bridge connections for K=5, A=3, L=3, V=3 shown on Figure 6-20) and the city was served by an odd number of

bridges. As Euler proved (J. R. Newman, 1953, p70), ‘If there are more than two regions which are approached by an odd number of bridges, no route satisfying the required conditions can be found.’ Only a fully connected graph with zero or two nodes of odd degree would allow Euler’s Walk; in Königsberg all four nodes were odd. Similar principles of ‘connectedness’ have been used in modern graph database software and applied to social network analysis (Russell, 2011, p288). Social networks comprise actors (Section 4.1, p118) who interact in some way, perhaps liking each other’s Facebook posts, mentioning each other on Twitter, or retweeting a given message.

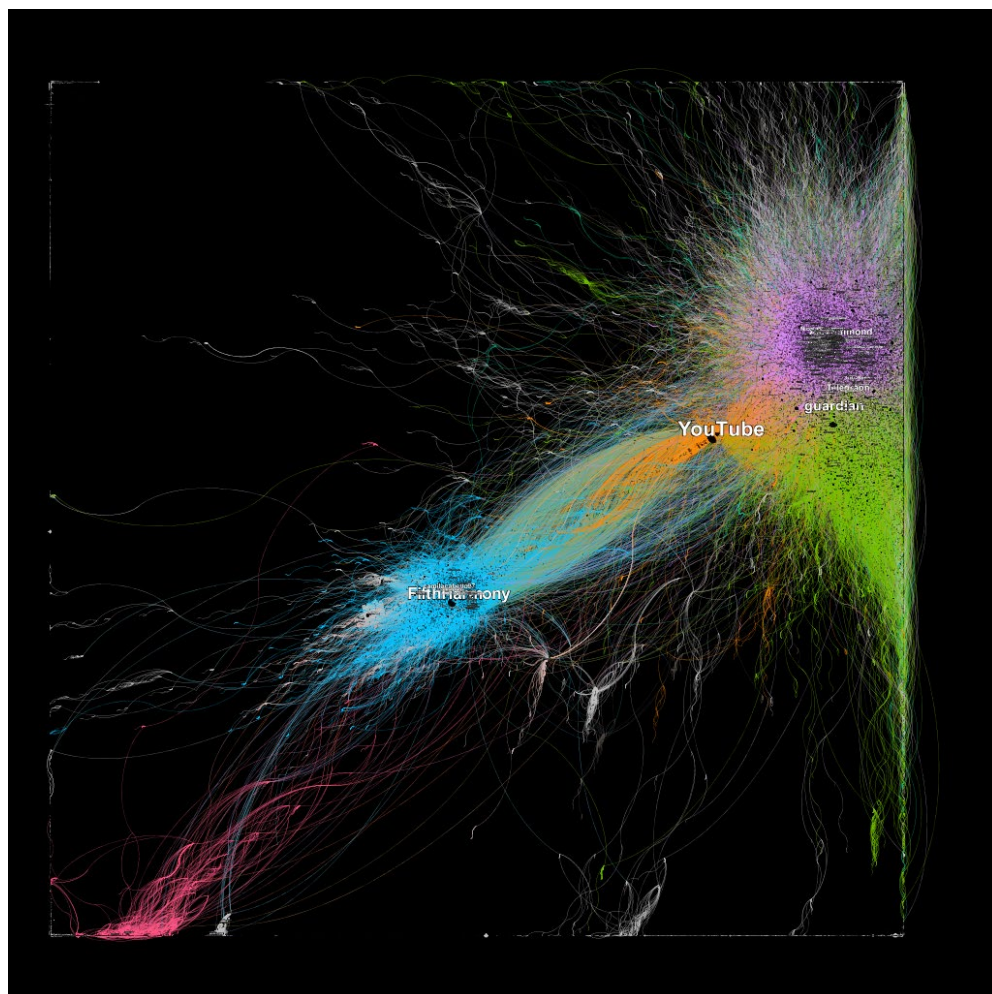


Figure 6-21 – SCOT2014: Gephi visualisation of ‘Twitter mentions’ (397,083 Nodes; 908,054 Edges) showing YouTube as the nexus between campaign-related interactions (in purple and green) and discussion of girl-band Fifth Harmony’s single ‘Better Together’ (in blue)

Social media actors may be considered nodes and the forms of interaction edges in this scenario (Figure 4-1, p120). Using Gephi (2018a) graph visualisation and analysis software, both on a Windows 10 laptop and on a Scientific Linux 6 HPC node of the SCIAMA supercomputer (Appendix 8, p436), several graphs have been computed using OSN data exported from Oracle 12c (Section 4.3.1.3, p145). Larger sets of OSN interaction relationships could be analysed on the more powerful SCIAMA supercomputer, including the 'Twitter mentions' graph from the SCOT2014 data set shown in Figure 6-21 (p277), consisting of 397,083 nodes and 908,054 edges.



Figure 6-22 – American girl-band Fifth Harmony's album, featuring title track Better Together, also the campaign slogan of the Unionist 'Vote No' coalition, was released on 18 October 2013 during data collection for the 2014 Scottish Independence Referendum

The spatialised graph visualisation shown in Figure 6-21 (p277), produced using the Force Atlas 2 algorithm (Jacomy et al., 2014) and colour-coded by modularity class (Blondel et al., 2008), shows YouTube as the nexus of interaction relationships linking political campaign groups and news organisations (to the right hand side) with American girl-band Fifth Harmony (2013), in the centre, whose album *Better Together* (Figure 6-22, p278) was released soon after OSN data collection commenced for the 2014 Scottish Independence Referendum. The album title, coincidentally, used the campaign slogan adopted by the Unionist 'Vote No' coalition in the referendum and was one of several filters used to sample OSN interactions at the time (Section 4.2.4.2, p129).

Graph analysis has proven useful in identifying communities in OSN data that are much less apparent when 'mining' data using SQL. A combination of graph analysis techniques and other indicators, e.g., identifying users making large numbers of one-sided communications, may also prove especially useful when identifying robotic posting (Marechal, 2016; Vosoughi et al., 2018). Altogether 153,637 OSN interactions in the SCOT2014 data set (2.37% of the total) mentioned 'Fifth Harmony' in message text and 1,206 of these (0.78%) were coordinate-geotagged. This reconfirms, once again, the typically low ~1-2% coordinate-geotagging rates observed more widely in this study (Table 4-8, p170) and by others (e.g., Leetaru et al., 2013), regardless of the terms used to 'select for inclusion' when filtering and recording OSN interactions.

6.4.6 Data skewness

Drawing on graph theory, Stefanidis, Crooks, et al. (2013, p329) point out that social media networks, especially Twitter, are 'highly skewed in the sense that the majority of nodes have a low degree of connectivity while there are a small number of nodes which have a high degree.' The highly connected nodes 'can be considered as hubs of information dispersal and to some extent key actors in the social media sphere.'

As Table 6-5 shows, many of these ‘key actors’ *really are* actors, singers, sports-stars or other modern-day media celebrities (TwitterCounter, 2017). Ranking the top 10 US2012 users posting on Twitter by ‘followers count’ (the number of others following the author), calculated using SQL (Appendix 11 listing 44, p492), reveals four celebrities in positions 1, 3, 9 and 10 with massive Twitter followings.

Table 6-5 – US2012: Top 10 users posting on Twitter by number of followers

Position	Author Name	Followers Count
1	Perez Hilton	5,716,500
2	OMG Facts	5,107,364
3	Stephen Fry	4,954,594
4	E! Online	4,899,137
5	TIME.com	3,907,791
6	CNN en Español	3,255,309
7	The White House	3,153,754
8	The New York Times	3,129,274
9	Pete Cashmore	3,022,966
10	MC HAMMER	2,885,523

While some major news organisations – e.g., *The New York Times* newspaper and *Time* magazine – and ‘The White House’ also appear, the number of celebrities in the top 10, and their reach in terms of followers, is striking. The influence of celebrity figures in social media networks, as measured here by followers count on Twitter, is in line with findings from other academic studies into online social media ‘opinion leaders’ (Karlsen, 2015). In a political context, Park, Lee, Ryu, & Hahn (2015, p246) have suggested that ‘the rise of networked media such as Twitter [has amplified] celebrities’ ability to speak on policy matters directly to the public’; a factor which may well have been influential in the election of ex-TV personality Donald Trump to the US Presidency in 2016.

Altering and re-running the query with the addition of a constraint to check for the presence of geographical coordinates (Appendix 11 listing 45, p492), it is possible to determine the top 10 users who have coordinate-geotagged at least one of their Twitter tweets. Coordinate-geotagging hip-hop musician, MC Hammer (also in

Table 6-5, p280), appears first in this list (with 2,885,523 average followers), joined by fellow singer Lea Salonga in second (1,236,081) and Joe Trippi (an American Democrat campaign worker, on 1,017,789) in third.

Table 6-6 – US2012: Top 10 users posting with coordinate-geotagged messages on Twitter by number of followers

Position	Author Name	Followers Count
1	MC HAMMER	2,885,523
2	Lea Salonga	1,236,081
3	Joe Trippi	1,017,789
4	B.J. Mendelson	766,765
5	Jason Squatriglia	384,269
6	Javed Akhtar	351,286
7	The Real Lil Dee	314,083
8	natalie nunn	305,429
9	WW2 Tweets from 1940	260,635
10	Kevin Nash	255,709

Only three coordinate-geotagging users with over one million followers tweeted on Twitter in the US2012 data set. Removing the condition for one million minimum followers altogether, Table 6-6 shows the top 10 users who have posted at any time with coordinate-geotags, together with their number of followers. The average number of followers for coordinate-geotagging users in the US2012 data set (Table 6-6), after the top three positions, is generally around one order of magnitude lower than that for the top 10 most followed users (Table 6-5, p280).

The pattern is broadly repeated, with larger orders of magnitude, when examining non-coordinate-geotagged and coordinate-geotagged tweets made by well-followed users during the 2014 Scottish Independence Referendum (Table 6-7, p282). It appears that coordinate-geotagging users, and the messages they create, have lower reach in terms of follower count than the most prominent users of Twitter, most of whom choose to tweet without coordinates. Not only do few users overall choose to coordinate-geotag their interactions (Section 5.2.1, p188) but

those who do so are much less widely-followed than their non-coordinate-geotagging peers.

Table 6-7 – SCOT2014: Top 10 users posting without/with coordinate geotagged messages on Twitter by number of followers

Position	Without coordinates	Followers	With coordinates	Followers
1	CNN	14,040,804	EL MUNDO	1,526,046
2	MTV	9,788,593	Stephanie Pratt	700,942
3	BBC News (World)	7,288,128	Khaled Abol Naga	694,501
4	E! Online	6,784,965	Antena3Noticias	681,692
5	TIME.com	6,161,366	Nigeria Newsdesk	600,726
6	The New York Times	6,035,089	Cadena SER	447,902
7	People magazine	5,384,668	Martin Lewis	308,775
8	Samsung Mobile US	5,155,299	Jon Snow	308,746
9	Reuters Top News	4,716,883	Eyewitness News	244,571
10	Perez Hilton	4,564,096	Marc 	213,458

The counts, and positions, of the most-followed users active during OSN recording are also highly influenced by contemporary events, with sometimes unexpected consequences. Perez Hilton, a gossip columnist who is widely-followed on Twitter, made just five tweets recorded during the 2014 Scottish Independence Referendum data collection exercise:

1. @KathieLGifford Fifth Harmony Tells Kathy Lee & Hoda Why They're Better Together On The Today Show! <http://t.co/TQsEuqqSeb>
2. Fifth Harmony Tells Kathy Lee & Hoda Why They're Better Together On The Today Show! <http://t.co/NZCYE7gDY3>
3. @FifthHarmony Fifth Harmony Tells Kathy Lee & Hoda Why They're Better Together On The Today Show! <http://t.co/TQsEuqqSeb>
4. @hodakotb Fifth Harmony Tells Kathy Lee & Hoda Why They're Better Together On The Today Show! <http://t.co/TQsEuqqSeb>
5. RT @hodakotb: Xo RT “@PerezHilton: @hodakotb Fifth Harmony Tells Kathy Lee & Hoda Why They're Better Together On The Today Show!
<http://t.co/x8e4229SuN>”

All of these tweets appeared in the SCOT2014 data set as their text contained the search phrase ‘Better Together’; the campaign slogan of the Unionist ‘Vote No’ alliance during the 2014 Scottish Independence Referendum (Appendix A7.3, p435). Hilton’s tweets, however, refer to American girl-band *Fifth Harmony* (2013), whose extended play album titled *Better Together* (Figure 6-22, p278) was released on 18 October 2013 (Section 6.4.5, p276) early in the data acquisition phase. Without this coincidence Perez Hilton’s interactions would not have appeared in the research data corpus at all and he would not have appeared at the head of Table 6-5 (p280).

Skewness in social media networks, and the potential for computerised misinterpretation of text based on string-matching rather than intelligent reading, highlights methodological problems (Section 3.3, p102) with machine-based analysis of large social media corpora, where search terms used for filtering may have multiple discursive meanings.

6.5 Summary

Exploratory spatiotemporal data analysis and visualisation techniques (Chapter 3, p94) have been used to examine large numbers of public-domain OSN interactions sampled and stored in this research (Chapter 4, p118) to answer the three research questions (Chapter 5, p186) set out in the introductory chapter of this thesis.

The 46.90GB of OSN data originally exported from DataSift in CSV and JSON formats in 2012 and 2014 (Section 4.2.5, p134), later stored in an Oracle 12c database (Section 4.3.1.3, p145), has been supplemented by another 5.58GB of augmented JSON file output from GATEcloud and CLAVIN-rest, together with a further 6.92GB of JSON data streamed directly from AlchemyAPI into the database. Table 6-8 (p284) shows that data augmentation operations conducted during this research (Section 4.4.1, p147) have necessitated the storage of an additional 21.39GB of data in Oracle 12c database tables; the OSNDATA database being over 150GB in size.

Table 6-8 – File output sizes and database table sizes (in GB) for data augmentations used in this research

System	File output (GB)	In database (GB)
GATEcloud	5.09	13.21
AlchemyAPI	NA	6.92
CLAVIN-rest	0.49	1.26
TOTAL	5.58	21.39

Very large amounts of data are intrinsically hard to digest, yet the ‘perception capabilities of the human cognitive system can be exploited by using the right visualizations [which can amplify] human cognitive capabilities in six basic ways’ (van der Aalst, 2014, p24):

1. by increasing cognitive resources, such as by using a visual resource to expand human working memory;
2. by reducing search, such as by representing a large amount of data in a small space;
3. by enhancing the recognition of patterns, such as when information is organized in space by its time relationships;
4. by supporting the easy perceptual inference of relationships that are otherwise more difficult to induce;
5. by perceptual monitoring of a large number of potential events, and;
6. by providing a manipulable medium that, unlike static diagrams, enables the exploration of a space of parameter values.

All of the data analysis and visualisation methods used in this research were designed, after van der Aalst (2014), to exploit and amplify the ‘perception capabilities’ of the researcher. Not all of them, as van der Aalst notes in his sixth point above, are easy to reproduce on paper. Nonetheless, using NLP/geoparsing software to conduct text-mining, and SQL in Oracle 12c to data-mine outputs, this thesis has demonstrated – through tables, figures and statistics – clear differences

between coordinate-geotagging and non-coordinate-geotagging users of two popular social media platforms.

The final chapter, overleaf, concludes this thesis and summarises this work. Data sourced from Online Social Networks, more often studied by Computer Scientists than Geographers (Figure 2-5, p62), offers many new opportunities for social science and geographical research. Several avenues for further and future research are highlighted in the concluding chapter, alongside an assessment of the validity of the research presented above.

7 CONCLUSION

7.1 Introduction

Geography matters! Just as Massey & Allen (1984) reaffirmed the relevance of geography in socio-spatial, environmental, political and economic spheres, a conception of place clearly matters when individuals interact online using social media platforms. In the 2012 US Presidential Election and the 2014 Scottish Independence Referendum case studies examined here, ~3.5-4.5 million toponymic mentions have been identified in around one quarter of the ~8 million interactions in the research data corpus. Around one quarter of the ~7 million entities identified in ~650,000 distinct URLs – posted, tweeted or retweeted ~3.5 million times – also contained toponymically identifiable content. Elections are peculiarly geographic, as well as political, events. Voters' affiliations, and the many attempts made to influence or predict them, are often highly correlated with spatially unevenly distributed factors such as levels of income or wealth, access to education, age and other demographic characteristics, as well as micro-economic and familial effects (Johnston & Pattie, 2006). It is, therefore, both unsurprising and reassuring to find that electorates and commentators make frequent geographical references online during electoral campaigns, and that many of these mentions refer to the 'swing' states or constituencies whose ballot results typically shape wider political outcomes.

What then of *geotagging*; does it *matter*? Geotagging is a relatively recent socio-technological phenomenon, primarily enabled by the worldwide proliferation and usage of GPS-equipped mobile, or smartphone, devices. The increasingly large volumes of Ambient (and/or Volunteered) Geospatial Information now available (Elwood et al., 2012; Goodchild, 2007; Stefanidis, Crooks, et al., 2013) offer new research opportunities for scholars in geography and related social science disciplines. Increased scrutiny of 'Geo-social Networks' (Bahir & Peled, 2013), and

the possibilities they afford for wider geographical analysis, are demonstrated by the growing number of academic papers and specialist journals published in the last decade or so, many cited and listed as references (p319) to this thesis. Geotagged photographic images publicly posted on Flickr have been used to combat wildlife poaching in protected areas and in criminological research (Lemieux, 2015).

Geotagged social media and other 'Big Data' have been used to monitor natural disaster situations (Burns, 2018; Goodchild & Glennon, 2010). OpenStreetMap has been used in the study of the production and 'prosumption' of user-generated geographic Big Data (Cockayne, 2016). Human interaction data sourced from Twitter and, to a lesser extent, Facebook have been used, seemingly, to 'do everything'; from monitoring earthquakes (Crooks et al., 2013) to tracking riots (Bonilla & Rosa, 2015; Crampton et al., 2013), helping to demarcate urban areas (Yin et al., 2017) and much else besides (see Kapoor et al., 2017, for a recent and comprehensive summary of application areas). This proliferation of research activity, for example, '[delineating] city cores, [gaining] insights into travel plans and tourism, [characterizing] urban landscapes, [studying] global migrations or [identifying] mobility patterns', has also been identified by Rzeszewski & Beluch (2017, p2), who go on to note that:

[...] there has been a growing interest in filling the gap in our knowledge about the demographics of both the Twitter user population as a whole and the subgroup of users that produce (or rather contribute since they may not be aware of it) an ambient geospatial information (AGI). The former has been addressed on many spatial scales by a range of papers with the general conclusion that Twitter users are younger than the general population and derive predominantly from urban areas, with gender and ethnic biases still visible but becoming less pronounced over time. The latter, however, has been given much less attention.

(Rzeszewski & Beluch, 2017, p2)

The present research focuses attention on this comparatively under-researched ‘subgroup’ of users, who consume, produce and share messages and links containing ambient ‘geospatial’ information. The thesis contributes to knowledge through a comprehensive analysis and cross-comparison of toponymic mentions in the message text and URL link shares of coordinate-geotagging and non-coordinate-geotagging users interacting online during two data-rich political case study events. From this research it is possible to conclude (Figure 7-1, p289) that:

1. Coordinate-geotagging users make fewer toponymic mentions in message text than non-coordinate-geotagging users of two popular OSN platforms;
2. Coordinate-geotagging users make far fewer URL link shares than non-coordinate-geotagging users, and;
3. The content of URLs shared by coordinate-geotagging users makes fewer mentions of place than content shared by non-coordinate-geotagging users.

These conclusions are at odds with the research hypothesis, that *coordinate-geotagging users are the most geographically expressive of all OSN users*. Although they do actively (or accidentally) coordinate-geotag their Twitter Tweets or Facebook Posts, this small group of social media users are not, in the three important respects above, representative of all OSN users. Of course, OSN users in general (Diaz et al., 2016), and geotagging users in particular (Sloan & Morgan, 2015), are not thought to be representative of the general population. During elections, they are likely to be even less so; probably being younger and living in urban areas (Section 6.4.4, p262) and often, according to Barberá & Rivero (2015), exhibiting ‘extreme ideological preferences’.

The research findings presented here imply that geographical outputs (point maps, counts or aggregations to larger areal units such as constituencies or states) based on searches for specific words, toponyms, #hashtags or @mentions in message text or URL link shares, which may readily be mapped using the interaction Latitude and Longitude coordinates of Twitter tweets or Facebook posts deposited by

coordinate-geotagging users (or aggregated to wider areas using a GIS), are unlikely to be representative of the spread of all such content within OSNs.

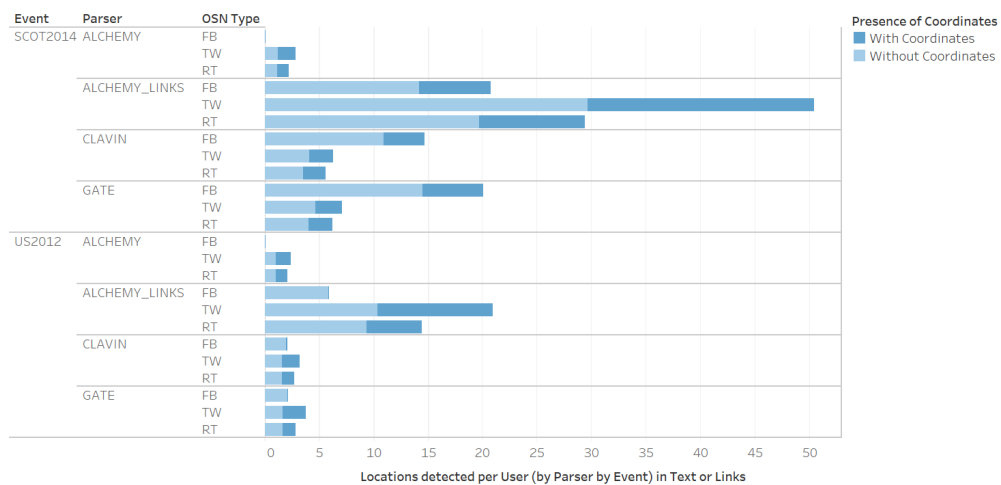


Figure 7-1 – US2012/SCOT2014: Number of toponymic mentions/user identified in message text (FB=Facebook, TW=Tweet, RT=Retweet) and linked/shared URL content

The importance of this conclusion to the academic community and to society, along with suggestions for further research in this area, are discussed in the remaining sections of this chapter.

7.2 New opportunities

Twitter's introduction of tweet geotagging functionality in 2009, closely followed by Facebook's broadly similar system for posts in 2010, offered new opportunities for researchers and others (governments and their surveillance agencies chief amongst them) seeking to mine potentially vast depositories of digital, time-stamped, textual, audio-visual and, in some cases, explicitly geospatial human-made content. However, research into OSN usage, and geographical OSN usage in particular, is still in its relative infancy. Advances in platform operators' systems, massively increased usage of social media websites and the development of technologies better able to store and analyse available digital 'Big Data' have enabled this research. These developments are thought to have created 'a paradigm shift to computational social science' (R. M. Chang et al., 2014), criticised by some as a form of 'digital

positivism' (Fuchs, 2017a), in many ways echoing academic Geography's much earlier battles (I. Burton, 1963; Harvey, 1973) with its 'quantitative revolution' (Cresswell, 2013, 2014; Johnston et al., 2014; Wyly, 2014).

Graham & Shelton (2013, p257) have argued that 'geographers have long struggled over what the appropriate ends of our scholarship should be, how we should be doing it, and how to accommodate competing claims to truth, especially in the context of new technologies opening up new methodological possibilities.' New forms of geographical data, such as the coordinate-geotags or toponymic mentions deposited on OSN websites and applications by billions of individuals, do offer new possibilities for research. By necessity these investigations tend to rely very heavily upon computerised analytical methods, as 'massive datasets of communication are challenging traditional, human-driven approaches to content analysis' (S. C. Lewis et al., 2013, p34). Nonetheless, while alert to the possibility that 'Big data give us a quickly expanding, shallow view of the vast horizontal landscape of the desert of the present real' (Wyly, 2014, p28), the research presented here does successfully move 'beyond the geotag', as Crampton et al. (2013) have advocated, by:

1. going beyond social media that is explicitly geographic;
2. going beyond spatialities of the 'here and now';
3. going beyond the proximate;
4. going beyond the human to data produced by bots and automated systems;
5. going beyond the geoweb itself, by leveraging these sources against ancillary data, such as news reports and census data.

It is well-known, and has been known for some time (Z. Cheng et al., 2010; Leetaru et al., 2013), that only a small percentage (typically ~1-2% of publicly-available Twitter tweets, and a lower percentage of Facebook posts) are geotagged with Latitude and Longitude coordinates. Much has been learnt by studying these records, as even small percentages of the enormous number of social media messages generated every day yield very large absolute numbers of geotagged

interactions. It has been suggested that ‘this one percent [of coordinate geotagged records] is already large enough’ for meaningful analyses (Jiang et al., 2016, p349), and in some application domains (e.g., using geotagged messages or images as proxies for population location and movement) this may well be true. Somewhat surprisingly, however, comparatively little is known about the characteristics of these coordinate-geotagging users. Are their messages, for example, more widely shared than those of their non-coordinate-geotagging peers? Is their personal interest (or mistake) in sharing their precise location also manifested in high levels of geographicality in written text and shared links? Does this small class of coordinate geotagging users, at a more fundamental level, *matter*; either by offering a representative, but uniquely spatialised, view of OSN users in general, or by forming an important subset of OSN users who, through their precisely stated locations, might be held in particularly high or low esteem by other users of social network sites?

Developments in Geographic Information Retrieval (Purves et al., 2018), combined with an exploratory case study methodology and comprehensive technical approach (Chapters 3, p94 and 4, p118), have helped to answer the first of these two questions. Coordinate geotagging users are much less widely followed (Section 6.4.6, p279) than others on OSNs and, somewhat counter-intuitively, express themselves less geographically than others in their messages and through the choice of URLs they link to and choose to share (Section 5.2.3, p205). For the first time, the comprehensive analysis of social media interactions presented here, has revealed important differences in the posting behaviour of coordinate-geotagging and non-coordinate-geotagging users during two political case study events. The more fundamental question, whether *geotagging matters*, is not so easily answered. To a professional geographer, the vast number of coordinate pairs now deposited online by social media users appears highly propitious. On a massive scale, arguably for the first time in human history, it is possible to know *who* is saying *what*, *when* and *where*. Unfortunately though, as detailed earlier (Section

5.2.1, p188) and remarked upon by Paraskevopoulos & Palpanas (2016, p1), ‘only a very small percentage of [OSN] posts are geotagged, which significantly restricts the applicability and utility of [many] applications.’ Low rates of coordinate-geotagging in OSN data, and the unrepresentativeness of coordinate-geotagging users, do limit the ‘applicability’ of any analyses based solely upon geotagging users’ message text, metadata or spatial location in political contexts. When examining politicised communications made on social media networks, or determining how political opinion or (mis)information may be geographically tracked across these systems, it appears that *geo* matters, but *tagging* matters much less. The following section discusses these limitations in more detail.

7.3 [Geo]tagging, politics, prediction and tracking

The research detailed in this thesis was partly inspired by much earlier work (Section 1.5.1, p22) to develop and publish an electoral information website covering the first UK General Election of the Internet era (Tear, 1997). It was apparent even then, when operating the site, that patterns of online political information consumption were not geographically uniform. The extraordinary revelation that a third-party state, Russia, or Russian-backed ‘trolls’, attempted to influence the outcome of the US 2016 Presidential Election by geo-targeting social media sites with a variety of inflammatory messages (Schrage, 2017; Stretch, 2017), even aiming to promote a secessionist California (BBC News, 2017c) and reaching up to 126 million Facebook users (BBC News, 2017b), aptly demonstrates the originality of this research and the continuing relevance of geography in modern politics. As the Internet developed, and Web 2.0 *participation* succeeded Web 1.0 *publication* (O’Reilly, 2005), it had appeared increasingly likely that techniques widely-used by marketing professionals (geodemographic profiling, geographical targeting, bespoke messaging) could, and probably would, be deployed online in an attempt to influence turnout or voting behaviour during democratic elections. The

Facebook / Cambridge Analytica data misuse and political targeting scandal, which broke early in 2018, clearly reveals that this has, indeed, occurred.

Using new techniques in text-mining and ‘Big Data’ analysis this research examines how geography and geotagging might aid the understanding of politically discursive information sharing during elections, and whether OSN-enabled socio-geo-technological developments can be used to accurately track the geographical spread of political opinion or (mis)information online. The detailed analysis and evaluation of case study data presented here, from two recent electoral events, suggests that tracking of this type using coordinate-geotagged OSN interactions alone is not possible. In both the 2012 US Presidential Election and the 2014 Scottish Independence Referendum there are too few coordinate-geotagged records to map political behaviour, information consumption or sharing at a granular level. While geographical tracking of opinion or URL link/sharing may be attempted using coordinate-geotagged OSN interactions the ‘toponymic unrepresentativeness’ of geotagging users in message text and link/shares hinders such efforts. Coordinate-geotagging users link to much less external 3rd party content than non-coordinate-geotagging users of either of the two OSN platforms, Twitter and Facebook, examined during the two political case study events and also make fewer references to place in their message text. In addition, while enhanced geo-referencing using NLP and geoparsing software (Chapter 5, p186) may successfully find locations in text these approaches cannot easily, or accurately, determine *locational meaning*; whether the locations mentioned are ‘lived in’, ‘visited’, ‘worked at’ or ‘talked about’ more generally. Human behaviour exhibits a strong diurnal influence (S. A. Golder & Macy, 2011) which is also visible in social media data (Morales et al., 2017). Figure 7-2 (p294) shows the number of interactions recorded by time of day and day of week during the 2014 Scottish Independence Referendum. Most days, most posts are made during day time, increasing late afternoon into evening before tailing off in the small hours of the morning, coinciding with human periods of activity and sleep.

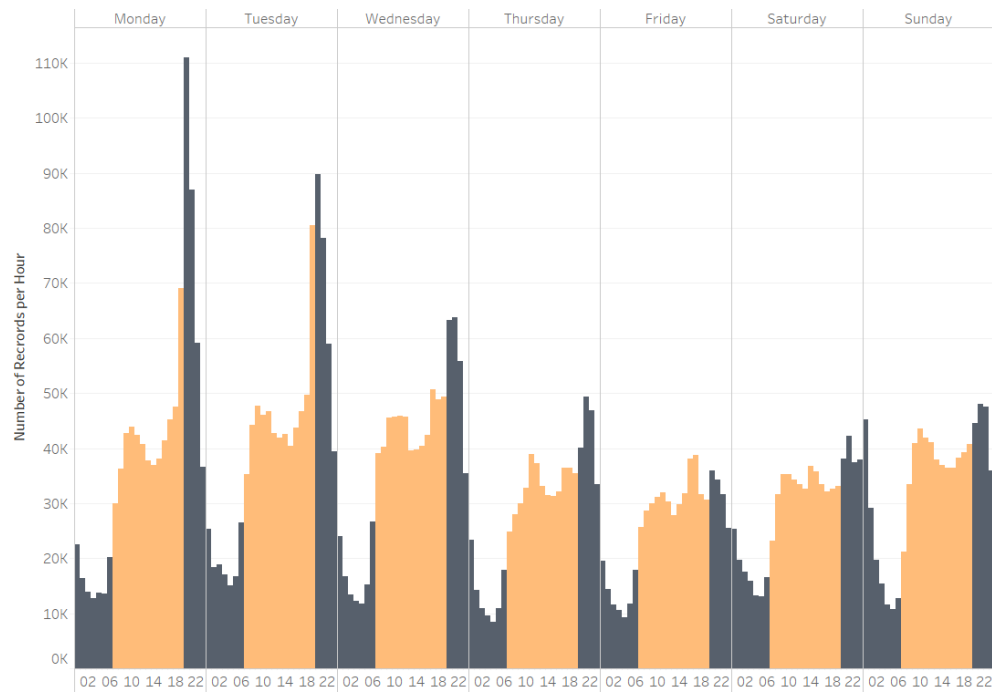


Figure 7-2 – SCOT2014: Number of social media posts per hour by times of day (light=day time; dark=night time) and day of week (smoothed by excluding counts from election day Thursday 18 September into Friday 19 September 2014)

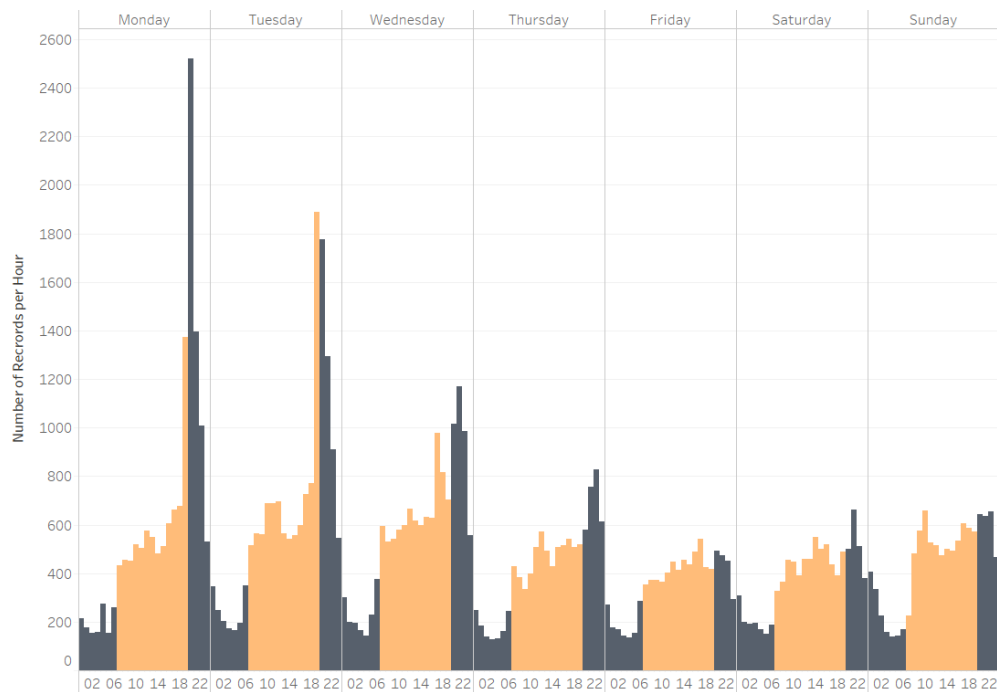


Figure 7-3 – SCOT2014: Number of coordinate-geotagged social media posts per hour by times of day (light=day time; dark=night time) and day of week (smoothed by excluding counts from election day Thursday 18 September into Friday 19 September 2014)

Coordinate-geotagged interactions (Figure 7-3, p294) exhibit a similar pattern, raising the obvious need to add 'home' and 'away' attributes to geo-locations, as voting normally takes place somewhere near home. A significant body of work is now devoted to this research question, analysing local commuting patterns (McNeill, Bright, & Hale, 2017), using 'end to end neural networks' (Lau, Chi, Tran, & Cohn, 2017), 'grid-based classifications' (Ajao, P, & Hong, 2017), 'geo-temporal' data (Longley & Adnan, 2016), NLP and network approaches (Alonso-Lorenzo et al., 2016), 'geo-location history' (Poulston et al., 2017) and even 'local' and 'global' references to celebrities in message text (Ebrahimi, ShafieiBavani, Wong, & Chen, 2017) to more accurately infer the positional meaning of any given spatial or detected locational reference found in users' OSN communications. Advances in addressing the 'home/away' problem may be particularly useful in several application domains (e.g., transport planning) and would assist in the tracking of politicised opinion and (mis)information sharing on social media sites. However, even these advances are unlikely to enable any form of accurate electoral prediction based upon social media data sources; especially when Twitter is used as the only, or major, source of such interactions.

Jungherr, Schoen, Posegga, & Jürgens, (2016, p1) have stated that in 'all tested metrics, indicators based on Twitter mentions of political parties differed strongly from parties' results in elections or opinion polls.' The authors advise using 'caution' in any attempt to infer political outcomes from Twitter data, the most commonly used source of OSN 'digital trace data', and 'question the power of Twitter to infer levels of political support of political actors.' This advice, and a similar note of caution, has been strongly expressed by Gayo-Avello (2012a), in a provocatively titled paper, which identifies repeated failings in many of the works attempting to predict electoral outcomes using Twitter data. Reviewing 17 papers, Gayo-Avello outlines eight 'Flaws in Current Research regarding Electoral Predictions using Twitter Data' which may be summarised as follows:

1. Research presents a *post-hoc* analysis not a prediction
2. Incumbency plays a ‘major role’ in most elections
3. There is no common currency for ‘counting [Twitter] votes’
4. There is no accepted way to compare predictions with reality
5. Sentiment analysis is ‘applied as a black-box and with naïveté’
6. All Tweets are presumed ‘trustworthy’ when they are not
7. Demographics are neglected
8. Self-selection bias is ignored

Results from a detailed ‘*post-hoc*’ analysis of the 2012 US Presidential Election and 2014 Scottish Independence Referendum data collected as part of this research programme reasserts the continuing presence of Gayo-Avello’s ‘Flaws’. In addition, there are a) significant difficulties in ascribing geographical meaning to coordinates or detected geo-locations, and; b) further, and deeper, difficulties in deriving meaning from short, unstructured text using computerised methods.

Phillips, Dowling, Shaffer, Hodas, & Volkova (2017, p1) have stated that ‘[social media] forecasting is limited by data biases, noisy data, lack of generalizable results, a lack of domain-specific theory, and underlying complexity in many prediction tasks.’ The conclusion, that geo-referenced OSN sentiment does *not* offer any locally predictive power for electoral outcomes, is not a new contribution to knowledge. However, the findings which show (Chapter 5, p186) that the message text and link/shares of coordinate-geotagging users are somewhat atypical of those produced by the majority of OSN users are both original and significant. We now know that, during elections, geographical and behavioural targeting of social media users has been conducted by political parties (Moore, 2016), companies (Albright, 2017; Cadwalladr, 2017) and even third-party states (Howard et al., 2018). These

recent revelations confirm the validity of the geographical perspective adopted in this research and suggest that much further research work (Section 7.5, p299), alongside several changes in policy (some suggested in Section 6.3, p238), are required to more effectively track the spread of politicised (mis)information online. Before these topics are discussed in greater depth, later in this chapter, the following section offers some reflections and criticisms of the work presented above.

7.4 Reflections and criticisms

This thesis presents results from a study explicitly designed to test a clear hypothesis – that coordinate-geotagging users are the most geographically expressive of all OSN users – by addressing three research questions:

1. How can baseline ‘geographicality’ be assessed and categorised in OSN data?
2. Does NLP-detectable ‘geographicality’ in message text increase in line with ‘spatiality’?
3. Does NLP-detectable ‘geographicality’ in linked/shared 3rd party content increase in line with ‘spatiality’?

Large volumes of social media data collected during two recent electoral events have been analysed to answer RQ1-3. The classification of PGI metadata fields, particularly from Twitter interactions, answers the first question (Section 5.2.1, p188). In both RQ2 (Section 5.2.2, p190) and RQ3 (Section 5.2.3, p205) results have shown that there is no support for the assertion that highly spatialised, coordinate-geotagging, users of social media sites are particularly geographically expressive whether in message text or in the choice of material they link to and share. In the two case study events examined here, the reverse is true.

Two main criticisms may be levelled at this research:

1. The data are old, and;
2. Research methods rely heavily on computerised techniques.

Addressing the first criticism is straightforward. New data could be collected from OSN interactions deposited around the time of another major election (or more generally) and the analyses re-run to determine whether the results reported here still hold true. While appealing, doing so would further extend the period of study; making the existing data older still and delaying submission of the thesis. Collecting and examining new data to reaffirm these results is, hence, either a task for the future or for other researchers.

The second criticism is more intimately bound with the research design (Chapter 3, p94) and research methods (Chapter 4, p118) used in this study. A hybrid case study-exploratory design using large amounts of digital data visualised and analysed using computational techniques is bound to encounter the types of criticisms eloquently expressed by Wyly (2014) and Fuchs (2017a, 2017b). These centre around the dangers of a ‘new quantitative revolution’ where ‘models arrive brandishing people – collections of thousands and millions of socially networked digital individuals in an expanding neoliberal noösphère’ (Wyly, 2014, p36). Models exist, it is suggested, in an atheoretical vacuum where social media are not adequately located ‘more generally within a model of society’ (Fuchs & Trottier, 2015, p121). Fuchs (2017a) argues in his paper, so-titled, for a move ‘*From digital positivism and administrative big data analytics towards critical digital and social media research!*’ Fuchs (2017a, pp38-39) points out that research by Peng, Zhang, Zhong, & Zhu (2013) has shown that ‘only 31% [of 27,340 Internet Studies articles published between 2000 and 2009] cited theoretical works [leading to a tendency to] engage with theory only on the micro- and middle-range [neglecting] the larger picture of society as a totality.’

While this thesis does not present a theoretical piece of research work, the work of many theorists is referenced here (Castells, 2009; Dahlgren, 2005; Deleuze et al.,

2004; G. Goldberg, 2010; Habermas, 2011; Kuhn, 1970; Papacharissi, 2002, 2010; Stahl, 2004). The identification of conceptually key political (Section 2.2.2.4, p60), communications (Section 2.5, p72), geographical (Section 2.6, p77) and technical (Section 2.7, p83) themes is deliberately designed to introduce cross-disciplinary theoretical elements to the research. Many other references (Barberá et al., 2015; Bimber, 2014; Diehl et al., 2016; Fuchs, 2017b; Harris & Harrigan, 2015) deal with the perplexing question of whether exposure to, or sharing of, online social media content actually has any great impact on the outcome of electoral events, and yet more (References, p319) detail technical approaches used to help answer these questions. As the field of *Internet Studies* or *Web Science* (Berners-Lee, Hall, Hendler, O'Hara, et al., 2006) matures theory can be expected to build. It seems likely, as Internet activity is human-made, even including the pernicious 'bots' designed to sway public opinion (Marechal, 2016), that theories will be based around several of the inter-disciplinary social science and technological strands detailed in this thesis. From the standpoint of academic geography, this thesis therefore makes an important contribution to work in a much wider subject area. Further, and future, research ideas based on this grounding are detailed in the following section.

7.5 Further (and future) research

The earlier recommendations of Crampton et al. (2013), that OSN research should move *beyond the geotag*, have been adopted here. In doing so, this study has examined the production of explicitly spatial and geographically referenced material in OSN interactions and drawn some important conclusions regarding differences between coordinate-geotagging and non-coordinate-geotagging users. Through a detailed examination, not only of geographical coordinates but of heavily linked websites and toponymic mentions of place in message text and linked/shared content, the results go some way towards meeting the call for a 'critical digital social and media research' recently advocated by Fuchs (2017a).

Fuchs (2017a, p43) has warned that 'Big data analytics' positivism [...] fails to understand users' motivations, experiences, interpretations, norms and values' and suggests that 'We do not just have to understand what people do on the Internet but also why they do it, what the broader implications are, and how power structures frame and shape online activities.' This should be a primary focus for any future research. It has become clear during the course of this study that simply analysing the 'digital trace data' of OSN interactions made on Twitter or Facebook does not provide sufficient insight into the motivations behind individuals' production or consumption of social media content, whether (or why) this content contains implicit geographical or explicit spatial references, or not.

Tasse et al. (2017, p1) have noted that 'there is currently little understanding of what people geotag on [...] popular social media sites not centered around location [including Facebook, Instagram, Twitter, Snapchat, and Flickr], and why.' While suggesting that 'people geotag consciously and intentionally, they geotag in uncommon places, they primarily do so to communicate and show where they've been, and they geotag soon after being at the place', Tasse et al.'s (2017) study usefully draws conclusions by mixing large-scale data analysis with a qualitative 'free-response' survey drawn from 4,119 'prolific [geotagging] tweeters'. Ethical decisions made during the course of this research (Section 3.4, p111; Appendix 4, p419), together with limited funding availability, precluded such a mixed approach here. However, the integration of 'Big Data' analysis with 'traditional methods' should be encouraged. As Fuchs (2017a, p43) has noted, 'Digital methods do not outdate but require traditional methods in order to avoid the pitfall of digital positivism. Traditional sociological methods, such as semi-structured interviews, participant observation, surveys, content and critical discourse analysis, focus groups, experiments, creative methods, participatory action research, statistical analysis of secondary data and so on, have not lost importance.' Other suggested areas for further or future research are more narrowly focused on specific problems encountered here. These may be grouped into three broad themes:

1. **Localness** – A key assumption of many investigations using spatially referenced, or referenceable, OSN data is the premise that interactions have been deposited locally, that ‘a unit of social media VGI always represents the perspective or experience of a person who is local to the region of the corresponding geotag’ (I. L. Johnson et al., 2016, p515). This, of course, is not always the case. In the two case study data sets examined here the most prolific coordinate-geotagging user, Mulder1981, later ‘unmasked’ by *The Herald* (2017) newspaper as a Scottish Tory Councillor and ‘influential BritNat Twitter troll who boasts about his manhood online’, posted from 2,503 locations, largely in Scotland but from as far afield as Turkey and the West Coast of America. There are 2,353,010 non-coordinate-geotagging and just 122,253 coordinate-geotagging users in the research data corpus, a figure and percentage (4.94%) which is elevated by the exclusively spatially sampled US2012_GEO Stream (Section 4.2.4.1, p126), without which there would be only 51,475 coordinate-geotagging users (2.19%) across the two events. Amongst all coordinate-geotagging users, in both case study events, 82,771 have posted from only one location and 39,482 from multiple locations, the median number of geotagged posts being 2 in this latter group. Research published by I. L. Johnson et al. (2016) has confirmed that it is not always correct to assume that geotagging users are spatially proximal to ‘home’ locations, with a rate of 75% (lower in rural areas) suggested. When most geotagging users make only one post, or a small number of posts, during reasonably extended periods of data collection (2 months or more in both of the case studies examined in this research) imputing home locations with accuracy is difficult. If ~25% of geotaggers are not at ‘home’ locations when making their posts, and coordinate-geotagged interactions, as shown here, make fewer mentions of geographical entities in text or link shares than non-coordinate-geotagged interactions, then further difficulties arise. The problem of lower NLP-detectable toponymic geographicality amongst the most spatialised of coordinate-geotagging users, reported here

(Chapter 5, p186), suggests that yet more work on ‘localness’ and ‘locational meaning’ in space or place-based OSN data is required.

