

Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

5-2023

Examining Political Discourse on Online 8Kun and Reddit Forums

Braden Mindrum

Utah State University

Follow this and additional works at: <https://digitalcommons.usu.edu/etd>



Part of the [Statistics and Probability Commons](#)

Recommended Citation

Mindrum, Braden, "Examining Political Discourse on Online 8Kun and Reddit Forums" (2023). *All Graduate Theses and Dissertations*. 8733.

<https://digitalcommons.usu.edu/etd/8733>

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.



EXAMINING POLITICAL DISCOURSE ON ONLINE 8KUN AND REDDIT FORUMS

by

Braden Mindrum

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Jürgen Symanzik, Ph.D.
Major Professor

Brennan Bean, Ph.D.
Committee Member

Jeannie Johnson, Ph.D.
Committee Member

D. Richard Cutler, Ph.D.
Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

Copyright © Braden Mindrum 2023

All Rights Reserved

ABSTRACT

Examining Political Discourse on Online 8kun and Reddit Forums

by

Braden Mindrum, Master of Science

Utah State University, 2023

Major Professor: Jürgen Symanzik, Ph.D.

Department: Mathematics and Statistics

A recent example of political violence in the United States was that of the January 6, 2021, Capitol attack in connection with the certification of Joseph R. Biden's victory over Donald J. Trump in the 2020 US presidential election. This thesis analyzes online forum discourse surrounding the events of January 6, 2021. By utilizing various Python packages, this thesis presents a workflow that acquired data from 13 political and apolitical 8kun and Reddit forums. In total, over 5 million posts are acquired that span the three months preceding and following the events at the US Capitol on January 6, 2021. These data are then analyzed with R text mining and clustering packages.

Various text vectorization schemes, distance calculations, linkage algorithms, and visualizations identify five main clusters of forum discourse corresponding to three apolitical clusters and two political clusters. These methods are also used to identify eight dates with online discourse that consistently clustered with the discourse of January 6, 2021. The discourse on these dates are then analyzed for feelings of isolation and displacement from society. The analysis shows that there is no clear connection between the discourse surrounding January 6, 2021, and online feelings of isolation and displacement.

(219 pages)

PUBLIC ABSTRACT

Examining Political Discourse on Online 8kun and Reddit Forums

Braden Mindrum

A recent example of political violence in the United States was that of the January 6, 2021, Capitol attack in connection with the certification of Joseph R. Biden's victory over Donald J. Trump in the 2020 US presidential election. This thesis analyzes the events of January 6, 2021, through the lens of social media discourse. This thesis presents a workflow that acquired over 5 million 8kun and Reddit posts from various apolitical and political forums in the three months preceding and following the Capitol attack on January 6, 2021. Techniques from text analysis are then used to group forums according to the similarities of their posting patterns. Five main groups of forums are identified. Finally, this thesis analyzes these forums for feelings of isolation and displacement from society in connection with the events of January 6, 2021. Such feelings were not clearly identified. This thesis demonstrates the challenges and opportunities of scraping and analyzing social media data.

ACKNOWLEDGMENTS

I would like to express my gratitude for Utah State University and its Mathematics and Statistics Department. Financial support as well as an excellent faculty has made my time here well worth it. In particular, I would like to thank Gary Tanner for his continued support and help in navigating the graduate program. Likewise, Dr. Ian Anderson — who encouraged me to pursue a graduate degree — deserves my eternal gratitude. His mentorship has helped shape me into the person I am today.

I would also like to express gratitude to my committee. Dr. Jeannie Johnson and Dr. Brennan Bean helped shape my time and work at Utah State University into something great. I owe a special thanks to Dr. Jeannie Johnson for her contributions. Additionally, her mentorship when I was an undergraduate has had lasting impacts on my life. Likewise, Dr. Brennan Bean's coursework and contributions to this thesis were of great meaning to me. My advisor, Dr. Jürgen Symanzik, also provided immeasurable aid throughout this daunting process. Without his efforts, I would not be where I am. These three individuals are owed a debt of gratitude that I cannot possibly pay.

Most importantly, I would like to express my infinite gratitude to my friends, my family, and God who loved me always.

CONTENTS

	Page
ABSTRACT	iii
PUBLIC ABSTRACT	iv
ACKNOWLEDGMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	xv
1 Introduction	1
1.1 The Events of January 6, 2021	1
1.2 Thesis Motivation	2
1.3 Related Works	3
1.4 Online Social Media Forums	5
1.4.1 8kun and Reddit	5
1.4.2 Forum Structure	6
1.5 Thesis Overview	6
2 Data	8
2.1 Data Sources	8
2.2 Data Acquisition	9
2.2.1 8kun	9
2.2.2 Reddit	11
2.2.3 Reddit Data Acquisition Addendum	15
2.3 Data Preprocessing, Data Cleaning, and Data Storage	15
2.3.1 Data Preprocessing	15
2.3.2 Data Cleaning	16
2.3.3 Data Storage	18
2.4 The Final Data Sets	18
3 Methods	20
3.1 Software Products	20
3.1.1 Software Products and Packages for 8kun Data Acquisition	20
3.1.2 Software Products and Packages for Reddit Data Acquisition	21
3.1.3 Packages for Data Preprocessing, Data Cleaning, and Data Storage	21
3.1.4 Packages for Computation and Data Extraction	22
3.1.5 Packages for word2vec Analysis	22
3.1.6 Packages for Cluster Analysis	22
3.1.7 Packages for Visualizations	22
3.2 Data Acquisition	23
3.2.1 8kun	23

3.2.2	Reddit	26
3.3	Data Preprocessing, Data Cleaning, and Data Storage	28
3.4	Introduction of Mock-Documents	28
3.5	Important Statistics of Text Documents	31
3.6	Vectorizations of Text Documents	34
3.6.1	Token Frequency - Inverse Document Frequency	35
3.6.2	word2vec and doc2vec	40
3.7	Distance Measures of Text Documents	44
3.7.1	Choices of Distance Measures	44
3.7.2	Distance Measure / Vectorization Pairs	45
3.8	Clustering of Text Documents	48
3.8.1	Hierarchical Clustering	48
3.8.2	Choices of Hierarchical Clustering Algorithms	49
3.8.3	Quantities of Clusters	50
3.9	Visualization Methods	62
3.10	Alternative Methods	63
3.10.1	Software Products for Data Acquisition	63
3.10.2	Software Products for Data Analysis	64
3.10.3	Important Statistics of Text Documents	65
3.10.4	Vectorizations of Text Documents	66
3.10.5	Distance Measures of Text Documents	67
3.10.6	Clustering Algorithms	68
3.10.7	Visualization Methods	69
3.10.8	Qualitative Analysis	70
4	Results	71
4.1	Data Summary Statistics	71
4.2	The R <i>NbClust</i> Package Clustering Results	75
4.3	Heatmaps	87
4.3.1	Heatmaps With $\eta = 50, 100, 150, 200$ Tokens	87
4.3.2	Heatmaps With $\eta > 200$ Tokens	101
4.4	Individual Dates and Feelings of Displacement	105
4.4.1	Selecting Key Dates	105
4.4.2	Identifying and Quantifying Key Tokens	109
5	Discussion	112
5.1	Discussion of the R <i>NbClust</i> Results	112
5.1.1	Five Clusters of Forums — R <i>NbClust</i> Results	113
5.1.2	Refinement One of the Five Clusters — R <i>NbClust</i> Results	114
5.1.3	Refinement Two of the Five Clusters — R <i>NbClust</i> Results	116
5.2	Discussion of the Heatmap Results	117
5.2.1	Five Clusters of Forums — Heatmap Results	117
5.2.2	Refinement One of the Five Clusters — Heatmap Results	120
5.2.3	Refinement Two of the Five Clusters — Heatmap Results	121
5.3	Summary of Clusters	123
5.4	Key Token Analysis on Key Dates	124
5.4.1	Selecting Key Dates	125

5.4.2	Analysis of Key Tokens	125
5.5	Limitations	128
5.5.1	Hindsight Analysis, Anticipatory Intelligence, and the Limitations of Quantitative Research	128
5.5.2	Multi-Token Analysis	129
5.5.3	Token Conversions Before Data Analysis	129
5.5.4	Data Privacy	130
6	Conclusion	131
6.1	Summary of Results and Discussions	131
6.2	Data Acquisition	133
6.3	Future Work	133
APPENDICES		135
A	Reddit Data Acquisition Process	136
B	Regular Expression Conversions in Data Cleaning	137
C	List of Stop Tokens	139
D	Terminology and Notation of this Thesis	140
E	Details of the word2vec Neural Network	143
F	Scaled Squared Euclidean Distance versus Cosine Distance	144
G	Criteria from the R <i>NbClust</i> Package (Charrad et al., 2022)	145
H	Data Summary Statistics for Alternative Forums	146
I	Analogous Results to Table 4.2 and Table 4.7	149
J	Analogous Results to Figure 4.3 – Figure 4.6	168
K	Analogous Results to Figure 4.7 – Figure 4.10	173
L	Analogous Results to Figure 4.11 – Figure 4.14 and Figure 4.15	178
M	The First 100 Tokens of Randomly Selected Forum Documents	181
References		186

LIST OF TABLES

Table	Page
2.1 Scraped variables for 8kun data, as obtained from the Internet Archive’s Wayback Machine (Internet Archive, ndc). The bolded variables were those that were used in analysis.	11
2.2 Scraped variables for Reddit data, as obtained from the Pushshift data set (Baumgartner et al., 2020a) and Reddit’s application programming interface (API) (Reddit, 2022b). The bolded variables were those that were used in analysis.	12
2.3 A demonstration of the data cleaning process described in Section 2.3.2.	18
2.4 The final variables associated with the 8kun and Reddit data. The bolded variables were those that were used in analysis.	19
3.1 Mock-documents (created by the author) to exemplify the methods in Section 3.5 – Section 3.8. The notation $d_{i,j}$ indicates the j^{th} mock-document from the i^{th} mock-forum, f_i . More details are provided in the main text.	30
3.2 As in Procedure 3.6.1.b: the selection of the tokens for the tf-idf vectorizations of the monthly aggregated mock-documents. For each document $\bar{d}_{f,m}^p$ and token t , the tf-idf value is shown and arranged in descending order. Highlighted rows correspond to the top tf-idf token(s) for each document. The final tokens are bolded (repeat tokens are not bolded twice). More details are provided in the main text. The original mock-documents are displayed in Table 3.1. The monthly aggregation of these mock-documents are shown in the example of Procedure 3.6.1.a.	39
3.3 The tf-idf vectorizations of the monthly aggregated mock-documents produced by Procedure 3.6.1.b, with $\hat{\eta} = 21$. Tokens are arranged in alphabetical order. Missing values correspond to a 0 tf-idf value. The original mock-documents are displayed in Table 3.1. The monthly aggregation of these mock-documents are shown in the example of Procedure 3.6.1.a.	40
3.4 The doc2vec vectorizations of the mock-documents in Table 3.1 produced by Procedure 3.6.2. The doc2vec vectorizations are not interpreted. These vectorizations optimally minimized the objective function prescribed by the word2vec machine learning software (see Appendix E for details).	43
3.5 The Jaccard distances between the monthly aggregated mock-document tf-idf vectorizations shown in Table 3.3. Distances range from 0 to 1.	46

3.6	The Euclidean distances between the mock-document doc2vec vectorizations shown in Table 3.4.	46
3.7	The optimal document clustering for each criteria of the R <i>NbClust</i> package (criteria are arranged as in the R <i>NbClust</i> package). Each criteria could identify 1 – 8 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. Independent of whether the Ward linkage or average linkage clustering algorithm was used, the optimal clustering results were the same, with two exceptions: the <i>ball</i> and <i>mcclain</i> criteria. The differences are marked in light blue for Ward linkage and light green for average linkage.	59
3.8	The optimal document clustering for each criteria of the R <i>NbClust</i> package (criteria are arranged as in the R <i>NbClust</i> package). Each criteria could identify 3 – 8 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. Independent of whether the Ward linkage or average linkage clustering algorithm was used, the optimal clustering results were the same, with one exception: the <i>sdindex</i> criteria. The differences are marked in light blue for Ward linkage and light green for average linkage.	61
4.1	For the six main forums of study, summary statistics of the total number of posts, submissions, and comments, as well as summary statistics for the number and percent of dates with no posts.	72
4.2	Identified clusters determined by the various criteria of the R <i>NbClust</i> package. Each column corresponds to one of three randomly created documents from each forum (seed 14741). Each row corresponds to a criteria (arranged as in the R <i>NbClust</i> package), and each criteria was allowed to identify 1 – 35 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. For example, the first row corresponds to the <i>kl</i> criteria. This criteria clustered all random AskThe.Donald, conservative, conspiracy, pnd, conservatives, democrats, and Liberal documents together; it clustered all random qresearch documents together; it clustered all random climate and climatechange documents together; it clustered all random immigration documents together; and it clustered all random math documents together. Most criteria determined the same clusters independent of whether Ward linkage or average linkage was used. The differences are marked in light blue for Ward linkage and light green for average linkage. Except for the <i>duda</i> and <i>pseudot2</i> criteria, the quantity of identified clusters remains the same regardless of whether Ward linkage or average linkage was used.	77
4.3	A visual representation of the Ward linkage cluster results in Table 4.2, which allowed each criteria to identify 1 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R <i>NbClust</i> criteria are color-coded. For example, the <i>dunn</i> criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.	78

4.4	As in Table 4.2 and Table 4.3, the number of criteria which identified a specific quantity of clusters (Ward linkage). Each criteria could identify 1 – 35 clusters. . . .	78
4.5	A visual representation of the average linkage cluster results in Table 4.2, which allowed each criteria to identify 1 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R <i>NbClust</i> criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.	79
4.6	As in Table 4.2 and Table 4.5, the number of criteria which identified a specific quantity of clusters (average linkage). Each criteria could identify 1 – 35 clusters. .	79
4.7	Identified clusters determined by the various criteria of the R <i>NbClust</i> package, with five being the minimum number of clusters. Each column corresponds to one of three randomly created documents from each forum (seed 14741). Each row corresponds to a criteria (arranged as in the R <i>NbClust</i> package), and each criteria was allowed to identify 5 – 35 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. For example, the first row corresponds to the kl criteria. This criteria clustered all random Ask-The_Donald, conservative, conspiracy, pnd, conservatives, democrats, and Liberal documents together; it clustered all random qresearch documents together; it clustered all random climate and climatechange documents together; it clustered all random immigration documents together; and it clustered all random math documents together. Most criteria determined the same clusters independent of whether Ward linkage or average linkage was used. The differences are marked in light blue for Ward linkage and light green for average linkage. Except for the duda and pseudot2 criteria, the quantity of identified clusters remains the same regardless of whether Ward linkage or average linkage was used.	83
4.8	A visual representation of the Ward linkage cluster results in Table 4.7, which allowed each criteria to identify 5 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R <i>NbClust</i> criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.	84
4.9	As in Table 4.7 and Table 4.8, the number of criteria which identified a specific quantity of clusters (Ward linkage). Each criteria could identify 5 – 35 clusters. . .	84

4.10	A visual representation of the average linkage cluster results in Table 4.7, which allowed each criteria to identify 5 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R <i>NbClust</i> criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.	85
4.11	As in Table 4.7 and Table 4.10, the number of criteria which identified a specific quantity of clusters (average linkage). Each criteria could identify 5 – 35 clusters. .	85
4.12	Document label color coding for heatmaps. Specifically, if a document came from a specific forum, then the label for that document was colored as below.	88
5.1	The seven R <i>NbClust</i> exceptions (by criteria, random seed, linkage algorithm, and range of identifiable clusters) that do not separate the pnd/conspiracy documents from the conservative/conservatives/AskThe_Donald/democrats/Liberal documents. As discussed in the main text, the gap and ball rows are minor exceptions and are separated from the duda and pseudot2 rows (less minor exceptions) by a horizontal line.	116
B.1	Replacements made during the data cleaning process (Section 2.3.2).	138
D.1	Terminology of text documents and their associated meanings. Terms are organized alphabetically for ease of navigation.	140
D.2	Notation of text documents. Subscripts may modify the meanings of notation beyond what is shown in this table. When this happens, the meaning will be clear. This table reflects the generic use.	141
D.3	Mathematical notation used throughout this thesis. Mathematical terms are quoted. .	142
H.1	For the seven alternative forums, summary statistics of the total number of posts, submissions, and comments, as well as summary statistics for the number and percent of dates with no posts.	146
I.1	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the <code>kl</code> criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	150
I.2	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the <code>ch</code> criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	151

I.3	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the hartigan criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	152
I.4	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the cindex criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	153
I.5	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the db criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	154
I.6	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the silhouette criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	155
I.7	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the duda criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	156
I.8	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the pseudot2 criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	157
I.9	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the ball criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	158
I.10	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the ptbiserial criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	159
I.11	For each random seed, linkage algorithm, and minimum number of identifiable clusters (<code>min.nc</code> , in the notation of the R <i>NbClust</i> package), the identified clusters by the gap criteria of the R <i>NbClust</i> package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.	160

- I.12 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `mcclain` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 161
- I.13 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `gamma` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 162
- I.14 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `gplus` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 163
- I.15 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `tau` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 164
- I.16 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `dunn` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 165
- I.17 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `sdindex` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 166
- I.18 For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `sdbw` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria. 167

LIST OF FIGURES

Figure	Page
2.1 The size of the cumulative sum of submission data (in $\log_{10}(\text{bytes})$) available from July 2005 – July 2022 through the Pushshift data set (Baumgartner et al., 2022). The red line marks April 2019.	13
3.1 Mock HTML code (created by author) to exemplify the Python <i>Scrapy</i> package. Numbering is not part of the HTML code but is for reference only.	24
3.2 The Ward linkage clustering results for the mock-documents of Table 3.1. The tf-idf vectorizations and pairwise distances are contained in Table 3.3 and Table 3.5, respectively. The “ i Month” leaves correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a.	53
3.3 The average linkage clustering results for the mock-documents of Table 3.1. The tf-idf vectorizations and pairwise distances are contained in Table 3.3 and Table 3.5, respectively. The “ i Month” leaves correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a.	54
3.4 The Ward linkage clustering results for the mock-documents of Table 3.1. The doc2vec vectorizations and pairwise distances of the documents are contained in Table 3.4 and Table 3.6, respectively. The “ i, j ” leaves correspond to the mock-documents $\bar{d}_{i,j}$	54
3.5 The average linkage clustering results for the mock-documents of Table 3.1. The doc2vec vectorizations and pairwise distances of the documents are contained in Table 3.4 and Table 3.6, respectively. The “ i, j ” leaves correspond to the mock-documents $\bar{d}_{i,j}$	55
3.6 A monthly aggregated mock-document heatmap (Ward linkage) with the vectorizations and pairwise distances displayed in Table 3.3 and Table 3.5, respectively. The tokens on the right indicate the tokens of the tf-idf vectorizations, while the “ i Month” leaves at the bottom correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a. Dendrograms for both the tokens (on the left) and the mock-documents (at the top) are also displayed.	56
3.7 A monthly aggregated mock-document heatmap (average linkage) with the vectorizations and pairwise distances displayed in Table 3.3 and Table 3.5, respectively. The tokens on the right indicate the tokens of the tf-idf vectorizations, while the “ i Month” leaves at the bottom correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a. Dendrograms for both the tokens (on the left) and the mock-documents (at the top) are also displayed.	57

3.8	Counts of the optimal number of clusters, as determined by the various criteria of the R <i>NbClust</i> package, when each criteria could identify 1 – 8 clusters. The counts are the same independent of whether the Ward linkage or average linkage clustering algorithm was used.	60
3.9	Counts of the optimal number of clusters, as determined by the various criteria of the R <i>NbClust</i> package, when each criteria could identify 3 – 8 clusters. Histograms (a) and (b) correspond to the Ward linkage and average linkage clustering algorithms, respectively.	62
4.1	Posts, submissions, and comments per day for each of the six main forums of study (log 10 scale). The black and blue lines (corresponding to posts and comments, respectively) nearly overlap.	73
4.2	Posts, submissions, and comments per month for each of the six main forums of study (log 10 scale). The black and blue marks (corresponding to posts and comments, respectively) nearly overlap.	74
4.3	A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	89
4.4	A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	90
4.5	A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	91
4.6	A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	92
4.7	A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.	96
4.8	A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.	97

4.9	A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.	98
4.10	A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.	99
4.11	A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 300$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.	101
4.12	A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 500$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.	102
4.13	A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 750$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.	102
4.14	A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. All $\eta = 821$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.	103
4.15	A Ward linkage dendrogram of the devoted political Reddit forums, split across forums and months. All $\eta = 281$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.	103
4.16	For conspiracy, pnd, and qresearch as well as each R <i>NbClust</i> criteria, the points mark the vectorized dates which clustered with January 6, 2021. Each criteria could identify any number of clusters between one and the number of dates minus one. More details are provided in the main text.	108
4.17	For each forum, the grey line indicates the proportion of posts for each date that used at least one of the key tokens: <i>attack</i> , <i>cultural</i> , <i>culture</i> , <i>diversity</i> , <i>enemies</i> , <i>ethnic</i> , <i>invaders</i> , <i>racial</i> , and <i>replacement</i> . Nine key dates — 2020-10-14, 2020-10-24, 2020-11-03, 2021-01-06, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19 — are marked in blue. The orange line indicates the proportion of posts for each date that used the token <i>election</i> . A rescaled version of the conspiracy graphic is shown inside the conspiracy graphic obeying the small multiples principle.	111

A.1	The process by which Reddit data were acquired for each forum on each date. (1) The Python <i>PSAW</i> package was used to interact with the Pushshift data set and Pushshift API. Specifically the Python <i>PSAW</i> package subsets the Pushshift dataset and selects a unique identifier to each submission from the specified forum on the specified date. (2) The Python <i>PRAW</i> package was then used to locate each submission from the previous step on Reddit's API. (3) The Python <i>PRAW</i> package was then used to access the data associated with that submission. This includes the comment data associated with the submission.	136
H.1	Posts, submissions, and comments per day for each of the seven alternative forums (log 10 scale). The black and blue lines (corresponding to posts and comments, respectively) nearly overlap.	147
H.2	Posts, submissions, and comments per month for each of the seven alternative forums (log 10 scale). The black and blue marks (corresponding to posts and comments, respectively) nearly overlap.	148
J.1	An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	169
J.2	An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	170
J.3	An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	171
J.4	An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	172
K.1	An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	174
K.2	An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	175
K.3	An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	176
K.4	An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.	177

- L.1 An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 300$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text. . 178
- L.2 An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 500$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text. . 179
- L.3 An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 750$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text. . 179
- L.4 An average linkage dendrogram of all (main and alternative) forums, split across forums and months. All $\eta = 821$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text. 180
- L.5 An average linkage dendrogram of the devoted political Reddit forums, split across forums and months. All $\eta = 281$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text. 180

CHAPTER 1

Introduction

On January 6, 2021, the world witnessed a violent and unprecedented attack on the U.S. Capitol, the Vice President, Members of Congress, and the democratic process. Rioters, attempting to disrupt the Joint Session of Congress, broke into the Capitol building, vandalized and stole property, and ransacked offices. They attacked members of law enforcement and threatened the safety and lives of our nation’s elected leaders. Tragically, seven individuals, including three law enforcement officers, ultimately lost their lives (US Senate, 2021).

This thesis analyzes the events of January 6, 2021, through one lens: that of online forum discourse. Specifically, it analyzes (a) the extent to which the discourses on different online forums were distinguishable, and (b) the extent to which users of online forums exhibited feelings of isolation and displacement from society. In this introductory chapter, Section 1.1 summarizes the events of January 6, 2021. Section 1.2 describes the motivation of the thesis, and Section 1.3 provides an overview of related works. Section 1.4 introduces 8kun and Reddit — the forum websites this thesis studies — as well as the concept of online forums in general. Finally, Section 1.5 gives an outline of the rest of the thesis.

1.1 The Events of January 6, 2021

Landry et al. (2021) noted that the 2020 United States (US) presidential election was “particularly divisive.” Moreover, the 2020 presidential election “saw an unprecedented number of false claims alleging election fraud and arguing that Donald Trump was the actual winner of the election” (Pennycook and Rand, 2021). Such a divisive and counter-factual election climaxed on January 6, 2021, when rioters broke into the US Capitol building and disrupted (temporarily) the US Congress from certifying Joseph R. Biden (Biden)’s victory over then-President Donald J. Trump (Trump).

The following details were taken from the US Senate’s report on the events of January 6, 2021 ([US Senate, 2021](#)).

At 10:30am, on the morning of January 6, 2021, the United States Capitol Police reported that somewhere between 25,000 and 30,000 protestors were gathered at the Ellipse in Washington DC to hear then-President Trump speak at his *Save America* rally. Just before noon, he began his remarks in which he “continued his claims of election fraud and encouraged his supporters to go to the Capitol.” Thirty minutes before he finished his remarks around 1:15pm, demonstrators began making their way to the Capitol building.

The initial breach of United States Capitol Police barriers occurred at 12:53pm. By 2:11pm, rioters breached the Capitol building, delaying the certification of Biden’s victory. It wasn’t until 8:00pm that the Capitol building was deemed safe and the certification process continued. At 3:42am, January 7, 2021, Biden was the certified victor of the 2020 US presidential election. More details about the events of January 6, 2021, can be found in [US Senate \(2021\)](#). The events of January 6, 2021, serve as the background for this thesis.

1.2 Thesis Motivation

The events of January 6, 2021, continue to be investigated ([United States Department of Justice, 2021](#)). Undoubtedly, there is much to be studied. This thesis focuses upon one aspect: online forum discourse. In particular, this thesis seeks to understand:

- The extent to which the general discourse on different online political forums were distinguishable from one another in the three months preceding and following the events of January 6, 2021.
- The extent to which the University of Chicago’s *Chicago Project on Security & Threats* (CPOST)’s character profiling of the January 6, 2021, protestors manifested itself on online political forums. The body of work by the CPOST is extensive (and will be discussed in greater detail in Section [1.3](#)), but this thesis focuses on one key aspect of their profiling: that the January 6, 2021, protestors felt isolated and displaced from society.

Given the rise of political violence — including the Capitol attack on January 6, 2021, as well as violent attacks connected with online forums (see [Baele et al. \(2020\)](#)) — it has become increasingly important to understand how people discuss politics online.

1.3 Related Works

Much about the January 6, 2021, Capitol attack has been studied. For instance, [Pion-Berlin et al. \(2022\)](#) highlighted the similarities between then-President Trump’s actions and self-coups from history, and concluded: “[I]t is clear that the attack fits the definition [of a self-coup] closely.” Moreover: “What separates the successful self-coups from those that failed is the support of the armed forces. The US armed forces never came to the president’s defence in pursuing this perilous course of action, and ultimately intervened to stop it.” Similarly, [Bennett \(2022\)](#) reviewed three books that analyzed the connection between the 9/11 attacks* and the January 6, 2021, Capitol attack: “Put simply, together these works show how transformations initiated in the wake of 9/11 have unwittingly undermined public trust in key institutions while militarizing American society and normalizing warfare as a political solution.” From a technology standpoint, [Van Dijke and Wright \(2022\)](#) studied January 6, 2021, using mobile device data. They stated: “We find evidence that partisanship, socio-political isolation, proximity to chapters of the Proud Boys organization, and the local activity on Parler are robustly associated with protest participation. Evidence from communities with close election outcomes in 2020 suggests narrow losses by Trump triggered a spike in participation.” Many other studies can be found. [Setty \(2021\)](#) studied January 6, 2021, from the lens of criminal prosecution of domestic terrorism. [Dinulescu \(2021\)](#) studied January 6, 2021, by looking at the intersection of religion and politics in the US. [Reid and Craig \(2021\)](#) studied January 6, 2021, in the context of race relations in the US.

Likewise, much has been studied about political discourse online. For instance, [Formica \(2020\)](#), [Benkler et al. \(2018\)](#), and [Hussain et al. \(2018\)](#) all wrote about social media’s influence on the growth of domestic extremism and hate groups within the borders of the US. Other articles have

*On September 11, 2001, nineteen Al-Qaeda terrorists hijacked four commercial planes, two of which hit the Twin Towers at the World Trade Center in New York City, NY, one of which hit the Pentagon in Arlington, VA, and the third was crashed in Pennsylvania field following a passenger counter-attack while the plane was en route to Washington DC. See the 9/11 Commission Report for more details ([Kean and Hamilton, 2004](#)).

called into question the efficacy of social media in radicalizing individuals. For instance, [Hosseinmardi et al. \(2021\)](#) and [Ledwich and Zaitsev \(2020\)](#) both claimed that Youtube’s recommendation algorithm actually pushes people to mainstream and moderate sources. Still, the phenomenon is complicated and not fully understood — [Hosseinmardi et al. \(2021\)](#) stated: “Our results indicate that trends in video-based political news consumption are determined by a complicated combination of user preferences, platform features, and the supply-and-demand dynamics of the broader web.”

There is, however, less literature that studied online discourse in relation to the events of January 6, 2021. [Hitkul et al. \(2021\)](#) studied January 6, 2021, by comparing Twitter and Parler content (Twitter and Parler are both microblogging social media platforms, with Parler being known for its lack of content guidelines — see [Yardi et al. \(2009\)](#) and [Hitkul et al. \(2021\)](#)). They found that Twitter users mostly condemned the attack and then-President Trump, whereas “Parler was in support of undermining the veracity of the 2020 US Presidential Elections.” They also found a preponderance of violence and hate speech on Parler. [Ng et al. \(2021\)](#) likewise studied Parler activity in connection with the events of January 6, 2021. They found that “users that openly present a military or veteran affiliation, user [sic] that use the moniker ‘patriot’, and users that identify with QAnon-related terms” were the most influential. Finally, [Martin and Fournillier \(2022\)](#) analyzed Twitter discourse surrounding memes that came from the January 6, 2021, attack. They stated: “Twitter offered space to quickly communicate ideas that interrogate and/or reinforce racism, sexism, ableism, and classism. However, discourse reveals that this space is often mediated through humor that can undermine the complexity, gravity, and intersectional nuances of social events like the January 6th attempted coup.”

The Work by the University of Chicago’s Chicago Project on Security & Threats

As alluded to in Section 1.2, this thesis seeks to understand the extent to which online 8kun and Reddit users felt isolated and displaced from society. This motivation came from the CPOST’s research on American political violence ([CPOST, 2022](#)), which has been studying the January 6, 2021, Capitol attack. In particular, the CPOST researchers have sought to profile the protestors. The key findings from two of their reports ([Pape \(2022a\)](#) and [Pape \(2022b\)](#)) included:

- “[T]he insurrectionists are a cross-section of America. They closely reflect the US electorate on most socio-economic variables, with the vast majority employed in white- and blue-collar occupations, including many business owners. In short, they come from the mainstream, not just the fringe of society” (Pape, 2022a).
- “[C]ounties with higher rates of demographic change [...] sent more insurrectionists even when controlling for a host of competing factors” (Pape, 2022a).
- “The greater the decrease in [the] non-Hispanic white [population], the higher the rate of sending insurrectionists [to the Capitol on January 6, 2021]” Pape (2022b).
- “[B]elief in the “Great Replacement” — that Whites are losing rights to demographically more powerful minorities — is the key driver of violent support for Trump” (Pape, 2022b). For more information on the great replacement conspiracy theory, see Cosentino (2020), Obaidi et al. (2022), or Carlson and Harris (2022) who stated that the great replacement conspiracy theory “postulates [that] white European populations are being demographically and culturally replaced by non-white immigrants through policies enacted by ‘the global elites’.”

This body of work is mentioned because this thesis seeks to understand whether or not their findings (that Capitol protestors felt isolated and displaced from society) appeared on online forums.

1.4 Online Social Media Forums

This section introduces the online forum websites of study, as well as introduces their general structure. Online forums were the chosen data source because of their naturally focused content. That is, each online forum is tied to a central theme which allows for focused data sets. This thesis analyzed two forum websites in particular: 8kun and Reddit.

1.4.1 8kun and Reddit

8kun (formally known as 8chan) has “long been known for its far-right extremist, racist and violence-endorsing content, and its links to multiple mass shootings” (Zeng and Schäfer, 2021). Because of this history, 8kun has had a history of being removed from the publicly accessible

internet, and has been described as a “semi-dark web forum” (Zamani et al., 2019). As of January 2021, 8kun contained 416 public forums (8kun, 2021).

Reddit — “the front page of the internet” — on the other hand is one of the most visited websites worldwide (Tsou, 2016). Contrasting with 8kun, Reddit seems to be known for forums like Eyebleach (in the top 1% of reddit forums, as ranked by size (Reddit, 2022a)) where people post endearing pictures and videos of animals. As of January 2021, Reddit contained 100,000+ forums (Reddit, 2021).

1.4.2 Forum Structure

Forum style websites have main pages that can vary in structure (e.g., popular posts from the whole website, or content that a particular user selects to see), but a *forum* will always be the basic component of a forum style website. A forum is a subset of user-created content that is tied to some central theme. For example, a forum may be devoted to space or books, or (because it is the internet) just about anything one can imagine. Forums are user-created and user-maintained. It is on these forums that users actually post. As such, one does not post to 8kun/Reddit. Instead, one posts to a specific forum on 8kun/Reddit.

Forums are made up of *threads*. A thread starts with someone posting new content. In this thesis, this new content is called a *submission*. Users are then free to continue the thread by posting replies to this submission. In this thesis, these replies are called *comments* (though a comment may actually be question, critique, etc.). The general term *post* refers to either a submission or a comment. In this way, a thread consists of the original submission along with all of its comments.

1.5 Thesis Overview

Chapter 2 introduces the 8kun and Reddit forums chosen for analysis, and also discusses the motivations behind such choices. It also describes where and how data were acquired, as well as how data were preprocessed and cleaned before analysis. It concludes with a description of the final data sets for each forum.

Chapter 3 describes the methods used throughout this thesis. This includes an overview of the various software products used. It also includes fuller details on how data were acquired and

preprocessed. Likewise, it presents (and illustrates with small mock-examples) the methods used to analyze the data. Finally, it discusses various alternative methods that were not used in this thesis.

Chapter 4 describes the results of data analysis. Specifically, it presents results that demonstrate how the various 8kun and Reddit forums related to each other in terms of online discourse in the three months preceding and following January 6, 2021. Likewise, it also presents results that illustrate the extent of feelings of isolation and displacement that existed on the online forums in the three months preceding and following January 6, 2021.

Chapter 5 discusses the results of the prior chapter. It does so by providing connections and comparisons between the different results of this thesis, as well as providing connections and comparisons with related literature.

Finally, Chapter 6 concludes this thesis by summarizing the main results and points of discussion. It also provides opportunities for future research.

There are 13 appendices in this thesis, each of which show technical details that would otherwise disrupt the flow of the main text. Appendix A visually depicts the process for acquiring Reddit data. Appendix B shows the text conversions made in the data cleaning process. Appendix C shows a list of stop tokens (i.e., tokens that do not contain much meaning on their own). Appendix D displays the terms and notations used throughout this thesis. Appendix E provides details on how a particular machine learning software is trained. Appendix F proves that two different distance functions are proportional to one another under suitable conditions. Appendix G provides details about the various criteria used within a certain method of this thesis. Appendix H shows data summary statistics for alternative data sets. Appendix I displays the results from one of the methods used in this thesis. Appendix J, Appendix K, and Appendix L each show results that are analogous to results in the main text, but with a different implementation. Specifically, the main text shows the results when a certain clustering algorithm was used, and Appendix J, Appendix K, and Appendix L show the results when a different clustering algorithm was used. Finally, Appendix M shows — to give the reader an idea of typical forum posts — the first 100 tokens of a randomly created document for each forum. Derogatory (and otherwise offensive) forum tokens may appear in Appendix M and elsewhere throughout this thesis.

CHAPTER 2

Data

A non-trivial component of this thesis consisted of data acquisition. This chapter introduces the data, where they were acquired, and also explains how they were preprocessed and cleaned before analysis.

2.1 Data Sources

As discussed in Chapter 1, this thesis analyzes forum-style social medias. Three main forums of study were each taken from 8kun and Reddit, and an additional seven forums were taken from Reddit.

The main forums of study were 8kun’s pnd, newsplus, and qresearch forums, and Reddit’s AskThe.Donald, conservative, and conspiracy forums. In general, 8kun was chosen for its infamy, and Reddit for its popularity (as discussed in Section 1.4.1). The specific forums were chosen for the following reasons. The 8kun forums are among the most frequented on the website (as measured by 8kun’s total post numbers and Active ISP measure — see [8kun \(nda\)](#) and [8kun \(ndb\)](#), respectively) and therefore provided a larger data set. Additionally, pnd was chosen for its connection to violence — Brenton Tarrant, an active pnd user, posted: “I will carry out an attack against the invaders” as well as a link to his manifesto and a live stream of his attack to pnd on March 15, 2019, before attacking two mosques in Christchurch, New Zealand, killing 51 people and injuring 49 others ([Baele et al., 2020](#)). The newsplus forum was chosen as a more general political/news forum and qresearch was chosen to provide a glimpse into devoted conspiracy content. See [Baele et al. \(2021\)](#) for an article that analyzed “far-right” forum data from 8kun (and other forum websites). Likewise, AskThe.Donald was chosen for its explicit support of former-President Trump, conservative was chosen as a more general political/news forum, and conspiracy was chosen to provide a glimpse into devoted conspiracy content.

Alternative forums were also taken from Reddit. These include climate, climatechange, and

immigration which are tangentially political; conservatives, democrats, and Liberal which provide more political viewpoints; as well as the math forum for explicitly apolitical content. No alternative data sets were taken from 8kun. This was done because other 8kun forums had data insufficiency: already 8kun is not as well attended as Reddit, and the data source for 8kun exacerbated this problem (see Section 2.2.1). Moreover, other 8kun forums with substantive data quantity are inappropriate in that their content is pornographic.

Data from each forum were acquired through the time frame of October 1, 2020 – March 31, 2021. This time frame was chosen to look at the discourse near January 6, 2021 (roughly the middle of the time frame). For data summary statistics, see Section 4.1.

2.2 Data Acquisition

This section details how data were acquired. For both 8kun and Reddit, data acquisition erred on the side of excess. That is, some information was initially scraped without knowing whether or not it would ultimately be used in this thesis. This was done because the time commitment of the data acquisition processes were front-loaded. In other words, missing a variable — and then needing to re-scrape data to obtain it later — would take significantly more time than simply scraping extraneous information the first time.

2.2.1 8kun

The Internet Archive and the Wayback Machine

Put by the Internet Archive itself: “The Internet Archive, a 501(c)(3) non-profit, is building a digital library of Internet sites and other cultural artifacts in digital form” ([Internet Archive](#), *nda*). The web specific portion of this digital library is called the Internet Archive’s Wayback Machine (WBM). The 8kun data for this thesis were acquired through the WBM.

The decision to use the WBM, as opposed to 8kun itself, was made because (a) 8kun has a history of being removed from the publicly accessible internet ([Zamani et al., 2019](#)) (i.e., accessing data through the WBM is more stable), and (b) retrieval of dated information on 8kun’s website would have been difficult as there was no known way of easily filtering 8kun posts by date.

However, the Internet Archive does not archive everything: “Some sites may not be included because the automated crawlers were unaware of their existence at the time of the crawl” ([Internet Archive, ndb](#)). Because of this, data acquired through the WBM cannot contain everything. This gap is made even larger due to 8kun’s nature. Its fringe status implies that many contributors to the WBM may overlook 8kun in their crawlers because:

Every day hundreds of web crawls contribute to the web captures available via the Wayback Machine. Behind each, there is a story about factors like who, why, when and how ([Internet Archive, ndc](#)).

This is evident by the fact that 8kun was “[s]aved 4,143 times between November 16, 2019 and June 26, 2022” ([Internet Archive, 2022a](#)), for an average of roughly four saves per day, and Reddit was “[s]aved 12,647,337 times between July 18, 2002 and July 14, 2022” ([Internet Archive, 2022b](#)), for an average of roughly 1,700 saves per day.

Because of this limitation in data, every effort was made to access all data the WBM has regarding the forums of 8kun. Fortunately, the WBM is organized well, and the structure of 8kun (and its HyperText Markup Language (HTML) code) is easy to navigate.

The 8kun data, accessed through the WBM, were scraped and saved as comma-separated value (csv) files using the Python *Scrapy* package (see [Scrapy Developers \(2022a\)](#), as well as Section 3.1.1 and Section 3.2.1). For each forum, the Python *Scrapy* package saved six csv files which corresponded to six months of archival in the WBM (from October 2020 to March 2021).

Table 2.1 shows the variables of the scraped 8kun data. Note that the variables *submission_id* and *submission_no* in Table 2.1 were originally scraped as *post_id* and *post_no* (respectively). Likewise, the variables *comment_id* and *comment_no* in Table 2.1 were originally scraped as *reply_id* and *reply_no* (respectively). The changes are reflected on this table for clarity (as in Section 1.4.2, a submission is the initial post to a new thread, a comment is a continuation of a thread, and a post is a generic term for either a submission or a comment). Also note that 8kun’s anonymization process (see [Zamani et al. \(2019\)](#)) made it difficult to distinguish authors (i.e., a single user might have multiple different identifications in the *submission_id* and *comment_id* variables). Finally, note that not all of the scraped information was used in this thesis.

Table 2.1: Scraped variables for 8kun data, as obtained from the Internet Archive’s Wayback Machine ([Internet Archive](#), [ndc](#)). The bolded variables were those that were used in analysis.

<i>Variable</i>	Description
<i>date_time</i>	the time of post creation
<i>file_names</i>	file names attached to a post (if applicable)
<i>links</i>	links included with a post (if applicable)
<i>omitted_comments</i>	the number of unscraped comments in a thread (submissions only) — 8kun’s forums include the number of unseen comments on their main page; if the thread was not scraped by the WBM, then this entry will be non-zero with the number of missing comments
<i>submission_id</i>	a unique identifier of the author of a submission to which a post is a part of
<i>submission_no</i>	a unique identifier to the thread to which a post is a part of
<i>comment_id</i>	a unique identifier to the author of a post (will be the same as <i>submission_id</i> if the post is a submission)
<i>comment_no</i>	a unique identifier to a post (will be the same as <i>submission_no</i> if the post is a submission)
<i>text</i>	the text of a post
<i>title</i>	the title of a submission (submissions only)

2.2.2 Reddit

The Pushshift Data Set and Reddit’s Application Programming Interface

The Reddit data for this thesis were scraped from the Pushshift data set ([Baumgartner et al., 2020a](#)) and Reddit’s application programming interface (API) ([Reddit, 2022b](#)). In short, the Pushshift data set was used to acquire a unique identification string to each desired post. Then the Reddit API used this identification string to acquire the desired information (e.g., the text of the post, the time of posting — see Table 2.2).

Pushshift is a Reddit data set, which — as of April 2019, dating back to June 2005 — consisted of 500 million submissions and 5 billion comments. Since April 2019, the Pushshift data set has continually been updated as new content is posted. Figure 2.1 shows updated information extending to July 2022. Roughly, the size of the Pushshift data set has grown by a half order of magnitude since April 2019.

Almost certainly, (though not stated by the Pushshift authors) the Pushshift data set did not/does

Table 2.2: Scraped variables for Reddit data, as obtained from the Pushshift data set ([Baumgartner et al., 2020a](#)) and Reddit’s API ([Reddit, 2022b](#)). The bolded variables were those that were used in analysis.