2. **Geoparsing** – Earlier studies into the difficulties of successfully geoparsing microblog text (e.g., Gelernter & Mushegian, 2011; Shi & Barker, 2011) have been extended in many directions (e.g., Kordopatis-Zilos, Papadopoulos, & Kompatsiaris, 2017; Poulston et al., 2017; Purves et al., 2018) and usefully summarised by Gritta et al. (2018). Despite the development of a multiplicity of techniques aimed at extracting spatial meaning from text the challenge remains considerable; ‘dynamic’ names exist, local slang may be used and ambiguity in place naming creates difficulties in gazetteer-based approaches (Kordopatis-Zilos et al., 2017). Language-modelling techniques, based on NLP and machine learning, face other problems; the need for software able to detect possibly geographical terms in short or ungrammatical text and the availability of training datasets for use in probabilistic modelling. Other approaches, e.g., using mentions of local celebrities in users’ message text (Ebrahimi et al., 2017), have claimed some success but overall the picture remains somewhat confused. Crucially, and despite the efforts of several leading researchers (e.g., Smart, Jones, & Twaroch, 2010; Wei Zhang & Gelernter, 2014), it remains difficult to access or run the most up-to-date geoparsing systems, many of which have been developed against ‘laboratory data and unlike in wider NLP are often not cross-compared’ (Gritta et al., 2018, p603). Further advances in text-based geoparsing are inevitable, as this is a particularly active research area. However, unless teams publish and document their code comprehensively many non-specialists will continue to use older-generation geoparsing systems simply because they work dependably, even if their results may not be quite as good as the latest ‘bleeding edge’ technologies.
3. **Ethics** – Astonishingly, 2,436,167 individuals have unwittingly contributed to this research. It would have been unthinkable to consider using such a vast pool of respondents to survey research in a doctoral thesis twenty years

ago, and probably reasonably implausible ten years later. Large, individual-level data sets of this type (even if sparsely populated, Section 6.4.3, p255) have historically been the preserve of government departments or large and well-funded research agencies. Such data sets generally arose either from infrequent questioning (e.g., the decennial Census) or were created as by-products of administrative processes, e.g., crime or health records, subsequently aggregated to wider areal units to avoid the unwanted identification of individuals. The OSN interactions considered here, however, do not come in response to a questionnaire or more traditional fieldwork approach requiring informed consent (Section 3.4, p111). Researchers, and others, may now analyse huge numbers of social media messages available continually via platform operators' APIs or data aggregation companies. Messages and downloadable metadata bundles *may* have been created by users who are comfortable posting in the public domain or, more worryingly, by those who *may not* have adjusted their personal privacy settings sufficiently to prevent inadvertent public posting of their material (Woodfield, Morrell, Metzler, & Blank, 2013). Ethical issues surrounding the use of social media data in research have been discussed at University conferences (Sugira et al., 2016) and published in both the sociological (Boyd & Crawford, 2012; Halavais, 2015; Williams, 2015) and medical literature (S. Golder, Ahmed, Norman, & Booth, 2017), where ethical good practice is considered paramount. A general conclusion is that it 'remains unlikely that a consensus on the ethical considerations on using social media research will ever be reached. Each Internet research project requires an individual assessment of its ethical issues and selection of the most appropriate methodological approach' (S. Golder et al., 2017, p13). A similarly pragmatic ethical approach has been adopted in this study (Section 3.4, p111; Appendix 4, p419). Individuals, however, may be surprised by the amount of 'digital trace data' they deposit online, and no more so than in the case of geographical data where, for example, Tasse et al. (2017) have

reported that some ‘prolific geotaggers’ were completely ‘unaware that they were posting geotagged tweets’. Locational privacy appears to be highly-valued by Web and social media users in a world of spontaneous, and usually anonymous, commentary (Cottrill, 2011; Tsou, 2015). As increasingly sophisticated geoparsing approaches now attempt to locate individuals using techniques bordering on surveillance (e.g., all of Tony’s friends asked ‘How was the Hemingford Arms last night?’ and most mentioned other locations in North London; therefore, Tony probably lives in Islington), spatio-ethical policies will also require further research.

In addition to the above, several authors (Hutton & Henderson, 2018; Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017) have expressed concern regarding the reproducibility of OSN research; their arguments centring on data size and availability and continued access to the often specialised and frequently updated methods and systems employed in analysis (Zelenkauskaitė & Bucy, 2016). While many key workings, programmes and SQL queries are detailed in this thesis they cannot all be reproduced. The Oracle 12c database, backed up in binary format using `EXPDP` (Dietrich, 2014), requires ~150GB of storage. Six virtual machines used in production take another 150GB. Nearly 100 Oracle SQL scripts have been saved, and many of these incorporate multiple individual SQL statements. A Tableau repository of 233MB holds over 40 workbooks and references further sets of Microsoft Excel spreadsheets, database views and queries. QGIS map files and other data require an additional 833MB of storage. Iterative workings saved along the way, including raw data, database dumps and NLP/geoparsing output total a further ~1TB. Around ~8m rows of OSN data stored in ~50GB have created a ~2TB project. Some code developed in this research works now, e.g., the bespoke Ruby scripts used to call the Cloud-hosted AlchemyAPI system, requiring a software key provided specifically for this project, but may become obsolete in the future. The Twitter and Facebook OSN data themselves should not, according to the terms of the licence agreements adhered to in this

research (Appendix 5, p424), be passed on to third parties in their entirety but may be regenerated from source (at some greater cost, since the data are now historic) using unique identifiers embedded in interaction metadata. Difficulties in transferring, storing and re-running analyses to achieve reproducible results in social media Big Data research are a particular concern (Hutton & Henderson, 2018; Kinder-Kurlanda et al., 2017) and should be addressed in future work. At a practical level, solutions could include the use of ‘digital preservation’ techniques (Maemura, Moles, & Becker, 2017) as recently demonstrated by the European Union’s E-ARK system (Thirifays et al., 2018) for scientific computational archiving.

7.6 Contributions to knowledge

This research makes contributions to knowledge in three main areas, summarised in numbered lists, in turn, below.

7.6.1 Technological contributions

1. The work detailed in this thesis demonstrates that one, reasonably technically competent, researcher in geography can integrate and combine several complex software systems and pipelines to store and analyse large numbers of social media interactions. Many social media Big Data projects are the preserve of Departments of Computer Science. However, Geography Departments, and the staff that work in them, have much to offer in social media subject areas and should not feel discouraged from pursuing such investigations.
2. Many social media geo-investigatory Big Data projects also, typically, use data sourced from just one platform, Twitter, even though it is far from the largest online social network. Facebook, which is, has not been so widely-studied in the literature (Stock, 2018). This research examines coordinate spatiality, and compares and contrasts NLP-detectable geographicality, in OSN interactions sourced from both Twitter *and* Facebook platforms.

- a. The longer messages allowed on Facebook yield ~4 NLP-detectable toponymic mentions/interaction against ~1 in every Twitter tweet. Although increasingly difficult to access (Hogan, 2018), Facebook posts offer greater geographical value, once geoparsed, than Twitter tweets. This situation may change over time following Twitter's newly increased limit of 280 characters/tweet; up from the 140 character limit used since 2006 (Jackson, 2017).
3. Three separate NLP/geoparsing systems have been used in this research:
 - a. Two, TwitIE on GATEcloud and CLAVIN-rest, produce broadly comparable results against interaction message text;
 - b. One, AlchemyAPI, is particularly well-suited to Information Extraction of linked/shared URL content.
 - c. Code listings in this thesis show how these systems may be used and how to mine the augmented data they produce.
 - d. Much additional technical information, including working virtual machines etc., is available upon request.
4. Collaborative work with researchers at the University of Sheffield has enhanced the functionality of GATEcloud's data ingestion system:
 - a. Researchers may now easily retain unique identifiers in input and output cycles on GATEcloud;
 - b. Accuracy of results will be improved through the simplification of database joins enabled by this development.
5. Data-mining and sparsity analysis in SQL shows that Big Data are far from complete, with many `NULL` values observed. This finding, and the identification of well-populated fields in OSN data, should help to guide future research.

7.6.2 Substantive contributions

1. Published literature reviews show that few studies have examined the interplay between geography and politics in online social media (Steiger, de Albuquerque,

et al., 2015). Recent events suggest that this situation is bound change.

Evidence of Russian state-sponsored interference in the 2016 US Presidential election and the Facebook / Cambridge Analytica scandal, which broke in 2018, will require much new research into online geo-behavioural targeting. By evaluating the potential for geographically tracking politicised discourse over social media networks, using data sampled during two earlier electoral events, this research a) demonstrates significant contemporary relevance, and; b) may be used to guide further research and/or help formulate future policy responses aimed at safeguarding 'free and fair' democratic processes.

2. The calculation of median values, from ~8 million records, shows that coordinate-geotagging users of Facebook are 'liked' slightly less (1.33 vs. 150) than corresponding non-coordinate-geotagging users. On Twitter, coordinate-geotagging users have fewer 'friends' (325 vs. 345) and fewer 'followers' (275 vs. 348) than non-coordinate-geotagging users of the platform.
3. Calculating maximum values from ~8 million records, the top 10 non-coordinate-geotagging users on Twitter posting during the two case study events have a following, typically, one order of magnitude or more greater than that of the corresponding top 10 coordinate-geotagging Twitter users.
4. Few users, on either of the two OSN platforms investigated during the two political case study events, coordinate-geotagged their social media messages and most coordinate-geotagging users made only one coordinate-geotagged interaction in the sampled data sets.
5. Analysis of the case study interactions, using three NLP/geoparsing systems, data-mining in SQL and statistical analysis in R, demonstrates that highly-spatialised, coordinate-geotagging, users of OSNs are not always the most geographically expressive:
 - a. Coordinate-geotagging users make fewer references to place in their message text than other users;
 - b. Coordinate-geotagging users link to and share external content containing fewer mentions of place in text, and;

- c. Coordinate-geotagging users, overall, make far fewer links to shared URL content than other users of these platforms.
- 6. In addition:
 - a. Fusion of coordinate-geotagged ONS interactions to US and UK Census data suggests that users posting spatially are likely to originate from areas with a generally more youthful and urban population profile.
 - b. In the US, geodemographic profiling suggests that OSN users may well originate from areas with a high proportion of non-institutionalised population, e.g., student halls of residence or nursing accommodation etc.
 - c. Analysis of patterns of geographical retweeting suggests that the re-posting of social media content may be comparatively localised during electoral events, particularly in the longer-running case of the 2014 Scottish Independence Referendum.
- 7. Taken together, the findings above imply that tracking or mapping the spread of places, news, views or opinion by searching for phrases or toponyms in message text created by coordinate-geotagging users alone, or searching for specific URL links shared alongside these messages, does not provide an adequate proxy for tracking the geographical spread of all politically discursive material created and shared online over social media networks.

7.6.3 Policy contributions

- 1. Senior politicians (e.g., US Senator Mark Warner, quoted in Charter's, 2018, report for *The Times*), and even an anonymous Guest Contributor (2015) to *Adweek* magazine, have both highlighted the unregulated 'Wild West' situation in social media advertising; something the senior politician expects will soon be 'coming to an end' and the advertising figure has said needs 'taming'. Writing in 2015 the Guest Contributor noted that while 'all other forms of advertising have strict metrics and accountability [s]ocial is such a new space that no one has

taken the initiative to create a system for it to thrive alongside the more traditional advertising mediums.’ Advertising ‘flight schedules’ have historically allowed regulators to track political marketing (and spending) on ‘traditional’ media, including television, radio and newspaper press. ‘Algorithmic advertising’ on Web or social media channels makes this harder (Eslami, Krishna Kumaran, Sandvig, & Karahalios, 2018) as advertisements, including political marketing, may only be shown to individual users if certain behavioural or geographical conditions are met. Electoral regulators in the US, UK and elsewhere will need to gain more understanding of how online or social media advertisements are targeted. Regulation or legislation may well be required to compel political advertisers, website or social media platform operators, to release this information.

2. Geographically, as this research has demonstrated, it is difficult to track the spread of political opinion, (mis)information or campaign advertising disseminated over online social media channels. As content may be targeted at *areas* as much as at individuals it is desirable to know, with a reasonable level of spatial precision, where content is being consumed or shared.
3. This research recommends encoding a lower-resolution Latitude and Longitude coordinate pair alongside all social media interaction metadata or, e.g., a lookup to a uniquely identifiable 1x1km grid square. Doing so would enable reasonably granular tracking of the geographical diffusion of online social media content without adversely affecting personal locational privacy. Availability of this additional data point could, e.g., be restricted to electoral regulators and/or accredited researchers, and would enable geographical analysis and mapping of patterns of online social media information consumption in much the same way that traditional ‘flight schedules’ allow the regional tracking of television, radio or newspaper advertising.

It is hoped, particularly in the latter category, that the policy recommendations offered alongside this research work (Section 6.3, p238) may prove useful when

considering regulatory, legal or technical responses aimed at ending the 'Wild West' phase of social media's early history. No-one, this author included, wants to live in a democracy subverted by power-hungry politicians or external agents; seeking, being promoted to, or winning office on false premises. Now we know just how much, collectively, we may have been manipulated by Cambridge Analytica, Russian state-sponsored agents or party campaign teams we must, collectively, devise better systems to track online geo-behaviourally targeted political material both for the good of society and to ensure future, free and fair democratic elections.

7.7 Contributions to debate

Attempts to 'segment' voters and micro-target political communications during electoral campaigns are not new. Blaemire (2018) has outlined a chronology in which messaging units staffed by 'backroom' politicians have been replaced by 'campaign professionals' using computerised databases, geodemographic targeting and, most recently, Internet technologies; the latter presenting 'enormous opportunity' for campaign teams. The shift to geodemographically targeted voter engagement occurred in the late 1990s and early 2000s in the UK, lagging earlier developments in the US. The UK's two major political parties, Conservative and Labour, have both made extensive and largely uncontroversial use of Experian's (2018) MOSAIC discriminator for well over a decade. The Labour Party's (2018, p14) *Campaigners' Handbook* states that MOSAIC is used to create 'broad segmentations of the electorate depending on electors life situations and a range of other modelled scores which are produced by the Targeting and Analysis team.' These scores are used to predict the likelihood of localised voter turnout, identify areas of policy interest and determine 'optimal times' for canvassing. The Conservatives' use of MOSAIC data are not recorded in a publicly accessible document but Mark Wallace (2015), writing on the Conservative Home website, has detailed the party's use of broadly similar geodemographic segmentation and computerised campaigning techniques during the 2010 and 2015 UK General Elections.

Elsewhere, in *The Guardian*, Sabbagh (2018) has described how MOSAIC's geodemographic classification scheme, segmenting UK population into 'metro high flyers', 'classic grandparents', 'disconnected youth' and other classes at postcode level, has been used by the Conservatives to help deliver targeted mailshots in more recent campaigns; a practice common to all main UK political parties.

Why then, as a greater proportion of campaign spending moves online (Pew Research Center, 2018; The Electoral Commission, 2018a), does the sort of targeting long-employed offline by political parties in the UK and US appear to prompt such concern when delivered over the Internet? Is online political campaigning so different, so advanced from previous generations of computerised targeting and messaging, that it truly represents a new *threat* to democracy? The following pages, comprising the penultimate section of this thesis, discuss these issues; contributing to an ongoing and, as yet, inconclusive debate surrounding the use of geo-behavioural micro-targeting in political campaigning.

Dommett & Temple (2018), citing earlier work by Gibson (2015), have noted that 'parties have become heavily dependent on digital technology' through the use of 'email, party websites, social media, online videos and gamification.' As the Labour Party's (2018, p14) *Campaigners' Handbook* proves, the use of digital does not comprise geodemographic targeting alone. An entire digital infrastructure is offered to local Labour party activists, including online print services, downloadable voter registration cards, templates for survey research, branded posters and stickers, ward level voter analysis and much more; all available from a dedicated Campaign Creator website. Labour's sophisticated campaign tools have been credited (Sabbagh, 2018) with boosting the party's performance in the 2017 UK General Election called, unexpectedly, by Conservative Prime Minister Theresa May. In both this, and the earlier 2016 UK EU Membership ('Brexit') Referendum, Sabbagh reports, the Conservatives' online efforts were outclassed by Labour's, despite significant £1.2 million spending made on Facebook by the Tories earlier during the

2015 UK General Election (Moore, 2016). While Sir John Holmes, writing in the forward to the UK Electoral Commission's (2018a) report, *Digital campaigning: Increasing transparency for voters*, correctly identifies 'an explosion in the use of digital tools in political campaigning' his main concern, and that of the Commission he chairs, is combatting 'serious allegations of misinformation, misuse of personal data, and overseas interference' that have recently emerged; in the UK, principally, during the Brexit Referendum and, in the US, during the 2016 US Presidential Election. The ethical use of data and messaging, together with accurate advertising attribution, are clearly key requirements in fair online electioneering (Section 6.3.1, p238). However, recent expressions of 'outrage' and upset in broadsheet newspaper commentary and, to some extent, academia may also reflect, at least in part, attitudinal discontent or discomfort with the victors in these two elections (W. Davies, 2018). Neither Leave nor Trump were widely popular 'establishment' choices, leaving aside any concerns more directly associated with the ethical or unethical modes of campaigning that brought Brexit or 'The Donald' to victory.

Dominic Cummings, architect of the successful Vote Leave campaign during the 2016 UK European Union Membership Referendum, has reportedly stated (Sabbagh, 2018) that 'It is hard to change people's minds. We are evolved creatures. If we were all dopey dupes we wouldn't be here, our ancestors would have all been killed.' Nonetheless, the 'Brexit' campaign he designed, according to his description of its genesis given at advertiser Ogilvy & Mather's Nudgestock 2017 event (YouTube, 2018b), was cleverly calculated both to play on electors' deep-seated emotions – the use of the 'Take back control' phrase was apparently intended to trigger feelings of 'loss aversion' – and to exploit data distributions, e.g., a non-normal 'third, a third and a fifth.' Here, Cummings has argued, one third of voters 'said the EU is rubbish and I want out and I'm not bothered about [and] not frightened of the consequences', a second third 'said the EU is a positive force and I definitely like it and I definitely want to be in and I'm definitely going to vote to stay in' and a fifth of voters 'said the EU is rubbish I would like to be out of it but

getting out is scary and I'll probably end up voting to stay in.' The Vote Leave campaign, designed quickly by 'a team of physicists who essentially looked at campaigning from complete first principles' (Dominic Cummings, recorded on YouTube, 2018b), was intended to discourage the first third from voting, ensure the second third turned out to vote Leave and to 'persuade enough of that fraction of the fifth not to be frightened and to [turn out and] vote with us.' While the small team of physicists and young online advertising experts employed by Vote Leave certainly used demographics and geographical targeting the key breakthrough, according to Cummings, involved understanding how Facebook's Likes and Interests data could be harnessed and played back to target particular groups of voters. The campaign was clever, but its unpopularity may owe more to inaccuracies in messaging (e.g., the £350m/week bus-side NHS funding 'pledge') and the outcome of the result rather than its detailed mode of operation. The fact Cummings' Vote Leave campaign was sparked off by a seemingly undemocratic set of 'phone calls from a combination of a few MPs and some campaigners and a few Tory party donor billionaires' has also done little to cement its place in popular affection.

Similar opprobrium has been directed at Cambridge Analytica, and Canadian associate AggregateIQ, in their support and work for elements of the Leave side during the 2016 Brexit Referendum and for the candidacy of Donald J. Trump during the 2016 US Presidential Election (Cadwalladr, 2018b; Cadwalladr & Graham-Harrison, 2018; Frenkel et al., 2018). Part-owned by a right-wing American hedge-fund billionaire (Robert Mercer), and with close ties to campaign manager Steve Bannon, Cambridge Analytica is supposed to have worked both for the Leave.eu and Trump election campaigns (Osborne, 2017). If Cummings' work for Vote Leave mainly drew criticism for lack of attribution and misinformational messaging, Cambridge Analytica's involvement in both campaigns proved controversial for two reasons. Firstly, because data used in campaigning appeared to have been misappropriated from Facebook and its users. Secondly, because the extent of the geo-psychologically targeted messaging laid bare the possibilities offered by Big

Data driven politicking (Section 1.1, p1 and Section 6.3.1, p238). The ‘harvesting’ of many millions of Facebook users’ personal details by Aleksandr Kogan’s personality quiz app, *Thisisyourdigitallife*, certainly constituted a major breach of trust and has led to the closure of Cambridge Analytica (BBC News, 2018a) and ongoing problems for Facebook, including government inquiries (U.S. House of Representatives, 2018b) and substantial fines from data protection regulators (BBC News, 2018d; Embury-Dennis, 2018). While sophisticated the campaigns delivered by Cambridge Analytica do not, however, appear to have diverged hugely – except, perhaps, in extent – from other similar attempts to influence public opinion made online during 2016 US and UK elections, several earlier and other overseas events including Obama’s 2012 US Presidential Election campaign and the 2014 Scottish Independence Referendum, examined here. Clever analysis and the use of Facebook, and its campaign management tools, are not unique to Cambridge Analytica or to Trump or to Vote Leave. Dommett & Temple (2018) have suggested that Facebook advertising is ‘the new normal’ in political campaigning and its use is certainly not confined only to the UK or US.

W. Davies (2018), in the *London Review of Books*, has expressed doubts over Cambridge Analytica’s self-proclaimed successes in changing political opinion, stating that ‘Cambridge Analytica looks conveniently like a smoking gun, primarily because it has repeatedly bragged that it is one.’ Others have used targeted political communications previously, and the question as to whether geographical targeting of marginal seats – as opposed to targeting of specific types of voters – is an intriguing one. In ‘first past the post’ electoral systems, such as that used in Britain and many countries with historical British ties, winning marginal seats matters enormously. In referenda, single transferable vote or additional member systems (the last two used in Scottish local and parliamentary elections, respectively), winning vote share is most important. Johnston & Pattie (2006) maintain that voting behaviours, and electoral results, arise from an interplay of socio-spatial factors operating at a range of scales; from the family home to the

workplace, local town or city through to the region and the state. In the Preface to *Putting Voters in their Place*, Johnston & Pattie (2006, pVII) state that their ‘perspective – basically, highlighting the importance of space and, especially, place in the understanding of voting and the operation of electoral systems – has become part of the contemporary discourse of UK electoral studies, rather than a separately identifiable sub-sub-discipline.’ Place, therefore, widely mentioned in the Facebook and Twitter interaction message text and linked/shared URL content analysed here (Chapter 5, p186), may be expected to have special significance in electoral campaigning; even if modern online campaigns are now more likely to be predicated on a mixture of ‘demographic data such as age, postcode, religion, and gender, combined with indicators of users’ interests’ (Dommett & Temple, 2018, p190). Such modern campaigns may target ‘swing’ constituencies or, as Cummings’ work for Vote Leave shows, ‘swing’ voters, who may or may not live and vote in the most marginal of constituencies. As targeting methods enabled by Facebook and other OSNs have developed, it is clear that the social does indeed – as Johnston & Pattie have argued – interact with the spatial. Targeting highly-educated voters would effectively target London and the South East in the UK, while targeting unemployed voters would favour the de-industrialised areas of Northern England and parts of Wales. Conversely, targeting London, the South East, Northern England or Wales would, of course, also effectively target voter groups with different educational, un/employment and other characteristics as the distribution of socio-economic traits (let alone Likes or Interests expressed on Facebook) are not uniformly distributed geographically throughout the country. While both social and spatial targeting are easily enabled on Facebook, Instagram etc. (Section 1.2, p11) it is currently much harder to determine whether political advertising based on socio-spatial targeting actually alters electoral outcomes or, using the available data, to determine how such material is promoted. Dimitrova & Matthes (2018, p333) note that while ‘social media have clearly affected our understanding of political communication and its effects on the public, it is difficult to see clear monolithic effects’ partly because ‘comprehensive aggregate studies offer evidence

that the effects of social media consumption and use are hardly uniform across different contexts and groups.’ Those in politics who spend significant sums on social media advertising might well disagree...

It has been suggested that the Cambridge Analytica ‘circus risks distracting from the real institutional and political questions, in this case concerning companies such as Facebook and the model of capitalism that tolerates, facilitates and even celebrates their extensive and sophisticated forms of data harvesting and analysis’ (W. Davies, 2018). Davies continues to argue that the major social media companies should no longer be allowed to totally exploit users’ personal data, suggesting that the thing they ‘fear most’ – anti-trust actions that would lead to their break-up, a suggestion also echoed by Reich (2018) – would be most effective in protecting privacy and preventing data-driven abuses. As Davies concludes, ‘Broken into smaller pieces, these companies would still be able to monitor us, but from disparate perspectives that couldn’t easily (or secretly) be joined up. Better a world full of snake-oil merchants like Cambridge Analytica, who eventually get caught out by their own bullshit, than a world of vast corporate monopolies such as Amazon and Facebook, gradually taking on the functions of government, while remaining eerily quiet about what they’re doing.’ Vote Leave campaign director, Dominic Cummings, has stated (YouTube, 2018b) that ‘most communications companies’, traditionally called upon by political parties to run campaigns, ‘are populated by bullshitting charlatans’ and the future of electoral campaigning will be driven by ‘experimental psychology and data scientists.’ He is probably correct. As this thesis has shown, however, tracking the downstream diffusion of geo-behaviourally micro-targeted communications is not straightforward. While Facebook (2018c) has recently announced ‘a new initiative to help provide independent, credible research about the role of social media in elections, as well as democracy more generally’ a new debate must now open; how much individual privacy must needfully be sacrificed to allow regulators sufficient oversight of digital campaigning to prevent democratic abuses? The data users gift to Facebook and other OSNs has been used by political marketers to

target them, possibly leading to manipulation of electoral results and changing policies (such as leaving the European Union) that will have long-lasting and far-reaching repercussions for societies. Are these users, or the platform operators themselves, also prepared to allow electoral officials, government agencies, researchers or others fuller access to these data for monitoring, even if in anonymised or aggregated form, to understand how this is being done?

7.8 Summary

There are no accepted 'gold standards' for OSN research. Scholars from many separate disciplines have adopted different methodological designs and technological approaches to filter huge volumes of social media data in search of meaning. Some of these data-driven approaches have been criticised for an over-reliance on computational methods at the expense of theoretical reasoning. The approach followed in this research, a hybrid exploratory-case study methodology, has allowed hypothesis testing to determine whether coordinate-geotagging users behave similarly, or differently, to non-coordinate-geotagging users of Facebook and Twitter social media platforms during pre-electoral periods.

While OSN coordinate-geotags have been widely studied for around ten years comparatively little is known about the online posting behaviour of the small percentage of OSN users who geotag their social media messages (Rzeszewski & Beluch, 2017). This thesis makes an original contribution to knowledge by showing that coordinate-geotagging users a) make fewer references to place in their messages, b) link to articles making fewer mentions of place in their text, and c) make fewer links to third-party material than other, non-coordinate-geotagging, users of OSN sites. In these respects, and in the two case study events under examination, geotagging users cannot be considered representative of all OSN users. Future research should test these conclusions against current data. As coordinate-geotagged OSN interactions are used in so many different application domains it is important both to reconfirm the validity and currency of these results

and to understand their ongoing relevance. In political contexts, however, it is clearly not enough simply to follow the spread of a word, toponym or URL link share in social media and to map geographical distributions by plotting the coordinates of those who have chosen to share that term or URL in geotagged messages; the results of such an exercise would be both unrepresentative and, therefore, in all likelihood, inaccurate.

REFERENCES

- Abbasi, M. A., Zafarani, R., Tang, J., & Liu, H. (2014). Am I More Similar to My Followers or Followees? Analyzing Homophily Effect in Directed Social Networks. In *Proceedings of the 25th ACM conference on Hypertext and social media - HT '14* (pp. 200–205). New York, New York, USA: ACM Press.
<https://doi.org/10.1145/2631775.2631828>
- Agarwal, R., & Dhar, V. (2014). Editorial —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, 25(3), 443–448. <https://doi.org/10.1287/isre.2014.0546>
- Agarwal, R., & Prasad, J. (2000). A field study of the adoption of software process innovations by information systems professionals. *IEEE Transactions on Engineering Management*, 47(3), 295–308. <https://doi.org/10.1109/17.865899>
- Agence France-Presse. (2016). Populist surge on several continents. Retrieved November 13, 2016, from <https://www.afp.com/en/news/23/populist-surge-several-continents>
- Agnew, J. A. (2011). Space and Place. In J. A. Agnew & D. N. Livingstone (Eds.), *Sage Handbook of Geographical Knowledge*. London: SAGE Publications.
- Agnew, J. A. (2013). Territory, Politics, Governance. *Territory, Politics, Governance*, 1(1), 1–4. <https://doi.org/10.1080/21622671.2013.765754>
- Agnew, J. A. (2014). *Place and Politics: The Geographical Mediation of State and Society*. Routledge.
- Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, 41(6), 855–864.
<https://doi.org/10.1177/0165551515602847>

- Ajao, O., P, D., & Hong, J. (2017). Location Inference from Tweets using Grid-based Classification. *ArXiv*, (2013). Retrieved from <http://arxiv.org/abs/1701.03855>
- Aladwani, A. M. (2015). Facilitators, characteristics, and impacts of Twitter use: Theoretical analysis and empirical illustration. *International Journal of Information Management*, 35(1), 15–25.
<https://doi.org/10.1016/j.ijinfomgt.2014.09.003>
- Albright, J. (2017). Cambridge Analytica: the Geotargeting and Emotional Data Mining Scripts. Retrieved March 19, 2018, from <https://medium.com/tow-center/cambridge-analytica-the-geotargeting-and-emotional-data-mining-scripts-bcc3c428d77f>
- Alex, B., Byrne, K., Grover, C., & Tobin, R. (2014). A web-based geo-resolution annotation and evaluation tool. *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII), COLING 2014*, 59–63. Retrieved from <http://aclweb.org/anthology/W14-4908>
- Alexander, P. A. (1992). Domain Knowledge: Evolving Themes and Emerging Concerns. *Educational Psychologist*, 27(1), 33–51.
https://doi.org/10.1207/s15326985ep2701_4
- Allen, M. (2017). Web 2.0: An Argument Against Convergence. In S. Sparviero, C. Peil, & G. Balbi (Eds.), *Media Convergence and Deconvergence* (pp. 177–196). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-51289-1_9
- Alonso-Lorenzo, J., Costa-Montenegro, E., & Fernandez-Gavilanes, M. (2016). Language independent big-data system for the prediction of user location on Twitter. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 2437–2446. <https://doi.org/10.1109/BigData.2016.7840880>
- Althaus, S. L., & Tewksbury, D. (2000). Patterns of Internet and Traditional News

- Media Use in a Networked Community. *Political Communication*, 17(1), 21–45.
<https://doi.org/10.1080/105846000198495>
- An, L., Tsou, M.-H., Crook, S. E. S., Chun, Y., Spitzberg, B., Gawron, J. M., & Gupta, D. K. (2015). Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions. *Annals of the Association of American Geographers*, 105(5), 891–914. <https://doi.org/10.1080/00045608.2015.1064510>
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 16(7), 7–16.
- Andersson, H. (2018). Social media apps are “deliberately” addictive to users. Retrieved July 4, 2018, from <https://www.bbc.co.uk/news/technology-44640959>
- Andreassen, C. S., Pallesen, S., & Griffiths, M. D. (2017). The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey. *Addictive Behaviors*, 64, 287–293.
<https://doi.org/10.1016/j.addbeh.2016.03.006>
- Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science and Engineering*, 15(3), 72–82.
<https://doi.org/10.1109/MCSE.2013.70>
- Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., ... Tanski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Space , Time and Visual Analytics*, 24(10), 1577–1600.
<https://doi.org/10.1080/13658816.2010.508043>
- Andrienko, G., Andrienko, N., Fuchs, G., & Wood, J. (2017). Revealing Patterns and Trends of Mass Mobility Through Spatial and Temporal Abstraction of Origin-Destination Movement Data. *IEEE Transactions on Visualization and Computer*

Graphics, 23(9), 2120–2136. <https://doi.org/10.1109/TVCG.2016.2616404>

Andrienko, G., Andrienko, N., & Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, 9(2), 38–46. <https://doi.org/10.1145/1345448.1345455>

Andrienko, N., Andrienko, G., Fuchs, G., Rinzivillo, S., & Betz, H.-D. (2015). Detection, tracking, and visualization of spatial event clusters for real time monitoring. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–10). IEEE. <https://doi.org/10.1109/DSAA.2015.7344880>

Andrienko, N., Andrienko, G., & Gatalisky, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6), 503–541. [https://doi.org/10.1016/S1045-926X\(03\)00046-6](https://doi.org/10.1016/S1045-926X(03)00046-6)

Apache Software Foundation. (2014). Apache Drill. Retrieved October 31, 2014, from <http://incubator.apache.org/drill/>

Apache Software Foundation. (2017). Maven – Welcome to Apache Maven. Retrieved June 10, 2017, from <https://maven.apache.org/>

Archer, K. (1995). A Folk Guide to Geography as a Holistic Science. *Journal of Geography*, 94(3), 404–411. <https://doi.org/10.1080/00221349508979343>

archive.today. (2017). `GitSampleCode/GeoLocation.py` at 85880cba53dd21372cb8ae96ae2ff21491eb8679 · MichaelPhillipsData/GitSampleCode · GitHub. Retrieved August 28, 2018, from <http://archive.is/pR9pj>

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49, 45–53. <https://doi.org/10.1016/j.apgeog.2013.09.012>

- Arthur, R., & Williams, H. (2017). Scaling laws in geo-located Twitter data. *ArXiv*. Retrieved from <http://arxiv.org/abs/1711.09700>
- Backstrom, L., Sun, E., & Marlow, C. (2010). Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *Proceedings of the 19th International Conference on World Wide Web*, 61–70. <https://doi.org/10.1145/1772690.1772698>
- Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons.
- Bahir, E., & Peled, A. (2013). Identifying and Tracking Major Events Using Geo-Social Networks. *Social Science Computer Review*, 31(4), 458–470. <https://doi.org/10.1177/0894439313483689>
- Barassi, V. (2016). Contested visions: Digital discourses as empty signifiers from the ‘network’ to ‘big data.’ *Communication and the Public*, 1(4), 423–435. <https://doi.org/10.1177/2057047316680220>
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10), 1531–1542. <https://doi.org/10.1177/0956797615594620>
- Barberá, P., & Rivero, G. (2015). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review*, 33(6), 712–729. <https://doi.org/10.1177/0894439314558836>
- Barr, R. (1985). Thematic mapping on microcomputers: the hardware and software environments. *Computers & Geosciences*, 11(3), 283–289. [https://doi.org/10.1016/0098-3004\(85\)90047-0](https://doi.org/10.1016/0098-3004(85)90047-0)
- Barr, R. (1996). A comparison of aspects of the US and UK censuses of population.

Transactions in GIS, 1(1), 49–60. <https://doi.org/10.1111/j.1467-9671.1996.tb00033.x>

Barreneche, C., & Wilken, R. (2015). Platform specificity and the politics of location data extraction. *European Journal of Cultural Studies*, 18(4–5), 497–513. <https://doi.org/10.1177/1367549415577386>

Bartlett, J., & Miller, C. (2013). *THE STATE OF THE ART: A LITERATURE REVIEW OF SOCIAL MEDIA INTELLIGENCE CAPABILITIES FOR COUNTER-TERRORISM*. London. Retrieved from http://www.demos.co.uk/files/DEMOS_Canada_paper.pdf

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. *ICWSM*, 8, 361–362. Retrieved from <https://gephi.org/publications/gephi-bastian-feb09.pdf>

Batista, F., & Figueira, Á. (2017). The Complementary Nature of Different NLP Toolkits for Named Entity Recognition in Social Media. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10423 LNAI, pp. 803–814). https://doi.org/10.1007/978-3-319-65340-2_65

Batty, M., Hudson-Smith, A., Milton, R., & Crooks, A. (2010). Map mashups, Web 2.0 and the GIS revolution. *Annals of GIS*, 16(1), 1–13. <https://doi.org/10.1080/19475681003700831>

BBC News. (2012). US election: Battleground states. Retrieved September 25, 2013, from <http://www.bbc.co.uk/news/world-us-canada-17306282>

BBC News. (2014). Facebook emotion experiment sparks criticism. Retrieved June 5, 2018, from <http://www.bbc.co.uk/news/technology-28051930>

BBC News. (2016). Can social media be used to predict election results? Retrieved February 15, 2017, from <http://www.bbc.co.uk/news/election-us-2016-37942842>

BBC News. (2017a). Facebook founding president Sean Parker sounds alarm. Retrieved November 9, 2017, from <http://www.bbc.co.uk/news/av/technology-41937476/facebook-founding-president-sean-parker-sounds-alarm>

BBC News. (2017b). Russia-linked posts “reached 126m Facebook users in US.” Retrieved November 6, 2017, from <http://www.bbc.co.uk/news/world-us-canada-41812369>

BBC News. (2017c). “Russian trolls” promoted California independence. Retrieved November 6, 2017, from <http://www.bbc.co.uk/news/blogs-trending-41853131>

BBC News. (2018a). Cambridge Analytica starts bankruptcy proceedings in US. Retrieved May 21, 2018, from <http://www.bbc.co.uk/news/technology-44167000>

BBC News. (2018b). Corbyn: Tech firm tax could fund journalism. Retrieved August 29, 2018, from <https://www.bbc.com/news/uk-politics-45271286>

BBC News. (2018c). Ex-GCHQ boss: “Social media firms sit above democracy.” Retrieved January 23, 2018, from <http://www.bbc.co.uk/news/av/uk-42780802/ex-gchq-boss-social-media-firms-sit-above-democracy>

BBC News. (2018d). Facebook fined £500,000 for Cambridge Analytica scandal. Retrieved October 25, 2018, from <https://www.bbc.co.uk/news/technology-45976300>

BBC News. (2018e). Facebook scandal “hit 87 million users.” Retrieved April 12,

2018, from <http://www.bbc.co.uk/news/technology-43649018>

BBC News. (2018f). New internet laws pledged as social media firms snub talks.
Retrieved September 4, 2018, from <https://www.bbc.co.uk/news/uk-politics-44188805>

BBC News. (2018g). Trump warns Google, Facebook and Twitter in row over bias.
Retrieved August 29, 2018, from <https://www.bbc.co.uk/news/technology-45331210>

BBC News. (2018h). Vote Leave spending investigation. Retrieved September 4, 2018, from <https://www.bbc.co.uk/news/topics/cpz72d0xmm7t/vote-leave-spending-investigation>

Berico-Technologies. (2014). Berico-Technologies/CLAVIN · GitHub. Retrieved October 31, 2014, from <https://github.com/Berico-Technologies/CLAVIN>

Berico-Technologies. (2017). About CLAVIN | CLAVIN Home. Retrieved June 10, 2017, from <https://clavin.bericotechnologies.com/about-clavin/>

Berico-Technologies. (2018). Try CLAVIN online | CLAVIN Home. Retrieved June 28, 2018, from <https://clavin.berico.us/clavin-web/>

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A Framework for Web Science. *Foundations and Trends in Web Science*, 1(1), 1–130. <https://doi.org/10.1561/18000000001>

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D. (2006). Creating a Science of the Web. *Science*, 313(5788), 769–771.

Bertin, J. (1967). *Semiology of Graphics: Diagrams, Networks, Maps*.

Bertino, E., & Matei, S. A. (2015). *Roles, Trust, and Reputation in Social Media Knowledge Markets*. (E. Bertino & S. A. Matei, Eds.). Cham: Springer

International Publishing. <https://doi.org/10.1007/978-3-319-05467-4>

Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A High Resolution Spatial and Temporal View. *ArXiv*, 1–12. Retrieved from <http://arxiv.org/abs/1308.5010>

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21(11). <https://doi.org/10.5210/fm.v21i11.7090>

Bessi, A., Scala, A., Rossi, L., Zhang, Q., & Quattrociocchi, W. (2014). The economy of attention in the age of (mis)information. *Journal of Trust Management*, 1(1), 12. <https://doi.org/10.1186/s40493-014-0012-y>

Bimber, B. (2014). Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the personalized political communication environment. *Journal of Information Technology & Politics*, 11(2), 130–150. <https://doi.org/10.1080/19331681.2014.895691>

Birkin, M., Malleson, N., Hudson-Smith, A., Gray, S., & Milton, R. (2011). Calibration of a spatial simulation model with volunteered geographical information. *International Journal of Geographical Information Science*, 25(8), 1221–1239. <https://doi.org/10.1080/13658816.2011.559169>

Blaemire, R. (2018). The Evolution of Microtargeting. In J. Thurber (Ed.), *Campaigns and Elections American Style* (4th ed., pp. 217–236). Routledge.

Blank, G., & Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61(7), 741–756. <https://doi.org/10.1177/0002764217717559>

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding

- of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 1–12. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Bond, R., & State, O. (2015). Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook. *American Political Science Review*, 109(1), 62–78. <https://doi.org/10.1017/S0003055414000525>
- Bonilla, Y., & Rosa, J. (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*, 42(1), 4–17. <https://doi.org/10.1111/amet.12112>
- Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 83–90). Hissar, Bulgaria. Retrieved from <http://derczynski.com/sheffield/papers/twitie-ranlp2013.pdf>
- Bontcheva, K., & Greenwood, M. A. (2014). GATE/Twitter related question.
- Bontcheva, K., & Rout, D. (2014). Making sense of social media streams through semantics: a survey. *Semantic Web*, 5(5), 373–403. <https://doi.org/10.3233/SW-130110>
- Borah, P. (2017). Emerging communication technology research: Theoretical and methodological variables in the last 16 years and future directions. *New Media & Society*, 19(4), 616–636. <https://doi.org/10.1177/1461444815621512>
- Borthakur, D., Rash, S., Schmidt, R., Aiyer, A., Gray, J., Sarma, J. Sen, ... Menon, A. (2011). Apache hadoop goes realtime at Facebook. In *Proceedings of the 2011 international conference on Management of data - SIGMOD '11* (p. 1071). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1989323.1989438>

- Bowcott, O. (2018, August 14). British expats in EU launch Brexit legal challenge | Politics | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/politics/2018/aug/14/british-expats-in-eu-launch-brexit-legal-challenge>
- Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>
- Bray, M. (2015). Ethical Review.
- Broich, J. (2017). 2017 isn't "1984" – it's stranger than Orwell imagined. Retrieved May 24, 2018, from <http://theconversation.com/2017-isnt-1984-its-stranger-than-orwell-imagined-71971>
- Buchanan, M. (2016). 'Liked', 'Shared', Re-tweeted: The Referendum Campaign on Social Media. In *Scotland's Referendum and the Media: National and International Perspectives* (pp. 70–82). Edinburgh University Press. Retrieved from <http://www.jstor.org/stable/10.3366/j.ctt1bgzd06.10>
- Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the Spread of Misinformation in Social Networks. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 665–674). New York, NY, USA: ACM. <https://doi.org/10.1145/1963405.1963499>
- Burkell, J., Fortier, A., Wong, L. (Lola) Y. C., & Simpson, J. L. (2014). Facebook: public space, or private space? *Information, Communication & Society*, 0(0), 1–12. <https://doi.org/10.1080/1369118X.2013.870591>
- Burns, R. (2018). Datafying Disaster: Institutional Framings of Data Production Following Superstorm Sandy. *Annals of the American Association of Geographers*, 108(2), 569–578. <https://doi.org/10.1080/24694452.2017.1402673>

- Burton, G. (2017). Sciama HPC Cluster. Retrieved June 4, 2017, from <http://www.sciama.icg.port.ac.uk/>
- Burton, I. (1963). The Quantitative Revolution and Theoretical Geography. *Canadian Geographer / Le Géographe Canadien*, 7(4), 151–162.
<https://doi.org/10.1111/j.1541-0064.1963.tb00796.x>
- Cadwalladr, C. (2017, May 7). The great British Brexit robbery: how our democracy was hijacked. *The Observer*. Retrieved from <https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy>
- Cadwalladr, C. (2018a, March 18). ‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower | News | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump>
- Cadwalladr, C. (2018b, March 31). AggregateIQ: the obscure Canadian tech firm and the Brexit data riddle | UK news | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/uk-news/2018/mar/31/aggregateiq-canadian-tech-brexit-data-riddle-cambridge-analytica>
- Cadwalladr, C., & Graham-Harrison, E. (2018, March 19). Facebook and Cambridge Analytica face mounting pressure over data scandal | News | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/news/2018/mar/18/cambridge-analytica-and-facebook-accused-of-misleading-mps-over-data-breach>
- Calhoun, C. J. (Ed.). (1992). *Habermas and the public sphere*. Cambridge, Mass.: MIT Press.
- Cambridge Analytica. (2018). CA Political - CA Political. Retrieved September 4, 2018, from <https://ca-political.com>

- Campbell, S. W., & Kwak, N. (2011). Political Involvement in “Mobilized” Society: The Interactive Relationships Among Mobile Communication, Network Characteristics, and Political Participation. *Journal of Communication*, 61(6), 1005–1024. <https://doi.org/10.1111/j.1460-2466.2011.01601.x>
- Castells, M. (1996). *The Rise of the Network Society, The Information Age: Economy, Society and Culture Vol. I*. Cambridge, MA Oxford, UK: Blackwell.
- Castells, M. (1997). *The Power of Identity, The Information Age: Economy, Society and Culture Vol. II*. Cambridge, MA; Oxford, UK: Blackwell.
- Castells, M. (1998). *End of Millennium, The Information Age: Economy, Society and Culture Vol. III*. Cambridge, MA; Oxford, UK: Blackwell.
- Castells, M. (2009). *Information Age Series : The Rise of the Network Society, With a New Preface : The Information Age: Economy, Society, and Culture Volume I (2)*. Hoboken, GB: Wiley-Blackwell. Retrieved from <http://site.ebrary.com/lib/portsmouth/docDetail.action?docID=10355273>
- CentOS. (2014). Download CentOS. Retrieved October 31, 2014, from <http://www.centos.org/download/>
- Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France. *New Media & Society*, 16(2), 340–358. <https://doi.org/10.1177/1461444813480466>
- Ceron, A., & Memoli, V. (2016). Flames and Debates: Do Social Media Affect Satisfaction with Democracy? *Social Indicators Research*, 126(1), 225–240. <https://doi.org/10.1007/s11205-015-0893-x>
- Chambers, S. (2003). Deliverative Democratic Theory. *Annual Review of Political Science*, 6(1), 307–326.

<https://doi.org/10.1146/annurev.polisci.6.121901.085538>

Chamley, C., Scaglione, A., & Li, L. (2013). Models for the diffusion of beliefs in social networks: An overview. *IEEE Signal Processing Magazine*, 30(3), 16–29. <https://doi.org/10.1109/MSP.2012.2234508>

Chang, C. W., & Chen, G. M. (2014). College students' disclosure of location-related information on Facebook. *Computers in Human Behavior*, 35, 33–38. <https://doi.org/10.1016/j.chb.2014.02.028>

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63, 67–80. <https://doi.org/10.1016/j.dss.2013.08.008>

Charter, D. (2018, September 6). A relentless threat to our democracy. *The Times*. Retrieved from <https://www.thetimes.co.uk/article/russian-style-cyberattackers-are-a-relentless-threat-to-western-democracy-say-facebook-and-twitter-35gtbmhwt>

Chen, H., Vasardani, M., & Winter, S. (2017). Geo-referencing Place from Everyday Natural Language Descriptions. *ArXiv*, V, 1–29. Retrieved from <http://arxiv.org/abs/1710.03346>

Cheng, G., & Du, Q. (2008). Ontology for Geographical Names management and retrieval. *Proceedings - 2008 International Symposium on Knowledge Acquisition and Modeling, KAM 2008*, 784–788. <https://doi.org/10.1109/KAM.2008.37>

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10* (pp. 759–768). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1871437.1871535>

- Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. *ArXiv Preprint ArXiv:1211.0938*. Retrieved from <http://arxiv.org/abs/1211.0938>
- Chu, K.-H., Wipfli, H., & Valente, T. W. (2013). Using Visualizations to Explore Network Dynamics. *Journal of Social Structure : JOSS*, 14. Retrieved from <http://www.cmu.edu/joss/content/articles/volume14/ChuWipfliValente.pdf>
- Cios, K. J., & Kurgan, L. a. (2006). Advances in Knowledge Discovery and Data Mining. *Data Mining and Knowledge Discovery*, 3918(Dm), 1–26. <https://doi.org/10.1007/11731139>
- Claburn, T. (2018). Twitter API overhaul threatens to seriously shaft apps... again. Retrieved June 1, 2018, from https://www.theregister.co.uk/2018/04/06/twitter_api_changes_threaten_to_bork_thirdparty_apps_again/
- Clark, J., & Jones, A. (2013). The great implications of spatialisation: Grounds for closer engagement between political geography and political science? *Geoforum*, 45, 305–314. <https://doi.org/10.1016/j.geoforum.2012.11.020>
- Clemm, J. (2015). A Brief History of Scaling LinkedIn. Retrieved February 1, 2018, from <https://engineering.linkedin.com/architecture/brief-history-scaling-linkedin>
- Cockayne, D. G. (2016). Affect and value in critical examinations of the production and 'prosumption of Big Data. *Big Data & Society*, 3(2), 1–11. <https://doi.org/10.1177/2053951716640566>
- Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6), 377–387. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=5221525&site=eds-live>

- Codd, E. F. (1972). RELATIONAL COMPLETENESS OF DATA BASE SUBLANGUAGES. *Computer*.
- Cohen, R. (2016, July 25). Trump and the End of Truth. *New York Times*. Retrieved from http://www.nytimes.com/2016/07/26/opinion/trump-and-the-end-of-truth.html?_r=1
- Collins, K. (2017). Google collects Android users' locations even when location services are disabled. Retrieved September 4, 2018, from <https://qz.com/1131515/google-collects-android-users-locations-even-when-location-services-are-disabled/>
- Collins, K. (2018). See Which Facebook Ads Russians Targeted to People Like You. Retrieved August 29, 2018, from <https://www.nytimes.com/interactive/2018/05/14/technology/facebook-ads-congress.html>
- Compton, R., Jurgens, D., & Allen, D. (2014). Geotagging one hundred million Twitter accounts with total variation minimization. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 393–401). IEEE. <https://doi.org/10.1109/BigData.2014.7004256>
- Cottrill, C. D. (2011). Location Privacy: Who Protects? *URISA Journal-Urban and Regional ...*, 23(2), 49–59. Retrieved from <http://ares.lids.mit.edu/fm/papers/Cottrill.URISA.pdf>
- Cox, K. R. (1969). The voting decision in a spatial context. In C. Board, R. J. Chorley, P. Haggett, & D. R. Stoddart (Eds.), *Progress in Geography, Vol. 1* (pp. 81–118). London: Edward Arnold.
- Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2), 130–

139. <https://doi.org/10.1080/15230406.2013.777137>

Cresswell, T. (2013). *Geographic Thought A Critical Introduction*. Wiley-Blackwell.

Cresswell, T. (2014). Deja vu all over again: Spatial science, quantitative revolutions and the culture of numbers. *Dialogues in Human Geography*, 4(1), 54–58.
<https://doi.org/10.1177/2043820614525715>

Croitoru, A., Crooks, A., Radzikowski, J., & Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, 27(12), 2483–2508.
<https://doi.org/10.1080/13658816.2013.825724>

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1), 124–147.
<https://doi.org/10.1111/j.1467-9671.2012.01359.x>

Crunchbase. (2018). Yatown | Crunchbase. Retrieved July 4, 2018, from
<https://www.crunchbase.com/organization/yatown>

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, 9(2), e1002854.
<https://doi.org/10.1371/journal.pcbi.1002854>

Cuthbertson, A. (2018, January 2). Facebook tracks Android users even if they don't have a Facebook account, study reveals. *The Independent*. Retrieved from
<https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-android-privacy-data-tracking-skyscanner-duolingo-a8708071.html>

Cutting, D. (2013). The Apache Hadoop Ecosystem. Retrieved January 30, 2014, from [http://assets.en.oreilly.com/1/event/75/The Apache Hadoop Ecosystem Presentation.pdf](http://assets.en.oreilly.com/1/event/75/The%20Apache%20Hadoop%20Ecosystem%20Presentation.pdf)

- Dahlberg, S. (2013). Does context matter – The impact of electoral systems, political parties and individual characteristics on voters’ perceptions of party positions. *Electoral Studies*, 32(4), 670–683.
<https://doi.org/10.1016/j.electstud.2013.02.003>
- Dahlgren, P. (2005). The Internet, public spheres, and political communication: Dispersion and deliberation. *POLITICAL COMMUNICATION*, 22(2), 147–162.
<https://doi.org/10.1080/10584600590933160>
- Dahlgren, P., & Sparks, C. (1991). *Communication and citizenship : journalism and the public sphere in the new media age*. London; New York: Routledge.
- Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial “Big Data” and geodemographics. *Big Data & Society*, 2(2), 1–15.
<https://doi.org/10.1177/2053951715601144>
- Dardel, É. (1952). *L’homme et la terre: nature de la réalité géographique*. Presses Universitaires de France. Retrieved from
<https://books.google.co.uk/books?id=x-YHAAAAMAAJ>
- Darmon, D., Omodei, E., & Garland, J. (2014). Followers Are Not Enough: Beyond Structural Communities in Online Social Networks. *ArXiv Preprint ArXiv:1404.0300*, 1–20. Retrieved from <http://arxiv.org/abs/1404.0300>
- Datasift. (2018). Datasift Developer Site | salience.content.sentiment. Retrieved June 8, 2018, from
<https://dev.datasift.com/docs/products/stream/features/augmentations/augmentation-salience/salience-content-sentiment>
- DataSift. (2013a). Language Guide | DataSift Developers. Retrieved September 23, 2013, from <http://dev.datasift.com/csdl>
- DataSift. (2013b). Twitter Data | DataSift Developers. Retrieved June 20, 2013, from

<http://dev.datasift.com/docs/getting-started/data/twitter>

DataSift. (2018a). interaction.sample | DataSift Developers. Retrieved September 23, 2018, from <https://dev.datasift.com/docs/products/stream/features/augmentations/common-interaction/interaction-sample>

DataSift. (2018b). STREAM for Human Data | DataSift. Retrieved September 11, 2018, from <https://datasift.com/products/stream/>

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. Retrieved from <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Davies, J. (2018). Word Cloud Generator. Retrieved May 2, 2018, from <https://www.jasondavies.com/wordcloud/>

Davies, W. (2018, April). Cambridge Analytica · LRB 5 April 2018. *The London Review of Books*. Retrieved from <https://www.lrb.co.uk/v40/n07/william-davies/short-cuts>

DB Browser for SQLite. (2016). DB Browser for SQLite. Retrieved September 2, 2016, from <http://sqlitebrowser.org/>

De Nooy, W., & Kleinnijenhuis, J. (2013). Polarization in the Media During an Election Campaign: A Dynamic Network Model Predicting Support and Attack Among Political Actors. *Political Communication*, 30(1), 117–138. <https://doi.org/10.1080/10584609.2012.737417>

de Souza e Silva, A. (2013). Location-aware mobile technologies: Historical, social and spatial approaches. *Mobile Media & Communication*, 1(1), 116–121. <https://doi.org/10.1177/2050157912459492>

de Zúñiga, H. G., Copeland, L., & Bimber, B. (2014). Political consumerism: Civic

- engagement and the social media connection. *New Media & Society*, 16(3), 488–506. <https://doi.org/10.1177/1461444813487960>
- de Zúñiga, H. G., Veenstra, A., Vraga, E., & Shah, D. (2010). Digital Democracy: Reimagining Pathways to Political Participation. *Journal of Information Technology & Politics*, 7(1), 36–51. <https://doi.org/10.1080/19331680903316742>
- Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., ... Horrocks, I. (2000). Semantic Web: The roles of XML and RDF. *IEEE Internet Computing*, 4(5), 63–74. <https://doi.org/10.1109/4236.877487>
- Defence Science and Technology Laboratory. (2015). Dstl adds to open source software - GOV.UK. Retrieved June 10, 2017, from <https://www.gov.uk/government/news/dstl-adds-to-open-source-software>
- Deitrick, W., & Hu, W. (2013). Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks. *Journal of Data Analysis and Information Processing*, 01(03), 19–29. <https://doi.org/10.4236/jdaip.2013.13004>
- Delboni, T. M., Borges, K. A. V., & Laender, A. H. F. (2005). Geographic web search based on positioning expressions. In *Proceedings of the 2005 workshop on Geographic information retrieval - GIR '05* (pp. 61–64). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1096985.1097000>
- Deleuze, G., Guattari, F. Ì., & Hurley, R. (2004). *Anti-Oedipus*. Bloomsbury Academic. Retrieved from <http://books.google.co.uk/books?id=4KCfPtKu4qAC>
- Deluliis, D. (2015). Gatekeeping Theory from Social Fields to Social Networks. *Communication Research Trends*, 34(1), 4–23. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=101594304&site=ehost-live>

- Demchenko, Y., de Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 104–112). IEEE.
<https://doi.org/10.1109/CTS.2014.6867550>
- Derczynski, L., Maynard, D., Aswani, N., & Bontcheva, K. (2013). Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13* (pp. 21–30). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2481492.2481495>
- DeSilver, D. (2015). *U.S. voter turnout trails most developed countries*. Retrieved from <http://www.pewresearch.org/fact-tank/2016/08/02/u-s-voter-turnout-trails-most-developed-countries/>
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73.
- Diamond, L. J., & Morlino, L. (2004). An Overview. *Journal of Democracy*, 15(4), 20–31. <https://doi.org/10.1353/jod.2004.0060>
- Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and Social Media Data As an Imperfect Continuous Panel Survey. *PLOS ONE*, 11(1), e0145406. <https://doi.org/10.1371/journal.pone.0145406>
- Diehl, T., Weeks, B. E., & de Zúñiga, H. G. (2016). Political persuasion on social media: Tracing direct and indirect effects of news use and social interaction. *New Media & Society*, 18(9), 1875–1895.
<https://doi.org/10.1177/1461444815616224>
- Dietrich, M. (2014). What's New with Oracle Data Pump in Oracle Database 12c. Oracle Corporation. Retrieved from <http://www.oracle.com/technetwork/database/upgrade/overview/what-new-datapump-12c-2197042.pdf>

- Digital Culture Media and Sport Committee. (2018). *Disinformation and 'fake news': Interim Report*. Retrieved from <https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf>
- Dimitrova, D. V., & Matthes, J. (2018). Social Media in Political Campaigning Around the World: Theoretical and Methodological Challenges. *Journalism & Mass Communication Quarterly*, 95(2), 333–342. <https://doi.org/10.1177/1077699018770437>
- Dimmick, J., Chen, Y., & Li, Z. (2004). Competition Between the Internet and Traditional News Media: The Gratification-Opportunities Niche Dimension. *Journal of Media Economics*, 17(1), 19–33. https://doi.org/10.1207/s15327736me1701_2
- Dinan, S. (2018, March 26). Facebook under fire as prosecutors, Congress confirm investigations. *The Washington Times*. Retrieved from <https://www.washingtontimes.com/news/2018/mar/26/facebook-under-fire-prosecutors-congress/>
- Dommett, K., & Temple, L. (2018). Digital Campaigning: The Rise of Facebook and Satellite Campaigns. *Parliamentary Affairs*, 71(suppl_1), 189–202. <https://doi.org/10.1093/pa/gsx056>
- Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text Mining: Finding Nuggets in Mountains of Textual Data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 398–401). New York, NY, USA: ACM. <https://doi.org/10.1145/312129.312299>
- Dunleavy, P. (2003). *Authoring a PhD thesis : how to plan, draft, write and finish a doctoral dissertation*. Basingstoke: Palgrave Macmillan.
- Ebrahimi, M., ShafieiBavani, E., Wong, R., & Chen, F. (2017). Exploring Celebrities on

- Inferring User Geolocation in Twitter. In J. Kim, K. Shim, L. Cao, J.-G. Lee, X. Lin, & Y.-S. Moon (Eds.), *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I* (pp. 395–406). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57454-7_31
- Echeverría, J., & Zhou, S. (2017). Discovery, Retrieval, and Analysis of “Star Wars” botnet in Twitter. *CoRR*, *abs/1701.0*. Retrieved from <http://arxiv.org/abs/1701.02405>
- ECMA International. (2013). *ECMA-404 The JSON Data Interchange Format*. Geneva. Retrieved from <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf>
- ECMA International. (2017). *The JSON Data Interchange Syntax. Standard ECMA-404* (Vol. 2nd Editio). Retrieved from <http://www.ecma-international.org/publications/standards/Ecma-404.htm>
- Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, *16*(3), 245–260. <https://doi.org/10.1080/13645579.2013.774185>
- Egelman, S., Felt, A. P., & Wagner, D. (2013). Choice Architecture and Smartphone Privacy: There’s a Price for That. In R. Bohme (Ed.), *The Economics of Information Security and Privacy* (pp. 211–236). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39498-0_10
- Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Management Review*, *14*(4), 532–550. <https://doi.org/10.5465/AMR.1989.4308385>
- Elden, S. (2005). Missing the point: globalization, deterritorialization and the space

of the world. *Transactions of the Institute of British Geographers*, 30(1), 8–19.
Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1475-5661.2005.00148.x/full>

Elisa Omodei, Manlio De Domenico, & Alex Arenas. (2015). Characterizing interactions in online social networks during exceptional events. *ArXiv E-Prints*, 1–11. <https://doi.org/10.3389/fphy.2015.00059>

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information : Spatial Data , Geographic Research , and New Social Practice. *Annals of the Association of American Geographers*, 102(3), 571–590. <https://doi.org/10.1080/00045608.2011.595657>

Elwood, S., & Leszczynski, A. (2013). New spatial media, new knowledge politics. *Transactions of the Institute of British Geographers*, 38(4), 544–559. <https://doi.org/10.1111/j.1475-5661.2012.00543.x>

Embury-Dennis, T. (2018, December 7). Facebook fined £8.9m by Italy for misleading users over data use. *The Independent*. Retrieved from <https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-fine-italy-user-data-scandal-privacy-settings-cambridge-analytica-social-media-a8673376.html>

Enli, G. S., & Skogerbø, E. (2013). PERSONALIZED CAMPAIGNS IN PARTY-CENTRED POLITICS. *Information, Communication & Society*, 16(5), 757–774. <https://doi.org/10.1080/1369118X.2013.782330>

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–13). New York, New York, USA: ACM Press. <https://doi.org/10.1145/3173574.3174006>

- ESRI. (2018). ArcGIS | Main. Retrieved May 29, 2018, from <https://www.arcgis.com/features/index.html>
- Ethington, P. J., & McDaniel, J. A. (2007). Political Places and Institutional Spaces: The Intersection of Political Science and Political Geography. *Annual Review of Political Science*, 10(1), 127–142.
<https://doi.org/10.1146/annurev.polisci.10.080505.100522>
- Etter, L., & Frier, S. (2018). Facebook App Developer Kogan Defends His Actions With User Data - Bloomberg. Retrieved April 12, 2018, from <https://www.bloomberg.com/news/articles/2018-03-21/facebook-app-developer-kogan-defends-his-actions-with-user-data>
- Experian. (2018). Mosaic - The consumer classification solution for consistent cross-channel marketing. Retrieved from https://www.experian.co.uk/assets/marketing-services/brochures/mosaic_uk_brochure.pdf
- Faber, A., Matthes, F., & Michel, F. (2016). *Digital Mobility Platforms and Ecosystems State of the Art Report* (Project Consortium TUM Living Lab Connected Mobility). <https://doi.org/10.14459/2016md1324021>
- Facebook. (2013). JSON with Unity. Retrieved January 28, 2014, from <https://developers.facebook.com/docs/unity/reference/current/Json/>
- Facebook. (2018a). APIs and SDKs - App Development - Documentation - Facebook for Developers. Retrieved January 31, 2018, from <https://developers.facebook.com/docs/apis-and-sdks>
- Facebook. (2018b). Company Info | Facebook Newsroom. Retrieved January 31, 2018, from <https://newsroom.fb.com/company-info/>
- Facebook. (2018c). Facebook Launches New Initiative to Help Scholars Assess Social

- Media's Impact on Elections | Facebook Newsroom. Retrieved July 9, 2018, from <https://newsroom.fb.com/news/2018/04/new-elections-initiative/>
- Facebook. (2018d). Location targeting | Facebook for Business. Retrieved September 4, 2018, from <https://en-gb.facebook.com/business/a/location-targeting>
- Farrell, D. M., McAllister, I., & Studlar, D. T. (1998). Sex, money and politics: sleaze and the Conservative party in the 1997 election. *British Elections & Parties Yearbook*, 8(1), 80–94. <https://doi.org/DOI: 10.1080/13689889808413006>
- Feder, A. (2006). BibTeX. Retrieved October 28, 2016, from <http://www.bibtex.org/>
- Feezell, J. T. (2016). Predicting Online Political Participation. *Political Research Quarterly*, 69(3), 495–509. <https://doi.org/10.1177/1065912916652503>
- Feinerer, I., Hornik, K., & Artifex Software Inc. (2016). Text Mining Package [R package tm version 0.6-2]. Retrieved October 7, 2016, from <https://cran.r-project.org/web/packages/tm/index.html>
- Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., ... Zamir, O. (1998). Text mining at the term level. In J. M. Żytkow & M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery* (pp. 65–73). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, 3(1), 1–15. <https://doi.org/10.1177/2053951716645828>
- Ferrara, E., JafariAsbagh, M., Varol, O., Qazvinian, V., Menczer, F., & Flammini, A. (2013). Clustering memes in social media. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 548–555. <https://doi.org/10.1145/2492517.2492530>

Feuerstein, S., & Pribyl, B. (2005). *Oracle PL/SQL Programming*. O'Reilly Media, Inc.

Fifth Harmony. (2013). *Better Together*. YouTube. Retrieved from

<https://www.youtube.com/watch?v=ish5x8SVqls>

Fildes, N. (2018, September 2). UK media groups call for oversight of social networks. *Financial Times*. Retrieved from

<https://www.ft.com/content/87e025b4-aea2-11e8-99ca-68cf89602132>

Fink, A. (2005). *Conducting research literature reviews: from the Internet to paper* (2nd ed.). Sage Publications.

Flickr. (2018). Flickr Services. Retrieved September 13, 2018, from

<https://www.flickr.com/services/api/>

Foer, F. (2017). *World Without Mind: The Existential Threat of Big Tech*. Penguin.

Retrieved from <https://www.penguinrandomhouse.com/books/533937/world-without-mind-by-franklin-foer/9781101981122/>

Foley, J. (2013). OracleVoice: Extreme Big Data: Beyond Zettabytes And Yottabytes - Forbes. Retrieved January 29, 2014, from

<http://www.forbes.com/sites/oracle/2013/10/09/extreme-big-data-beyond-zettabytes-and-yottabytes/>

Forbrukerrådet. (2018). *EVERY STEP YOU TAKE: How deceptive design lets Google*

track users 24/7. Retrieved from <https://fil.forbrukerradet.no/wp-content/uploads/2018/11/27-11-18-every-step-you-take.pdf>

Franch, F. (2013). (Wisdom of the Crowds)2: 2010 UK Election Prediction with Social Media. *Journal of Information Technology & Politics*, 10(1), 57–71.

<https://doi.org/10.1080/19331681.2012.705080>

Frenkel, S., Rosenberg, M., & Confessore, N. (2018, April 10). Facebook Data

Collected by Quiz App Included Private Messages - The New York Times. *The*

New York Times. Retrieved from
<https://www.nytimes.com/2018/04/10/technology/facebook-cambridge-analytica-private-messages.html>

Fu, P.-W., Wu, C.-C., & Cho, Y.-J. (2017). What makes users share content on facebook? Compatibility among psychological incentive, social capital focus, and content type. *Computers in Human Behavior*, 67, 23–32.
<https://doi.org/10.1016/j.chb.2016.10.010>

Fuchs, C. (2017a). From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication*, 32(1), 37–49. <https://doi.org/10.1177/0267323116682804>

Fuchs, C. (2017b). *Social Media: A Critical Introduction* (2nd ed.). SAGE Publications. Retrieved from <https://uk.sagepub.com/en-gb/eur/social-media/book250870>

Fuchs, C. (2017c). *Written Evidence Submitted to the House of Commons-Digital, Culture, Media and Sport Select Committee's Inquiry on Fake News*. Retrieved from
<http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/fake-news/written/73970.pdf>

Fuchs, C., & Trottier, D. (2015). Towards a theoretical model of social media surveillance in contemporary society. *Communications*, 40(1), 113–135.
<https://doi.org/10.1515/commun-2014-0029>

Gallup. (2012). Romney 49%, Obama 48% in Gallup's Final Election Survey. Retrieved from <http://www.gallup.com/poll/158519/romney-obama-gallup-final-election-survey.aspx>

Gapper, J. (2018, April 11). Mark Zuckerberg cannot control his own creation. *Financial Times*. Retrieved from <https://www.ft.com/content/a8d6762a-3cc8->

11e8-b9f9-de94fa33a81e?segmentId=aac5ef6f-cbfa-4a55-3109-33e655534e06

GATE. (2014). GATE.ac.uk - download/index.html. Retrieved October 31, 2014, from <https://gate.ac.uk/download/>

GATE. (2017). GATE.ac.uk - overview.html. Retrieved June 5, 2017, from <https://gate.ac.uk/overview.html>

GATE. (2018). Try sample services - GATE Cloud. Retrieved June 28, 2018, from <https://cloud.gate.ac.uk/shopfront/sampleServices>

Gayo-Avello, D. (2012a). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data. *ArXiv Preprint ArXiv:1204.6441*, 13. <https://doi.org/10.1234/12345678>

Gayo-Avello, D. (2012b). No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, 16(6), 91–94. <https://doi.org/10.1109/MIC.2012.137>

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, 31(6), 649–679. <https://doi.org/10.1177/0894439313493979>

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, 15(6), 753–773. <https://doi.org/10.1111/j.1467-9671.2011.01294.x>

GeoNames. (2016). About GeoNames. Retrieved April 15, 2017, from <http://www.geonames.org/about.html>

George Washington University Libraries. (2016). Social Feed Manager. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.597278>

Gephi. (2018a). Gephi - The Open Graph Viz Platform. Retrieved June 17, 2018, from <https://gephi.org/>

- Gephi. (2018b). Modularity · gephi/gephi Wiki · GitHub. Retrieved from <https://github.com/gephi/gephi/wiki/Modularity>
- Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2012). Disagreement and the Avoidance of Political Discussion: Aggregate Relationships and Differences across Personality Traits. *American Journal of Political Science*, 56(4), 849–874. <https://doi.org/10.1111/j.1540-5907.2011.00571.x>
- Gerring, J. (2006). *Case study research: Principles and practices*. Cambridge University Press.
- Giardullo, P. (2016). Does ‘bigger’ mean ‘better’? Pitfalls and shortcuts associated with big data for social research. *Quality & Quantity*, 50(2), 529–547. <https://doi.org/10.1007/s11135-015-0162-8>
- Gibson, R. K. (2015). Party change, social media and the rise of ‘citizen-initiated’ campaigning. *Party Politics*, 21(2), 183–197. <https://doi.org/10.1177/1354068812472575>
- Giddens, A. (1985). *A Contemporary Critique of Historical Materialism: The nation-state and violence*. University of California Press. Retrieved from <http://books.google.co.uk/books?id=qqJ753lp-FAC>
- Gilbert, E., Bergstrom, T., & Karahalios, K. (2009). Blogs are Echo Chambers: Blogs are Echo Chambers. In *2009 42nd Hawaii International Conference on System Sciences* (pp. 1–10). IEEE. <https://doi.org/10.1109/HICSS.2009.91>
- Giridhar, P., Wang, S., Abdelzaher, T., Amin, T. Al, & Kaplan, L. (2017). Social Fusion: Integrating Twitter and Instagram for Event Monitoring. In *2017 IEEE International Conference on Autonomic Computing (ICAC)* (pp. 1–10). IEEE. <https://doi.org/10.1109/ICAC.2017.46>
- Gittelman, S. H., Thomas, R. K., Lavrakas, P. J., & Lange, V. (2015). Quota Controls in

- Survey Research. *Journal of Advertising Research*, 55(4), 368–379.
<https://doi.org/10.2501/JAR-2015-020>
- Giuliani, M. (2018). Making sense of pollsters' errors. An analysis of the 2014 second-order European election predictions. *Journal of Elections, Public Opinion and Parties*, 1–17. <https://doi.org/10.1080/17457289.2018.1466786>
- Glynn, C. J., Huge, M. E., & Hoffman, L. H. (2012). All the news that's fit to post: A profile of news use on social networking sites. *Computers in Human Behavior*, 28(1), 113–119. <https://doi.org/10.1016/j.chb.2011.08.017>
- Goldberg, G. (2010). Rethinking the public/virtual sphere: The problem with participation. *New Media & Society*, 13(5), 739–754.
<https://doi.org/10.1177/1461444810379862>
- Goldberg, R. (1974). Survey of virtual machine research. *Computer*, 7(6). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6323581
- Golder, S. A., & Macy, M. W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333(6051), 1878–1881.
<https://doi.org/10.1126/science.1202775>
- Golder, S., Ahmed, S., Norman, G., & Booth, A. (2017). Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review. *Journal of Medical Internet Research*, 19(6), e195. <https://doi.org/10.2196/jmir.7082>
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211–221. <https://doi.org/10.1007/s10708-007-9111-y>
- Goodchild, M. F. (2013). The quality of big (geo)data. *Dialogues in Human Geography*, 3(3), 280–284. <https://doi.org/10.1177/2043820613513392>
- Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*,

3(3), 231–241. <https://doi.org/10.1080/17538941003759255>

Google. (2018a). Google+ API | Google+ Platform for Web | Google Developers.

Retrieved January 31, 2018, from

<https://developers.google.com/+/web/api/rest/>

Google. (2018b). Google Analytics Solutions - Marketing Analytics & Measurement.

Retrieved July 8, 2018, from <https://www.google.com/analytics/>

Google. (2018c). Reverse Geocoding | Maps JavaScript API | Google Developers.

Retrieved January 7, 2019, from

<https://developers.google.com/maps/documentation/javascript/examples/geocoding-reverse>

Google. (2018d). Target ads to geographic locations - Google Ads Help. Retrieved

September 4, 2018, from [https://support.google.com/google-](https://support.google.com/google-ads/answer/1722043?hl=en-GB)

[ads/answer/1722043?hl=en-GB](https://support.google.com/google-ads/answer/1722043?hl=en-GB)

Grabher, G., & König, J. (2017). Performing Network Theory? Reflexive Relationship

Management on Social Network Sites. In B. Hollstein, W. Matiaske, & K.-U.

Schnapp (Eds.), *Networked Governance: New Research Perspectives* (pp. 121–

140). Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-](https://doi.org/10.1007/978-3-319-50386-8_8)

[319-50386-8_8](https://doi.org/10.1007/978-3-319-50386-8_8)

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data

and the future of geography. *Dialogues in Human Geography*, 3(3), 255–261.

<https://doi.org/10.1177/2043820613513121>

Graham, M., Stephens, M., & Hale, S. (2013). Featured graphic: Mapping the

geoweb: a geography of Twitter. *Environment and Planning A*, 45(1), 100–102.

<https://doi.org/10.1068/a45349>

Granka, L. (2010). Measuring Agenda Setting with Online Search Traffic: Influences

of Online and Traditional Media. In *2010 Annual Meeting of the American Political Science Association*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1658172

Gray, M., & Caul, M. (2000). Declining voter turnout in advanced industrial democracies, 1950 to 1997 - The effects of declining group mobilization. *COMPARATIVE POLITICAL STUDIES*, 33(9), 1091–1122. <https://doi.org/10.1177/0010414000033009001>

Gray, S., Milton, R., & Hudson-Smith, A. (2015). Advances in Crowdsourcing: Surveys, Social Media and Geospatial Analysis: Towards a Big Data Toolkit. In F. J. Garrigos-Simon, I. Gil-Pechuán, & S. Estelles-Miguel (Eds.), *Advances in Crowdsourcing* (pp. 163–179). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18341-1_13

Gribkovskaia, I., Halskau, Ø., & Laporte, G. (2007). The bridges of Königsberg—A historical perspective. *Networks*, 49(3), 199–203. <https://doi.org/10.1002/net.20159>

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2), 603–623. <https://doi.org/10.1007/s10579-017-9385-8>

Groat, S., Dunlop, M., Marchanyy, R., & Tront, J. (2011). IPv6: Nowhere to run, nowhere to hide. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1–10. <https://doi.org/10.1109/HICSS.2011.258>

Groom, R., & Booth, S. (2016). AGI GeoCom 2015: Day 2. *GIS Professional*, (68), 10–11.

Guest Contributor. (2015, August). Has the Social Media Wild West Been Tamed? *Adweek*. Retrieved from <https://www.adweek.com/digital/has-the-social-media-wild-west-been-tamed/>

- Habermas, J. (2011). *The structural transformation of the public sphere : an inquiry into a category of bourgeois society*. Cambridge: Polity.
- Habermas, J. (2016). Core Europe To The Rescue : A Conversation With Jürgen Habermas About Brexit And The EU Crisis. Retrieved November 10, 2016, from <https://www.socialeurope.eu/2016/07/core-europe-to-the-rescue/>
- Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, 9(9), 1–36. <https://doi.org/10.5311/JOSIS.2014.9.185>
- Halavais, A. (2015). Bigger sociological imaginations: framing big social data theory and methods. *Information, Communication & Society*, 18(5), 583–594. <https://doi.org/10.1080/1369118X.2015.1008543>
- Hammerschmidt, B. (2015). The new SQL/JSON Query operators (Part5: JSON_TABLE, Nested Path, Ordinality Column) | JSON in the Oracle database. Retrieved December 6, 2017, from <https://blogs.oracle.com/jsondb/the-new-sqljson-query-operators-part5:-jsontable,-nested-path,-ordinality-column>
- Han, B., Cook, P., & Baldwin, T. (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 49, 451–500. <https://doi.org/10.1613/jair.4200>
- Hardy, D., Frew, J., & Goodchild, M. F. (2012). Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science*, 26(7), 1191–1212. <https://doi.org/10.1080/13658816.2011.629618>
- Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 63–76. <https://doi.org/10.1177/0002716215570866>

- Hargittai, E., Neuman, W. R., & Curry, O. (2012). Taming the Information Tide: Perceptions of Information Overload in the American Home. *The Information Society*, 28(3), 161–173. <https://doi.org/10.1080/01972243.2012.669450>
- Harris, L., & Harrigan, P. (2015). Social Media in Politics: The Ultimate Voter Engagement Tool or Simply an Echo Chamber? *Journal of Political Marketing*, 14(3), 251–283. <https://doi.org/10.1080/15377857.2012.693059>
- Harvey, D. (1973). *Social justice and the city*. Baltimore: Johns Hopkins University Press.
- Hasell, A., & Weeks, B. E. (2016). Partisan Provocation: The Role of Partisan News Use and Emotional Responses in Political Information Sharing in Social Media. *Human Communication Research*, 42(4), 641–661. <https://doi.org/10.1111/hcre.12092>
- Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), 260–271. <https://doi.org/10.1080/15230406.2014.890072>
- Healey, R. G. (2011). A Full-Scale Implementation of the NAPP 1880 U.S. Census Data Set Using Dimensional Modeling and Data-Warehousing Technology. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(2), 95–105. <https://doi.org/10.1080/01615440.2010.506417>
- Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber’s heart. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11* (pp. 237–246). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1978942.1978976>
- Helles, R. (2013). Mobile communication and intermediality. *Mobile Media & Communication*, 1(1), 14–19. <https://doi.org/10.1177/2050157912459496>

- Hemsley, J., & Eckert, J. (2014). Examining the role of “Place” in Twitter Networks through the Lens of Contentious Politics. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1844–1853.
<https://doi.org/10.1109/HICSS.2014.233>
- Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Wietzner, D. (2008). Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51(7), 60–69.
<https://doi.org/http://doi.acm.org/10.1145/1364782.1364798>
- Henneberg, S. C., & O’Shaughnessy, N. J. (2009). Political Relationship Marketing: some macro/micro thoughts. *Journal of Marketing Management*, 25(1–2), 5–29. <https://doi.org/10.1362/026725709X410016>
- Hermida, A., Fletcher, F., Korell, D., & Logan, D. (2012). Share, Like, Recommend. *Journalism Studies*, 13(5–6), 815–824.
<https://doi.org/10.1080/1461670X.2012.664430>
- Hern, A. (2014). Twitter buys Gnip, one of only four companies with “firehose” access. Retrieved January 14, 2017, from
<https://www.theguardian.com/technology/2014/apr/16/twitter-buys-gnip-firehose-analytics-apple-toppsy>
- Heylighen, F. (2007). The Global Superorganism : an evolutionary-cybernetic model of the emerging network society. *Journal of Social and Evolutionary Systems*, 6(1), 1–37.
- Himmelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter Topic-Networks Using Social Network Analysis. *Social Media + Society*, 3(1), 1–13. <https://doi.org/10.1177/2056305117691545>
- Hoffman, L. H., Jones, P. E., & Young, D. G. (2013). Does my comment count? Perceptions of political participation in an online environment. *Computers in*

Human Behavior, 29(6), 2248–2256.

<https://doi.org/10.1016/j.chb.2013.05.010>

Hogan, B. (2018). Social Media Giveth , Social Media Taketh Away : Facebook , Friendships , and APIs. *International Journal of Communication*, 12, 592–611.

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, 29(4), 455–461. <https://doi.org/10.1016/j.giq.2012.06.004>

Hornick, M. (2010). To sample or not to sample... Part 2 | Oracle Data Mining (ODM) Blog. Retrieved November 1, 2017, from <https://blogs.oracle.com/datamining/to-sample-or-not-to-sample-part-2>

Hough, M. G. (2009). Keeping it to ourselves: Technology, privacy, and the loss of reserve. *Technology in Society*, 31(4), 406–413. <https://doi.org/10.1016/j.techsoc.2009.10.005>

Howard, P. N., Kollanyi, B., Bradshaw, S., & Neudert, L.-M. (2018). Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States? *ArXiv*, 1–6. Retrieved from <http://arxiv.org/abs/1802.03573>

Huang, Q., Cao, G., & Wang, C. (2014). From where do tweets originate? - A GIS approach for user location inference. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2014 - Held in Conjunction with the 22nd ACM SIGSPATIAL GIS 2014*, 1–8. <https://doi.org/10.1145/2755492.2755494>

Huang, Q., & Wong, D. W. S. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of American Geographers*, 105(6), 1179–1197.

<https://doi.org/10.1080/00045608.2015.1081120>

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, 30(9), 1873–1898.

<https://doi.org/10.1080/13658816.2016.1145225>

Hubbard, P., & Kitchin, R. (Eds.). (2011). *Key Thinkers on Space and Place* (2nd ed). SAGE Publications Ltd.

Humphreys, L. (2013). Mobile social media: Future challenges and opportunities. *Mobile Media & Communication*, 1(1), 20–25.

<https://doi.org/10.1177/2050157912459499>

Hutton, L., & Henderson, T. (2018). Toward Reproducibility in Online Social Network Research. *IEEE Transactions on Emerging Topics in Computing*, 6(1), 156–167.

<https://doi.org/10.1109/TETC.2015.2458574>

Iacus, S. M. (2014). Big Data or Big Fail? The Good, the Bad and the Ugly and the missing role of Statistics. *Electronic Journal of Applied Statistical Analysis*, 5(11), 4–11. <https://doi.org/10.1285/i2037-3627v5n1p4>

IBM. (2017a). AlchemyLanguage. Retrieved February 15, 2017, from <http://www.ibm.com/watson/developercloud/alchemy-language.html>

IBM. (2017b). IBM Watson - AlchemyAPI. Retrieved from <https://www.ibm.com/watson/alchemy-api.html>

IBM. (2017c). Natural Language Understanding - IBM Bluemix. Retrieved November 1, 2017, from <https://console.stage1.bluemix.net/catalog/services/natural-language-understanding>

IBM. (2018). Natural Language Understanding Demo. Retrieved June 28, 2018, from <https://natural-language-understanding-demo.ng.bluemix.net/>

- ICANN. (2018). ICANN WHOIS. Retrieved July 4, 2018, from <https://whois.icann.org/en/lookup?name=dk.pairsonnalites.org>
- Instagram. (2018a). Advertising on Instagram | Instagram for Business. Retrieved September 4, 2018, from https://business.instagram.com/advertising?locale=en_GB
- Instagram. (2018b). Instagram Developer Documentation. Retrieved September 11, 2018, from <https://www.instagram.com/developer/>
- Internet Assigned Numbers Authority. (2017). IANA — Time Zone Database. Retrieved July 11, 2017, from <http://www.iana.org/time-zones>
- Iosifidis, P., & Wheeler, M. (2016). Social Media, Public Sphere and Democracy. In *Public Spheres and Mediated Social Networks in the Western Context and Beyond* (pp. 13–37). London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-41030-6_2
- Jackson, D. (2017). Know Your Limit: The Ideal Length of Every Social Media Post. Retrieved June 1, 2018, from <https://sproutsocial.com/insights/social-media-character-counter/>
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, 9(6), 1–12. <https://doi.org/10.1371/journal.pone.0098679>
- Jain, V. K., & Kumar, S. (2017). Towards Prediction of Election Outcomes Using Social Media. *International Journal of Intelligent Systems and Applications*, 9(12), 20–28. <https://doi.org/10.5815/ijisa.2017.12.03>
- Jenkins, J. C., Slomczynski, K. M., & Dubrow, J. K. (2016). GUEST EDITORS' INTRODUCTION Political Behavior and Big Data. *International Journal of*

-
- Sociology*, 46(1), 1–7. <https://doi.org/10.1080/00207659.2016.1130409>
- Jiang, B., Ma, D., Yin, J., & Sandberg, M. (2016). Spatial Distribution of City Tweets and Their Densities. *Geographical Analysis*, 48(3), 337–351. <https://doi.org/10.1111/gean.12096>
- JISC. (2012). *The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure*. Retrieved from <http://bit.ly/jisc-textm>
- Johnson, I. L., Sengupta, S., Schöning, J., & Hecht, B. (2016). The Geography and Importance of Localness in Geotagged Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 515–526. <https://doi.org/10.1145/2858036.2858122>
- Johnson, T. J., & Kaye, B. K. (2014). Site Effects: How Reliance on Social Media Influences Confidence in the Government and News Media. *Social Science Computer Review*, 33(2), 127–144. <https://doi.org/10.1177/0894439314537029>
- Johnston, R. (2009). The Concept of Constituency: Political Representation, Democratic Legitimacy, and Institutional Design. *Political Geography*, 28(8), 511–512. <https://doi.org/10.1016/j.polgeo.2009.10.001>
- Johnston, R., Harris, R., Jones, K., Manley, D., Sabel, C. E., & Wang, W. W. (2014). Mutual misunderstanding and avoidance, misrepresentations and disciplinary politics: spatial science and quantitative analysis in (United Kingdom) geographical curricula. *Dialogues in Human Geography*, 4(1), 3–25. <https://doi.org/10.1177/2043820614525706>
- Johnston, R., & Pattie, C. (2006). *Putting Voters in Their Place : Geography and Elections in Great Britain*. Oxford, GBR: Oxford University Press.
- Jung, J.-K. (2015). Code clouds: Qualitative geovisualization of geotweets. *The*
-

Canadian Geographer / Le Géographe Canadien, 59(1), 52–68.

<https://doi.org/10.1111/cag.12133>

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital Trace Data in the Study of Public Opinion. *Social Science Computer Review*, 35(3), 336–356.

<https://doi.org/10.1177/0894439316631043>

Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* (pp. 273–282).

Retrieved from

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6067>

Kallet, R. H. (2004). How to Write the Methods Section of a Research Paper.

Respiratory Care, 49(10), 1229–1232. Retrieved from

<http://www.ncbi.nlm.nih.gov/pubmed/15447808>

Kamath, K. Y., Caverlee, J., Cheng, Z., & Sui, D. Z. (2012). Spatial Influence vs.

Community Influence: Modeling the Global Spread of Social Media. In

Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12 (pp. 962–971). New York, New York, USA:

ACM Press. <https://doi.org/10.1145/2396761.2396883>

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics.

Journal of Parallel and Distributed Computing, 74(7), 2561–2573.

<https://doi.org/10.1016/j.jpdc.2014.01.003>

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68.

<https://doi.org/10.1016/j.bushor.2009.09.003>

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018).

Advances in Social Media Research: Past, Present and Future. *Information*

Systems Frontiers, 20(3), 531–558. <https://doi.org/10.1007/s10796-017-9810-y>

Karlsen, R. (2015). Followers are opinion leaders: The role of people in the flow of political communication on and beyond social networking sites. *European Journal of Communication*, 30(3), 301–318.

<https://doi.org/10.1177/0267323115577305>

Keane, S. (2018). EU consumer watchdogs demand action against Google location tracking. Retrieved January 7, 2019, from <https://www.cnet.com/news/eu-consumer-watchdogs-demand-action-against-google-location-tracking/>

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 154–175). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

Kim, J., Zhang, A. X., Kim, J., Miller, R. C., & Gajos, K. Z. (2014). Content-aware Kinetic Scrolling for Supporting Web Page Navigation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 123–127). New York, NY, USA: ACM.

<https://doi.org/10.1145/2642918.2647401>

Kim, K.-S., Kojima, I., & Ogawa, H. (2016). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science*, 30(9), 1899–1922.

<https://doi.org/10.1080/13658816.2016.1146956>

Kim, Y., Hsu, S.-H., & de Zúñiga, H. G. (2013). Influence of Social Media Use on Discussion Network Heterogeneity and Civic Engagement: The Moderating Role of Personality Traits. *Journal of Communication*, 63(3), 498–516.

<https://doi.org/10.1111/jcom.12034>

- Kim, Y., Russo, S., & Amnå, E. (2017). The longitudinal relation between online and offline political participation among youth at two different developmental stages. *New Media & Society*, 19(6), 899–917.
<https://doi.org/10.1177/1461444815624181>
- Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, 4(2), 1–14. <https://doi.org/10.1177/2053951717736336>
- Kirkby, E. J. (2016). The city getting rich from fake news. Retrieved February 25, 2017, from <http://www.bbc.co.uk/news/magazine-38168281>
- Kiyohara, S. (2009). A Study on How Technological Innovation Affected the 2008 U.S. Presidential Election: Young Voters’ Participation and Obama’s Victory. *2009 Ninth Annual International Symposium on Applications and the Internet*, (2003), 223–226. <https://doi.org/10.1109/SAINT.2009.51>
- Klofstad, C. a., & Bishin, B. G. (2014). Do Social Ties Encourage Immigrant Voters to Participate in Other Campaign Activities? *Social Science Quarterly*, 95(2), 295–310. <https://doi.org/10.1111/ssqu.12040>
- Knobbe, A. J., Siebes, A., & Marseille, B. (2002). Involving Aggregate Functions in Multi-relational Search. In T. Elomaa, H. Mannila, & H. Toivonen (Eds.), *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19--23, 2002 Proceedings* (pp. 287–298). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45681-3_24
- Knowles, E. (2006). *What They Didn’t Say: A Book of Misquotations*. OUP Oxford. Retrieved from <https://books.google.co.uk/books?id=jxFQqDLav6wC>
- Kohn, B. (2016). The New Reality Of Location Targeting And Presidential Political

Campaigns | . Retrieved December 14, 2016, from
<http://www.geomarketing.com/the-new-reality-of-location-targeting-and-presidential-political-campaigns>

Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, I. (2017). Geotagging Text Content With Language Models and Feature Mining. *Proceedings of the IEEE*, 105(10), 1971–1986. <https://doi.org/10.1109/JPROC.2017.2688799>

Koylu, C. (2018). Uncovering Geo-Social Semantics from the Twitter Mention Network: An Integrated Approach Using Spatial Network Smoothing and Topic Modeling. In S.-L. Shaw & D. Sui (Eds.), *Human Dynamics Research in Smart and Connected Communities* (pp. 163–179). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73247-3_9

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kulshrestha, J., Zafar, M. B., Espin-Noboa, L., Gummadi, K. P., & Ghosh, S. (2017). Characterizing Information Diets of Social Media Users. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 218–227). Oxford. Retrieved from
<https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10595/10505>

Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter Data Analytics*. New York, NY: Springer New York. <https://doi.org/10.1007/978-1-4614-9372-3>

Küpper, A. (2005). *Location-based services: fundamentals and operation*. John Wiley

& Sons.

Kwon, K. H., Wang, H., Raymond, R., & Xu, W. W. (2015). A Spatiotemporal Model of Twitter Information Diffusion: An Example of Egyptian Revolution 2011. In *Proceedings of the 2015 International Conference on Social Media & Society - SMSociety '15* (pp. 1–7). New York, New York, USA: ACM Press.
<https://doi.org/10.1145/2789187.2789205>

Labaree, R. V. (2017). Organizing Your Social Sciences Research Paper. Retrieved February 7, 2017, from <http://libguides.usc.edu/writingguide>

Labour Party. (2018). *CAMPAIGNERS' HANDBOOK: Your guide to elections and year-round campaigning*. Retrieved from [https://action.labour.org.uk/page/-/Campaigner%27s Handbook.pdf](https://action.labour.org.uk/page/-/Campaigner%27s%20Handbook.pdf)

Lam, K. P. (2016). Re: MapR & Keele.ac.uk.

Lane, K. (2017). History of APIs. Retrieved January 31, 2018, from <https://history.apievangelist.com/>

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Retrieved from <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>

Langer, G. (2012). Dead Heat in Vote Preferences Presages an Epic Battle Ahead. Retrieved from <http://abcnews.go.com/blogs/politics/2012/07/dead-heat-in-vote-preferences-presages-an-epic-battle-ahead/>

Language Technology Group. (2014). The Edinburgh Geoparser | Language Technology Group. Retrieved from <https://www.ltg.ed.ac.uk/software/geoparser/>

Lau, J. H., Chi, L., Tran, K.-N., & Cohn, T. (2017). End-to-end Network for Twitter

- Geolocation Prediction and Hashing. *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*, 10. Retrieved from <http://arxiv.org/abs/1710.04802v1>
- Lazer, D., Brewer, D., Christakis, N., Fowler, J., & King, G. (2009). Life in the network: the coming age of computational social science. *Science*, 323(5915), 721–723. <https://doi.org/10.1126/science.1167742>.Life
- Lee, B. (2014). Window Dressing 2.0: Constituency-Level Web Campaigns in the 2010 UK General Election. *Politics*, 34(1), 45–57. <https://doi.org/10.1111/1467-9256.12029>
- Lee, C.-H. (2012). Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications*, 39(10), 9623–9641. <https://doi.org/10.1016/j.eswa.2012.02.136>
- Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, 28(2), 331–339. <https://doi.org/10.1016/j.chb.2011.10.002>
- Lees-Marshment, J., & Lilleker, D. G. (2012). Knowledge sharing and lesson learning: consultants' perspectives on the international sharing of political marketing strategy. *Contemporary Politics*, 18(3), 343–354. <https://doi.org/10.1080/13569775.2012.702976>
- Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5). Retrieved from <http://firstmonday.org/ojs/index.php/fm/article/view/4366>
- Lemieux, A. M. (2015). Geotagged photos: a useful tool for criminological research? *Crime Science*, 4(3), 1–11. <https://doi.org/10.1186/s40163-015-0017-6>
- Lerman, K., Marin, L. G., Arora, M., de Lima, L. H. C., Ferrara, E., & Garcia, D. (2018).

Language, demographics, emotions, and the structure of online social networks. *Journal of Computational Social Science*, 1(1), 209–225.
<https://doi.org/10.1007/s42001-017-0001-x>

Leszczynski, A., & Crampton, J. (2016). Introduction: Spatial Big Data and everyday life. *Big Data & Society*, 3(2), 1–6. <https://doi.org/10.1177/2053951716661366>

Leventhal, B. (2016). *Geodemographics for Marketers: Using Location Analysis for Research and Marketing*. Kogan Page.

Lewin, K. (1947). Frontiers in group dynamics II. Channels of group life; social planning and action research. *Human Relations*, 1(2), 143–153.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4), 330–342. <https://doi.org/10.1016/j.socnet.2008.07.002>

Lewis, P., & Hilder, P. (2018, March 23). Leaked: Cambridge Analytica’s blueprint for Trump victory | UK news | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory>

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1), 34–52.
<https://doi.org/10.1080/08838151.2012.761702>

Lexalytics. (2018). Salience™ | Lexalytics. Retrieved June 8, 2018, from <https://www.lexalytics.com/salience/server>

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40(2), 61–77.

<https://doi.org/10.1080/15230406.2013.777139>

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... Cheng, T.