<i>Variable</i>	Description
<i>author</i>	a unique identifier to the author of a post
<i>created_utc</i>	the time of post creation
<i>distinguished</i>	an identifier if the post was made by a distinguished user (e.g., a moderator)
<i>edited</i>	(unclear) a boolean indicating if the author of the post changed their original post
<i>id</i>	a unique identifier to a post
<i>is_original_content</i>	(unclear) a boolean indicating if a submission is original content
<i>is_self</i>	(unclear) unknown
<i>is_submitter</i>	a boolean indicating if a comment was made by the same author as the submission (comments only)
<i>link_flair_text</i>	(unclear) a string if the author was a flaired user (e.g., the author may be marked as a new user)
<i>link_id</i>	(unclear) unknown
<i>locked</i>	(unclear) a boolean indicating if a submission has been locked from future replies (submissions only)
<i>name</i>	(unclear) unknown
<i>num_comments</i>	the number of comments to a submission (submissions only)
<i>over_18</i>	a boolean indicating if the author of a submission has marked the thread as mature content (submissions only)
<i>parent_id</i>	a unique identifier to the post to which a comment was made (comments only)
<i>score</i>	the number of upvotes a post has
<i>stickied</i>	a boolean indicating if a comment will appear first in a thread (for comments) or first on a forum (for submissions)
<i>self_text</i>	the text of a post (for comments, this variable name was scraped as <i>body</i>)
<i>spoiler</i>	a boolean indicating if the author of a submission has marked the thread as containing spoilers
<i>title</i>	the title of a submission (submissions only)
<i>upvote_ratio</i>	the upvote ratio of a submission (submissions only)

not/will not contain everything posted to Reddit. It may also contain posts that were later deleted by users ([Proferes et al., 2021](#)). Nevertheless:

The Pushshift Reddit dataset has attracted a substantial research community. As of late

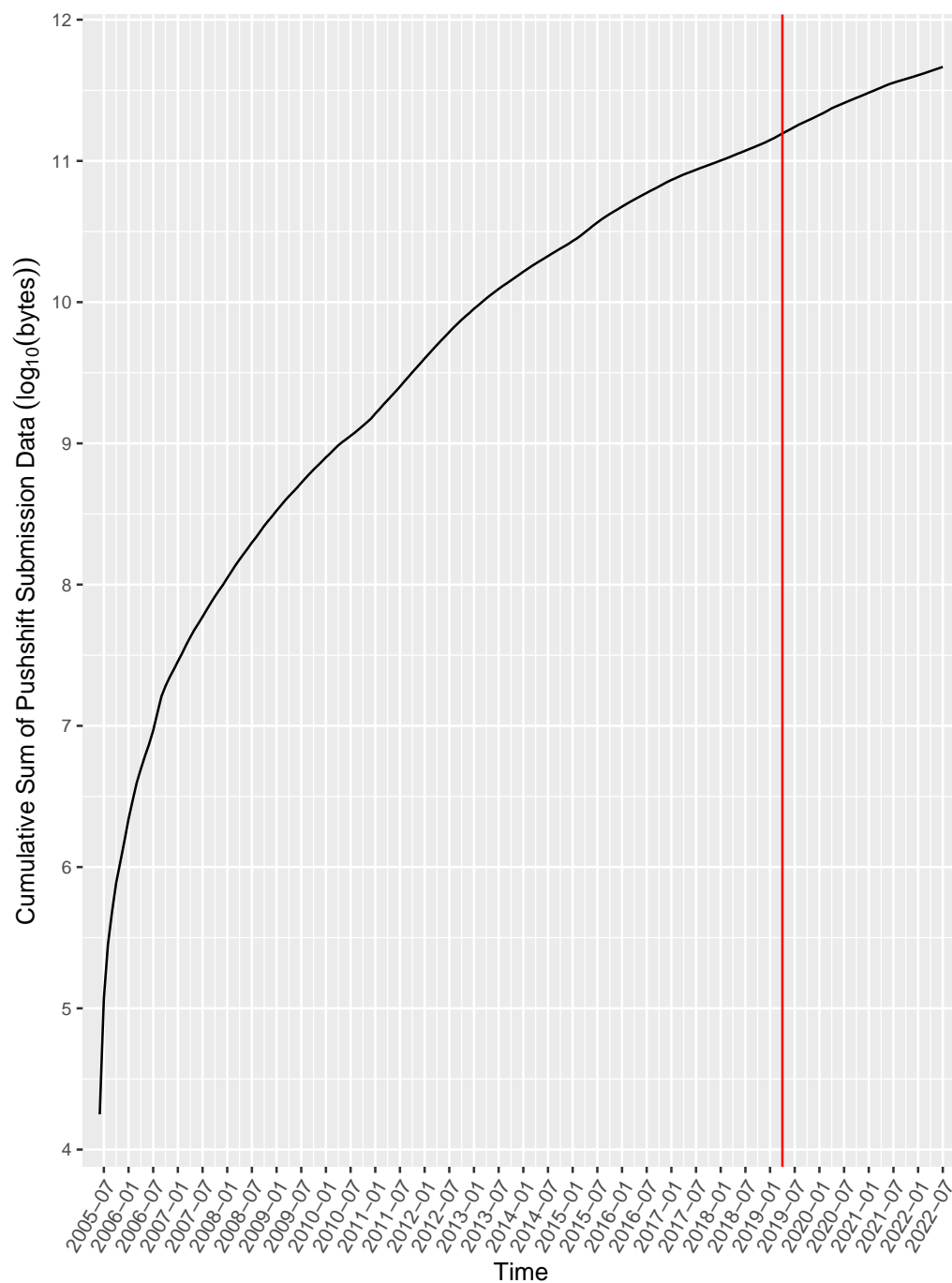


Fig. 2.1: The size of the cumulative sum of submission data (in $\log_{10}(\text{bytes})$) available from July 2005 – July 2022 through the Pushshift data set ([Baumgartner et al., 2022](#)). The red line marks April 2019.

2019, Google Scholar indexes over 100 peer-reviewed publications that used Pushshift data. (Baumgartner et al., 2020a)

For a comprehensive overview of the use of Reddit data and the Pushshift data set in the literature, see Proferes et al. (2021).

The authors of the Pushshift data set also provided an API to access focused portions of their data set (see Baumgartner et al. (2019)). This API (in conjunction with the Python *PSAW* package (Marx, 2018) — see Section 3.1.2 and Section 3.2.2) acquired a unique posting identification string for each post of interest.

From these unique post identification strings, Reddit’s API (in conjunction with the Python *PRAW* package (Boe, 2022) — see Section 3.1.2 and Section 3.2.2) acquired the relevant data and saved them as csv files. For each forum, the Python *PRAW* package saved roughly 180×2 csv files corresponding to roughly 180 days of submission and comment data.

Table 2.2 shows the variables of the scraped Reddit data. Note that not all of the variables displayed in Table 2.2 had clear meanings (as mentioned previously, data acquisition erred on the side of excess). Such variables are indicated in Table 2.2, and no unclear variables were used in the final analysis. Also note that (like the 8kun data) not all of the scraped information was used in this thesis.

Note on Pushshift Terminology

The terms ‘Pushshift data set’ and ‘Pushshift API’ (and other derivatives) are ambiguous. Baumgartner et al. (2020a) also wrote another article detailing a data set associated with the messaging platform, Telegram (Baumgartner et al., 2020b). The ‘Pushshift’ term is also used there. However, because this thesis does not deal with Telegram, the terms ‘Pushshift data set’ and ‘Pushshift API’ (and other derivatives) exclusively refer to Reddit (in this thesis).

Note on the Reddit Data Acquisition Process

The mode of Reddit data acquisition was somewhat onerous. First, the Pushshift data set and API (in conjunction with the Python *PSAW* package) were used to acquire a unique identification

string to each post of interest. Then Reddit’s API (in conjunction with the Python *PRAW* package) was used to acquire the data of Table 2.2. This process is visualized in Figure A.1 in Appendix A. While onerous, this choice was intentional.

As mentioned above, Pushshift has a “substantial research community” (Baumgartner et al., 2020a). Moreover: “Pushshift makes it much easier for researchers to query and retrieve historical Reddit data” (again, see Baumgartner et al. (2020a)). This made it a good option to acquire the dated information of this thesis. However, the Pushshift API is under active development (see Baumgartner et al. (2019)), and the Python *PSAW* package has not been updated since 2020 (see Marx (2018)).

The Python *PRAW* package, however, has been updated regularly (most recently, May 2022 — see Boe (2022)). This made the Python *PRAW* package a good option for retrieving the information associated with the unique identification string acquired by the Python *PSAW* package.

2.2.3 Reddit Data Acquisition Addendum

The Python *PSAW* package has since become obsolete. A similar package — the Python *PMAW* package (Podolak, 2022) — can be used in its place. The flow of Reddit data acquisition remains unchanged. Also note that, as of March 1, 2023, Reddit data pre-dating November 3, 2022, can no longer be acquired as outlined in this thesis.

2.3 Data Preprocessing, Data Cleaning, and Data Storage

2.3.1 Data Preprocessing

The Reddit submission/comment data and 8kun data described in Section 2.2 all contained coinciding information, but were formatted differently. For ease of analysis, data were preprocessed into the same format.

Data preprocessing began with reformatting Reddit submission/comment data to the same format. This included dropping decidedly irrelevant information (e.g., the unknown *is_self* variable in Table 2.2), adding derived variables (e.g., not just the time of posting, but month of posting), and reordering variables and changing their names into agreeance. For each forum, the 180×2 csv

files (corresponding to roughly 180 days of submission and comment data) were then merged into a single csv file with the information described in Section 2.4.

Preprocessing continued with the 8kun data and deleting repeated posts which were scraped multiple times by the WBM. Such repeated posts were identified with the *comment_no* variable, which contains a unique identifier to each post. Then, like the Reddit data, the process continued by reformatting these data into the same format: again, dropping decidedly irrelevant information, adding derived variables, and reordering variables and changing their names into agreeance. Finally, for each forum, the six csv files (corresponding to six months of archival in the WBM from October 2020 to March 2021) were merged into a single csv file with the information described in Section 2.4.

2.3.2 Data Cleaning

The scraped text data for this thesis were messy. For example, the acquired information contained irrelevant American Standard Code for Information Interchange (ASCII) characters (e.g., the newline delimiter ‘\n’) or non user-created data (e.g., when a Reddit user had deleted their post, the text simply stated ‘[deleted]’). Other times, non alpha-numeric characters were used to communicate unique information. For example, triple parenthesized nouns are identified by the Anti-Defamation League (ADL) — an organization intended to “stop the defamation of the Jewish people and to secure justice and fair treatment to all” ([Anti-Defamation League](#), *nda*) — as “a typographical practice used by some anti-Semites on-line” to identify the noun with Judaism ([Anti-Defamation League](#), *ndb*). Likewise, computers read capital letters different from their lowercase counterparts. For these reasons and others, the text data were cleaned in the following manner:

1. Remove website and uniform resource locator (URL) text. Without such a step, tokens like ‘https’ will obfuscate more relevant tokens.
2. Remove ‘[removed]’ and ‘[deleted]’. Without such a step, the tokens ‘removed’ and ‘deleted’ will obfuscate more relevant tokens.

3. Replace ‘(((’ and ‘)))’ with ‘*TRIPLEXO*’ and ‘*TRIPLEXC*’, respectively. Without such a step, the anti-semitic typographical practice will disappear when non alpha-numeric characters are removed in a subsequent step. The obscure ‘*TRIPLEXO*’ and ‘*TRIPLEXC*’ replacement strings were chosen so that they would not coincide with scraped forum data from actual users. Without knowing the end from the beginning, this typographical practice could have been important. As it turned out, this typographical practice did not play an important role in this thesis.
4. Remove ASCII characters ‘\n’, ‘\r’, and ‘\t’ which represent the newline character, carriage return character, and tab character, respectively.
5. Remove all non alpha-numeric characters.
6. Remove excess white space.
7. Convert all alphabetical characters to lowercase. Without such a step, the tokens ‘*WE*’, ‘*We*’, and ‘*we*’ are all treated differently, despite conveying the same information.
8. Fix common errors from the prior seven steps (e.g., replacing ‘*you re*’ to ‘*you are*’), consolidate relevant colloquials (e.g., replace ‘*corona*’, ‘*coronavirus*’, ‘*covid 19*’ all to ‘*covid*’), replace relevant full forms with their abbreviations (e.g., ‘*make america great again*’ becomes ‘*maga*’ — note that this was done for ease of counting: ‘*maga*’ is a single entity, whereas ‘*make america great again*’ counts as four separate tokens). A full list of conversions can be found in Appendix B.
9. (Occasionally) Remove stop tokens (e.g., ‘*the*’, ‘*and*’, ‘*of*’, etc.) which do not hold much meaning. For a full set of stop tokens, see Appendix C.

The process described above is not perfect. But it nonetheless captures the original meaning well. Table 2.3 shows an example of the cleaning process.

After cleaning, the data were saved as additional csv files. The original data were saved separately in case raw data were ever desired (e.g., analyzing the posted URLs). All told, this means each forum of study had three associated csv files: one for raw text data, and two for cleaned text

Table 2.3: A demonstration of the data cleaning process described in Section 2.3.2.

Raw Data	Cleaned Data	
	with Stop Tokens	without Stop Tokens
<i>An antisemetic World is through the back door A Chinese economist found the truth: corrupt (((Pharma))). https://www.revolver.news/chinese-professor-reveals-plot-in-shock-video-us-elites-and-china-have-teamed-up-to-take-control-of-america/</i>	<i>an antisemetic world is through the back door a chinese economist found the truth corrupt triplexo pharma triplexc</i>	<i>antisemetic world back door chinese economist found truth corrupt triplexo pharma triplexc</i>

data (both with and without the removal of stop tokens). See Section 2.4 for a thorough description of the final saved files for each forum.

2.3.3 Data Storage

In total, this thesis scraped approximately 5 million posts (see Section 4.1). Therefore, efficient data storage and the reading of such was necessary. The csv files were therefore converted into feather files: “[A] lightweight binary columnar data store designed for maximum speed” (Wickham et al., 2019). See Apache Software Foundation (2021), the R *feather* package (Wickham et al., 2019), and Section 3.1.3 and Section 3.3 for more information.

2.4 The Final Data Sets

For each forum of study, the posts of that forum were saved into three feather files containing the information of Table 2.4. The three feather files corresponded to the level of text data cleaning in the *text* variable. Specifically, one file contained the text as it was scraped, another file had the text with steps 1 – 8 of Section 2.3.2 applied, and the last file had the text with steps 1 – 9 of Section 2.3.2 applied.

Table 2.4: The final variables associated with the 8kun and Reddit data. The bolded variables were those that were used in analysis.

<i>Variable</i>	Description
<i>doc_id</i>	a unique identifier of the post
<i>text</i>	all text associated with a particular post (for submissions, this combines the title and text of a post)
<i>site</i>	either ‘8kun’ or ‘Reddit’
<i>forum</i>	the forum of the associated data
<i>type</i>	either ‘submissions’ or ‘comments’
<i>author</i>	a unique identifier of the author (may or may not identify 8kun authors due to 8kun’s anonymization)
<i>submission_id</i>	a unique identifier to the associated submission (if the post is a submission, then this is the same as <i>doc_id</i>)
<i>parent_id</i>	a unique identifier that indicates the post to which someone replied (for 8kun posts, this is the same as <i>submission_id</i> ; for 8kun and Reddit, if the post is a submission, then this is the same as <i>doc_id</i>)
<i>year</i>	the year of posting
<i>month</i>	the month of posting
<i>day</i>	the day of the month of posting
<i>date_time</i>	the time of posting (precise to the minute)
<i>title</i>	the title of the post (for submissions)
<i>body</i>	the rest of the text of the post
<i>num_comments</i>	numeric indicating how many people commented to a submission
<i>score</i>	the number of upvotes to a post (Reddit only)
<i>upvote_ratio</i>	the upvote ratio to a post (Reddit only)
<i>omitted_comments</i>	the number of unscraped comments to a submission (8kun only)

CHAPTER 3

Methods

This chapter can be broken down into five main pieces. Section 3.1 details the software products used throughout the thesis. Section 3.2 and Section 3.3 detail how the data were acquired and how they were preprocessed, cleaned, and stored. Then Section 3.4 introduces mock-documents which exemplify the data analysis methods used throughout this thesis. Section 3.5 – Section 3.9 introduce the data analysis methods. Finally, Section 3.10 discusses alternative methods not used in this thesis.

3.1 Software Products

This section details the various software products used in this thesis. Data acquisition was done primarily in the Python programming language, an “easy to learn, powerful programming language [with] efficient high-level data structures and a simple but effective approach to object-oriented programming” (van Rossum and the Python Core Development Team, 2020). Data analysis was done primarily in the R programming language, a “software environment for statistical computing and graphics” (R Core Team, 2020).

3.1.1 Software Products and Packages for 8kun Data Acquisition

- (1) **The Internet Archive’s Wayback Machine** (Internet Archive, ndc). “The Internet Archive Wayback Machine is a service that allows people to visit archived versions of Web sites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web.”
- (2) **The Python *Scrapy* Package** (Scrapy Developers, 2022a). As stated by the developers: “Scrapy is a fast high-level web crawling and webscraping framework, used to crawl web-sites and extract structured data from their pages.”

3.1.2 Software Products and Packages for Reddit Data Acquisition

- (1) **The Pushshift API** ([Baumgartner et al., 2019](#)). The Pushshift application programming interface (API) “gives full functionality for searching Reddit data and also includes the capability of creating powerful data aggregations. With this API, you can quickly find the data that you are interested in.”
- (2) **The Python *PSAW* Package** ([Marx, 2018](#)). The Python *PSAW* package is a “minimalist wrapper for searching public reddit comments/submissions via the pushshift.io API.”
- (3) **Reddit’s API** ([Reddit, 2022b](#)). Reddit’s API allows one to access publicly available Reddit data in a systematic way.
- (4) **The Python *PRAW* Package** ([Boe, 2022](#)). The Python *PRAW* package is a wrapper to Reddit’s own API.

3.1.3 Packages for Data Preprocessing, Data Cleaning, and Data Storage

- (1) **The R *dplyr* Package** ([Wickham et al., 2022b](#)). During preprocessing, data files needed to be reformatted and “dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges” ([dplyr, nd](#)).
- (2) **The R *stringr* Package** ([Wickham, 2019](#)). The main type of data in this thesis were text strings, and the “stringr package provide[s] a cohesive set of functions designed to make working with strings as easy as possible” ([stringr, nd](#)).
- (3) **The R *tm* Package** ([Feinerer and Hornik, 2020](#)). Likewise, the R *tm* package provides a “framework for text mining.”
- (4) **The R *feather* Package** ([Wickham et al., 2019](#)). In total, this thesis acquired approximately 5 million posts (see Table 4.1). Therefore, speed in data storage and import were imperative. The R *feather* package provides “a lightweight binary columnar data store designed for maximum speed.”

3.1.4 Packages for Computation and Data Extraction

- (1) **The R *dplyr* Package** (Wickham et al., 2022b). As above, “dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges” (dplyr, nd).
- (2) **The R *stringr* Package** (Wickham, 2019). As above, the “stringr package provide[s] a cohesive set of functions designed to make working with strings as easy as possible” (stringr, nd).
- (3) **The R *tidytext* Package** (De Queiroz et al., 2022). The R *tidytext* package, like the R *stringr* package, “can make many text mining tasks easier, more effective, and consistent with tools already in wide use.”

3.1.5 Packages for word2vec Analysis

- (1) **The R *word2vec* Package** (Wijffels, 2021). The R *word2vec* package implements the word2vec machine learning software (and other derivatives). See Section 3.6.2 for more information.

3.1.6 Packages for Cluster Analysis

- (1) **The R *NbClust* Package** (Charrad et al., 2022). This thesis used cluster analysis to analyze document similarity (see Section 3.8), and the R *NbClust* package provides “30 indexes for determining the optimal number of clusters in a data set.”

3.1.7 Packages for Visualizations

- (1) **The R *ggplot2* Package** (Wickham et al., 2022a). The methods used throughout this thesis were visualized primarily using the R *ggplot2* package: “A system for ‘declaratively’ creating graphics.”
- (2) **The R *dendextend* Package** (Galili et al., 2022). The R *dendextend* package supplies “a set of functions for extending ‘dendrogram’ objects in R.” Specifically, it supplies functionality to modify dendrogram visualizations.

- (3) **The R *ggtext* Package** (Wilke, 2020). Graphics were enhanced by the R *ggtext* package: “A ‘ggplot2’ extension that enables the rendering of complex formatted plot labels (titles, subtitles, facet labels, axis labels, etc.).”

3.2 Data Acquisition

8kun and Reddit data were both acquired using Python (see [van Rossum and the Python Core Development Team \(2020\)](#)). However, the exact mode and software products used were different.

3.2.1 8kun

8kun data were scraped using the Python *Scrapy* package ([Scrapy Developers, 2022a](#)). This Python package was able to acquire all the relevant and available 8kun data by working directly with the HyperText Markup Language (HTML) and Cascading Style Sheets (CSS) code of 8kun accessed through the Internet Archive’s Wayback Machine (WBM) (see Section 2.2.1 and Section 3.1.1 for information on the WBM), as in the following example.

Python *Scrapy* Example with Mock Webpage Code

The small demonstration shown below is only meant to be a minimal working example of the Python *Scrapy* package. As such, please refer to [Duckett \(2011\)](#) for an expansive overview of HTML and CSS code, as well as the structure of webpages. Likewise, refer to the Python *Scrapy* documentation ([Scrapy Developers, 2022a](#)) and Python *Scrapy* tutorial ([Scrapy Developers, 2022b](#)) for more details related to the Python *Scrapy* package and its implementation.

The HTML code of 8kun (accessed through the WBM) was well organized and easy to navigate. However, the website’s structure had HTML tags nested — at times — ten units deep. Therefore, actual 8kun HTML code would not be conducive to illustration. Consider instead the mock HTML code (created by author) contained in Figure 3.1.

```

<!DOCTYPE html>
<html>
(1)     <head>
(2)         <title>Fake Web Page Title (Not Displayed to User)</title>
(3)     </head>
(4)     <body class="this is a css attribute" id="another css attribute">
(5)         <div class="first div environment">
(6)             <p id="hello">
(7)                 Displayed content (like this) is contained inside the HTML body tag.
(8)             </p>
(9)             <p id="hello">
(10)                HTML div tags compartemantalize/organize a webpage.
(11)            </p>
(12)            <span id="hello">
(13)                The HTML p and span tags stand for "paragraph" and "span", respectively.
(14)            </span>
(15)        </div>
(16)        <div class="second div environment">
(17)            <p id="unique id">
(18)                A paragraph with a unique CSS id attribute.
(19)            </p>
(20)            <div class="nested div">
(21)                <p>
(22)                    This paragraph tag has no css attribute.
(23)                </p>
(24)                <p class="goodbye" id="friend">
(25)                    This paragraph tag has two css attributes.
(26)                </p>
(27)            </div>
(28)        </div>
(29)    </body>
</html>

```

Fig. 3.1: Mock HTML code (created by author) to exemplify the Python *Scrapy* package. Numbering is not part of the HTML code but is for reference only.

The fundamental unit of scraping within the Python *Scrapy* package is a `selector` object — effectively, a portion of HTML code. Scraping begins with a `selector` object named `response` which contains the entire page. From there, new `selector` objects are created by navigating the

tree structure of the HTML code in a similar manner as one would navigate document files on a computer.

For instance, suppose one wanted the information contained within the three `p` tags in lines 6 – 8, 9 – 11, and 17 – 19. Running:

```
response.xpath('./body/div/p')
```

would create the three desired `selector` objects: one for each `p` tag contained within a `div` tag contained within the `body` tag. To obtain only the first two of these (i.e., lines 6 – 8 and 9 – 11, the ones contained within the first `div` tag), one could run:

```
response.xpath('./body/div/p')[0:2]
```

where indexing starts at 0 and 2 `selectors` are chosen. One could also specify which `div` tag to consider. That is:

```
response.xpath('./body/div[@class="first div environment"]/p')
```

would also result in the same two selectors (lines 6 – 8 and 9 – 11). Likewise, note that these same `p` tags are the only `p` tags with the CSS attribute of `id="hello"`. Therefore:

```
response.xpath('./body/div/p[@id="hello"]')
```

would also result in the same two selectors (lines 6 – 8 and 9 – 11).

Now suppose one wanted the information contained within the three `p` tags in lines 17 – 19, 21 – 23, and 24 – 26 (i.e., the `p` tags within the second `div` tag). Running:

```
response.xpath('./body/div[@class="second div environment"]/p')
```

would only result in exactly one selector: lines 17 – 19. The missing selectors are not included because they are nested within a second `div` tag. One can specify to include all `p` tags within a selector using `//`. That is:

```
response.xpath('./body/div[@class="second div environment"]//p')
```

would result in the three desired selectors corresponding to lines 17 – 19, 21 – 23, and 24 – 26.

To obtain the displayed text data, add `/text()` and `.getall()` in the following manner:

```
response.xpath('./body/div[@class="second div environment"]//p/text()).getall()
```

which will return a three element list containing the character strings of lines 18, 22, and 25.

The above commands are not comprehensive. For any piece of desired information, there are many corresponding commands. For instance, the following commands:

```
response.xpath('./body/div/span/text()').getall()

response.xpath('./body[@class="this is a css attribute"]/div/span/text()').getall()

response.xpath('./body[@id="another css attribute"]/div/span/text()').getall()

response.xpath('./body/div[@class="first div environment"]/span/text()').getall()

response.xpath('./div[@class="first div environment"]/span/text()').getall()

response.xpath('./span[@id="hello"]//text()').getall()

response.xpath('./span/text()').getall()
```

all return a one element list containing the character string of line 13 within the span HTML tag.

That is, each of the above commands returns the character string:

The HTML p and span tags stand for "paragraph" and "span", respectively.

Likewise, this simple example does not cover all that can be done with the Python *Scrapy* package (e.g., following uniform resource locator (URL)s). Again, additional information can be found at [Scrapy Developers \(2022a\)](#) and [Scrapy Developers \(2022b\)](#).

3.2.2 Reddit

As discussed in Section 2.2.2, Reddit data were scraped using the Pushshift API ([Baumgartner et al., 2019](#)) and the Python *PSAW* package ([Marx, 2018](#)), as well as Reddit's API ([Reddit, 2022b](#)) and the Python *PRAW* package ([Boe, 2022](#)). A visual representation of the Reddit data acquisition process is visualized in Figure A.1 in Appendix A. The process is also demonstrated in the following example.

Reddit Data Acquisition Example

The Python *PSAW* package was used to create a generator object of unique identification strings to desired submissions. This generator is created and assigned to the variable `gen` by running:

```
gen = PSAW.PushshiftAPI().search_submissions(
    after=...
```

```

        before=...
        subreddit=...
        filter=["id"]
        limit=None
    )

```

where the user supplies the `after`, `before`, and `subreddit` arguments detailing the desired time frame and subreddit (forum). The `filter` argument specifies that the unique `"id"` variable is desired, and the `limit` argument specifies to return all submissions in the generator (and not some fixed integer amount).

One then accesses Reddit's API — and assigns the access to the variable `reddit` — by running the following command:

```

reddit = PRAW.praw.Reddit(
    site_name=...,
    user_agent=...
)

```

The arguments to this function hold private credentials that allow one to access Reddit's API. See [Boe \(2022\)](#) for more information.

With the `gen` and `reddit` variables declared, one attains the data for the submissions (and their associated comments) by iterating through the `gen` object as follows:

```

for psaw_submission in gen:
    # Pull unique submission id from gen (PSAW)
    submission_id = psaw_submission.d_["id"]

    # Access data of that submission through Reddit's API (PRAW)
    submission = reddit.submission(id=submission_id)

    # Get desired attributes of submission. For example
    submission.author    # Get the author of the submission
    submission.selftext  # Get the text of the submission

    # Fetch all comments associated with the submission and iterate
    # through list of comments
    for comment in submission.comments.replace_more(limit=None).list():
        # Get desired attributes of comment. For example

```

```
comment.author # Get the author of the comment
comment.body   # Get the text of the comment
```

3.3 Data Preprocessing, Data Cleaning, and Data Storage

As discussed in Section 2.3.1, the scraped data for this thesis needed to be reformatted. This was done using the R *dplyr* package (Wickham et al., 2022b) and associated functionality such as `dplyr::mutate` for creating new variables, `dplyr::bind_rows` for combining data frames, and `dplyr::rename` for renaming data frame columns.

As discussed in Section 2.3.2, the data scraped for this thesis were messy. The R *stringr* package (Wickham, 2019) was used for text removal and text replacement according to the steps shown in Section 2.3.2. For example, the command:

```
stringr::str_replace_all(text, "[^[:alnum:]]", " ")
```

would modify `text` by replacing all non alpha-numeric characters with a space. Stop tokens were removed in a similar manner (see Appendix C for a full list of stop tokens).

Finally, as discussed in Section 2.3.3, the data for this thesis needed to be stored and read efficiently. The R *feather* package (Wickham et al., 2019) converted comma-separated value (csv) files into feather files (Apache Software Foundation, 2021) using the `feather::write_feather` function. The R *feather* package also read such files into R using the `feather::read_feather` function. A simple microbenchmark test suggested that feather files and the R *feather* package were faster than other alternatives (e.g., approximately 6.5 times faster than `utils::read.csv` and twice as fast as `data.table::fread` — see R Core Team and Contributors Worldwide (2022b) and Dowle and Srinivasan (2021) for more information on these functions).

3.4 Introduction of Mock-Documents

This thesis analyzes text documents. Formally, a text *document* is a single entity of text data united under a common facet (e.g., the text for a specific online post, or the combined text of a specific online submission with its associated comments). Documents are represented — in general — by d , \bar{d} , or $\bar{\bar{d}}$, which correspond to different levels of data cleaning. Specifically, d represents the text document as it was scraped, \bar{d} represents the text document with steps 1 – 8 of Section 2.3.2

applied, and $\bar{\bar{d}}$ represents the text document with steps 1 – 9 of Section 2.3.2 applied. See Table 2.3 for an example of the text cleaning process. Also see Appendix D for the terminology and notation used throughout this thesis.

To exemplify the methods described in Section 3.5 – Section 3.8, consider Table 3.1, which shows nine mock-documents coming from three mock-forums: f_1 , f_2 , and f_3 . The mock-documents in Table 3.1 were created to demonstrate the methods in Section 3.5 – Section 3.8. The documents analyzed in this thesis were considerably larger than those displayed in Table 3.1. Likewise this thesis analyzed far more than nine documents (approximately 5 million posts were scraped for this thesis — see Section 4.1).

Table 3.1: Mock-documents (created by the author) to exemplify the methods in Section 3.5 – Section 3.8. The notation $d_{i,j}$ indicates the j^{th} mock-document from the i^{th} mock-forum, f_i . More details are provided in the main text.

$d_{i,j}$	Raw Data	Cleaned Data	
	d	with Stop Tokens \bar{d}	without Stop Tokens $\bar{\bar{d}}$
$d_{1,1}$	<i>Hypothesis on the election and Trump: Trump will concede the election and admit Biden won, but Trump or Trump Jr. will run again.</i>	<i>hypothesis on the election and trump trump will concede the election and admit biden won but trump or trump jr will run again</i>	<i>hypothesis election trump trump will concede election admit biden won trump trump jr will run</i>
$d_{1,2}$	<i>The Biden victory was INTENSE and is one for the history books. More people voted for Trump in 2020 than in 2016, but he still lost.</i>	<i>the biden victory was intense and is one for the history books more people voted for trump in 2020 than in 2016 but he still lost</i>	<i>biden victory intense one history books people voted trump 2020 2016 still lost</i>
$d_{1,3}$	<i>Trump did concede and Biden won the 2020 presidential election last night. Biden will be inaugurated in 2021.</i>	<i>trump did concede and biden won the 2020 presidential election last night biden will be inaugurated in 2021</i>	<i>trump concede biden won 2020 presidential election last night biden will inaugurated 2021</i>
$d_{1,4}$	<i>I hope Trump and Biden do not run in the next presidential election.</i>	<i>i hope trump and biden do not run in the next presidential election</i>	<i>hope trump biden run next presidential election</i>
$d_{2,1}$	<i>Hypothesis: Earth earth will warm by 2050 all thanks to Trump.</i>	<i>hypothesis earth earth will warm by 2050 all thanks to trump</i>	<i>hypothesis earth earth will warm 2050 thanks trump</i>
$d_{2,2}$	<i>The warm Earth climate is all thanks to Trump.</i>	<i>the warm earth climate is all thanks to trump</i>	<i>warm earth climate thanks trump</i>
$d_{2,3}$	<i>The climate is a problem for this generation (thanks).</i>	<i>the climate is a problem for this generation thanks</i>	<i>climate problem generation thanks</i>
$d_{3,1}$	<i>A math professor just ended math (proved Riemann hypothesis).</i>	<i>a math professor just ended math proved riemann hypothesis</i>	<i>math professor just ended math proved riemann hypothesis</i>
$d_{3,2}$	<i>A math professor just proved the Taylor-Johnson conjecture.</i>	<i>a math professor just proved the taylor johnson conjecture</i>	<i>math professor just proved taylor johnson conjecture</i>

3.5 Important Statistics of Text Documents

This section begins the body of work that describes the methods used in data analysis (Section 3.5 – Section 3.9). First, this section introduces relevant notation and terminology. Then an important statistic of text documents is introduced.

Relevant Notation and Terminology

A *set* is a collection of objects (called *elements*) with no ordering and no repetitions. That is, $S_1 = \{1, 2\}$, $S_2 = \{1, 1, 2\}$, $S_3 = \{2, 2, 1\}$, and $S_4 = \{2, 1\}$ are all exactly the same set, which contains exactly two elements. Note that there is no need to specify that these sets contain exactly two *distinct* elements. They simply contain exactly two elements, despite how they may be written. The notation $|S|$ is the *cardinality* of a set S and is the number of elements in S . The notation $a \in S$ means that a is an element of the set S , whereas $a \notin S$ means that a is not an element of the set S . Finally, for sets S and S' , the set $S - S'$ is defined by:

$$S - S' = \{a \in S : a \notin S'\}.$$

A *token* t is a string of characters unbroken by any whitespace, which are “the empty space[s] between all the characters you can actually see” (e.g., letters, numbers, etc.) (Bartlett, 2020). Note that tokens are also commonly called *terms* and *words*. This thesis uses the token terminology because it is the most unique of the three. Then, mathematically, a document d is a set of ordered pairs (i, t) where i is a positive integer index specifying the location of a token t . For example, $\bar{d}_{3,1}$ in Table 3.1 is:

$$\bar{d}_{3,1} = \left\{ (1, a), (2, \text{math}), (3, \text{professor}), (4, \text{just}), (5, \text{ended}), (6, \text{math}), \right. \\ \left. (7, \text{proved}), (8, \text{riemann}), (9, \text{hypothesis}) \right\}.$$

In this way, documents are represented as sets with both an ordering index and repeated tokens. For clarity:

$$\begin{aligned} d_1 &= \{bob, is, my, name\} \\ d_2 &= \{my, name, is, bob\} \\ d_3 &= \{my, name, is, bob, bob\} \end{aligned}$$

are exactly the same set with cardinality $|d_1| = |d_2| = |d_3| = 3$. But:

$$\begin{aligned} d'_1 &= \{(1, bob), (2, is), (3, my), (4, name)\} \\ d'_2 &= \{(1, my), (2, name), (3, is), (4, bob)\} \\ d'_3 &= \{(1, my), (2, name), (3, is), (4, bob), (4, bob)\} \\ d'_4 &= \{(1, my), (2, name), (3, is), (4, bob), (5, bob)\} \end{aligned}$$

are three different sets with d'_2 and d'_3 being the same set, and $|d'_1| = |d'_2| = |d'_3| = 4$ and $|d'_4| = 5$. As a final example, consider the following three equivalent documents, represented as sets of ordered pairs.

$$\begin{aligned} d''_1 &= \{(4, litmus), (2, is), (1, this), (3, a), (5, test)\} \\ d''_2 &= \{(5, test), (2, is), (3, a), (4, litmus), (1, this)\} \\ d''_3 &= \{(2, is), (1, this), (5, test), (3, a), (4, litmus), (5, test)\} \end{aligned}$$

This formality means that things defined as a set are exactly a set, with no ordering or repetitions. It also means that certain definitions and notations are wholly unambiguous. For example, for a document d , $|d|$ is the number of tokens (counting repeated tokens) in the document. This works because the indexing variable i within the (i, t) ordered pairs is different for each token, as if being read by a reader.

The (slight abuse of) notation $t \in d$ indicates that there is some index i for which $(i, t) \in d$. And for documents d_1 and d_2 define:

$$d_1 \odot d_2 = d_1 \cup \{(i + |d_1|, t) : (i, t) \in d_2\}.$$

In other words, $d_1 \odot d_2$ is simply the concatenation of document d_1 followed by d_2 . This operation is associative. That is, for documents d_1 , d_2 , and d_3 , $(d_1 \odot d_2) \odot d_3 = d_1 \odot (d_2 \odot d_3)$. However, for absolute clarity, let repeated applications of \odot be read left to right (i.e., concatenations are applied as if being read by a reader). That is:

$$d_1 \odot d_2 \odot d_3 \odot \dots \odot d_{n-1} \odot d_n = \left((d_1 \odot d_2) \odot d_3 \right) \odot \dots \odot d_{n-1} \odot d_n.$$

A *corpus* C is a set of documents and the cardinality of C , $|C|$, is equal to the number of documents in C . A *forum* f is a set of documents coming from the same online source. In this way, a forum is a specific instance of a corpus, and a corpus is an abstraction of a forum. A set of tokens D is a *dictionary*.

The notation $\text{ntokens}(d, t)$ is the number of occurrences of the token t in the document d . That is:

$$\text{ntokens}(d, t) = |\{(i, \tau) \in d : \tau = t\}|.$$

From this, one defines the token frequency (tf) of t in d as:

$$\text{tf}(d, t) = \frac{\text{ntokens}(d, t)}{\sum_{t_i \in d} \text{ntokens}(d, t_i)} = \frac{\text{ntokens}(d, t)}{|d|}.$$

That is, $\text{tf}(d, t)$ is the proportion of tokens in d that are equal to t . Note that token frequency is also commonly called *term* frequency.

The document d^p is the document d where tokens t which do not satisfy $\text{tf}(d, t) \geq p$ are removed. The parameter p is a small proportion, commonly chosen as 0.001. This parameter is chosen big enough to filter out tokens that are uncommon enough to artificially inflate tf-idf values (defined in the next paragraph), but small enough to keep major themes and intended meanings in

tact.

For a corpus of documents C , and a fixed document $d \in C$ as well as a fixed token t , the token frequency - inverse document frequency (tf-idf) of the token t and document d inside C is:

$$\text{tf-idf}(C, d, t) = \text{tf}(d, t) \cdot \ln \left(\frac{|C|}{\sum_{d_i \in C} [t \in d_i]} \right),$$

where \ln is the natural logarithm \log_e , and $[\cdot]$ is the Iverson bracket. That is, for a Boolean proposition Q (e.g., $t \in d_i$), $[Q] = 1$ if Q is true, and $[Q] = 0$ if Q is false. Note that token frequency - inverse document frequency is also commonly called *term* frequency - inverse document frequency.

Finally, note that these definitions were based on raw text documents (as indicated by the absence of any bar(s) over the d 's). However, the definitions extend to cleaned text documents as well. See Appendix D for the terminology and notation used throughout this thesis.

Token Frequency - Inverse Document Frequency

Above, the definition of tf-idf may seem odd. However, as stated by [Singh and Shashi \(2019\)](#), tf-idf “reflects the prominence of a [token] in a collection of documents to [an] individual document.” Holistically, the larger $\text{tf-idf}(C, d, t)$ is, the more important the token t is to the document d when compared with the other documents of C . Viewing the definition, and noting that \ln is an increasing function, there are three ways to increase $\text{tf-idf}(C, d, t)$ (one does not actually have the ability to change $\text{tf-idf}(C, d, t)$ for a fixed C , d , and t , but it is informative to think about it this way). (1) Increase the prominence of t in d , and/or (2) Decrease the appearance of t in the other documents of the corpus, and/or (3) Increase the number of documents in the corpus. The values of $\text{tf-idf}(C, d, t)$ range from 0 (if $t \notin d$ or $t \in d_i$ for every $d_i \in C$) to $\text{tf}(d, t) \cdot \ln(|C|)$ (if t is only present in document d). In this thesis, tf-idf values were used for a document vectorization scheme from which to measure document similarities — see Section 3.6.1.

3.6 Vectorizations of Text Documents

To quantitatively analyze text documents, they must be represented numerically. This thesis does so by encoding the information of a text document as a vector, thus allowing analysis to be done

in the language of linear algebra. This section first introduces relevant notation and terminology, and then details the two ways this thesis encoded text documents as vectors.

Relevant Notation and Terminology

A *vectorization* of a text document is a way to represent a text document in a purely quantitative way — specifically, as a vector. If d is a document, then its vectorization is represented by \mathbf{d} . Notationally, if \mathbf{d} is a vectorization of a document, then $(\mathbf{d})_i$ is its i^{th} component (and in general, $(\mathbf{v})_i$ is the i^{th} component of an arbitrary vector \mathbf{v}). See Appendix D for the terminology and notation used throughout this thesis.

3.6.1 Token Frequency - Inverse Document Frequency

As mentioned in Section 3.5, tf-idf is one way in which a document may be vectorized, the importance of such a vectorization being emphasized by Singh and Shashi (2019):

[Token] Frequency-Inverse Document Frequency is the most commonly used method in NLP [natural language processing] for converting text documents into matrix representation of vectors.

The tf-idf vectorizations used in this thesis were applied to monthly aggregated documents (as in Procedure 3.6.1.a). From these monthly aggregated documents, Procedure 3.6.1.b creates the vectorizations.

Procedure 3.6.1.a Let F be the set of forums of interest. For each $f \in F$ and each month $m \in M = \{10, 11, 12, 01, 02, 03\}$, suppose that $\bar{d}_{f,m,1}, \dots, \bar{d}_{f,m,N_m^f}$ are the N_m^f (cleaned, with stop tokens removed) posts from forum f in month m . For each f and m , let:

$$\bar{d}_{f,m} = \bar{d}_{f,m,1} \odot \dots \odot \bar{d}_{f,m,N_m^f}.$$

Then filter out tokens which are not used with frequency at least $p = 0.001$, giving $\bar{d}_{f,m}^p$. Let C^{Month} be the resulting corpus containing $6 \times |F|$ documents (each forum of F creates six documents corresponding to the six months of data).

Procedure 3.6.1.b Let F be the set of forums of interest. Let η be the desired number of tokens for the vectorizations. Let C^{Month} be the corpus of monthly aggregated documents as created by Procedure 3.6.1.a with the forums F .

For each $\bar{d}_{f,m}^p \in C^{\text{Month}}$, let $D_{f,m}$ be the set of tokens $t_{f,m}$ such that:

$$\text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, t_{f,m}) \geq \text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, t)$$

for every other $t \in \bar{d}_{f,m}^p$. In other words, $D_{f,m}$ contains the token(s) with the largest tf-idf value for the document $\bar{d}_{f,m}^p$ in the corpus C^{Month} . Let $D^{\text{Ensure}} = \bigcup_{f,m} D_{f,m}$. The set D^{Ensure} is a set to make sure that each document of C^{Month} contributes at least one token to the final set of tokens from which the vectorizations are built. If a document did not contribute at least one token, then that document may have a zero tf-idf value for every token produced by this procedure, and would have to be excluded from subsequent clustering results (Section 3.8).