(2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 119–133. <https://doi.org/10.1016/j.isprsjprs.2015.10.012>

Licoppe, C. (2013). Merging mobile communication studies and urban research:

Mobile locative media, “onscreen encounters” and the reshaping of the interaction order in public places. *Mobile Media & Communication*, 1(1), 122–128. <https://doi.org/10.1177/2050157912464488>

Lilleker, D. G., Tenscher, J., & Štětka, V. (2015). Towards hypermedia campaigning?

Perceptions of new media’s importance for campaigning by party strategists in comparative perspective. *Information, Communication & Society*, 18(7), 747–765. <https://doi.org/10.1080/1369118X.2014.993679>

Lin, J., & Ryaboy, D. (2013). Scaling big data mining infrastructure: the twitter

experience. *ACM SIGKDD Explorations Newsletter*, 14(2), 6–19. Retrieved from <http://dl.acm.org/citation.cfm?id=2481247>

Liu, F., Vasardani, M., & Baldwin, T. (2014). Automatic Identification of Locative

Expressions from Social Media Text. In *Proceedings of the 4th International Workshop on Location and the Web - LocWeb '14* (pp. 9–16). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2663713.2664426>

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social Sensing: A

New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, 105(3), 512–530.

<https://doi.org/10.1080/00045608.2015.1018773>

Loader, B. D., & Dutton, W. H. (1998). Editorial introduction. *Information,*

Communication & Society, 1(1), 5–6.

<https://doi.org/10.1080/13691189809358950>

Logan, D. C. (2009). Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, 60(3), 712.
<https://doi.org/10.1093/jxb/erp043>

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30(2), 369–389.
<https://doi.org/10.1080/13658816.2015.1089441>

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of twitter usage. *Environment and Planning A*, 47(2), 465–484.
<https://doi.org/10.1068/a130122p>

Lunden, I. (2015). Twitter Cuts Off DataSift To Step Up Its Own Big Data Business. Retrieved May 11, 2015, from <https://techcrunch.com/2015/04/11/twitter-cuts-off-datasift-to-step-up-its-own-b2b-big-data-analytics-business/>

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography*, 70, 11–25.
<https://doi.org/10.1016/j.apgeog.2016.03.001>

Ma, L., Lee, C. S., & Goh, D. H.-L. (2012). Sharing in Social News Websites: Examining the Influence of News Attributes and News Sharers. In *2012 Ninth International Conference on Information Technology - New Generations* (pp. 726–731). IEEE. <https://doi.org/10.1109/ITNG.2012.143>

Macafee, T. (2013). Some of these things are not like the others: Examining motivations and political predispositions among political Facebook activity. *Computers in Human Behavior*, 29(6), 2766–2775.
<https://doi.org/10.1016/j.chb.2013.07.019>

- Macagba, J. (2017). The “Intolerable Image” and New Modes of Circulation: Perpetual Revolution at the ICP. *American Quarterly*, 69(4), 967–986. <https://doi.org/10.1353/aq.2017.0076>
- Madowo, L. (2018, March 20). How Cambridge Analytica poisoned Kenya’s democracy - The Washington Post. *The Washington Post*. Retrieved from <https://www.washingtonpost.com/news/global-opinions/wp/2018/03/20/how-cambridge-analytica-poisoned-kenyas-democracy>
- Maemura, E., Moles, N., & Becker, C. (2017). Organizational assessment frameworks for digital preservation: A literature review and mapping. *Journal of the Association for Information Science and Technology*, 68(7), 1619–1637. <https://doi.org/10.1002/asi.23807>
- Magoulas, R., & Lorica, B. (2009). Big data: Technologies and techniques for large scale data. *Release 2.0*, 11, 1–39. Retrieved from <https://www.oreilly.com/data/free/release-2-issue-11.csp?intcmp=il-npa-ebooks-radar-release-2>
- Mahmud, J., Nichols, J., & Drews, C. (2012). Where Is this tweet from? Inferring home locations of Twitter users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 511–514). Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4605/5045>
- Mahmud, J., Nichols, J., & Drews, C. (2014). Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3), 1–21. <https://doi.org/10.1145/2528548>
- Mallig, N. (2010). A relational database for bibliometric analysis. *Journal of Informetrics*, 4(4), 564–580. <https://doi.org/10.1016/j.joi.2010.06.007>

- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60). Retrieved from <http://www.aclweb.org/anthology/P/P14/P14-5010>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
- MapR. (2014). Quick Installation Guide - Latest Documentation - doc.mapr.com. Retrieved October 31, 2014, from <http://doc.mapr.com/display/MapR/Quick+Installation+Guide>
- MapR. (2018). Connecting Using Hive MapR-DB JSON Connector. Retrieved June 15, 2018, from <https://maprdocs.mapr.com/home/Hive/ConnectingToMapR-DB.html>
- Maram, M. (2017). JSON in the Oracle/ Validate JSON in Oracle column – Java in & Out Blog. Retrieved July 30, 2017, from <https://manismaran.wordpress.com/json-in-the-oracle-validate-json-in-oracle-column/>
- Marechal, N. (2016). When bots tweet : Toward a normative framework for bots on social networking sites. *International Journal of Communication*, 10, 5022–5031. Retrieved from <http://ijoc.org/index.php/ijoc/article/download/6180/1811>
- MarkLogic. (2014). Enterprise NoSQL Database | MarkLogic. Retrieved January 30, 2014, from <http://www.marklogic.com/>
- Marshall, D., & Tear, A. (2016). Result!!! Portsmouth.

- Marten, K. (2017). Tackle your geospatial analysis with ease in Tableau 10.2 | Tableau Software. Retrieved May 6, 2017, from <https://www.tableau.com/about/blog/2017/2/tackle-your-geospatial-analysis-ease-tableau-102-66018>
- Martin, R. M. (2010). *Epistemology: A Beginner's Guide*. Oneworld Publications. Retrieved from <https://books.google.co.uk/books?id=czoqAQAAMAAJ>
- Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. In *Proc. SPIE 7868, Visualization and Data Analysis 2011* (pp. 786–806). <https://doi.org/10.1117/12.871402>
- Maruyama, M., Robertson, S., & Douglas, S. (2014). Hybrid Media Consumption: How Tweeting During a Televised Political Debate Influences the Vote Decision. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1422–1432). Baltimore, Maryland, USA. Retrieved from <http://hdl.handle.net/10125/34239>
- Massey, D., & Allen, J. (1984). *Geography matters!: a reader*. Cambridge University Press.
- MaxMind. (2012a). GeoIP City Accuracy for Selected Countries. Retrieved from http://www.maxmind.com/app/city_accuracy
- MaxMind. (2012b). What is GeoIP? Retrieved from <http://www.maxmind.com/app/ip-locate>
- Mayer-Kress, G., & Barczys, C. (1995). The global brain as an emergent structure from the worldwide computing network, and its implications for modeling. *The Information Society*, 11(1), 1–27. <https://doi.org/10.1080/01972243.1995.9960177>
- Maynard, D., Roberts, I., Greenwood, M. A., Rout, D., & Bontcheva, K. (2017). A

framework for real-time semantic social media analysis. *Web Semantics: Science, Services and Agents on the World Wide Web*, 44, 75–88.
<https://doi.org/10.1016/j.websem.2017.05.002>

McCarthy, T. (2018, June 14). Trump-Russia investigation explained: what we know and what happens next | US news | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/us-news/2018/jun/14/trump-russia-investigation-explained-latest-news-charges>

McElhatton, N. (2004). Secrets of my Success: Richard Webber, Founder of geodemographics. Retrieved May 6, 2017, from <http://www.campaignlive.co.uk/article/secrets-success-richard-webber-founder-geodemographics/226747>

McGreal, C. (2012, October 17). Obama deflects Romney's challenge on Benghazi attack during Hofstra debate | World news | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/world/2012/oct/17/romney-obama-benghazi-defeated-debate>

McGregor, R. (2011, December 9). Obama campaign sharpens tech edge. *Financial Times*. Retrieved from <http://www.ft.com/cms/s/0/b2e7043c-2284-11e1-923d-00144feabdc0.html>

McKenzie, G., & Janowicz, K. (2014). Coerced Geographic Information: The Not-so-voluntary Side of User-generated Geo-content. *Extended Abstract Proceedings of the 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014.*, 231–233.

McKinnon, J. D., & Seetharaman, D. (2018, April 11). In Facebook Hearings, Lawmakers Ramp Up Talk of Regulation - WSJ. *The Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/congressional-hearing-on-facebook-turns-up-heat-on-mark-zuckerberg-1523464332>

- McKirdy, E., Smith-Spark, L., & Robertson, N. (2014). Results of Scotland independence referendum: “No” campaign victorious. Retrieved July 31, 2017, from <http://edition.cnn.com/2014/09/18/world/europe/scotland-independence-vote/index.html>
- McNaught, C., & Lam, P. (2010). Using wordle as a supplementary research tool. *Qualitative Report*, 15(3), 630–643. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-77953058705&partnerID=tZOtx3y1>
- McNeill, G., Bright, J., & Hale, S. A. (2017). Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, 6(1), 24. <https://doi.org/10.1140/epjds/s13688-017-0120-x>
- Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3), 1–9. <https://doi.org/10.1177/2053168017720008>
- Mendeley. (2016). Overview | Mendeley. Retrieved August 31, 2016, from <https://www.mendeley.com/features/>
- Mercurio, B. (2004). Democracy in decline: can internet voting save the electoral process. *John Marshall Journal of Computer & Information Law*, 22. Retrieved from <https://ssrn.com/abstract=590441>
- Messing, S., & Westwood, S. (2012). How Social Media Introduces Biases in Selecting and Processing News Content. Retrieved from https://www.researchgate.net/profile/Solomon_Messing/publication/265673993_How_Social_Media_Introduces_Biases_in_Selecting_and_Processing_News_Content/links/54d8e9620cf2970e4e7a399b.pdf
- Metaxas, P. T., & Mustafaraj, E. (2012). Social Media and the Elections. *Science*,

338(6106), 472–473. <https://doi.org/10.1126/science.1230456>

Microsoft. (2013). Microsoft Download Center. Retrieved January 30, 2014, from <http://www.microsoft.com/en-us/download/details.aspx?id=36843>

Microsoft. (2014a). Hyper-V. Retrieved October 31, 2014, from <http://technet.microsoft.com/en-us/windowsserver/dd448604.aspx>

Microsoft. (2014b). Microsoft DreamSpark. Retrieved October 31, 2014, from <https://www.dreamspark.com/>

Microsoft. (2018). Microsoft Office | Productivity Tools for Home & Office. Retrieved June 1, 2018, from <https://products.office.com/en-GB/>

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4), 449–461. <https://doi.org/10.1007/s10708-014-9602-6>

Min, J., & Kim, B. (2015). How are people enticed to disclose personal information despite privacy concerns in social network sites? The calculus between benefit and cost. *Journal of the Association for Information Science and Technology*, 66(4), 839–857. <https://doi.org/10.1002/asi.23206>

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, 11, 554–557.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. <https://doi.org/10.1177/2053951716679679>

Moeller, J., de Vreese, C., Esser, F., & Kunz, R. (2014). Pathway to Political Participation: The Influence of Online and Offline News Media on Internal Efficacy and Turnout of First-Time Voters. *American Behavioral Scientist*, 58(5), 689–700. <https://doi.org/10.1177/0002764213515220>

- MongoDB. (2014). MongoDB. Retrieved January 30, 2014, from <http://www.mongodb.org/>
- Moore, M. (2016). Facebook, the Conservatives and the Risk to Fair and Open Elections in the UK. *The Political Quarterly*, 87(3), 424–430. <https://doi.org/10.1111/1467-923X.12291>
- Morales, A. J., Vavilala, V., Benito, R. M., & Bar-Yam, Y. (2017). Global patterns of synchronization in human communications. *Journal of The Royal Society Interface*, 14(128), 11. <https://doi.org/10.1098/rsif.2016.1048>
- Moreno, M. a, Goniou, N., Moreno, P. S., & Diekema, D. (2013). Ethics of social media research: common concerns and practical considerations. *Cyberpsychology, Behavior and Social Networking*, 16(9), 708–713. <https://doi.org/10.1089/cyber.2012.0334>
- Morgan, J. (2015). DataSift tweets.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. *Proceedings of ICWSM*. Retrieved from <http://arxiv.org/abs/1306.5204>
- Mousavi, R., & Gu, B. (2014). The Role of Online Social Networks in Political Polarization of Elite Politicians. In *Twentieth Americas Conference on Information Systems* (pp. 1–11).
- Muhammad, S. S., Dey, B. L., & Weerakkody, V. (2018). Analysis of Factors that Influence Customers’ Willingness to Leave Big Data Digital Footprints on Social Media: A Systematic Review of Literature. *Information Systems Frontiers*, 20(3), 559–576. <https://doi.org/10.1007/s10796-017-9802-y>
- Müller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN Electronic Journal*, 31. <https://doi.org/10.2139/ssrn.3082972>

- Mummery, J., & Rodan, D. (2013). The role of blogging in public deliberation and democracy. *Discourse, Context & Media*, 2(1), 22–39.
<https://doi.org/10.1016/j.dcm.2012.12.003>
- Murnion, S., & Healey, R. G. (1998). Modeling Distance Decay Effects in Web Server Information Flows. *Geographical Analysis - An International Journal of Theoretical Geography*, 30(4), 19.
- Murray, S. (2013). Import UTF-8 Unicode Special Characters with SQL Server Integration Services. Retrieved January 30, 2014, from
<http://www.mssqltips.com/sqlservertip/3119/import-utf8-unicode-special-characters-with-sql-server-integration-services/>
- Murthy, D. (2015). Twitter and elections: are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society*, 18(7), 816–831.
<https://doi.org/10.1080/1369118X.2015.1006659>
- Murthy, D., Gross, A., & Pensavalle, A. (2016). Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities. *Journal of Computer-Mediated Communication*, 21(1), 33–49. <https://doi.org/10.1111/jcc4.12144>
- National Records of Scotland. (2013). *Distribution comparison of Output Areas, 2001 and 2011 Censuses*. Retrieved from
<https://www.nrscotland.gov.uk/files/geography/2011-census/oa-distribution-comparison-2001-2011.pdf>
- New York Times. (2018, April 10). Mark Zuckerberg Testimony : Senators Question Facebook’s Commitment to Privacy. *New York Times*. Retrieved from
<https://www.nytimes.com/2018/04/10/us/politics/mark-zuckerberg-testimony.html>
- Newman, J. R. (1953). Leonhard Euler and the Konigsberg Bridges. *Scientific American*, 189, 66–70.

- Newman, N., Fletcher, R., Levy, D. A. L., & Nielsen, R. K. (2016). *Reuters Institute Digital News Report 2016*. Oxford. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital-News-Report-2016.pdf>
- Nissen, T. E. (2015). *# TheWeaponizationOfSocialMedia*. Royal Danish Defence College. Retrieved from <https://www.stratcomcoe.org/thomas-nissen-weaponization-social-media>
- Noble, J., & Lockett, H. (2016, November 16). 'Post-truth' made word of the year by Oxford Dictionaries | Financial Times. *Financial Times*. Retrieved from <https://www.ft.com/content/85cbb2f8-abdd-11e6-9cb3-bb8207902122>
- Nokogiri. (2017). Tutorials - Nokogiri 鋸. Retrieved June 6, 2017, from <http://www.nokogiri.org/>
- O'Brien, O., & Cheshire, J. (2014). The Mapping London Blog | Highlighting the best London maps. Retrieved October 30, 2014, from <http://mappinglondon.co.uk/>
- O'Reilly, T. (2005). What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software. Retrieved from <http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>
- Ó Tuathail, G. (1998). Political geography III: dealing with deterritorialization. *Progress in Human Geography*, 22(1), 81–93. <https://doi.org/10.1191/030913298673827642>
- Oeldorf-Hirsch, A., & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior*, 44, 240–249. <https://doi.org/10.1016/j.chb.2014.11.024>
- Office for National Statistics. (2012). 2011 Census: Population and Household Estimates for Small Areas in England and Wales, March 2011. Retrieved May

13, 2017, from

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuspopulationandhouseholdestimatesforsmallareasinenglandandwales/2012-11-23>

Ohri, A. (2014). Cheat sheets for data scientists. Retrieved November 29, 2016, from <http://www.slideshare.net/ajayohri/cheat-sheets-for-data-scientists/>

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Working Papers on Information Systems*, 10(26), 1–51. <https://doi.org/10.2139/ssrn.1954824>

Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, 4(1), 19. <https://doi.org/10.1186/s40537-017-0063-x>

OpenStreetMap. (2017). OpenStreetMap. Retrieved May 6, 2017, from <https://www.openstreetmap.org/about>

OpenStreetMap. (2018). HOT Export Tool. Retrieved June 26, 2018, from <https://export.hotosm.org/en/v3/>

Oracle. (2012). New Features in Oracle Text with Oracle Database12c, (June). Retrieved from <http://www.oracle.com/technetwork/database/information-management/oracletext12cfeatures-1932547.pdf>

Oracle. (2014a). JSON in Oracle Database. Retrieved from <http://docs.oracle.com/database/121/ADXDB/json.htm#ADXDB6246>

Oracle. (2014b). News – Oracle VM VirtualBox. Retrieved October 31, 2014, from <https://www.virtualbox.org/wiki/News>

Oracle. (2014c). Oracle Database Software Downloads | Oracle Technology Network | Oracle. Retrieved October 31, 2014, from

<http://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html>

Oracle. (2016a). How to split comma separated column of clob datatype and insert distinct rows into another table? Retrieved July 26, 2017, from https://asktom.oracle.com/pls/apex/f?p=100:11:::NO:RP:P11_QUESTION_ID:9529975800346436673

Oracle. (2016b). Indexes and index-organized tables. Retrieved April 1, 2017, from <https://docs.oracle.com/database/121/CNCPT/indexiot.htm#CNCPT1162>

Oracle. (2017a). Oracle SQL Developer Downloads. Retrieved July 18, 2017, from <http://www.oracle.com/technetwork/developer-tools/sql-developer/downloads/index.html#close>

Oracle. (2017b). STATS_MODE. Retrieved July 21, 2017, from https://docs.oracle.com/database/122/SQLRF/STATS_MODE.htm#SQLRF06320

Oracle. (2018a). ANSI Standards. Retrieved June 20, 2018, from https://docs.oracle.com/database/121/SQLRF/ap_standard_sql001.htm#SQLRF55514

Oracle. (2018b). CASE Statement. Retrieved June 22, 2018, from https://docs.oracle.com/database/121/LNPLS/case_statement.htm#LNPLS01304

Oracle. (2018c). CAST. Retrieved July 4, 2018, from <https://docs.oracle.com/database/121/SQLRF/functions024.htm#SQLRF00613>

Oracle. (2018d). Specifying Command-Line Parameters in the Control File. Retrieved June 15, 2018, from <https://docs.oracle.com/database/121/SUTIL/GUID-34A050B6-3FD7-4B77-97D2-04C03D359D16.htm#SUTIL1051>

- Osborne, S. (2017, February 26). Hedge-fund billionaire and Donald Trump backer “played key role in Brexit campaign.” *The Independent*. Retrieved from <https://www.independent.co.uk/news/uk/politics/robert-mercere-brexit-nigel-farage-donald-trump-breitbart-facebook-advertisement-cambridge-analytica-a7600041.html>
- Ostman, J. (2012). Information, expression, participation: How involvement in user-generated content relates to democratic engagement among young people. *New Media & Society*, 14(6), 1004–1021. <https://doi.org/10.1177/1461444812438212>
- Paltrinieri, R., & Esposti, P. (2013). Processes of Inclusion and Exclusion in the Sphere of Prosumerism. *Future Internet*, 5(1), 21–33. <https://doi.org/10.3390/fi5010021>
- Panagopoulos, C., Gueorguieva, V., Slotnick, A., Gulati, G., & Williams, C. (2009). *Politicking online: The transformation of election campaign communications*. Rutgers University Press.
- Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society*, 4(1), 9–27. <https://doi.org/10.1177/14614440222226244>
- Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *NEW MEDIA & SOCIETY*, 6(2), 259–283. <https://doi.org/10.1177/1461444804041444>
- Papacharissi, Z. (2010). *A private sphere : democracy in a digital age*. Cambridge, UK; Malden, MA: Polity.
- Paraskevopoulos, P., & Palpanas, T. (2016). Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social Network Analysis and Mining*, 6(1), 89. <https://doi.org/10.1007/s13278-016-0400-7>

- Park, S., Lee, J., Ryu, S., & Hahn, K. S. (2015). The Network of Celebrity Politics: Political Implications of Celebrity Following on Twitter. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 246–258. <https://doi.org/10.1177/0002716215569226>
- Parr, B. (2010). Facebook Launches Its Location Features. Retrieved August 31, 2016, from <https://mashable.com/2010/08/18/facebook-launches-its-location-features-live>
- Pavalanathan, U., & Eisenstein, J. (2015). Confounds and Consequences in Geotagged Twitter Data. *Emnlp*, (September), 2138–2148. Retrieved from <http://arxiv.org/abs/1506.02275>
- Peng, T.-Q., Zhang, L., Zhong, Z.-J., & Zhu, J. J. (2013). Mapping the landscape of Internet Studies: Text mining of social science journal articles 2000–2009. *New Media & Society*, 15(5), 644–664. <https://doi.org/10.1177/1461444812462846>
- Pennacchiotti, M., & Popescu, A. (2011). A Machine Learning Approach to Twitter User Classification. In *Fifth International AAAI Conference on Weblogs and Social Media* (pp. 281–288). Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2886>
- Pennington, N., Winfrey, K. L., Warner, B. R., & Kearney, M. W. (2015). Liking Obama and Romney (on Facebook): An experimental evaluation of political engagement and efficacy during the 2012 general election. *Computers in Human Behavior*, 44, 279–283. <https://doi.org/10.1016/j.chb.2014.11.032>
- Perkins, C. (2014). Plotting practices and politics: (im)mutable narratives in OpenStreetMap. *Transactions of the Institute of British Geographers*, 39(2), 304–317. <https://doi.org/10.1111/tran.12022>
- Persily, N. (2017). Can Democracy Survive the Internet? *Journal of Democracy*, 28(2), 63–76. <https://doi.org/10.1353/jod.2017.0025>

- Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3), 441–461.
- Pew Research Center. (2018). *The Public, the Political System and American Democracy*. Retrieved from <http://www.people-press.org/wp-content/uploads/sites/4/2018/04/4-26-2018-Democracy-release1.pdf>
- Phillips, L., Dowling, C., Shaffer, K., Hodas, N., & Volkova, S. (2017). Using Social Media to Predict the Future: A Systematic Literature Review. *ArXiv*, (June 2016), 1–55. Retrieved from <http://arxiv.org/abs/1706.06134>
- Pitney Bowes. (2018). MapInfo® Pro - Desktop GIS | GIS Mapping | Pitney Bowes. Retrieved June 24, 2018, from <https://www.pitneybowes.com/us/location-intelligence/geographic-information-systems/mapinfo-pro.html>
- Pitts, R., & Venzl, G. (2015). Oracle JSON create index fails.
- Polat, R. K. (2005). The Internet and Political Participation: Exploring the Explanatory Links. *European Journal of Communication*, 20(4), 435–459. <https://doi.org/10.1177/0267323105058251>
- PostgreSQL. (2014). PostgreSQL: Downloads. Retrieved October 31, 2014, from <http://www.postgresql.org/download/>
- Poulston, A., Stevenson, M., & Bontcheva, K. (2017). Hyperlocal Home Location Identification of Twitter Profiles. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 45–54). New York, NY, USA: ACM. <https://doi.org/10.1145/3078714.3078719>
- Pradeepa, S., & Manjula, K. R. (2016). Construction of gazetteers from geo big data using machine learning technique on Hadoop. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp.

- 1619–1622). Retrieved from <https://ieeexplore.ieee.org/document/7724541/>
- Puglisi, P. L., Montanari, D., Petrella, A., Picelli, M., & Rossetti, D. (2014). From news to facts: An Hadoop-based social graphs analysis. In *2014 International Conference on High Performance Computing & Simulation (HPCS)* (pp. 315–322). IEEE. <https://doi.org/10.1109/HPCSim.2014.6903702>
- Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*, 12(2–3), 164–318. <https://doi.org/10.1561/15000000034>
- Putnam, R. D. (1995). Bowling Alone: America's Declining Social Capital. *Journal of Democracy*, 6(1), 65–78. <https://doi.org/10.1353/jod.1995.0002>
- QGIS. (2018). Welcome to the QGIS project! Retrieved June 24, 2018, from <https://qgis.org/en/site/>
- Quercia, D., Capra, L., & Crowcroft, J. (2012). The Social World of Twitter: Topics, Geography, and Emotions. In *Sixth International AAAI Conference on Weblogs and Social Media*. Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4612/4996>
- Quintelier, E., & Theocharis, Y. (2012). Online Political Engagement, Facebook, and Personality Traits. *Social Science Computer Review*, 31(3), 280–290. <https://doi.org/10.1177/0894439312462802>
- Ram, A. (2018, June 1). AggregatIQ had data of thousands of Facebook users | Financial Times. *Financial Times*. Retrieved from <https://www.ft.com/content/737290d8-6584-11e8-90c2-9563a0613e56>
- Rao, A., Spasojevic, N., Li, Z., & Dsouza, T. (2015). Klout score: Measuring influence

across multiple social networks. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2282–2289). IEEE.

<https://doi.org/10.1109/BigData.2015.7364017>

Raphael, C., & Karpowitz, C. F. (2013). Good Publicity: The Legitimacy of Public Communication of Deliberation. *Political Communication*, 30(1), 17–41.

<https://doi.org/10.1080/10584609.2012.737412>

Ratcliffe, S. (2016). *Oxford Essential Quotations*. (S. Ratcliffe, Ed.), Oxford University Press (4th ed., Vol. 1). Oxford University Press.

<https://doi.org/10.1093/acref/9780191826719.001.0001>

RCUK. (2016). Excellence with impact - Research Councils UK. Retrieved from

<http://www.rcuk.ac.uk/innovation/impact/>

Rehfield, A. (2005). *The Concept of Constituency*. Cambridge University Press.

Retrieved from <http://books.google.co.uk/books?id=Y1GNBhfdGM4C>

Reich, R. (2018, November 20). Break up Facebook (and while we're at it, Google, Apple and Amazon). *The Guardian*. Retrieved from

<https://www.theguardian.com/commentisfree/2018/nov/20/facebook-google-antitrust-laws-gilded-age>

Relph, E. (1985). Geographical experiences and being-in-the-world: The phenomenological origins of geography. In D. Seamon & R. Mugerauer (Eds.), *Dwelling, Place and Environment: Towards a Phenomenology of Person and World* (pp. 15–31). Dordrecht: Springer Netherlands.

https://doi.org/10.1007/978-94-010-9251-7_2

Revelle, W. (2018). Procedures for Psychological, Psychometric, and Personality Research. Retrieved from <http://personality-project.org/r/psych-manual.pdf>

Roberts, I. (2016). Re: GATECloud.net: [Contact] Processing ~8m Datasift JSON

Tweets and Facebook Posts.

Roberts, I., & Tear, A. (2017). Question with GATEcloud jobs.

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases New Opportunities for Connected Data* (2nd ed.). Sebastopol: O'Reilly Media, Inc.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. Retrieved from <http://ije.oxfordjournals.org/content/38/2/337.short>

Rosi, A., Mamei, M., Zambonelli, F., Dobson, S., Stevenson, G., & Ye, J. (2011). Social sensors and pervasive services: Approaches and perspectives. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 525–530). IEEE. <https://doi.org/10.1109/PERCOMW.2011.5766946>

Rowe, I. (2015). Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media*, 59(4), 539–555. <https://doi.org/10.1080/08838151.2015.1093482>

Roy, S. D., & Zeng, W. (2015). *Social Multimedia Signals*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-319-09117-4>

Ruby. (2017). Ruby Programming Language. Retrieved June 6, 2017, from <https://www.ruby-lang.org/en/>

Rumsfeld, D. (2011). *Known and unknown: a memoir*. Penguin.

Russell, M. A. (2011). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol: O'Reilly Media, Inc.

Rzeszewski, M., & Beluch, L. (2017). Spatial Characteristics of Twitter Users—Toward the Understanding of Geosocial Media Production. *ISPRS International*

Journal of Geo-Information, 6(8), 236. <https://doi.org/10.3390/ijgi6080236>

Sabbagh, D. (2018, March 23). Rise of digital politics: why UK parties spend big on Facebook | Technology | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2018/mar/23/facebook-digital-politics-tories-labour-online-advertising-marketing>

Saif, H., He, Y., & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In *2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on the World Wide Web (WWW'12)* (Vol. 838, pp. 2–9). Retrieved from http://ceur-ws.org/Vol-838/paper_01.pdf

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, 52(1), 5–19. <https://doi.org/10.1016/j.ipm.2015.01.005>

Salus, P. H. (1995). *Casting the Net: From ARPANET to Internet and Beyond...* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Sandvoss, C. (2013). Toward an understanding of political enthusiasm as media fandom: Blogging, fan productivity and affect in American politics. *Participations*, 10(1), 252–296. Retrieved from [http://participations.org/Volume 10/Issue 1/12a Sandvoss 10 1.pdf](http://participations.org/Volume%2010/Issue%201/12a%20Sandvoss%2010%201.pdf)

Sanger, D. E., & Shane, S. (2016, December 9). Russian Hackers Acted to Aid Trump in Election, U.S. Says - The New York Times. *The New York Times*. Retrieved from <http://www.nytimes.com/2016/12/09/us/obama-russia-election-hack.html>

Sarver, R. (2009). Think Globally, Tweet Locally. Retrieved August 31, 2016, from <https://blog.twitter.com/2009/think-globally-tweet-locally>

SAS. (2014). Download SAS University Edition | SAS. Retrieved October 31, 2014, from http://www.sas.com/en_us/software/university-edition/download-software.html

Scharl, A. (2007). Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories. In A. Scharl & K. Tochtermann (Eds.), *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society* (pp. 3–14). London: Springer London. https://doi.org/10.1007/978-1-84628-827-2_1

Schrage, E. (2017). Hard Questions: Russian Ads Delivered to Congress. Retrieved November 6, 2017, from <https://newsroom.fb.com/news/2017/10/hard-questions-russian-ads-delivered-to-congress/>

Schwartz, M. J. (2013). Facebook Stalking Fears: 6 Geotagging Facts - InformationWeek. Retrieved March 21, 2016, from <http://www.informationweek.com/mobile/facebook-stalking-fears-6-geotagging-facts/d/d-id/1111161?>

Scotese, C. R. (2004). A Continental Drift Flipbook. *The Journal of Geology*, 112(6), 729–741. <https://doi.org/10.1086/424867>

Scott, J. (2017). *Social Network Analysis*. SAGE Publications. Retrieved from <https://uk.sagepub.com/en-gb/eur/social-network-analysis/book249668>

Seamon, D., & Lundberg, A. (2017). Humanistic Geography. In D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, & R. A. Marston (Eds.), *International Encyclopedia of Geography: People, the Earth, Environment and Technology* (pp. 1–11). Oxford, UK: John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118786352.wbieg0412>

Searles, K., Smith, G., & Sui, M. (2018). Partisan Media, Electoral Predictions, and Wishful Thinking. *Public Opinion Quarterly*, 82(Special Issue), 302–324.

<https://doi.org/10.1093/poq/nfy006>

Selvin, H. C. (1958). Durkheim's Suicide and problems of empirical research.

American Journal of Sociology, 63(6), 607–619. Retrieved from
<http://www.jstor.org/stable/10.2307/2772991>

Severance, C. (2012). Discovering JavaScript Object Notation. *Computer*, 45(4), 6–8.

<https://doi.org/10.1109/MC.2012.132>

Shavitt, Y., & Zilberman, N. (2010). A study of geolocation databases. *ArXiv Preprint ArXiv:1005.5674*.

Shehata, A., & Strömbäck, J. (2018). Learning Political News From Social Media: Network Media Logic and Current Affairs News Learning in a High-Choice Media Environment. *Communication Research*, 1–12.

<https://doi.org/10.1177/0093650217749354>

Shi, G., & Barker, K. (2011). Extraction of geospatial information on the Web for GIS applications. In *IEEE 10th International Conference on Cognitive Informatics and Cognitive Computing (ICCI-CC'11)* (pp. 41–48). IEEE.

<https://doi.org/10.1109/COGINF.2011.6016120>

Shields, R. (2012). Cultural Topology: The Seven Bridges of Konigsburg, 1736.

Theory, Culture & Society, 29(4–5), 43–57.

<https://doi.org/10.1177/0263276412451161>

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2017). Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society*, 19(8), 1214–1235.

<https://doi.org/10.1177/1461444816634054>

Shneiderman, B. (2007). Web science. *Communications of the ACM*, 50(6), 25.

<https://doi.org/10.1145/1247001.1247022>

Silva, S. S. C., Silva, R. M. P., Pinto, R. C. G., & Salles, R. M. (2013). Botnets: A survey.

Computer Networks, 57(2), 378–403.

<https://doi.org/10.1016/j.comnet.2012.07.021>

Skeen, B. (2017). *Including Geolocation Information in IPv6 Packet Headers (IPv6 GEO)* (Network Working Group). Retrieved from <https://tools.ietf.org/id/draft-skeen-6man-ipv6geo-03.html>

Sleight, P. (2004). *Targeting customers: How to use geodemographic and lifestyle data in your business*. World Advertising Research Center Henley-on-Thames.

Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLoS ONE*, 10(11), 1–15.

<https://doi.org/10.1371/journal.pone.0142209>

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE*, 10(3), 1–20.

<https://doi.org/10.1371/journal.pone.0115545>

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, 18(3), 1–11.

<https://doi.org/10.5153/sro.3001>

Small, H., Kasianovitz, K., Blanford, R., & Celaya, I. (2012). What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data. *International Journal of Digital Curation*, 7(1), 174–197.

<https://doi.org/10.2218/ijdc.v7i1.224>

Smart, P. D., Jones, C. B., & Twaroch, F. A. (2010). Multi-source toponym data integration and mediation for a meta-gazetteer service. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence*

and Lecture Notes in Bioinformatics), 6292 LNCS, 234–248.

https://doi.org/10.1007/978-3-642-15300-6_17

Smith, A. F., & Carlisle, J. (2015). Reviews, systematic reviews and Anaesthesia. *Anaesthesia*, 70(6), 644–650. <https://doi.org/10.1111/anae.13096>

Smith, A., & Gaur, M. (2018). What’s my age?: Predicting Twitter User’s Age using Influential Friend Network and DBpedia. *Arxiv Preprint*, 1–16. Retrieved from <https://arxiv.org/pdf/1804.03362.pdf>

Snapchat. (2018a). Snapchat Ads | Audiences. Retrieved from <https://forbusiness.snapchat.com/audiences/>

Snapchat. (2018b). Snapchat Developers - Ads. Retrieved from <https://developers.snapchat.com/ads/>

Social Data Science Laboratory. (2016). Lab Online Guide to Social Media Research Ethics. Retrieved February 25, 2017, from <http://socialdatalab.net/ethics-resources>

Southall, H., Mostern, R., & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2), 127–145. <https://doi.org/10.3366/ijhac.2011.0028>

Southall, H., von Lunen, A., & Aucott, P. (2009). On the organisation of geographical knowledge: Data models for gazetteers and historical GIS. In *2009 5th IEEE International Conference on E-Science Workshops* (pp. 162–166). IEEE. <https://doi.org/10.1109/ESCIW.2009.5407970>

Spector, P. (2018). Using t-tests in R | Department of Statistics. Retrieved August 3, 2018, from <https://statistics.berkeley.edu/computing/r-t-tests>

SQLite. (2016). SQLite. Retrieved September 2, 2016, from <https://sqlite.org/>

- Stackexchange. (2016). geolocation - How does Google know where I am? - Information Security Stack Exchange. Retrieved September 10, 2018, from <https://security.stackexchange.com/questions/137418/how-does-google-know-where-i-am>
- Stahl, B. C. (2004). Whose discourse? A comparison of the Foucauldian and Habermasian concepts of discourse in critical IS research. In *AMCIS 2004 Proceedings* (Vol. 538, pp. 4329–4336).
- Stanford University. (2017). Stanford CoreNLP – Core natural language software | Stanford CoreNLP. Retrieved June 10, 2017, from <https://stanfordnlp.github.io/CoreNLP/>
- Statista. (2018a). Most popular social networks worldwide as of April 2018, ranked by number of active users (in millions). Retrieved May 29, 2018, from <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Statista. (2018b). Twitter - Statistics & Facts | Statista. Retrieved January 31, 2018, from <https://www.statista.com/topics/737/twitter/>
- Stefanidis, A., Cotnoir, A., Croitoru, A., Crooks, A., Rice, M., & Radzikowski, J. (2013). Demarcating New Boundaries: Mapping Virtual Polycentric Communities through Social Media Content. *Cartography and Geographic Information Science*, 40(2), 116–129. <https://doi.org/10.1080/15230406.2013.776211>
- Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2), 319–338. <https://doi.org/10.1007/s10708-011-9438-2>
- Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, 19(6), 809–834. <https://doi.org/10.1111/tgis.12132>

- Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54, 255–265.
<https://doi.org/10.1016/j.compenvurbsys.2015.09.007>
- Stock, K. (2018). Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, 71, 209–240.
<https://doi.org/10.1016/j.compenvurbsys.2018.05.007>
- Stolte, C., Tang, D., & Hanrahan, P. (2002). Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 52–65.
<https://doi.org/10.1109/2945.981851>
- Stone, B. (2006). Introducing the Twitter API. Retrieved from
https://blog.twitter.com/official/en_us/a/2006/introducing-the-twitter-api.html
- Stretch, C. (2017). Facebook to Provide Congress With Ads Linked to Internet Research Agency. Retrieved November 6, 2017, from
<https://newsroom.fb.com/news/2017/09/providing-congress-with-ads-linked-to-internet-research-agency/>
- Sugira, L., Carpenter, D., Evans, H., & Parry, J. (2016). The Ethical Challenges of Internet Mediated Research (IMR). In *Biennial University Research Ethics and Governance Conference*. Portsmouth, UK, 24 June 2016: University of Portsmouth. Retrieved from <https://www.eventbrite.co.uk/e/research-ethics-and-governance-conference-2016-tickets-25406021093>
- Sui, D. Z. (2017). Understanding locational-based services: core technologies, key applications and major concerns. In B. Warf (Ed.), *Handbook on Geographies of Technology* (p. 85). Cheltenham, UK: Edward Elgar Publishing. Retrieved from

<http://www.e-elgar.com/shop/handbook-on-geographies-of-technology>

Sui, D. Z., & Goodchild, M. F. (2001). GIS as media? *International Journal of Geographical Information Science*, 5(15), 387–390.

<https://doi.org/10.1080/13658810110038924>

Sui, D. Z., & Goodchild, M. F. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737–1748. <https://doi.org/10.1080/13658816.2011.604636>

Sunstein, C. R. (2018). *#Republic : divided democracy in the age of social media*.

Princeton University Press. Retrieved from

<https://press.princeton.edu/titles/10935.html#.W0NBynt5QDw.mendeley>

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday.

Swaine, J. (2018, August 27). Roger Stone says he may soon be indicted in Trump-Russia investigation | US news | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/us-news/2018/aug/27/roger-stone-trump-russia-investigation-mueller-indictment-latest-news>

Swanson, R. A., & Holton, E. F. (2005). *Research in Organizations: Foundations and Methods in Inquiry*. Berrett-Koehler Publishers. Retrieved from

<https://books.google.co.uk/books?id=AyMZt9AodEEC>

Swirsky, E. S., Hoop, J. G., & Labott, S. (2014). Using Social Media in Research: New Ethics for a New Meme? *The American Journal of Bioethics*, 14(10), 60–61.

<https://doi.org/10.1080/15265161.2014.948302>

Tablan, V., Roberts, I., Cunningham, H., & Bontcheva, K. (2012). GATECloud.net: a platform for large-scale, open-source text processing on the cloud.

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1983), 1–13.

<https://doi.org/10.1098/rsta.2012.0071>

Tableau. (2017a). Academic Programs | Tableau Software. Retrieved March 26, 2017, from <https://www.tableau.com/academic>

Tableau. (2017b). Business Intelligence and Analytics | Tableau Software. Retrieved February 1, 2017, from <https://www.tableau.com/>

Tableau. (2017c). Tableau Technology | Tableau Software. Retrieved May 6, 2017, from <https://www.tableau.com/products/technology>

Tao, X., Zhou, X., Lau, C. H., & Li, Y. (2013). Personalised Information Gathering and Recommender Systems: Techniques and Trends. *ICST Transactions on Scalable Information Systems*, 13(1), e4. <https://doi.org/10.4108/trans.sis.2013.01-03.e4>

Tasse, D., Liu, Z., Sciuto, A., & Hong, J. I. (2017). State of the Geotags : Motivations and Recent Changes. In *Eleventh International AAAI Conference on Web and Social Media*. Retrieved from <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15588>

Taylor, D. (2018, September 2). Big Tech's double trouble: political heat from Trump and the left may signal reckoning ahead | Technology | The Guardian. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2018/sep/02/big-techs-double-trouble-bipartisan-criticism-may-signal-a-reckoning-ahead>

Tear, A. (1997). www.election.co.uk - the creation of a geographically referenced web site for the 1997 General Election. In *Geographic Information - exploiting the benefits*, AGI 97. London: The Association for Geographic Information and Miller Freeman.

- Tear, A. (2013). Geographic and non-geographic social media interactions around the time of the 2012 US Presidential Election. In *RGS-IBG Annual International Conference 2013 - Big, Open Data and the Practice of GIScience*. London. Retrieved from <http://conference.rgs.org/conference/sessions/View.aspx?heading=Y&session=37eb9b96-0812-45c0-a5f6-a61c3b88f2fc>
- Tear, A. (2014). SQL or NoSQL? Contrasting Approaches to the Storage, Manipulation and Analysis of Spatio-temporal Online Social Network Data. In B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, ... O. Gervasi (Eds.), *Computational Science and Its Applications -- ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30 -- July 3, 2014, Proceedings, Part I* (Vol. 8579 LNCS, pp. 221–236). Springer International Publishing. https://doi.org/10.1007/978-3-319-09144-0_16
- Tear, A. (2016). Drilling for Big Data. In *19th AGILE International Conference on Geographic Information Science - Workshop "GIS with NoSQL."* Retrieved from https://www.unibw.de/inf4/professors/geoinformatics/agile-2016-workshop-gis-with-nosql?set_language=en
- Tear, A. (2017). *How to get Hive working*. Portsmouth.
- Tear, A., & Healey, R. G. (2017). Wading through the swamp: filter systems for geospatial data science. In *GISRUK 2017*. Manchester, 18th - 21st April 2017. Retrieved from <http://manchester.gisruk.org/proceedings.php>
- Tear, A., & Maynard, D. (2018). quick question.
- Tear, A., & Southall, H. (2019). Social media data. In J. Evans, S. Ruane, & H. Southall (Eds.), *Data in Society: Challenging Statistics in an Age of Globalisation*. Bristol: Policy Press.
- Teppan, E. C., & Zanker, M. (2015). Decision Biases in Recommender Systems.

Journal of Internet Commerce, 14(2), 255–275.

<https://doi.org/10.1080/15332861.2015.1018703>

The Electoral Commission. (2014). *Scottish independence referendum: Report on the referendum held on 18 September 2014*. Retrieved from

http://www.electoralcommission.org.uk/__data/assets/pdf_file/0010/179812/Scottish-independence-referendum-report.pdf

The Electoral Commission. (2016). Electoral Commission | EU referendum results.

Retrieved November 13, 2016, from

<http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information>

The Electoral Commission. (2018a). *Digital campaigning: Increasing transparency for voters*. Retrieved from

https://www.electoralcommission.org.uk/__data/assets/pdf_file/0010/244594/Digital-campaigning-improving-transparency-for-voters.pdf

The Electoral Commission. (2018b). *Report of an investigation in respect of - Vote Leave Limited - Mr Darren Grimes - BeLeave - Veterans for Britain Concerning campaign funding and spending for the 2016 referendum on the UK's membership of the EU*. Retrieved from

https://www.electoralcommission.org.uk/__data/assets/pdf_file/0019/244900/Report-of-an-investigation-in-respect-of-Vote-Leave-Limited-Mr-Darren-Grimes-BeLeave-and-Veterans-for-Britain.pdf

The Guardian. (2018). Edward Snowden. Retrieved May 24, 2018, from

<https://www.theguardian.com/us-news/edward-snowden>

The Herald. (2017, May 10). New Tory councillor unmasked as influential arch-

'BritNat' Twitter troll who boasts about his manhood online | HeraldScotland.

The Herald. Retrieved from

http://www.heraldscotland.com/news/15276119.New_Tory_councillor_unmasked_as_influential_arch__BritNat__Twitter_troll_who_boasts_about_his_manhhood_online/

The Independent. (2018, April 11). Mark Zuckerberg admits his own personal Facebook data was harvested in data abuse scandal | The Independent. *The Independent*. Retrieved from <https://www.independent.co.uk/life-style/gadgets-and-tech/news/zuckerberg-hearing-facebook-data-latest-cambridge-analytica-mark-a8299836.html>

The New York Times. (2018). Trump and the Russians - The New York Times. Retrieved September 4, 2018, from <https://www.nytimes.com/spotlight/trump-russia>

The R Foundation. (2018). R: The R Project for Statistical Computing. Retrieved July 11, 2018, from <https://www.r-project.org/>

Thirifays, A., Lux, Z., Škofljanec, J., Završnik, G., Paulič, A., Bo Nielsen, A., ... Anderson, D. (2018). *E-ARK Dissemination Information Package (DIP) Final Specification*. Zenodo. <https://doi.org/10.5281/zenodo.1172968>

Thomassen, L. (2010). *Habermas : a guide for the perplexed*. London; New York, NY: Continuum.

Till, B. C., Longo, J., Dobell, a. R., & Driessen, P. F. (2014). Self-organizing maps for latent semantic analysis of free-form text in support of public policy analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1), 71–86. <https://doi.org/10.1002/widm.1112>

Tolbert, C. J., & McNeal, R. S. (2003). Unraveling the Effects of the Internet on Political Participation? *Political Research Quarterly*, 56(2), 175–185. Retrieved from <http://www.jstor.org/stable/3219896>

- Trafford, V., & Leshem, S. (2008). *Stepping stones to achieving your doctorate focusing on your viva from the start*. Maidenhead; New York: Open University Press ; McGraw-Hill. Retrieved from <http://site.ebrary.com/id/10274031>
- Tridimas, G. (2011). Cleisthenes' Choice: The Emergence of Direct Democracy in Ancient Athens. *The Journal of Economic Asymmetries*, 8(1), 39–59.
<https://doi.org/10.1016/j.jeca.2011.01.002>
- Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, 42(sup1), 70–74. <https://doi.org/10.1080/15230406.2015.1059251>
- Tsou, M.-H., & Leitner, M. (2013). Visualization of social media: seeing a mirage or a message? *Cartography and Geographic Information Science*, 40(2), 55–60.
<https://doi.org/10.1080/15230406.2013.776754>
- Tuan, Y.-F. (2001). *Space and Place: The Perspective of Experience* (2nd Ed). Minneapolis: University of Minnesota Press.
- Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... Nyhan, B. (2018). *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Retrieved from <https://hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf>
- Tucker, J. A., Theocharis, Y., Roberts, M. E., & Barberá, P. (2017). From Liberation to Turmoil: Social Media And Democracy. *Journal of Democracy*, 28(4), 46–59.
<https://doi.org/10.1353/jod.2017.0064>
- Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 505–514.
Retrieved from <http://arxiv.org/abs/1403.7400>

- Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Icwsn*, 178–185. <https://doi.org/10.1074/jbc.M501708200>
- Turing, A. M. (1950). COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, 49, 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turner, A. (2006). *Introduction to neogeography*. O'Reilly Media, Inc.
- Twitter. (2013a). How do I get firehose access? | Twitter Developers. Retrieved September 24, 2013, from <https://dev.twitter.com/discussions/2752>
- Twitter. (2013b). Overview: Version 1.1 of the Twitter API | Twitter Developers. Retrieved January 28, 2014, from <https://dev.twitter.com/docs/api/1.1/overview>
- Twitter. (2014). Geo Guidelines | Twitter Developers. Retrieved March 21, 2016, from <https://dev.twitter.com/overview/terms/geo-developer-guidelines>
- Twitter. (2017). Twitter Developer Documentation. Retrieved February 5, 2017, from <https://dev.twitter.com/overview/api/tweets#obj-coordinates>
- Twitter. (2018a). Geo-targeting by location, country, or language. Retrieved from <https://business.twitter.com/en/targeting/geo-and-language.html>
- Twitter. (2018b). Introduction to Tweet JSON - Twitter Developers. Retrieved from <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json>
- TwitterCounter. (2017). Top 100 Most Followed Users on Twitter. Retrieved June 3, 2017, from <https://twittercounter.com/pages/100>
- U.S. Census Bureau. (2017). Population and Housing Unit Estimates. Retrieved February 19, 2017, from <http://www.census.gov/programs->

surveys/popest.html

U.S. House of Representatives. (2018a). *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements*. Washington, DC. Retrieved from <https://democrats-intelligence.house.gov/social-media-content/>

U.S. House of Representatives. (2018b). *Facebook: Transparency and Use of Consumer Data*. Washington, DC. Retrieved from <https://energycommerce.house.gov/hearings/facebook-transparency-use-consumer-data/>

U.S. House of Representatives. (2018c). *Social Media Advertisements*. Washington, DC. Retrieved from <https://democrats-intelligence.house.gov/social-media-content/social-media-advertisements.htm>

U.S. Office of Special Counsel. (2018). U.S. Office of Special Counsel What We Do. Retrieved September 4, 2018, from <https://osc.gov/pages/whatwedo.aspx>

Ubuntu. (2014). Get Ubuntu | Download | Ubuntu. Retrieved October 31, 2014, from <http://www.ubuntu.com/download>

Uhlig, R., Neiger, G., Rodgers, D., Santoni, A. L., Martins, F. C. M., Anderson, A. V., ... Smith, L. (2005). Intel virtualization technology. *Computer*, 38(5), 48–56. <https://doi.org/10.1109/MC.2005.163>

Unwin, T. (2012). Social Media and Democracy: Critical Reflections. In *Background Paper for Commonwealth Parliamentary Conference* (pp. 1–8). Colombo. Retrieved from <http://www.cpahq.org/cpahq/cpadocs/Unwin CPA Social media and democracy.pdf>

US Census Bureau. (2010). 2010 American Community Survey / Puerto Rico Community Survey Group Quarters Definitions, 1–7. Retrieved from

http://www.census.gov/acs/www/Downloads/data_documentation/GroupDefinitions/2010GQ_Definitions.pdf

US Census Bureau. (2012a). 2010 Census Tallies of Census Tracts, Block Groups and Blocks. Retrieved May 13, 2017, from <https://www.census.gov/geo/maps-data/data/tallies/tractblock.html>

US Census Bureau. (2012b). TIGER/Line® with Demographic Data. Retrieved July 3, 2013, from <http://www.census.gov/geo/maps-data/data/tiger-data.html>

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: a systematic review via text mining. *Journal of Knowledge Management*, 22(7), 1471–1488. <https://doi.org/10.1108/JKM-11-2017-0517>

van der Aalst, W. M. P. (2014). Data Scientist: The Engineer of the Future. In K. Mertins, F. Bénaben, R. Poler, & J.-P. Bourrières (Eds.), *Enterprise Interoperability VI: Interoperability for Agility, Resilience and Plasticity of Collaborations* (pp. 13–26). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-04948-9_2

Van Diepen, C., Twigg, L., Ekinsmyth, C., & Moon, G. (2017). The use of Twitter to support public health activities surrounding adolescent smoking in Wales'. In *The Emerging New Researchers in the Geographies of Health and Impairment (ENRGHI)*. Glasgow.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197–208. <https://doi.org/10.24908/ss.v12i2.4776>

Van Grove, J. (2010). Are We All Asking to Be Robbed? Retrieved April 17, 2018, from <https://mashable.com/2010/02/17/pleaserobme/#k1O4XEG.Jqqt>

van Liere, D. (2010). How far does a tweet travel? In *Proceedings of the*

- International Workshop on Modeling Social Media - MSM '10* (pp. 1–4). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1835980.1835986>
- van Renesse, R. (2003). The Importance of Aggregation. In A. Schiper, A. A. Shvartsman, H. Weatherspoon, & B. Y. Zhao (Eds.), *Future Directions in Distributed Computing: Research and Position Papers* (pp. 87–92). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-37795-6_16
- Vassil, K., & Weber, T. (2011). A bottleneck model of e-voting: Why technology fails to boost turnout. *New Media & Society*, 13(8), 1336–1354. <https://doi.org/10.1177/1461444811405807>
- Venkatesh, V., Davis, F., & Morris, M. (2007). Dead Or Alive? The Development, Trajectory And Future Of Technology Adoption Research. *Journal of the Association for Information Systems*., 8(4), 267–286. Retrieved from <https://aisel.aisnet.org/jais/vol8/iss4/10>
- Vergeer, M. (2013). Politics, elections and online campaigning: Past, present . . . and a peek into the future. *New Media & Society*, 15(1), 9–17. <https://doi.org/10.1177/1461444812457327>
- Vissers, S., & Stolle, D. (2014). The Internet and new modes of political participation: online versus offline participation. *Information, Communication & Society*, 17(8), 937–955. <https://doi.org/10.1080/1369118X.2013.867356>
- Voas, D., & Williamson, P. (2001). The Diversity of Diversity: A Critique of Geodemographic Classification. *Area*, 33(1), 63–76. <https://doi.org/10.1111/1475-4762.00009>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>

- Wachowicz, M., & Liu, T. (2016). Finding spatial outliers in collective mobility patterns coupled with social ties. *International Journal of Geographical Information Science*, 30(9), 1–26.
<https://doi.org/10.1080/13658816.2016.1144887>
- Wallace, G., & Yoon, R. (2016). Voter turnout at 20-year low in 2016. Retrieved November 13, 2016, from
<http://edition.cnn.com/2016/11/11/politics/popular-vote-turnout-2016/>
- Wallace, M. (2015). The computers that crashed. And the campaign that didn't. The story of the Tory stealth operation that outwitted Labour last month | Conservative Home. Retrieved January 8, 2019, from
<https://www.conservativehome.com/thetorydiary/2015/06/the-computers-that-crashed-and-the-campaign-that-didnt-the-story-of-the-tory-stealth-operation-that-outwitted-labour.html>
- Wallace, M., & Wray, A. (2011). *Critical reading and writing for postgraduates*. London: SAGE.
- Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 115–120). Jeju, Republic of Korea. Retrieved from <http://delivery.acm.org/10.1145/2400000/2390490/p115-wang.pdf>
- Wang, S. (2013). CyberGIS: blueprint for integrated and scalable geospatial software ecosystems. *International Journal of Geographical Information Science*, 27(11), 2119–2121. <https://doi.org/10.1080/13658816.2013.841318>
- Ward, K. (2018). Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting. *Journal of Media Ethics*, 33(3), 133–148.

<https://doi.org/10.1080/23736992.2018.1477047>

Ware, M., & Mabe, M. (2015). *The STM Report: An overview of scientific and scholarly journal publishing*. Prins Willem Alexanderhof 5, The Hague, 2595BE, The Netherlands.

Warf, B., & Sui, D. Z. (2010). From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS*, 16(4), 197–209.
<https://doi.org/10.1080/19475683.2010.539985>

Webber, R. (2004). Designing geodemographic classifications to meet contemporary business needs. *Interactive Marketing*, 5(3), 219–237.
<https://doi.org/10.1057/palgrave.im.4340240>

Weber, L. M., Loumakis, A., & Bergman, J. (2003). Who Participates and Why?: An Analysis of Citizens on the Internet and the Mass Public. *Social Science Computer Review*, 21(1), 26–42. <https://doi.org/10.1177/0894439302238969>

Weber, M. S., & Monge, P. (2011). The flow of digital news in a network of sources, authorities, and hubs. *Journal of Communication*, 61(6), 1062–1081.
<https://doi.org/10.1111/j.1460-2466.2011.01596.x>

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii. Retrieved from <https://www.jstor.org/stable/4132319>

Webtrends. (2018). Website Measurement & Analytics | Web Optimization | Webtrends. Retrieved April 24, 2018, from <https://www.webtrends.com/>

Weeks, B. E., Ardèvol-Abreu, A., & de Zúñiga, H. G. (2015). Online Influence? Social Media Use, Opinion Leadership, and Political Persuasion. *International Journal of Public Opinion Research*, edv050. <https://doi.org/10.1093/ijpor/edv050>

Wei, R. (2013). Mobile media: Coming of age with a big splash. *Mobile Media &*

- Communication*, 1(1), 50–56. <https://doi.org/10.1177/2050157912459494>
- Westen, T. (1998). Can Technology Save Democracy? *National Civic Review*, 87(1), 47–56. <https://doi.org/10.1002/ncr.87103>
- Wiener, J., & Bronson, N. (2014). Facebook's Top Open Data Problems - Facebook Research. Retrieved January 31, 2018, from <https://research.fb.com/facebook-s-top-open-data-problems/>
- Wilken, R. (2012). Locative media: From specialized preoccupation to mainstream fascination. *Convergence: The International Journal of Research into New Media Technologies*, 18(3), 243–247. <https://doi.org/10.1177/1354856512444375>
- Wilkinson, L. (1999). *The Grammar of Graphics*. New York, NY: Springer New York. <https://doi.org/10.1007/978-1-4757-3100-2>
- Williams, M. L. (2015). Towards an Ethical Framework for Using Social Media Data in Social Research. In *Social Research Association Ethics Workshop*. Institute of Education, UCL, 15 June 2015. Retrieved from <http://socialdatalab.net/wp-content/uploads/2016/08/EthicsSM-SRA-Workshop.pdf>
- Wilson, M. W., & Graham, M. (2013). Situating neogeography. *Environment and Planning A*, 45(1), 3–9. <https://doi.org/10.1068/a444482>
- Wilson, S. (2018). UK in EU Challenge. Retrieved September 4, 2018, from <https://www.crowdjustice.com/case/ukineuchallenge/>
- Wolfram, D. (2006). Applications of SQL for informetric frequency distribution processing. *Scientometrics*, 67(2), 301–313. <https://doi.org/10.1007/s11192-006-0101-5>
- Woodfield, K., Morrell, G., Metzler, K., & Blank, G. (2013). *Blurring the boundaries? New social media, new social research: Developing a network to explore the*

issues faced by researchers negotiating the new research landscape of online social media. NCRM Networks for Methodological Innovation Report.

Retrieved from <http://eprints.ncrm.ac.uk/3168/>

Work, D. B., Blandin, S., Tossavainen, O. P., Piccoli, B., & Bayen, a. M. (2010). A Traffic Model for Velocity Data Assimilation. *Applied Mathematics Research EXpress*, 35. <https://doi.org/10.1093/amrx/abq002>

World Wide Web Consortium. (2018). Geolocation API Specification 2nd Edition. Retrieved January 7, 2019, from <https://www.w3.org/TR/geolocation-API/>

Worldometers. (2018). Worldometers - real time world statistics. Retrieved January 31, 2018, from <http://www.worldometers.info/>

Wyly, E. (2014). The new quantitative revolution. *Dialogues in Human Geography*, 4(1), 26–38. <https://doi.org/10.1177/2043820614525732>

Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the “geographical awareness” of individuals: an exploratory analysis of twitter data. *Cartography and Geographic Information Science*, 40(2), 103–115. <https://doi.org/10.1080/15230406.2013.776212>

Yergeau, F. (2003). *UTF-8, a transformation format of ISO 10646* (Network Working Group No. 3629). Retrieved from <http://tools.ietf.org/html/rfc3629>

Yin, J., Soliman, A., Yin, D., & Wang, S. (2017). Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, 31(7), 1293–1313. <https://doi.org/10.1080/13658816.2017.1282615>

Yousfi, S., Chiadmi, D., & Nafis, F. (2016). Toward a Big Data-as-a-Service for Social Networks Graphs Analysis. In A. El Oualkadi, F. Choubani, & A. El Moussati (Eds.), *Proceedings of the Mediterranean Conference on Information &*

Communication Technologies 2015: MedCT 2015 Volume 2 (pp. 593–598).

Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30298-0_63

YouTube. (2018a). About targeting for video campaigns - YouTube Help. Retrieved September 4, 2018, from <https://support.google.com/youtube/answer/2454017?hl=en-GB>

YouTube. (2018b). Cummings - Why Leave Won the Referendum. Retrieved January 8, 2019, from <https://www.youtube.com/watch?v=CDbRxH9Kiy4>

Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M.-. (2013). Who, where, when and what. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13* (p. 605). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2487575.2487576>

Zelenkauskaite, A., & Bucy, E. P. (2016). A scholarly divide: Social media, Big Data, and unattainable scholarship. *First Monday*, 21(5). <https://doi.org/10.5210/fm.v21i5.6358>

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, 9(9), 37–70. <https://doi.org/10.5311/JOSIS.2014.9.170>

Appendix 1 MAXMIND GEOIP CODING

A1.1 Introduction

The map in Figure 1-1 (p23) results from the processing of 426,747 rows of web server log data retrieved from a Digital Audio Tape (DAT) backup of the UK 1997 General Election website created on 26/03/1997. These data, the only recoverable fragment of a much larger original set of web server logs, were loaded into Microsoft Access. 6,593 distinct IP addresses were recorded, and these were passed through the MaxMind GeoIP RESTful API using the ColdFusion code below.

A1.2 ColdFusion code

A1.2.1 application.cfm

```
<!--- licence key for MaxMind CityIP webservice --->
<cfparam name="LicenceKey" default="80BYiNRaYo1y">

<!--- myDSN --->
<cfparam name="myDSN"
default="Election97_log_analysis">
```

A1.2.2 process_ips.cfm

```
<!--- get some records --->
<cfquery name="get_some" datasource="#myDSN#">
  select top 50 * from distinctips where processed=0
</cfquery>

<!--- some counters --->
<cfset found_counter = 0>
<cfset unfound_counter =0>

<!--- loop over the records to geocode them --->
<cfif get_some.recordcount is not 0>
  <cfloop query="get_some">
    <cfset this_ip = get_some.ipaddress>
```

```

        <cfoutput><br />Working on #this_ip#...<br
/></cfoutput>
        <cfinclude
template="_process_one_ip_address.cfm">
    </cfloop>
    <br />
    <cfoutput>Done #get_some.recordcount#! Found
#found_counter#, did not find
#unfound_counter#!</cfoutput>
</cfif>

```

A1.2.3 _process_one_ip_address.cfm

```

<!--- contributed by reinhard jung --->
<cfhttp method="get"
url="http://geoip1.maxmind.com/f?l=#LicenceKey#&i=#this
_ip#"></cfhttp>
<cfset resultMaxMind = cfhttp.FileContent>

<!---><cfoutput>#resultMaxMind#</cfoutput>--->

<!--- if string does not contain IP_NOT_FOUND --->
<cfif findnocase('IP_NOT_FOUND', resultMaxMind) is 0>

    <!--- create Array --->
    <cfset qMaxMindByID = structNew() />
    <cfset qMaxMindByName = structNew() />
    <cfset thisField =
"country,region,city,postal,latitude,longitude,metroCod
e,area,ISP,organization"/>
    <cfset thisPos = 1/>
    <cfset thisValue = ""/>
    <cfset stringField = "false"/>
    <cfloop from="1" to="#Len(resultMaxMind)#"
index="mmField">
        <cfif mid(resultMaxMind,mmField,1) IS ','
AND NOT stringField>
            <cfset qMaxMindByID[thisPos] =
thisValue>
            <cfset
qMaxMindByName['#ListgetAt(thisField,thisPos)#'] =
thisValue>
            <cfset thisPos = thisPos +1/>
            <cfset thisValue = ""/>

```

```

        <cfelse>
            <cfif mid(resultMaxMind,mmField,1)
IS ''>
                <cfset stringField =
iif(stringField,"false","true")/>
                <cfelse>
                    <cfset thisValue =
thisValue &mid(resultMaxMind,mmField,1)/>
                </cfif>
            <cfif Len(resultMaxMind) EQ mmField>
                <cfset qMaxMindByID[thisPos] =
thisValue/>
                <cfset
qMaxMindByName['#ListgetAt(thisField,thisPos)#'] =
thisValue>
            </cfif>
        </cfloop>

        <!--- access Array
<br /><cfoutput>#qMaxMindByID[3]#</cfoutput>
<br
/><cfoutput>#qMaxMindByName['city']#</cfoutput>--->

        <!--- dump Array for overview
<cfdump var="#qMaxMindByID#"
label="qMaxMindByID"><br />
        <cfdump var="#qMaxMindByName#"
label="qMaxMindByName"><br />--->

        <!--- update the database --->
        <cfquery name="update_row_with_data"
datasource="#myDSN#">
            update distinctips
            set
                country='#qMaxMindByName['country']#',
                region='#qMaxMindByName['region']#',
                city='#qMaxMindByName['city']#',
                postal='#qMaxMindByName['postal']#',
                latitude=#qMaxMindByName['latitude']#,

longitude=#qMaxMindByName['longitude']#,

metrocode='#qMaxMindByName['metrocode']#',
                area='#qMaxMindByName['area']#',
                isp='#qMaxMindByName['isp']#',

```

```
organisation='#qMaxMindByName['organization']#',
    processed=1,
    processed_datetime=getdate()
    where
    ipaddress='#this_ip#'
</cfquery>
<cfset found_counter = found_counter + 1>
...found!

<cfelse>

<!--- mark the row as done --->
<cfquery name="update_row_no_data"
datasource="#myDSN#">
    update distinctips set processed=1,
processed_datetime=getdate() where
ipaddress='#this_ip#'
</cfquery>
<cfset unfound_counter = unfound_counter + 1>
...not found!

</cfif>
```

Appendix 2 WORD CLOUD GENERATION

A2.1 Introduction

Word Clouds offer an attractive mechanism for visualising the relative importance, or weight, of terms appearing in a corpus of written material (McNaught & Lam, 2010). Over 1,250 academic journal, book, chapter and/or online references (summarised in Chapter 2, p51) have been collected and stored in Mendeley Desktop reference management software as part of this research. Computational analysis of document titles provides a useful overview of key terms in the corpus.

A2.2 BibTeX file-based processing

Mendeley provides the facility (under File -> Export...) to export selected references. By selecting all references (Edit -> Select All) and hitting File -> Export... all references will be saved to a file (`My Collection.bib`) in BibTeX format, a widely used standard for the interchange of academic references (Feder, 2006). As BibTeX is a format consisting of marked-up plain text (Figure A2-1) it is possible to search the backup file for all occurrences of the phrase `'title = {'` to build a list of document titles.

```
@article{Wilson2012,
author = {Wilson, Matthew W.},
doi = {10.1016/j.geoforum.2012.03.014},
file = {:C$\backslash$:/Users/Adrian
Tear/Documents/Mendeley Desktop/Wilson - 2012 -
Location-based services, conspicuous mobility, and the
location-aware future.pdf:pdf},
issn = {00167185},
journal = {Geoforum},
keywords = {location-based services},
month = {nov},
number = {6},
pages = {1266--1275},
publisher = {Elsevier Ltd},
```

```

title = {{Location-based services, conspicuous
mobility, and the location-aware future}},
url =
{http://linkinghub.elsevier.com/retrieve/pii/S001671851
2000747},
volume = {43},
year = {2012}
}

```

Figure A2-1 – Snippet of a BibTeX file

Using a Linux VM the following steps were executed:

- In Terminal run `grep 'title = {{\' My\ Collection.bib > 01titles.txt` to split title line text out into a separate file
- Edit `01titles.txt` to find/replace and remove the following text:
 - `title = {{`
 - `}},`
 - `{\\textregistered}`
 - `{\\textperiodcentered}`
 - `{\\ldots}`
 - `{\`
 - `}`
- In Terminal run `awk '{print tolower($0)}' 01titles.txt > 02titleslcase.txt` to convert all text to lowercase

Text in the resulting file may then be copied/pasted or uploaded to any one of the many Word Cloud generators on the Internet (e.g., Davies, 2018). The approach would work for any BibTeX-formatted file.

A2.3 SQLite database-based processing

As Mendeley Desktop is built around a SQLite database an easier alternative to BibTeX file-based processing involves opening, querying and saving results from the Mendeley SQLite database using the followings steps:

- Backup references using the Help -> Create Backup... option in Mendeley
- Extract the [username]@www.mendeley.com.sqlite file in the resulting ZIP archive to a working directory
- Open the .SQLITE file using DB Browser for SQLite or similar and execute the query `select lower(title) from documents`

The screen in DB Browser for SQLite should be similar to that shown in Figure A2-2.

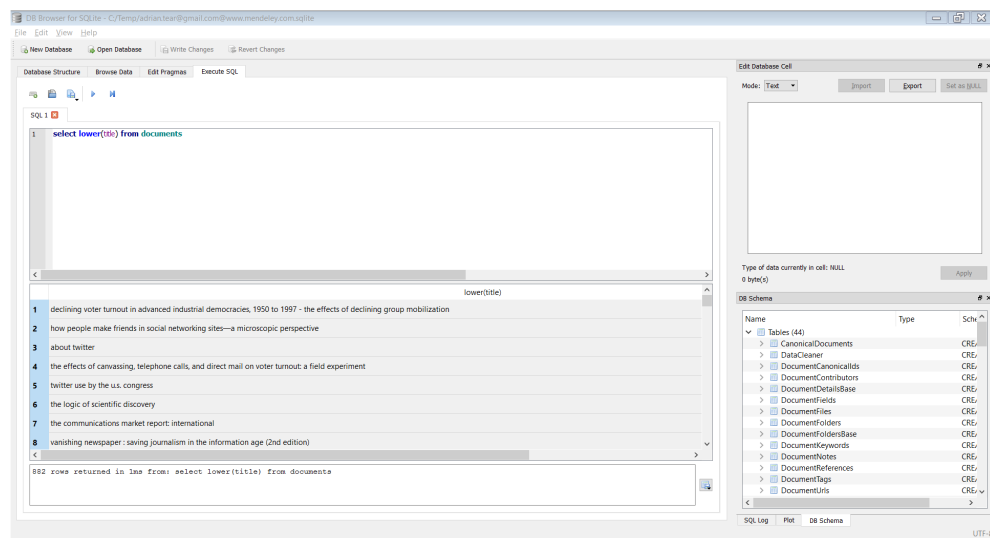


Figure A2-2 – Querying the Mendeley SQLite database to return lowercase article titles

Output from the query may be saved/uploaded or copied/pasted into one of the many Word Cloud generators found online (e.g., Davies, 2018) or read into R for further analysis.

Appendix 3 ACADEMIC LITERATURE TEXT-MINING

A3.1 Introduction

Programming techniques from several sources (academic, online and software training) have been combined to produce text-mining outputs detailed in Section 2.2.2 (p57). The methods are described below.

A3.2 Preparation

Mendeley stores saved documents (typically Adobe PDF files) in the Windows directory `C:\Users\[username]\Documents\Mendeley Desktop`. Files can be read from this location directly by a virtual machine.

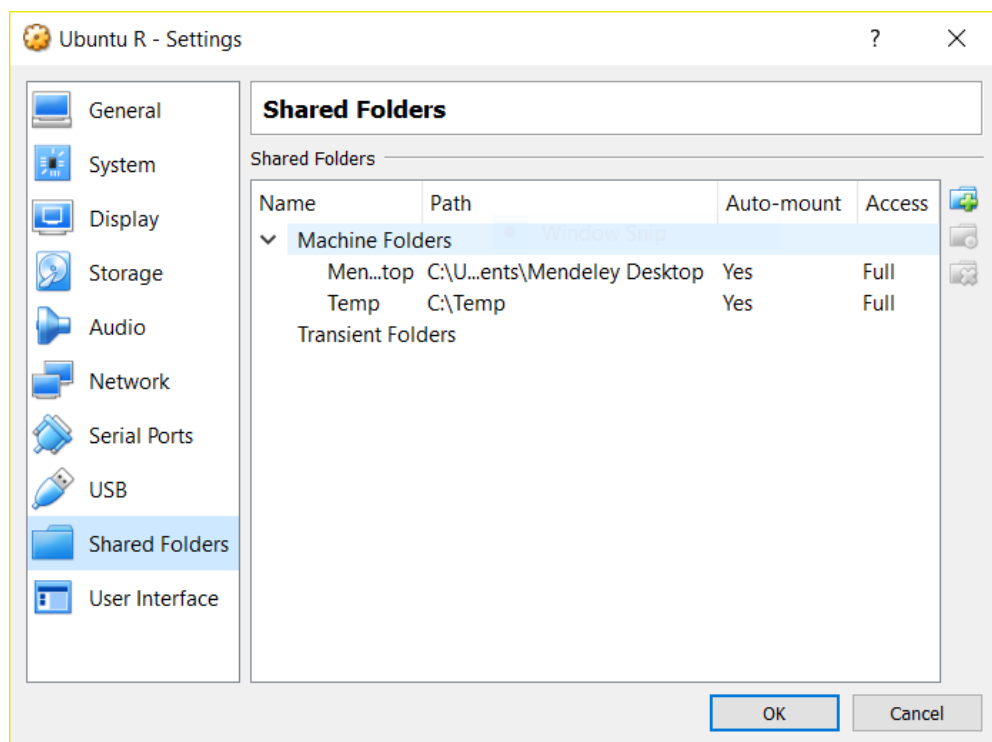


Figure A3-1 – Mapping the working directory in VirtualBox to a Shared Folder on the Ubuntu virtual machine

The directory should be mapped as a Shared Folder in Oracle VM VirtualBox Manager (Figure A3-1) to a Linux virtual machine running R, in this case an Ubuntu 16.04 LTS box configured with 4GB of Random Access Memory (RAM) and 2 processors. All file operations on the PDF files may then be completed in R on the VM, through the RStudio interface, using the scripts and required R libraries below.

A3.3 Corpus creation

The R script `001 - read LIVE corpus from PDF files.R` resolves a list of PDF files to read from `/media/sf_Mendeley_Desktop` (the live Mendeley PDF document repository) and uses the `readPDF` construct to create a corpus from these files.

```
#####
#####
#
# VM-UBUNTU-R
#
#####
#####

# load tm
library(tm)

# Set the working directory
setwd("/media/sf_Mendeley_Desktop")

# list of filenames
filenames = list.files(getwd(),pattern="\\.pdf")
cat(filenames)

# adapted from
http://data.library.virginia.edu/reading-pdf-files-
into-r-for-text-mining/
Rpdf <- readPDF(control = list(text = "-layout"))

# make the corpus
lit_corpus <- Corpus(URISource(filenames),
readerControl = list(reader = Rpdf))
```

The resulting data set, `lit_corpus`, is a Large VCorpus (Volatile Corpus) with n documents.

A3.4 Corpus analysis

The R script `002 - analyse LIVE corpus from PDF files.R` is used to process and analyse text in the PDF documents to produce counts and visualisations of word occurrences in the corpus. It depends upon several packages (`tm`, `ggplot2`, `wordcloud2`) which must be installed (with dependencies) in RStudio.

```
#####
#####
#
# VM-UBUNTU-R
#
#####
#####

# parts from https://github.com/juliasilge/tidytext

# install these packages using R Studio
library(tm)
library(ggplot2)
library(wordcloud2)

# remove punctuation, whitespace etc.
# use content_transformer(tolower) as
tm_map(lit_corpus, tolower) destroys the corpus!
doc.corpus <- tm_map(lit_corpus,
content_transformer(tolower))

# the rest of it
doc.corpus <- tm_map(doc.corpus, removePunctuation)
doc.corpus <- tm_map(doc.corpus, removeNumbers) # THIS
MAY BE UNHELPFUL WHEN TRYING TO FIND PAGE NUMBERS!!

# stopwords
# get the list of unwanted abundant characters by
running through without this removal first and looking
at findFreqTerms results
doc.corpus <- tm_map(doc.corpus, removeWords,
c(stopwords("english"), "figure", "table", "management", "p
roceedings", "research", "acm", "pages", "items", "topic", "j
```



```

p <- ggplot(subset(wf, freq>8000), aes(word, freq))
p <- p + geom_bar(stat="identity")
p <- p + theme(axis.text.x=element_text(angle=45,
hjust=1))
p

# as a wordcloud (adjust freq as required)
# pretty it up a bit, from https://cran.r-
project.org/web/packages/wordcloud2/vignettes/wordcloud
.html (random-light/random-dark etc.)
# black background, light text
wordcloud2(subset(wf, freq>=4000), color = "random-
light", backgroundColor = "black", size=0.4)

# white background, dark text
wordcloud2(subset(wf, freq>=4000), color = "random-
dark", backgroundColor = "white", size=0.4)

```

Selected output from this script is presented in the main body of the thesis in Section 2.2.2 (p57).

A3.5 Word frequency export

The R script 003 - export WF.R exports the data frame WF, containing word frequencies, to a CSV file. This file may be opened in Excel for thematic hand-coding of key words, histogram generation and so forth.

```

#####
#####
#
# VM-UBUNTU-R
#
#####
#####

write.csv(wf, file="WF.csv")

```



Appendix 4 ETHICAL REVIEW CORRESPONDENCE

A4.1 Introduction

Ethical review was sought on 18/05/2015, with a bundle of material (available upon request) sent to Dr Malcolm Bray, Department of Geography Ethics Co-ordinator at the University of Portsmouth,

A4.2 Initial response

Dr Bray responded to the Ethical review request with a favourable opinion (subject to conditions) on 29/05/2015 (Figure A4-1).

 	Department of Geography University of Portsmouth Buckingham Building Lion Terrace PORTSMOUTH PO1 3HE
Adrian Tear Department of Geography Date: 29 th May 2015	
<u>FAVOURABLE OPINION WITH CONDITIONS</u>	
Protocol Title: Social media, sentiment and location: the role of geography in politically charged online debate.	
Date Reviewed: 29 th May 2015	
Dear Adrian,	
Thank you for submitting your application for ethical review. The proposal was reviewed by Dr Malcolm Bray as a Departmental Review. M. Bray made the decision not to invoke full review by the Science Faculty Ethics Committee because:	
A. All of your data are obtained from public domain secondary sources and you have undertaken to abide by regulations of the data provider in the use of that data;	

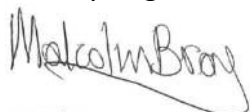
- B. There are no interactions between the researcher and the data subjects (persons sending messages publically on online social networks);
- C. Data analysis is aggregated as the samples involve several million messages, so that individual message contents would not be reported;
- D. The content filtering design does not aim to collect potentially sensitive information.

You have provided a full account of your proposals and a useful review of the emerging ethical considerations associated with research based on social media data sources. Nonetheless, there are some uncertainties as to whether all subjects engaged with social media fully understand the “public” nature of their engagement. Furthermore, there is a strong potential for disclosures within the messages of illegal activities or other items that could be harmful to the subjects and/or others. Also, it is theoretically possible to identify the originators of messages even though that is not the purpose of your research. Because of these uncertainties I list the following conditions to which you should comply:

- 1) It is important that the data are stored securely. Your application document provides appropriate details and I understand also that by using the Oracle database storage is especially secure;
- 2) Do not link message content to identifiable individuals (i.e. originators of messages) when writing up your Thesis;
- 3) Wherever possible do not filter message contents using terms that are likely to identify potentially sensitive information. If your research necessarily reports on potentially sensitive message contents it is important that it is impossible to trace specific content back to specific individuals i.e. wherever possible discuss aggregated results;
- 4) If you inadvertently encounter potentially sensitive information you should keep it confidential and secure unless you feel that there is an over-riding reason to report it. In cases where you are uncertain please consult your supervisor. I accept that your data is now “historic” which can reduce the potential sensitivity of some types of information;
- 5) If your data is ever to be used for a different research project then a new ethics review would be required.

I hope that you are able to incorporate these conditions into your research design. Please can you confirm back to me that the conditions are OK, similarly do let me know if any are problematic.

I wish you good luck with the study.



Dr Malcolm Bray

Geography Dept. Ethics Co-ordinator

Figure A4-1 – Dr Malcom Bray’s Ethical review response: ‘Favourable opinion with conditions’

A4.3 Supplementary question

Following the Annual Review meeting at the end of 2015 a supplementary question was sent by email to Dr Bray (Figure A4-2).

Hi Malcolm

At my annual review the other month Dr.Potts raised the question as to whether I should be able to report the content of individual Tweets etc.

Earlier when I submitted my bundle to you I had indicated that I was most interested in the aggregate data and counts etc. However, it may be useful to output content from 'established figures' in my corpus (e.g. Obama or Salmond) or indeed that of the 'most retweeted' users etc. (who may not be candidates but might be newspaper reporters or simply private individuals). All of the data are, effectively, public domain but - as I mentioned - some users might well be unaware just how public domain it all is!

I understand from Kim, with her Twitter/Smoking in Wales project, that David Carpenter thinks reporting the content of 'established figures' is basically OK but I thought I'd better run this question past you again.

The PhD will not be an exercise in repeating Tweets - I am much more interested in the aggregate behaviour of geo-tagging/non-geo-tagging users - but it probably would be useful to be able to reference particular messages, which is what I think Jonathan was thinking.

Following your earlier review/guidance do you have any suggestions?

Regards, Adrian Tear

Figure A4-2 – Email of 22/02/2016 (Adrian Tear - Ethical Review - Further Question)

A4.4 Supplementary response

Dr Bray replied to the supplementary question posed above on 25/02/2016 (Figure A4-3).

Adrian,

In my opinion it is reasonable to expect that "established figures" - who know that they are "in the public limelight" and who might expect media attention - should be aware that their publicly available Tweets really are public! Thus, they should be considering their comments carefully as they make them and they should accept full responsibility for comments made. That said I would still recommend not attributing comments that are obviously highly sensitive and/or which are clearly detrimental to named or identifiable third parties (Provision 1).

A possible uncertainty is how to judge whether a social media participant is an "established figure?" I'm guessing that you would only wish to quote from readily identifiable "established figures" rather than obscure or marginal ones? To be sure I feel that you need outline roughly how many "established figures" do you intend to quote in this manner? and (b) what is your process for identifying "established figures" giving examples (Provision 2). I would have a problem for example in establishing the status of a prolific tweeter who otherwise is not in the public eye, maybe that's not an issue for you?

I feel that your desire to name/identify a few select Tweeters as you request in your message is a minor, very justifiable and safe (with my provisions) variation to your original ethics application.

Please can you confirm that I have interpreted your intentions correctly and that you are happy to abide by Provisions 1 and 2. Please reply to the specific elements of Provision 2.

Please retain our exchanges of messages as evidence of continuing ethical consideration within your research.

Figure A4-3 – Email of 25/02/2016 in response to the supplementary question

The final exchange in this supplementary matter is reproduced below (Figure A4-4).

Thank you Malcolm. That sounds sensibly balanced to me. In the next week or three I will have more of an idea about whose individual messages I might like to report on and then, by looking at metrics such as numbers of followers or friends etc., it should be self evident in the data who is a 'figure' of any repute. Regards
Adrian

Figure A4-4 – Email of 25/02/2016 confirming acceptance of the opinion regarding the supplementary question

A4.5 Form UPR16

FORM UPR16**Research Ethics Review Checklist**

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information)



Postgraduate Research Student (PGRS) Information		Student ID:	UP634737
PGRS Name:	Adrian P.C.Tear		
Department:	Geography	First Supervisor:	Professor Richard G.Healey
Start Date: (or progression date for Prof Doc students)	2011		
Study Mode and Route:	Part-time <input checked="" type="checkbox"/>	MPhil <input type="checkbox"/>	MD <input type="checkbox"/>
	Full-time <input type="checkbox"/>	PhD <input checked="" type="checkbox"/>	Professional Doctorate <input type="checkbox"/>

Title of Thesis:	Geotagging matters? The role of space and place in politicised social media discourse
Thesis Word Count: (excluding ancillary data)	In the Declaration section of this thesis

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:

(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: <http://www.ukrio.org/what-we-do/code-of-practice-for-research/>)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES <input checked="" type="checkbox"/> NO <input type="checkbox"/>
b) Have all contributions to knowledge been acknowledged?	YES <input checked="" type="checkbox"/> NO <input type="checkbox"/>
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES <input checked="" type="checkbox"/> NO <input type="checkbox"/>
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES <input checked="" type="checkbox"/> NO <input type="checkbox"/>
e) Does your research comply with all legal, ethical, and contractual requirements?	YES <input checked="" type="checkbox"/> NO <input type="checkbox"/>

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	Letter from Dr Malcolm Bray of 29 May 2015
---	--

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed (PGRS):	<i>Adrian P.C.Tear</i>	Date: 30 September 2018
-----------------------	------------------------	--------------------------------

Appendix 5 DATA SIFT TWITTER LICENCE

A5.1 Introduction

The following licence terms were accepted 13/06/2012 upon registration at DataSift.com.

A5.2 Licence

Twitter

Your use of Datasift is subject to a licence from us, Mediasift Limited, and your use shall be limited to the scope agreed between us, based on the description provided by you (below) of the service or product you intend to provide, including any changes required by us. We are obliged by our own contractual commitments to ensure that Datasift is only used for Approved Purposes as described below.

Approved Purposes

Products and services created and marketed by you may only be used to display Content (a "**Display Product**") where they are provided to the End User through a Commercial Service. For example, it is not permissible for you to display Content on a publicly available website, including your own website.

You may create and market "**Data Products**" for analysis or statistical purposes, such as search engine ranking algorithms, sentiment analysis engines, ad targeting algorithms and malware analysis products provided such products do not involve the sale or resale of Content or the public display or curation of Content via a service whose primary purpose is display of Content for end user consumption.

You may not create any services or products, which facilitate the delivery of sponsored tweets or other advertising.

You must respect the privacy and sharing settings of Twitter Content. Do not share, or encourage or facilitate the sharing of protected Twitter Content. Promptly change your treatment of Twitter Content (for example, deletions, modifications, and sharing options) as changes are reported through the API.

We will pass on to you any Delete Messages that we receive from Twitter. When you receive the Delete Message, you must remove any Deleted Tweet that is the subject of the Delete Message and discard the Delete Message itself. You may not store or create any service displaying or publicizing any Deleted Tweets. Please see instructions for how to handle delete messages

<http://dev.datasift.com/docs/twitter-deletes>.

We also pass on to you any User Status Messages we receive from Twitter. When you receive a User Status Message, you must act according to the guidelines set out in our [Twitter User Status Messages](#) documentation to respect the privacy and sharing settings of Twitter's users.

In addition to the restrictions contained in this license from us, you are also required to follow the Twitter Rules.

In these terms:

"Commercial Service" means a paid for service provided on commercial terms by you to third parties, via paid subscription or behind a paywall. Services charged at a nominal or token rate are expressly excluded from this definition and Mediasift shall in its absolute discretion determine if a service is provided on commercial terms.

"Content" means all data provided to you via the Datasift API including the body of tweets, profile information and other metadata contained in the Datasift API.

"Delete Message" means a notification from Twitter of a Deleted Tweet.

"Deleted Tweet" means any Tweet deleted by a user.

"End User" means an individual user of any service or product created by you.

"Twitter Rules" means the Developer Rules of the Road, Display Guidelines and the Connect with Twitter Guidelines available on Twitter's Developer Site located at <http://dev.twitter.com>.

"User Status Message" means a notification from Twitter of a change to a user's account status.

"You" means you, the customer subscribing to the Datasift service subject to these terms and conditions, including your employees, directors, agents or any other party accessing the Datasift service on your behalf.

[If you are unsure whether your use of Datasift is for an Approved Purpose, please [contact us](#).]

For our full licence terms, see our [Terms and Conditions](#).

Appendix 6 2012 FRENCH PRESIDENTIAL ELECTION

A6.1 Technical proof of concept

On 6 May 2012, a technical proof of concept exercise was undertaken to record OSN interactions made during the final stages of the 2012 French Presidential Election using the CSDL statement below:

```
interaction.content CONTAINS_ANY "french election,
presidential election, sarkozy, hollande"
```

The CSDL was designed to record OSN interactions with message text containing any of the case-insensitive phrases (e.g., 'french election') shown within double quotes above. The recording started on Sunday, 6 May 2012 at 16:17:47 and was stopped at 17:31:03 on the same day, some 1 hour, 13 minutes and 16 seconds later.

A6.2 Outputs and analyses

Figure A6-1 shows the timeline of OSN interactions sampled.

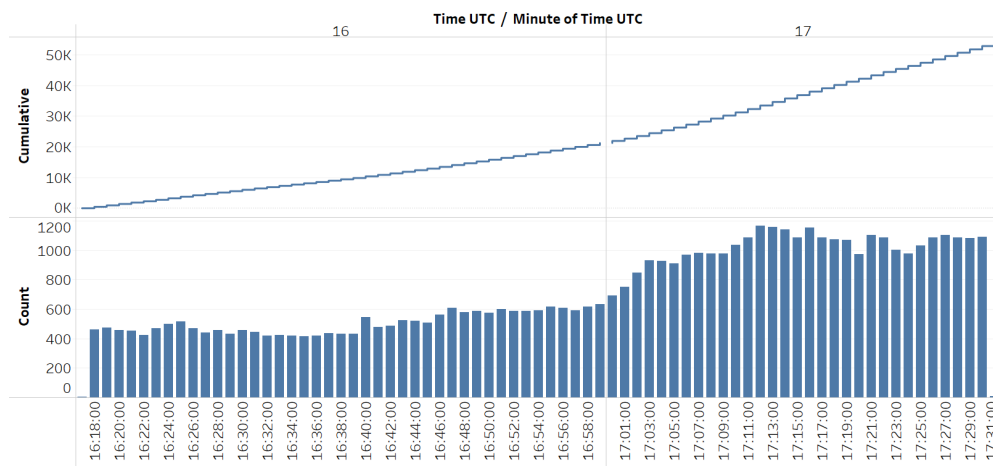


Figure A6-1 – Minute by minute Count and Cumulative total of OSN interactions recorded 4pm and 5pm on 6 May 2012 during the French Presidential Election second-round runoff

The peak flow of 1,166 interactions/minute was observed at 17:12 on 6 May 2012. In total, 52,914 OSN interactions were recorded in under an hour and a quarter. The resultant data set, downloadable in both CSV and JSON file formats, contained 163 fields ('columns', 'variables' or 'key/value pairs') for each observation. Most records came from Twitter (n=52,853), a few others came from Facebook (n=21) and Digg (n=40).

Analysis of the 76MB CSV file followed, using standard desktop applications including Microsoft's Excel spreadsheet and Access RDBMS software (Microsoft, 2018). One of the first results showed that OSN data are somewhat atypical of many other types of data, e.g., Census counts, often used in social science research. Most of the fields within the data set exhibited a high degree of row level sparsity. Only 6 fields were fully row-populated:

- `interaction.author.avatar`
- `interaction.author.name`
- `interaction.content`
- `interaction.created_at`
- `interaction.id`
- `interaction.type`

A further 8 fields were highly populated but most fields in the data set (72 of 163, Figure A6-2, p429) contained high percentages of `NULL` (no data) records in rows. Of the 6 fully-populated fields, `interaction.content` contained the message text and `interaction.created_at` contained a long date/time stamp with time zone offset (e.g., 'Sun, 06 May 2012 16:17:47 +0000'). This field has been used to calculate the number of interactions recorded/minute shown in Figure A6-1.

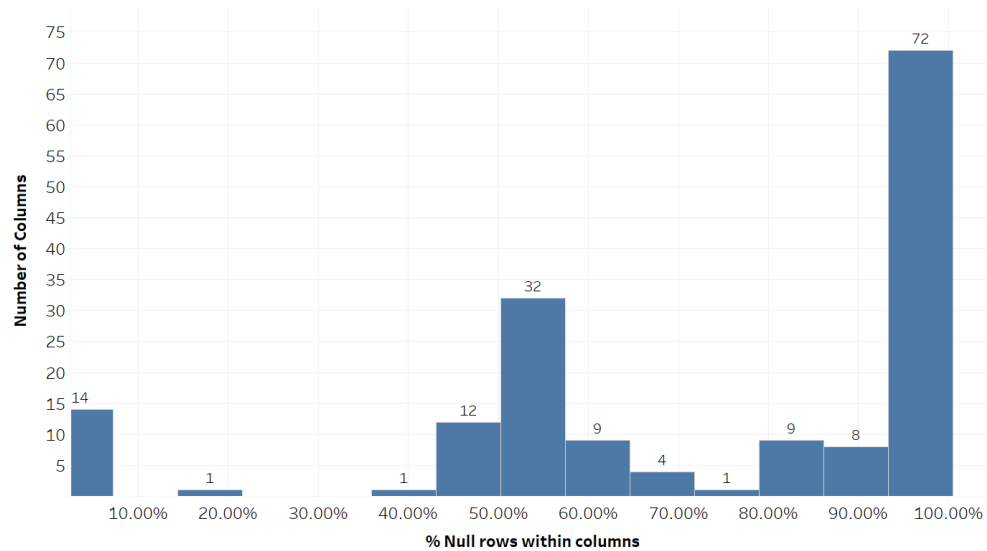


Figure A6-2 – % Null rows within columns in the French Presidential Election technical proof of concept data set; few columns/fields have fully populated rows

While many fields in the data set were not fully row-populated, others, e.g., `demographic.gender`, present in 60.89% of records, contained somewhat perplexing values (Figure A6-3).

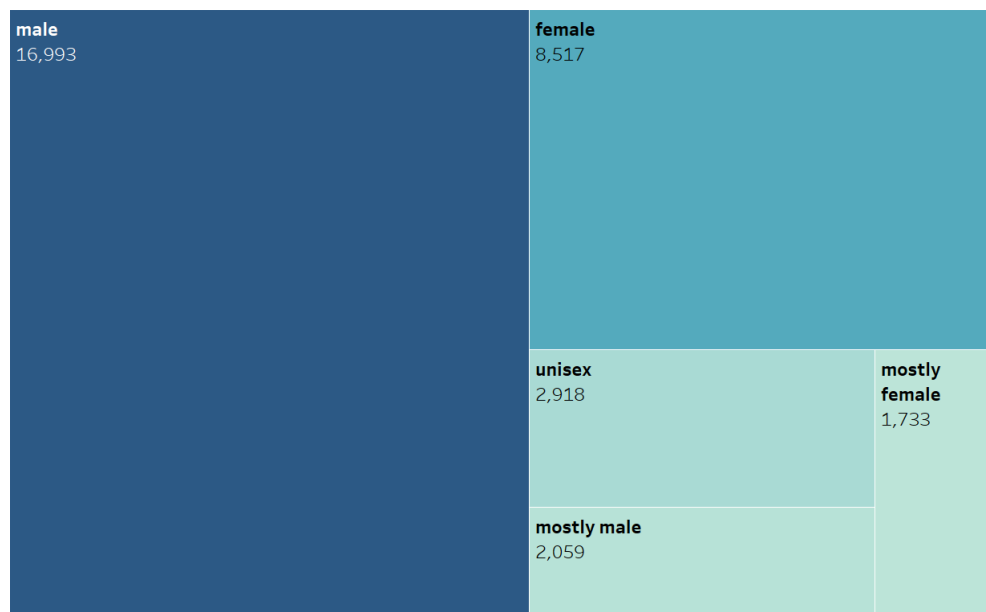


Figure A6-3 – Gender in the French Presidential Election technical proof of concept data set

Gender, it transpired, had been assigned to records based on DataSift’s internal modelling of language usage and was, hence, an approximation featuring more genders than is normal (‘mostly male’, ‘unisex’ etc.), rather than a reality based on actual and, for privacy reasons, unavailable user registration data.