For every document/token pair $(\bar{d}_{f,m}^p, t)$, compute $\text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, t)$. Suppose that there are A such document/token pairs: $e_1 = (\bar{d}_{f_1,m_1}^p, t_1), \dots, e_A = (\bar{d}_{f_A,m_A}^p, t_A)$. Further suppose that the first a of these document/token pairs are those with positive tf-idf values. This is simply to ensure that each token t has at least one corresponding document $\bar{d}_{f,m}^p$ such that $\text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, t)$ is positive (otherwise, a dendrogram of tokens could not be produced). Suppose:

$$\text{tf-idf}(C^{\text{Month}}, \bar{d}_{f_1,m_1}^p, t_1) \geq \dots \geq \text{tf-idf}(C^{\text{Month}}, \bar{d}_{f_a,m_a}^p, t_a) > 0. \quad (1)$$

Let η' be the positive integer such that $|D^{\text{Ensure}} \cup \{t_1, \dots, t_{\eta'}\}| = \eta$. Then define the set:

$$D^{\text{Ties}} = \{t_{\tilde{\eta}} : \tilde{\eta} > \eta' \text{ and } \text{tf-idf}(C^{\text{Month}}, \bar{d}_{f_{\tilde{\eta}},m_{\tilde{\eta}}}^p, t_{\tilde{\eta}}) = \text{tf-idf}(C^{\text{Month}}, \bar{d}_{f_{\eta'},m_{\eta'}}^p, t_{\eta'})\}.$$

The set D^{Ties} is merely to account for the fringe case when the η' th relation in (1) is

equality, not strictly greater than. This is unlikely for large text documents. Most often D^{Ties} is empty.

Let:

$$D^{\text{Ensure}} \cup \{t_1, \dots, t_{\eta'}\} \cup D^{\text{Ties}} = \{\hat{t}_1, \dots, \hat{t}_{\hat{\eta}}\}$$

be the final set of tokens from which to build the vectorizations. That is, for each $\bar{d}_{f,m}^p \in C^{\text{Month}}$, define its vectorization by $(\bar{\mathbf{d}}_{f,m}^p)_i = \text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, \hat{t}_i)$ for $i = 1, \dots, \hat{\eta}$.

Procedure 3.6.1.a Example with Mock-Documents

Consider the mock-documents in Table 3.1. Recall that $d_{i,j}$ represents the j^{th} document from the i^{th} forum, f_i . Note that Procedure 3.6.1.a considers six different months of data. However, because there are only nine mock-documents, this example will only consider two consecutive months of data. That is, in the notation of Procedure 3.6.1.a, $M = \{01, 02\}$, which correspond to January and February, respectively. Suppose that the mock-document $d_{i,j}$ occurred in January (respectively, February) if j is odd (respectively, if j is even).

Then C^{Month} consists of the following 2×3 documents (each of the three forums creates two documents corresponding to two months of data).

$\bar{d}_{f_1, \text{Jan}}^p$: *hypothesis election trump trump will concede election admit biden won trump trump jr will run trump concede biden won 2020 presidential election last night biden will inaugurated 2021*

$\bar{d}_{f_1, \text{Feb}}^p$: *biden victory intense one history books people voted trump 2020 2016 still lost hope trump biden run next presidential election*

$\bar{d}_{f_2, \text{Jan}}^p$: *hypothesis earth earth will warm 2050 thanks trump climate problem generation thanks*

$\bar{d}_{f_2, \text{Feb}}^p$: *warm earth climate thanks trump*

$\bar{d}_{f_3, \text{Jan}}^p$: *math professor just ended math proved riemann hypothesis*

$\bar{d}_{f_3, \text{Feb}}^p$: *math professor just proved taylor johnson conjecture*

Procedure 3.6.1.b Example with Mock-Documents

Consider the mock-documents in Table 3.1. Recall that $d_{i,j}$ represents the j^{th} document from the i^{th} forum, f_i . Fix $\eta = 20$. Let C^{Month} be the corpus of documents as created by Procedure 3.6.1.a (which can be seen in the preceding example of Procedure 3.6.1.a).

For every document/token pair $(\bar{d}_{f,m}^p, t)$, the $\text{tf-idf}(C^{\text{Month}}, \bar{d}_{f,m}^p, t)$ value needs to be computed. Such calculations are shown in Table 3.2. The sets $D_{f,m}$ are the highlighted rows. For example, $D_{f_3, \text{Jan}} = \{\text{math}\}$, whereas $D_{f_3, \text{Feb}} = \{\text{conjecture}, \text{johnson}, \text{taylor}\}$ to allow for ties. Then D^{Ensure} consists of all the tokens in the highlighted rows. This set consists of 11 tokens (13 highlighted rows with two repeated tokens). Then, to return $\eta = 20$ tokens, an additional nine tokens must be added. That is, $\eta' = 25$, because the first 25 rows (plus the highlighted rows) correspond to 20 unique tokens. Then, to account for ties in the 25th row, $D^{\text{Ties}} = \{\text{election}, \text{will}\}$. In total, the 21 bolded tokens (repeated tokens are not bolded twice) in Table 3.2 make up the final set of tokens from which to build the vectorizations. In the notation of Procedure 3.6.1.b, the $\hat{\eta} = 21$ bolded tokens correspond to $\hat{t}_1, \dots, \hat{t}_{\hat{\eta}}$. The tf-idf vectorizations are then shown in Table 3.3.

Table 3.2: As in Procedure 3.6.1.b: the selection of the tokens for the tf-idf vectorizations of the monthly aggregated mock-documents. For each document $\bar{d}_{f,m}^p$ and token t , the tf-idf value is shown and arranged in descending order. Highlighted rows correspond to the top tf-idf token(s) for each document. The final tokens are bolded (repeat tokens are not bolded twice). More details are provided in the main text. The original mock-documents are displayed in Table 3.1. The monthly aggregation of these mock-documents are shown in the example of Procedure 3.6.1.a.

	Document $\bar{d}_{f,m}^p$	Token t	tf-idf($C^{\text{Month}}, \bar{d}_{f,m}^p, t$)		Document $\bar{d}_{f,m}^p$	Token t	tf-idf($C^{\text{Month}}, \bar{d}_{f,m}^p, t$)
1	f ₃ , Jan	math	0.27	32	f ₁ , Feb	2016	0.09
2	f ₃ , Feb	conjecture	0.26	33	f ₁ , Feb	books	0.09
3	f ₃ , Feb	johnson	0.26	34	f ₁ , Feb	history	0.09
4	f ₃ , Feb	taylor	0.26	35	f ₁ , Feb	hope	0.09
5	f ₃ , Jan	ended	0.22	36	f ₁ , Feb	intense	0.09
6	f ₃ , Jan	riemann	0.22	37	f ₁ , Feb	lost	0.09
7	f ₂ , Feb	climate	0.22	38	f ₁ , Feb	next	0.09
8	f ₂ , Feb	earth	0.22	39	f ₁ , Feb	one	0.09
9	f ₂ , Feb	thanks	0.22	40	f ₁ , Feb	people	0.09
10	f ₂ , Feb	warm	0.22	41	f ₁ , Feb	still	0.09
11	f ₂ , Jan	earth	0.18	42	f ₁ , Feb	victory	0.09
12	f ₂ , Jan	thanks	0.18	43	f ₁ , Feb	voted	0.09
13	f ₃ , Feb	just	0.16	44	f ₃ , Jan	hypothesis	0.09
14	f ₃ , Feb	math	0.16	45	f ₂ , Feb	trump	0.08
15	f ₃ , Feb	professor	0.16	46	f ₁ , Jan	trump	0.07
16	f ₃ , Feb	proved	0.16	47	f ₁ , Jan	2021	0.06
17	f ₂ , Jan	2050	0.15	48	f ₁ , Jan	admit	0.06
18	f ₂ , Jan	generation	0.15	49	f ₁ , Jan	inaugurated	0.06
19	f ₂ , Jan	problem	0.15	50	f ₁ , Jan	jr	0.06
20	f ₃ , Jan	just	0.14	51	f ₁ , Jan	last	0.06
21	f ₃ , Jan	professor	0.14	52	f ₁ , Jan	night	0.06
22	f ₃ , Jan	proved	0.14	53	f ₂ , Jan	hypothesis	0.06
23	f ₁ , Jan	concede	0.13	54	f ₁ , Feb	2020	0.05
24	f ₁ , Jan	won	0.13	55	f ₁ , Feb	election	0.05
25	f ₁ , Jan	biden	0.12	56	f ₁ , Feb	presidential	0.05
26	f ₁ , Jan	election	0.12	57	f ₁ , Feb	run	0.05
27	f ₁ , Jan	will	0.12	58	f ₁ , Feb	trump	0.04
28	f ₁ , Feb	biden	0.11	59	f ₁ , Jan	2020	0.04
29	f ₂ , Jan	climate	0.09	60	f ₁ , Jan	presidential	0.04
30	f ₂ , Jan	warm	0.09	61	f ₁ , Jan	run	0.04
31	f ₂ , Jan	will	0.09	62	f ₂ , Jan	trump	0.03
				63	f ₁ , Jan	hypothesis	0.02

Table 3.3: The tf-idf vectorizations of the monthly aggregated mock-documents produced by Procedure 3.6.1.b, with $\hat{\eta} = 21$. Tokens are arranged in alphabetical order. Missing values correspond to a 0 tf-idf value. The original mock-documents are displayed in Table 3.1. The monthly aggregation of these mock-documents are shown in the example of Procedure 3.6.1.a.

\hat{t}	$\bar{\mathbf{d}}_{f_1, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_1, \text{Feb}}^p$	$\bar{\mathbf{d}}_{f_2, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_2, \text{Feb}}^p$	$\bar{\mathbf{d}}_{f_3, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_3, \text{Feb}}^p$
<i>2050</i>			0.15			
<i>biden</i>	0.12	0.11				
<i>climate</i>			0.09	0.22		
<i>concede</i>	0.13					
<i>conjecture</i>						0.26
<i>earth</i>			0.18	0.22		
<i>election</i>	0.12	0.05				
<i>ended</i>					0.22	
<i>generation</i>			0.15			
<i>johnson</i>						0.26
<i>just</i>					0.14	0.16
<i>math</i>					0.27	0.16
<i>problem</i>			0.15			
<i>professor</i>					0.14	0.16
<i>proved</i>					0.14	0.16
<i>riemann</i>					0.22	
<i>taylor</i>						0.26
<i>thanks</i>			0.18	0.22		
<i>warm</i>			0.09	0.22		
<i>will</i>	0.12		0.09			
<i>won</i>	0.13					

3.6.2 word2vec and doc2vec

Originally patented in 2013, the word2vec machine learning software, developed by a team

of Google researchers, embeds tokens as vectors (see [Mikolov et al. \(2021\)](#) for the current patent) using neural network techniques. [Goldberg and Levy \(2014\)](#) stated:

The word2vec software of Tomas Mikolov and colleagues has gained a lot of traction lately, and provides state-of-the-art [token] embeddings.

The word2vec software can be extended to embed entire documents — not just tokens — as vectors (see [Le and Mikolov \(2014\)](#)). Such document vectorizations are referred to as doc2vec vectorizations. This thesis used the R *word2vec* package ([Wijffels, 2021](#)) for doc2vec document vectorizations.

The procedure by which doc2vec vectorizations were created is as follows.

Procedure 3.6.2 Let d_1, \dots, d_V be the documents to be vectorized.

Consider the corpus $C = \{\bar{d}_1, \dots, \bar{d}_V\}$ of cleaned text documents (where stop tokens are not removed). Train the word2vec neural network on the documents in C (see Appendix E for details on how the word2vec neural network is trained). This is automated by passing the documents $\bar{d}_1, \dots, \bar{d}_V$ into the `word2vec::word2vec` function from the R *word2vec* package. Then create the document vectorizations $\bar{\mathbf{d}}_1, \dots, \bar{\mathbf{d}}_V$ using the trained model and the `word2vec::doc2vec` function from the R *word2vec* package.

Note that the word2vec neural network (and its `word2vec::word2vec` implementation) has many tunable parameters. For example, there are parameters to determine the dimension of resulting vectorizations, or to set the initial learning rate. Changing these parameters would result in different vectorizations because the structure of the neural network would have changed. These parameters, however, were left to their defaults as implemented by the `word2vec::word2vec` function. This was done because (a) the defaults fit within the recommendations provided by [Mikolov et al. \(2013a\)](#) and [Mikolov et al. \(2013b\)](#); (b) the defaults still give positive results, even within the small clustering example of mock-documents in Section 3.8; and (c) this thesis is not meant to be a detailed study of the word2vec software. Because these parameters were left to their defaults, the resulting vectorizations are 50-dimensional.

Finally, note a contrast to the tf-idf vectorization scheme described in Procedure 3.6.1.b. There, a vectorized document had a clear interpretation to each of its components because each component was exactly defined as a tf-idf value. By contrast, the components of this vectorization scheme are not interpretable. These vectorizations are simply the optimal (in the sense that the vectorizations minimize the objective function prescribed by the word2vec machine learning software) way to encode the mock-documents of Table 3.1 as vectors.

Procedure 3.6.2 Example with Mock-Documents

Consider the mock-documents in Table 3.1. Recall that $d_{i,j}$ represents the j^{th} document from the i^{th} forum, f_i . To match the notation of Procedure 3.6.2, let a single subscript describe the mock-documents in the natural ordering of Table 3.1 (e.g., $d_1 = d_{1,1}, \dots, d_4 = d_{1,4}, d_5 = d_{2,1}, \dots, d_7 = d_{2,3}, d_8 = d_{3,1}, d_9 = d_{3,2}$). Suppose each of these documents are to be vectorized. Table 3.4 shows vectorizations resulting from Procedure 3.6.2. As a reminder, the doc2vec vectorizations are not interpreted. The vectorizations are displayed for completeness only (subsequent examples build upon these vectorizations).

Table 3.4: The doc2vec vectorizations of the mock-documents in Table 3.1 produced by Procedure 3.6.2. The doc2vec vectorizations are not interpreted. These vectorizations optimally minimized the objective function prescribed by the word2vec machine learning software (see Appendix E for details).

\mathbf{d}_i (or, $\mathbf{d}_{i,j}$)								
i (or, i, j)								
1 (1,1)	2 (1,2)	3 (1,3)	4 (1,4)	5 (2,1)	6 (2,2)	7 (2,3)	8 (3,1)	9 (3,2)
-0.537	0.285	-0.100	-0.228	0.102	2.129	1.188	-1.043	-0.115
0.554	0.926	0.345	0.149	2.281	2.096	1.610	0.340	1.130
-0.373	-0.283	-0.654	-0.533	0.629	1.454	1.665	0.857	1.371
0.876	-0.599	0.264	-0.255	0.982	0.361	0.083	0.729	1.184
0.440	0.477	0.564	0.290	2.122	1.013	1.094	0.940	0.947
1.074	-0.711	0.395	0.083	1.185	0.690	-0.243	-0.516	-0.420
0.951	1.181	0.563	0.608	1.863	1.387	1.576	2.226	1.554
-1.726	-1.470	-1.757	-1.497	-0.872	-0.563	-1.364	-0.868	-0.525
0.530	-0.129	0.453	0.450	-1.509	-2.160	-1.189	-0.272	-0.950
0.962	0.359	1.284	0.585	3.090	1.965	-0.172	0.490	0.358
1.025	0.726	0.461	0.981	0.248	0.708	1.012	0.870	0.933
-1.124	-0.933	0.227	-0.315	-0.102	-1.187	-1.223	-0.213	-0.964
-0.863	-0.252	-0.929	-0.788	-0.477	-0.762	0.517	-0.867	-0.363
0.589	1.550	0.958	0.603	1.540	1.056	1.365	0.370	0.786
-0.057	-0.343	-0.118	-0.430	0.308	0.746	1.421	2.173	1.434
1.392	0.745	2.292	1.779	0.837	1.230	-0.427	-0.076	0.123
-0.193	0.276	0.014	-0.111	0.076	0.940	0.876	-0.368	0.314
-0.822	-0.955	-1.142	-0.948	-0.574	-1.635	-1.988	-1.349	-1.489
-0.075	0.574	0.832	0.771	0.806	0.310	0.337	0.242	0.040
1.043	1.579	1.564	1.155	0.786	-0.070	1.035	1.119	0.541
-1.511	-0.477	-0.975	-0.912	-0.856	-0.299	-0.972	-2.153	-1.719
-1.703	-2.273	-1.242	-1.556	-1.550	-0.927	-0.647	-1.997	-1.802
-1.482	-0.758	-1.098	-1.696	0.361	-0.155	0.263	-0.979	-1.061
-0.097	-0.214	-0.490	-0.134	-0.376	-0.101	-0.002	-0.009	0.332
-2.545	-2.012	-2.375	-2.951	-1.392	-0.921	-0.664	-1.154	-1.717
0.101	0.459	-0.012	0.235	0.077	-0.046	0.660	-0.249	-0.117
0.916	0.944	0.950	0.459	1.200	0.921	1.726	0.374	0.723
0.570	0.049	0.228	0.714	0.619	0.455	0.267	0.664	0.624
0.077	-0.350	-0.187	0.081	-0.122	-0.381	-0.192	-0.645	-0.569
0.163	0.765	-0.133	0.400	0.671	1.341	0.407	0.435	1.064
-0.635	-0.013	-0.180	0.360	-0.769	-1.319	-0.422	0.404	-0.716
0.307	0.129	0.811	0.288	-0.848	-0.140	0.326	0.197	0.315
-0.608	-0.450	-0.408	-0.474	0.502	0.584	-0.202	-0.907	-0.981
-2.340	-2.563	-2.874	-2.203	-0.889	-1.033	-1.338	-0.249	-0.862
-1.180	-1.193	-0.904	-0.504	-1.058	-0.698	0.097	-0.383	-0.779
1.366	2.113	1.601	2.066	0.225	0.932	1.149	0.173	0.419
-0.421	-1.499	-0.330	-0.312	-0.460	-0.469	-1.443	-1.762	-1.867
0.359	-0.232	-0.646	-0.307	1.068	1.172	0.960	-0.411	-0.554
-0.148	-0.387	0.618	0.244	0.070	-0.694	-0.758	-0.176	0.058
0.536	0.085	0.395	0.477	-0.130	0.639	-0.096	-0.787	-0.066
-1.209	-0.982	-1.611	-1.971	-1.090	0.435	-0.283	-0.560	-0.279
-1.024	-0.878	-1.108	-1.334	-0.937	-1.565	-2.845	-1.868	-1.626
0.904	0.910	0.483	0.811	-0.366	-0.488	0.699	1.143	1.757
-0.384	0.924	-0.065	-0.516	0.149	0.268	0.517	0.718	0.795
-1.024	-0.888	-0.451	0.043	-0.050	-0.346	-0.650	-0.980	-1.502
0.236	0.293	0.561	0.055	0.306	0.746	0.796	-2.224	-1.547
-0.099	-0.101	-0.101	1.008	0.171	1.165	0.152	-0.080	1.001
1.704	1.886	0.987	1.661	-0.041	0.370	0.167	0.549	0.695
-1.054	0.063	-1.022	-0.429	-0.826	-0.347	0.694	-0.151	-0.366
1.006	0.988	1.064	1.207	0.910	-0.388	-0.555	1.212	1.012

3.7 Distance Measures of Text Documents

From vectorizations of text documents, this section describes how relative similarities between text documents were established. It does so by first introducing relevant notation and terminology, and then describing the two document distance measures used in this thesis.

Relevant Notation and Terminology

Matching the notation for vectorized documents, vectors are also represented by boldface letters. For vectors \mathbf{x} and \mathbf{y} , their distance is represented by $\text{distance}_X(\mathbf{x}, \mathbf{y})$ with the X subscript specifying the distance measure chosen. For vectors \mathbf{x} and \mathbf{y} , define their *inner-product* as:

$$\mathbf{x}^T \mathbf{y} = \sum_i (\mathbf{x})_i (\mathbf{y})_i.$$

Likewise, define the α -norm as:

$$\|\mathbf{x}\|_\alpha = \left(\sum_i ((\mathbf{x})_i)^\alpha \right)^{1/\alpha},$$

where $\alpha > 0$. See Appendix D for the terminology and notation used throughout this thesis.

3.7.1 Choices of Distance Measures

The two distance measures used in this thesis were the Jaccard distance and the Euclidean distance, both of which are common in analyzing text data (Huang, 2008). For vectors \mathbf{x} and \mathbf{y} , their Jaccard distance is defined as:

$$\text{distance}_{\text{Jaccard}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{(\|\mathbf{x}\|_2)^2 + (\|\mathbf{y}\|_2)^2 - \mathbf{x}^T \mathbf{y}}.$$

The Jaccard distance is a weighted measure of the similarity in vector components, on a scale from 0 to 1. That is, $\text{distance}_{\text{Jaccard}}(\mathbf{x}, \mathbf{y}) = 0$ when $\mathbf{x} = \mathbf{y}$, and $\text{distance}_{\text{Jaccard}}(\mathbf{x}, \mathbf{y}) = 1$ when \mathbf{x} and \mathbf{y} are orthogonal.

Likewise, for vectors \mathbf{x} and \mathbf{y} , their Euclidean distance is defined as:

$$\text{distance}_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2.$$

The Euclidean distance is the usual straight line distance between points.

3.7.2 Distance Measure / Vectorization Pairs

Distance measures were only applied to specific vectorizations: the Jaccard distance was used exclusively with the tf-idf vectorizations (Section 3.6.1), and the Euclidean distance was used exclusively with the doc2vec vectorizations (Section 3.6.2). This was done because the Jaccard distance is a measure of weighted shared information (Huang, 2008) and therefore requires interpretability in vector components (which the doc2vec vectorizations do not have — see Section 3.6.2).

Distance Calculation Example with Mock-Documents

Consider again the mock-documents in Table 3.1, as well as the tf-idf vectorizations and doc2vec vectorizations in Table 3.3 and Table 3.4, respectively. Recall that the vectorizations in Table 3.3 are of the monthly aggregated mock-documents $\bar{d}_{f,m}^p$ (see the example of Procedure 3.6.1.a), whereas the vectorizations in Table 3.4 are of the mock-documents separately. As discussed in Section 3.7.2, the Jaccard distance applies to the vectorizations of Table 3.3, and the Euclidean distance applies to the vectorizations of Table 3.4. The results of such distance calculations are displayed in Table 3.5 and Table 3.6, respectively.

Table 3.5: The Jaccard distances between the monthly aggregated mock-document tf-idf vectorizations shown in Table 3.3. Distances range from 0 to 1.

	$\bar{\mathbf{d}}_{f_1, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_1, \text{Feb}}^p$	$\bar{\mathbf{d}}_{f_2, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_2, \text{Feb}}^p$	$\bar{\mathbf{d}}_{f_3, \text{Jan}}^p$	$\bar{\mathbf{d}}_{f_3, \text{Feb}}^p$
$\bar{\mathbf{d}}_{f_1, \text{Jan}}^p$		0.723	0.952	1.000	1.000	1.000
$\bar{\mathbf{d}}_{f_1, \text{Feb}}^p$	0.723		1.000	1.000	1.000	1.000
$\bar{\mathbf{d}}_{f_2, \text{Jan}}^p$	0.952	1.000		0.479	1.000	1.000
$\bar{\mathbf{d}}_{f_2, \text{Feb}}^p$	1.000	1.000	0.479		1.000	1.000
$\bar{\mathbf{d}}_{f_3, \text{Jan}}^p$	1.000	1.000	1.000	1.000		0.743
$\bar{\mathbf{d}}_{f_3, \text{Feb}}^p$	1.000	1.000	1.000	1.000	0.743	

Table 3.6: The Euclidean distances between the mock-document doc2vec vectorizations shown in Table 3.4.

	$\bar{\mathbf{d}}_1$	$\bar{\mathbf{d}}_2$	$\bar{\mathbf{d}}_3$	$\bar{\mathbf{d}}_4$	$\bar{\mathbf{d}}_5$	$\bar{\mathbf{d}}_6$	$\bar{\mathbf{d}}_7$	$\bar{\mathbf{d}}_8$	$\bar{\mathbf{d}}_9$
$\bar{\mathbf{d}}_1$		0.100	0.056	0.060	0.227	0.314	0.323	0.241	0.189
$\bar{\mathbf{d}}_2$	0.100		0.101	0.092	0.267	0.293	0.213	0.243	0.179
$\bar{\mathbf{d}}_3$	0.056	0.101		0.045	0.241	0.355	0.367	0.309	0.275
$\bar{\mathbf{d}}_4$	0.060	0.092	0.045		0.313	0.387	0.376	0.287	0.249
$\bar{\mathbf{d}}_5$	0.227	0.267	0.241	0.313		0.130	0.261	0.307	0.254
$\bar{\mathbf{d}}_6$	0.314	0.293	0.355	0.387	0.130		0.148	0.426	0.247
$\bar{\mathbf{d}}_7$	0.323	0.213	0.367	0.376	0.261	0.148		0.260	0.180
$\bar{\mathbf{d}}_8$	0.241	0.243	0.309	0.287	0.307	0.426	0.260		0.060
$\bar{\mathbf{d}}_9$	0.189	0.179	0.275	0.249	0.254	0.247	0.180	0.060	

The four smallest values in Table 3.5 are:

$$\begin{aligned}
 \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_2, \text{Jan}}^p, \bar{\mathbf{d}}_{f_2, \text{Feb}}^p) &= 0.479 = \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_2, \text{Feb}}^p, \bar{\mathbf{d}}_{f_2, \text{Jan}}^p) \\
 \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_1, \text{Jan}}^p, \bar{\mathbf{d}}_{f_1, \text{Feb}}^p) &= 0.723 = \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_1, \text{Feb}}^p, \bar{\mathbf{d}}_{f_1, \text{Jan}}^p) \\
 \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_3, \text{Jan}}^p, \bar{\mathbf{d}}_{f_3, \text{Feb}}^p) &= 0.743 = \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_3, \text{Feb}}^p, \bar{\mathbf{d}}_{f_3, \text{Jan}}^p) \\
 \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_1, \text{Jan}}^p, \bar{\mathbf{d}}_{f_2, \text{Jan}}^p) &= 0.952 = \text{distance}_{\text{Jaccard}}(\bar{\mathbf{d}}_{f_2, \text{Jan}}^p, \bar{\mathbf{d}}_{f_1, \text{Jan}}^p).
 \end{aligned}$$

In the first three cases, the vectorizations came from the same mock-forum (in order, mock-forum f_2, f_1, f_3). The fourth case is of vectorizations coming from mock-forums f_1 and f_2 , but the distance between them is almost the maximal 1. The largest value in Table 3.5 is 1, and many document pairs have this distance. This happens because tf-idf vectorizations are often sparse (i.e., they contain many zero components), which in turn implies that their inner product will be the sum of many zeros. When sparse vectorizations have relatively few components, it is increasingly likely that their inner product will be the sum of only zeros, as is often the case with the vectorizations in Table 3.3. If the inner product of two vectorizations is zero, then the Jaccard distance between them will be 1.

Likewise, the four smallest values in Table 3.6 are:

$$\begin{aligned}
 \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_3, \bar{\mathbf{d}}_4) &= 0.045 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_4, \bar{\mathbf{d}}_3) \\
 \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_3) &= 0.056 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_3, \bar{\mathbf{d}}_1) \\
 \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_4) &= 0.060 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_4, \bar{\mathbf{d}}_1) \\
 \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_8, \bar{\mathbf{d}}_9) &= 0.060 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_9, \bar{\mathbf{d}}_8).
 \end{aligned}$$

Again, in each case, the vectorizations came from mock-documents coming from the same mock-forum (in order, mock-forum f_1, f_1, f_1, f_3). On the other hand, the four largest values in Table 3.6

are:

$$\begin{aligned} \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_4, \bar{\mathbf{d}}_6) &= 0.387 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_6, \bar{\mathbf{d}}_4) \\ \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_4, \bar{\mathbf{d}}_7) &= 0.376 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_7, \bar{\mathbf{d}}_4) \\ \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_3, \bar{\mathbf{d}}_7) &= 0.367 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_7, \bar{\mathbf{d}}_3) \\ \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_3, \bar{\mathbf{d}}_6) &= 0.355 = \text{distance}_{\text{Euclidean}}(\bar{\mathbf{d}}_6, \bar{\mathbf{d}}_3). \end{aligned}$$

In each of these, the vectorizations came from mock-documents coming from different mock-forums (mock-forum f_1 and f_2 in all four cases).

3.8 Clustering of Text Documents

From document vectorizations and their pairwise distances, this section introduces the concept of document clustering. Let $\mathbf{d}_1, \dots, \mathbf{d}_V$ be a collection of vectorized documents. In this thesis, these vectorized documents come from Procedure 3.6.1.b and Procedure 3.6.2 of Section 3.6. Given all the pairwise distances $\text{distance}_X(\mathbf{d}_i, \mathbf{d}_j)$, clustering is an approach that groups similar documents into groupings. There are many such ways to do so (see Estivill-Castro (2002) which discussed the large extent of clustering algorithms). This thesis, however, only dealt with hierarchical clustering algorithms. The algorithms were implemented using the `stats::hclust` function from the R *stats* package (R Core Team and Contributors Worldwide, 2022a), and the results were visualized with dendrograms and heatmaps (see Figure 3.2 – Figure 3.5 for illustrations of dendrograms, see Figure 3.6 – Figure 3.7 for illustrations of heatmaps, and see Section 3.9 on visualization methods in general). Note that heatmaps cluster not only vectors, but the components of the vectors. As such, vector component interpretability is necessary for heatmaps. Therefore, as discussed in Section 3.6.1 and Section 3.6.2, heatmaps can only visualize clusters of tf-idf vectorizations. This likewise implies — as discussed in Section 3.7.2 — that only the Jaccard distance will be used in conjunction with heatmaps.

3.8.1 Hierarchical Clustering

Broadly, hierarchical clustering algorithms fall into two categories: agglomerative (bottom-up)

and divisive (top-down). Agglomerative procedures begin with each data point in its own cluster, and then aggregates the two closest. These two are then considered as a new data point. The process continues iteratively by aggregating the two closest data points until all are clustered together. On the other hand, divisive procedures begin with all data points together. It then splits them into two disjoint groups. The algorithm recursively does this to each of the disjoint groups until each data point is in a cluster alone. See [Roux \(2018\)](#) for a more extensive comparison on the differences between these two types. See also [Carlsson and Mémoli \(2010\)](#) for a more comprehensive overview of hierarchical clustering in general.

Mathematically, there is no universally best hierarchical clustering algorithm ([Kleinberg, 2002](#)). Therefore, like vectorizations and distance measures, two different hierarchical clustering algorithms were used in this thesis.

3.8.2 Choices of Hierarchical Clustering Algorithms

This thesis utilized two well established clustering methods — Ward linkage ([Ward, 1963](#)) and average linkage ([Sokal and Michener, 1958](#)) — following the pattern in [El-Hamdouchi and Willett \(1989\)](#) whose work compared four of the most popular agglomerative hierarchical clustering methods on text based data. Their work compared the two algorithms used in this thesis, as well as single linkage ([McQuitty, 1957](#)) and complete linkage ([Sneath, 1957](#)).

The Ward, average, single, and complete linkage algorithms aggregate clusters of vectors by computing a distance between each pair of clusters, and then combining the two closest clusters. For clusters $X = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_X}\}$ and $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_{n_Y}\}$, the Ward, average, single, and complete linkage algorithms compute the distance between X and Y as:

$$\begin{aligned} \text{distance}_{\text{Ward}}(X, Y) &= \text{ESS}(X \cup Y) - \text{ESS}(X) - \text{ESS}(Y), \\ \text{distance}_{\text{average}}(X, Y) &= \frac{1}{n_X \cdot n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \|\mathbf{x}_i - \mathbf{y}_j\|_2, \\ \text{distance}_{\text{single}}(X, Y) &= \min_{\mathbf{x}_i \in X, \mathbf{y}_j \in Y} (\|\mathbf{x}_i - \mathbf{y}_j\|_2), \\ \text{distance}_{\text{complete}}(X, Y) &= \max_{\mathbf{x}_i \in X, \mathbf{y}_j \in Y} (\|\mathbf{x}_i - \mathbf{y}_j\|_2), \end{aligned}$$

where, for a set $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_{n_Z}\}$:

$$\text{ESS}(Z) = \sum_{i=1}^{n_Z} \left(\left\| \mathbf{z}_i - \frac{1}{n_Z} \sum_{j=1}^{n_Z} \mathbf{z}_j \right\|_2 \right)^2$$

is the *error sum of squares*.

[El-Hamdouchi and Willett \(1989\)](#) concluded that, in general, average linkage performs best, followed by Ward linkage, then single linkage, and lastly complete linkage. This is the basis for the two clustering algorithms used in this thesis. In addition, the exclusion of single linkage was further motivated by its tendency to produce chaining clusters (i.e., if a, b, c, d , and e are data points to be grouped, then a chained cluster would be a and b grouped together, then c grouped into that cluster, then d grouped into that cluster, then e grouped into that cluster). This is a well known issue with single linkage and is acknowledged by [El-Hamdouchi and Willett \(1989\)](#) and [Aggarwal and Zhai \(2012\)](#).

3.8.3 Quantities of Clusters

Hierarchical clustering algorithms in general, and the two used in this thesis specifically, produce clusters of clusters. That is, each data point is considered as its own cluster, which then group with other data points to form a larger cluster, and so on. Therefore, hierarchical clustering algorithms do not finish by concluding an optimal quantity of clusters.

The R *NbClust* package ([Charrad et al., 2022](#)) can extend these hierarchical clustering algorithms by concluding the optimal quantity of clusters based upon 30 different criteria, of which this thesis used 18. Certain criteria were excluded because they required user choice (e.g., graphical methods to determine the optimal number of clusters). Others were excluded because they were incompatible with the document vectorizations (see Appendix [G](#) for more details about the criteria of the R *NbClust* package and those that were used for this thesis).

Note that the Jaccard distance is only compatible with four of the 18 R *NbClust* criteria used throughout this thesis (specifically, the mcclain, cindex, silhouette and dunn criteria). As such, only doc2vec vectorizations (and hence the Euclidean distance) were considered in conjunction with the R *NbClust* package (as discussed in Section [3.7.2](#), doc2vec vectorizations were paired with the

Euclidean distance).

The following two procedures were used in conjunction with the R *NbClust* package. Procedure 3.8.3.a creates documents from randomly selected forum posts. This was done to strip away time dependencies to determine the general patterns of how the various forums of this thesis clustered together. Procedure 3.8.3.b creates documents of forum posts according to date. This was done to establish daily similarities within a forum.

Procedure 3.8.3.a Let F be a set of forums. Let a and b be positive integers. The integers a and b specify that for each $f \in F$, a documents will be created, each from b randomly selected posts. For each $f \in F$, create a documents by:

- Randomly selecting ab posts from f : $d_{f,1}, \dots, d_{f,ab}$ and then concatenating the cleaned documents $\bar{d}_{f,1}, \dots, \bar{d}_{f,ab}$ as follows:

$$\begin{aligned}
 \tilde{d}_{f,1} &= \bar{d}_{f,1} \odot \dots \odot \bar{d}_{f,b} && \text{(first group of } b \text{ documents)} \\
 \tilde{d}_{f,2} &= \bar{d}_{f,b+1} \odot \dots \odot \bar{d}_{f,2b} && \text{(second group of } b \text{ documents)} \\
 &\vdots && \\
 \tilde{d}_{f,a} &= \bar{d}_{f,(a-1)b+1} \odot \dots \odot \bar{d}_{f,ab} && \text{(last group of } b \text{ documents).}
 \end{aligned}$$

Notationally, suppose $\tilde{d}_1, \dots, \tilde{d}_V$ are the resulting $a|F|$ documents (a documents for each forum of F). Produce doc2vec vectorizations of $\tilde{d}_1, \dots, \tilde{d}_V$ according to Procedure 3.6.2. Pass such vectorizations on to the `NbClust::NbClust` function with `distanceEuclidean` and Ward linkage or average linkage specified. Also specify a range of identifiable clusters.

Procedure 3.8.3.b Let f be a forum to be split and aggregated according to date. Suppose that S^f is the set of all documents associated with the fixed forum f . Let

$Y = \{2020-10-01, \dots, 2021-03-31\}$ be the dates of interest. For each $y \in Y$, suppose $d_{f,1}^y, \dots, d_{f,N_y}^y$ are the N_y documents of S^f that occurred on the date y . Let

$$d_{f,y} = d_{f,1}^y \odot \dots \odot d_{f,N_y}^y.$$

Produce doc2vec vectorizations of $d_{f,2020-10-01}, \dots, d_{f,2021-03-31}$ according to Procedure 3.6.2. Pass such vectorizations on to the `NbClust::NbClust` function with `distanceEuclidean` and Ward linkage or average linkage specified. Also specify a range of identifiable clusters.

Hierarchical Clustering Example with Mock-Documents

Consider again the mock-documents in Table 3.1, as well as the tf-idf vectorizations and doc2vec vectorizations in Table 3.3 and Table 3.4, respectively. Recall that the vectorizations in Table 3.3 are of the monthly aggregated mock-documents $\bar{d}_{f,m}^p$ (see the example of Procedure 3.6.1.a), whereas the vectorizations in Table 3.4 are of the mock-documents separately. Then Table 3.5 and Table 3.6 show the pairwise distances between these document vectorizations, respectively. From these vectorizations and pairwise distances, Figure 3.2 – Figure 3.7 show the results of the hierarchical clustering algorithms. Figure 3.2 – Figure 3.5 show dendrograms and Figure 3.6 – Figure 3.7 show heatmaps.

Figure 3.2 and Figure 3.3 show the clustering of the tf-idf vectorizations (and Jaccard distance) with the Ward linkage and average linkage clustering algorithms, respectively. Likewise, Figure 3.4 and Figure 3.5 show the clustering of the doc2vec vectorizations (and Euclidean distance) with the Ward linkage and average linkage clustering algorithms, respectively. As a reminder, the monthly aggregated tf-idf vectorizations were produced with cleaned text documents with stop words removed, and the doc2vec vectorizations were produced with cleaned text documents with stop words not removed (see Procedure 3.6.1.b and Procedure 3.6.2, respectively). Each of Figure 3.2 – Figure 3.5 show documents from the same forum clustering together well.

Figure 3.6 and Figure 3.7 show the clustering of the monthly aggregated tf-idf vectorizations (and Jaccard distance) with the Ward linkage and average linkage clustering algorithms, respec-

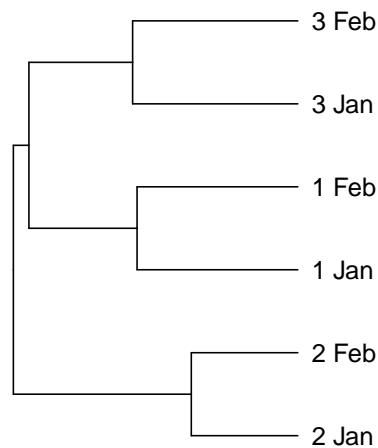


Fig. 3.2: The Ward linkage clustering results for the mock-documents of Table 3.1. The tf-idf vectorizations and pairwise distances are contained in Table 3.3 and Table 3.5, respectively. The “ i Month” leaves correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a.

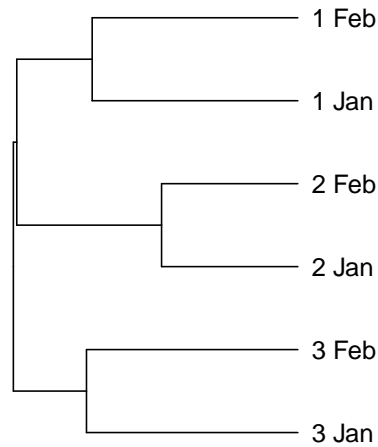


Fig. 3.3: The average linkage clustering results for the mock-documents of Table 3.1. The tf-idf vectorizations and pairwise distances are contained in Table 3.3 and Table 3.5, respectively. The “ i Month” leaves correspond to the monthly aggregated mock-documents $\bar{\bar{d}}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a.

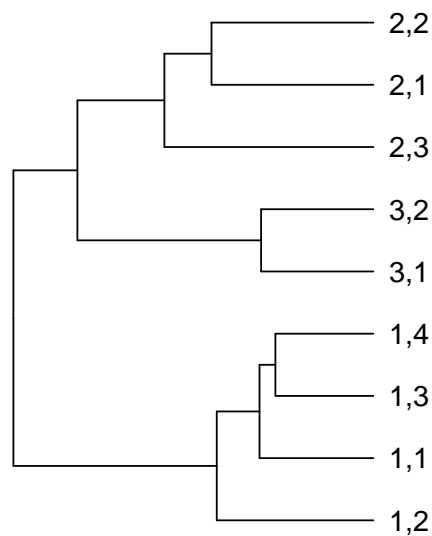


Fig. 3.4: The Ward linkage clustering results for the mock-documents of Table 3.1. The doc2vec vectorizations and pairwise distances of the documents are contained in Table 3.4 and Table 3.6, respectively. The “ i, j ” leaves correspond to the mock-documents $\bar{d}_{i,j}$.

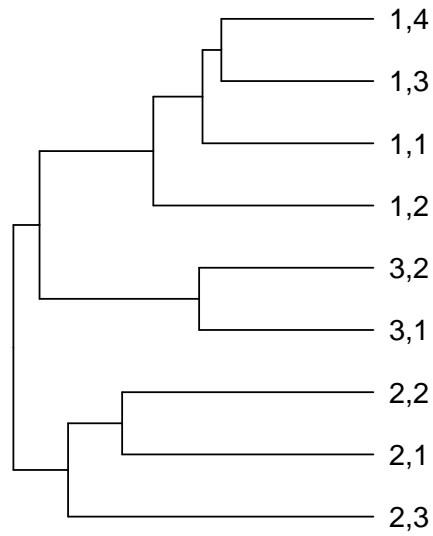


Fig. 3.5: The average linkage clustering results for the mock-documents of Table 3.1. The doc2vec vectorizations and pairwise distances of the documents are contained in Table 3.4 and Table 3.6, respectively. The “ i, j ” leaves correspond to the mock-documents $\bar{d}_{i,j}$.

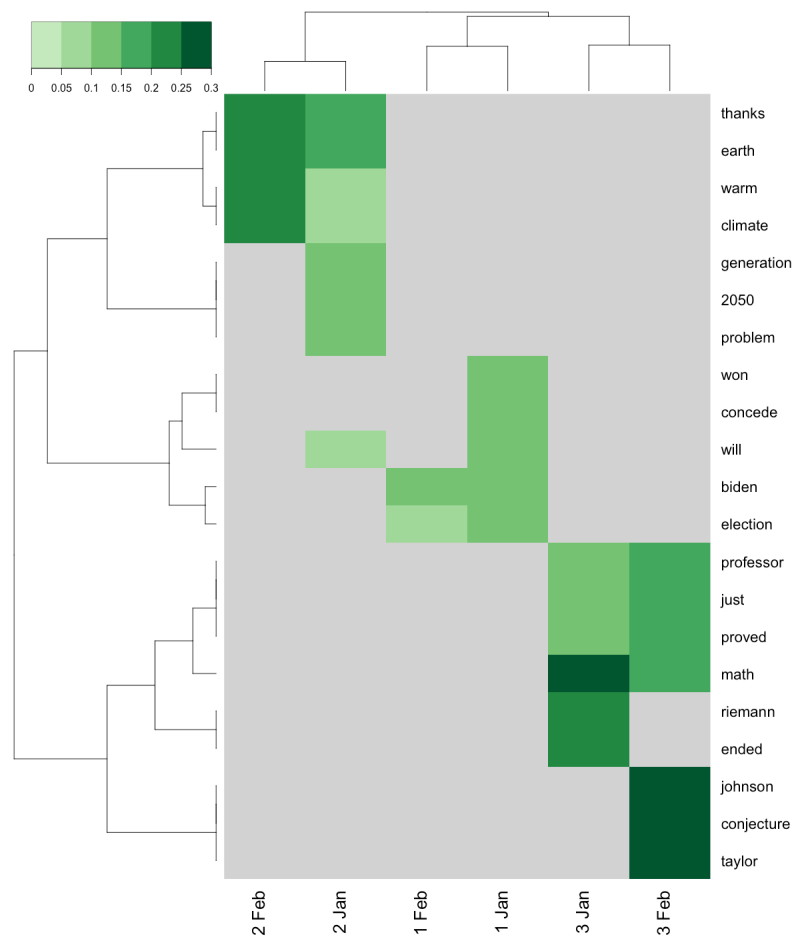


Fig. 3.6: A monthly aggregated mock-document heatmap (Ward linkage) with the vectorizations and pairwise distances displayed in Table 3.3 and Table 3.5, respectively. The tokens on the right indicate the tokens of the tf-idf vectorizations, while the “ i Month” leaves at the bottom correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a. Dendrograms for both the tokens (on the left) and the mock-documents (at the top) are also displayed.