Just 736 records (1.39% total) with non-null `interaction.geo.latitude` could easily be mapped (Figure A6-4), showing a European and predictably French bias in the geographical distribution of OSN interactions. This low percentage of coordinate-geotagged interactions proved to be in line with similarly low percentages subsequently reported by Leetaru et al. (2013).

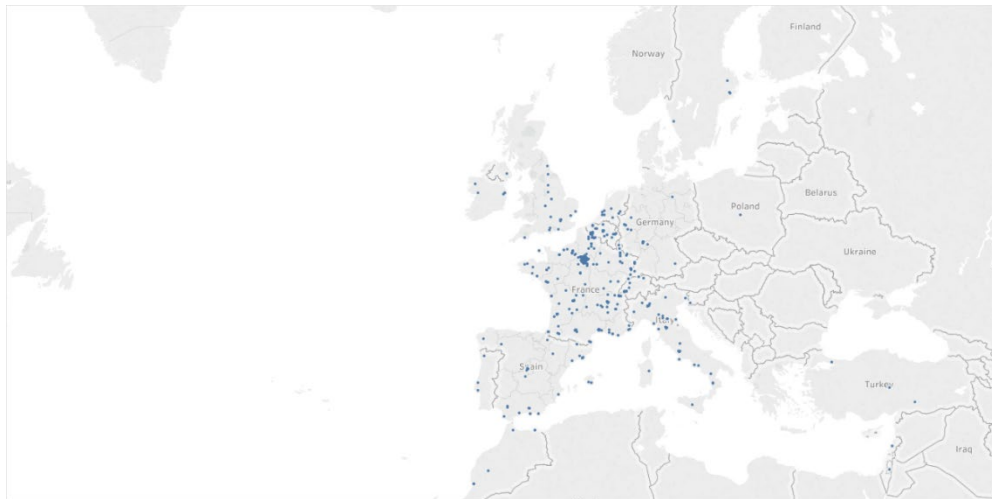


Figure A6-4 – European distribution of explicitly geotagged interactions (n=736, 1.39% of the total) in the French Presidential Election technical proof of concept data set

Further DataSift ‘augmentations’, e.g., the `salience.content.sentiment` field, a pre-computed sentiment score, allowed other types of analyses, e.g., graphing the positive/negative opinions surrounding a given candidate (Figure A6-5, p431). The score is one of several augmentations to social media data provided by Datasift (2018) as part of their service offering and has been calculated using ‘black box’ NLP software provided by Lexalytics (2018).

It appeared that messages mentioning Hollande exhibited somewhat more favourable sentiment at the upper (positive) ends of the salience scale, although

many messages sent by highly-followed Twitter users (>10,000 followers) were only just positive or negative in tone (+4 to -4 in Figure A6-5).

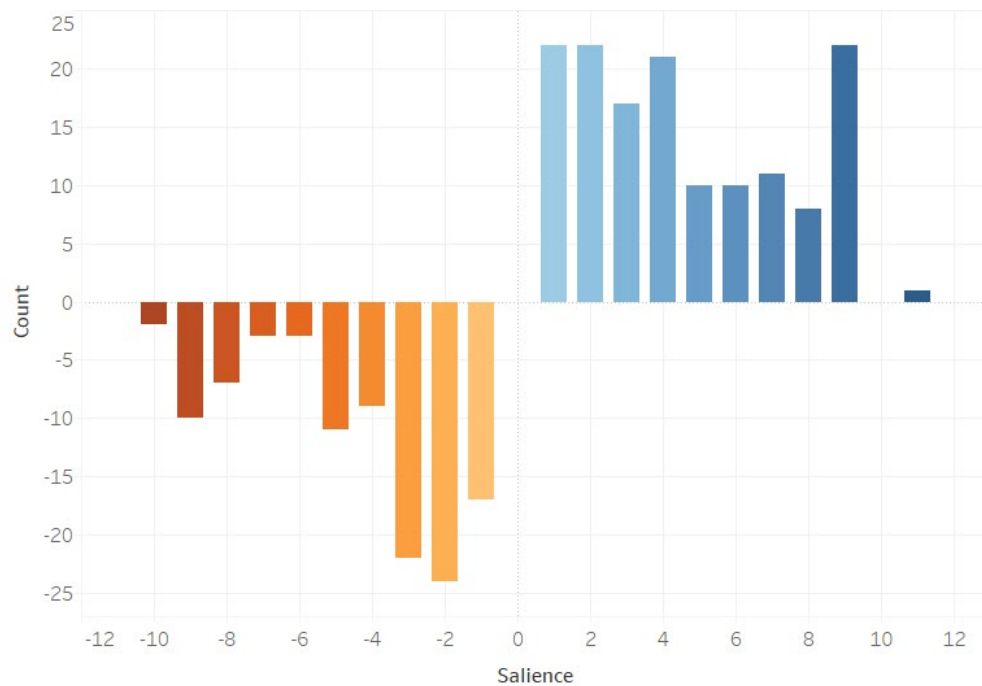


Figure A6-5 – Salience score and count of interactions mentioning ‘Hollande’ sent by Twitter users (with over 10,000 followers) in the French Presidential Election technical proof of concept data set

Outputs from the technical proof of concept demonstrated that OSN interactions could be filtered, recorded, saved and downloaded using the DataSift platform at speed, in large volume and with controllable costs. Analyses of downloaded data further suggested that larger numbers of OSN interactions could be used to answer the research questions (Section 1.7, p34) set out in this thesis. The CSDL definitions used to collect larger data volumes, for the 2012 US Presidential Election and the 2014 Scottish Independence Referendum, are given in Appendix 7 (p432).

Appendix 7 DATA SIFT STREAM DEFINITIONS

A7.1 Introduction

DataSift allows for the extraction of filtered Online Social Network (OSN) data using a proprietary Curated Stream Definition Language, CSDL (DataSift, 2013a). The CSDL used to record each Stream is reproduced below.

The United States Stream definitions are designed to:

- Tag Tweets or Posts ‘Democratic Party’ or ‘Republican Party’;
- Tag Tweets or Posts ‘Barack Obama’ or ‘Mitt Romney’;
- Tag Tweets or Posts ‘Positive’, ‘Neutral’ or ‘Negative’;
- Filtered on a set of keywords listed in `interaction.content`, and
- On language, and;
- On the existence of salience (sentiment), and;
- On the existence of geography (if explicitly required);
- Sampled from the universe in some proportion.

All three United States Stream definitions are identical save for differences in language, sample size and/or the explicit requirement for geographically referenced data. An ‘internally generated floating-point random number between 0 and 100’ (DataSift, 2018a) controls sample size.

The Scottish Stream definition differs slightly in that only one Stream was used (for a much longer time period), and no sampling or geographic filtering was used.

A7.2 2012 United States Presidential Election

The three Stream definitions are given below each with a summary table detailing start and end dates, language, geographical scope and so forth.

A7.2.1 US2012_GEO

Table A7-1 – Summary of the US2012_GEO Stream

Start	04/09/2012	End	06/11/2012	Duration	64 days
Language	EN	Scope	Worldwide	Sample	1:5
Geo filter	Yes	Records	146,424	CSV/JSON	0.17/0.26GB

```

tag "Democratic Party" { interaction.content any
  "democrat,democrats,democratic,barack obama,obama,joe biden,biden" }
tag "Republican Party" { interaction.content any
  "republican,republicans,mitt romney,romney,paul ryan,ryan" }
tag "Positive" { salience.content.sentiment > 0 }
tag "Neutral" { salience.content.sentiment == 0 }
tag "Negative" { salience.content.sentiment < 0 }
tag "Barack Obama" { interaction.content any "barack
obama,barack,obama" }
tag "Mitt Romney" { interaction.content contains "mitt
romney,mitt,romney" }

return {
  interaction.content any "2012 election,presidential election,US
  president,president of the united states,republican,republicans,mitt
  romney,mitt,romney,democrat,democrats,democratic,barack
  obama,barack,obama"
  AND language.tag == "en"
  AND salience.content.sentiment exists
  AND interaction.geo exists
  AND interaction.sample < 20
}

```

A7.2.2 US2012_NON_GEO

Table A7-2 – Summary of the US2012_NON_GEO Stream

Start	06/09/2012	End	06/11/2012	Duration	62 days
Language	EN	Scope	Worldwide	Sample	1:50
Geo filter	No	Records	1,560,967	CSV/JSON	2.38/2.87GB

```

tag "Democratic Party" { interaction.content any
"democrat,democrats,democratic,barack obama,obama,joe biden,biden" }
tag "Republican Party" { interaction.content any
"republican,repUBLICans,mitt romney,romney,paul ryan,ryan" }
tag "Positive" { salience.content.sentiment > 0 }
tag "Neutral" { salience.content.sentiment == 0 }
tag "Negative" { salience.content.sentiment < 0 }
tag "Barack Obama" { interaction.content any "barack
obama,barack,obama" }
tag "Mitt Romney" { interaction.content contains "mitt
romney,mitt,romney" }

return {
interaction.content any "2012 election,presidential election,US
president,president of the united states,repUBLICan,repUBLICans,mitt
romney,mitt,romney,democrat,democrats,democratic,barack
obama,barack,obama"
AND language.tag == "en"
AND salience.content.sentiment exists
AND interaction.sample < 2
}

```

A7.2.3 US2012_NON_GEO_HISPANIC

Table A7-3 – Summary of the US2012_NON_GEO_HISPANIC Stream

Start	05/10/2012	End	06/11/2012	Duration	33 days
Language	ES	Scope	Worldwide	Sample	1:50
Geo filter	No	Records	11,276	CSV/JSON	0.01/0.02GB

```

tag "Democratic Party" { interaction.content any
"democrat,democrats,democratic,barack obama,obama,joe biden,biden" }
tag "Republican Party" { interaction.content any
"republican,repUBLICans,mitt romney,romney,paul ryan,ryan" }
tag "Positive" { salience.content.sentiment > 0 }
tag "Neutral" { salience.content.sentiment == 0 }
tag "Negative" { salience.content.sentiment < 0 }
tag "Barack Obama" { interaction.content any "barack

```

```

obama,barack,obama" }
tag "Mitt Romney" { interaction.content contains "mitt
romney,mitt,romney" }

return {
interaction.content any "2012 election,presidential election,US
president,president of the united states,repulican,repulicans,mitt
romney,mitt,romney,democrat,democrats,democratic,barack
obama,barack,obama"
AND language.tag == "es"
AND salience.content.sentiment exists
AND interaction.sample < 2
}

```

A7.3 2014 Scottish Independence Referendum

The Stream definition is given below together with a summary table.

A7.3.1 SCOT2014

Table A7-4 – Summary of the SCOT2014 Stream

Start	18/09/2013	End	30/09/2014	Duration	378 days
Language	EN	Scope	Worldwide	Sample	1:1
Geo filter	No	Records	6,477,713	CSV/JSON	21.1/19.9GB

```

interaction.content contains_any "Scottish independence,Scottish
referendum,independence referendum,independance
referundum,independence vote,independance vote,Scotland
independence,Scotland referendum,Scottish independent,Scotland
independent,Scottish vote,Scotland vote,Scottish union,Scotland
union,Scottish secession,Scotland secession,Scottish
separation,Scotland separation,Scottish nationalist,Scotland
nationalist,SNP,Salmond,Alec Salmond,Alistair Darling,Better
Together,Leave the UK,Yes Scotland"

```

Appendix 8 COMPUTING ENVIRONMENT

A8.1 Background

Recent advances in desktop-based computers and computer virtualisation have provided significant platform flexibility in the conduct of this research. Over thirty years ago Barr (1985) argued that ‘in a rapidly changing environment, with hardware regularly offering more power for less money, portable machine-independent software needs to be developed for mapping applications. This needs to take full advantage of the powerful interactive capabilities of microcomputers to provide both skilled and naive users with opportunities for interacting with maps, both at the design and the end-user stages, and in new forms.’ The current generation of desktop mapping and visualisation tools such as QGIS and Tableau, used here, have gone quite some way towards satisfying these requirements.

Separately, and over forty years ago, R. Goldberg (1974, p34) noted that ‘Virtual machines have finally arrived. Dismissed for years as academic curiosities, they are now seen as cost-effective techniques for organizing computer system resources to provide extraordinary system flexibility and support for certain unique applications.’ Only comparatively recently, however, has virtualisation technology moved out of back-office, high-end, managed IT infrastructure onto the desktop; first with the release in 2007 of VirtualBox (Oracle, 2014b), closely followed in 2008 by the release of Microsoft Hyper-V (Microsoft, 2014a).

A8.2 Physical and Virtual computing environment

The combination of improved hardware virtualisation support with modern multi-core 64-bit processors (Uhlig et al., 2005) and larger and cheaper amounts of Random Access Memory (RAM) has enabled individual researchers, as opposed to centralised University IT or Computer Science departments, to provision multiple operating systems or software stacks tailored to specific application requirements.

Table A8-1 – Matrix of Host/Guest Operating System and Software Installations

Host OS	Windows Server 2012	Windows Server 2012R2	Windows Server 2012R2	Windows 10 Professional
Hardware	IBM System x3850 M2	Dell Power Edge 2950	Dell Vostro 410	Dell Latitude E7440
Class	Server	Server	Desktop	Laptop
Hypervisor	Hyper-V	VirtualBox	Hyper-V	VirtualBox
Host-based Software Installation(s)				
Software		Oracle 12.1.0.2.0		Oracle 12.1.0.2.0 Gephi 0.9.2 Tableau 10.5
Guest-based Software Installation(s)				
CentOS Desktop 6		SAS University Edition		SAS University Edition
CentOS Server 6.7				MapR Sandbox for Apache Drill
CentOS Desktop 7			CLAVIN	Ruby Nokogiri AlchemyAPI CLAVIN
Oracle Linux Server 5.11				Oracle Endeca Information Discovery
Scientific Linux 6.3				Edinburgh GeoParser (03/2016)
Ubuntu Desktop 13.10			PostgreSQL 9.3	
Ubuntu Desktop 14.04LTS			CLAVIN GATE 8.0 R/RStudio	
Ubuntu Server 14.04.1 LTS	MapR Hadoop 4.01 M3 (5 nodes)		MapR Hadoop 4.01 M3 (3 nodes)	MapR Hadoop 4.01 M5 (3 nodes)
Ubuntu Desktop 16.04 LTS				R/RStudio
Windows Server 2012			MarkLogic 7 Oracle 12c 12.1.0.2.0	

Table A8-1 (p437) shows the matrix of Host and Guest Operating Systems (OS), hardware, hypervisors and software installed during this research programme. Many Virtual Machines have been ‘spun-up’, ‘cloned’ and, in some cases, ‘torn-down’. Each Virtual Machine (VM) has been kept as ‘clean’ as possible, with minimal software installed, to fulfil a specific task. While this collection of physical and virtual infrastructure may seem somewhat excessive, it results from:

- a) hardware and/or Internet (non-)availability;
- b) the difficulty of installing two hypervisors simultaneously on one host when any one hypervisor requires dedicated and exclusive access to the Intel VTx chipset;
- c) the requirement to travel with data and software available on a laptop, and;
- d) the iterative and exploratory nature of the research project itself (Section 3.3.2, p107).

VM, OS, and software installations have been made on commodity desktop (Intel i7 8-core processor, 16GB memory) and laptop hardware (Intel i7/i5 4-core/2-core processors, 16GB memory), upgraded with 256GB or 512GB SSDs. Installations of the Oracle 12c RDBMS and the MapR Hadoop distribution have also been performed on physical server-class hardware:

- Oracle 12.1.0.2.0 (with support for JSON) running on a 2U Dell PowerEdge 2950 with 2*4-core Xeon processors, 32GB memory, 2*73GB and 3*146GB 15K SAS drives in a Redundant Array of Inexpensive Disks (RAID0).
- MapR 4.01 Hadoop distribution running on a 4U IBM System x3850 M2 with 4*6-core Xeon processors, 128GB memory, 4*146GB 10K SAS drives (later upgraded to 4*200GB SAS Enterprise SSDs) in RAID0.

While the server hardware is somewhat dated and, in the case of the Dell PowerEdge 2950 extremely noisy, it has been purchased cheaply (<£500 per server from eBay) to provide specific processing capabilities. For example, the larger

amounts of server RAM were designed to allow either better throughput in Oracle 12c (which proved, in fact, to be Input/Output bound) or, in the case of the IBM System x3850 M2 machine with its 24 processor cores and 128GB of memory, the creation of a personal Private Cloud (Figure A8-1) multi-node Hadoop cluster consisting of five virtual Ubuntu Servers each configured with 2/4 processors and 16GB/32GB of memory (more in the cluster controller).



Figure A8-1 – ‘Shed-hosted’ personal Private Cloud (1*2U Dell PowerEdge 2950, 2*4U IBM System x3850 M2)

As this proved unreliable, and with invaluable assistance from University of Portsmouth staff (G. Burton, 2017; Marshall & Tear, personal communication, 2016), a parallel initiative used supercomputer resources in the University’s Data Centre:

- Five High Performance Compute (HPC) nodes each with 12 core CPUs, 24GB of RAM and 2TB of disk space running Scientific Linux 6 (Figure A8-2, p440) were configured to form a MapR 5.0.0.32987 Hadoop cluster running Drill

and Hive amongst other ‘ecosystem’ tools (Cutting, 2013). The cluster offered 60 cores, 120GB RAM and 7.5TB storage; available space is lower than total disk space due to replication in the distributed file system.

- Two further HPC nodes (specified as above) were configured, one running Oracle 11g, the other Oracle 12c. A large existing Data Warehouse application (Healey, 2011) was deployed on the 2-node Oracle 11g RDBMS instance with significant speed improvements. Queries against the OSN research database `OSNDATA` likewise ran appreciably faster on the 2-node supercomputer Oracle 12c instance.

To handle large amounts of data and evaluate different software systems, frequently from locations in Sussex and Scotland with poor Internet access, it has become necessary to consolidate, and/or acquire, skills in virtualisation, operating system and software application installation.



Figure A8-2 – Maintenance activities on the University of Portsmouth’s SCIAMA supercomputer

The need for portability, identified by Barr (1985), has won out over superior computing resources sometimes inaccessible over a poor Internet connection and is discussed in the following section.

A8.3 System architecture

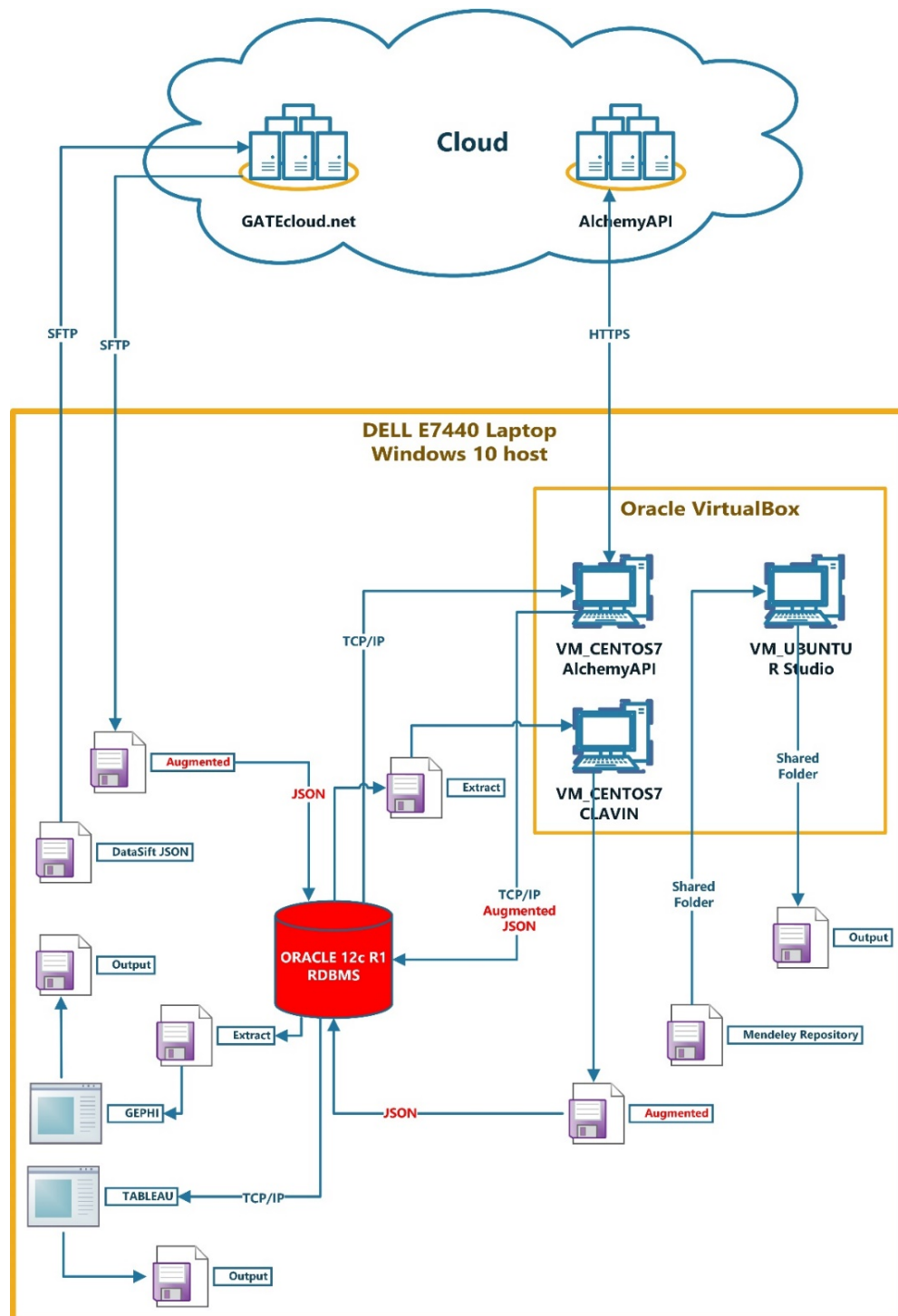


Figure A8-3 – Schematic representation of containers (i.e., physical host, virtual machines, Cloud), data flows, data transfer mechanisms and software applications employed in this research

Figure A8-3 (p441) shows a schematic representation of the the containers (physical hosts and virtual machines and Cloud-hosted services) and systems architecture arrived at during this research. Along with Table A8-1 (p437) this indicates that the Dell Latitude E7440 laptop has been the most heavily used computer throughout, running:

- Oracle 12c (version 12.1.0.2.0);
- Oracle SQL Developer (version 4.1.2);
- Tableau (up to version 10.5);
- Gephi (version 0.9.2), and;
- Oracle VirtualBox (up to version 5.2.6).

VirtualBox has been used on the laptop asynchronously to run three VMs, dedicated to R statistical analysis (Ubuntu 16.04 LTS), AlchemyAPI processing (CentOS 7) and CLAVIN-rest geoparsing (CentOS 7). All VMs access files (over a VirtualBox Shared Folder) or data (TCP/IP over the network to Oracle 12c) stored on the host. While it is possible to locally test, and then remotely scale-up, environments of this type either to Private ('on premise') or Public ('off premise') Cloud computing environments, e.g., Amazon Web Services (AWS) or Windows Azure, these can quickly become expensive, particularly if multiple well-specified VM instances running experimental applications are created or, worse, accidentally left running. The computing environment(s) that have been tested, developed, used and sometimes destroyed during this research owe much to free academic software licences provided by Microsoft (2014b) and Tableau (2017a). Oracle's developer licensing has proved invaluable when evaluating and using the JSON capabilities built into the Oracle 12c RDBMS (Oracle, 2014c). Applications from open sources have been widely used (Berico-Technologies, 2014; CentOS, 2014; GATE, 2014; Gephi, 2018a; MongoDB, 2014; PostgreSQL, 2014; The R Foundation, 2018; Ubuntu, 2014) together with several commercial systems used under 'developer licensing' terms (MapR, 2014; MarkLogic, 2014; SAS, 2014).

Appendix 9 SPARSITY IN THE INTERACTIONS TABLE

A9.1 Introduction

The `INTERACTIONS` table stores 8,196,380 rows (records) with 149 fields, 3 created as part of the ETL process (`STREAM`, `STREAMID`, `UUID`) and 146 fields common to all DataSift CSV files imported from source data. The table consists of a matrix of 1,196,671,480 data points in all.

A9.2 `INTERACTIONS` table definition and metadata

The following query may be used against Oracle 12c's data dictionary to return table metadata:

```
SELECT *
FROM USER_TAB_COLUMNS
WHERE TABLE_NAME = 'INTERACTIONS'
ORDER BY COLUMN_ID
```

The query returns field names, data type definitions and a number of other metadata items including number of null values. Salient metadata fields for the `INTERACTIONS` table are reproduced below (Table A9-1). Type size has been reduced to 8pt to fit the listing across the page.

Table A9-1 – Columns in the `INTERACTIONS` table

Column name	Data type	N Distinct	N Null	Avg Col Len
<code>UUID</code>	RAW (16)	8196380	0	17
<code>STREAMID</code>	NUMBER (22)	6489088	0	6
<code>STREAM</code>	VARCHAR2 (20)	4	0	11
<code>DEMOGRAPHIC_GENDER</code>	VARCHAR2 (20)	5	3925328	5
<code>FB_APPLICATION</code>	VARCHAR2 (100)	1886	7531547	3
<code>FB_AUTHOR_AVATAR</code>	VARCHAR2 (60)	319008	7353878	6
<code>FB_AUTHOR_ID</code>	NUMBER (22 ,19 ,0)	316992	7353878	2
<code>FB_AUTHOR_LINK</code>	VARCHAR2 (200)	313344	7353878	7
<code>FB_AUTHOR_NAME</code>	VARCHAR2 (200)	283648	7353878	3

FB_CAPTION	CLOB (4000)	0	146424	113
FB_DESCRIPTION	CLOB (4000)	0	146424	120
FB_ID	VARCHAR2 (40)	849792	7353878	5
FB_LIKES_COUNT	NUMBER (22 ,19 ,0)	23	8194253	2
FB_LIKES_IDS	VARCHAR2 (4000)	7725	8186904	2
FB_LIKES_NAMES	VARCHAR2 (2000)	7734	8186904	2
FB_LINK	VARCHAR2 (4000)	160928	7544057	8
FB_MESSAGE	CLOB (4000)	0	146424	125
FB_NAME	VARCHAR2 (400)	86664	7592440	4
FB_OG_BY	VARCHAR2 (1000)	6110	7997632	2
FB_OG_LENGTH	VARCHAR2 (20)	507	8148764	2
FB_OG_PAGE	VARCHAR2 (40)	125	8196231	2
FB_SOURCE	VARCHAR2 (4000)	19756	7353878	5
FB_TO_IDS	VARCHAR2 (4000)	25864	8125827	2
FB_TO_NAMES	VARCHAR2 (4000)	25560	8125827	2
FB_TYPE	VARCHAR2 (10)	5	7353878	2
INTERACTION_AUTHOR_AVATAR	VARCHAR2 (800)	2821120	0	79
INTERACTION_AUTHOR_ID	NUMBER (22 ,19 ,0)	2425600	0	7
INTERACTION_AUTHOR_LINK	VARCHAR2 (200)	2628352	0	34
INTERACTION_AUTHOR_NAME	VARCHAR2 (200)	2126336	17025	14
INTERACTION_AUTHOR_USERNAME	VARCHAR2 (50)	2180096	842502	11
INTERACTION_CONTENT	CLOB (4000)	0	0	356
INTERACTION_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	5216256	0	13
INTERACTION_GEO_LATITUDE	FLOAT (22 ,126)	183120	7941778	2
INTERACTION_GEO_LONGITUDE	FLOAT (22 ,126)	183040	7941778	2
INTERACTION_ID	VARCHAR2 (40)	8196380	0	33
INTERACTION_LINK	VARCHAR2 (200)	8168960	57265	58
INTERACTION_SCHEMA_VERSION	NUMBER (22 ,19 ,0)	1	10	3
INTERACTION_SOURCE	VARCHAR2 (4000)	28718	0	18
INTERACTION_TAGS	VARCHAR2 (200)	24	6477713	10
INTERACTION_TITLE	VARCHAR2 (400)	86664	7592440	4
INTERACTION_TYPE	VARCHAR2 (20)	2	0	9
KLOUT_SCORE	NUMBER (22 ,19 ,0)	90	1076963	3
LANGUAGE_CONFIDENCE	NUMBER (22 ,19 ,0)	84	117911	3
LANGUAGE_TAG	VARCHAR2 (2)	138	117911	3
LINKS_CREATED_AT	VARCHAR2 (2000)	840576	4915230	16
LINKS_RT_COUNT	VARCHAR2 (200)	6374	6754734	2
LINKS_TITLE	CLOB (4000)	0	0	153
LINKS_URL	VARCHAR2 (4000)	640256	4915230	39
SALIENCE_CONTENT_SENTIMENT	NUMBER (22 ,19 ,0)	73	301486	3
SALIENCE_TITLE_SENTIMENT	NUMBER (22 ,19 ,0)	40	7658404	2
TRENDS_CONTENT	VARCHAR2 (200)	36948	5599508	7

TRENDS_SOURCE	VARCHAR2 (20)	1	5599508	5
TRENDS_TYPE	VARCHAR2 (4000)	35912	5599508	30
TW_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	2755328	4483533	7
TW_DOMAINS	VARCHAR2 (200)	35400	6692194	4
TW_GEO_LATITUDE	FLOAT (22 ,126)	182016	7943009	2
TW_GEO_LONGITUDE	FLOAT (22 ,126)	182000	7943009	2
TW_ID	NUMBER (22 ,19 ,0)	7339008	842502	10
TW_IN_RE_TO_SCREEN_NAME	VARCHAR2 (40)	219504	7380199	3
TW_IN_RE_TO_STATUS_ID	NUMBER (22 ,19 ,0)	552768	7551801	2
TW_IN_RE_TO_USER_ID	NUMBER (22 ,19 ,0)	215120	7380178	2
TW_LINKS	VARCHAR2 (4000)	803712	6692184	9
TW_MENTION_IDS	VARCHAR2 (200)	483328	6937228	4
TW_MENTIONS	VARCHAR2 (200)	477792	6937228	5
TW_PLACE_ATT_LOCALITY	VARCHAR2 (20)	5	8196374	2
TW_PLACE_ATT_REGION	VARCHAR2 (100)	5	8196374	2
TW_PLACE_ATT_ST_ADDRESS	VARCHAR2 (200)	209	8194932	2
TW_PLACE_COUNTRY	VARCHAR2 (100)	213	7947715	2
TW_PLACE_COUNTRY_CODE	VARCHAR2 (2)	163	7947717	2
TW_PLACE_FULL_NAME	VARCHAR2 (200)	20270	7947713	2
TW_PLACE_ID	VARCHAR2 (20)	19220	7947713	2
TW_PLACE_NAME	VARCHAR2 (100)	15224	7947713	2
TW_PLACE_PLACE_TYPE	VARCHAR2 (40)	5	7947713	2
TW_PLACE_URL	VARCHAR2 (100)	21874	7947713	3
TW_RT_COUNT	NUMBER (22 ,19 ,0)	18258	4555349	2
TW_RT_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	2631936	4555349	7
TW_RT_DOMAINS	VARCHAR2 (200)	14746	7038612	4
TW_RT_ID	NUMBER (22 ,19 ,0)	3723776	4555349	6
TW_RT_LINKS	VARCHAR2 (4000)	199904	7038594	9
TW_RT_MENTION_IDS	VARCHAR2 (400)	214816	5945525	5
TW_RT_MENTIONS	VARCHAR2 (400)	228384	5724787	7
TW_RT_SOURCE	VARCHAR2 (400)	2186	4555349	33
TW_RT_TEXT	VARCHAR2 (4000)	921920	4555349	52
TW_RT_USER_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	1185536	4555349	7
TW_RT_USER_DESCRIPTION	VARCHAR2 (800)	1102208	5149682	36
TW_RT_USER_FOLLOWERS_COUNT	NUMBER (22 ,19 ,0)	35816	4556823	3
TW_RT_USER_FRIENDS_COUNT	NUMBER (22 ,19 ,0)	18664	4555897	3
TW_RT_USER_GEO_ENABLED	VARCHAR2 (40)	2	5100815	3
TW_RT_USER_ID	NUMBER (22 ,19 ,0)	1162112	4555349	4
TW_RT_USER_ID_STR	VARCHAR2 (10)	1166336	4555349	6
TW_RT_USER_LANG	VARCHAR2 (20)	56	4555349	2
TW_RT_USER_LISTED_COUNT	NUMBER (22 ,19 ,0)	4102	4920218	2
TW_RT_USER_LOCATION	VARCHAR2 (200)	393664	5663542	6

TW_RT_USER_NAME	VARCHAR2 (200)	1058688	4561421	7
TW_RT_USER_SCREEN_NAME	VARCHAR2 (20)	1192192	4555349	6
TW_RT_USER_STATUSES_COUNT	NUMBER (22 ,19 ,0)	153520	4555376	3
TW_RT_USER_TIME_ZONE	VARCHAR2 (50)	228	5854493	5
TW_RT_USER_URL	VARCHAR2 (4000)	329312	7239231	5
TW_RT_USER_UTC_OFFSET	NUMBER (22 ,19 ,0)	35	5867254	2
TW_RT_USER_VERIFIED	VARCHAR2 (40)	2	5349580	3
TW_RTED_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	764992	4555349	7
TW_RTED_GEO_LATITUDE	FLOAT (22 ,126)	5335	8094037	2
TW_RTED_GEO_LONGITUDE	FLOAT (22 ,126)	5330	8094037	2
TW_RTED_ID	NUMBER (22 ,19 ,0)	944192	4555349	6
TW_RTED_PLACE_ATT_ST_ADDRESS	VARCHAR2 (100)	9	8194972	2
TW_RTED_PLACE_COUNTRY	VARCHAR2 (100)	137	8094023	2
TW_RTED_PLACE_COUNTRY_CODE	VARCHAR2 (2)	105	8094023	2
TW_RTED_PLACE_FULL_NAME	VARCHAR2 (200)	5525	8094023	2
TW_RTED_PLACE_ID	VARCHAR2 (40)	4978	8094023	2
TW_RTED_PLACE_NAME	VARCHAR2 (100)	4419	8094023	2
TW_RTED_PLACE_PLACE_TYPE	VARCHAR2 (20)	5	8094023	2
TW_RTED_PLACE_URL	VARCHAR2 (200)	5464	8094023	2
TW_RTED_SOURCE	VARCHAR2 (400)	2270	4555349	30
TW_RTED_USER_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	361248	4555349	7
TW_RTED_USER_DESCRIPTION	VARCHAR2 (800)	377888	4731661	45
TW_RTED_USER_FOLLOWERS_COUNT	NUMBER (22 ,19 ,0)	448000	4555498	3
TW_RTED_USER_FRIENDS_COUNT	NUMBER (22 ,19 ,0)	42096	4564818	3
TW_RTED_USER_GEO_ENABLED	VARCHAR2 (40)	2	5169141	3
TW_RTED_USER_ID	NUMBER (22 ,19 ,0)	365568	4555349	4
TW_RTED_USER_ID_STR	VARCHAR2 (20)	362880	4555349	5
TW_RTED_USER_LANG	VARCHAR2 (20)	40	4555349	2
TW_RTED_USER_LISTED_COUNT	NUMBER (22 ,19 ,0)	33464	4667975	3
TW_RTED_USER_LOCATION	VARCHAR2 (200)	156400	5359761	6
TW_RTED_USER_NAME	VARCHAR2 (200)	351872	4559850	7
TW_RTED_USER_SCREEN_NAME	VARCHAR2 (40)	368032	4555349	6
TW_RTED_USER_STATUSES_COUNT	NUMBER (22 ,19 ,0)	155648	4555351	3
TW_RTED_USER_TIME_ZONE	VARCHAR2 (100)	178	5271660	6
TW_RTED_USER_URL	VARCHAR2 (4000)	138608	5978922	10
TW_RTED_USER_UTC_OFFSET	NUMBER (22 ,19 ,0)	35	5285072	2
TW_RTED_USER_VERIFIED	VARCHAR2 (40)	2	5194291	3
TW_SOURCE	VARCHAR2 (400)	7681	4483533	30
TW_TEXT	VARCHAR2 (4000)	3396864	4483533	48
TW_USER_CREATED_AT	TIMESTAMP(9) WITH TIME ZONE (13 ,9)	1213312	4483533	7
TW_USER_DESCRIPTION	VARCHAR2 (4000)	1170560	5111785	36
TW_USER_FOLLOWERS_COUNT	NUMBER (22 ,19 ,0)	63452	4494095	3

TW_USER_FRIENDS_COUNT	NUMBER (22,19,0)	27036	4496739	3
TW_USER_GEO_ENABLED	VARCHAR2 (40)	2	4997402	3
TW_USER_ID	NUMBER (22,19,0)	1220608	4483533	4
TW_USER_ID_STR	VARCHAR2 (10)	1193984	4483533	6
TW_USER_LANG	VARCHAR2 (20)	53	4483533	2
TW_USER_LISTED_COUNT	NUMBER (22,19,0)	8068	4845836	2
TW_USER_LOCATION	VARCHAR2 (200)	406880	5671841	6
TW_USER_NAME	VARCHAR2 (200)	1118848	4494487	7
TW_USER_SCREEN_NAME	VARCHAR2 (40)	1240704	4483533	7
TW_USER_STATUSES_COUNT	NUMBER (22,19,0)	271232	4483991	3
TW_USER_TIME_ZONE	VARCHAR2 (40)	230	5696823	6
TW_USER_URL	VARCHAR2 (800)	420032	6780283	7
TW_USER_UTC_OFFSET	NUMBER (22,19,0)	35	5715812	2
TW_USER_VERIFIED	VARCHAR2 (40)	2	5346570	3

A9.3 PL/SQL programme to count and store Zero Length Strings or NULLs into table ZERO_LEN_SPARSITY_INT_COLS

The SQL in the previous section reveals many `NULL` values (no data) in cells. Counts of `NULL` values in rows across columns by Stream (Appendix 7, p432) have been recorded programmatically.

The PL/SQL script below was developed to count (zero length strings OR `NULL` values) in rows across columns for each of the four Streams held in the `INTERACTIONS` table.

```
-- COUNT NULLS from
http://stackoverflow.com/questions/11642079/query-each-column-of-a-table-in-a-loop-oracle-database
-- FIXED 01/04/2017 to cater for (zero length string OR null)

-- see the output!
set serveroutput on

declare
```

```

mytable varchar(32) := 'INTERACTIONS';

cursor s1 (mytable varchar2) is
    select column_name
    from user_tab_columns
    where table_name = mytable
    order by column_name;

mycolumn varchar2(100);
query_str varchar2(1000);
mycount number(19);

-- arrays for handling streams from
http://www.tutorialspoint.com/plsql/plsql\_arrays.htm
type streamsarray IS VARRAY(4) OF VARCHAR2(20);
streams streamsarray;
total integer;
type allcountsarray is VARRAY(4) OF NUMBER(19);
allcounts allcountsarray;

begin

    streams :=
streamsarray('US2012_GEO','US2012_NON_GEO','US2012_NON_
GEO_HISP','SCOT2014');
    allcounts := allcountsarray(146424, 1560967, 11276,
6477713);
    total := streams.count;

-- loop over streams
for i in 1 .. total loop

    open s1 (mytable);

    loop
        fetch s1 into mycolumn;
        exit when s1%NOTFOUND;

        query_str := 'select count(*) from ' || mytable
|| ' where STREAM = ''' || streams(i) || ''' and (
length(trim(' || mycolumn || ')) = 0 or ' || mycolumn
|| ' is null)';

        execute immediate query_str into mycount;

```

```

        dbms_output.put_line('Column ' || mycolumn || '
has ' || mycount || ' null values in stream ' ||
streams(i));

        -- save data to table
        INSERT INTO ZERO_LEN_SPARSITY_INT_COLS(STREAM,
COLUMN_NAME, NULL_COUNT, ALL_COUNT) VALUES(streams(i),
mycolumn, mycount, allcounts(i));
        COMMIT;

    end loop;

    close s1;

end loop;

-- done
dbms_output.put_line('DONE!');
end;
```

A9.4 SQL query to merge/update records into table

ZERO_LEN_SPARSITY_INTERACTIONS

The programme listing above inserts into ZERO_LEN_SPARSITY_INT_COLS, a staging table with STREAM, COLUMN_NAME, NULL_COUNT, ALL_COUNT for all combinations of Streams (4) and Column Names (149). SPARSITY_INT_COLS has 596 rows (counts for 4*149 columns, three of which – STREAM, STREAMID, UUID – arise from ETL processes).

The following SQL query, edited in turn for each Stream, is used to transpose the resultant 596 rows into 149 rows with 9 columns (COLUMN_NAME and NULL/ALL counts for each of the 4 Streams).

```

-- CREATE A WIDE SPARSITY TABLE with some updates from
http://stackoverflow.com/questions/2446764/update-
statement-with-inner-join-on-oracle
-- CREATE TABLE
"ADRIANT"."ZERO_LEN_SPARSITY_INTERACTIONS"
```

```

--      ("COLUMN_NAME" VARCHAR2(100 BYTE),
-- "GEO_NULL" NUMBER(19,0),
-- "GEO_ALL" NUMBER(19,0),
-- "NONGEO_NULL" NUMBER(19,0),
-- "NONGEO_ALL" NUMBER(19,0),
-- "HISP_NULL" NUMBER(19,0),
-- "HISP_ALL" NUMBER(19,0),
-- "SCOT_NULL" NUMBER(19,0),
-- "SCOT_ALL" NUMBER(19,0)
--  )

-- INSERT LIST OF COLUMNS
-- insert into ZERO_LEN_SPARSITY_INTERACTIONS
(COLUMN_NAME) SELECT DISTINCT COLUMN_NAME FROM
ZERO_LEN_SPARSITY_INT_COLS ORDER BY COLUMN_NAME

merge into ZERO_LEN_SPARSITY_INTERACTIONS t1
using (select * from ZERO_LEN_SPARSITY_INT_COLS where
STREAM='SCOT2014') t2 -- change STREAM;
US2012_GEO/GEO_, US2012_NON_GEO/NONGEO_,
US2012_NON_GEO_HISP/HISP_, SCOT2014/SCOT_
on (t1.COLUMN_NAME = t2.COLUMN_NAME)
when matched then update set t1.SCOT_NULL =
t2.NULL_COUNT, t1.SCOT_ALL = t2.ALL_COUNT -- change
DESTINATION fields

```

The resulting output has been used in Tableau to graph Sparsity (Section 6.4.3, p255).

Appendix 10 ALCHEMYAPI PROCESSING

A10.1 Introduction

Entity extraction, sentiment analysis etc. in AlchemyAPI are achieved by sending text over HTTP to a Cloud-hosted service, with authentication by API key. Successful calls result in a response, again received over HTTP, consisting of JSON data, which may be stored in the database. The code developed to achieve this result, which also had to work within daily rate limit constraints, is reproduced below.

A10.2 Job controllers

The Ruby code below is called by two simple, executable, shell scripts every 10/15 minutes, respectively, by `cron` (the Unix/Linux job scheduler).

A10.2.1 `run_job.sh`

```
#!/bin/bash
cd ~/alchemyapi_ruby
/usr/bin/ruby process_recs.rb
```

A10.2.2 `run_url_job.sh`

```
#!/bin/bash
cd ~/alchemyapi_ruby
/usr/bin/ruby process_url_recs.rb
```

A10.3 INTERACTION_CONTENT processing

These scripts run through 311,575 records of sampled social media text content.

A10.3.1 process_recs.rb

The main program calls AlchemyAPI to determine how many daily transactions are available; if > 0 remain the program calculates a `process_target` number, and sets an `actual_target` number before calling the next program.

```
# script name
puts "\nPROCESS_RECS.RB"

# read xml and parse to determine how many API
# transactions have been consumed today
# broadly from
http://www.nokogiri.org/tutorials/searching_a_xml_html_
document.html

require 'rubygems'
require 'net/http'
require 'uri'
require 'json'
require 'open-uri'

require 'nokogiri'

# url for the check of how many api calls have been
# made
url =
'http://access.alchemyapi.com/calls/info/GetAPIKeyInfo?
apikey=[KEY GOES HERE]'

# the full url (could otherwise have more see
http://stackoverflow.com/questions/32483783/error-
after-trying-to-access-xml-zlibbuferror-ruby
fullurl = url

# use hint at
http://stackoverflow.com/questions/32483783/error-
after-trying-to-access-xml-zlibbuferror-ruby to alter
Accept-Encoding (bizarrely uses an AlchemyAPI example!!
# opener = open(fullurl) {|f| f.read }
xmlreturned = open(fullurl, 'Accept-Encoding' => '')
{|f| f.read }

#puts xmlreturned
=begin
# should output along lines of
```

```

<?xml version="1.0" encoding="UTF-8"?>
<results>
  <status>OK</status>

  <consumedDailyTransactions>182</consumedDailyTransactions>

  <dailyTransactionLimit>30000</dailyTransactionLimit>
</results>
=end

#####
#####
# put the xmlreturned into nokogiri xml structure (as
# Slop!)          #
#####
#####
doc = Nokogiri::Slop(xmlreturned)
# FINALLY, we know how many API calls have been used
# today!!
api_calls_used =
doc.results.consumedDailyTransactions.content
api_calls_limit =
doc.results.dailyTransactionLimit.content
print api_calls_used, " API calls used today...", "\n"

#####
#####
# control how many to process and pass control to main
# program          #
#####
#####
remaining_today = api_calls_limit.to_i -
api_calls_used.to_i
if remaining_today > 0
  print remaining_today, " API calls remaining
  today...", "\n"
  process_target = 150
  if remaining_today > process_target
    @actual_target = process_target
    print @actual_target, " is the target!", "\n"
  else
    @actual_target = (remaining_today/9).ceil #### seems
    to use 9 api calls/record
    print @actual_target, " is the target!", "\n"
  end
  #### @actual_target = 0 # for testing

```

```
if @actual_target >= 1
  load './mytest04.rb'
end
end
```

A10.3.2 mytest04.rb

This program fetches records from Oracle 12c on the VM host, if any remain to be processed on run date, loops over them passing data to AlchemyAPI, using a series of UPDATE statements to post responses back into the database.

```
#script name
puts "\nMYTEST04.RB"

print @actual_target, " is my target...\n"

# start time
starttime = Time.now

# set NLS_LANG from https://www.ruby-
forum.com/topic/188066
ENV['NLS_LANG']='AMERICAN_AMERICA.UTF8'

# database connection gem
require 'oci8'

# alchemy api
require './alchemyapi'

# get some data from the db (IP ADDRESS WILL CHANGE
DEPENDING UPON NETWORK CONNECTION)
@oci = OCI8.new('adriant/[PWD GOES
HERE]@//192.168.47.1:1521/OSNDATA')

# check we have less than 30,000 records processed
today (Alchemy rate limit)
# ruby date formats from
http://apidock.com/ruby/DateTime/strftime
oradate_today = Time.now.strftime("%d-^b-%y")
print "Today is ", oradate_today, "\n"
results_array = []
```



```

@oci.exec("select count(*) from alchemy_api where
trunc(date_processed) = '#{oradate_today}')" { |row|
results_array << row }
print results_array[0][0].to_i, " Records processed
today.\n"

# if we have fewer than 30k (and the API calls
remaining checker has got us to this point) then we're
off
if results_array[0][0].to_i <= 30000

  # results into array from
http://stackoverflow.com/questions/15581388/whats-the-
most-concise-way-to-get-an-oracle-resultset-into-a-
printable-array-in
  # ruby/oracle oci8 integration from
http://www.oracle.com/webfolder/technetwork/tutorials/o
be/db/11g/r2/prod/appdev/opensrclang/rubyrails/rubyrail
s.htm

  # put the query results into an array - use rownum <=
x to restrict number of recs to process
  results_array = []
  # select statement must be in double quotes when
passing a variable in!
  @oci.exec("select cast(a.uuid as VARCHAR(40)),
cast(b.interaction_content as VARCHAR(140)) fred from
alchemy_api a, interactions b where a.uuid = b.uuid and
a.date_processed is null and rownum <=
#{@actual_target}") { |row| results_array << row }

  # how many records selected
  print results_array.length, " RECORDS SELECTED\n\n\n"

  # output the records
  ##puts results_array.join("\n")

  # loop over the records processing with the AlchemyAPI
as required files
  # the AlchemyAPI integration
  # use @ to make it accessible from included file from
http://stackoverflow.com/questions/8334684/how-to-
share-variables-across-my-rb-files
  @alchemyapi = AlchemyAPI.new()
  # LOOP OVER DB RECORDS
  for ss in 0...results_array.length

```

```

# loop
# IN THE RESULTS ARRAY FIELDS FROM QUERY ARE ARRAY
NUMBERS [0]..[N]
#print results_array[ss][0], ":",
results_array[ss][1], "\n"

# set up this_uuid and the demo_text (used in the
included)
# use @ to make it accessible from included file from
http://stackoverflow.com/questions/8334684/how-to-
share-variables-across-my-rb-files
@this_uuid = @demo_text = results_array[ss][0].to_s
@demo_text = results_array[ss][1].to_s
print @this_uuid, "\n"

# setup a sql statement
@sqlstatement = 'update alchemy_api set '

# ***NOTE*** replace single quotes in the returned
response with two single quotes to allow insert in
Oracle
# loosely from
http://stackoverflow.com/questions/2180322/ruby-gsub-
doesnt-escape-single-quotes/2180375#2180375

# ENTITY EXTRACTION
@entity_response = 'NODATA'
load './_entity_extraction_example.rb'
#puts @entity_response, "\n"
if @entity_response != 'NODATA' && @entity_response
!= 'ERROR'
  #@oci.exec("update alchemy_api set entity_json =
'#{@entity_response}' where uuid = '#{@this_uuid}'")
  @sqlstatement << "entity_json =
'#{@entity_response.gsub("'", "'')}'',"
  end

# KEYWORD EXTRACTION
@keyword_response = 'NODATA'
load './_keyword_extraction_example.rb'
#puts @keyword_response, "\n"
if @keyword_response != 'NODATA' && @keyword_response
!= 'ERROR'
  #@oci.exec("update alchemy_api set keyword_json =
'#{@keyword_response}' where uuid = '#{@this_uuid}'")

```

```

    @sqlstatement << "keyword_json =
'#{@keyword_response.gsub("'", "'')}'',"
    end

    # CONCEPT TAGGING
    @concept_response = 'NODATA'
    load './_concept_tagging_example.rb'
    #puts @concept_response, "\n"
    if @concept_response != 'NODATA' && @concept_response
    != 'ERROR'
        #@oci.exec("update alchemy_api set concept_json =
'#{@concept_response}' where uuid = ' #{@this_uuid}'")
        @sqlstatement << "concept_json =
'#{@concept_response.gsub("'", "'')}'',"
        end

    # SENTIMENT ANALYSIS
    @sentiment_response = 'NODATA'
    load './_sentiment_analysis_example.rb'
    #puts @sentiment_response, "\n"
    if @sentiment_response != 'NODATA' &&
    @sentiment_response != 'ERROR'
        #@oci.exec("update alchemy_api set sentiment_json =
'#{@sentiment_response}' where uuid = ' #{@this_uuid}'")
        @sqlstatement << "sentiment_json =
'#{@sentiment_response.gsub("'", "'')}'',"
        end

    # RELATIONS EXTRACTION
    @relations_response = 'NODATA'
    load './_relations_extraction_example.rb'
    #puts @relations_response, "\n"
    if @relations_response != 'NODATA' &&
    @relations_response != 'ERROR'
        #@oci.exec("update alchemy_api set relations_json =
'#{@relations_response}' where uuid = ' #{@this_uuid}'")
        @sqlstatement << "relations_json =
'#{@relations_response.gsub("'", "'')}'',"
        end

    # TEXT CATEGORIZATION
    @category_response = 'NODATA'
    load './_text_categorization_example.rb'
    #puts @category_response, "\n"
    if @category_response != 'NODATA' &&
    @category_response != 'ERROR'

```

```

    #@oci.exec("update alchemy_api set textcat_json =
'#{@category_response}' where uuid = ' #{@this_uuid}'")
    @sqlstatement << "textcat_json =
'#{@category_response.gsub("'", "'')}',"
    end

    # TAXONOMY EXAMPLE
    @taxonomy_response = 'NODATA'
    load './_taxonomy_example.rb'
    #puts @taxonomy_response, "\n"
    if @taxonomy_response != 'NODATA' &&
@taxonomy_response != 'ERROR'
        #@oci.exec("update alchemy_api set taxonomy_json =
'#{@taxonomy_response}' where uuid = ' #{@this_uuid}'")
        @sqlstatement << "taxonomy_json =
'#{@taxonomy_response.gsub("'", "'')}',"
        end

        @sqlstatement << "date_processed = sysdate where uuid
= ' #{@this_uuid}'"

        #print "\n\n", @sqlstatement, "\n\n"

        # update date_processed and commit changes
        @oci.exec(@sqlstatement)
        @oci.commit

    end

end

# end time
endtime = Time.now
difftime = endtime - starttime
print "\nRun in: ", difftime, " seconds\n\n"

```

A10.3.3 Included files

The program mytest04.rb includes several scripts, each calling AlchemyAPI in a specific way, with text passed in by the controlling program, and a JSON response or error returned.

```
puts '# Entity Extraction Example #'

# API
response = @alchemyapi.entities('text', @demo_text, {
  'sentiment'=>1 })

if response['status'] == 'OK'
  #puts '## Response Object ##'
  #puts JSON.pretty_generate(response) # prettified
  ##puts JSON(response) # standard on one line
  @entity_response = JSON(response)
else
  puts 'Error in entity extraction call: ' +
response['statusInfo']
  @entity_response = 'ERROR'
End
```

```
puts '# Keyword Extraction Example #'

# API
response = @alchemyapi.keywords('text', @demo_text, {
  'sentiment'=>1 })

if response['status'] == 'OK'
  @keyword_response = JSON(response)
else
  puts 'Error in keyword extraction call: ' +
response['statusInfo']
  @keyword_response = 'ERROR'
end
```

```
puts '# Concept Tagging Example #'

# API
response = @alchemyapi.concepts('text', @demo_text)

if response['status'] == 'OK'
  @concepts_response = JSON(response)
else
  puts 'Error in concept tagging call: ' +
response['statusInfo']
  @concepts_response = 'ERROR'
end
```

```
puts '# Sentiment Analysis Example #'

# API
response = @alchemyapi.sentiment('text', @demo_text)
```

```

if response['status'] == 'OK'
  @sentiment_response = JSON(response)
else
  puts 'Error in sentiment analysis call: ' +
response['statusInfo']
  @sentiment_response = 'ERROR'
end

puts '# Relation Extraction Example #'

# API
response = @alchemyapi.relations('text', @demo_text)

if response['status'] == 'OK'
  @relations_response = JSON(response)
else
  puts 'Error in relation extraction call: ' +
response['statusInfo']
  @relations_response = 'ERROR'
end

puts '# Text Categorization Example #'

# API
response = @alchemyapi.category('text', @demo_text)

if response['status'] == 'OK'
  @category_response = JSON(response)
else
  puts 'Error in text categorization call: ' +
response['statusInfo']
  @category_response = 'ERROR'
end

puts '# Taxonomy Example #'

# API
response = @alchemyapi.taxonomy('text', @demo_text)

if response['status'] == 'OK'
  @taxonomy_response = JSON(response)
else
  puts 'Error in taxonomy call: ' +
response['statusInfo']
  @taxonomy_response = 'ERROR'
end

```

A10.4 LI_LINKS_URLS_DISTINCT processing

These scripts run through a queue of 641,472 distinct link URLs.

A10.4.1 process_url_recs.rb

In a similar way to `process_recs.rb`, this script determines how many transactions remain today before calling the next program.

```
# script name
puts "\nPROCESS_URL_RECS.RB"

# read xml and parse to determine how many API
transactions have been consumed today
# broadly from
http://www.nokogiri.org/tutorials/searching_a_xml_html_
document.html

require 'rubygems'
require 'net/http'
require 'uri'
require 'json'
require 'open-uri'

require 'nokogiri'

# url for the check of how many api calls have been
made
url =
'http://access.alchemyapi.com/calls/info/GetAPIKeyInfo?
apikey=[KEY GOES HERE]'

# the full url (could otherwise have more see
http://stackoverflow.com/questions/32483783/error-
after-trying-to-access-xml-zlibbuferror-ruby
fullurl = url

# use hint at
http://stackoverflow.com/questions/32483783/error-
after-trying-to-access-xml-zlibbuferror-ruby to alter
Accept-Encoding (bizarrely uses an AlchemyAPI example!!
# opener = open(fullurl) {|f| f.read }
xmlreturned = open(fullurl, 'Accept-Encoding' => '')
{|f| f.read }
```

```

#puts xmlreturned
=begin
# should output along lines of
<?xml version="1.0" encoding="UTF-8"?>
<results>
    <status>OK</status>

<consumedDailyTransactions>182</consumedDailyTransactio
ns>

<dailyTransactionLimit>30000</dailyTransactionLimit>
</results>
=end

#####
#####
# put the xmlreturned into nokogiri xml structure (as
Slop!) #
#####
#####
doc = Nokogiri::Slop(xmlreturned)
# FINALLY, we know how many API calls have been used
today!!
api_calls_used =
doc.results.consumedDailyTransactions.content
api_calls_limit =
doc.results.dailyTransactionLimit.content
print api_calls_used, " API calls used today...", "\n"

#####
#####
# control how many to process and pass control to main
program #
#####
#####
remaining_today = api_calls_limit.to_i -
api_calls_used.to_i
if remaining_today > 0
    print remaining_today, " API calls remaining
today...", "\n"
    process_target = 250
    if remaining_today > process_target
        @actual_target = process_target
        print @actual_target, " is the target!", "\n"
    else

```



```

    @actual_target = (remaining_today/2).ceil #### seems
to use 1 api calls/record on URLs
    print @actual_target, " is the target!", "\n"
end
#### @actual_target = 0 # for testing
if @actual_target >= 1
    load './url_test04.rb'
end
end
end

```

A10.4.2 urltest04.rb

This program fetches records from Oracle 12c on the VM host, if any remain to be processed on run date, loops over them passing data to AlchemyAPI, using an UPDATE statement to post responses back into the database.

```

#script name
puts "\nURL_TEST04.RB"

####@actual_target=10

print @actual_target, " is my target...\n"

# start time
starttime = Time.now

# set NLS_LANG from https://www.ruby-
forum.com/topic/188066
ENV['NLS_LANG']='AMERICAN_AMERICA.UTF8'

# database connection gem
require 'oci8'

# alchemy api
require './alchemyapi'

# get some data from the db (IP ADDRESS WILL CHANGE
DEPENDING UPON NETWORK CONNECTION)
@oci = OCI8.new('adriant/[PWD GOES
HERE]@//192.168.47.1:1521/OSNDATA')

# check we have less than 30,000 records processed
today (Alchemy rate limit)

```

```

# ruby date formats from
http://apidock.com/ruby/DateTime/strftime
oradate_today = Time.now.strftime("%d-%^b-%y")
print "Today is ", oradate_today, "\n"
results_array = []
@oci.exec("select count(*) from li_links_urls_distinct
where trunc(date_processed) = '#{oradate_today}'") {
|row| results_array << row }
print results_array[0][0].to_i, " Records processed
today.\n"

# if we have fewer than 30k (and the API calls
remaining checker has got us to this point) then we're
off
if results_array[0][0].to_i <= 30000

  # results into array from
http://stackoverflow.com/questions/15581388/whats-the-
most-concise-way-to-get-an-oracle-resultset-into-a-
printable-array-in
  # ruby/oracle oci8 integration from
http://www.oracle.com/webfolder/technetwork/tutorials/o
be/db/11g/r2/prod/appdev/opensrclang/rubyrails/rubyrail
s.htm

  # put the query results into an array - use rownum <=
x to restrict number of recs to process
  results_array = []
  # select statement must be in double quotes when
passing a variable in!
  @oci.exec("select uuid, link_url from
li_links_urls_distinct where date_processed is null and
rownum <= #{@actual_target}") { |row| results_array <<
row }

  # how many records selected
  print results_array.length, " RECORDS SELECTED\n\n\n"

  # output the records
  ##puts results_array.join("\n")

  # loop over the records processing with the AlchemyAPI
as required files
  # the AlchemyAPI integration

```

```

# use @ to make it accessible from included file from
http://stackoverflow.com/questions/8334684/how-to-
share-variables-across-my-rb-files
@alchemyapi = AlchemyAPI.new()
# LOOP OVER DB RECORDS
for ss in 0...results_array.length

  # loop
  @this_uuid = @demo_text = results_array[ss][0].to_s
  @this_url = results_array[ss][1].to_s
  print @this_url, "\n"

  # ENTITY EXTRACTION
  @entity_response = 'NODATA'

  # API
  response = @alchemyapi.entities('url', @this_url)

  if response['status'] == 'OK'
    # the JSON returned by the API
    ##puts JSON.pretty_generate(response) # prettified
    ##puts JSON(response) # standard on one line
    @entity_response = JSON(response)
  else
    # store the error as a snippet of JSON to store in
    the db
    @entity_response = '{"Error":"' +
response['statusInfo'] + '"'
    puts @entity_response
  end

  # UPDATE the clob with a bind parameter (or with the
  error message if it has failed, e.g. 404)
  # from https://learncodeshare.net/2016/11/04/update-
  crud-using-ruby-oci8/
  statement = "update li_links_urls_distinct set
entity_json = :this_response, date_processed = sysdate
where uuid = :this_uuid"
  cursor = @oci.parse(statement)
  cursor.bind_param(:this_response,@entity_response)
  cursor.bind_param(:this_uuid,@this_uuid)
  cursor.exec
  @oci.commit

end

```

```

end

# end time
endtime = Time.now
difftime = endtime - starttime
print "\nRun in: ", difftime, " seconds\n\n"

```

A10.5 Sample output

AlchemyAPI produces verbose output in JSON. The following Entities detected by AlchemyAPI against CNN's Scottish Independence Referendum results page (URL=<http://edition.cnn.com/2014/09/18/world/europe/scotland-independence-vote/index.html>) are shown below.

```

{
  "status": "OK",
  "usage": "By accessing AlchemyAPI or using
information generated by AlchemyAPI, you are agreeing
to be bound by the AlchemyAPI Terms of Use:
http://www.alchemyapi.com/company/terms.html",
  "url":
"http://edition.cnn.com/2014/09/18/world/europe/scotland-independence-vote/index.html",
  "language": "english",
  "entities": [{
    "type": "Country",
    "relevance": "0.813336",
    "count": "22",
    "text": "Scotland",
    "disambiguated": {
      "subType": ["Location",
"AdministrativeDivision", "GovernmentalJurisdiction"],
      "name": "Scotland",
      "website": "http://www.scotland.org/",
      "dbpedia":
"http://dbpedia.org/resource/Scotland",
      "freebase":
"http://rdf.freebase.com/ns/m.01xk6b",

```

```

        "geonames":
        "http://sws.geonames.org/2643576/",
        "opencyc":
        "http://sw.opencyc.org/concept/Mx4rvViIm5wpEbGdrcN5Y29ycA",
        "yago": "http://yago-knowledge.org/resource/Scotland"
    }, {
        "type": "City",
        "relevance": "0.436839",
        "count": "6",
        "text": "Glasgow",
        "disambiguated": {
            "subType": ["AdministrativeDivision",
"ScottishCouncilArea"],
            "name": "Glasgow",
            "website":
"http://www.glasgow.gov.uk/",
            "dbpedia":
"http://dbpedia.org/resource/Glasgow",
            "freebase":
"http://rdf.freebase.com/ns/m.0hyxv",
            "geonames":
"http://sws.geonames.org/2648579/",
            "yago": "http://yago-knowledge.org/resource/Glasgow"
        }
    }, {
        "type": "City",
        "relevance": "0.408424",
        "count": "5",
        "text": "Edinburgh",
        "disambiguated": {
            "subType": ["AdministrativeDivision",
"PlaceWithNeighborhoods",
"AwardPresentingOrganization"],
            "name": "Edinburgh",
            "website":
"http://www.edinburgh.gov.uk/",
            "dbpedia":
"http://dbpedia.org/resource/Edinburgh",
            "freebase":
"http://rdf.freebase.com/ns/m.02m77",
            "geonames":
"http://sws.geonames.org/2650225/",

```

```

        "yago": "http://yago-
knowledge.org/resource/Edinburgh"
    }, {
        "type": "Company",
        "relevance": "0.321778",
        "count": "4",
        "text": "CNN",
        "disambiguated": {
            "subType": ["Broadcast", "AwardWinner",
"RadioNetwork", "TVNetwork"],
            "name": "CNN",
            "website": "http://www.cnn.com/",
            "dbpedia":
"http://dbpedia.org/resource/CNN",
            "freebase":
"http://rdf.freebase.com/ns/m.0gsgr",
            "yago": "http://yago-
knowledge.org/resource/CNN"
        }
    }, {
        "type": "Person",
        "relevance": "0.321343",
        "count": "3",
        "text": "Alex Salmond",
        "disambiguated": {
            "subType": ["Politician",
"OfficeHolder", "TVActor"],
            "name": "Alex Salmond",
            "website": "http://www.snp.org",
            "dbpedia":
"http://dbpedia.org/resource/Alex Salmond",
            "freebase":
"http://rdf.freebase.com/ns/m.0lk0jf",
            "yago": "http://yago-
knowledge.org/resource/Alex Salmond"
        }
    }, {
        "type": "Country",
        "relevance": "0.288505",
        "count": "3",
        "text": "United Kingdom",
        "disambiguated": {
            "subType": ["Location",
"AdministrativeDivision", "GovernmentalJurisdiction",
"Kingdom", "MeteorologicalService"],
            "name": "United Kingdom",

```

```

        "geo": "51.5 -0.11666666666666667",
        "website": "http://www.royal.gov.uk/",
        "dbpedia":
"http://dbpedia.org/resource/United_Kingdom",
        "freebase":
"http://rdf.freebase.com/ns/m.07ssc",
        "ciaFactbook": "http://www4.wiwiss.fu-
berlin.de/factbook/resource/United_Kingdom",
        "opencyc":
"http://sw.opencyc.org/concept/Mx4rvViRhJwpEbGdrcN5Y29y
cA",
        "yago": "http://yago-
knowledge.org/resource/United_Kingdom"
    }, {
        "type": "Person",
        "relevance": "0.284499",
        "count": "2",
        "text": "Prime Minister David Cameron"
    }, {
        "type": "City",
        "relevance": "0.260409",
        "count": "2",
        "text": "Edinburgh"
    }, {
        "type": "StateOrCounty",
        "relevance": "0.233592",
        "count": "2",
        "text": "Aberdeenshire"
    }, {
        "type": "Region",
        "relevance": "0.227693",
        "count": "2",
        "text": "Northern Ireland"
    }, {
        "type": "Country",
        "relevance": "0.222848",
        "count": "2",
        "text": "Wales",
        "disambiguated": {
            "subType": ["Location",
"AdministrativeDivision", "GovernmentalJurisdiction",
"FilmScreeningVenue"],
            "name": "Wales",
            "website": "http://www.visitwales.com",
            "dbpedia":
"http://dbpedia.org/resource/Wales",

```

```

        "freebase":
        "http://rdf.freebase.com/ns/m.0j5g9",
        "geonames":
        "http://sws.geonames.org/2636718/",
        "opencyc":
        "http://sw.opencyc.org/concept/Mx4rvVitivJwpEbGdrcN5Y29ycA",
        "yago": "http://yago-knowledge.org/resource/Wales"
    }, {
        "type": "Organization",
        "relevance": "0.222649",
        "count": "1",
        "text": "Scottish Parliament",
        "disambiguated": {
            "subType": ["GovernmentalBody"],
            "name": "Scottish Parliament",
            "geo": "55.95194 -3.17513",
            "website":
            "http://www.scottish.parliament.uk",
            "dbpedia":
            "http://dbpedia.org/resource/Scottish Parliament",
            "freebase":
            "http://rdf.freebase.com/ns/m.0glvp",
            "yago": "http://yago-knowledge.org/resource/Scottish Parliament"
        }
    }, {
        "type": "Country",
        "relevance": "0.220393",
        "count": "2",
        "text": "England",
        "disambiguated": {
            "subType": ["Location",
            "PoliticalDistrict", "AdministrativeDivision",
            "GovernmentalJurisdiction"],
            "name": "England",
            "website": "http://www.direct.gov.uk/",
            "dbpedia":
            "http://dbpedia.org/resource/England",
            "freebase":
            "http://rdf.freebase.com/ns/m.02jx1",
            "geonames":
            "http://sws.geonames.org/3333218/",

```



```

        "opencyc":
"http://sw.opencyc.org/concept/Mx4rvViWaZwpEbGdrcN5Y29ycA",
        "yago": "http://yago-knowledge.org/resource/England"
    }, {
        "type": "JobTitle",
        "relevance": "0.219891",
        "count": "1",
        "text": "Prime minister"
    }, {
        "type": "Person",
        "relevance": "0.218625",
        "count": "2",
        "text": "Sue Bruce"
    }, {
        "type": "JobTitle",
        "relevance": "0.21754",
        "count": "2",
        "text": "officer"
    }, {
        "type": "Person",
        "relevance": "0.212998",
        "count": "1",
        "text": "Prime Minister Gordon Brown",
        "disambiguated": {
            "subType": ["Politician",
"PoliticalAppointer", "TVActor"],
            "name": "Gordon Brown",
            "website":
"http://www.number10.gov.uk/",
            "dbpedia":
"http://dbpedia.org/resource/Gordon Brown",
            "freebase":
"http://rdf.freebase.com/ns/m.03f77",
            "yago": "http://yago-knowledge.org/resource/Gordon Brown"
        }
    }, {
        "type": "City",
        "relevance": "0.208436",
        "count": "1",
        "text": "Dundee",
        "disambiguated": {
            "subType": ["AdministrativeDivision",
"ScottishCouncilArea"],

```

```

        "name": "Dundee",
        "dbpedia":
        "http://dbpedia.org/resource/Dundee",
        "freebase":
        "http://rdf.freebase.com/ns/m.02fvv",
        "geonames":
        "http://sws.geonames.org/2650752/",
        "yago": "http://yago-
knowledge.org/resource/Dundee"
    }, {
        "type": "FieldTerminology",
        "relevance": "0.207396",
        "count": "1",
        "text": "oil-rich city"
    }, {
        "type": "Organization",
        "relevance": "0.206967",
        "count": "1",
        "text": "Glasgow City Council"
    }, {
        "type": "City",
        "relevance": "0.196357",
        "count": "1",
        "text": "Aberdeen",
        "disambiguated": {
            "subType": ["AdministrativeDivision",
"ScottishCouncilArea"],
            "name": "Aberdeen",
            "website":
            "http://www.aberdeencity.gov.uk/",
            "dbpedia":
            "http://dbpedia.org/resource/Aberdeen",
            "freebase":
            "http://rdf.freebase.com/ns/m.0rng",
            "geonames":
            "http://sws.geonames.org/2657832/",
            "yago": "http://yago-
knowledge.org/resource/Aberdeen"
        }
    }, {
        "type": "Person",
        "relevance": "0.191401",
        "count": "1",
        "text": "Phil MacHugh",
        "disambiguated": {
            "name": "Phil MacHugh",

```

```

        "dbpedia":
        "http://dbpedia.org/resource/Phil MacHugh",
        "freebase":
        "http://rdf.freebase.com/ns/m.09gdw11"
    }, {
        "type": "Person",
        "relevance": "0.188891",
        "count": "1",
        "text": "Alistair Darling",
        "disambiguated": {
            "subType": ["Politician",
"Chancellor"],
            "name": "Alistair Darling",
            "dbpedia":
            "http://dbpedia.org/resource/Alistair Darling",
            "freebase":
            "http://rdf.freebase.com/ns/m.01zgx3",
            "yago": "http://yago-
knowledge.org/resource/Alistair Darling"
        }
    }, {
        "type": "Person",
        "relevance": "0.187902",
        "count": "1",
        "text": "Nic Robertson",
        "disambiguated": {
            "subType": [],
            "name": "Nic Robertson",
            "dbpedia":
            "http://dbpedia.org/resource/Nic Robertson",
            "freebase":
            "http://rdf.freebase.com/ns/m.08783w",
            "yago": "http://yago-
knowledge.org/resource/Nic Robertson"
        }
    }, {
        "type": "Organization",
        "relevance": "0.18525",
        "count": "1",
        "text": "EU"
    }, {
        "type": "Person",
        "relevance": "0.184526",
        "count": "1",
        "text": "Mary Pitcaithly"
    }, {

```

```

        "type": "City",
        "relevance": "0.183566",
        "count": "1",
        "text": "Hong Kong",
        "disambiguated": {
            "subType": ["HumanLanguage",
"AdministrativeDivision", "Country",
"GovernmentalJurisdiction", "BodyOfWater", "Cuisine"],
            "name": "Hong Kong",
            "geo": "22.27833333333332
114.15888888888889",
            "website": "http://www.gov.hk/en/",
            "dbpedia":
"http://dbpedia.org/resource/Hong Kong",
            "freebase":
"http://rdf.freebase.com/ns/m.03h64",
            "geonames":
"http://sws.geonames.org/1819727/",
            "ciaFactbook": "http://www4.wiwiss.fu-berlin.de/factbook/resource/Hong Kong",
            "opencyc":
"http://sw.opencyc.org/concept/Mx4rvVipapwpEbGdrcN5Y29ycA",
            "yago": "http://yago-knowledge.org/resource/Hong Kong"
        }
    }, {
        "type": "City",
        "relevance": "0.183279",
        "count": "1",
        "text": "Dumfries",
        "disambiguated": {
            "subType": [],
            "name": "Dumfries",
            "dbpedia":
"http://dbpedia.org/resource/Dumfries",
            "freebase":
"http://rdf.freebase.com/ns/m.0zc6f",
            "geonames":
"http://sws.geonames.org/2650798/",
            "yago": "http://yago-knowledge.org/resource/Dumfries"
        }
    }, {
        "type": "Person",
        "relevance": "0.181844",
        "count": "1",

```

```

        "text": "Angus"
    }, {
        "type": "City",
        "relevance": "0.179558",
        "count": "1",
        "text": "London",
        "disambiguated": {
            "subType": ["AdministrativeDivision",
"GovernmentalJurisdiction", "OlympicHostCity",
"PlaceWithNeighborhoods"],
            "name": "London",
            "geo": "51.50805555555556 -
0.1247222222222222",
            "website": "http://www.london.gov.uk/",
            "dbpedia":
"http://dbpedia.org/resource/London",
            "freebase":
"http://rdf.freebase.com/ns/m.04jpl",
            "geonames":
"http://sws.geonames.org/2643743/",
            "yago": "http://yago-
knowledge.org/resource/London"
        }
    }, {
        "type": "Region",
        "relevance": "0.177794",
        "count": "1",
        "text": "East Dunbartonshire"
    }, {
        "type": "Crime",
        "relevance": "0.177649",
        "count": "1",
        "text": "fraud"
    }, {
        "type": "Person",
        "relevance": "0.174821",
        "count": "1",
        "text": "Euan McKirdy"
    }, {
        "type": "City",
        "relevance": "0.174539",
        "count": "1",
        "text": "Kirkcaldy",
        "disambiguated": {
            "subType": [],
            "name": "Kirkcaldy",

```

```

        "dbpedia":
        "http://dbpedia.org/resource/Kirkcaldy",
        "freebase":
        "http://rdf.freebase.com/ns/m.01zrs ",
        "geonames":
        "http://sws.geonames.org/2645298/",
        "yago": "http://yago-
knowledge.org/resource/Kirkcaldy"
    }
}, {
    "type": "City",
    "relevance": "0.173872",
    "count": "1",
    "text": "Galloway",
    "disambiguated": {
        "subType": [],
        "name": "Galloway, West Virginia",
        "dbpedia":
        "http://dbpedia.org/resource/Galloway, West Virginia",
        "freebase":
        "http://rdf.freebase.com/ns/m.041799d"
    }
}, {
    "type": "City",
    "relevance": "0.172553",
    "count": "1",
    "text": "Strichen",
    "disambiguated": {
        "subType": [],
        "name": "Strichen",
        "geo": "57.5865 -2.0904",
        "dbpedia":
        "http://dbpedia.org/resource/Strichen",
        "freebase":
        "http://rdf.freebase.com/ns/m.0271hd1",
        "geonames":
        "http://sws.geonames.org/2636654/",
        "yago": "http://yago-
knowledge.org/resource/Strichen"
    }
}, {
    "type": "Person",
    "relevance": "0.170637",
    "count": "1",
    "text": "Laura Smith-Spark"
}, {
    "type": "Person",

```

```

        "relevance": "0.163373",
        "count": "1",
        "text": "Richard Allen Greene"
    }, {
        "type": "Person",
        "relevance": "0.153864",
        "count": "1",
        "text": "Greg Botelho"
    }, {
        "type": "Person",
        "relevance": "0.144819",
        "count": "1",
        "text": "Lindsay Isaac"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "17-year"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "46%"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "54%"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "75%"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "80%"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",
        "text": "86%"
    }, {
        "type": "Quantity",
        "relevance": "0.144819",
        "count": "1",

```

```
    "text": "8%"  
  }  
]  
}
```


Appendix 11 SQL STATEMENTS

Many Structured Query Language (SQL) statements have been designed and executed during this research; to create and populate tables, check data consistency upon loading and to query resultant data sets.

The full set of over 100 statements is too voluminous to reproduce here. However, several key SQL statements, referenced in the text by the item number adjacent to each SQL statement, are shown below.

Item	SQL
1.	<pre>SELECT TRUNC(INTERACTION_CREATED_AT) AS INT_DATE, COUNT(*) AS N_OBAMA FROM INTERACTIONS WHERE STREAM <> 'SCOT2014' AND CONTAINS (INTERACTION_CONTENT, 'OBAMA') > 0 GROUP BY TRUNC(INTERACTION_CREATED_AT) ORDER BY TRUNC(INTERACTION_CREATED_AT)</pre>
2.	<pre>SELECT COUNT(*) FROM INTERACTIONS WHERE STREAM LIKE 'US%' AND INTERACTION_CONTENT LIKE '%OH%'</pre>
3.	<pre>SELECT YEAR, TYPE, COUNT(*) FROM DOCUMENTS GROUP BY YEAR, TYPE ORDER BY YEAR</pre>
4.	<pre>SELECT JSON_DATAGUIDE(JSON_DOC) FROM JSON_INTERACTIONS</pre>
5.	<pre>CREATE TABLE scot2014_jdump_00001 (json_document CLOB) ORGANIZATION EXTERNAL (TYPE ORACLE_LOADER DEFAULT DIRECTORY scot2014_entry_dir ACCESS PARAMETERS (RECORDS DELIMITED BY '\n' READSIZE 1048576 CHARACTERSET 'utf8' DISABLE_DIRECTORY_LINK_CHECK BADFILE scot2014_output_dir: 'JSONDumpFile_00001.bad' LOGFILE scot2014_output_dir: 'JSONDumpFile_00001.log' FIELDS (json_document CHAR(1048576))) LOCATION (scot2014_entry_dir:'part-r-00001.json')) PARALLEL</pre>

	REJECT LIMIT UNLIMITED;
6.	CREATE TABLE SCOT2014_JSON (JSON_DOC CLOB, CONSTRAINT ENSURE_SCOT_JSON CHECK (JSON_DOC IS JSON))
7.	INSERT INTO SCOT2014_JSON (JSON_DOC) SELECT JSON_DOC FROM scot2014_jdump_00001
8.	SELECT INTERACTION_AUTHOR_NAME, INTERACTION_CONTENT FROM INTERACTIONS WHERE INTERACTION_AUTHOR_NAME = 'FM Alex Salmond'
9.	SELECT COUNT(*) FROM INTERACTIONS
10.	SELECT STREAM, COUNT(*) FROM INTERACTIONS GROUP BY STREAM
11.	SELECT INTERACTION_AUTHOR_ID, COUNT(*) AS N FROM INTERACTIONS GROUP BY INTERACTION_AUTHOR_ID ORDER BY N DESC
12.	-- USERS MAKING LTE 5 INTERACTIONS, HOW MANY INTERACTIONS IN TOTAL = 3171447 SELECT SUM(USER_N_INTS) FROM (SELECT INTERACTION_AUTHOR_ID, COUNT(*) AS USER_N_INTS FROM INTERACTIONS GROUP BY INTERACTION_AUTHOR_ID HAVING COUNT(*) <= 5) -- USERS MAKING GTE 6 INTERACTIONS, HOW MANY INTERACTIONS IN TOTAL = 5024933 SELECT SUM(USER_N_INTS) FROM (SELECT INTERACTION_AUTHOR_ID, COUNT(*) AS USER_N_INTS FROM INTERACTIONS GROUP BY INTERACTION_AUTHOR_ID HAVING COUNT(*) >= 6)
13.	/***** * * ALL

```

*
*****/
SELECT
  'ALL' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
/*****
*
* US2012
*
*****/
-- US2012
SELECT
  'US2012' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  B.STREAM <> 'SCOT2014'
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
/*****
*
* BY STREAM (WILL PICK UP SCOT2014)
*
*****/
-- stream
SELECT
  STREAM AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
GROUP BY
  STREAM,
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
/*****

```

```

*
* BY SOURCE (ALL)
*
*****/
-- facebook
SELECT
  'FB - ALL' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  B.FB_ID IS NOT NULL
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
-- twitter
SELECT
  'TW - ALL' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  ((B.TW_RT_ID IS NULL OR B.TW_RT_ID = ''))
AND
  (B.TW_ID IS NOT NULL))
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
-- retweet
SELECT
  'RT - ALL' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  ((B.TW_RT_ID IS NOT NULL OR B.TW_RT_ID <> ''))
AND
  (B.TW_ID IS NOT NULL))
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
/*****

```

```

*
* BY SOURCE (GEO)
*
*****/
-- facebook geo
SELECT
  'FB - GEO' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  B.FB_ID IS NOT NULL
AND
  INTERACTION_GEO_LATITUDE IS NOT NULL
AND
  INTERACTION_GEO_LATITUDE <> 0
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
-- twitter tweets geo
SELECT
  'TW - GEO' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND
  ((B.TW_RT_ID IS NULL OR B.TW_RT_ID = ''))
AND
  (B.TW_ID IS NOT NULL))
AND
  INTERACTION_GEO_LATITUDE IS NOT NULL
AND
  INTERACTION_GEO_LATITUDE <> 0
GROUP BY
  INTERACTION_AUTHOR_ID
-- keep going
UNION ALL
-- twitter retweets geo
SELECT
  'RT - GEO' AS COLDESC,
  INTERACTION_AUTHOR_ID,
  COUNT(*) AS N
FROM
  VW_INT_GEO_SCORING_TOTAL A,
  INTERACTIONS B
WHERE
  A.UUID = B.UUID
AND

```

	<pre> ((B.TW_RT_ID IS NOT NULL OR B.TW_RT_ID <> '') AND (B.TW_ID IS NOT NULL)) AND TW_RTED_GEO_LATITUDE IS NOT NULL AND TW_RTED_GEO_LATITUDE <> 0 GROUP BY INTERACTION_AUTHOR_ID /***** * * DONE ALL PERMUTATIONS * *****/ </pre>
14.	<pre> SELECT COUNT(*) FROM VW_USER_COUNT_INTS WHERE N >= 1000 ORDER BY N DESC </pre>
15.	<pre> SELECT TW_PLACE_COUNTRY, TW_PLACE_COUNTRY_CODE, TW_PLACE_FULL_NAME, TW_PLACE_ID, TW_PLACE_NAME, TW_PLACE_PLACE_TYPE FROM INTERACTIONS WHERE TW_PLACE_ID IS NOT NULL AND LENGTH(TRIM(TW_PLACE_ID)) <> 0 </pre>
16.	<pre> CREATE OR REPLACE VIEW VW_INT_GEO_SCORING_TOTAL AS SELECT UUID, (case when (TW_RT_USER_GEO_ENABLED is not null) then 1 else 0 end) as P_TW_RT_USER_GEO_ENABLED, (case when (TW_RTED_USER_GEO_ENABLED is not null) then 1 else 0 end) as P_TW_RTED_USER_GEO_ENABLED, (case when (TW_RTED_USER_TIME_ZONE is not null) then 1 else 0 end) as P_TW_RTED_USER_TIME_ZONE, (case when (TW_RTED_USER_UTC_OFFSET is not null) then 1 else 0 end) as P_TW_RTED_USER_UTC_OFFSET, (case when (TW_RTED_USER_LOCATION is not null) then 1 else 0 end) as P_TW_RTED_USER_LOCATION, (case when (TW_RT_USER_LOCATION is not null) then 1 else 0 end) as P_TW_RT_USER_LOCATION, (case when (TW_USER_LOCATION is not null) then 1 else 0 end) as P_TW_USER_LOCATION, (case when (TW_USER_TIME_ZONE is not null) then 1 else 0 end) as P_TW_USER_TIME_ZONE, </pre>

```

(case
  when (TW_USER_UTC_OFFSET is not null) then 1
  else 0 end) as P_TW_USER_UTC_OFFSET,
(case
  when (TW_RT_USER_TIME_ZONE is not null) then 1
  else 0 end) as P_TW_RT_USER_TIME_ZONE,
(case
  when (TW_RT_USER_UTC_OFFSET is not null) then 1
  else 0 end) as P_TW_RT_USER_UTC_OFFSET,
(case
  when (INTERACTION_GEO_LATITUDE is not null) then 50
  else 0 end) as P_INTERACTION_GEO_LATITUDE,
(case
  when (INTERACTION_GEO_LONGITUDE is not null) then 50
  else 0 end) as P_INTERACTION_GEO_LONGITUDE,
(case
  when (TW_PLACE_FULL_NAME is not null) then 1
  else 0 end) as P_TW_PLACE_FULL_NAME,
(case
  when (TW_RTED_PLACE_FULL_NAME is not null) then 1
  else 0 end) as P_TW_RTED_PLACE_FULL_NAME,
(case
  when (TW_RTED_GEO_LATITUDE is not null) then 100
  else 0 end) as P_TW_RTED_GEO_LATITUDE,
(case
  when (TW_RTED_GEO_LONGITUDE is not null) then 100
  else 0 end) as P_TW_RTED_GEO_LONGITUDE,
(case
  when (TW_PLACE_ATT_ST_ADDRESS is not null) then 1
  else 0 end) as P_TW_PLACE_ATT_ST_ADDRESS,
(case
  when (TW_RTED_PLACE_ATT_ST_ADDRESS is not null) then 1
  else 0 end) as P_TW_RTED_PLACE_ATT_ST_ADDRESS,
(case
  when (TW_PLACE_ATT_LOCALITY is not null) then 1
  else 0 end) as P_TW_PLACE_ATT_LOCALITY,
(case
  when (TW_PLACE_ATT_REGION is not null) then 1
  else 0 end) as P_TW_PLACE_ATT_REGION,
0 as P_TOTAL FROM INTERACTIONS

```

```

17. CREATE OR REPLACE VIEW VW_INT_GEO_SCORING_TOTAL AS
SELECT UUID,
(P_TW_RT_USER_GEO_ENABLED+
P_TW_RTED_USER_GEO_ENABLED+
P_TW_RTED_USER_TIME_ZONE+
P_TW_RTED_USER_UTC_OFFSET+
P_TW_RTED_USER_LOCATION+
P_TW_RT_USER_LOCATION+
P_TW_USER_LOCATION+
P_TW_USER_TIME_ZONE+
P_TW_USER_UTC_OFFSET+
P_TW_RT_USER_TIME_ZONE+
P_TW_RT_USER_UTC_OFFSET+
P_INTERACTION_GEO_LATITUDE+
P_INTERACTION_GEO_LONGITUDE+
P_TW_PLACE_FULL_NAME+

```

	P_TW_RTED_PLACE_FULL_NAME+ P_TW_RTED_GEO_LATITUDE+ P_TW_RTED_GEO_LONGITUDE+ P_TW_PLACE_ATT_ST_ADDRESS+ P_TW_RTED_PLACE_ATT_ST_ADDRESS+ P_TW_PLACE_ATT_LOCALITY+ P_TW_PLACE_ATT_REGION+ 0) as P_TOTAL FROM VW_INT_GEO_SCORING
18.	SELECT MOST_FREQ_AUTHOR_P_TOTAL, COUNT(*) FROM (SELECT B.INTERACTION_AUTHOR_ID, STATS_MODE(A.P_TOTAL) AS MOST_FREQ_AUTHOR_P_TOTAL FROM VW_INT_GEO_SCORING_TOTAL A, INTERACTIONS B WHERE A.UUID = B.UUID GROUP BY B.INTERACTION_AUTHOR_ID) GROUP BY MOST_FREQ_AUTHOR_P_TOTAL ORDER BY MOST_FREQ_AUTHOR_P_TOTAL
19.	SELECT TW_USER_UTC_OFFSET, COUNT(*) FROM INTERACTIONS WHERE STREAM <> 'SCOT2014' -- FOR US2012 OR = TO GET SCOT2014 GROUP BY TW_USER_UTC_OFFSET ORDER BY TW_USER_UTC_OFFSET
20.	SELECT "GATEUS2012_JSON", "TEXT", "INTERACTION_ID", "LOCTYPE" FROM GATE_NER_US2012 T, JSON_TABLE (T.GATEUS2012_JSON, '\$' COLUMNS (TEXT VARCHAR2(100) PATH '\$.text', INTERACTION_ID VARCHAR2(100) PATH '\$.id_str', NESTED PATH '\$.entities.Location[*]' COLUMNS (LOCTYPE VARCHAR2(100) PATH '\$.locType'))) D WHERE GATEUS2012_JSON IS JSON STRICT AND D.LOCTYPE IS NOT NULL
21.	select b.* from interactions a, Gate_Ner_Us2012 b where a.interaction_id=b.GATEUS2012_JSON.id_str and a.interaction_id = '1e227914e2f4ac80e0740cf699462aae'
22.	select A.UUID, COUNT(*) AS N_GATE_LOC_ENTITIES from

	<pre> INTERACTIONS A, VW_GATE_NER_US2012_LOC B where A.INTERACTION_ID = B.INTERACTION_ID group by A.UUID </pre>
23.	<pre> insert into alchemy_api (uuid, tranche, date_loaded) select uuid, 'US2012_GEO Stream', sysdate from interactions where stream = 'US2012_GEO' </pre>
24.	<pre> insert into alchemy_api (uuid, tranche, date_loaded) select uuid, 'US2012_NON_GEO 1% sample tweets', sysdate from interactions sample(1) where stream = 'US2012_NON_GEO' and tw_id is not null </pre>
25.	<pre> insert into alchemy_api (uuid, tranche, date_loaded) select uuid, 'SCOT2014 geo-tagged tweets', sysdate from interactions where stream = 'SCOT2014' and tw_id is not null and (interaction_geo_latitude is not null and interaction_geo_latitude <> 0) and (interaction_geo_longitude is not null and interaction_geo_longitude <> 0) </pre>
26.	<pre> insert into alchemy_api (uuid, tranche, date_loaded) select uuid, 'SCOT2014 1% sample tweets', sysdate from interactions sample(1) where stream = 'SCOT2014' and tw_id is not null </pre>
27.	<pre> SELECT TRANCHE, COUNT(*) FROM VW_ALCHEMY_INTERACTIONS C </pre>

	<pre> WHERE (C.TYPE IN ('COUNTRY', 'CITY', 'STATEORCOUNTY', 'CONTINENT', 'GEOGRAPHICFEATURE', 'REGION') OR C.GEO IS NOT NULL) GROUP BY TRANCHE </pre>
28.	<pre> SELECT CAST(A.UUID AS VARCHAR2(40)) AS UUID, CAST(A.INTERACTION_CONTENT AS VARCHAR2(4000)) AS INT_CONTENT_4000, TRUNC(A.INTERACTION_CREATED_AT) AS DATE_CREATED_AT, JT.* FROM INTERACTIONS A INNER JOIN GEO_CLAVIN_030_MINJSON_OUT B ON A.UUID = B.UUID, JSON_TABLE(clavin_json, '\$.resolvedLocationsMinimum[*]' COLUMNS (row_number FOR ORDINALITY, GEONAMEID NUMBER(19,0) PATH '\$.geonameID', NAME VARCHAR2(200) PATH '\$.name', COUNTRYCODE VARCHAR2(2) PATH '\$.countryCode', LATITUDE NUMBER(19,8) PATH '\$.latitude', LONGITUDE NUMBER(19,8) PATH '\$.longitude')) AS JT </pre>
29.	<pre> CREATE TABLE LI_LINKS_URLS AS (SELECT UUID, STREAM, STREAMID, TRIM(REGEXP_SUBSTR(REPLACE(REPLACE(REPLACE(LINKS_URL, ['', ''], ']', ''), '"', '') , '[^,]+' , 1, LEVELS.COLUMN_VALUE)) AS LINK_URL FROM INTERACTIONS T , TABLE(CAST(MULTISET(SELECT LEVEL FROM DUAL CONNECT BY LEVEL <= LENGTH (REGEXP_REPLACE(T.LINKS_URL, '[^,]+')) + 1) AS SYS.ODCINUMBERLIST)) LEVELS WHERE LINKS_URL IS NOT NULL) </pre>
30.	<pre> SELECT SUBSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), 1, INSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), '/', 1)-1) AS DOMAIN, COUNT(*) FROM LI_LINKS_URLS A, INTERACTIONS B </pre>

	<pre> WHERE A.UUID = B.UUID AND LINK_URL IS NOT NULL AND SUBSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), 1, INSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), '/', 1)-1) IS NOT NULL AND A.STREAM = 'SCOT2014' /* ALTER THIS <>/= TO GET US2012/SCOT2014 */ --AND (b.interaction_geo_latitude is null or b.tw_rted_geo_latitude is null) -- Get non-geographic posters (whether retweeted or not) AND (B.INTERACTION_GEO_LATITUDE IS NOT NULL OR B.TW_RTED_GEO_LATITUDE IS NOT NULL) -- Get geographic posters (whether reretweeted or not) GROUP BY SUBSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), 1, INSTR((REPLACE(REPLACE(LINK_URL, 'http://', ''), 'https://', '')), '/', 1)-1) ORDER BY COUNT(*) DESC </pre>
31.	<pre> CREATE TABLE LI_LINKS_URLS_DISTINCT AS (SELECT DISTINCT LINK_URL FROM LI_LINKS_URLS) </pre>
32.	<pre> SELECT T.LINK_URL, SUBSTR(D.GEO, 1, INSTR(D.GEO, ' ')-1) AS LAT, SUBSTR(D.GEO, INSTR(D.GEO, ' ')+1) AS LON, D."URL", D."LANG", D."TYPE", D."RELEVANCE", D."COUNT", D."TEXT", D."GEO" FROM LI_LINKS_URLS_DISTINCT T, json_table (T.ENTITY_JSON, '\$' COLUMNS (url VARCHAR2(100) PATH '\$.url', lang VARCHAR2(20) PATH '\$.language', NESTED PATH '\$.entities[*]' COLUMNS (type VARCHAR2(100) PATH '\$.type', relevance NUMBER PATH '\$.relevance', count NUMBER PATH '\$.count', text VARCHAR2(100) PATH '\$.text', geo VARCHAR2(20) PATH '\$.disambiguated.geo'))) D WHERE ENTITY_JSON IS JSON STRICT </pre>
33.	<pre> CREATE OR REPLACE VIEW VW_ALCHEMY_LINKS_N_INT AS SELECT A.UUID, COUNT(*) AS N_GEOMENTIONSINLINKS_INT FROM </pre>

	<pre> INTERACTIONS A, LI_LINKS_URLS B, VW_ALCHEMY_LINKS_URLS C, VW_INT_GEO_SCORING_TOTAL D WHERE A.UUID = B.UUID AND B.LINK_URL = C.LINK_URL AND A.UUID = D.UUID AND (C.TYPE IN ('Country', 'City', 'StateOrCounty', 'Continent', 'GeographicFeature', 'Region') OR C.GEO IS NOT NULL) GROUP BY A.UUID </pre>
34.	<pre> CREATE OR REPLACE VIEW VW_ALCHEMY_LINKS_N_USR AS SELECT A.INTERACTION_AUTHOR_ID, COUNT(*) AS N_GEOMENTIONSINLINKS_USR FROM INTERACTIONS A, LI_LINKS_URLS B, VW_ALCHEMY_LINKS_URLS C, VW_INT_GEO_SCORING_TOTAL D WHERE A.UUID = B.UUID AND B.LINK_URL = C.LINK_URL AND A.UUID = D.UUID AND (C.TYPE IN ('Country', 'City', 'StateOrCounty', 'Continent', 'GeographicFeature', 'Region') OR C.GEO IS NOT NULL) GROUP BY A.INTERACTION_AUTHOR_ID </pre>
35.	<pre> SELECT B.P_TOTAL, AVG(A.N_GEOMENTIONSINLINKS_INT) AS AVG_GEOMENTIONSINLINKS_INT FROM VW_ALCHEMY_LINKS_N_INT A, VW_INT_GEO_SCORING_TOTAL B WHERE </pre>

	A.UUID = B.UUID GROUP BY B.P_TOTAL
36.	SELECT B.MOST_FREQ_AUTHOR_P_TOTAL AS P_TOTAL, AVG(A.N_GEOMENTIONSINLINKS_USR) AS AVG_GEOMENTIONSINLINKS_USR FROM VW_ALCHEMY_LINKS_N_USR A, VW_USER_GEO_SCORING_MOSTFREQ B WHERE A.INTERACTION_AUTHOR_ID = B.INTERACTION_AUTHOR_ID GROUP BY B.MOST_FREQ_AUTHOR_P_TOTAL
37.	CREATE OR REPLACE VIEW VW_STATS_GT_US_I_TW_NOTGEO AS SELECT A.UUID, A.N_GATE_LOC_ENTITIES AS NUMB FROM VW_GATE_NER_US2012_LOC_N_INT A, INTERACTIONS B WHERE A.UUID = B.UUID AND ((B.TW_RT_ID IS NULL OR B.TW_RT_ID = '') AND (B.TW_ID IS NOT NULL)) AND B.INTERACTION_GEO_LATITUDE IS NULL
38.	SELECT STREAM, COUNT(*) FROM INTERACTIONS WHERE LOWER(INTERACTION_CONTENT) LIKE '%perth%' GROUP BY STREAM
39.	SELECT COUNT(*) FROM INTERACTIONS WHERE TW_ID IN (SELECT TW_RTED_ID FROM INTERACTIONS)
40.	SELECT COUNT(*) FROM INTERACTIONS WHERE TW_RTED_GEO_LATITUDE IS NOT NULL
41.	SELECT COUNT(*) FROM INTERACTIONS A, INT_TW_RTED_GEO_NONNULL B WHERE A.TW_ID = B.TW_RTED_ID
42.	SELECT COUNT(*)

	<pre> FROM INTERACTIONS A, INT_TW_RTED_GEO_NONNULL B WHERE A.TW_ID = B.TW_RTED_ID AND A.INTERACTION_GEO_LATITUDE <> B.TW_RTED_GEO_LATITUDE </pre>
43.	<pre> SELECT COUNT(*) FROM (SELECT DISTINCT(A.TW_ID) FROM INTERACTIONS A, INT_TW_RTED_GEO_NONNULL B WHERE A.TW_ID = B.TW_RTED_ID AND A.INTERACTION_GEO_LATITUDE <> B.TW_RTED_GEO_LATITUDE) </pre>
44.	<pre> SELECT INTERACTION_AUTHOR_NAME, ROUND(AVG(TW_USER_FOLLOWERS_COUNT)) FROM INTERACTIONS WHERE STREAM <> 'SCOT2014' GROUP BY INTERACTION_AUTHOR_NAME HAVING AVG(TW_USER_FOLLOWERS_COUNT) > 1000000 ORDER BY AVG(TW_USER_FOLLOWERS_COUNT) DESC </pre>
45.	<pre> SELECT INTERACTION_AUTHOR_NAME, ROUND(AVG(TW_USER_FOLLOWERS_COUNT)) FROM INTERACTIONS WHERE STREAM <> 'SCOT2014' AND INTERACTION_GEO_LATITUDE IS NOT NULL GROUP BY INTERACTION_AUTHOR_NAME HAVING AVG(TW_USER_FOLLOWERS_COUNT) > 1000000 ORDER BY AVG(TW_USER_FOLLOWERS_COUNT) DESC </pre>

Appendix 12 STATISTICAL ANALYSIS IN R

A12.1 R scripts

Several scripts written in R (The R Foundation, 2018) have been used to test the statistical significance of results reported in this thesis.