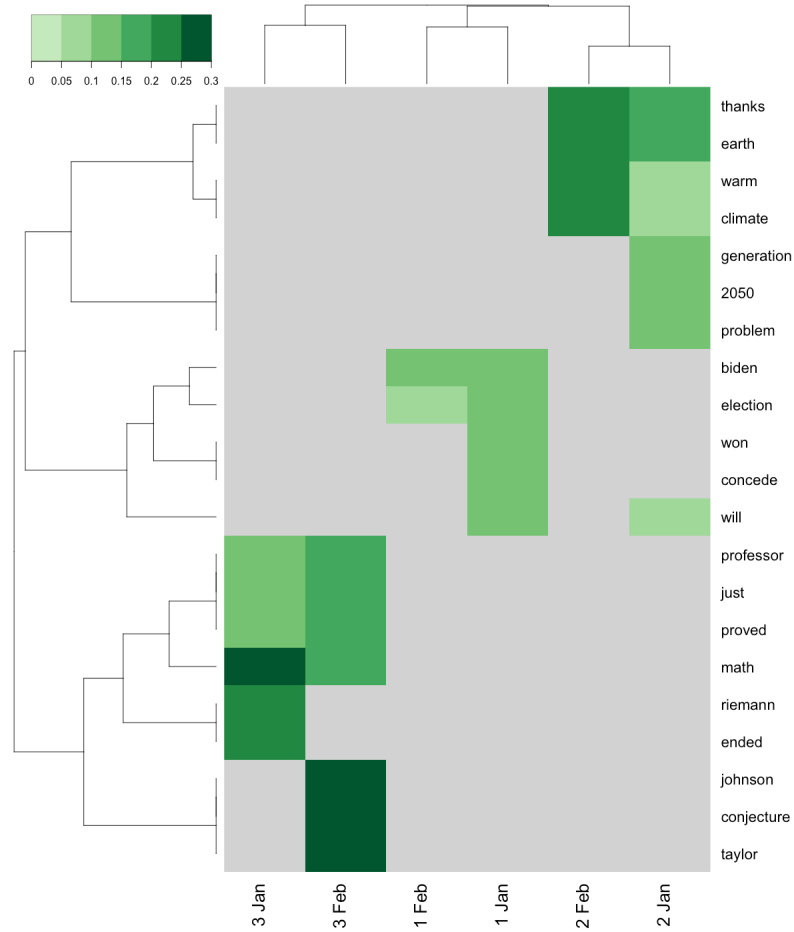


Fig. 3.7: A monthly aggregated mock-document heatmap (average linkage) with the vectorizations and pairwise distances displayed in Table 3.3 and Table 3.5, respectively. The tokens on the right indicate the tokens of the tf-idf vectorizations, while the “ i Month” leaves at the bottom correspond to the monthly aggregated mock-documents $\bar{d}_{f_i, \text{Month}}^p$, as in the example of Procedure 3.6.1.a. Dendrograms for both the tokens (on the left) and the mock-documents (at the top) are also displayed.

From the mock-documents in Table 3.1, as well as their doc2vec vectorizations in Table 3.4, the R *NbClust* package (Charrad et al., 2022) determines the optimal clustering of documents based upon 18 different criteria, as described in Section 3.8.3. Table 3.7 shows the results of such. Specifically, the left-most column specifies the relevant criteria from the R *NbClust* package, the middle columns describe which cluster the relevant mock-document was clustered into, and the final column states the optimal number of clusters, as determined by the criteria. The optimal number of clusters ranges from 1 cluster to 8 clusters (the R *NbClust* package requires that at least two docu-

ments be clustered together, and so the max number of clusters is the number of documents minus one). Figure 3.8 shows a histogram of the optimal number of clusters, as determined by the various different criteria of the R *NbClust* package. The histogram is bimodal with one (larger) peak at three clusters (the number of forums), and another peak at eight clusters (the number of documents minus one).

Table 3.8 and Figure 3.9 mirror Table 3.7 and Figure 3.8, except for the fact that Table 3.8 and Figure 3.9 show the R *NbClust* results where the minimum number of clusters was specified as three. As in Figure 3.8, the two histograms in Figure 3.9 are bimodal with one (larger) peak at three clusters (the number of forums), and another peak at eight clusters (the number of documents minus one).

Table 3.7: The optimal document clustering for each criteria of the R *NbClust* package (criteria are arranged as in the R *NbClust* package). Each criteria could identify 1 – 8 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. Independent of whether the Ward linkage or average linkage clustering algorithm was used, the optimal clustering results were the same, with two exceptions: the *ball* and *mcclain* criteria. The differences are marked in light blue for Ward linkage and light green for average linkage.

Criteria	Mock-Document										# Clusters
	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	$d_{3,1}$	$d_{3,2}$		
kl	1	1	1	1	2	2	2	3	3	3	
ch	1	2	1	1	3	4	5	6	6	6	
hartigan	1	2	3	3	4	5	6	7	8	8	
cindex	1	1	1	1	2	2	2	3	3	3	
db	1	2	3	3	4	5	6	7	8	8	
silhouette	1	2	3	3	4	5	6	7	8	8	
duda	1	1	1	1	2	2	2	3	3	3	
pseudot2	1	1	1	1	2	2	2	3	3	3	
ball (Ward)	1	1	1	1	2	2	2	2	2	2	
ball (average)	1	1	1	1	2	2	2	1	1	2	
ptbserial	1	1	1	1	2	2	3	4	4	4	
gap	1	1	1	1	1	1	1	1	1	1	
mcclain (Ward)	1	1	1	1	2	2	2	2	2	2	
mcclain (average)	1	1	1	1	2	2	2	1	1	2	
gamma	1	1	1	1	2	2	3	4	4	4	
gplus	1	1	1	1	2	2	3	4	4	4	
tau	1	1	1	1	2	2	2	3	3	3	
dunn	1	2	1	1	3	4	5	6	6	6	
sdindex	1	1	1	1	2	3	4	5	5	5	
sdbw	1	2	3	3	4	5	6	7	8	8	

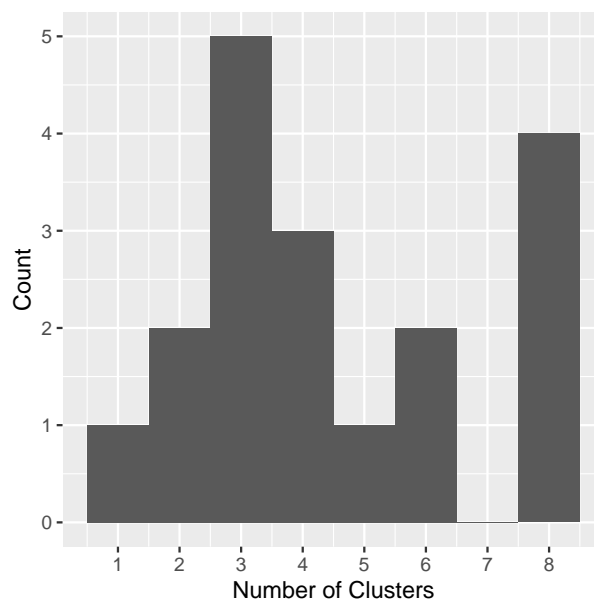


Fig. 3.8: Counts of the optimal number of clusters, as determined by the various criteria of the R *NbClust* package, when each criteria could identify 1 – 8 clusters. The counts are the same independent of whether the Ward linkage or average linkage clustering algorithm was used.

Table 3.8: The optimal document clustering for each criteria of the R *NbClust* package (criteria are arranged as in the R *NbClust* package). Each criteria could identify 3 – 8 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. Independent of whether the Ward linkage or average linkage clustering algorithm was used, the optimal clustering results were the same, with one exception: the *sdindex* criteria. The differences are marked in light blue for Ward linkage and light green for average linkage.

Criteria	Mock-Document									# Clusters
	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	$d_{3,1}$	$d_{3,2}$	
kl	1	1	1	1	2	2	2	3	3	3
ch	1	1	1	1	2	2	2	3	3	3
hartigan	1	2	3	3	4	5	6	7	8	8
cindex	1	1	1	1	2	2	2	3	3	3
db	1	2	3	3	4	5	6	7	8	8
silhouette	1	2	3	3	4	5	6	7	8	8
duda	1	1	1	1	2	2	2	3	3	3
pseudot2	1	1	1	1	2	2	2	3	3	3
ball	1	1	1	1	2	2	3	4	4	4
ptbiserial	1	1	1	1	2	2	3	4	4	4
gap	1	1	1	1	2	2	2	3	3	3
mcclain	1	1	1	1	2	2	2	3	3	3
gamma	1	1	1	1	2	2	3	4	4	4
gplus	1	1	1	1	2	2	3	4	4	4
tau	1	1	1	1	2	2	2	3	3	3
dunn	1	2	1	1	3	4	5	6	7	7
sdindex (Ward)	1	1	1	1	2	2	3	4	4	4
sdindex (average)	1	1	1	1	2	3	4	5	5	5
sdbw	1	2	3	3	4	5	6	7	8	8

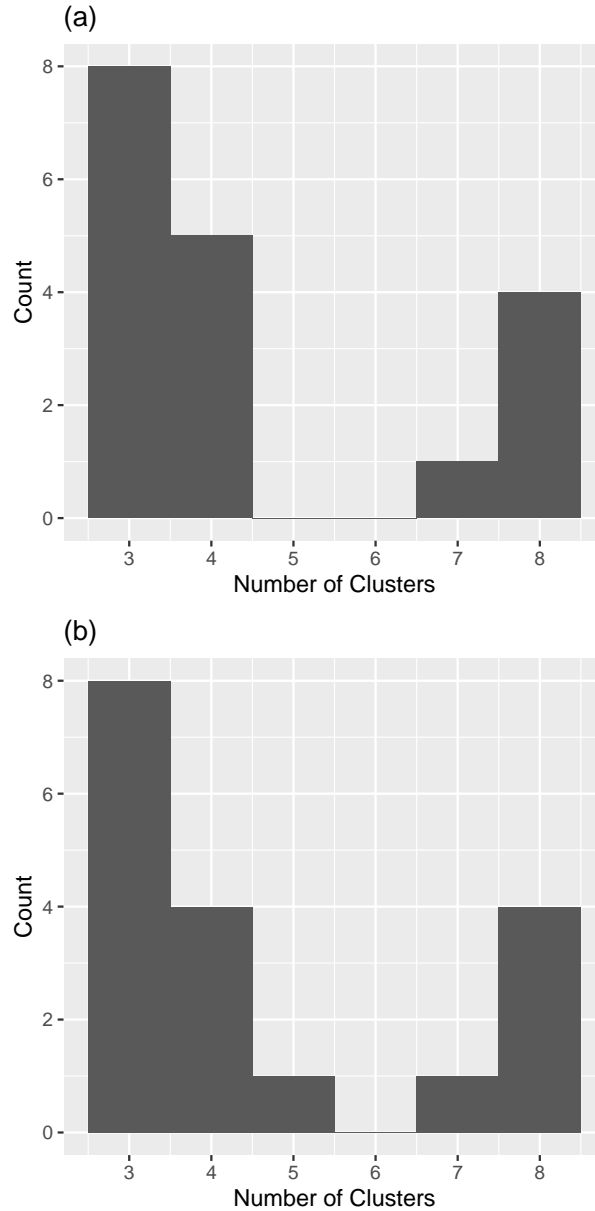


Fig. 3.9: Counts of the optimal number of clusters, as determined by the various criteria of the R *NbClust* package, when each criteria could identify 3 – 8 clusters. Histograms (a) and (b) correspond to the Ward linkage and average linkage clustering algorithms, respectively.

3.9 Visualization Methods

Various different visualizations were used throughout this thesis. Histograms were used as in Figure 3.8 and Figure 3.9 to determine the optimal number of groupings in a hierarchical clustering, and time series plots (Wang et al., 2020a) were used to visualize post frequencies through time.

Likewise, Section 3.8 discussed hierarchical clusterings of text documents. These hierarchical clusterings were visualized with *dendrograms*, as in Figure 3.3 – Figure 3.4. Such visualizations show a hierarchical structure of the mock-documents of Table 3.1. For example, Figure 3.4 shows two main clusters. The first of these were formed from two clusters which consisted of the mock-documents from the first and third mock-forums, respectively. The second of the main clusters were formed from the mock-documents of the second mock-forum. Dendrograms further visualize ‘when’ clusters were formed. For example, the two most similar mock-documents in Figure 3.4 were $\bar{d}_{1,3}$ and $\bar{d}_{1,4}$ as these two documents combined together furthest to the right in the dendrogram. Likewise, $\bar{d}_{1,1}$ was close to that cluster as it aggregated with $\bar{d}_{1,3}$ and $\bar{d}_{1,4}$ shortly later. The documents $\bar{d}_{3,1}$ and $\bar{d}_{3,2}$ were likewise quite similar. On the other hand, the documents from the second mock-forum showed the most intra-dissimilarity, as evident by the fact that they clustered together late (one can see that the cluster of documents from the first mock-forum was fully formed shortly before any two documents from the second mock-forum were clustered together).

Dendrograms and *cluster heatmaps* (see Wilkinson and Friendly (2009)) enhance each other. A cluster heatmap (i.e., a heatmap) shows dendrogram clusters and encodes two-dimensional numeric information as colors/hues, with more extreme numbers corresponding to more extreme colors/hues — see Figure 3.7 and Figure 3.6. Such visualizations allow one to see why clusters were formed as they were. For example, the tokens *professor*, *just*, *proved*, *math*, *riemann*, *ended*, *johnson*, *conjecture*, and *taylor* (found at the bottom right of Figure 3.6) are quite useful in identifying mock-documents coming from the third mock-forum f_3 , as indicated by the darker color/hue at the corresponding token/document intersections.

There are many other ways to visualize text data. The most useful ones were used in this thesis, but alternative visualizations can be found in Section 3.10.7.

3.10 Alternative Methods

3.10.1 Software Products for Data Acquisition

8kun data were acquired via the Internet Archive’s Wayback Machine (Internet Archive, ndc) with the Python *Scrapy* package (Scrapy Developers, 2022a). As discussed in Section 2.2.1, the

use of an archive (as opposed to working with the 8kun website directly) was necessary because (a) 8kun has a history of being removed from the publicly accessible internet ([Zamani et al., 2019](#)), and (b) the retrieval of dated information on 8kun’s website would have been difficult as there was no known way of easily filtering 8kun posts by date. Many internet archives exist (e.g., [archive.today](#) ([nd](#)) and [Perma.cc](#) ([nd](#))), but the choice of the WBM for an archive was made because it is “the largest publicly accessible archive in the world” ([Bowyer, 2021](#)). Likewise, many alternatives to the Python *Scrapy* package exist (e.g., the Python *BeautifulSoup* ([Richardson, 2022](#)), Python *lxml* ([lxml Dev Team, 2022](#)), R *RCrawler* ([Khalil and Fakir, 2017](#)), R *RCurl* ([Temple Lang, 2022](#)), and R *rvest* ([Wickham, 2022](#)) packages. However, the Python *Scrapy* package was chosen over these alternatives because of its reputation (see [Scrapy](#) ([nd](#)) which lists over 50 companies currently using the Python *Scrapy* package).

Reddit data were acquired with the Pushshift data set ([Baumgartner et al., 2022](#)) and API ([Baumgartner et al., 2019](#)) with the associated Python *PSAW* package ([Marx, 2018](#)), as well as Reddit’s API ([Reddit, 2022b](#)) with the associated Python *PRAW* package ([Boe, 2022](#)). These software products were used because no viable alternatives could be found.

3.10.2 Software Products for Data Analysis

There are many alternatives to the chosen data analysis packages. For example, the R *languageR* ([Baayen and Shafaei-Bajestan, 2019](#)) and the R *koRpus* ([Michalke, 2021](#)) packages both provide text analysis support. However, the R *stringr* and R *tidytext* packages were chosen over these alternatives for their intentional compatibility with the tidyverse — “an opinionated collection of R packages designed for data science [which] share an underlying design philosophy, grammar, and data structures” ([Tidyverse, nd](#)).

Likewise, the R *cluster* package ([Maechler et al., 2022](#)) is an alternative to the base functionality found within the `stats::hclust` function. However, the base functionality within R was defaulted to because R is a well established and dependable software for data analysis. The R *clusterCrit* package ([Desgraupes, 2018](#)) is an alternative to the R *NbClust* package ([Charrad et al., 2022](#)). The R *NbClust* package was chosen because it has been tested in its intended usage (see

[Charrad et al. \(2014\)](#)).

3.10.3 Important Statistics of Text Documents

Section 3.5 discussed text document statistics which isolate key tokens within a document. There are, however, other ways in which to identify key tokens. For instance, one could identify tokens with consistent presences within a corpus of documents. That is, for a set of tokens D and corpus C , identify tokens that satisfy $t \in d$ for every document in C (or some subset of C). However, such a process finds common tokens that are not necessarily of interest (e.g., stop tokens or other uninteresting tokens, such as *always*, *can*, and *more*). As such, it is difficult to determine tokens that are common, but also special or unique.

Likewise, one could determine key tokens by simply identifying tokens with large unweighted token counts — $\text{ntokens}(d, t)$ — within a document. This, however, is a poor choice when there is a corpus of documents to consider, especially when documents are of much different sizes. That is, the number of occurrences of a specified token in one document may be drastically smaller than another, but only because the first document contains far fewer tokens in total. This may prompt one to instead determine key tokens by simply looking at token proportions. This too is ineffective when a larger corpus of documents is considered. Key tokens are ideally those that can be used to identify differences in documents. But just because a certain token appears frequently in one document, that does not mean that that token is special to that document (each document may use that token frequently). See [Schwartz et al. \(2013\)](#) for more pitfalls to these simple, but rudimentary, approaches.

Likewise, one could modify the definition of tf into a weighted tf. For example, [Manning et al. \(2008\)](#) defined:

$$\text{w-tf}(d, t) = \begin{cases} 1 + \log(\text{tf}(d, t)), & \text{if } \text{tf}(d, t) > 0 \\ 0, & \text{otherwise} \end{cases}$$

for a document d and token t . [Manning et al. \(2008\)](#) also defined a normalized tf: for a document d and token t :

$$\text{n-tf}(d, t) = a + (1 - a) \cdot \frac{\text{tf}(d, t)}{\max_{\tau \in d}(\text{tf}(d, \tau))},$$

where a “is a *smoothing* term whose role is to damp the contribution of the second term.” In total, Manning et al. (2008) presented five modifications to tf.

Likewise, within the definition of tf-idf presented in Section 3.5, one could modify the

$$\ln \left(\frac{|C|}{\sum_{d_i \in C} [t \in d_i]} \right)$$

scaling term (commonly called the *inverse document frequency*, or *idf*, term). Manning et al. (2008) presented three such modifications which scale the tf of a token. One modification is to scale tf values uniformly (essentially returning raw token frequencies), and the others modify the threshold for when the idf term becomes zero (essentially zeroing out the modified tf-idf value of a token if the token appears in a specified number of documents within the corpus).

However, Manning et al. (2008) noted that Zobel and Moffat (1998) “failed to lead to any conclusions on the best [of these] weighting schemes” and Zobel and Moffat (1998) originally stated: “It is obvious from these results that there is no measure that is a clear winner.”

This failure to determine the best scheme — along with an application of Occam’s Razor* — was the basis for keeping tf-idf in the unmodified form presented in this thesis.

3.10.4 Vectorizations of Text Documents

This thesis considered two vectorization schemes. Procedure 3.6.1.b produces vectorizations of tf-idf scores and Procedure 3.6.2 produces doc2vec vectorizations. One could modify underlying definitions and hyper-parameters to produce different vectorizations. For instance, the modifications to tf and tf-idf presented in Section 3.10.3 would result in different vectorizations. This was not done for the same reasons as described in Section 3.10.3. Likewise, the word2vec machine learning software — from which doc2vec vectorizations are produced — has changeable hyper-parameters. These hyper-parameters were not tuned in this thesis because (a) the default hyper-parameters follow the advice of Mikolov et al. (2013a) and Mikolov et al. (2013b), (b) the defaults give positive results, and (c) this thesis is not meant to research the word2vec software in depth.

*“Given two explanations of the data, all other things being equal, the simpler explanation is preferable” (Blumer et al., 1987).

Outside of these alternatives, documents may be vectorized in many different ways. For example, rather than vectors of tf-idf scores, one could have a vector of tf scores, or a Boolean vector of token presences (i.e., for a specified dictionary $D = \{t_1, \dots, t_r\}$ of r tokens and a document d , a vectorization \mathbf{d} could be defined by $(\mathbf{d})_i = 1$ if $t_i \in d$ and 0 otherwise, for $i = 1, \dots, r$). More complicated vectorizations can also be defined. For example, [Wu et al. \(2018\)](#) defined a vectorization scheme inspired by the word mover’s distance in [Kusner et al. \(2015\)](#).

These alternative methods, however, were not used in this thesis because tf-idf and doc2vec are ubiquitous and well-known throughout the literature. Indeed, they are a standard by which other vectorizations are often compared: [Wang et al. \(2016\)](#) compared their vectorization scheme of linked documents (i.e., a combined document which considers both a base document to be vectorized, as well as the documents to which the base document links) to a tf-idf scheme, and [Wu et al. \(2018\)](#) compared their vectorization scheme to the word2vec software.

3.10.5 Distance Measures of Text Documents

As discussed in Section 3.7, this thesis utilized two primary distance measures (the Jaccard distance and the Euclidean distance). Both of these are common in text analysis ([Huang, 2008](#)). In addition to the two distance measures of this thesis, [Huang \(2008\)](#) also considered the cosine distance, the Pearson Correlation Coefficient and the Averaged Kullback-Leibler Divergence. The definitions of these distances are more complex than the distance measures used in this thesis. Interested readers may refer to [Huang \(2008\)](#) for more information on these definitions.

The choice of the Jaccard distance was made based on the results of [Huang \(2008\)](#) who stated: “On average, the Jaccard and Pearson measures are slightly better in generating more coherent clusters.” The Euclidean distance did not perform well according to [Huang \(2008\)](#). However, the cosine distance did perform well and the squared Euclidean distance is proportional to the cosine distance when vectors are normalized (see Appendix F). That is, the Euclidean distance is a decent (but imperfect, since this thesis did not work with normalized vectors) substitute for the cosine distance. Moreover, the interpretability of the Euclidean distance makes it a strong choice (the Euclidean distance is the usual straight line distance between points).

Still, many other distance measures of text documents exist: for example, the word mover’s distance presented in [Kusner et al. \(2015\)](#) or [Wang et al. \(2020b\)](#) which proposed the Wasserstein-Fisher-Rao document metric. Interested readers may refer to these papers for the details to these distance measures. The respective papers to these document distance measures conclude positive results. However, this thesis did not use these alternative distance measures because of their more complicated nature, whereas both $\text{distance}_{\text{Jaccard}}$ and $\text{distance}_{\text{Euclidean}}$ are easily computable and interpretable.

3.10.6 Clustering Algorithms

As discussed in Section 3.8.2, this thesis used two primary clustering algorithms: Ward linkage ([Ward, 1963](#)) and average linkage ([Sokal and Michener, 1958](#)). The motivation behind this design choice is discussed in Section 3.8.2. To reiterate: [El-Hamdouchi and Willett \(1989\)](#) compared four of the most popular agglomerative hierarchical clustering algorithms — the Ward, average, single, and complete linkage algorithms — on text based data and found that Ward linkage and average linkage performed best. However, other agglomerative hierarchical clustering algorithms exist. For example, the `stats::hclust` supports — in addition to Ward linkage and average linkage — the single, complete, centroid, McQuitty, and median linkage algorithms, as well as a derivative of the Ward linkage algorithm.

Still, the work of [El-Hamdouchi and Willett \(1989\)](#) only compared agglomerative hierarchical clustering algorithms. Divisive clustering algorithms are less common and more time consuming. [Roux \(2018\)](#) stated: “Most papers using hierarchical clusterings employ one of the four popular agglomerative methods” (the four being those tested by [El-Hamdouchi and Willett \(1989\)](#)). The R *cluster* package ([Maechler et al., 2022](#)) also acknowledged the predominance of agglomerative procedures:

[The `cluster::diana` function] is probably unique in computing a divisive hierarchy, whereas most other software for hierarchical clustering is agglomerative.

Moreover, divisive procedures can be computationally intensive. [Roux \(2018\)](#) noted that for n data points, there are $2^{n-1} - 1$ ways to split them into two non-empty and disjoint sets. These are the

reasons that divisive procedures were excluded from analysis.

Non-hierarchical Clustering Algorithms

This thesis only considered hierarchical clustering algorithms. Non-hierarchical clustering algorithms (such as the k -means algorithm — see [MacQueen \(1967\)](#)) were not considered in this thesis, despite having certain advantages (e.g., non-hierarchical clustering algorithms tend to be quicker). This was done because it was possible that natural clusters of clusters would form. For example, there could be a broad cluster of political forums, and within that cluster there could be clusters of forums spanning different political leanings. This hierarchical structure would be lost if non-hierarchical algorithms were used.

3.10.7 Visualization Methods

By no means does Section [3.9](#) encompass all ways to visualize the data of this thesis. For instance, *token clouds* are a quick way to represent the relative usage of tokens in a document. They do this by arranging tokens into a circle with common tokens near the center and in large typefaces. Less common tokens are arranged near the edges in small typefaces. Token clouds also have many variations. For example, [Li and Zhou \(2016\)](#) incorporated geographic location to their token clouds, [Cui et al. \(2010\)](#) arranged tokens by semantic and contextual relationships, and both [Lohmann et al. \(2015\)](#) and [Burch et al. \(2014\)](#) extended token clouds to visualize multiple text documents at once. However, token clouds were not used in this thesis for their ease of misinterpretation. For instance, the scale is ambiguous: important tokens could be depicted by area or height of letters. Likewise, there is no ordering to token clouds as all words appear in a circle. Token clouds were not used because the quantitative information to be visualized can be depicted by heatmaps (and be depicted more accurately). That is, the coloring scheme of a heatmap clarifies the value to be depicted over the ambiguity in a token cloud. The coloring scheme likewise prescribes an ordering that lacked in token clouds. Note that token clouds are more commonly known as *word clouds*. But to keep in line with the terminology of this thesis, they will be referred to as token clouds.

Likewise, [Cleveland \(1993\)](#) introduced Cleveland dot plots as an alternative to bar charts where only the top height of the bar is shown. As such, they can display any quantitative information

similar to that of a bar chart (e.g., token frequencies, tf-idf scores, etc.), but are often preferable to bar charts because they are not as susceptible to the effects of axis truncation. These were not used, however, because displaying such quantitative information can once again be efficiently depicted in heatmaps.

Other techniques include a variation on stem and leaf plots in which a root word is selected and then the distribution of following or antecedent words is depicted (Brath, 2018), or calendar plots which show counts (e.g., post frequencies) through the year in a calendar format (Wang et al., 2020a). Scatterplots can also be used to visualize word embeddings, as in Kessler (2017). For a good overview of alternate text-visualization techniques, see Yang et al. (2008). These alternatives were not used because the visualizations described in Section 3.9 effectively visualized the relevant information.

3.10.8 Qualitative Analysis

This thesis analyzed text documents using a quantitative approach. However, qualitative approaches exist with equal validity (e.g., critical discourse analysis (CDA) (Wodak, 2011) or linguistic anthropology (Salzmann et al., 2012)). Such methods were not used in this thesis because (a) they would require the analysis of individual posts (of which this thesis had millions), and (b) biases can influence the analysis (Norris, 1997).

CHAPTER 4

Results

This chapter presents the results of this thesis for six main forums of study — three from 8kun (newsplus, pnd, and qresearch) and three from Reddit (AskThe_Donald, conservative, and conspiracy) — as well as seven alternative forums of study (conservatives, democrats, Liberal, climate, climatechange, immigration, and math from Reddit). See Section 2.1 for more information on these forums. Results were derived from all data that could be acquired from October 1, 2020, through March 31, 2021, which roughly covers the three months preceding and following January 6, 2021, the date of the US Capitol attack in relation to the 2020 US presidential election. See Section 1.1 for more information about the events of January 6, 2021, and see Section 2.2 and Section 3.2 for more information on data acquisition.

Section 4.1 details summary statistics for each forum. Section 4.2 and Section 4.3 both analyze how different subsets of forum data clustered together. Specifically, Section 4.2 analyzes random subsets of forum data, and Section 4.3 analyzes forum data aggregated by month. The results are based on the tf-idf and doc2vec vectorization procedures described in Section 3.6, the Jaccard and Euclidean distance calculations between such vectorizations described in Section 3.7, and the Ward linkage and average linkage clustering algorithms described in Section 3.8. Then Section 4.4 analyzes specific dates and tokens associated with feelings of isolation and displacement.

4.1 Data Summary Statistics

Table 4.1, Figure 4.1, and Figure 4.2 show summary statistics for each of the six main forums of study. Figure 4.1 and Figure 4.2 distinguish between submissions (original posts), comments (replies to a submission), and posts (a generic term for both submissions and posts). See Section 1.4.2 for the exact definitions of these terms. Analogous results for the alternative forums of study can be found in Appendix H.

Table 4.1, Figure 4.1, and Figure 4.2 are all clear in that newsplus data was scant (likely, this is

due to data inaccessibility and not data absence — see Section 2.2.1 for information on the source of 8kun data). As such, newsplus was excluded from further analysis.

Otherwise, Table 4.1 shows that the quantity of posts from the remaining forums spanned three orders of magnitude: AskThe_Donald and pnd were on the order of 10,000, qresearch was on the order of 100,000, and conservative and conspiracy were on the order of 1,000,000. Most posting data came from comments, not submissions. This is apparent in Table 4.1. It is also apparent due to the fact that the posts and comments lines in Figure 4.1 nearly overlap (likewise for Figure 4.2). Furthermore, (newsplus aside) no forum had zero post dates until after January 31, 2021. Lastly, note that there are spikes to zero in data quantity in Figure 4.1. Within the 8kun forums (newsplus, pnd, and qresearch), any spikes to zero are likely due to data inaccessibility and not data absence. For the Reddit forums (AskThe_Donald, conservative, and conspiracy), these spikes to zero indicate a limitation in the Pushshift data set — namely, that the Pushshift data set did not archive data on these dates.

Table 4.1: For the six main forums of study, summary statistics of the total number of posts, submissions, and comments, as well as summary statistics for the number and percent of dates with no posts.

	Total Number of:			Number of Dates	Percent of Dates
	Posts	Submissions	Comments	with No Posts	with No Posts
AskThe_Donald	20,335	3,831	16,504	5	2.75 %
conservative	1,944,455	102,001	1,842,454	1	0.55 %
conspiracy	2,655,905	97,219	2,558,686	0	0.00 %
newsplus	3,147	314	2,833	111	60.99 %
pnd	22,503	623	21,880	16	8.79 %
qresearch	326,716	549	326,167	0	0.00 %

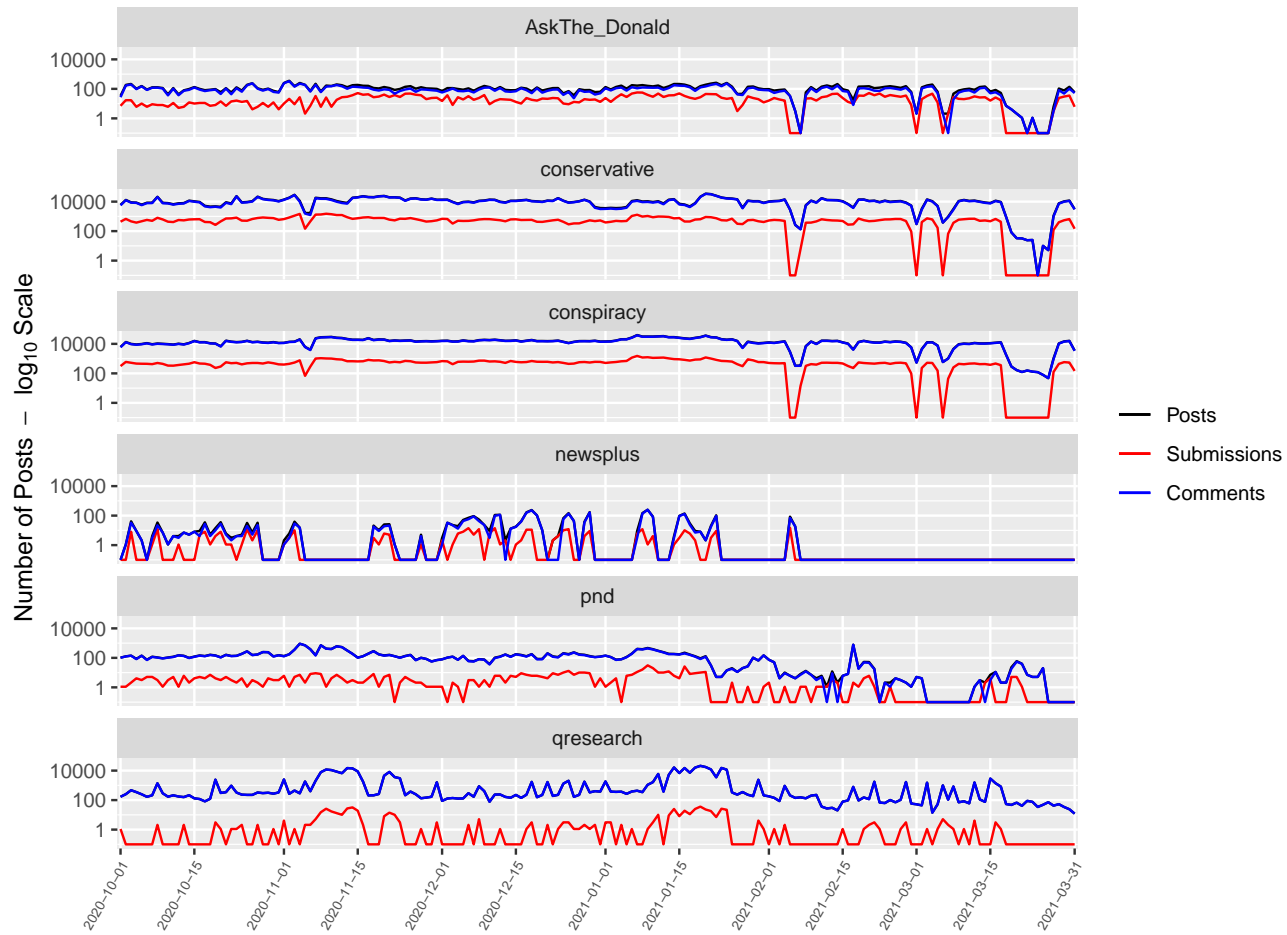


Fig. 4.1: Posts, submissions, and comments per day for each of the six main forums of study (log 10 scale). The black and blue lines (corresponding to posts and comments, respectively) nearly overlap.

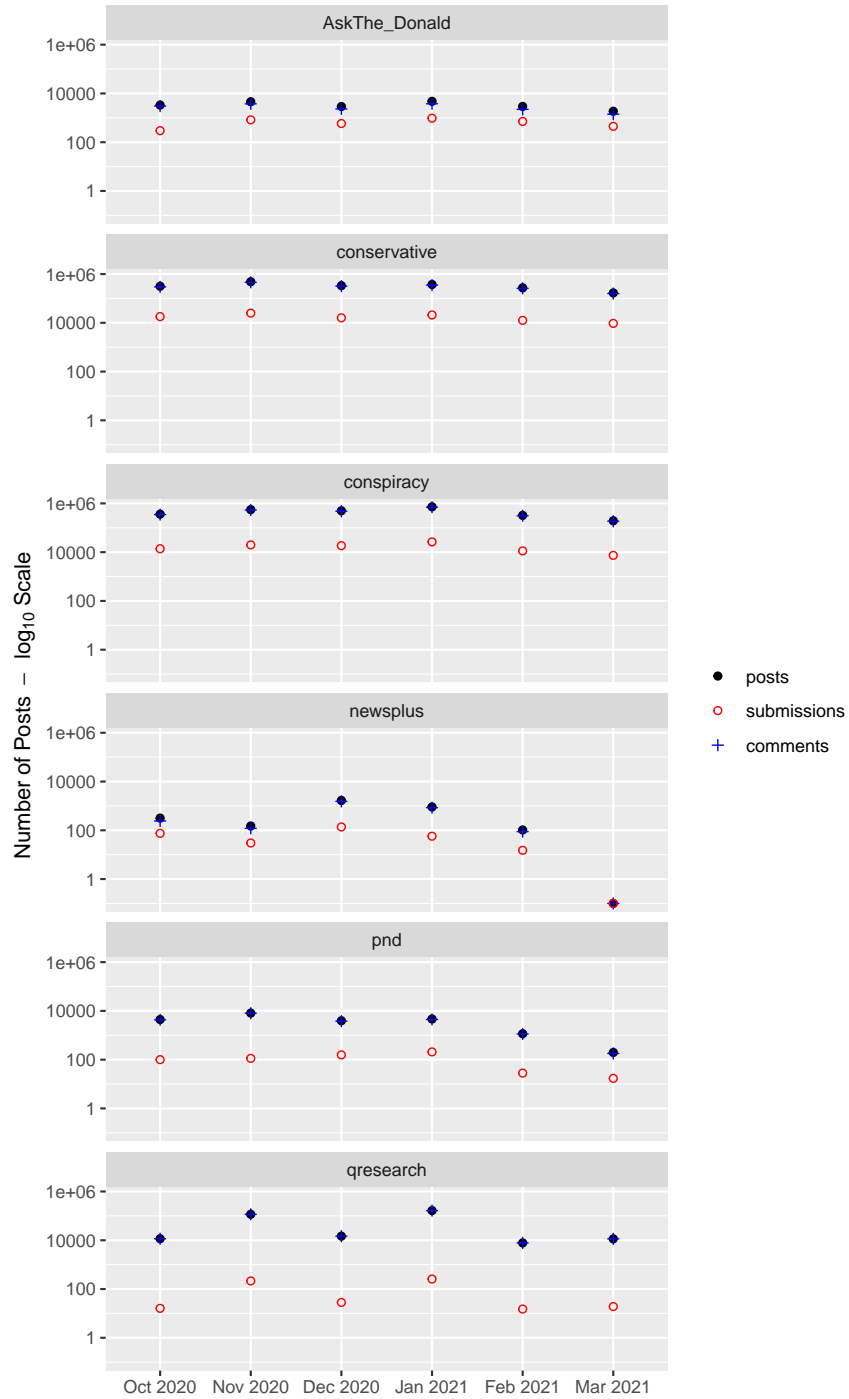


Fig. 4.2: Posts, submissions, and comments per month for each of the six main forums of study (log 10 scale). The black and blue marks (corresponding to posts and comments, respectively) nearly overlap.

4.2 The R *NbClust* Package Clustering Results

This section presents results on how random subsets of forum data intra and inter clustered. Specifically, the results based on Procedure 3.8.3.a with $a = 3$, $b = 1000$, and F being the set of all forums, i.e., the five remaining main forums of study and the seven alternative forums:

$$F = \{\text{AskThe_Donald, conservative, conspiracy, pnd, qresearch,} \\ \text{climate, climatechange, immigration, conservatives,} \\ \text{democrats, Liberal, math}\}.$$

Procedure 3.8.3.a:

- (1) Creates $a = 3$ documents from each forum of F , each of which is the aggregation of $b = 1000$ randomly selected posts. This is done by selecting 3,000 random posts (without replacement and a starting seed of 14741), and breaking such posts into three groups. This ensures that there is no post repetition within the created documents. With three documents per forum, and 12 forums, 36 documents are created.
- (2) Vectorizes these 36 documents according to the doc2vec algorithm (see Section 3.6.2 for information on the doc2vec algorithm).
- (3) Passes such vectorizations onto the `NbClust::NbClust` function, with the Euclidean distance specified, either the Ward linkage or average linkage clustering algorithm specified, and also specifying the minimum number of identifiable clusters as either one or five. This returns 18 possible best groupings based upon the 18 different criteria in the R *NbClust* package. See Section 3.7 for information on the Euclidean distance function, and see Section 3.8 for information on these clustering algorithms, as well as for more information on the `NbClust::NbClust` function. The results when the minimum number of clusters was specified as one (Table 4.2 – Table 4.6) justify the alternate choice of five (Table 4.7 – Table 4.11).

This third step has considerable run time, which is why the first step takes a sample of data from each forum. As stated in (1), there is no possibility of post repetition within the created documents.

This is essential so that the groupings returned by (3) are not due to such post repetition. The $a = 3$ and $b = 1000$ numbers were chosen to provide both non-trivial (and easily readable) results, and also to keep the program run time reasonable.

Table 4.2 – Table 4.6 show both of the results when either Ward linkage or average linkage was used, and each criteria could identify 1 – 35 clusters. In particular, Table 4.2 shows the identified clusters for each criteria (for both Ward linkage and average linkage). Table 4.3 and Table 4.5 are restructurings and visual representations of Table 4.2 (for Ward linkage and average linkage, respectively). Table 4.4 and Table 4.6 show the different quantities of identified clusters, and how many criteria determined that quantity of clusters (for Ward linkage linkage and average linkage, respectively).

Table 4.2: Identified clusters determined by the various criteria of the R *NbClust* package. Each column corresponds to one of three randomly created documents from each forum (seed 14741). Each row corresponds to a criteria (arranged as in the R *NbClust* package), and each criteria was allowed to identify 1 – 35 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. For example, the first row corresponds to the kl criteria. This criteria clustered all random AskThe_Donald, conservative, conspiracy, pnd, conservatives, democrats, and Liberal documents together; it clustered all random qresearch documents together; it clustered all random climate and climatechange documents together; it clustered all random immigration documents together; and it clustered all random math documents together. Most criteria determined the same clusters independent of whether Ward linkage or average linkage was used. The differences are marked in light blue for Ward linkage and light green for average linkage. Except for the duda and pseudot2 criteria, the quantity of identified clusters remains the same regardless of whether Ward linkage or average linkage was used.

Criteria	AskThe_Donald			conservative			conspiracy			pnd			qresearch			climate			climatechange			immigration			conservatives			democrats			Liberal			math			# Clusters
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
kl	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	5	5	5	5	
ch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
hartigan (Ward)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
hartigan (average)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
cindex (Ward)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35
cindex (average)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
db	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	19	20	20	20	21	22	23	24	25	26	27	28	29	30	31	32	
silhouette	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	19	20	20	20	21	22	23	24	25	26	27	28	29	30	31	32	
duda (Ward)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	1	10	10	10	11	11	11	12	12	12	
duda (average)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	9	9	9	10	10	10	
pseudot2 (Ward)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	1	10	10	10	11	11	11	12	12	12	
pseudot2 (average)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	9	9	9	10	10	10	
ball (Ward)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	
ball (average)	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	
ptbiserial	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
gap	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
mcclain (Ward)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	
mcclain (average)	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	
gamma	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
gplus	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
tau	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
dunn	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
sdindex	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	
sdbw (Ward)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
sdbw (average)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	

Table 4.3: A visual representation of the Ward linkage cluster results in Table 4.2, which allowed each criteria to identify 1 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R *NbClust* criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.

	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal		
Criteria	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3			
gap																																				
ball																																				
mcclain																																				
dunn																																				
gamma																																				
gplus																																				
kl																																				
ptbiserial																																				
sdindex																																				
tau																																				
duda																																				
pseudot2																																				
db	*	*	*	*	*	*							*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*				
silhouette	*	*	*	*	*	*							*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*				
ch	*	*	*	*	*	*				*			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*				
cindex	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
hartigan	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
sdbw	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			

Table 4.4: As in Table 4.2 and Table 4.3, the number of criteria which identified a specific quantity of clusters (Ward linkage). Each criteria could identify 1 – 35 clusters.

# Clusters	1	2	5	12	> 30
Count	1	2	7	2	6

Table 4.5: A visual representation of the average linkage cluster results in Table 4.2, which allowed each criteria to identify 1 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R *NbClust* criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.

	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal														
Criteria	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3															
gap																																																
ball																																																
mcclain																																																
dunn																																																
gamma																																																
gplus																																																
kl																																																
ptbiserial																																																
sdindex																																																
tau																																																
duda																																																
pseudot2																																																
db	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*										
silhouette	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									
ch	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									
cindex	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									
hartigan	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									
sdbw	*	*	*	*	*	*													*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*									

Table 4.6: As in Table 4.2 and Table 4.5, the number of criteria which identified a specific quantity of clusters (average linkage). Each criteria could identify 1 – 35 clusters.

# Clusters	1	2	5	10	> 30
Count	1	2	7	2	6

The identified clusters determined by the 18 various R *NbClust* criteria break into roughly five categories (I) – (V), corresponding to the five different quantities of clusters depicted in either Table 4.4 or Table 4.6 (Ward linkage and average linkage, respectively).

(I Ward and average) The gap criteria identified a single cluster containing all documents.

(II Ward) The ball and mcclain criteria identified two clusters, both of which grouped the math, climate, climatechange documents together, and everything else into a separate cluster.

(II average) The ball and mcclain criteria identified two clusters, both of which grouped the qresearch documents together, and everything else into a separate cluster.

(III Ward and average) Seven criteria — the dunn, gamma, gplus, kl, ptbiserial, sdindex, and tau criteria — identified the same five clusters. These five clusters correspond to (a) the documents from the math forum, (b) the documents from the climate and climatechange forums, (c) the documents from the immigration forum, (d) the documents from the qresearch forum, and (e) the documents from pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal forums.

(IV Ward) The duda and pseudot2 criteria identified 12 clusters (the number of forums). Both criteria clustered documents from the same forum into the same grouping, with the exception of the third document from the conservatives forum which was grouped with the AskThe_Donald documents.

(IV average) The duda and pseudot2 criteria identified 10 clusters. The math, climate, climatechange, immigration, qresearch, pnd, and conspiracy documents each formed their own clusters. Of the three remaining clusters: the first consisted of all documents from the conservative forum and two documents from the conservatives forum, the second consisted of all documents from

the AskThe_Donald forum and one document from the conservatives forum, and the third consisted of all documents from both the democrats and Liberal forums.

(V Ward and average) Finally, six criteria — the db, silhouette, ch, cindex, hartigan, and sdbw criteria — identified each document, more or less, as its own cluster (a singleton cluster). The documents that were clustered together (e.g., documents from the climatechange forum, or documents from the immigration forum) came from the same forum. Of these six, the cindex, hartigan, and sdbw criteria only grouped two documents together, which is minimally required by the criteria of the R *NbClust* package.

Overall, while different criteria (and different clustering algorithms) identified different clusters, each criteria still clustered documents in a reasonable manner. For both Ward linkage and average linkage, there was exact cluster agreement for the gap criteria (which clustered everything together), and for the dunn, gamma, gplus, kl, ptbserial, sdindex, and tau criteria, which identified five clusters. These five clusters corresponded to (a) the documents from the AskThe_Donald, conservative, conspiracy, pnd, conservatives, democrats, and Liberal forums, (b) the documents from the qre-search forum, (c) the documents from the climate and climatechange forums, (d) the documents from the immigration forum, and (e) the documents from the math forum.

The previously described results allowed each criteria to identify any number of clusters between 1 – 35. As stated, the dunn, gamma, gplus, kl, ptbserial, sdindex, and tau criteria all identified the exact same five clusters for both Ward linkage and average linkage. Moreover, no other criteria identified a different set of five clusters, and five was the mode quantity of identified clusters. As such, the results were compiled again, this time allowing each criteria to identify any number of clusters between 5 – 35. Table 4.7 – Table 4.11 show these results when either Ward linkage or average linkage was used. In particular, Table 4.7 shows the identified clusters for each criteria (for both Ward linkage and average linkage). Table 4.8 and Table 4.10 are restructurings and visual representations of Table 4.7 for Ward linkage and average linkage, respectively. Table 4.9 and Table 4.11 show the different quantities of identified clusters, and how many criteria determined that

quantity of clusters (for Ward linkage and average linkage, respectively).

Table 4.7: Identified clusters determined by the various criteria of the R *NbClust* package, with five being the minimum number of clusters. Each column corresponds to one of three randomly created documents from each forum (seed 14741). Each row corresponds to a criteria (arranged as in the R *NbClust* package), and each criteria was allowed to identify 5 – 35 clusters. Within each row, documents sharing the same tabular value are clustered together according to that criteria. For example, the first row corresponds to the kl criteria. This criteria clustered all random AskThe.Donald, conservative, conspiracy, pnd, conservatives, democrats, and Liberal documents together; it clustered all random qresearch documents together; it clustered all random climate and climatechange documents together; it clustered all random immigration documents together; and it clustered all random math documents together. Most criteria determined the same clusters independent of whether Ward linkage or average linkage was used. The differences are marked in light blue for Ward linkage and light green for average linkage. Except for the duda and pseudot2 criteria, the quantity of identified clusters remains the same regardless of whether Ward linkage or average linkage was used.

Criteria	AskThe.Donald			conservative			conspiracy			pnd			qresearch			climate			climatechange			immigration			conservatives			democrats			Liberal			math			# Clusters
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
kl	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	5	5	5	5			
ch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
hartigan	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
cindex	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
db	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	19	20	20	20	21	22	23	24	25	26	27	28	29	30	31	32	32
silhouette	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	19	20	20	20	21	22	23	24	25	26	27	28	29	30	31	32	32
duda (Ward)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	10	10	10	11	11	11	11
duda (average)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	9	9	9	10	10	10	10
pseudot2 (Ward)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	10	10	10	11	11	11	11
pseudot2 (average)	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	2	2	1	9	9	9	9	9	9	10	10	10	10
ball (Ward)	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	4	4	4	5	5	5	1	1	1	1	1	1	1	1	1	6	6	6	6
ball (average)	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	1	1	1	1	1	1	1	1	1	6	6	6	6
ptbiserial	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
gap	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
mcclain	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
gamma	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
gplus	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
tau	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
dunn	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
sdindex	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	4	4	4	1	1	1	1	1	1	1	1	1	5	5	5	5
sdbw	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

Table 4.8: A visual representation of the Ward linkage cluster results in Table 4.7, which allowed each criteria to identify 5 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R *NbClust* criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.

	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal		
Criteria	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3			
gap																																				
mcclain																																				
dunn																																				
gamma																																				
gplus																																				
kl																																				
ptbiserial																																				
sdindex																																				
tau																																				
ball																																				
duda																																				
pseudot2																																				
db																																				
silhouette																																				
ch																																				
cindex	*	*	*	*	*	*				*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*				
hartigan	*	*	*	*	*	*				*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
sdbw	*	*	*	*	*	*				*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		

Table 4.9: As in Table 4.7 and Table 4.8, the number of criteria which identified a specific quantity of clusters (Ward linkage). Each criteria could identify 5 – 35 clusters.

# Clusters	5	6	11	> 30
Count	10	1	2	6

Table 4.10: A visual representation of the average linkage cluster results in Table 4.7, which allowed each criteria to identify 5 – 35 clusters. Columns have been reordered to place documents in the same cluster near each other. Rows also have been reordered from the smallest quantity of clusters to the largest quantity of clusters. For each row, the clusters determined by the corresponding R *NbClust* criteria are color-coded. For example, the dunn criteria identified 5 clusters, which have been marked by black, green, yellow, brown, and tan. Asterisks mark singleton clusters.

	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal		
Criteria	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3			
gap																																				
mcclain																																				
dunn																																				
gamma																																				
gplus																																				
kl																																				
ptbiserial																																				
sdindex																																				
tau																																				
ball																																				
duda																																				
pseudot2																																				
db																																				
silhouette																																				
ch																																				
cindex	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
hartigan	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			
sdbw	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*			

Table 4.11: As in Table 4.7 and Table 4.10, the number of criteria which identified a specific quantity of clusters (average linkage). Each criteria could identify 5 – 35 clusters.

# Clusters	5	6	10	> 30
Count	10	1	2	6

The results are not drastically different from when each criteria could identify 1 – 35 clusters. Now, when each criteria could identify 5 – 35 clusters:

- For both Ward linkage and average linkage, the gap and mcclain criteria now identified five clusters, corresponding to the same five clusters in (III Ward and average) above.
- The ball criteria now identified six clusters, which varied slightly between Ward linkage and average linkage.
- For Ward linkage, the duda and pseudot2 criteria identified one less cluster than before (11 now, instead of 12). The one less cluster came from two of the conservatives documents which now clustered with the conservative documents (as opposed to clustering alone). The clusters identified by the duda and pseudot2 criteria remain unchanged for average linkage.
- For Ward linkage, the cindex, hartigan, and sdbw criteria only clustered two documents together. Previously, they clustered two documents from the climatechange forum. Now they clustered two documents from the immigration forum. The cindex, hartigan, and sdbw criteria remain unchanged for average linkage.

Setting Random Seeds

The results of this section were based upon three random samples of data from each forum. These random documents were created with the random seed 14741. The simulations of this section were rerun nine additional times with the random seeds 35753, 10601, 12421, 72227, 15451, 98689, 18181, 97079, and 73037. Naturally, this changed the exact clustering results somewhat. However, these new results qualitatively matched the results presented with the 14741 random seed. The results for the different random seeds are displayed in Appendix I. The full extent of the similarities and differences will be discussed in Section 5.1.

4.3 Heatmaps

The results of this section are based upon heatmaps, as discussed in Section 3.8. Specifically, the heatmaps consist of forum documents aggregated by month, as in Procedure 3.6.1.a, with












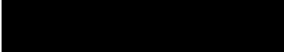
$$F = \{\text{AskThe.Donald, conservative, conspiracy, pnd, qresearch,} \\ \text{climate, climatechange, immigration, conservatives,} \\ \text{democrats, Liberal, math}\}.$$

This creates a corpus consisting of 72 documents, where each document corresponds to one forum during one specific month (six months of data for 12 individual forums). These monthly forum documents were created from all (cleaned with stop tokens removed) forum posts during a specific month, and then tokens which were not used with frequency at least $p = 0.001$ were filtered out. Note that comments do not repeat submission data. The heatmap label for the March qresearch document appears as “qresearch 03” (and likewise for the other forum/month documents). The vectorizations of these 72 documents were created with Procedure 3.6.1.b, which selects η unique tokens with the highest tf-idf values (see Section 3.5 for information on tf-idf values) for the components of the vectorizations. Section 4.3.1 shows heatmap results with up to 200 tokens, and Section 4.3.2 shows heatmap results for token quantities larger than 200.

4.3.1 Heatmaps With $\eta = 50, 100, 150, 200$ Tokens

Figure 4.3 – Figure 4.6 show heatmaps for these 72 monthly documents and $\eta = 50, 100, 150, 200$ unique tokens with the highest tf-idf values, respectively. This implies that if a token appears in Figure 4.3 then it also appears in Figure 4.4 – Figure 4.6 (and likewise, if a token appears in Figure 4.4 then it also appears in Figure 4.5 and Figure 4.6, or if a token appears in Figure 4.5 then it also appears in Figure 4.6). The heatmaps are colored from yellows to reds as tf-idf values increase, and zero tf-idf values are colored grey. The 72 documents are colored according to the forum from which they came, as in Table 4.12. For each of Figure 4.3 – Figure 4.6 the clustering was determined by the Ward linkage algorithm. Appendix J shows the results when average linkage was used, and the results are similar.

Table 4.12: Document label color coding for heatmaps. Specifically, if a document came from a specific forum, then the label for that document was colored as below.

Forum	Color
AskThe_Donald	
conservative	
conservatives	
democrats	
Liberal	
climate	
climatechange	
conspiracy	
pnd	
qresearch	
immigration	
math	

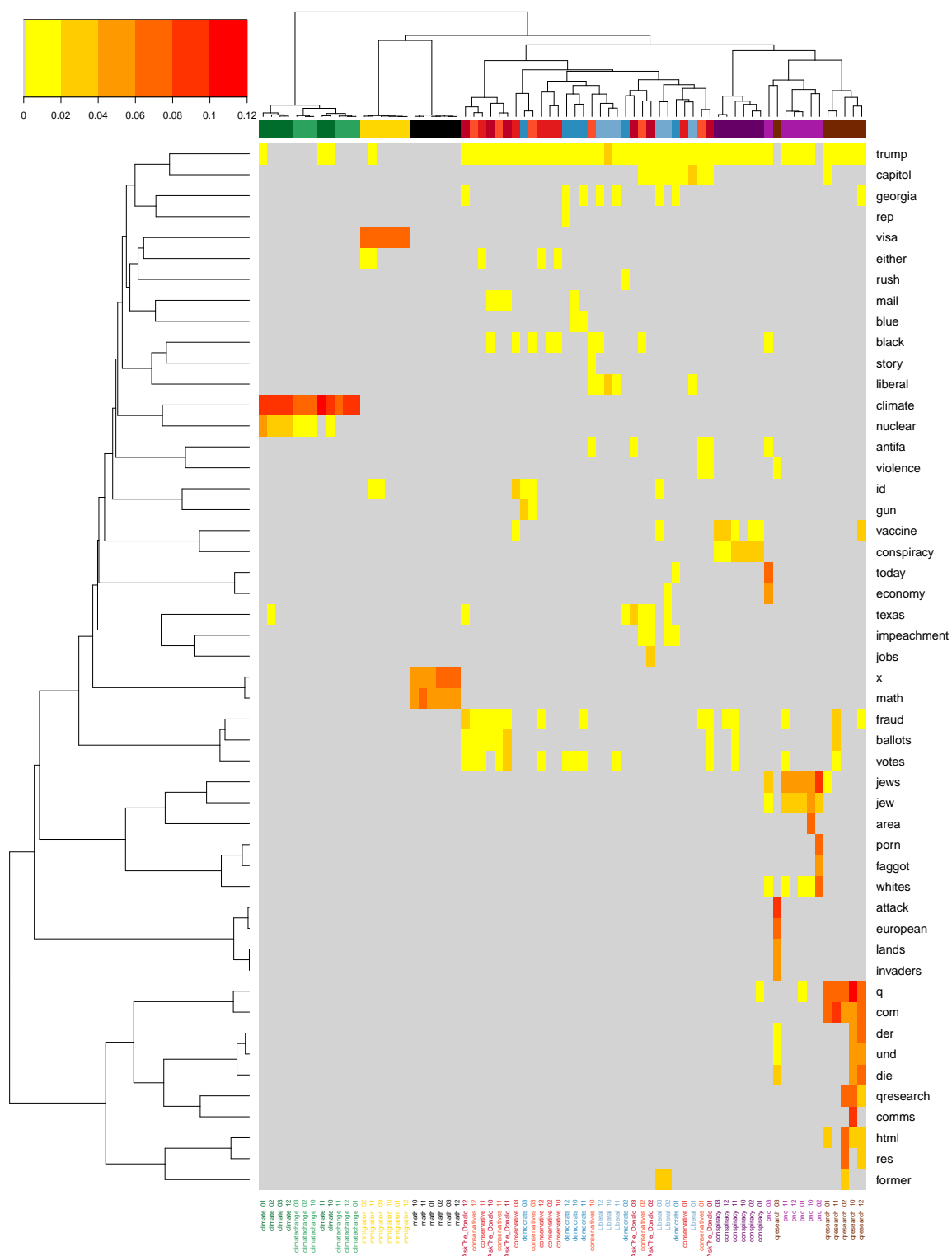


Fig. 4.3: A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

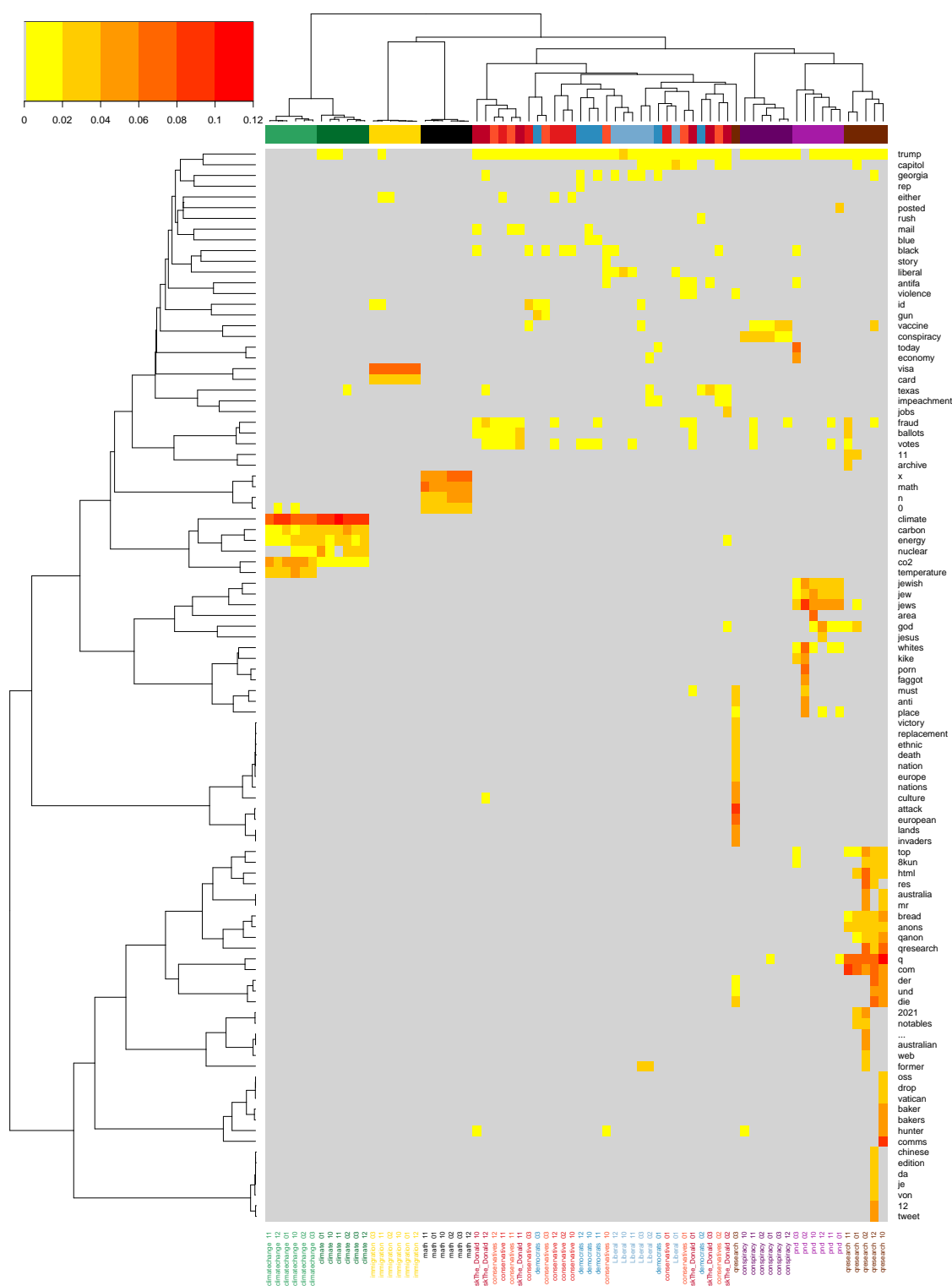


Fig. 4.4: A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

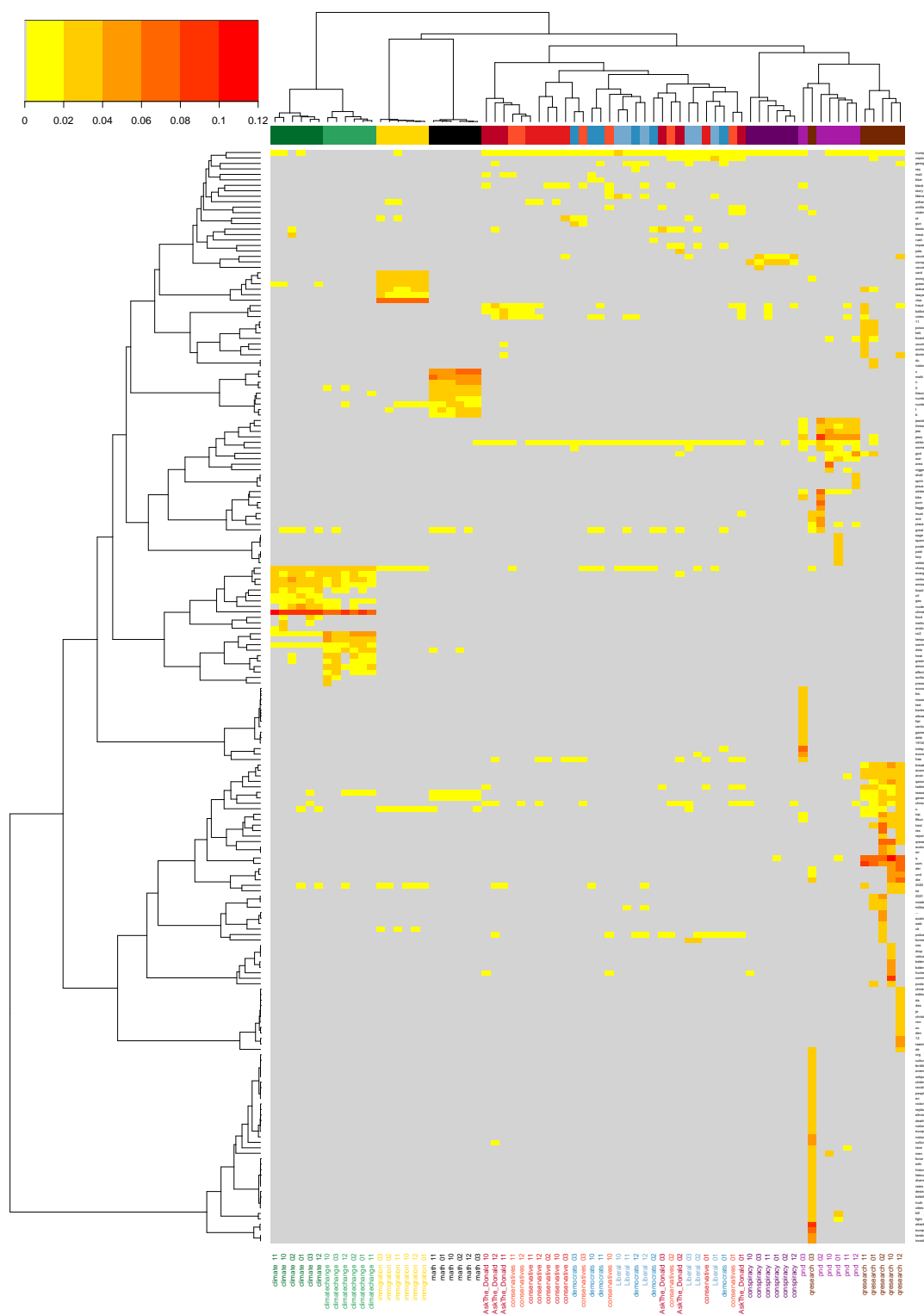


Fig. 4.6: A Ward linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

Figure 4.3 communicates why certain documents were clustered together. For instance, on the left, and about 3/4 up, there is a block of color in the heatmap that corresponds to the climate and climatechange documents and the *climate* and *nuclear* tokens. The heatmap clarifies that the *climate* and *nuclear* tokens were unique to the climate and climatechange documents, as the *climate* and *nuclear* row is completely grey (i.e., zero tf-idf values) aside from the columns corresponding to the climate and climatechange documents. Likewise, there are few other tokens with positive tf-idf values within the climate and climatechange documents (i.e., the columns corresponding to the climate and climatechange documents are mostly grey: only four non-zero tf-idf values can be found in those columns that do not correspond to the *climate* and *nuclear* tokens). Such uniqueness clarifies why the climate and climatechange documents were quickly to be clustered together, but last to be clustered with all other documents, as indicated by the dendrogram at the top. The uniqueness also clarifies why the *climate* and *nuclear* tokens were clustered together before being clustered with any other tokens.

Figure 4.4 – Figure 4.6 show the same documents as Figure 4.3 with a progressively larger set of tokens with the highest tf-idf value ($\eta = 100, 150, 200$ tokens, respectively). The results with η values larger than 200 are shown in Section 4.3.2. With only 50 tokens (Figure 4.3), the math (and immigration) documents are already clearly distinct from documents of different forums. Figure 4.3 does not distinguish the climate and climatechange documents, but the figures with $\eta = 100, 150, 200$ tokens (Figure 4.4 – Figure 4.6, respectively) do distinguish the climate and climatechange documents. However, even with $\eta = 200$ tokens (Figure 4.6), the political forum documents from Reddit (i.e., the documents from the AskThe_Donald, conservative, conservatives, democrats, and Liberal forums) were mostly indistinguishable.

In general, the documents from the 8kun forums (pnd and qresearch) and the non-devoted political Reddit forums (conspiracy, climate, climatechange, immigration, and math) are all distinguishable from the documents coming from the devoted political Reddit forums (AskThe_Donald, conservative, conservatives, democrats, and Liberal). Moreover, these documents from the 8kun and non-devoted political Reddit forums show a great deal of intra-clustering. That is, for example, math documents cluster with math documents. The only exceptions being the March 2021 docu-

ment from the qresearch forum, and the climate/climatechange documents in Figure 4.3. However, the March qresearch document still joined the cluster of pnd documents which then formed with the qresearch documents. Likewise, the climate and climatechange documents still inter-cluster in Figure 4.3, and Figure 4.4 – Figure 4.6 each distinguish the climate documents from the climatechange documents. Among these easily identifiable documents (again, the documents coming from the 8kun forums and the non-devoted political Reddit forums), each had token(s) which were mostly unique to that forum (e.g., *climate* for climate/climatechange, *visa* for immigration, *x* for math, *conspiracy* for conspiracy, *jews* for pnd, *q* for qresearch, and *attack* for March of qresearch).

The documents from the devoted political Reddit forums (AskThe_Donald, conservative, conservatives, democrats, and Liberal) are less distinguishable. However, consistencies can be found through Figure 4.3 – Figure 4.6 in regard to these documents. For example, each of these figures shows a cluster consisting of the October, November, and December documents from AskThe_Donald; the November and December documents from conservatives; and the December document from conservative.

The climate, climatechange, immigration, and math forums each had respective documents that formed into clusters quickly. Likewise, the climate and climatechange documents combined into a single cluster quickly.

Tokens cluster well in Figure 4.3 – Figure 4.6. For example, Figure 4.3 shows the tokens *x* and *math* clustering closely together (and far separate from other tokens). The subsequent figures — Figure 4.4, Figure 4.5, and Figure 4.6 — show the same thing, but with additional tokens in the cluster (e.g., *n*, *0*, and *number*). Likewise, the tokens *fraud*, *ballots*, and *votes* cluster together in each of Figure 4.3 – Figure 4.6. Other examples can be found throughout. Additionally, certain tokens with connotations to race, culture, and the great replacement conspiracy theory* are depicted in Figure 4.3 – Figure 4.6 (e.g., *replacement*, *ethnic*, *culture*, *invaders*). These tokens are mostly isolated to the somewhat anomalous March qresearch document.

*As discussed in Section 1.3, the great replacement conspiracy theory is “a conspiracy theory that postulates white European populations are being demographically and culturally replaced by non-white immigrants through policies enacted by ‘the global elites’” (Carlson and Harris, 2022). Also see Obaidi et al. (2022) or Cosentino (2020) for more information.

Figure 4.3 and Figure 4.4 show two major clusters of tokens: one formed from qresearch specific tokens, and the other formed from all other tokens. Likewise, Figure 4.5 and Figure 4.6 show two major clusters of tokens: one formed from tokens specific to the March qresearch document, and the other formed from all other tokens. This second of the major clusters formed from two clusters: one formed from other qresearch specific tokens, and the other formed from all other tokens. As such, qresearch tokens are quite distinct from other tokens. Most clear clusters of tokens involve tokens that are highly related. For example, consider the major heatmap block located roughly at middle height and on the left of Figure 4.6. The associated tokens cluster together. These tokens include *oil*, *nuclear*, *climate*, *co2*, *warming*, and *greenhouse*.

As stated, the documents from the devoted political Reddit forums (AskThe_Donald, conservative, conservatives, democrats, and Liberal) did not cluster well in Figure 4.3 – Figure 4.6 (or the analogous figures in Appendix J). As such, heatmaps were also created where the underlying corpus only included the documents from these devoted political Reddit forums (Figure 4.7 – Figure 4.10). Figure 4.7 – Figure 4.10 still show the $\eta = 50, 100, 150, 200$ tokens with the highest tf-idf value, respectively (again, results with η values larger than 200 are shown in Section 4.3.2). However, because the underlying corpus changed (from including six document from all forums, to only including six documents from each of AskThe_Donald, conservative, conservatives, democrats, and Liberal), the tf-idf values changed. Therefore, the set of $\eta = 50, 100, 150, 200$ tokens with the highest tf-idf value also changed. Note that the different color scheme in Figure 4.7 – Figure 4.10 (as compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. Document labels are still colored according to the forum from which they came, as in Table 4.12. Figure 4.7 – Figure 4.10 show the results when the Ward linkage clustering algorithm was used. Appendix K shows the results when average linkage was used. There are differences between the Ward linkage and average linkage results, as discussed below.

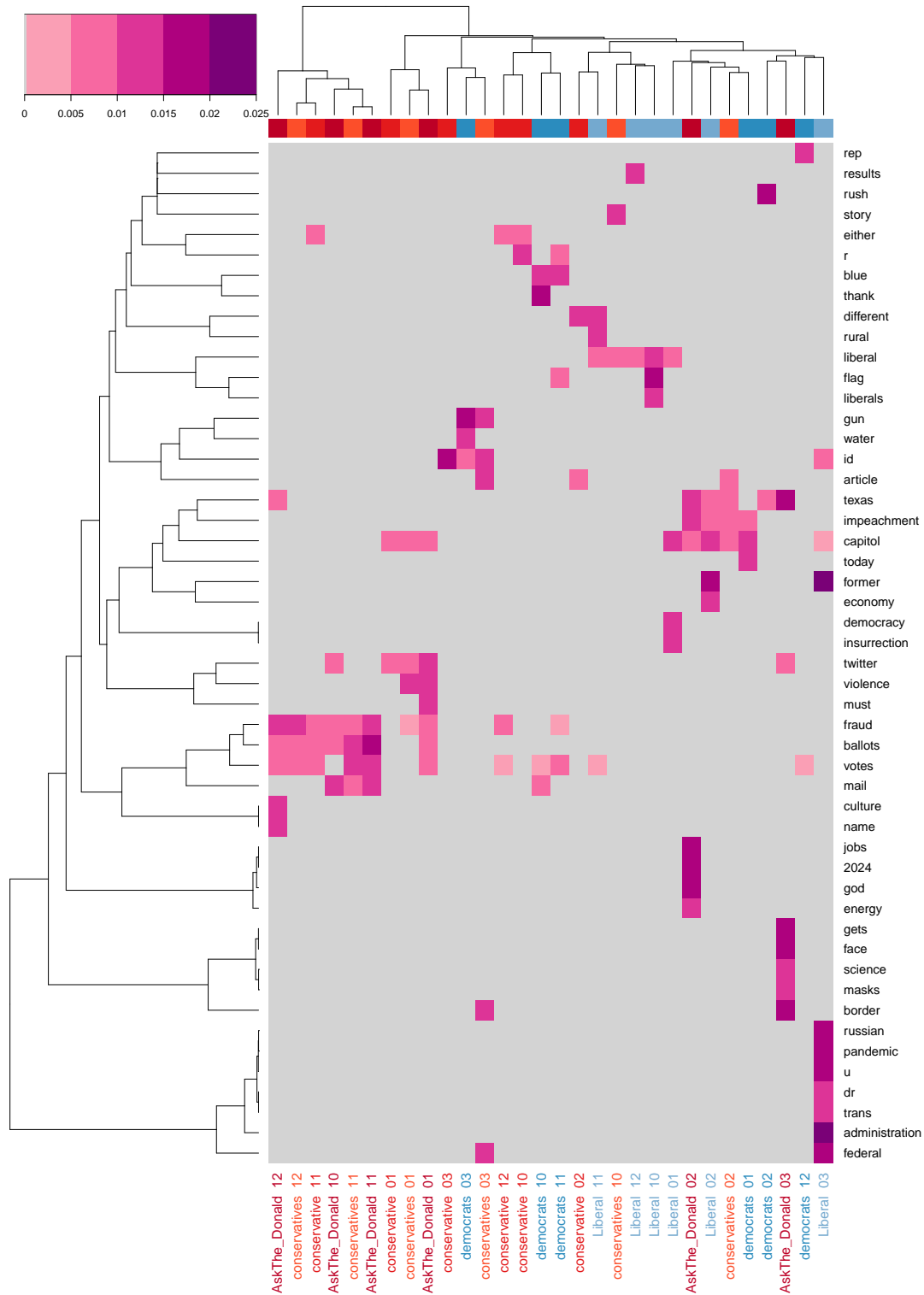


Fig. 4.7: A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.



Fig. 4.8: A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.

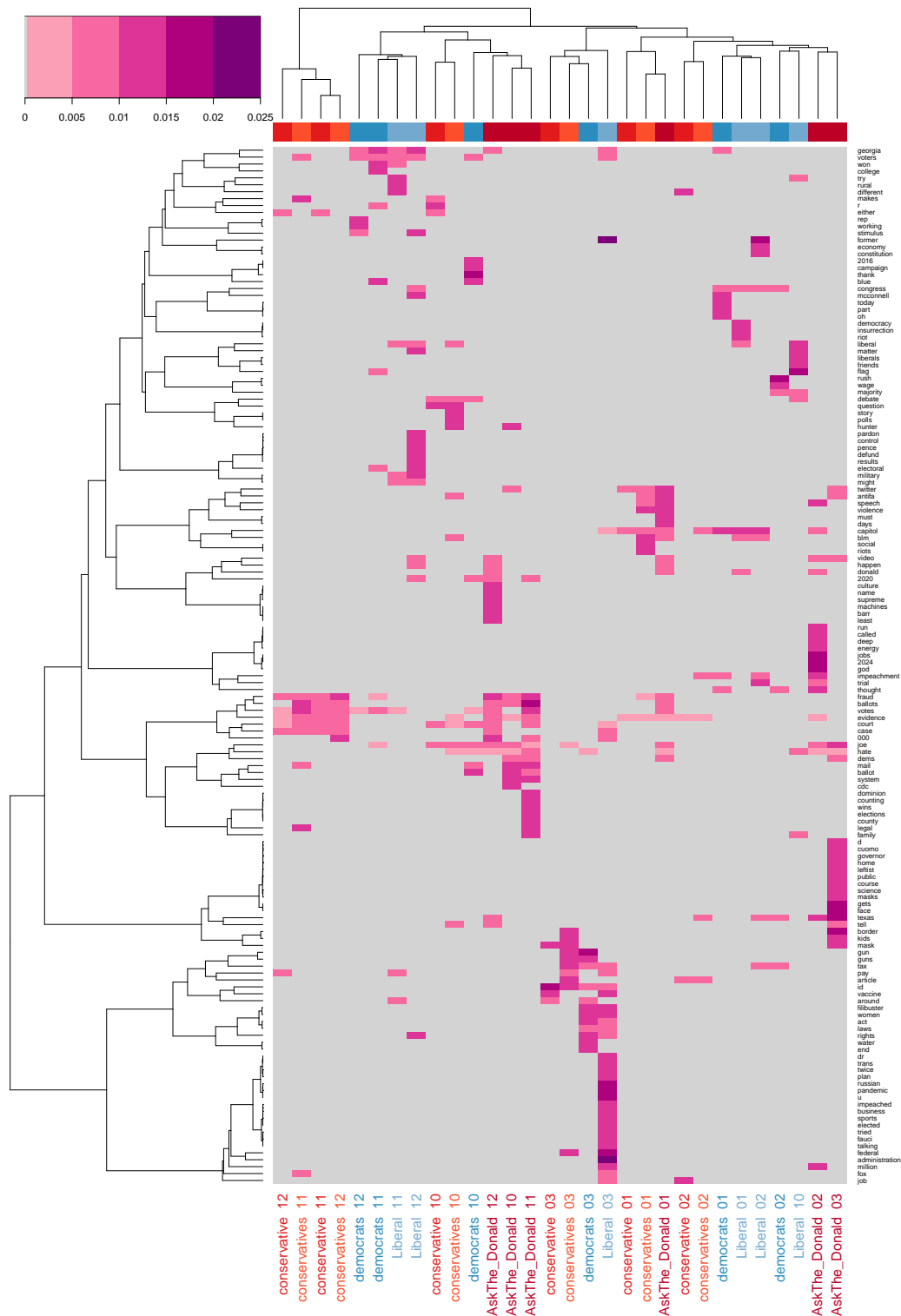


Fig. 4.9: A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.

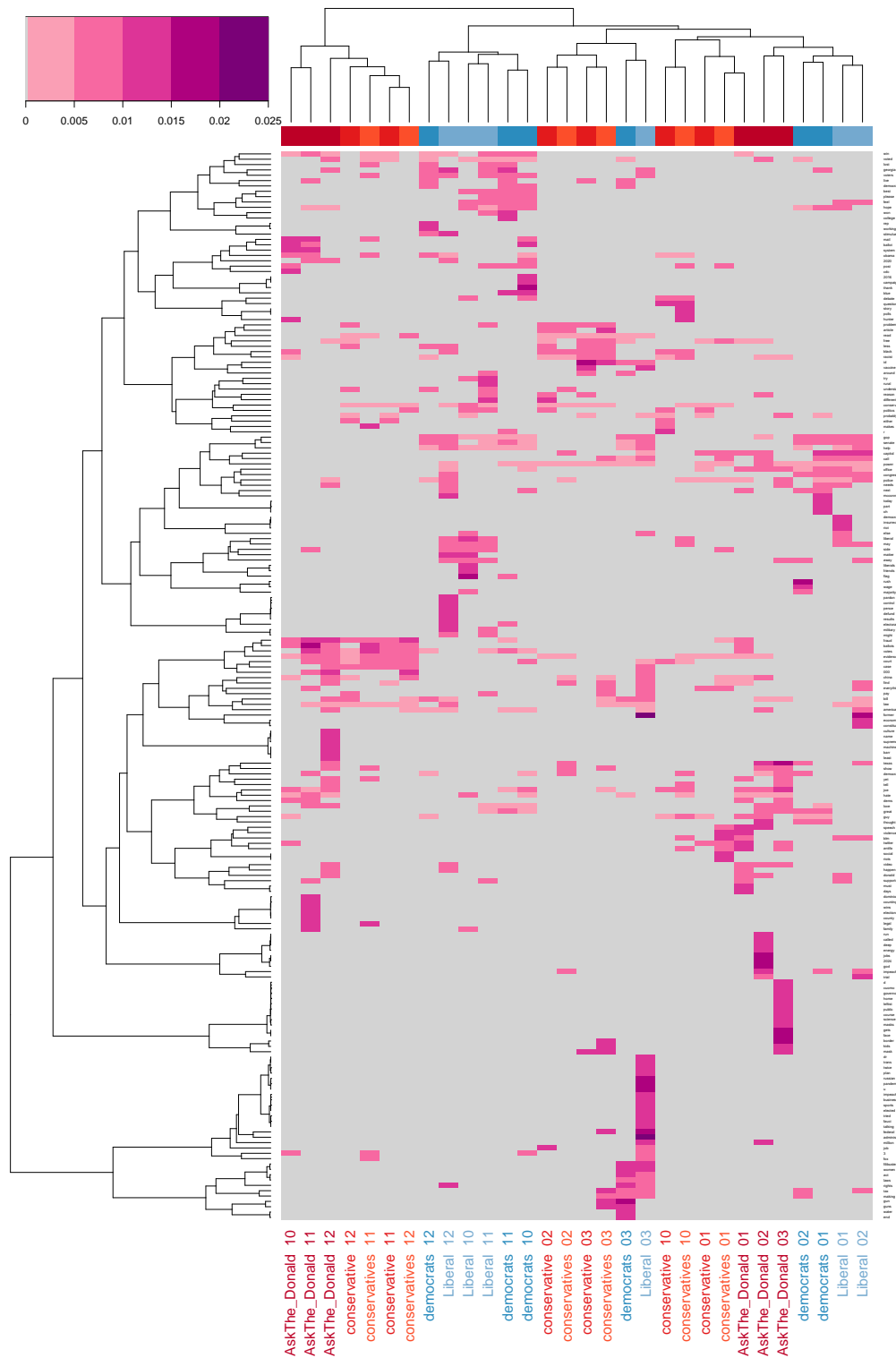


Fig. 4.10: A Ward linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. The different color scheme (compared to Figure 4.3 – Figure 4.6) is used to emphasize the different scale in tf-idf values. More details are provided in the main text.

Though Figure 4.7 – Figure 4.10 only portray documents from the devoted political Reddit forums, the depicted clustering is not considerably more coherent than when all forums were clustered (Figure 4.3 – Figure 4.6 and Appendix J). Documents from one forum still cluster with documents from a different forum, though some consistencies can be found (e.g., the October, November, and December AskThe_Donald documents cluster closely through Figure 4.7 – Figure 4.10). Notably, Figure 4.3 – Figure 4.6 each clustered together the October, November, and December documents from AskThe_Donald; the November and December documents from conservatives; and the December document from conservative. But this does not hold as consistently in Figure 4.7 – Figure 4.10.

The tokens in Figure 4.7 – Figure 4.10 cluster better than the documents. For example, the tokens *cuomo*, *govenor*, *leftist*, *science*, and *masks* all cluster together, as do tokens dealing with discussions of voter fraud (e.g., *fraud*, *ballots*, *votes*, and *counting*) or tokens discussing the events of January 6, 2021 (e.g., *riot*, *insurrection*, and *democracy*). Tokens which could be in reference to feelings of isolation and/or displacement from society (e.g., *replacement*, *ethnic*, *culture*, and *invaders*) did appear in Figure 4.3 – Figure 4.6. However, these tokens were mostly isolated to the March qresearch document.

As mentioned above, there are differences in these heatmaps when Ward linkage was used (Figure 4.7 – Figure 4.10) versus when average linkage was used (Appendix K). For example, Figure 4.10 shows one cluster of seven politically-right documents, another cluster of six politically-left documents, and a final cluster of 11 politically-right documents and six politically-left documents. However, the analogous figure with average linkage (Figure K.4) shows a different story. In Figure K.4, there is a cluster consisting of 10 politically-left documents and one politically-right document, another cluster of 12 politically-right documents, and a final cluster of two politically-left documents and five politically-right documents. In other words, the linkage algorithm influenced the clusters in a considerable way. As such, the Ward linkage and average linkage results together paint a picture that these political documents are difficult to distinguish.

4.3.2 Heatmaps With $\eta > 200$ Tokens

All of the above heatmaps (Figure 4.3 – Figure 4.10 and the analogous figures in Appendix J and Appendix K) utilized $\eta = 50, 100, 150, 200$ tokens. As η increased, more coherent clusters seemed to form. For example, Figure 4.3 (with $\eta = 50$ tokens) did not distinguish the monthly climate documents from the monthly climatechange documents. But with $\eta = 100, 150, 200$ tokens (Figure 4.4 – Figure 4.6, respectively) the climate and climatechange documents were distinguishable. As such, η was increased and heatmaps were reconstructed.