These R scripts, referenced in the text by the item number adjacent to each script, are shown below.

Item	R script
1.	<pre># Read CSV into R MyData <- read.csv('~/Desktop/R_STATS/VW_USER_COUNT_INTS_SRC_TYP_GEO .CSV', header=TRUE, sep=',') # normal summary summary(MyData\$N) # by the combinations s_all <- MyData [which (MyData\$COLDESC == 'ALL'),] s_us2012 <- MyData [which (MyData\$COLDESC == 'US2012'),] s_us2012_geo <- MyData [which (MyData\$COLDESC == 'US2012_GEO'),] s_us2012_non_geo <- MyData [which (MyData\$COLDESC == 'US2012_NON_GEO'),] s_us2012_non_geo_hisp <- MyData [which (MyData\$COLDESC == 'US2012_NON_GEO_HISP'),] s_scot2014 <- MyData [which (MyData\$COLDESC == 'SCOT2014'),] s_fb_all <- MyData [which (MyData\$COLDESC == 'FB - ALL'),] s_tw_all <- MyData [which (MyData\$COLDESC == 'TW - ALL'),] s_rt_all <- MyData [which (MyData\$COLDESC == 'RT - ALL'),] s_fb_geo <- MyData [which (MyData\$COLDESC == 'FB - GEO'),] s_tw_geo <- MyData [which (MyData\$COLDESC == 'TW - GEO'),] s_rt_geo <- MyData [which (MyData\$COLDESC == 'RT - GEO'),] # PACKAGE from https://cran.r- project.org/web/packages/psych/index.html library(psych) describe(s_all\$N) describe(s_us2012\$N) describe(s_scot2014\$N) describe(s_fb_all\$N) describe(s_tw_all\$N) describe(s_rt_all\$N) describe(s_fb_geo\$N) describe(s_tw_geo\$N)</pre>

	<code>describe(s_rt_geo\$N)</code>
2.	<pre>#### compute psych descriptive statistics #### # set working directory setwd('/media/sf_R_DATA/FOR_T_TESTS') #my_files <- Sys.glob("*GT*US*_I*.CSV") # to get only GT etc. my_files <- list.files(pattern = "*.CSV\$") # to get all files into a list # read the files my_data <- lapply(my_files, read.csv) # names the files names(my_data) <- gsub("*.CSV\$", "", my_files) # what have we go names(my_data) lapply(my_data, head) # vector of col name I want to extract my_col = c("NUMB") # get at that column for each list element (ie file read in) my_col_numb = lapply(my_data, "[", , my_col) # output it for each list element lapply(my_col_numb, head) # calculate means for each list element my_means <- lapply(my_col_numb, mean) my_means ####boxplot(my_col_numb) # kills R # psych library(psych) my_describes <- lapply(my_col_numb, describe) # run describe on NUMB in each file my_describes # what have we got names(my_describes) # store results in a data frame my_describes_df <- data.frame(t(sapply(my_describes,c))) # save output my_describes_df_char <- apply(my_describes_df,2,as.character) write.csv(my_describes_df_char, file='output/all_psych_describe_output.csv') #### the above works but it's easiest to copy/paste results from data frame viewer into LibreOffice</pre>
3.	<pre>#### R SCRIPT DRIVEN FROM 002 - Excel control for t- tests.xlsx ####</pre>


```

# set working directory
setwd('/media/sf_R_DATA/FOR_T_TESTS')

# empty list to store t test output in
outlist <- list()

# read CSV files into R; this is driven by the spreadsheet
to match like with like
VW_STATS_GT_SC_I_FB_ISGEO <-
read.csv('VW_STATS_GT_SC_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_SC_I_FB_NOTGEO <-
read.csv('VW_STATS_GT_SC_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_I_FB_NOTGEO$NUMB,VW_STATS_GT_SC
_I_FB_ISGEO$NUMB)))
VW_STATS_GT_SC_I_RT_ISGEO <-
read.csv('VW_STATS_GT_SC_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_SC_I_RT_NOTGEO <-
read.csv('VW_STATS_GT_SC_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_I_RT_NOTGEO$NUMB,VW_STATS_GT_SC
_I_RT_ISGEO$NUMB)))
VW_STATS_GT_SC_I_TW_ISGEO <-
read.csv('VW_STATS_GT_SC_I_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_SC_I_TW_NOTGEO <-
read.csv('VW_STATS_GT_SC_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_I_TW_NOTGEO$NUMB,VW_STATS_GT_SC
_I_TW_ISGEO$NUMB)))
VW_STATS_GT_SC_U_FB_ISGEO <-
read.csv('VW_STATS_GT_SC_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_SC_U_FB_NOTGEO <-
read.csv('VW_STATS_GT_SC_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_U_FB_NOTGEO$NUMB,VW_STATS_GT_SC
_U_FB_ISGEO$NUMB)))
VW_STATS_GT_SC_U_RT_ISGEO <-
read.csv('VW_STATS_GT_SC_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_SC_U_RT_NOTGEO <-
read.csv('VW_STATS_GT_SC_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_U_RT_NOTGEO$NUMB,VW_STATS_GT_SC
_U_RT_ISGEO$NUMB)))
VW_STATS_GT_SC_U_TW_ISGEO <-
read.csv('VW_STATS_GT_SC_U_TW_ISGEO.CSV', header=TRUE,
sep=',')

```

```

VW_STATS_GT_SC_U_TW_NOTGEO <-
read.csv('VW_STATS_GT_SC_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_SC_U_TW_NOTGEO$NUMB,VW_STATS_GT_SC
_U_TW_ISGEO$NUMB)))
VW_STATS_GT_US_I_FB_ISGEO <-
read.csv('VW_STATS_GT_US_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_I_FB_NOTGEO <-
read.csv('VW_STATS_GT_US_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_I_FB_NOTGEO$NUMB,VW_STATS_GT_US
_I_FB_ISGEO$NUMB)))
VW_STATS_GT_US_I_RT_ISGEO <-
read.csv('VW_STATS_GT_US_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_I_RT_NOTGEO <-
read.csv('VW_STATS_GT_US_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_I_RT_NOTGEO$NUMB,VW_STATS_GT_US
_I_RT_ISGEO$NUMB)))
VW_STATS_GT_US_I_TW_ISGEO <-
read.csv('VW_STATS_GT_US_I_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_I_TW_NOTGEO <-
read.csv('VW_STATS_GT_US_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_I_TW_NOTGEO$NUMB,VW_STATS_GT_US
_I_TW_ISGEO$NUMB)))
VW_STATS_GT_US_U_FB_ISGEO <-
read.csv('VW_STATS_GT_US_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_U_FB_NOTGEO <-
read.csv('VW_STATS_GT_US_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_U_FB_NOTGEO$NUMB,VW_STATS_GT_US
_U_FB_ISGEO$NUMB)))
VW_STATS_GT_US_U_RT_ISGEO <-
read.csv('VW_STATS_GT_US_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_U_RT_NOTGEO <-
read.csv('VW_STATS_GT_US_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_U_RT_NOTGEO$NUMB,VW_STATS_GT_US
_U_RT_ISGEO$NUMB)))
VW_STATS_GT_US_U_TW_ISGEO <-
read.csv('VW_STATS_GT_US_U_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_GT_US_U_TW_NOTGEO <-
read.csv('VW_STATS_GT_US_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')

```

```

outlist <- append(outlist,
list(t.test(VW_STATS_GT_US_U_TW_NOTGEO$NUMB,VW_STATS_GT_US
_U_TW_ISGEO$NUMB)))

VW_STATS_CL_SC_I_FB_ISGEO <-
read.csv('VW_STATS_CL_SC_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_I_FB_NOTGEO <-
read.csv('VW_STATS_CL_SC_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_I_FB_NOTGEO$NUMB,VW_STATS_CL_SC
_I_FB_ISGEO$NUMB)))
VW_STATS_CL_SC_I_RT_ISGEO <-
read.csv('VW_STATS_CL_SC_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_I_RT_NOTGEO <-
read.csv('VW_STATS_CL_SC_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_I_RT_NOTGEO$NUMB,VW_STATS_CL_SC
_I_RT_ISGEO$NUMB)))
VW_STATS_CL_SC_I_TW_ISGEO <-
read.csv('VW_STATS_CL_SC_I_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_I_TW_NOTGEO <-
read.csv('VW_STATS_CL_SC_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_I_TW_NOTGEO$NUMB,VW_STATS_CL_SC
_I_TW_ISGEO$NUMB)))
VW_STATS_CL_SC_U_FB_ISGEO <-
read.csv('VW_STATS_CL_SC_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_U_FB_NOTGEO <-
read.csv('VW_STATS_CL_SC_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_U_FB_NOTGEO$NUMB,VW_STATS_CL_SC
_U_FB_ISGEO$NUMB)))
VW_STATS_CL_SC_U_RT_ISGEO <-
read.csv('VW_STATS_CL_SC_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_U_RT_NOTGEO <-
read.csv('VW_STATS_CL_SC_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_U_RT_NOTGEO$NUMB,VW_STATS_CL_SC
_U_RT_ISGEO$NUMB)))
VW_STATS_CL_SC_U_TW_ISGEO <-
read.csv('VW_STATS_CL_SC_U_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_SC_U_TW_NOTGEO <-
read.csv('VW_STATS_CL_SC_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')

```

```

outlist <- append(outlist,
list(t.test(VW_STATS_CL_SC_U_TW_NOTGEO$NUMB,VW_STATS_CL_SC
_U_TW_ISGEO$NUMB)))
VW_STATS_CL_US_I_FB_ISGEO <-
read.csv('VW_STATS_CL_US_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_I_FB_NOTGEO <-
read.csv('VW_STATS_CL_US_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_I_FB_NOTGEO$NUMB,VW_STATS_CL_US
_I_FB_ISGEO$NUMB)))
VW_STATS_CL_US_I_RT_ISGEO <-
read.csv('VW_STATS_CL_US_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_I_RT_NOTGEO <-
read.csv('VW_STATS_CL_US_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_I_RT_NOTGEO$NUMB,VW_STATS_CL_US
_I_RT_ISGEO$NUMB)))
VW_STATS_CL_US_I_TW_ISGEO <-
read.csv('VW_STATS_CL_US_I_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_I_TW_NOTGEO <-
read.csv('VW_STATS_CL_US_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_I_TW_NOTGEO$NUMB,VW_STATS_CL_US
_I_TW_ISGEO$NUMB)))
VW_STATS_CL_US_U_FB_ISGEO <-
read.csv('VW_STATS_CL_US_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_U_FB_NOTGEO <-
read.csv('VW_STATS_CL_US_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_U_FB_NOTGEO$NUMB,VW_STATS_CL_US
_U_FB_ISGEO$NUMB)))
VW_STATS_CL_US_U_RT_ISGEO <-
read.csv('VW_STATS_CL_US_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_U_RT_NOTGEO <-
read.csv('VW_STATS_CL_US_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_U_RT_NOTGEO$NUMB,VW_STATS_CL_US
_U_RT_ISGEO$NUMB)))
VW_STATS_CL_US_U_TW_ISGEO <-
read.csv('VW_STATS_CL_US_U_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_CL_US_U_TW_NOTGEO <-
read.csv('VW_STATS_CL_US_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_CL_US_U_TW_NOTGEO$NUMB,VW_STATS_CL_US
_U_TW_ISGEO$NUMB)))

```

```

VW_STATS_AL_SC_I_TR2_ISGEO <-
read.csv('VW_STATS_AL_SC_I_TR2_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_AL_SC_I_TR2_NOTGEO <-
read.csv('VW_STATS_AL_SC_I_TR2_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_AL_SC_I_TR2_NOTGEO$NUMB,VW_STATS_AL_S
C_I_TR2_ISGEO$NUMB)))
VW_STATS_AL_SC_U_TR2_ISGEO <-
read.csv('VW_STATS_AL_SC_U_TR2_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_AL_SC_U_TR2_NOTGEO <-
read.csv('VW_STATS_AL_SC_U_TR2_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_AL_SC_U_TR2_NOTGEO$NUMB,VW_STATS_AL_S
C_U_TR2_ISGEO$NUMB)))
VW_STATS_AL_US_I_TR1_ISGEO <-
read.csv('VW_STATS_AL_US_I_TR1_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_AL_US_I_TR1_NOTGEO <-
read.csv('VW_STATS_AL_US_I_TR1_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_AL_US_I_TR1_NOTGEO$NUMB,VW_STATS_AL_U
S_I_TR1_ISGEO$NUMB)))
VW_STATS_AL_US_U_TR1_ISGEO <-
read.csv('VW_STATS_AL_US_U_TR1_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_AL_US_U_TR1_NOTGEO <-
read.csv('VW_STATS_AL_US_U_TR1_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_AL_US_U_TR1_NOTGEO$NUMB,VW_STATS_AL_U
S_U_TR1_ISGEO$NUMB)))

VW_STATS_LI_SC_I_FB_ISGEO <-
read.csv('VW_STATS_LI_SC_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_I_FB_NOTGEO <-
read.csv('VW_STATS_LI_SC_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_I_FB_NOTGEO$NUMB,VW_STATS_LI_SC
_I_FB_ISGEO$NUMB)))
VW_STATS_LI_SC_I_RT_ISGEO <-
read.csv('VW_STATS_LI_SC_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_I_RT_NOTGEO <-
read.csv('VW_STATS_LI_SC_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_I_RT_NOTGEO$NUMB,VW_STATS_LI_SC
_I_RT_ISGEO$NUMB)))

```

```

VW_STATS_LI_SC_I_TW_ISGEO <-
read.csv('VW_STATS_LI_SC_I_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_I_TW_NOTGEO <-
read.csv('VW_STATS_LI_SC_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_I_TW_NOTGEO$NUMB,VW_STATS_LI_SC
_I_TW_ISGEO$NUMB)))
VW_STATS_LI_SC_U_FB_ISGEO <-
read.csv('VW_STATS_LI_SC_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_U_FB_NOTGEO <-
read.csv('VW_STATS_LI_SC_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_U_FB_NOTGEO$NUMB,VW_STATS_LI_SC
_U_FB_ISGEO$NUMB)))
VW_STATS_LI_SC_U_RT_ISGEO <-
read.csv('VW_STATS_LI_SC_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_U_RT_NOTGEO <-
read.csv('VW_STATS_LI_SC_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_U_RT_NOTGEO$NUMB,VW_STATS_LI_SC
_U_RT_ISGEO$NUMB)))
VW_STATS_LI_SC_U_TW_ISGEO <-
read.csv('VW_STATS_LI_SC_U_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_SC_U_TW_NOTGEO <-
read.csv('VW_STATS_LI_SC_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_SC_U_TW_NOTGEO$NUMB,VW_STATS_LI_SC
_U_TW_ISGEO$NUMB)))
VW_STATS_LI_US_I_FB_ISGEO <-
read.csv('VW_STATS_LI_US_I_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_US_I_FB_NOTGEO <-
read.csv('VW_STATS_LI_US_I_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_I_FB_NOTGEO$NUMB,VW_STATS_LI_US
_I_FB_ISGEO$NUMB)))
VW_STATS_LI_US_I_RT_ISGEO <-
read.csv('VW_STATS_LI_US_I_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_US_I_RT_NOTGEO <-
read.csv('VW_STATS_LI_US_I_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_I_RT_NOTGEO$NUMB,VW_STATS_LI_US
_I_RT_ISGEO$NUMB)))
VW_STATS_LI_US_I_TW_ISGEO <-
read.csv('VW_STATS_LI_US_I_TW_ISGEO.CSV', header=TRUE,
sep=',')

```

```

VW_STATS_LI_US_I_TW_NOTGEO <-
read.csv('VW_STATS_LI_US_I_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_I_TW_NOTGEO$NUMB,VW_STATS_LI_US
_I_TW_ISGEO$NUMB)))
VW_STATS_LI_US_U_FB_ISGEO <-
read.csv('VW_STATS_LI_US_U_FB_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_US_U_FB_NOTGEO <-
read.csv('VW_STATS_LI_US_U_FB_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_U_FB_NOTGEO$NUMB,VW_STATS_LI_US
_U_FB_ISGEO$NUMB)))
VW_STATS_LI_US_U_RT_ISGEO <-
read.csv('VW_STATS_LI_US_U_RT_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_US_U_RT_NOTGEO <-
read.csv('VW_STATS_LI_US_U_RT_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_U_RT_NOTGEO$NUMB,VW_STATS_LI_US
_U_RT_ISGEO$NUMB)))
VW_STATS_LI_US_U_TW_ISGEO <-
read.csv('VW_STATS_LI_US_U_TW_ISGEO.CSV', header=TRUE,
sep=',')
VW_STATS_LI_US_U_TW_NOTGEO <-
read.csv('VW_STATS_LI_US_U_TW_NOTGEO.CSV', header=TRUE,
sep=',')
outlist <- append(outlist,
list(t.test(VW_STATS_LI_US_U_TW_NOTGEO$NUMB,VW_STATS_LI_US
_U_TW_ISGEO$NUMB)))

# TURN APPENDED LIST INTO A DATA FRAME
my_outlist_df <- data.frame(t(sapply(outlist,c)))

# SAVE OUTPUT
my_outlist_df_char <- apply(my_outlist_df,2,as.character)
write.csv(my_outlist_df_char,
file='output/all_t_test_X_NOTGEO_Y_ISGEO.csv')

```

A12.2 Detailed statistical results and commentary

The following pages present detailed results of statistical tests comparing numbers of NLP-detectable toponymic place names found at interaction and user levels in the varying sources and subtypes of social media data collected during the two case study events examined in this research (Section 4.2.4, p126). Results based upon statistical and other analyses, derived by text-mining social media data using three NLP systems described in Section 4.4.1 (p147), are presented throughout Chapter 5 (p186) with summary statistics given in Section 5.3 (p219).

Multiple *levels* of statistical analysis are reported in detail in the tables which follow, as illustrated graphically in Figure A12-1.

Event	Level	Geotagged	Source/Subtype
US2012 / SCOT2014	Interaction	Yes	FB
			TW
			RT
		No	FB
			TW
			RT
	User	Yes	FB
			TW
			RT
		No	FB
			TW
			RT

Figure A12-1 – Hierarchical levels for statistical analysis of toponymic place name detection in case study social media data

In each case statistics are presented for TwitIE on GATEcloud (GT), AlchemyAPI (AL against message text; LI against links) and CLAVIN-rest (CL). Statistics report T and P scores for numbers of NLP-detectable toponymic mentions in Facebook (FB), Twitter tweet (TW) and retweet (RT) message text or linked/shared URL content, whether coordinate-geotagged (GEO=Y) or not (GEO=N). Descriptive statistics, at interaction and user levels respectively, for the US2012 event are shown in Table

A12-1 (p504) and Table A12-2 (p505). Similar tables for the SCOT2014 event are shown in Table A12-3 (p506) and Table A12-4 (p507). In each case tables show NLP/geoparser used, OSN source (src), coordinate-geotagged status (geo) together with numbers of records (n), mean, standard deviation (sd), median, trimmed mean, median absolute deviation (mad), minimum (min), maximum (max), range, skew, kurtosis and standard error (se) calculated using R's `psych` package (Revelle, 2018). Table A12-5 (p508) and Table A12-6 (p509) show T statistics and P values computed for US2012 and SCOT2014 events by geoparser, OSN source, level (I=interaction, U=user) and is/is-not coordinate-geotagged permutations.

Table A12-1 – US2012 geoparsing descriptive statistics (interaction level)

nlp	src	geo	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
GT	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
GT	FB	N	13,341	1.90	2.79	1	1.39	0.00	1	161	160	19.01	847.07	0.02
GT	TW	Y	21,455	1.41	0.82	1	1.25	0.00	1	20	19	4.17	40.66	0.01
GT	TW	N	104,303	1.22	0.57	1	1.09	0.00	1	14	13	3.92	25.40	0.00
GT	RT	Y	1,103	1.20	0.53	1	1.07	0.00	1	6	5	3.40	14.57	0.02
GT	RT	N	124,892	1.21	0.54	1	1.08	0.00	1	14	13	4.19	30.35	0.00
AL	TW	Y	18,321	1.17	0.52	1	1.05	0.00	1	10	9	4.67	35.06	0.00
AL	TW	N	2,231	1.16	0.45	1	1.05	0.00	1	5	4	3.36	14.80	0.01
CL	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
CL	FB	N	12,199	1.80	2.42	1	1.37	0.00	1	159	158	26.08	1,495.61	0.02
CL	TW	Y	16,958	1.17	0.49	1	1.05	0.00	1	9	8	4.49	34.81	0.00
CL	TW	N	84,706	1.19	0.51	1	1.06	0.00	1	9	8	3.87	23.61	0.00
CL	RT	Y	895	1.14	0.44	1	1.02	0.00	1	7	6	4.89	40.18	0.01
CL	RT	N	100,704	1.18	0.51	1	1.06	0.00	1	10	9	4.59	35.09	0.00
LI	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
LI	FB	N	12,554	4.97	4.25	4	4.31	2.97	1	50	49	1.99	8.36	0.04
LI	TW	Y	6,637	5.58	5.37	4	4.60	4.45	1	51	50	2.11	6.81	0.07
LI	TW	N	95,812	5.66	5.15	4	4.80	4.45	1	54	53	1.90	5.74	0.02
LI	RT	Y	393	4.99	4.32	4	4.29	4.45	1	24	23	1.41	1.93	0.22
LI	RT	N	92,225	5.48	5.10	4	4.59	4.45	1	53	52	2.09	6.82	0.02

Table A12-2 – US2012 geoparsing descriptive statistics (user level)

nlp	src	geo	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
GT	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
GT	FB	N	17,224	2.47	3.73	1	1.73	0.00	1	161	160	11.33	278.53	0.03
GT	TW	Y	62,923	7.33	12.63	2	4.18	1.48	1	108	107	3.58	15.91	0.05
GT	TW	N	366,345	5.44	12.88	2	3.22	1.48	1	306	305	13.58	264.67	0.02
GT	RT	Y	4,474	4.02	6.58	1	2.42	0.00	1	79	78	4.11	23.84	0.10
GT	RT	N	359,233	5.41	8.51	2	3.47	1.48	1	177	176	4.88	47.88	0.01
AL	TW	Y	13,614	1.58	1.96	1	1.20	0.00	1	49	48	9.91	147.11	0.02
AL	TW	N	2,195	1.18	0.48	1	1.07	0.00	1	5	4	3.24	13.61	0.01
CL	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
CL	FB	N	16,096	2.62	6.23	1	1.76	0.00	1	425	424	30.31	1,590.71	0.05
CL	TW	Y	82,617	4.23	8.36	2	2.57	1.48	1	370	369	14.61	423.09	0.03
CL	TW	N	283,606	5.77	15.62	2	3.28	1.48	1	1,099	1,098	19.04	706.27	0.03
CL	RT	Y	2,592	4.61	10.41	2	3.05	1.48	1	427	426	26.80	1,050.20	0.20
CL	RT	N	313,564	4.87	10.93	2	3.27	1.48	1	1,281	1,280	58.68	6,110.18	0.02
LI	FB	Y	0	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
LI	FB	N	15,583	8.65	12.51	5	6.35	4.45	1	389	388	7.42	116.19	0.10
LI	TW	Y	32,887	61.15	170.35	14	25.05	17.79	1	5,301	5,300	8.11	122.82	0.94
LI	TW	N	213,075	64.15	193.08	18	31.13	22.24	1	9,011	9,010	16.07	487.13	0.42
LI	RT	Y	1,740	44.23	77.41	20	30.34	23.72	1	1,937	1,936	10.39	213.31	1.86
LI	RT	N	241,875	44.71	118.00	19	29.30	23.72	1	13,877	13,876	64.42	7,128.63	0.24

Table A12-3 – SCOT2014 geoparsing descriptive statistics (interaction level)

nlp	src	geo	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
GT	FB	Y	481	4.83	17.53	2	2.34	1.48	1	299	298	13.12	197.49	0.80
GT	FB	N	370,293	5.66	12.01	2	2.90	1.48	1	706	705	8.83	201.62	0.02
GT	TW	Y	23,037	1.29	0.63	1	1.15	0.00	1	9	8	2.92	11.83	0.00
GT	TW	N	810,198	1.31	0.63	1	1.17	0.00	1	10	9	2.42	7.07	0.00
GT	RT	Y	26,670	1.25	0.55	1	1.13	0.00	1	6	5	2.31	5.59	0.00
GT	RT	N	858,109	1.28	0.59	1	1.16	0.00	1	10	9	2.97	14.48	0.00
AL	TW	Y	19,914	1.20	0.49	1	1.09	0.00	1	6	5	2.78	10.22	0.00
AL	TW	N	14,677	1.19	0.45	1	1.08	0.00	1	7	6	2.77	10.99	0.00
CL	FB	Y	508	3.05	4.29	1	2.01	0.00	1	34	33	3.66	15.62	0.19
CL	FB	N	394,604	3.98	5.67	2	2.53	1.48	1	119	118	2.90	10.43	0.01
CL	TW	Y	18,048	1.24	0.53	1	1.12	0.00	1	6	5	2.61	8.37	0.00
CL	TW	N	649,215	1.25	0.58	1	1.11	0.00	1	10	9	2.82	9.28	0.00
CL	RT	Y	21,295	1.21	0.48	1	1.10	0.00	1	6	5	2.36	5.52	0.00
CL	RT	N	679,272	1.22	0.51	1	1.11	0.00	1	10	9	2.85	12.42	0.00
LI	FB	Y	144	4.76	7.36	2	3.22	1.48	1	66	65	4.93	33.46	0.61
LI	FB	N	276,409	5.43	9.17	3	3.88	2.97	1	374	373	10.10	159.60	0.02
LI	TW	Y	10,719	8.65	6.51	7	8.11	7.41	1	51	50	0.71	-0.15	0.06
LI	TW	N	749,541	6.76	5.87	5	5.81	4.45	1	72	71	2.04	7.00	0.01
LI	RT	Y	9,864	6.32	4.81	5	5.69	4.45	1	45	44	1.26	1.82	0.05
LI	RT	N	649,704	6.49	5.33	5	5.67	4.45	1	62	61	1.53	3.09	0.01

Table A12-4 – SCOT2014 geoparsing descriptive statistics (user level)

nlp	src	geo	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
GT	FB	Y	681	31.18	89.36	5	12.63	5.93	1	1,298	1,297	7.12	71.43	3.42
GT	FB	N	631,764	389.52	2,043.08	26	63.45	35.58	1	24,688	24,687	9.43	99.08	2.57
GT	TW	Y	76,191	196.93	664.07	11	41.41	14.83	1	3,667	3,666	4.73	21.59	2.41
GT	TW	N	2,200,470	374.57	1,309.35	23	70.23	32.62	1	11,549	11,548	5.40	32.86	0.88
GT	RT	Y	82,517	203.24	393.76	48	110.63	68.20	1	3,667	3,666	3.97	21.18	1.37
GT	RT	N	2,343,135	184.38	395.68	31	90.44	44.48	1	3,938	3,937	4.37	25.31	0.26
AL	TW	Y	11,177	2.14	9.33	1	1.36	0.00	1	535	534	39.33	1,806.47	0.09
AL	TW	N	11,576	1.50	1.60	1	1.21	0.00	1	48	47	11.83	222.11	0.01
CL	FB	Y	672	25.25	60.29	4	13.38	4.45	1	853	852	6.83	67.14	2.33
CL	FB	N	626,440	325.24	1,599.81	23	54.04	31.13	1	16,781	16,780	8.07	69.69	2.02
CL	TW	Y	68,473	160.68	507.82	10	36.79	13.34	1	2,667	2,666	4.42	18.76	1.94
CL	TW	N	2,136,945	256.28	797.54	17	56.25	23.72	1	5,291	5,290	4.39	19.46	0.55
CL	RT	Y	81,301	154.19	289.47	37	86.04	51.89	1	2,987	2,986	3.76	19.23	1.02
CL	RT	N	2,264,303	141.37	293.24	25	71.43	35.58	1	3,822	3,821	4.16	23.08	0.19
LI	FB	Y	411	34.83	77.32	10	19.18	11.86	1	940	939	6.19	53.40	3.81
LI	FB	N	555,655	252.64	1,272.49	28	68.46	38.55	1	21,934	21,933	11.34	140.19	1.71
LI	TW	Y	54,837	1,608.95	4,517.96	61	221.60	83.03	1	17,345	17,344	2.97	7.05	19.29
LI	TW	N	1,695,265	965.41	3,643.01	88	293.47	124.54	1	49,278	49,277	9.20	100.01	2.80
LI	RT	Y	74,192	791.07	1,671.53	171	400.81	241.66	1	49,278	49,277	5.13	47.94	6.14
LI	RT	N	2,090,245	745.97	1,781.68	115	331.64	161.60	1	49,278	49,277	5.53	49.19	1.23

Table A12-5 – US2012 geoparsing T-test results (x=NOTGEO, y=ISGEO) at interaction and user levels

nlp	lev	src	t	t > ± 2	df	p.value	p	p < .05	conf.int (x)	conf.int (y)	mean (x)	mean (y)
GT	I	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
GT	I	TW	-32.54	✓	25,862	1.1598E-227	p < .001	✓	-0.20	-0.18	1.22	1.41
GT	I	RT	0.44	✗	1,122	0.65959844	p > .05	✗	-0.02	0.04	1.21	1.20
AL	I	TW	-1.04	✗	3,008	0.296346447	p > .05	✗	-0.03	0.01	1.16	1.17
CL	I	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CL	I	TW	3.92	✓	24,686	8.735E-05	p < .001	✓	0.01	0.02	1.19	1.17
CL	I	RT	2.63	✓	915	0.008667996	p < .05	✓	0.01	0.07	1.18	1.14
LI	I	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LI	I	TW	1.29	✗	7,507	0.196869928	p > .05	✗	-0.05	0.22	5.66	5.58
LI	I	RT	2.21	✓	397	0.02795643	p < .05	✓	0.05	0.91	5.48	4.99
GT	U	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
GT	U	TW	-34.48	✓	86,934	7.7341E-259	p < .001	✓	-1.99	-1.78	5.44	7.33
GT	U	RT	14.01	✓	4,662	1.08119E-43	p < .001	✓	1.20	1.59	5.41	4.02
AL	U	TW	-20.24	✓	13,845	9.39088E-90	p < .001	✓	-0.44	-0.36	1.18	1.58
CL	U	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CL	U	TW	37.18	✓	258,221	9.4758E-302	p < .001	✓	1.46	1.62	5.77	4.23
CL	U	RT	1.25	✗	2,638	0.212312129	p > .05	✗	-0.15	0.66	4.87	4.61
LI	U	FB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LI	U	TW	2.92	✓	46,936	0.003482679	p < .05	✓	0.99	5.02	64.15	61.15
LI	U	RT	0.25	✗	1,798	0.800902614	p > .05	✗	-3.20	4.14	44.71	44.23

Table A12-6 – SCOT2014 geoparsing T-test results (x=NOTGEO, y=ISGEO) at interaction and user levels

nlp	lev	src	t	t > ± 2	df	p.value	p	p < .05	conf.int (x)	conf.int (y)	mean (x)	mean (y)
GT	I	FB	1.03	✗	481	0.301849547	p > .05	✗	-0.74	2.40	5.66	4.83
GT	I	TW	3.37	✓	24,336	0.000756031	p < .001	✓	0.01	0.02	1.31	1.29
GT	I	RT	7.01	✓	28,661	2.43692E-12	p < .001	✓	0.02	0.03	1.28	1.25
AL	I	TW	-3.55	✓	32,935	0.000392362	p < .001	✓	-0.03	-0.01	1.19	1.20
CL	I	FB	4.91	✓	509	1.20166E-06	p < .001	✓	0.56	1.31	3.98	3.05
CL	I	TW	2.99	✓	19,289	0.002775675	p < .05	✓	0.00	0.02	1.25	1.24
CL	I	RT	4.68	✓	22,841	2.91914E-06	p < .001	✓	0.01	0.02	1.22	1.21
LI	I	FB	1.10	✗	143	0.272798881	p > .05	✗	-0.54	1.89	5.43	4.76
LI	I	TW	-29.85	✓	10,969	2.7156E-188	p < .001	✓	-2.01	-1.77	6.76	8.65
LI	I	RT	3.44	✓	10,234	0.000591145	p < .001	✓	0.07	0.26	6.49	6.32
GT	U	FB	83.69	✓	1,662	0	p < .001	✓	349.94	366.73	389.52	31.18
GT	U	TW	69.32	✓	98,020	0	p < .001	✓	172.62	182.67	374.57	196.93
GT	U	RT	-13.52	✓	88,485	1.24271E-41	p < .001	✓	-21.60	-16.13	184.38	203.24
AL	U	TW	-7.17	✓	11,807	8.05429E-13	p < .001	✓	-0.82	-0.47	1.50	2.14
CL	U	FB	97.36	✓	2,066	0	p < .001	✓	293.95	306.03	325.24	25.25
CL	U	TW	47.42	✓	79,707	0	p < .001	✓	91.65	99.55	256.28	160.68
CL	U	RT	-12.41	✓	87,397	2.49296E-35	p < .001	✓	-14.85	-10.80	141.37	154.19
LI	U	FB	52.12	✓	591	3.8717E-223	p < .001	✓	209.59	226.01	252.64	34.83
LI	U	TW	-33.01	✓	57,166	9.8355E-237	p < .001	✓	-681.75	-605.33	965.41	1,608.95
LI	U	RT	-7.20	✓	80,291	5.87117E-13	p < .001	✓	-57.36	-32.83	745.97	791.07

Descriptive statistics for geoparsing operations (Section 5.2, p188) against US2012 OSN message text and linked/shared URL content are given in Table A12-1 and Table A12-2. The statistics show that:

- At interaction level (Table A12-1, p504):
 - The two geoparsers (GT=GATEcloud TwitIE, CL=CLAVIN-rest) run against all three sources of OSN message text (FB=Facebook, TW=Twitter tweets, RT=Twitter retweets) and AlchemyAPI (AL), run against samples of coordinate-geotagged (GEO=Y) and non-coordinate-geotagged (GEO=N) TW message text, all report similar median numbers of 1 detected toponymic mention/interaction. There is, however, considerable variability in numbers of NLP-detected toponymic mentions with a maximum of 161 mentions/interaction detected by GT in non-coordinate-geotagged FB message text, corresponding well with a similar maximum number of 159 geo-entities detected in the same source by CL. This FB post (UUID= D827A23004B04B26AC44893AE63627C6) starts ‘What's on your mind? What's on your mind?’ before mentioning the Libyan city of Benghazi 161 times (cf. Figure 5-9, p216). Skewness in toponymic mentions/interaction is low in TW and RT message text, whether coordinate-geotagged or not, and higher in FB text.
 - AlchemyAPI, run against linked/shared URL content (LI), detects a median number of 4 toponymic mentions/URL at interaction level, rising to a maximum of 54 mentions/URL for a link shared in a non-coordinate-geotagged TW interaction.
 - Three of the permutations tested at interaction level for both events are not present in the US2012 data set; there are no coordinate-geotagged Facebook records in the US2012 Streams (Table 4-8, p170) and, hence, R is unable to calculate descriptive statistics for GT/FB, CL/FB and LI/FB where GEO=Y. As there are no records

for these three permutations two-sided T-tests comparing the results of geoparsing operations for non-coordinate and coordinate-geotagged FB message text cannot be computed (below).

- At user level (Table A12-2, p505):
 - Median numbers of NLP-detectable toponymic mentions/user for GT, AL and CL geoparsers are in the range 1-2, with a maximum number of 306 detectable toponymic mentions for one user making non-coordinate-geotagged TW interactions. This maximum is lower than the maximums for SCOT2014 data at user level (below) where the data set contains many more records (n=6,477,713 vs. 1,718,667) and correspondingly larger numbers of interactions/user.
 - AlchemyAPI, run against linked/shared URL content (LI), detects variable mean numbers of toponymic mentions for all of the URLs shared by users, with higher means ranging 14-20/user for TW and RT data and a lower mean of 5 for non-coordinate-geotagged FB data, of which there are comparatively few records (n=57,265; 3.33%) in the US2012 data set.
 - The lack of coordinate-geotagged Facebook data in the US2012 data set is again highlighted by three permutations at user level (GT-CL-LI/FB where GEO=Y) with a 0 record count. As noted above, 0-record counts for one side of a two-sided T-test will not compute in R, although distributions clearly differ.

Descriptive statistics for geoparsing operations (Section 5.2, p188) against SCOT2014 OSN message text and linked/shared URL content are given in Table A12-3 and Table A12-4. The statistics show that:

- At interaction level (Table A12-3, p506):
 - Median numbers of toponymic entities detected/interaction in message text range from 1-2 depending upon NLP/geoparser and

-
- OSN source. Maximum numbers are higher than for the US2012 data set and 706 geo-entities have been detected in one non-coordinate-geotagged FB post. This Facebook post (UUID=5ABAB239EC7B4B5CB42EFC0666087EDD) contains a long list of parliamentarian's names and email addresses, many ending @parliament.uk, coded as a country ('UK') by TwitIE on GATEcloud.
- AlchemyAPI, run against linked/shared URL content (LI), detects higher median numbers of toponymic mentions ranging from 2-7 with a maximum of 374 geo-entities detected in 40 linked URLs shared from one non-coordinate-geotagged FB post (UUID=731E8FA2DDC74170AEB08BA6ACFEEB04). These URLs, from a range of news or comment sites including bbc.com, bloomberg.com, informationclearinghouse.com and reuters.com amongst others, contain many toponymic mentions in HTML content.
 - All of the permutations of geoparser and OSN source, at both levels, whether coordinate-geotagged or not, have non-0 record counts allowing for calculation of paired T-tests for all like-for-like NLP/GEO=N/GEO=Y permutations using R (below).
 - At user level (Table A12-4, p507):
 - Median numbers of toponymic entities detected in message text by GT, AL and CL range from 1-48/user, reflecting larger numbers of interactions/user (albeit with significant skewness, Section 4.6.1, p164) recorded in the SCOT2014 data set, sampled using one long-running Stream (Appendix A7.3, p435).
 - In linked/shared content median numbers of detected geo-entities are higher, with maximums significantly higher, reflecting larger numbers of links shared by repeatedly observed users in this data set (Table 5-6, p207) which, as noted above, was recorded over a much longer time-period than the 1:5 or 1:50 sampled US2012 data set,
-

using just one DataSift Stream, that was bound to record more linked/shared URLs per user.

Welch Two Sample T-tests, calculated in R (Spector, 2018), compare distributions of numbers of NLP-detected toponymic mentions per interaction, or per user, for non-coordinate-geotagged (x variable) and coordinate-geotagged (y variable) interactions or users. These are shown above in Table A12-5 (US2012) and Table A12-6 (SCOT2014).

In the US2012 event (Table A12-5, p508):

- In 9 out of 20 cases like-for-like comparisons of geoparser, OSN source and level for non-coordinate-geotagged message text or linked/shared URL content against coordinate-geotagged corollaries are statistically significant with >95% confidence. The null hypothesis, that there is no difference in the distribution of numbers of toponymic mentions detected in OSN message text or linked/shared URL content by OSN source for the given NLP/geoparsers for GEO=N and GEO=Y at interaction and user levels, can be rejected. In most (n=6) of these statistically significant comparisons non-coordinate-geotagged interactions or non-coordinate-geotagging users, mean(x) in Table A12-5, make more NLP/geoparser-detectable toponymic mentions in message text, or link to and share URLs having more detectable toponymic mentions in content, than their coordinate-geotagged or geotagging, mean(y), corollaries.
- Statistics for 6 cases (FB message text parsed by GT and CL and linked/shared URL content parsed by AlchemyAPI LI at interaction and user levels) cannot be calculated in R for the US2012 data set as no coordinate-geotagged Facebook interactions are present. While this prevents execution of a two-sided T-test of GEO=N against GEO=Y, like-for-like distributions clearly differ (being totally absent on one side in each case) and are

significant in their own right, probably arising from DataSift's changing access to Facebook data over time (below).

- In the remaining 5 cases statistical significance is not demonstrated. It is not possible to reject the null hypothesis for coordinate-geotagged/non-coordinate-geotagged data NLP/geoparsed by GT in RT message text or byAlchemyAPI in TW message text or TW links at interaction level, or by CL in RT message text or by AlchemyAPI LI in RT linked/shared URL content at user level. Similar patterns of insignificance are not present in the SCOT2014 comparisons (below) suggesting that the sampling strategies used in US2012 (a mixture of one exclusively coordinate-geotagged Stream in 1:5 and two agnostically sampled Streams in 1:50 ratios) may have affected results. Further research (Section 7.5, p299) is required to repeat these tests against other OSN data sets recorded at different times, during similar electoral events, to reach decisive conclusions.

In the SCOT2014 event (Table A12-6, p509):

- Statistical significance with >95% confidence is found in 18 of 20 like-for-like cases comparing numbers of toponymic detections by NLP/geoparser in message text and linked/shared URLs for FB, TW and RT data at interaction and user levels. In over half (n=11) of these statistically significant comparisons non-coordinate-geotagged interactions or non-coordinate-geotagging users, mean(x) in Table A12-6, make more NLP/geoparser-detectable toponymic mentions in message text, or link to and share URLs having more detectable toponymic mentions in content, than their coordinate-geotagged or coordinate-geotagging, mean(y), corollaries.
- In 2 cases, FB message text NLP/geoparsed by GT and linked/shared FB URL content parsed by AlchemyAPI LI, no statistical significance is found at interaction level. This is likely to reflect the disparity in numbers of non-

coordinate (n=784,006) and coordinate-geotagged (n=1,231) Facebook posts recorded in the SCOT2014 data set (Table 4-8, p170).

Overall, comparison of distributions of toponymic detections in the message text and linked/shared URL content of coordinate-geotagged and non-coordinate geotagged interactions, or for coordinate-geotagging and non-coordinate-geotagging users, are statistically significant ($t > \pm 2$) with >95% confidence in 27 out of 40 cases. Statistics for 6 other cases could not be calculated owing to a lack of coordinate-geotagged Facebook posts in the US2012 data set which is itself significant, probably reflecting DataSift's access to Facebook-sourced OSN interactions over time. In the SCOT2014 data set, sampled by one consistent and long-running 1:1 DataSift Stream, most like-for-like comparisons are statistically significant.