Figure 4.11 – Figure 4.14 and Figure 4.15 show the results as η increased further. In particular, Figure 4.11 – Figure 4.14 show the results with all (main and alternative) forums and $\eta = 300, 500, 750, 821$ tokens, respectively (the set of 821 tokens was maximal in the sense that any other token had zero tf-idf values for every document in the heatmap). Likewise, Figure 4.15 shows the results with just the devoted political Reddit forums and $\eta = 281$ tokens (again, the set of 281 tokens was maximal in the sense that any other token had zero tf-idf values for every document in the heatmap). Because the number of tokens has exceeded beyond what is reasonable to display, the heatmaps and tokens have been suppressed, leaving only the dendrogram of documents.

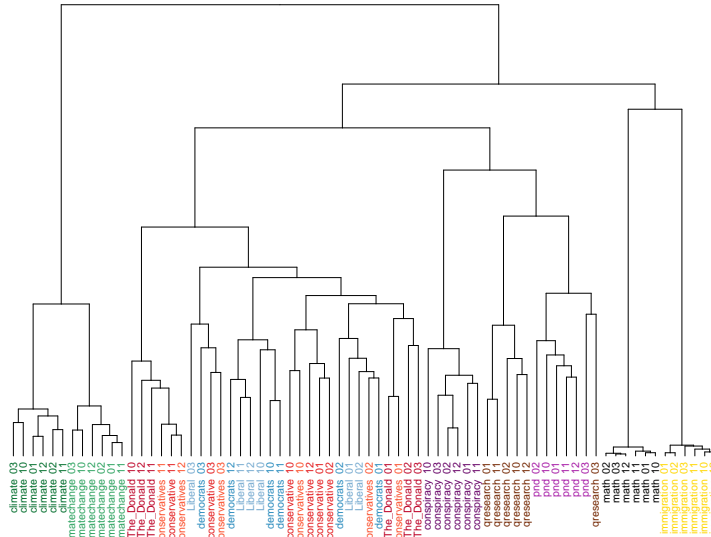


Fig. 4.11: A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 300$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

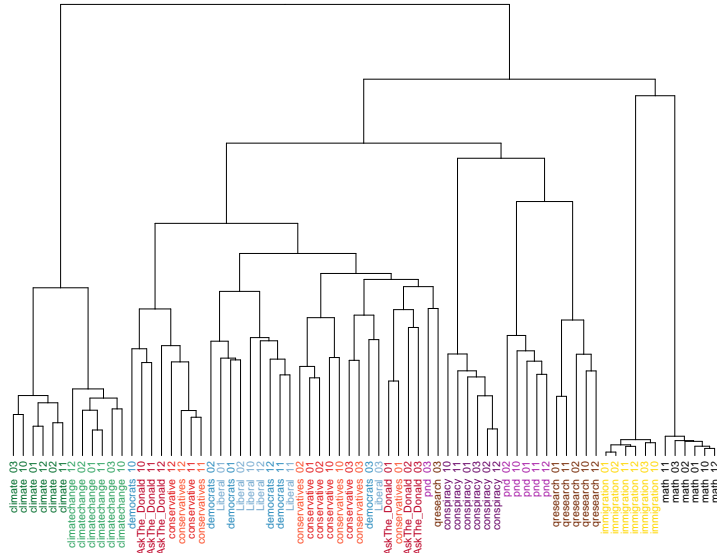


Fig. 4.12: A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 500$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

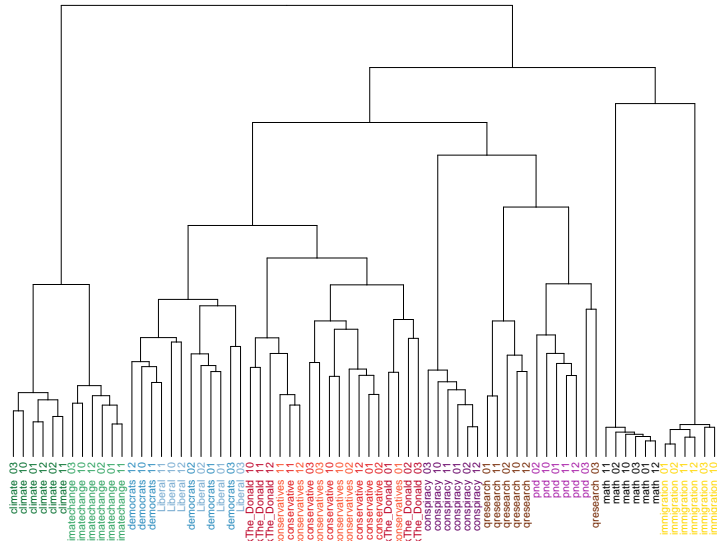


Fig. 4.13: A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 750$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

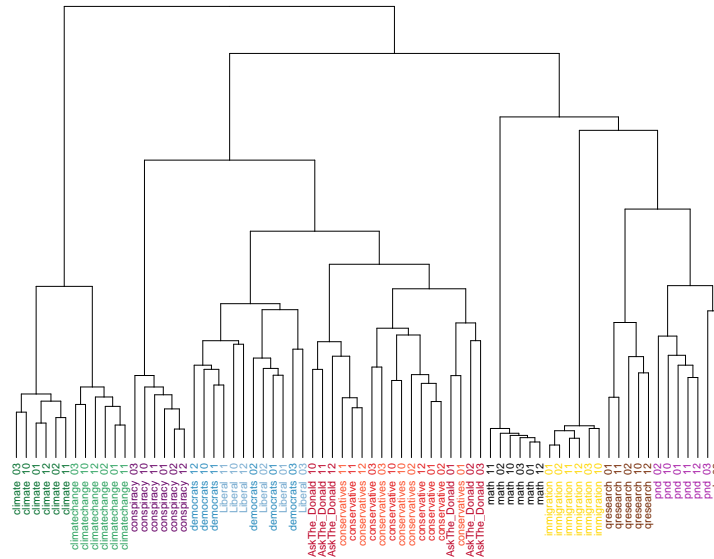


Fig. 4.14: A Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. All $\eta = 821$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.

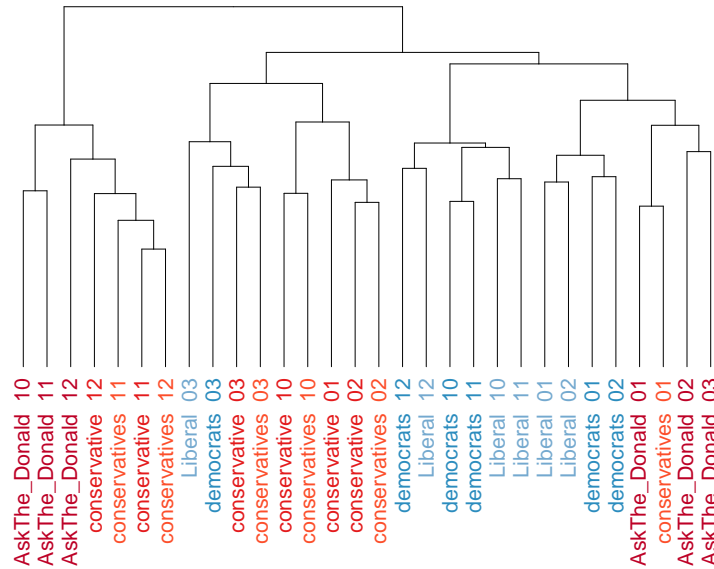


Fig. 4.15: A Ward linkage dendrogram of the devoted political Reddit forums, split across forums and months. All $\eta = 281$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.

Figure 4.11 – Figure 4.14 show that increasing the number of tokens (η) seems to improve

clustering. Both Figure 4.13 and Figure 4.14 ($\eta = 750, 821$, respectively) show similar results as those in Figure 4.3 – Figure 4.6 with the added result that the documents from the democrats and Liberal forums are better distinguished from the documents from the AskThe_Donald, conservative, and conservatives forums (though documents from the democrats and Liberal forums still inter-cluster, as do the documents from the AskThe_Donald, conservative, and conservatives forums). Interestingly, when η is increased with just the devoted political Reddit forums (Figure 4.15), this added result disappears and documents once again inter-cluster (as in Figure 4.10), though some time-based patterns can be found. For example, within Figure 4.15, there is a cluster consisting of October, November, and December democrats and Liberal documents; a cluster consisting of January and February democrats and Liberal documents; a cluster consisting of November and December Ask_TheDonald, conservative, and conservatives documents, as well as the October Ask_TheDonald document; and a cluster consisting of March democrats, Liberal, conservative, and conservatives documents. Similar (but not as strong) time-based patterns can be seen in Figure L.5 in Appendix L.

Figure 4.11 – Figure 4.14 and Figure 4.15 used the Ward linkage clustering algorithm. The analogous figures with the average linkage algorithm can be found in Appendix L. The results are similar, but not quite as strong. In particular, the average linkage analogs to Figure 4.13 and Figure 4.14 do not as well distinguish the documents from the democrats and Liberal forums from the documents from the AskThe_Donald, conservative, and conservatives forums. However, larger quantities of tokens still seem to perform better.

A Note on Heatmaps

Results analogous to Section 4.2 which determine the optimal number of clusters from a dendrogram were not applied to the heatmaps of this section. This is because the Jaccard distance is only compatible with four of the 18 R *NbClust* criteria used throughout this thesis (specifically, the mc-clain, cindex, silhouette and dunn criteria). See the documentation for the `NbClust : :NbClust` function for more details (Charrad et al., 2022).

4.4 Individual Dates and Feelings of Displacement

This section attempts to identify key dates that exhibited feelings of isolation and displacement on politically-right online forums in connection with the January 6, 2021, Capitol attack. However, this section only analyzes three of the forums of study: conspiracy, pnd, and qresearch. This was done because:

- The math, climate, climatechange, and immigration forums were selected as apolitical control forums. Moreover, the prior two sections (Section 4.2 and Section 4.3) determined that these forums were easily distinguishable from all others.
- The prior two sections also determined that the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums were largely indistinguishable. Because these forums span both sides of the political aisle, they were excluded from further analysis.

First, Section 4.4.1 identifies key dates related to the events of January 6, 2021. Then Section 4.4.2 analyzes these key dates for tokens which deal with feelings of isolation and displacement.

4.4.1 Selecting Key Dates

This section identifies key dates related to the events of January 6, 2021, by using the method described in Section 3.8.3, which, from a hierarchical clustering, identifies 18 possible best document groupings based upon 18 different criteria in the R *NbClust* package. Specifically, this identification of best document groupings was done for the conspiracy, pnd, and qresearch forums using Procedure 3.8.3.b, which, for each fixed forum f :

- (1) Splits and aggregates posts according to date, thus resulting in roughly 180 documents corresponding to roughly 180 dates.
- (2) Vectorizes these documents according the doc2vec algorithm (Procedure 3.6.2). Recall that Procedure 3.6.2 cleans — but does not remove stop tokens from — the daily aggregated forum documents before vectorizing them (see Section 2.3.2 for details on the data cleaning process).

- (3) Passes these vectorizations onto the `NbClust::NbClust` function, with the Euclidean distance and Ward linkage algorithm specified, as well as allowing each criteria to identify any number of clusters from one to the number of dates minus one. This returns 18 possible best groupings based upon the 18 different criteria in the R *NbClust* package. Note that these results were not run with the average linkage algorithm or with a different range of identifiable clusters. This was done because (a) with roughly 180 documents, this final step had extremely long run time, and (b) there were overall consistencies in the R *NbClust* results presented in Section 4.2 (also see the discussion of Section 4.2 in Section 5.1).

From these 18 possible best groupings, the daily aggregated forum documents which clustered with the vectorized date of January 6, 2021, were recorded. Figure 4.16 shows the results for conspiracy, pnd, and qresearch. In particular, within each forum block, a dot indicates that that vectorized date clustered with the vectorized date of January 6, 2021, according to the R *NbClust* criteria marked at the bottom. For example, the dot in the top right corner of the qresearch block indicates that the qresearch vectorized date of October 1, 2020, clustered with (according to the *sdbw* criteria) the qresearch vectorized date of January 6, 2021.

Within each row (spanning across all three forum blocks) grey scaling is used to indicate the proportion of columns marked with a dot (the darker the dot, the higher the proportion). Hence, dark rows show vectorized dates that consistently clustered with January 6, 2021, regardless of the forum from which the vectorizations came or the criteria used to determine the clusters. Additionally, for a fixed row (again, spanning across all three forum blocks), if the proportion of columns that are marked with a dot exceeds 0.85, then the corresponding date is marked in blue. The value of 0.85 was chosen so as to capture a modest (between 5 and 10) number of dates.

These blue dates consistently clustered with January 6, 2021. That is, the online discourses of the blue dates were similar to that of January 6, 2021. Among the dates shown in blue, most are intuitively similar to January 6, 2021 — November 3, 2020, was election day, January 6 (and 7), 2021, was the date of (and following) the Capitol attack, and January 16, 17, 18, and 19, 2021, were the four dates preceding inauguration day. However, the two October blue dates make less immediate sense. But part of Amy Coney Barrett’s Supreme Court nominee hearing took place on

October 14, 2020, and her nomination was quite controversial. For one, she was both nominated and confirmed shortly before the 2020 US presidential election. Moreover, Justice Barrett's conservative views oppose that of the late Justice Ginsberg, whom Justice Barrett replaced (see [Chang and Freeman \(2021\)](#) for more information on Justice Barrett's nomination and confirmation to the Supreme Court). Likewise, October 24, 2020, was just two days after the final 2020 US presidential debate between then-President Trump and his democratic challenger, Biden. See *The Week's* daily briefings of October 14, 2020, and October 24, 2020, at [Maass \(2020\)](#) and [O'Donnell \(2020\)](#), respectively, for more information on these dates.

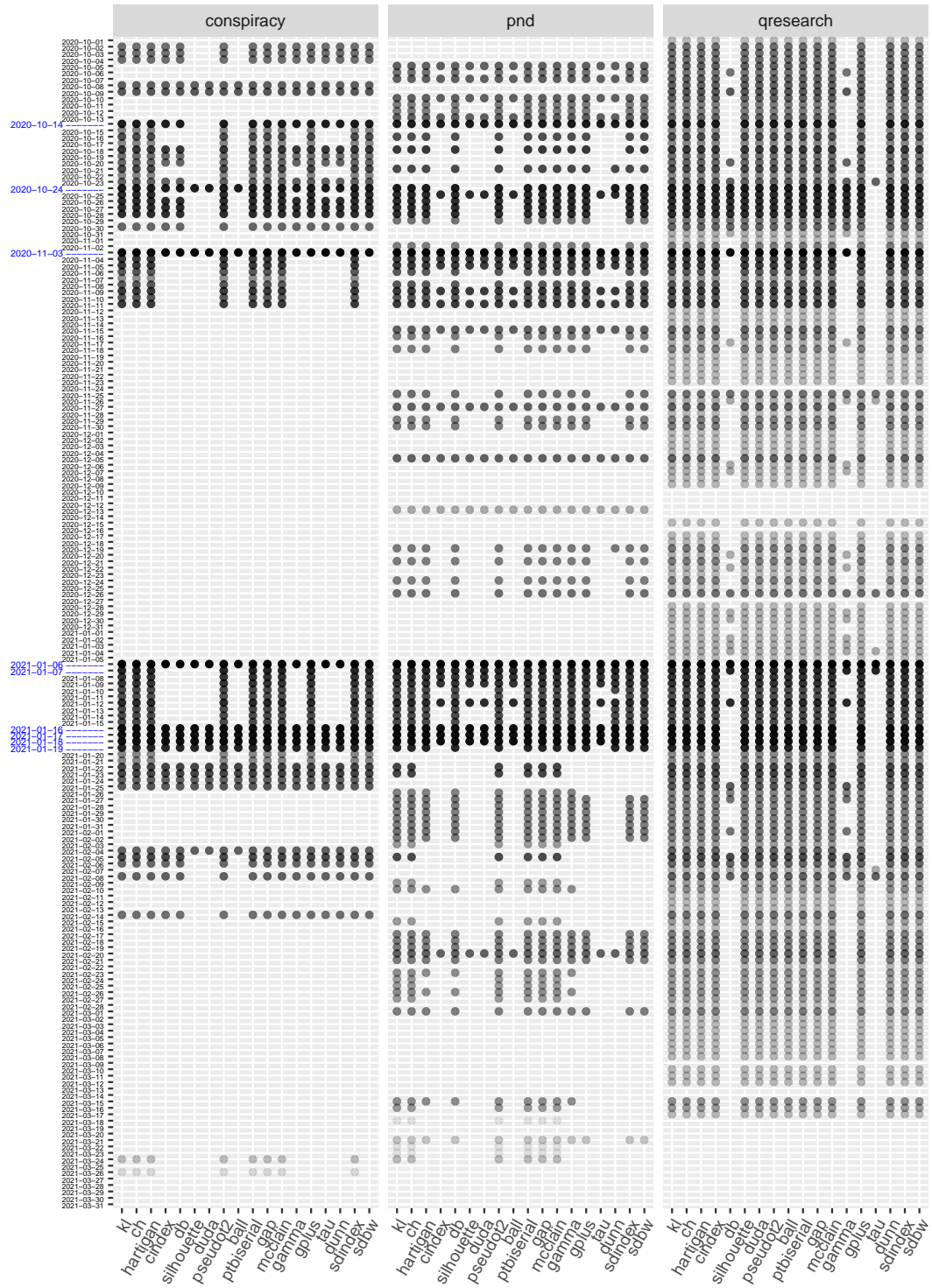


Fig. 4.16: For conspiracy, pnd, and qresearch as well as each R *NbClust* criteria, the points mark the vectorized dates which clustered with January 6, 2021. Each criteria could identify any number of clusters between one and the number of dates minus one. More details are provided in the main text.

4.4.2 Identifying and Quantifying Key Tokens

The prior section (Section 4.4.1) identified nine key dates for the conspiracy, pnd, and qresearch forums — 2020-10-14, 2020-10-24, 2020-11-03, 2021-01-06, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19 — which each had online discourse that consistently clustered (according to the various criteria in the R *NbClust* package) with the online discourse of January 6, 2021, the date of the US Capitol attack. This section determines the extent to which the conspiracy, pnd, and qresearch forums exhibited feelings of isolation and displacement on these key dates.

Consider the following set of key tokens:

$$D^{\text{Replacement}} = \{\textit{attack}, \textit{cultural}, \textit{culture}, \textit{diversity}, \textit{enemies}, \\ \textit{ethnic}, \textit{invaders}, \textit{racial}, \textit{replacement}\}.$$

These key tokens were chosen because (a) they appeared in the heatmaps of Section 4.3, and (b) they have possible connotations to race, culture, and the great replacement conspiracy theory.*

For conspiracy, pnd, and qresearch, Figure 4.17 shows the proportion of forum posts (i.e., a submission or a comment — see Section 1.4.2) that used at least one of these key tokens over time. Note that comments do not repeat submission data. For comparison with anticipated tokens, Figure 4.17 also shows the proportion of posts using the token *election*. The key dates identified in Section 4.4.1 are marked in blue. For each forum, there does not appear to be any proportional increase in key token usage on the key dates marked in blue. There does appear to be a proportional increase in key token usage in the latter months (February and March 2021). However, this could be due to the fact that posting data became more scant through these months (see Figure 4.2). Moreover, Figure 4.17 shows that there was more overall proportional key token usage for the pnd and qresearch forums (both of which are from 8kun) than there was for the Reddit conspiracy forum.

Note that Figure 4.17 shows the proportion of forum *posts* that used at least one of these key

*As discussed in Section 1.3, the great replacement conspiracy theory is “a conspiracy theory that postulates white European populations are being demographically and culturally replaced by non-white immigrants through policies enacted by ‘the global elites’” (Carlson and Harris, 2022). Also see Obaidi et al. (2022) or Cosentino (2020) for more information.

tokens over time. The results are qualitatively the same for the proportion of forum *tokens* used over time (i.e., a single post containing multiple key tokens counts multiple times).

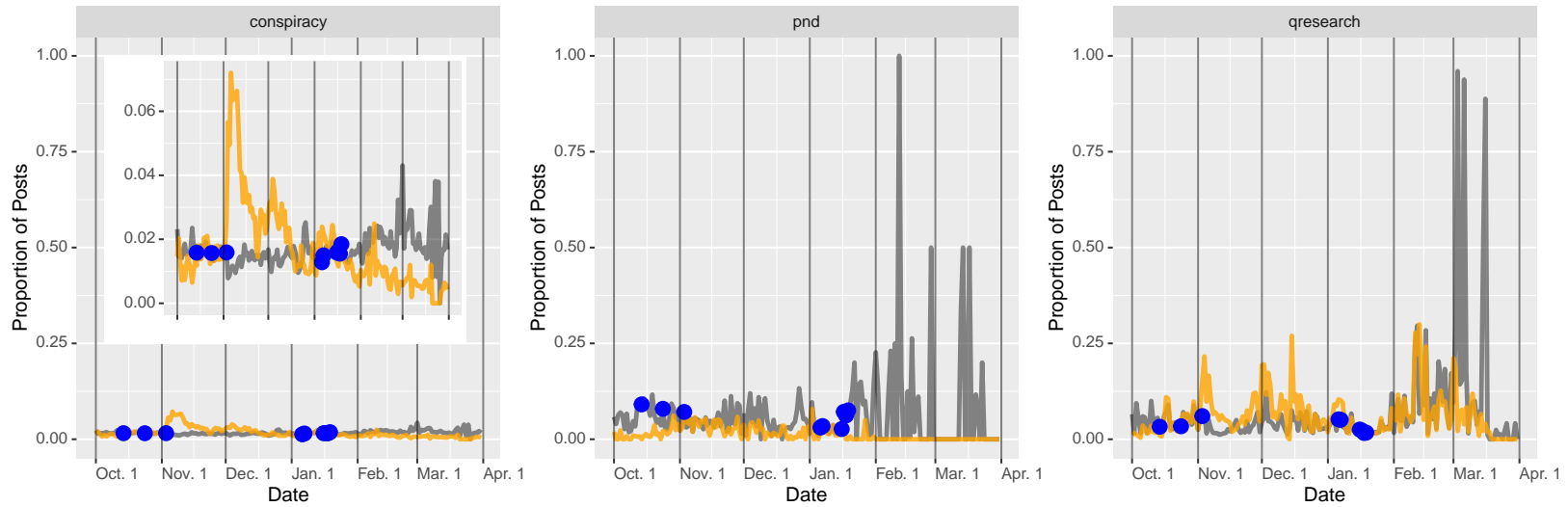


Fig. 4.17: For each forum, the grey line indicates the proportion of posts for each date that used at least one of the key tokens: *attack*, *cultural*, *culture*, *diversity*, *enemies*, *ethnic*, *invaders*, *racial*, and *replacement*. Nine key dates — 2020-10-14, 2020-10-24, 2020-11-03, 2021-01-06, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19 — are marked in blue. The orange line indicates the proportion of posts for each date that used the token *election*. A rescaled version of the conspiracy graphic is shown inside the conspiracy graphic obeying the small multiples principle.

CHAPTER 5

Discussion

This chapter discusses the results from Chapter 4 by highlighting the key takeaways, as well as providing connections and comparisons between sections, different methods, and related work. Section 5.1 discusses the most prevalent patterns found in the R *NbClust* results of Section 4.2. Likewise, Section 5.2 discusses the most prevalent patterns found in the heatmap results of Section 4.3. Section 5.3 summarizes the clustering results found. Section 5.4 then discusses the extent to which feelings of isolation and displacement were observed online. Finally, Section 5.5 discusses some limitations of this thesis.

5.1 Discussion of the R *NbClust* Results

This section discusses the most prevalent patterns of document clustering found in the R *NbClust* results of Section 4.2, where forum documents were created from random samples of data. In particular, Section 4.2 took three random samples of data from each (main and alternative) forum — i.e., AskThe_Donald, conservative, conspiracy, pnd, qresearch, climate, climatechange, immigration, conservatives, democrats, Liberal, and math — and sought to identify clusters of such documents using 18 different criteria from the R *NbClust* package. Random samples of data were taken to both allow the R *NbClust* package to run in a reasonable time, and also to determine the general patterns of how forums intra and inter cluster. As a reminder, the R *NbClust* results are based upon Procedure 3.8.3.a which utilizes the doc2vec algorithm for document vectorizations. The components of these vectorizations have no interpretation. Therefore, tokens cannot not be discussed in this section (but will be discussed in Section 5.2).

The detailed results of Section 4.2 are displayed in Appendix I. The 18 tables in Appendix I correspond to the 18 criteria of the R *NbClust* package. That is, each table fixes a single criteria from the R *NbClust* package, and shows the clustering results for the various random seeds, linkage algorithms, and ranges of identifiable clusters. Section 5.1.1 discusses one clear and consistent

clustering pattern found throughout these R *NbClust* results. Then Section 5.1.2 and Section 5.1.3 discuss refinements of that clear clustering.

5.1.1 Five Clusters of Forums — R *NbClust* Results

This section discusses the most prevalent pattern of document clustering found in the R *NbClust* results of Section 4.2, where forum documents were created from random samples of data. The prevalent pattern is depicted in (★) and consists of five clusters.

- (★) {
- The documents from the math forum.
 - The documents from the climate/climatechange forums.
 - The documents from the immigration forum.
 - The documents from the qresearch forum.
 - The documents from the pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal forums.

Any time a particular criteria identified five clusters, those five clusters corresponded exactly to (★). Moreover, of the 18 criteria, six (the ptbserial, gamma, gplus, tau, dunn, and sdindex criteria) universally agreed with (★), regardless of the random seed, linkage algorithm, or range of identifiable clusters (see Table I.10, Table I.13, Table I.14, Table I.15, Table I.16, and Table I.17 in Appendix I, respectively). No other criteria had perfectly stable results through the various seeds, linkage algorithms, and ranges of identifiable clusters. The kl criteria (Table I.1 in Appendix I) was also remarkably stable: of 40 instances (for the varying random seeds, linkage algorithms, and ranges of identifiable clusters), only three did not identify (★). Likewise, the gap and mcclain criteria (Table I.11 and Table I.12 in Appendix I, respectively) would identify one and two clusters (respectively) when allowed, but when they were required to identify at least five clusters, they would identify (★) with only three exceptions for the gap criteria and no exceptions for the mcclain criteria.

The ball criteria (Table I.9 in Appendix I) behaved somewhat uniquely. When allowed to

identify any number of clusters, the ball criteria would always identify two clusters. When forced to identify at least five clusters, the ball criteria would identify six or seven clusters.

The *duda* and *pseudot2* criteria (Table I.7 and Table I.8 in Appendix I, respectively) always identified between seven and 12 clusters. Five of the clusters always came from the five forums: *math*, *climate*, *climatechange*, *immigration*, and *qresearch*. The remaining clusters would split the fifth cluster of (★).

The *ch* criteria (Table I.2 in Appendix I) was the most unstable of all criteria. Of the 40 instances (for the varying random seeds, linkage algorithms, and ranges of identifiable clusters), 10 identified either nine or 10 clusters, and the remaining 30 instances identified 33, 34, or 35 clusters. In short, the *ch* criteria was very sensitive to document changes, linkage algorithms, and the specified range of identifiable clusters.

The remaining criteria — the *hartigan*, *cindex*, *db*, *silhouette*, and *sdbw* criteria (see Table I.3, Table I.4, Table I.5, Table I.6, and Table I.18 in Appendix I, respectively) — always identified 32+ clusters. That is, most documents were separated into singleton clusters. Importantly, when 32+ clusters were identified (for these criteria, and also for the unstable *ch* criteria), the non-singleton clusters always formed from documents of the same forum.

Notably, no criteria substantially disagreed with (★). More precisely (barring the results that identified only one or two clusters), no criteria ever clustered a document from one cluster of (★) with a document from a different cluster of (★). Therefore, all told, the clustering in (★) is a consistent result. As such, the random *math*, *climate/climatechange*, *immigration*, and *qresearch* documents — and hence (likely) the forums in general — were the most unique and distinguishable. This agrees with [Tran and Ostendorf \(2016\)](#) who clustered Reddit forum data and found that intuitively distinct forums are easily identifiable. This also mostly agrees with the heatmaps of Section 4.3 (see the discussion in Section 5.2.1).

5.1.2 Refinement One of the Five Clusters — R *NbClust* Results

This section describes a refinement of the (★) clustering found in the R *NbClust* results of Section 4.2, where forum documents were created from random samples of data. In (★) above,

the pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal random documents formed a single cluster together. As discussed in Section 5.1.1, (*) was prevalent through the R *NbClust* results of Section 4.2. This implies a degree of similarity between these forums in the fifth cluster of (*). However, there is an additional refinement of this cluster within the R *NbClust* results. Namely, a split between the pnd and conspiracy forums from the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums. Consider the following conditional statement:

If the pnd/conspiracy/conservative/conservatives/AskThe_Donald/democrats/Liberal documents were identified as two or more clusters (e.g., the duda criteria in Table 4.3), then no pnd/conspiracy documents clustered with conservative/conservatives/AskThe_Donald/democrats/Liberal documents.

Among all R *NbClust* results (shown in Appendix I), the above conditional statement only had seven exceptions, as depicted in Table 5.1. Of these seven exceptions, one came from the gap criteria and four came from the ball criteria. When allowed to identify any number of clusters, the gap criteria always identified one cluster and the ball criteria always identified two clusters. The gap and ball exceptions in Table 5.1 came when the criteria were forced to identify at least five clusters, when they otherwise would not have. As such, it seems that these exceptions are minor. The remaining two exceptions are somewhat different. They came from the duda and pseudot2 criteria for the random seed 10601, average linkage, and requiring 5 – 35 clusters. When requiring 1 – 35 clusters, these criteria (still for the random seed 10601 and average linkage) both identified 10 clusters. Hence — because these criteria ‘normally’ identified non-trivial clusters — these two exceptions are less minor. Still, these are the only two non-minor exceptions of all 720 possibilities in Appendix I, and both occurred for the same random seed with average linkage and requiring a minimum of five clusters. As such, the above conditional statement that separates the pnd and conspiracy forums from the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums holds well, with only isolated exceptions. Therefore, there is a reasonable split between the pnd and conspiracy forums and the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums. This agrees with the heatmaps of Section 4.3 (see the discussion in Section 5.2.2).

Table 5.1: The seven R *NbClust* exceptions (by criteria, random seed, linkage algorithm, and range of identifiable clusters) that do not separate the pnd/conspiracy documents from the conservative/conservatives/AskThe_Donald/democrats/Liberal documents. As discussed in the main text, the gap and ball rows are minor exceptions and are separated from the duda and pseudot2 rows (less minor exceptions) by a horizontal line.

Criteria	Seed	Linkage	Possible Clusters
gap	12421	Ward	5 – 35
ball	12421	Ward	5 – 35
ball	12421	average	5 – 35
ball	97079	Ward	5 – 35
ball	73037	Ward	5 – 35
duda	10601	average	5 – 35
pseudot2	10601	average	5 – 35

5.1.3 Refinement Two of the Five Clusters — R *NbClust* Results

This section describes yet another refinement of the (★) clustering found in the R *NbClust* results of Section 4.2, where forum documents were created from random samples of data. The prior section (Section 5.1.2) discussed that the fifth cluster shown in (★) can be further broken down into a cluster of pnd and conspiracy documents and a cluster of conservative, conservatives, AskThe_Donald, democrats, and Liberal documents. An additional refinement can be made of this second cluster consisting of conservative, conservatives, AskThe_Donald, democrats, and Liberal documents. In this vein, consider the following conditional statement:

If the conservative/conservatives/AskThe_Donald/democrats/Liberal documents were identified as two or more clusters (e.g., the duda criteria in Table 4.3), then no conservative/conservatives/AskThe_Donald documents clustered with democrats/Liberal documents.

For all criteria, random seeds, linkage algorithms, and ranges of identifiable clusters:

- There were 720 possible best groupings (18 criteria \times 10 random seeds \times 2 linkage algorithms \times 2 ranges of identifiable clusters).

- Of these 720 groupings, 84 had no singleton clusters and satisfied the hypothesis of the above conditional statement.
- Of these 84 groupings, 29 were an exception to the above conditional statement.

In other words, roughly 1/3 of the time, the R *NbClust* results had difficulty distinguishing political leanings in the conservative, conservatives, AskThe.Donald, democrats, and Liberal forums (or, 2/3 of the time, the R *NbClust* results could distinguish political leanings). In this way, there is some ability to distinguish political leanings, but by no means is this ability strong.

This agrees with [Tran and Ostendorf \(2016\)](#) who stated: Reddit “[forums] that were most confusable are intuitively similar: politics and worldnews, askmen and askwomen.” In the case of this thesis, the conservative, conservatives, AskThe.Donald, democrats, and Liberal forums are intuitively similar (but are opposed in a sense — like askmen and askwomen). This also agrees with the heatmaps of Section 4.3 (see the discussion in Section 5.2.3).

5.2 Discussion of the Heatmap Results

This section discusses the most prevalent patterns found in the heatmaps of Section 4.3 (and associated appendices), where forum documents were aggregated monthly. As a reminder, the heatmaps are based upon Procedure 3.6.1.b which vectorizes the monthly forum documents by selecting η tokens with the highest tf-idf values. As such, tokens will be discussed throughout this section. Section 5.2.1 discusses the extent to which the heatmaps of Section 4.3 agree with the R *NbClust* (\star) clustering discussed in Section 5.1.1. Likewise, Section 5.2.2 and Section 5.2.3 discuss refinements of (\star) in the context of heatmaps.

5.2.1 Five Clusters of Forums — Heatmap Results

This section discusses how the heatmaps of monthly forum documents in Section 4.3 relate to the five clusters of (\star) discussed in Section 5.1.1. As a reminder, the five clusters of (\star) were determined by the R *NbClust* package with forum documents that were created from random samples of

data. The (\star) clustering was:

- $$(\star) \left\{ \begin{array}{l} \bullet \text{ The documents from the math forum.} \\ \bullet \text{ The documents from the climate/climatechange forums.} \\ \bullet \text{ The documents from the immigration forum.} \\ \bullet \text{ The documents from the qresearch forum.} \\ \bullet \text{ The documents from the pnd, conspiracy, conservative, conservatives,} \\ \text{AskThe_Donald, democrats, and Liberal forums.} \end{array} \right.$$

As discussed in Section 5.1.1, this clustering of randomly sampled documents implies the uniqueness and distinguishability of the math, climate/climatechange, immigration, and qresearch forums. For all but the qresearch forum, this uniqueness holds exactly for the heatmaps in Section 4.3 (and associated appendices: Appendix J, Appendix K, and Appendix L). More precisely:

The monthly documents of the math (or climate/climatechange, or immigration) forum were all aggregated together before clustering with any other monthly document from a different forum.

Because the math, climate/climatechange, and immigration forums were chosen as apolitical control forums, this agrees with logic. This also agrees with Tran and Ostendorf (2016) who clustered Reddit forum data and found that intuitively distinct forums are easily identifiable.

However, the clear uniqueness and distinguishability of the qresearch forum implied by (\star) somewhat disagrees with the heatmaps of Section 4.3 (and associated appendices). Unlike the math, climate/climatechange, and immigration forums,

It is *not* the case that the monthly documents of the qresearch forum were all aggregated together before clustering with any other monthly document from a different forum.

As such, it seems that the monthly qresearch documents had more cross-forum similarity than the five cluster pattern in (\star) suggests. However, (a) only the March qresearch document clustered away from the other monthly qresearch documents in the heatmaps of Section 4.3 (and associated

appendices), and (b) this March qresearch document was always (through each heatmap of Section 4.3 as well as the associated appendices) the last monthly document to be aggregated into a cluster. Therefore, it seems that the qresearch forum still had a large amount of similarity between monthly documents (excluding March), as was suggested by (*). Likewise, it seems that the anomalous March document did not stem from the fact that it was actually closer to monthly documents from different forums, but that it was considerably different from *all* other monthly documents, including the other qresearch documents. This anomaly was not seen in the R *NbClust* results with random documents. Likely, this is due to the fact that a randomly created qresearch document does not contain many posts from March (roughly 1/6 would be from March if posting quantity was uniform, but less than 1/6 given that March had less posting data — see Figure 4.2).

This considerable difference in the March qresearch document is seen in the heatmaps of monthly documents in Section 4.3 (and associated appendices). For example, consider Figure 4.6. Directly above the March qresearch label, there is a column of 35 positive tf-idf values that correspond to 35 tokens* which are mostly unique to the March qresearch document — only four other documents used one of these tokens: the December AskThe_Donald document used *culture*, the October pnd document used *men*, the November pnd document used *race*, and the January pnd document used both *kill* and *fight*. The uniqueness of these 35 tokens is also seen in the token dendrogram where they formed a single cluster together before clustering with all other tokens in the final aggregation of the dendrogram. Note that qresearch tokens in general (not just those specific to March) were rather unique as well (see the heatmap in Figure 4.6 — on the right edge about 1/3 up there is a large block of positive tf-idf values that correspond to 53 tokens† which are mostly unique to the qresearch forum). The fact that the qresearch forum used unique tokens agrees with prior research. Hannah (2021) noted that QAnon posts (of which the qresearch forum is a hub) are “cryptic.”

*Alphabetically, these 35 tokens are: *attack, beliefs, children, cities, cultural, culture, death, destroy, diversity, en, enemies, ethnic, europe, european, fertility, fight, force, history, invaders, kill, labour, lands, men, nation, nations, org, peoples, race, racial, rates, replacement, truth, victory, wiki, and wikipedia.*

†Alphabetically, these 53 tokens are: *..., 12, 2020, 2021, 8kun, anon, anons, australia, australian, baker, bakers, bread, china, chinese, christmas, com, comms, da, das, de, den, der, die, drop, edition, former, general, html, hunter, ist, je, military, mr, notables, oss, police, posts, q, qanon, qresearch, report, res, research, top, tweet, twitter, u, uk, und, vatican, von, web, and zu.* Note that the first token in this list (...) represents an error in the data cleaning process. The token should have been removed when non-American Standard Code for Information Interchange (ASCII) characters were removed, but inexplicably stayed.

These unique tokens make it difficult to cluster the March qresearch document in a reasonable fashion. This is seen by the fact that the March qresearch document most often immediately clustered with the March pnd document, which had similarly unique tokens (see Figure 4.6, above the March pnd label, and about halfway up the heatmap — there is a column of 14 positive tf-idf values that correspond to 14 tokens* which are mostly unique to the March pnd document). In other words, these two March documents are fairly disparate from all others, but they needed to cluster somewhere and so they clustered together in an ‘island of misfit toys’ fashion. Therefore, barring these anomalous March documents, the five cluster pattern of (★) detected in the R *NbClust* results (with randomly created forum documents) holds well in the heatmaps of monthly aggregated documents in Section 4.3 (and associated appendices). In particular, this implies that temporal differences in the monthly documents were subsidiary to the forum itself. Or, in other words, the forum from which a document came was the primary factor in determining where it clustered (aside from the anomalous March qresearch and March pnd documents).

5.2.2 Refinement One of the Five Clusters — Heatmap Results

Section 5.1.2 discussed that the fifth cluster of (★) — for the R *NbClust* results with randomly created forum documents — could be refined further. Specifically, there was a separation between the pnd and conspiracy forums and the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums.

This agrees with the heatmaps of Section 4.3 (and associated appendices), which clustered forum data by month. In particular, each heatmap (except Figure 4.4 and Figure 4.12) contained a cluster of all (and only) conservative, conservatives, AskThe_Donald, democrats, and Liberal documents. Moreover, the two exceptions (again, Figure 4.4 and Figure 4.12) still contained a cluster of all conservative, conservatives, AskThe_Donald, democrats, and Liberal documents, and only added the anomalous March qresearch and/or March pnd document to that cluster. As discussed in Section 5.2.1, these anomalous documents must be clustered somewhere, and they tended to aggregate into clusters late. In other words, the Figure 4.4 and Figure 4.12 exceptions are some-

*Alphabetically, these 14 tokens are: 191491, allowed, banks, btc, central, debt, economic, economy, fiat, free, game, massive, text, and today.

what minor. In this way, there is a reasonable cluster consisting of the conservative, conservatives, AskThe_Donald, democrats, and Liberal documents. Or, there is a separation of the pnd and conspiracy documents from the conservative, conservatives, AskThe_Donald, democrats, and Liberal documents.

5.2.3 Refinement Two of the Five Clusters — Heatmap Results

Section 5.1.3 discussed that the fifth cluster of (★) — for the R *NbClust* results with randomly created forum documents — could again be refined further. Specifically, there was some minor separation of the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums along political leanings.

This agrees with the heatmaps of Section 4.3 (and associated appendices), which clustered forum data by month. Each heatmap identified a cluster which contained all the conservative, conservatives, AskThe_Donald, democrats, and Liberal monthly documents. However, most heatmaps did not distinguish political leanings between these documents. Figure 4.13 and Figure 4.14 were the only two to show a clear separation along political lines. Figure 4.13 and Figure 4.14 used Ward linkage. Their average linkage results are shown in Figure L.3 and Figure L.4 in Appendix L, respectively. The average linkage results show good, but not perfect, separation along political leanings. In this way, the heatmaps of monthly aggregated documents showed some separation along political leanings, but not a consistent separation along political leanings. Moreover, separation along political leanings required large quantities of tokens — not just more than 200 tokens, but almost all tokens ($\eta = 750,821$ tokens, as in Figure 4.13 and Figure 4.14, respectively). This implies that these Reddit political forums differ most in somewhat less common tokens.

This agrees with Tran and Ostendorf (2016) who stated: Reddit “[forums] that were most confusable are intuitively similar: politics and worldnews, askmen and askwomen.” In the case of this thesis, the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums are intuitively similar (but are opposed in a sense — like askmen and askwomen).

Notably, the figures which show a clear separation along political leanings (Figure 4.13 and Figure 4.14) clustered monthly documents from all forums, not just the Reddit political forums:

conservative, conservatives, AskThe.Donald, democrats, and Liberal. That is, in order to observe a clear separation along political lines, all (main and alternative) forums needed to be in the corpus. This again implies that these Reddit political forums differ most in somewhat less common tokens. In particular, consider Figure 4.14 which depicts a Ward linkage dendrogram of all (main and alternative) forums, split across forums and months. This dendrogram came from a heatmap (not pictured) that utilized all $\eta = 821$ tokens with positive tf-idf values. The heatmap clustered together 30 tokens that tended to have higher tf-idf values for politically-left forums,^{*} and clustered together 30 tokens that tended to have higher tf-idf values for politically-right forums.[†] These 60 tokens together contain general political tokens (e.g., *democrats*, *republicans*, *votes*, *party*, *biden*, *trump*, etc.), but also contain tokens that are understandably one-sided. For example, *georgia* appeared in politically-left forums (Biden won the state of Georgia in the 2020 US presidential election, making it the first time since 1992 that the democratic nominee for president won the state) and *fraud* appeared in politically-right forums. Recall that these are aggregated results. In other words, no conclusion can be drawn about individual users of these forums. The aggregated data simply suggest that *georgia* and *fraud* were discussed more on politically-left and politically-right forums (respectively), in general (and likewise for the other tokens). The 30 tokens that tended to have higher tf-idf values for politically-right forums have no obvious connotations to feelings of isolation and displacement. However, there were tokens dealing with voter fraud and other politically-contentious topics (e.g., *covid*).

Figure 4.15, which clustered only the Reddit political forums, shows some time-based patterns: a cluster consisting of October, November, and December democrats and Liberal documents; a cluster consisting of January and February democrats and Liberal documents; a cluster consisting of November and December Ask_TheDonald, conservative, and conservatives documents, as well as the October Ask_TheDonald document; and a cluster consisting of March democrats, Liberal, conservative, and conservatives documents. However, the heatmap and dendrogram of tokens (not

^{*}Alphabetically, these 30 tokens are: *2016*, *americans*, *best*, *bill*, *blue*, *campaign*, *college*, *democratic*, *democrats*, *feel*, *georgia*, *gop*, *hope*, *law*, *lost*, *love*, *party*, *please*, *rep*, *republican*, *republicans*, *senate*, *stimulus*, *vote*, *voted*, *voter*, *voters*, *voting*, *win*, and *won*.

[†]Alphabetically, these 30 tokens are: *ballot*, *ballots*, *biden*, *cdc*, *china*, *counting*, *court*, *covid*, *dems*, *election*, *elections*, *evidence*, *fraud*, *hate*, *house*, *joe*, *mail*, *media*, *news*, *obama*, *police*, *president*, *says*, *state*, *states*, *trump*, *u*, *video*, *votes*, and *wins*.

pictured in Figure 4.15) do not provide much clarity as to why these time-based patterns emerged. For example, the October, November, and December democrats and Liberal documents tended to use 19 tokens more frequently than other documents,* and the March democrats, Liberal, conservative, and conservatives documents tended to use 11 tokens more frequently than other documents.† However, these patterns are rather weak. This ambiguity agrees with the fact that the heatmap results had difficulty distinguishing political leanings when only clustering documents from the Reddit political forums: conservative, conservatives, AskThe_Donald, democrats, and Liberal.

5.3 Summary of Clusters

This thesis considered cluster results coming from two different methods. Section 5.1 discussed the R *NbClust* results which were derived from randomly created forum documents, and Section 5.2 discussed the heatmap results derived from monthly forum documents. Overall, the two methods showed a good deal of agreement. For one, both methods showed that documents coming from the same forum tended to cluster together. Moreover, both methods showed that certain forums had documents (either randomly created forum documents, or monthly forum documents) that tended to cluster with documents coming from a different forum. In particular, both methods identified the (★) clustering depicted below. Within the fifth cluster of (★), both methods also showed two refinements. First, the pnd and conspiracy documents separated from the conservative, conservatives, AskThe_Donald, democrats, and Liberal documents. For both methods, this refinement was consistent but not perfect. Secondly, the conservative, conservatives, AskThe_Donald, democrats, and Liberal documents separated along political lines. For both methods, this refinement was less consistent but still apparent.

*Alphabetically, these 19 tokens are: 4, best, blue, change, college, democratic, feel, georgia, gop, help, hope, live, lost, please, senate, voted, voters, win, and won.

†Alphabetically, these 11 tokens are: around, article, black, id, less, pay, problem, racist, read, understand, and vaccine.

- (*) {
- The documents from the math forum.
 - The documents from the climate/climatechange forums.
 - The documents from the immigration forum.
 - The documents from the qresearch forum.
 - The documents from the pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal forums.

5.4 Key Token Analysis on Key Dates

Section 4.4 analyzed the extent to which feelings of isolation and displacement manifested themselves on the conspiracy, pnd, and qresearch forums. Note that only these three forums were studied in this context because:

- The math, climate, climatechange, and immigration forums were selected as apolitical control forums.
- The conservative, conservatives, AskThe_Donald, democrats, and Liberal forums — which span both sides of the political aisle — were mostly indistinguishable in Section 4.2 and Section 4.3.

This analysis of feelings of isolation and displacement first identified nine key dates (2020-10-14, 2020-10-24, 2020-11-03, 2021-01-06, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19) which had online discourse that consistently clustered (according to the various R *NbClust* criteria) with the discourse of January 6, 2021, the date of the US Capitol attack. The analysis then determined the prevalence of certain key tokens dealing with feelings of isolation and displacement. These key tokens (*attack*, *cultural*, *culture*, *diversity*, *enemies*, *ethnic*, *invaders*, *racial*, and *replacement*) were selected because they were a subset of the tokens in the heatmaps of Section 4.3 and also have connotations to race, culture, and the great replacement conspiracy

theory.* Section 5.4.1 discusses the selection process of the nine key dates. Then Section 5.4.2 discusses the key token analysis.

5.4.1 Selecting Key Dates

The key dates identified in Section 4.4.1 (2020-10-14, 2020-10-24, 2020-11-03, 2021-01-06, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19) were those that consistently clustered with January 6, 2021, according to the 18 criteria of the R *NbClust* package (see Figure 4.16). In this analysis, Ward linkage was used and each criteria was allowed to identify any number of clusters between one and the number of dates minus one. Most of these key dates have connections to January 6, 2021. November, 3, 2020, was election day, January 6 (and 7), 2021, was the date of (and following) the January 6, 2021, Capitol attack, and January 16, 17, 18, and 19, 2021, were the four dates preceding inauguration day. It is encouraging that these were dates that were identified as having discourse similar to that of January 6, 2021. However, October 14, 2020, and October 24, 2020, have less immediate connections to January 6, 2021. But part of Amy Coney Barrett’s Supreme Court nominee hearing took place on October 14, 2020, and October 24, 2020, was two days after the final 2020 US presidential debate (see *The Week*’s daily briefings of October 14, 2020, and October 24, 2020, at [Maass \(2020\)](#) and [O’Donnell \(2020\)](#), respectively). If feelings of isolation and displacement existed online in connection with the events of January 6, 2021, it seems that those feelings would appear on these key dates.

5.4.2 Analysis of Key Tokens

The key tokens selected in Section 4.4.2 for further analysis were: *attack*, *cultural*, *culture*, *diversity*, *enemies*, *ethnic*, *invaders*, *racial*, and *replacement*. These tokens were selected because they all appeared in the heatmaps of Section 4.3 and also each have connotations to race, culture, and the great replacement conspiracy theory. Figure 4.17 showed that these key tokens did not increase in proportional usage on the key dates. However, Figure 4.17 did show that the overall usages of

*As discussed in Section 1.3, the great replacement conspiracy theory is “a conspiracy theory that postulates white European populations are being demographically and culturally replaced by non-white immigrants through policies enacted by ‘the global elites’” ([Carlson and Harris, 2022](#)). Also see [Obaidi et al. \(2022\)](#) or [Cosentino \(2020\)](#) for more information.

these key tokens on the 8kun forums (pnd and qresearch) were more than that of the Reddit forum (conspiracy).

No Apparent Increase in Feelings of Isolation and Displacement

As discussed in Section 1.3, the works of the University of Chicago’s *Chicago Project on Security & Threats* (CPOST) (see [Pape \(2022a\)](#) and [Pape \(2022b\)](#)) suggested that a common feature among Capitol protestors on January 6, 2021, was that of feeling isolated and/or displaced from society. However, the results of this thesis did not measure any proportional increase in these feelings in connection with January 6, 2021. Nonetheless, this does not necessarily contradict the works of the CPOST. In particular, there is an age discrepancy between that of the Capitol protestors on January 6, 2021, and that of Reddit users. [Pape \(2022a\)](#) found that roughly half of the Capitol protestors were between the ages of 35 and 54, and only one in ten Capitol protestors were under 25. Contrast this with a 2021 Pew Research study ([Auxier and Anderson, 2021](#)) which concluded that only 22% of US adults aged 30 – 49 used Reddit (compared with 36%, 10%, and 3% for US adults aged 18 – 29, 50 – 64, and 65+, respectively, and also compared with the 77% of US adults aged 30 – 49 who used Facebook). Note that no data concerning the ages of 8kun users could be found (recall that 8kun is “semi-dark” ([Zamani et al., 2019](#))).

As such, the population characteristics of Capitol protestors were fundamentally different than that of Reddit users. In this way, the results of this thesis — which did not identify considerable feelings of isolation and displacement on Reddit (and 8kun) in connection with the events of January 6, 2021 — do not contradict the CPOST’s findings. Rather, this thesis simply implies that the CPOST’s findings do not carry over to the users of Reddit and 8kun in any conclusive way.

As a comparison with outside work, [Bortolon \(2022\)](#) studied the qresearch forum and did not identify feelings of isolation and displacement as a key theme of the forum. This agrees with this thesis inasmuch that tokens with connotations to isolation and displacement did not manifest themselves strongly on the qresearch forum.

8kun Shows More Feelings of Displacement

As mentioned above, the overall usages of key isolation/displacement tokens on the 8kun forums (pnd and qresearch) were more than that of the Reddit forum (conspiracy). Or, in other words, there were more feelings of displacement and isolation on 8kun than that of Reddit (these feelings just did not coincide with the events of January 6, 2021). This aligns with prior knowledge about 8kun.

First of all, as discussed in Section 2.1, the pnd forum has hosted people like Brenton Tarrant who posted “I will carry out an attack against the invaders” as well as a link to his manifesto and a live stream of his attack on two mosques in Christchurch, New Zealand, in 2019 (see Baele et al. (2020)). Other attackers have likewise posted similar manifestos and committed other xenophobic atrocities (again, see Baele et al. (2020)). Second of all, there is a sense in which 8kun users believe they are distinct from society. Hannah (2021) stated: “[8kun users] believe they are actually *more* informed than the regular population” and:

Because of QAnon’s networked structure, the theory is spread off-line, beyond the chans, to family, friends, co-workers, and colleagues through a pedagogical front known as “redpilling.” Used as a verb, redpilling refers to the process of revealing, explaining, and contextualizing secret information to the broader, off-line community.

In this way, it makes sense that 8kun users exhibited feelings of isolation and displacement. On the whole, they believe that they are people who have taken the “red pill” and left normal society. Taking the “red pill” refers to the 1999 movie *The Matrix*, where the main character (Neo) is offered a choice between a blue pill and red pill — the blue pill returning Neo to his normal life, the red pill revealing the truth about the Matrix: that humanity lost a war against machines and now (unknowingly) lives in a simulated reality known as the Matrix. Still, despite these feelings of isolation and displacement that existed on 8kun, there were not any noticeable differences in these feelings in terms of the January 6, 2021, Capitol attack.

Selected Posts

The following three text strings show selected forum posts (in their entirety and uncleaned form) that utilized one of these key tokens related to feelings of isolation and displacement.

conspiracy: I'm not sure why ''**diversity**'' is important. I don't visit India for a taste of Norway, or France so that I may enjoy the hodgepodge of **culture**.

pnd: >>217310 >the great **replacement** It's a genocide and Jews are responsible. Fuck off faggot kike shill.

qresearch: I'm sick and tired of all the bi-**racial** commercials on tv now since the BLM summer.

5.5 Limitations

This thesis analyzed social media posts from various 8kun and Reddit forums. Specifically, it sought to analyze how different forums clustered together, as well as determine the extent to which feelings of isolation and displacement existed on online forums in connection with the January 6, 2021, Capitol attack. There are limitations to the analysis described in this thesis, and this section discusses those limitations.

5.5.1 Hindsight Analysis, Anticipatory Intelligence, and the Limitations of Quantitative Research

This thesis was a hindsight analysis of the events at the US Capitol on January 6, 2021. Even with the benefit of hindsight, there was difficulty in detecting the feelings of isolation and displacement discussed by the CPOST (see Section 1.3 on the related works of this thesis). As such, there would be even more difficulty in anticipating the events of January 6, 2021. This represents a limitation of purely quantitative approaches to complex social problems. Such a limitation has been discussed before (e.g., [Queirós et al. \(2017\)](#)) and has also been exemplified by the fact that US government agencies have hired qualitative researchers in the past (e.g., the Central Intelligence

Agency (CIA) used anthropological research during World War II and the Cold War — see [Price \(2011\)](#)). Even in a day of big data, [González \(2015\)](#) stated:

The 21st century is likely to be one in which leaders of powerful institutions will continue seeking sociocultural expertise to help them accomplish their technological goals, whether for purposes of warfare, commerce, surveillance, or social control. If recent history is any guide, they will ‘crave anthropological knowledge ... they need our spirits – our ability to symbolically and humanly process the human environments these machines dominate.’

5.5.2 Multi-Token Analysis

The analysis of this thesis focused largely on single tokens (i.e., strings of characters unbroken by any whitespace), as opposed to token phrases. This strips tokens away from the context in which they were given. To exemplify this limitation, consider the two documents d_1 and d_2 :

d_1 : *Trump sucks, Biden is awesome.*

d_2 : *Biden sucks, Trump is awesome.*

Clearly, d_1 and d_2 are similar documents but also represent two quite different meanings. Nevertheless, these two documents share the same tf-idf value for every token. Admittedly, d_1 and d_2 are contrived documents, but such similarity demonstrates the limitations of single token methods.

As such, one could extend token methods to consider strings of multiple tokens, in which case the token *Trump sucks* would appear only in document d_1 above, and *Biden sucks* would appear only in document d_2 above. This extension, however, would require considerably more computational power since a dictionary of z tokens corresponds to z^2 possible strings that are two tokens long.

5.5.3 Token Conversions Before Data Analysis

Throughout the data cleaning process, certain concepts with multiple different token references were consolidated to a single common token. For example, different references to the COVID-19 virus (e.g., *rona* and *coronavirus*) were consolidated to the common token *covid* (see Appendix

B). This was done to ensure that different references to the same concept were counted as such: as references to the same concept. This could have unintended consequences. It is reasonable to suppose that the population of people who refer to the COVID-19 virus as *rona* is different than the population of people who refer to the COVID-19 virus as *coronavirus*. Likewise, it is reasonable to suppose that the token *rona* is used primarily in a derogatory manner, whereas *coronavirus* is used primarily in a scientific manner. This subtlety is lost when various references to the same concept are consolidated to a common token. This demonstrates a limitation of the analysis in this thesis. However, note that the concepts that were consolidated to a single token did not appear to be important in this thesis.

5.5.4 Data Privacy

This thesis analyzed social media data. Such an analysis comes with many ethical considerations, especially when dealing with contentious topics (as was the case for this thesis). For example, there are “risk of harm” considerations. [Townsend and Wallace \(2018\)](#) stated:

This risk of harm is most likely where a social media user’s privacy and anonymity have been breached, and is also great when dealing with more sensitive data which when revealed to new audiences might expose a social media user to the risk of embarrassment, reputational damage, or prosecution (to name a few examples).

One must also consider particularly vulnerable populations, such as children or those who are not capable of understanding the accessibility of their data (see [Townsend and Wallace \(2018\)](#) and [Ravn et al. \(2020\)](#)).

These ethical challenges represent a grey area in terms of ‘publicly accessible’ data versus ‘public’ data, especially considering the importance of reproducible research (see [Gentleman and Temple Lang \(2007\)](#)). This thesis dealt with these challenges by favoring data privacy. In particular, this thesis only dealt with data in aggregate. Likewise, to limit the potential risk of harm to the 8kun and Reddit users whose posts were used in this thesis, neither data nor code were published or made publicly available.

CHAPTER 6

Conclusion

This thesis analyzed various political and apolitical forums from 8kun and Reddit (from 8kun: the political newsplus, pnd, and qresearch forums, and from Reddit: the political AskThe_Donald, conservative, conservatives, conspiracy, democrats, and Liberal forums, as well as the apolitical climate, climatechange, immigration, and math forums). Specifically, this thesis analyzed the extent to which one forum's online discourse (in the three months preceding and following January 6, 2021, the date of the US Capitol attack) was distinguishable from another's. It also analyzed the extent to which feelings of isolation and displacement existed on online forums. This chapter concludes the thesis by giving a summary of the results and main points of discussion in Section 6.1, as well as providing opportunities for future research in Section 6.3.

6.1 Summary of Results and Discussions

Of the 13 forums analyzed by this thesis, the newsplus forum had little data (likely this was due to data inaccessibility and not data absence) and was excluded from further analysis. Otherwise, the remaining forums consistently clustered (regardless of individual methods) into five groups as in:

- $$(\star) \left\{ \begin{array}{l} \bullet \text{ The math forum.} \\ \bullet \text{ The climate/climatechange forums.} \\ \bullet \text{ The immigration forum.} \\ \bullet \text{ The qresearch forum.} \\ \bullet \text{ The pnd, conspiracy, conservative, conservatives, AskThe_Donald,} \\ \text{democrats, and Liberal forums.} \end{array} \right.$$

Because this happened so consistently, these are clear groupings. These groupings also largely agree with intuition and prior research. For one, the math, climate/climatechange, immigration forums

were intentionally chosen as apolitical forums. Likewise, QAnon posts (of which the qresearch forum is a hub) are known to be “cryptic” ([Hannah, 2021](#)). The remaining fifth cluster consisting of the pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal forums contains somewhat more standard political forums.

Still, this fifth cluster showed evidence of breaking down further. The pnd and conspiracy forums were distinguished from the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums. Likewise, there was some ability to distinguish the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums along political leanings, though this ability was not strong. In other words, this fifth cluster consisting of the pnd, conspiracy, conservative, conservatives, AskThe_Donald, democrats, and Liberal forums showed topical similarities with pnd and conspiracy being the most unique among them. Then, the conservative, conservatives, AskThe_Donald, democrats, and Liberal forums showed more topical similarity with some minor differences along political leanings.

This further break down largely agrees with intuition and prior knowledge. For one, the pnd forum comes from 8kun, a website known for its extremist content ([Zeng and Schäfer, 2021](#)). Likewise, conspiracy discusses more than just politics. Hence, these two forums were the most distinguishable in this fifth cluster. Similarly, it makes sense that the explicitly political Reddit forums (conservative, conservatives, AskThe_Donald, democrats, and Liberal) had some, but not a drastic, separation along political leanings. That is, they discussed very similar topics, but perhaps from slightly different viewpoints, as one would expect.

This (★) clustering and subsequent refinings were comprehensive results that held for different subsets of data. Specifically, these results held for randomly sampled forum documents (as in Section 4.2), and also for monthly aggregated forum documents (as in Section 4.3).

Given these results, the distinguishable qresearch, pnd, and conspiracy political forums were chosen for a deeper look at individual dates and feelings of isolation and displacement in connection with the January 6, 2021, Capitol attack. Nine dates with discourse similar to January 6, 2021, were analyzed for each of these forums. The eight additional dates were: 2020-10-14, 2020-10-24, 2020-11-03, 2021-01-07, 2021-01-16, 2021-01-17, 2021-01-18, and 2021-01-19. Among these key dates,

there was no proportional increase in posts that used key tokens associated with race, culture, and the great replacement conspiracy theory* (these key tokens were: *attack*, *cultural*, *culture*, *diversity*, *enemies*, *ethnic*, *invaders*, *racial*, and *replacement*). In this way — for the qresearch, pnd, and conspiracy forums from October 1, 2020, to March 31, 2021 — this thesis identified no online feelings of isolation and displacement in connection with the January 6, 2021, Capitol attack.

However, feelings of isolation and displacement still manifested themselves online, particularly on 8kun (there just was not any increase in these feelings on dates with discourse similar to January 6, 2021). This agrees with prior knowledge about 8kun as a host for xenophobic attackers (Baele et al., 2020) and users who believe they have taken the “red pill”† and left normal society (Hannah, 2021).

6.2 Data Acquisition

In total, this thesis scraped over 5 million forum posts from 8kun and Reddit using various Python packages. 8kun data were scraped using the Python *Scrapy* package (Scrapy Developers, 2022a), and an example was given that demonstrates how the package can be used to interact with the HyperText Markup Language (HTML) code of a website. Reddit data were acquired with the Python *PSAW* package (Marx, 2018) and Python *PRAW* package (Boe, 2022). A schematic diagram of the more complicated Reddit data acquisition process was presented, as well as the basic code structure of acquiring Reddit data.

6.3 Future Work

This thesis provides many opportunities for future work. In terms of partitioning forums into clusters, the methods of this thesis did well, but could be improved. In particular, the political Reddit forums (conservative, conservatives, AskThe_Donald, democrats, and Liberal) were largely

*As discussed in Section 1.3, the great replacement conspiracy theory is “a conspiracy theory that postulates white European populations are being demographically and culturally replaced by non-white immigrants through policies enacted by ‘the global elites’” (Carlson and Harris, 2022). Also see Obaidi et al. (2022) or Cosentino (2020) for more information.

†As discussed in Section 5.4.2, taking the “red pill” refers to the 1999 movie *The Matrix*, where the main character (Neo) is offered a choice between a blue pill and red pill — the blue pill returning Neo to his normal life, the red pill revealing the truth about the Matrix: that humanity lost a war against machines and now (unknowingly) lives in a simulated reality known as the Matrix.

indistinguishable. Determining alternate vectorization schemes that can dissect these forums along political leanings presents an opportunity for future research.

In terms of isolation and displacement, this thesis did not identify these feelings online in connection with the events of January 6, 2021. However, this thesis only analyzed the three months preceding and following January 6, 2021. Extending this time frame — perhaps even to a pre-Trump time — could reveal that online feelings of isolation and displacement have been long-term rising, and the events of January 6, 2021, were just the culmination of them. Likewise, the 8kun and Reddit populations are considerably younger than that of the Capitol protestors on January 6, 2021. A social media with an older user base (such as Facebook) may show more feelings of isolation and displacement in connection with the January 6, 2021, Capitol attack. As such, analyzing alternate social media websites (similar to the work of [Butler \(2022\)](#) who analyzed Islamophobic tweets) presents an area of future research.

APPENDICES

APPENDIX A

Reddit Data Acquisition Process

As described in Section 2.2.2, the acquisition process of Reddit data was somewhat onerous. This appendix visually depicts the process in Figure A.1.

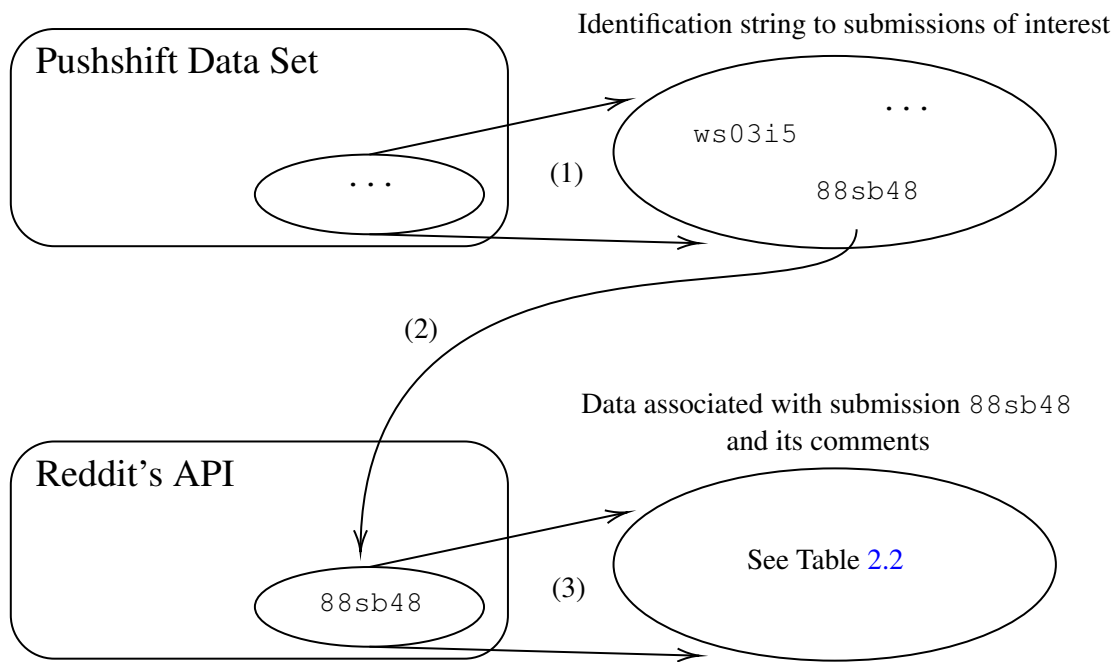


Fig. A.1: The process by which Reddit data were acquired for each forum on each date. (1) The Python *PSAW* package was used to interact with the Pushshift data set and Pushshift API. Specifically the Python *PSAW* package subsets the Pushshift dataset and selects a unique identifier to each submission from the specified forum on the specified date. (2) The Python *PRAW* package was then used to locate each submission from the previous step on Reddit's API. (3) The Python *PRAW* package was then used to access the data associated with that submission. This includes the comment data associated with the submission.

Note that this process may scrape data that occurred after March 31, 2021, the final date considered in this thesis. This is because a submission may have been posted before or on March 31, 2021, but a comment may have occurred after March 31, 2021. However, these comments were always omitted from analysis.

APPENDIX B

Regular Expression Conversions in Data Cleaning

Table B.1 shows the various regular expressions and their replacements used in the data cleaning process (Section 2.3.2). Before converting the regular expressions to their replacements, each text string was all lower case and padded with a space at the beginning and end. Moreover, non alpha-numeric characters were removed and replaced with spaces. Note that there is overlap in usage (e.g., ‘*trump s*’ could have been possessive or a contraction of ‘*trump is*’). The replacements aimed to capture the predominant intended usage.

Table B.1: Replacements made during the data cleaning process (Section 2.3.2).

Regular Expression	Replacement
'black lives matter '	'blm '
'make america great again '	'maga '
'(main stream media) (mainstream media)'	'msm '
'vice president '	'vp '
'(aint) (ain t)'	'is not '
'gonna '	'going to '
'wanna '	'want to '
'(arent) (aren t)'	'are not '
'(cant) (can t) (can not)'	'cannot '
'(didnt) (didn t)'	'did not '
'(doesnt) (doesn t)'	'does not '
'(dont) (don t)'	'do not '
'(hasnt) (hasn t)'	'has not '
'(hes) (he s)'	'he is '
'he ll '	'he will '
'(im) (i m)'	'i am '
'(ive) (i ve)'	'i have '
'i ll '	'i will '
'(isnt) (isn t)'	'is not '
'(its) (it s)'	'it is '
'(itll) (it ll)'	'it will '
'(shes) (she s)'	'she is '
'she ll '	'she will '
'(shouldnt) (shouldn t)'	'should not '
'(thats) (that s)'	'that is '
'(theres) (there s)'	'there is '
'(theyre) (they re)'	'they are '
'(theyve) (they ve)'	'they have '
'(theyll) (they ll)'	'they will '
'(theyd) (they d)'	'they would '
'(wasnt) (wasn t)'	'was not '
'we re '	'we are '
'(werent) (weren t)'	'were not '
'(wont) (won t)'	'will not '
'(wouldnt) (wouldn t)'	'would not '
'(youre) (you re)'	'you are '
'(youve) (you ve)'	'you have '
'(youll) (you ll)'	'you will '
'(bidens) (biden s)'	'biden '
'(trumps) (trump s)'	'trump '
'(harriss) (harris s)'	'harris '
'(pences) (pence s)'	'pence '
'(joes) (joe s)'	'joe '
'(donalds) (donald s)'	'donald '
'(kamalas) (kamala s)'	'kamala '
'(mikes) (mike s)'	'mike '
'(corona) (rona) (coronavirus) (covid 19)'	'covid '

APPENDIX C

List of Stop Tokens

As mentioned in Section 2.3.2 on data cleaning and Section 3.3 on data preprocessing, cleaning, and storage, stop tokens were occasionally removed from text documents. The set of stop tokens used in this thesis can be found in the R *tm* package (Feinerer and Hornik, 2020). The set of 174 tokens are contained in the vector `tm::stopwords()`. Alphabetically, they are:

a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can't, cannot, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, him, himself, his, how, how's, i, i'd, i'll, i'm, i've, if, in, into, is, isn't, it, it's, its, itself, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours, ourselves, out, over, own, same, shan't, she, she'd, she'll, she's, should, shouldn't, so, some, such, than, that, that's, the, their, theirs, them, themselves, then, there, there's, these, they, they'd, they'll, they're, they've, this, those, through, to, too, under, until, up, very, was, wasn't, we, we'd, we'll, we're, we've, were, weren't, what, what's, when, when's, where, where's, which, while, who, who's, whom, why, why's, with, won't, would, wouldn't, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves.

APPENDIX D

Terminology and Notation of this Thesis

Table D.1 shows the various terms with their corresponding definition as related to text documents. Table D.2 shows the various notations with their corresponding meaning as related to text documents. Table D.3 shows the various mathematical notations used throughout this thesis.

Table D.1 will be useful to those unfamiliar with text analysis. Table D.2 will be useful to be comfortable with the notation of this thesis. For those with familiarity of basic set theoretic and linear algebra notation Table D.3 will not strictly be necessary — the only perhaps unusual piece of notation being the Iverson bracket notation $[Q]$ for a Boolean propositions Q .

Back references: Section 3.4, Section 3.5, Section 3.6, and Section 3.7.

Table D.1: Terminology of text documents and their associated meanings. Terms are organized alphabetically for ease of navigation.

<i>Term</i>	Definition
<i>corpus</i>	A set of documents.
<i>dictionary</i>	A set of tokens.
<i>document</i>	A single entity of text data united under a common facet.
<i>token</i>	A single string of characters unbroken by whitespace.
<i>tf</i>	Token frequency. See Table D.2 — $\text{tf}(d, t)$.
<i>tf-idf</i>	Token frequency - inverse document frequency. See Table D.2 — $\text{tf-idf}(C, d, t)$.
<i>vectorization</i>	(of a document) A way to represent a document as a numeric vector.

Table D.2: Notation of text documents. Subscripts may modify the meanings of notation beyond what is shown in this table. When this happens, the meaning will be clear. This table reflects the generic use.

<i>Notation</i>	<i>Meaning</i>
C	A corpus.
d	A document with no data cleaning done.
\bar{d}	A cleaned text document (i.e., applying steps 1 - 8 as in Section 2.3.2).
$\bar{\bar{d}}$	A cleaned text document with stop words removed (i.e., applying steps 1 - 9 as in Section 2.3.2).
\mathbf{d}	A vectorization of the document d .
$\text{ntokens}(d, t)$	The number of occurrences of the token t in the document d .
t	A token.
$t \in d$	The token t appears in the document d .
$\text{tf}(d, t)$	The token frequency (tf) of the token t in the document d . Precisely: $\text{tf}(d, t) = \frac{\text{ntokens}(d, t)}{\sum_{t_i \in d} \text{ntokens}(d, t_i)}$.
$\text{tf-idf}(C, d, t)$	The token frequency - inverse document frequency (tf-idf) of a token t and document d inside of a corpus C . Precisely: $\text{tf-idf}(C, d, t) = \text{tf}(d, t) \cdot \ln \left(\frac{ C }{\sum_{d_i \in C} [t \in d_i]} \right)$.

Table D.3: Mathematical notation used throughout this thesis. Mathematical terms are quoted.

Notation	Meaning
‘set’	A collection of objects, usually denoted with $\{\dots\}$. Sets have no ordering and no repetition.
‘element’	(of a set) An object in a set.
$a \in S$	For a set S , $a \in S$ states that a is an element of S .
$ S $	The cardinality of a set S . For a set S , $ S $ represents the number of elements in S . Because sets do not have repeated elements, $ \{1, 2, 2\} = 2 = \{1, 2\} $.
$S_1 - S_2$	The set of elements that are in set S_1 but not in set S_2 .
$S_1 \cup S_2$	For sets S_1 and S_2 , their union $S_1 \cup S_2$ is the set of elements that are in at least one of S_1 and S_2 .
$S_1 \cap S_2$	For sets S_1 and S_2 , their intersection $S_1 \cap S_2$ is the set of elements that are in both S_1 and S_2 .
\cos	The usual cosine function.
\log_b and \ln	The logarithmic function in base b . When b is Euler’s constant $e = 2.71\dots$, \log_b is denoted as \ln . In graphics, \log_b appears as <code>log_b</code> or <code>logb</code> .
$[Q]$	For a boolean proposition Q , $[Q]$ is 1 if Q is true and 0 if Q is false.
$\sum_{i \in S} f(i)$	For a set of elements $S = \{s_1, \dots, s_n\}$ and an expression f of such elements, $\sum_{i \in S} f(i) = f(s_1) + f(s_2) + \dots + f(s_n)$. When a and b are integers with $a < b$ and $S = \{a, a + 1, \dots, b - 1, b\}$, this is denoted as $\sum_{i=a}^b f(i)$.
\mathbf{x}, \mathbf{y} , etc.	Vectors.
$(\mathbf{x})_i$	The i^{th} component of a vector. That is, for an n -dimensional vector $\mathbf{x} = (x_1, \dots, x_n)$, $(\mathbf{x})_i = x_i$.
$\mathbf{x} + \mathbf{y}, \mathbf{x} - \mathbf{y}$	For n dimensional vectors \mathbf{x} and \mathbf{y} , $\mathbf{x} + \mathbf{y}$ (respectively, $\mathbf{x} - \mathbf{y}$) is the vector whose i^{th} component is $(\mathbf{x})_i + (\mathbf{y})_i$ (respectively, $(\mathbf{x})_i - (\mathbf{y})_i$)
$\ \mathbf{x}\ _\alpha$	For an n -dimensional vector \mathbf{x} and real number $\alpha \geq 1$, $\ \mathbf{x}\ _\alpha = (\sum_{i=1}^n ((\mathbf{x})_i)^\alpha)^{1/\alpha}$.
$\mathbf{x}^T \mathbf{y}$	The usual inner product of vectors. That is, for n -dimensional vectors \mathbf{x} and \mathbf{y} , $\mathbf{x}^T \mathbf{y} = (\mathbf{x})_1 \cdot (\mathbf{y})_1 + \dots + (\mathbf{x})_n \cdot (\mathbf{y})_n$.
distance_X	A generic distance function for vectors \mathbf{x} and \mathbf{y} .
$\text{distance}_{\text{Jaccard}}$	The Jaccard distance function defined by $\text{distance}_{\text{Jaccard}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{(\ \mathbf{x}\ _2^2 + \ \mathbf{y}\ _2^2 - \mathbf{x}^T \mathbf{y})}$.
$\text{distance}_{\text{Euclidean}}$	The Euclidean distance function defined by $\text{distance}_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \ \mathbf{x} - \mathbf{y}\ _2$.

APPENDIX E

Details of the word2vec Neural Network

As discussed in Section 3.6.2, the word2vec machine learning software embeds tokens as vectors. This appendix provides an overview of the training process of the software. Exact details can be found in Mikolov et al. (2013a) and Mikolov et al. (2013b), as well as in the word2vec patent (Mikolov et al., 2021).

The word2vec neural network learns token embeddings in a ‘guess the missing token’ fashion. In particular:

- (1) Assign each token an initial embedding.
- (2) For each token t in the corpus, consider the preceding w and following w tokens. Average the embeddings of those $2w$ tokens.
- (3) Determine the relative probabilities of the missing token t by determining which token embeddings are closest to the average embedding produced by step (2).
- (4) Update token embeddings based upon the cross-entropy loss function. Updates are made using stochastic gradient descent with an adaptive learning rate (specifically, Adagrad as in Duchi et al. (2011)) and backpropagation.

APPENDIX F

Scaled Squared Euclidean Distance versus Cosine Distance

As mentioned in Section 3.10.5, the squared Euclidean distance is proportional to the cosine distance when vectors are normalized. Note that, for vectors \mathbf{x} and \mathbf{y} the cosine distance is defined as

$$\text{distance}_{\text{cosine}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

As such, for normalized vectors \mathbf{x} and \mathbf{y} (i.e. $\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2$), their squared Euclidean distance is

$$\begin{aligned} (\text{distance}_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}))^2 &= (\|\mathbf{x} - \mathbf{y}\|_2)^2 \\ &= (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) \\ &= (\mathbf{x}^T - \mathbf{y}^T) (\mathbf{x} - \mathbf{y}) \\ &= \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{y} - \mathbf{y}^T \mathbf{x} + \mathbf{y}^T \mathbf{y} \\ &= (\|\mathbf{x}\|_2)^2 - \mathbf{x}^T \mathbf{y} - \mathbf{y}^T \mathbf{x} + (\|\mathbf{y}\|_2)^2 \\ &= 2 - 2\mathbf{x}^T \mathbf{y} \\ &= 2 \left(1 - \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \right) \\ &= 2 \cdot \text{distance}_{\text{cosine}}(\mathbf{x}, \mathbf{y}) \\ &\propto \text{distance}_{\text{cosine}}(\mathbf{x}, \mathbf{y}). \end{aligned}$$

APPENDIX G

Criteria from the R *NbClust* Package ([Charrad et al., 2022](#))

The following list shows all available criteria in the R *NbClust* package for determining the optimal number of clusters from a hierarchical clustering, as discussed in Section 3.8.3. Those bolded were used in analysis. Those marked with a ‘★’ were excluded because they required graphical interpretation. Those marked with a ‘†’ were excluded because the method was incompatible with the vectorizations (specifically, the matrix of vectorizations were often rank deficient and the method required full rank). Those marked with a ‘?’ were excluded because they resulted in unknown errors.

kl, ch, hartigan, ccc[†], scott[†], marriot[†], trcovw[†], tracew[†], friedman[†], rubin[†], cindex, db, silhouette, duda, pseudot2, beale[?], ratkowsky[?], ball, ptbiserial, gap, frey[?], mcclain, gamma, gplus, dunn, hubert★, sdindex, dindex★, sdbw

The details of these criteria are not presented here. Those wishing to see the details may refer to the R *NbClust* package.

APPENDIX H

Data Summary Statistics for Alternative Forums

As in Section 4.1, Table H.1, Figure H.1, and Figure H.2 show summary statistics for each of the seven alternative forums used in this thesis. As with the main forums of study, most posting data came from comments, not submissions. The alternative forums had much more similar posting quantities than did the main forums of study (compare the alternative forum statistics in Table H.1, Figure H.1, and Figure H.2 to the main forum statistics in Table 4.1, Figure 4.1, and Figure 4.2). As in Table H.1, the quantity of posts for the democrats forum was on the order of 100,000, whereas all other alternative forums were on the order of 10,000.

Table H.1: For the seven alternative forums, summary statistics of the total number of posts, submissions, and comments, as well as summary statistics for the number and percent of dates with no posts.

	Number of:			Number of Dates	Percent of Dates
	Posts	Submissions	Comments	with No Posts	with No Posts
climate	14,113	3,134	10,979	7	3.85 %
climatechange	13,293	1,830	11,463	2	1.10 %
conservatives	92,351	12,698	79,653	7	3.85 %
democrats	102,073	16,067	86,006	3	1.65 %
immigration	43,338	6,399	36,939	0	0.00 %
Liberal	20,148	3,817	16,331	4	2.20 %
math	80,308	11,220	69,088	0	0.00 %

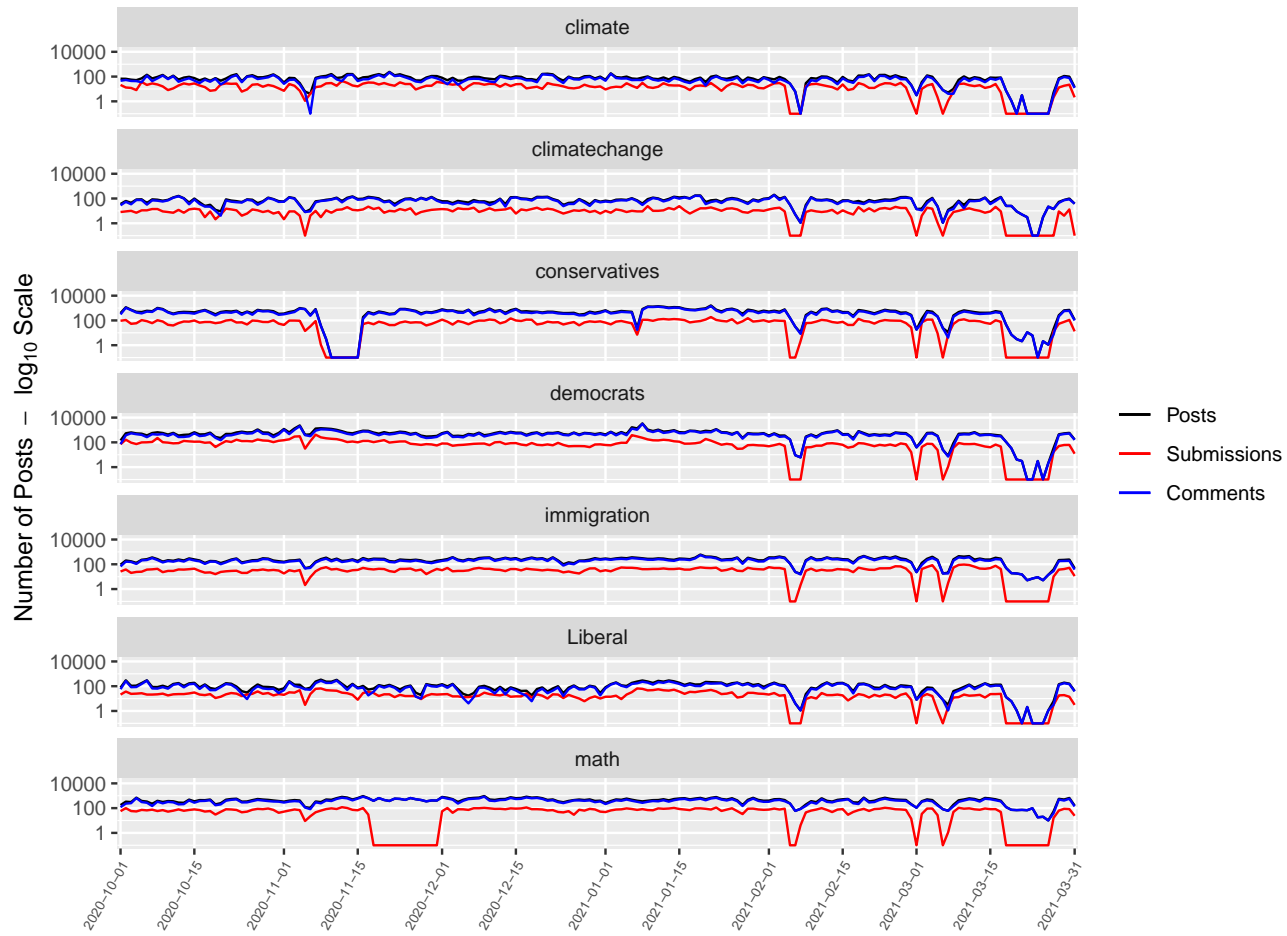


Fig. H.1: Posts, submissions, and comments per day for each of the seven alternative forums (log 10 scale). The black and blue lines (corresponding to posts and comments, respectively) nearly overlap.

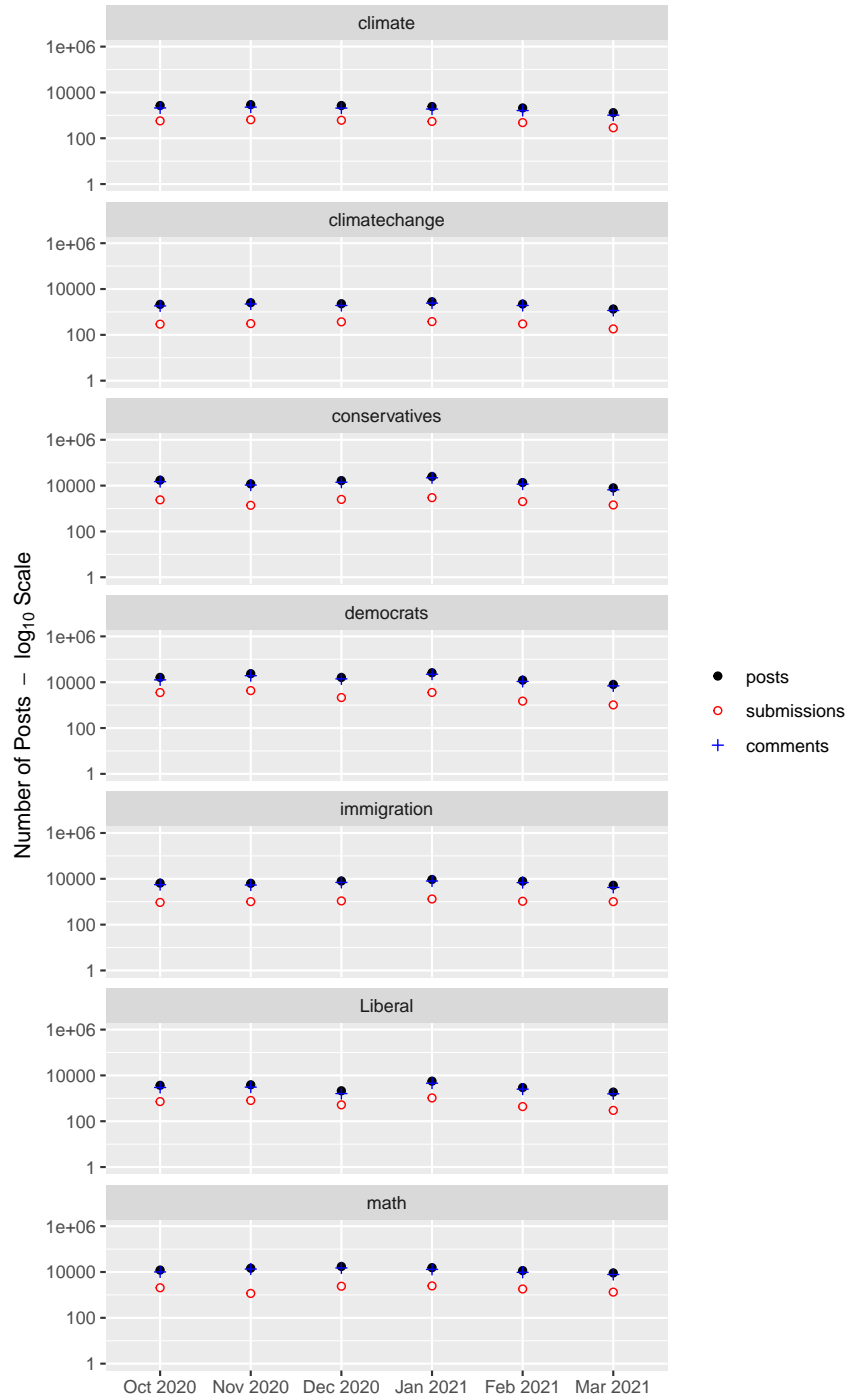


Fig. H.2: Posts, submissions, and comments per month for each of the seven alternative forums (log 10 scale). The black and blue marks (corresponding to posts and comments, respectively) nearly overlap.

APPENDIX I

Analogous Results to Table 4.2 and Table 4.7

Section 4.2 took three random samples of data from each (main and alternative) forum — i.e., AskThe_Donald, conservative, conspiracy, pnd, qresearch, climate, climatechange, immigration, conservatives, democrats, Liberal, and math — and sought to identify clusters of such documents using 18 different criteria from the R *NbClust* package. Random samples of data were taken to both allow the R *NbClust* package to run in a reasonable time frame, and also to determine the general patterns of how forums intra and inter cluster. This was done with 10 different random seeds. Table I.1 – Table I.18 show the results for the 18 different criteria. Each table fixes a single criteria from the R *NbClust* package, and shows the results for the various random seeds, linkage algorithms, and ranges of identifiable clusters. Section 5.1.1 discusses the major features of these tables.

Table I.1: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the kl criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7			
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7			
72227	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7			
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7			
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			

Table I.2: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the ch criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe.Donald			democrats			Liberal			# Clusters	
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3					
14741	Ward	1	1	2	3	4	5	6	7	7	8	9	10	9	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34	
37573	Ward	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34	34	
12421	Ward	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
15451	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
98689	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
73037	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
14741	average	1	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
37573	average	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
12421	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
15451	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9		
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
73037	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
14741	Ward	5	1	2	3	4	5	6	7	7	8	9	10	9	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34	34
37573	Ward	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	21	22	23	24	25	26	26	27	28	29	30	31	32	33	33	33
12421	Ward	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
15451	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
98689	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
73037	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		
14741	average	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
37573	average	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25	26	26	27	28	29	30	31	32	33	34	34	
12421	average	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
15451	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9		
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
73037	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10		

Table I.3: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the hartigan criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
14741	Ward	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	1	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	Ward	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	5	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	25	27	28	29	30	31	32	33	34	35	35
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

Table I.4: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the cindex criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
14741	Ward	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	1	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	Ward	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	5	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	25	27	28	29	30	31	32	33	34	35	35
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

Table I.5: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the db criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			#Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
14741	Ward	1	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	Ward	1	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	1	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	1	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	average	1	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	1	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	Ward	5	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	Ward	5	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	5	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	5	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	5	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	average	5	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	5	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	25	27	28	29	30	31	32	33	34	35	35
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

Table I.6: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the silhouette criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
14741	Ward	1	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	Ward	1	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	1	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	1	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	average	1	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	1	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	Ward	5	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	Ward	5	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	Ward	5	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	5	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	5	1	2	3	4	5	6	7	7	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	32
37573	average	5	1	2	3	4	5	6	7	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	5	1	2	3	4	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	34
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	25	27	28	29	30	31	32	33	34	35	35
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

Table I.7: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the duda criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	10	10	10	10	10	10	10	10	12		
37573	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	11	11	11	
10601	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	9	9	9	9	10	10	10	11	12	12	11	11	11	
12421	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	7	7	8	8	8	9	9	9	9	10	10	10	11	11	11	9	9	9
72227	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	8	8	10	10	10	9	9	9	9	9	10	
15451	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	11	10	11	11	11	11	
98689	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
18181	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
97079	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	8	10	10	10	11	11	11	12	12	12	
73037	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	9	10	10	10	10	9	9	9	9	9	10	
37573	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
10601	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	7	7	8	8	8	8	8	8	8	8	8	8	8	8	8	
12421	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	
72227	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	
15451	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9	
98689	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	8	8	9	
18181	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
97079	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
73037	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	10	10	10	10	10	10	10	10	10	12	
37573	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	11	11	11	
10601	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	9	9	9	9	10	10	10	11	12	12	11	11	11	
12421	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9	9	9	9	9	9	9	9	9		
72227	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	8	8	9	9	9	9	10	10	10	10	10	
15451	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	11	10	11	11	11	
98689	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
18181	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
97079	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	8	10	10	10	11	11	11	12	12	12	
73037	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	9	10	10	10	10	9	9	9	9	9	10	
37573	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7		
10601	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	8	8	8	8	8	9		
12421	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7		
72227	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7		
15451	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9	
98689	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	8	8	9	
18181	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7		
97079	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
73037	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	

Table I.8: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the pseudot2 criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	10	10	10	10	10	10	10	10	12		
37573	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	11	11	11	11	
10601	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	9	9	9	10	10	10	10	10	10	11	11	11	12	
12421	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	7	7	8	8	8	9	9	9	10	10	10	11	11	11	9	9	9	11
72227	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	8	8	10	10	10	9	9	9	9	9	10	
15451	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	11	10	11	11	11	11	
98689	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
18181	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
97079	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	8	10	10	10	11	11	11	12	12	12	
73037	Ward	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	9	10	10	10	10	9	9	9	9	9	10	
37573	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
10601	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	7	7	8	8	8	8	8	8	8	8	8	8	8	8	8	
12421	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	8	8	8	8	8	8	8	8	8	8	8	8	
72227	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
15451	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9	
98689	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	8	8	9	
18181	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
97079	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	10	10	10	10	
73037	average	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	10	10	10	10	10	10	10	10	10	10	
37573	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	10	10	10	11	11	11	11	
10601	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	9	9	9	9	10	10	10	11	12	12	11	11	11	
12421	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9	9	9	9	9	9	9	9	9	10	
72227	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	8	8	9	9	9	9	10	10	10	10	10	
15451	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	11	10	11	11	11	
98689	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
18181	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
97079	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	8	10	10	10	11	11	11	12	12	12	
73037	Ward	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	
14741	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	9	10	10	10	10	9	9	9	9	9	10	
37573	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
10601	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	8	9	9	8	8	8	9	
12421	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	8	
72227	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
15451	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	9	9	9	
98689	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	8	8	8	8	8	9	
18181	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	
97079	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	9	9	9	10	10	10	10	10	10	
73037	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	8	8	9	9	9	9	10	10	10	10	10	10	

Table I.9: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the ball criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
37573	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
10601	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
12421	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
72227	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
15451	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
98689	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
18181	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
97079	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
73037	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
14741	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
37573	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
10601	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
12421	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
72227	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
15451	average	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
98689	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
18181	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2			
97079	average	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
73037	average	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
14741	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
37573	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
10601	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
12421	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
72227	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
15451	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
98689	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
18181	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
97079	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			
73037	average	5	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6			

Table I.10: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the ptbserial criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			

Table I.11: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the gap criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
37573	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
10601	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
12421	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
72227	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
15451	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
98689	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
18181	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
97079	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
73037	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
14741	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
37573	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
10601	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
12421	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
72227	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
15451	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
98689	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
18181	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
97079	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
73037	average	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6				
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				

Table I.12: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the mcclain criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
37573	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
10601	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
12421	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
72227	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
15451	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
98689	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
18181	Ward	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
97079	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
73037	Ward	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
14741	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
37573	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
10601	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
12421	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
72227	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
15451	average	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
98689	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
18181	average	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2				
97079	average	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
73037	average	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2				
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				

Table I.13: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the gamma criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				

Table I.14: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the gplus criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				

Table I.15: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the tau criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5				

Table I.16: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the dunn criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			

Table I.17: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `sdindex` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe_Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	Ward	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
14741	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
37573	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
10601	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
12421	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
72227	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
15451	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
98689	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
18181	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
97079	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			
73037	average	5	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5			

Table I.18: For each random seed, linkage algorithm, and minimum number of identifiable clusters (`min.nc`, in the notation of the R *NbClust* package), the identified clusters by the `sdbw` criteria of the R *NbClust* package. Within each row, documents sharing the same tabular value are clustered together according to that criteria.

Seed	Linkage	min.nc	math			climate			climatechange			immigration			qresearch			pnd			conspiracy			conservative			conservatives			AskThe.Donald			democrats			Liberal			# Clusters
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3				
14741	Ward	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
12421	Ward	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	1	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	1	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	1	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	1	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	1	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	average	1	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	1	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	1	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	Ward	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	Ward	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	Ward	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35	
12421	Ward	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	Ward	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	Ward	5	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
98689	Ward	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	Ward	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	Ward	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	Ward	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
14741	average	5	1	2	3	4	5	6	7	8	9	10	11	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
37573	average	5	1	2	3	4	5	6	7	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
10601	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
12421	average	5	1	2	3	4	5	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
72227	average	5	1	2	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
15451	average	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	25	27	28	29	30	31	32	33	34	35	35
98689	average	5	1	2	3	4	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
18181	average	5	1	2	3	4	5	6	7	8	9	10	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
97079	average	5	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35
73037	average	5	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	35

APPENDIX J

Analogous Results to Figure 4.3 – Figure 4.6

Figure 4.3 – Figure 4.6 in Section 4.3 show heatmap visualizations for each (main and alternative) forum, over the six months of data. That is, each individual document in Figure 4.3 – Figure 4.6 corresponds to all the posts from a single forum over one month (see Procedure 3.6.1.a). The tokens were chosen because they had the $\eta = 50, 100, 150, 200$ largest tf-idf values among all token/document pairs, respectively (see Procedure 3.6.1.b). The heatmaps were clustered with the Ward linkage algorithm. The same figures as Figure 4.3, Figure 4.4, Figure 4.5, and Figure 4.6 — but with the average linkage algorithm — are displayed in Figure J.1, Figure J.2, and Figure J.3, and Figure J.4, respectively. The results are not appreciably different than those that are described in Section 4.3.

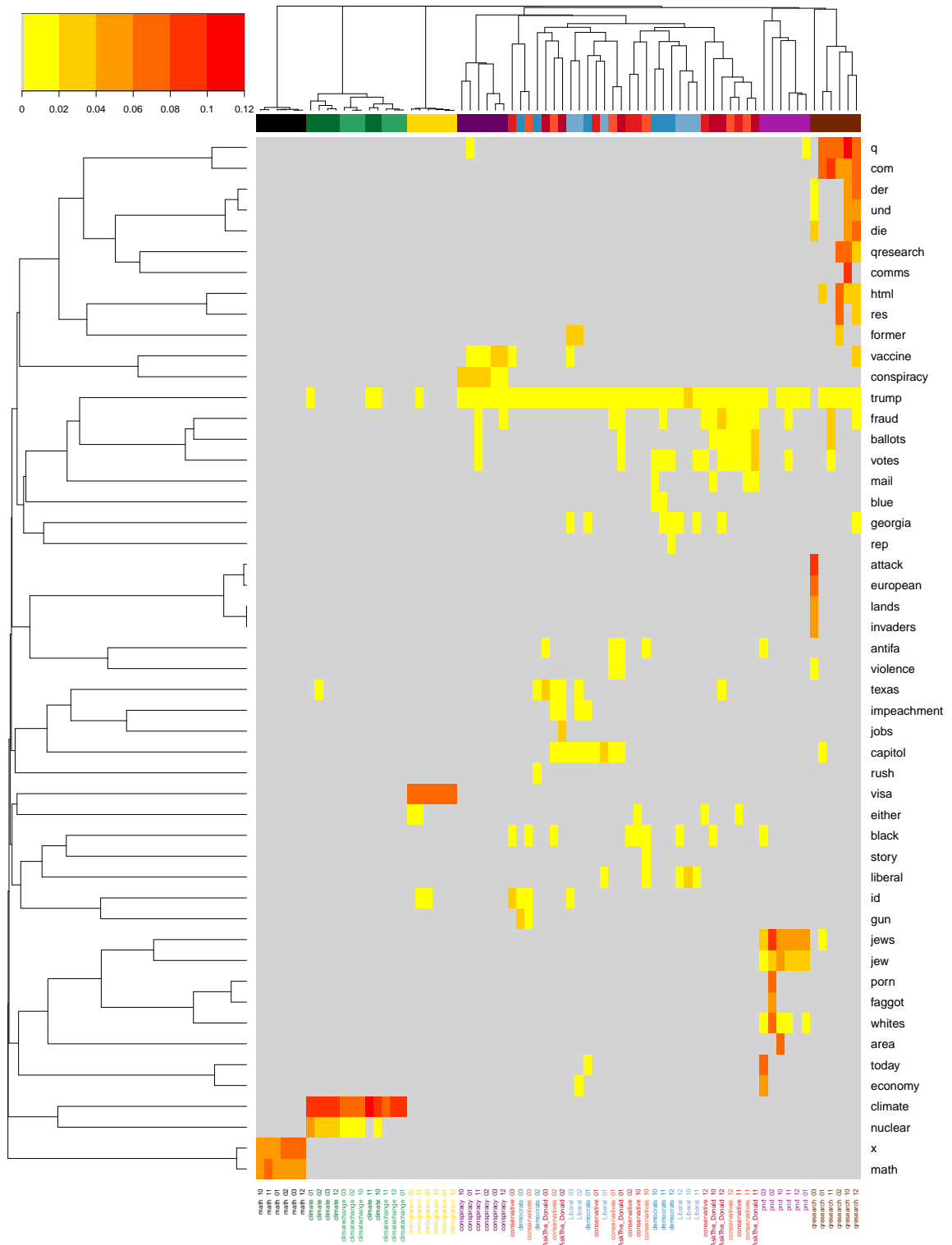


Fig. J.1: An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

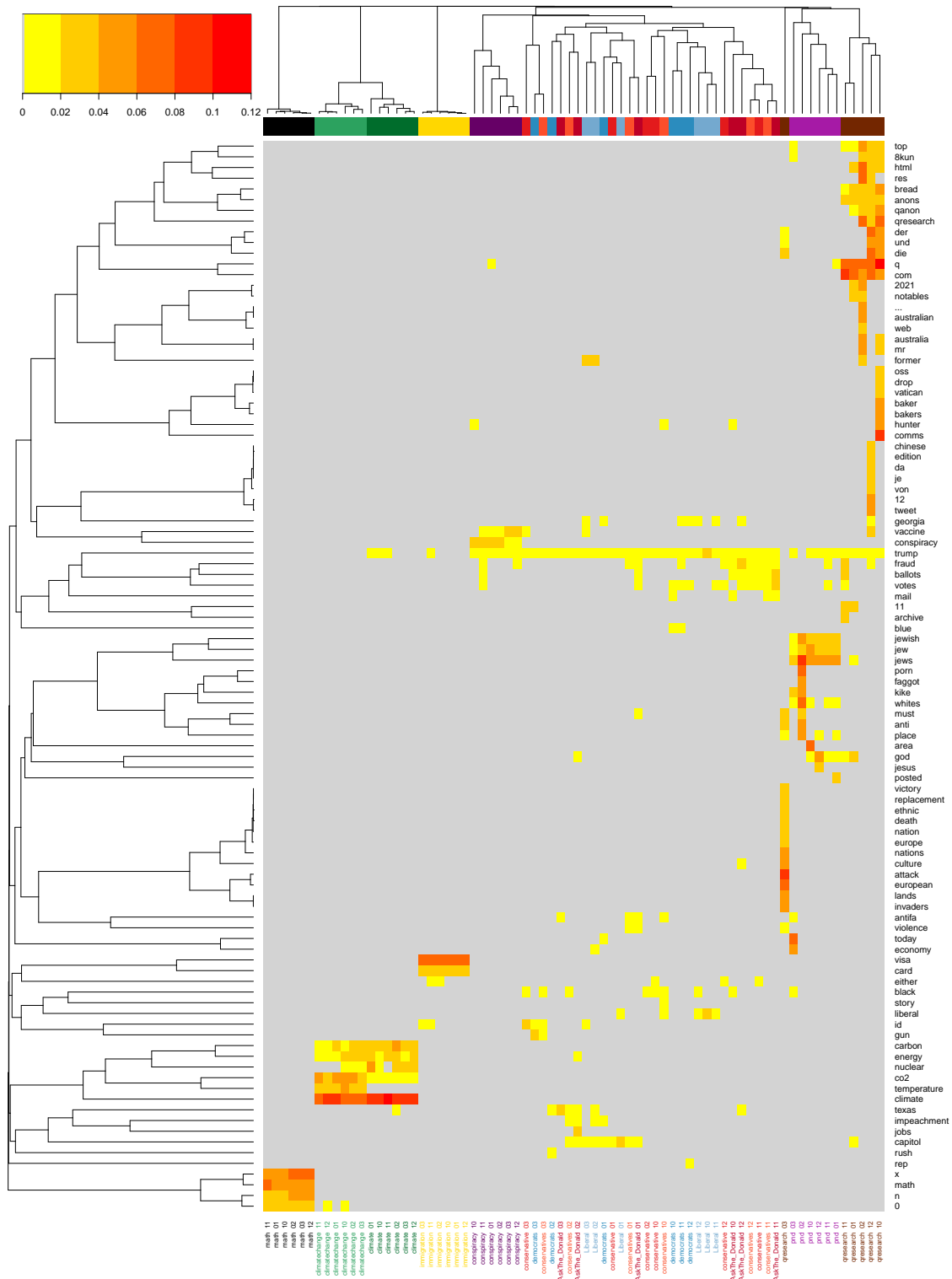


Fig. J.2: An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.



Fig. J.3: An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

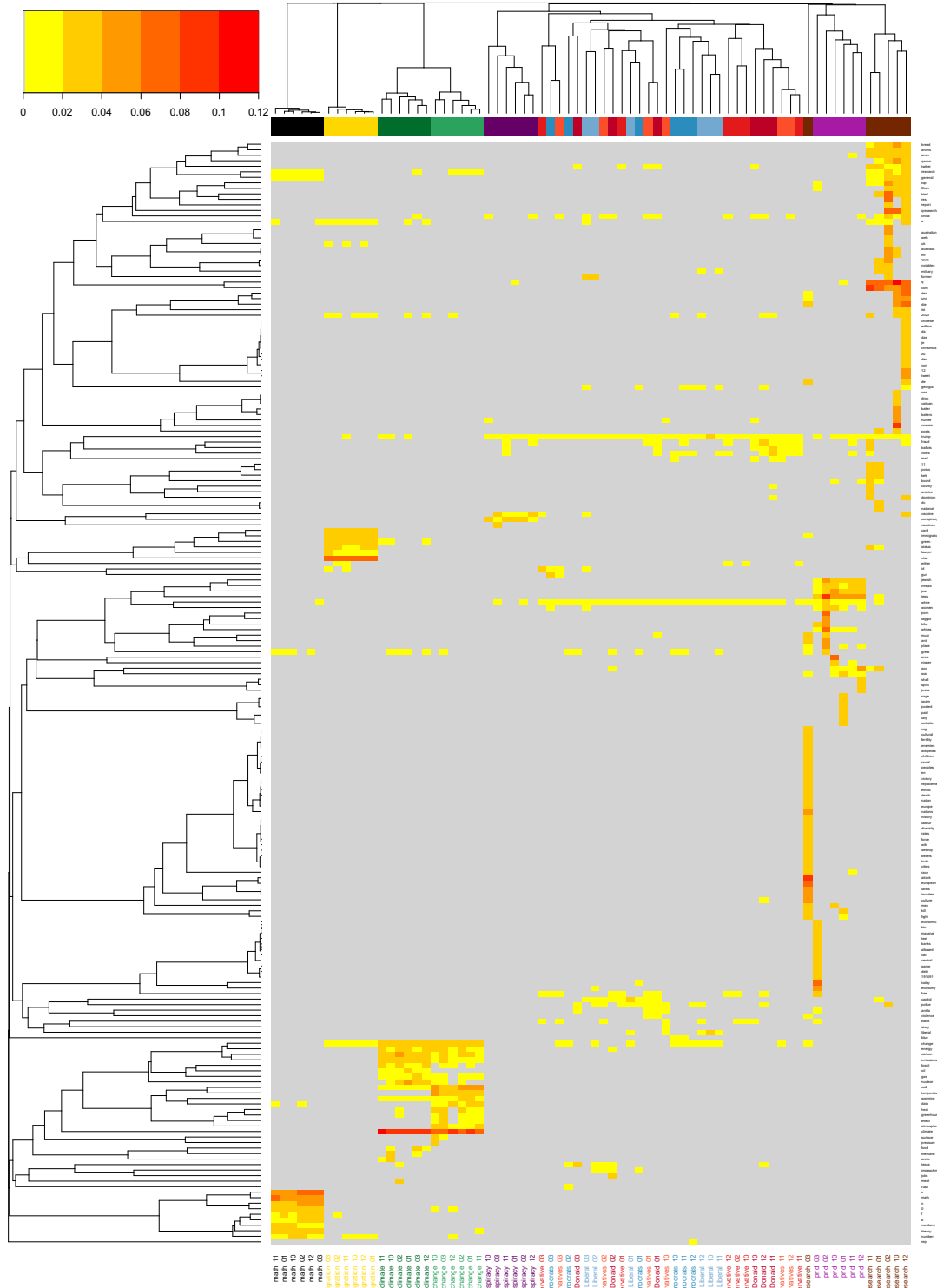


Fig. J.4: An average linkage heatmap of all (main and alternative) forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

APPENDIX K

Analogous Results to Figure 4.7 – Figure 4.10

Figure 4.7 – Figure 4.10 in Section 4.3 show heatmap visualizations for each devoted political Reddit forum, over the six months of data. That is, each individual document in Figure 4.7 – Figure 4.10 corresponds to all the posts from a single forum over one month (see Procedure 3.6.1.a). The tokens were chosen because they had the $\eta = 50, 100, 150, 200$ largest tf-idf values among all token/document pairs, respectively (see Procedure 3.6.1.b). The heatmaps were clustered with the Ward linkage algorithm. The same figures as Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10 — but with the average linkage algorithm — are displayed in Figure K.1, Figure K.2, Figure K.3, and Figure K.4, respectively. The results are not appreciably different than those that are described in Section 4.3.

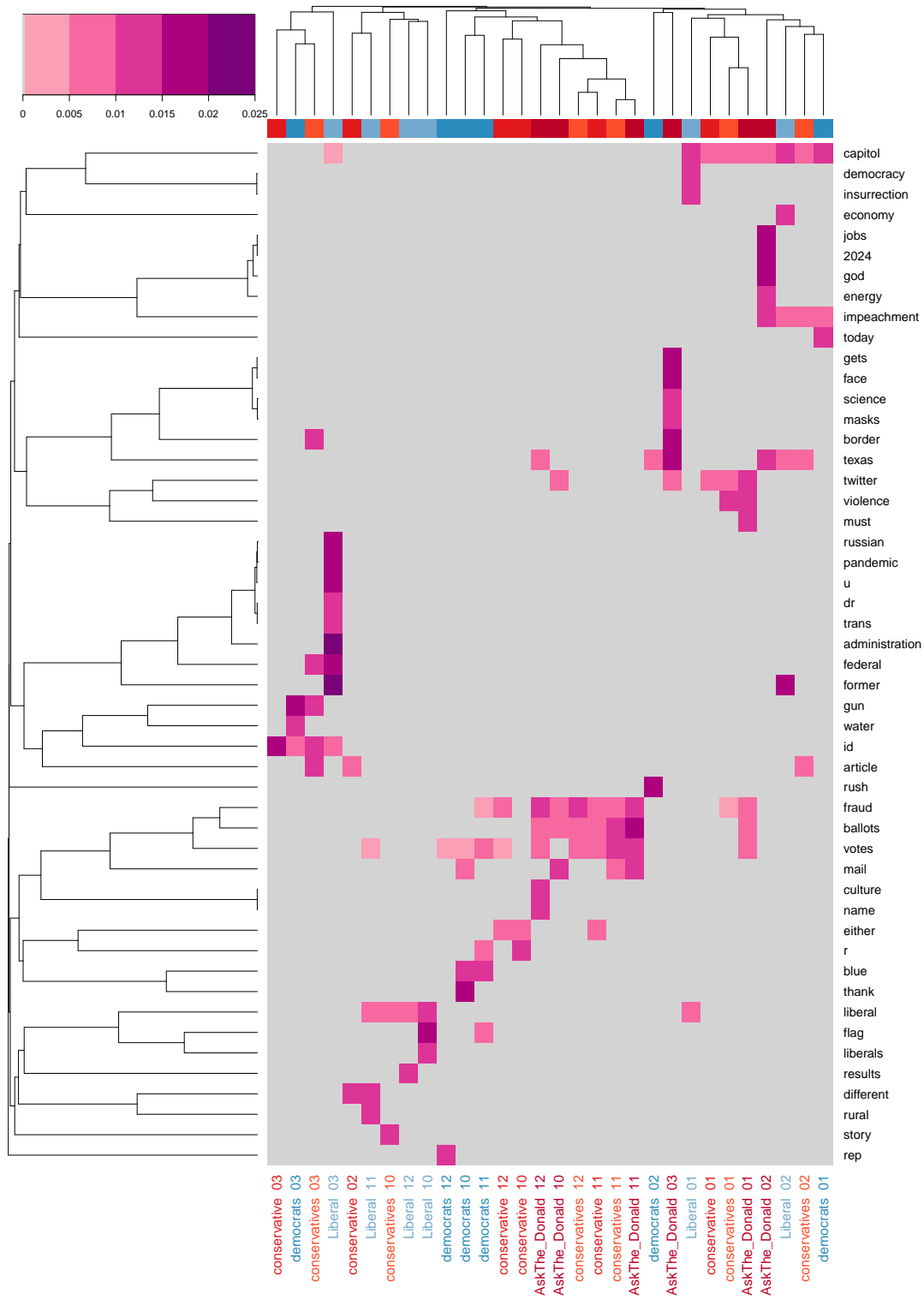


Fig. K.1: An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 50$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.



Fig. K.2: An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 100$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

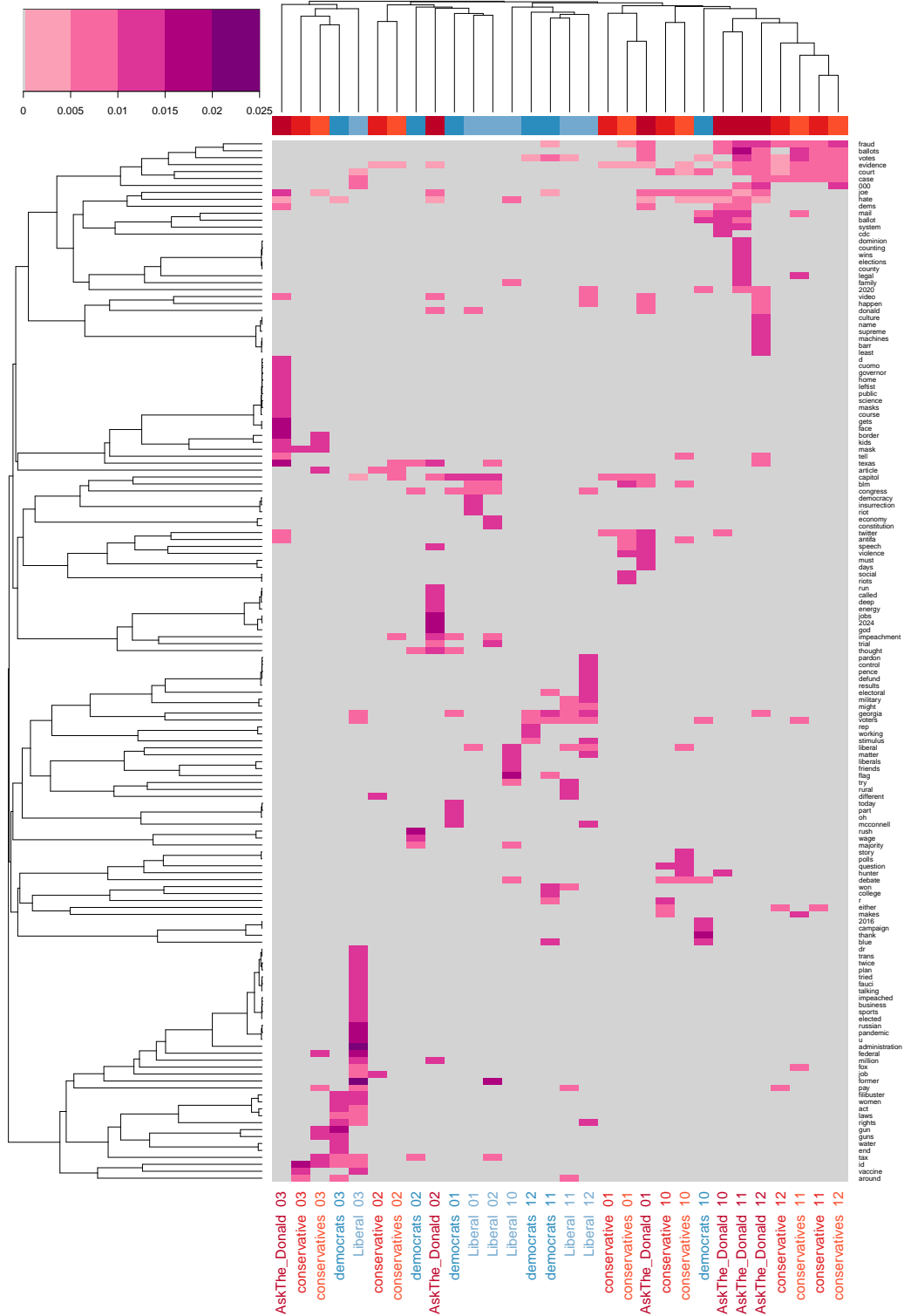


Fig. K.3: An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 150$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

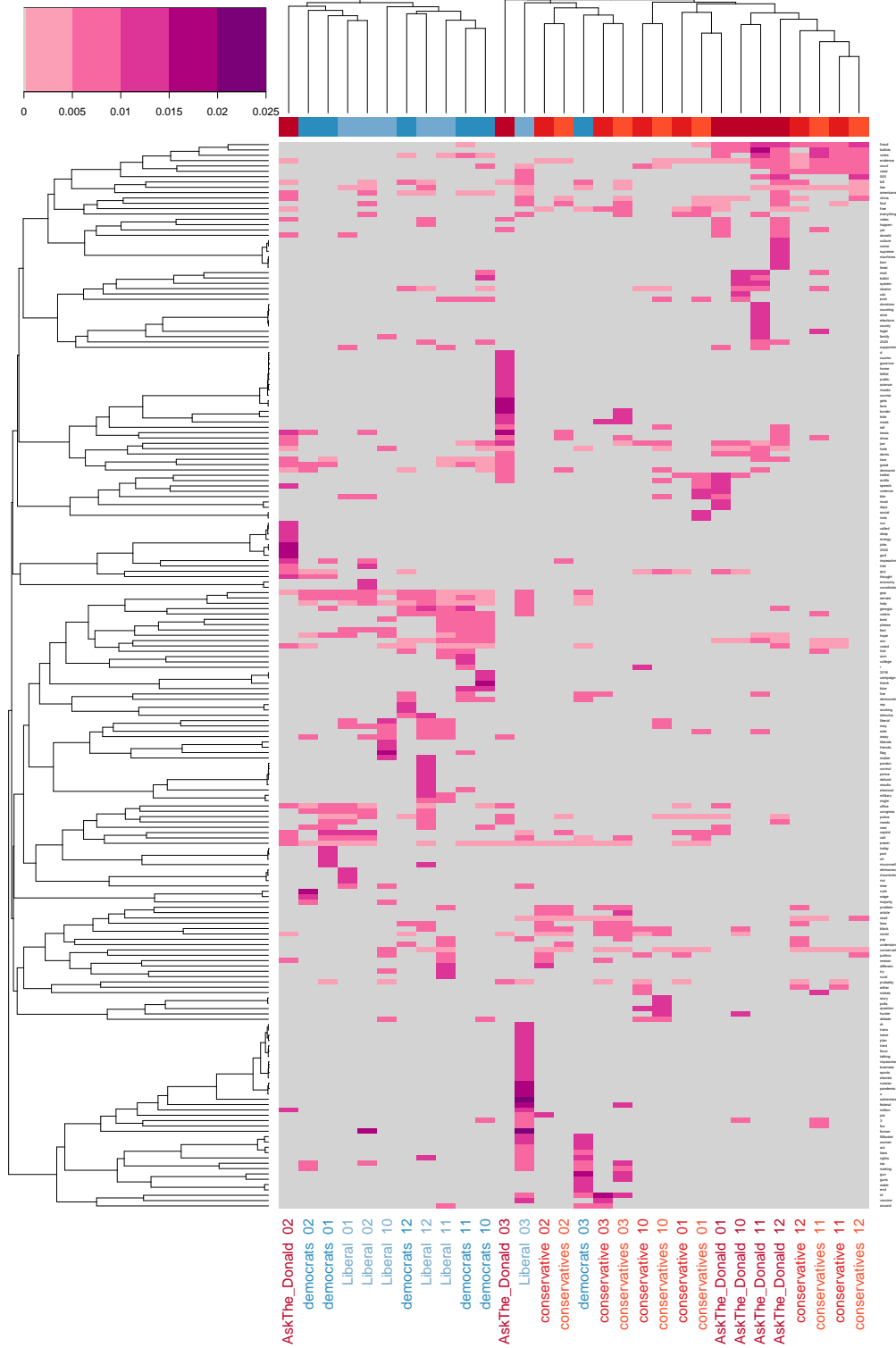


Fig. K.4: An average linkage heatmap of the devoted political Reddit forums, split across forums and months. The top $\eta = 200$ tokens with the highest tf-idf value were chosen for the heatmap. More details are provided in the main text.

APPENDIX L

Analogous Results to Figure 4.11 – Figure 4.14 and Figure 4.15

Figure 4.11 – Figure 4.14 and Figure 4.15 in Section 4.3 show heatmap clustering results with a larger number of tokens ($\eta > 200$) than those displayed in Figure 4.3 – Figure 4.10 and the analogous figures in Appendix J and Appendix K. Each of Figure 4.11, Figure 4.12, Figure 4.13, Figure 4.14, and Figure 4.15 used the Ward linkage clustering algorithm. The analogous plots — using the average linkage clustering algorithm — are displayed in Figure L.1, Figure L.2, Figure L.3, Figure L.4, and Figure L.5, respectively. The results are similar (but not quite as strong) as those that are described in Section 4.3.

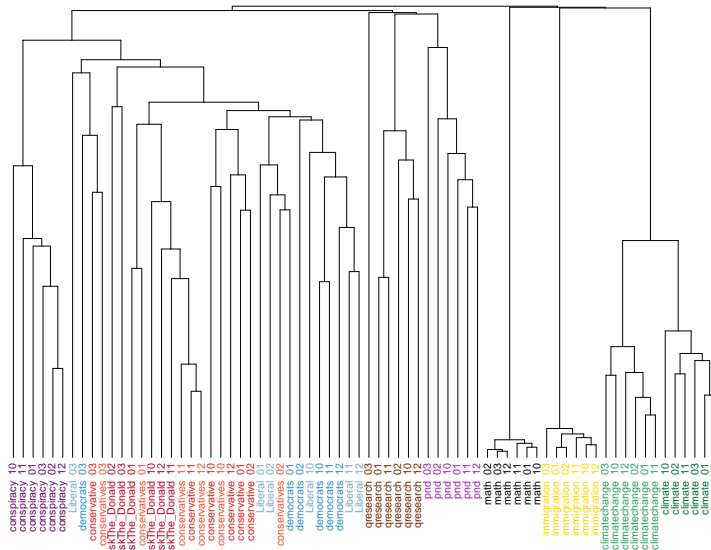


Fig. L.1: An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 300$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

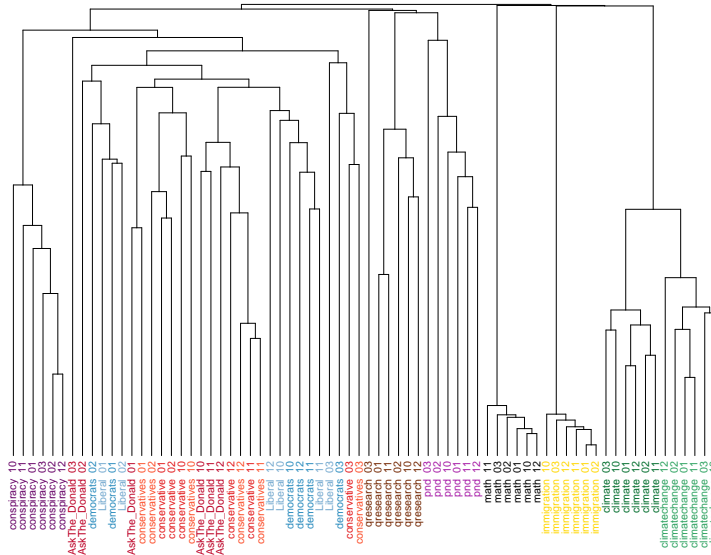


Fig. L.2: An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 500$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

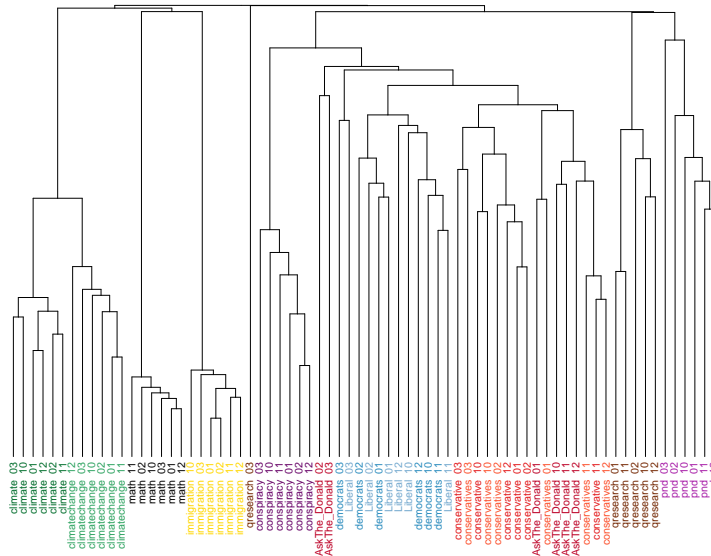


Fig. L.3: An average linkage dendrogram of all (main and alternative) forums, split across forums and months. The top $\eta = 750$ tokens with the highest tf-idf value were chosen for the heatmap (not pictured). More details are provided in the main text.

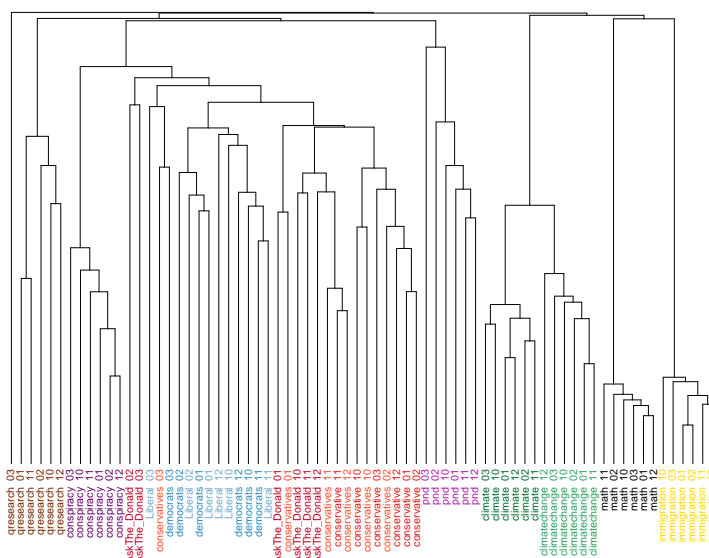


Fig. L.4: An average linkage dendrogram of all (main and alternative) forums, split across forums and months. All $\eta = 821$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.

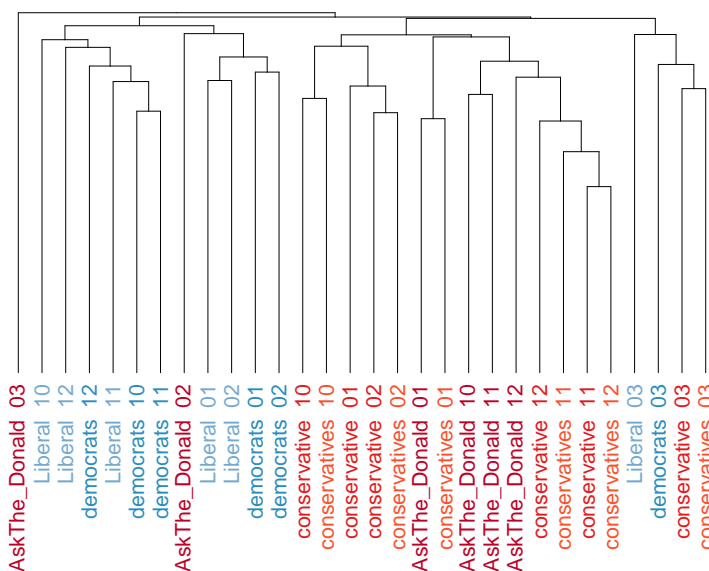


Fig. L.5: An average linkage dendrogram of the devoted political Reddit forums, split across forums and months. All $\eta = 281$ tokens with positive tf-idf values were chosen for the heatmap (not pictured). More details are provided in the main text.

APPENDIX M

The First 100 Tokens of Randomly Selected Forum Documents

To give the reader an idea of typical posts from each forum, the following text strings show the first 100 tokens of a randomly created document from the AskThe_Donald, conservative, conspiracy, pnd, qresearch, climate, climatechange, immigration, conservatives, democrats, Liberal, and math forums. Note that the newsplus forum was excluded from analysis due to lack of data (see Section 4.1). These documents have been cleaned (and so there is no punctuation or capital letters) and consist of more than just one post each. These things make the strings difficult to read, but the overall meaning of each can still be understood. Large integer tokens exist for the pnd and qresearch forums (both of which are from 8kun). 8kun users use these tokens to refer to other posts (which are numbered). Derogatory (or otherwise offensive) tokens may appear.

AskThe_Donald: and joe biden was molesting taking bribes
 this is not even true lmao gas is people dependent in my
 area it will be 6 in southern delaware by the 4th joe is
 being propped up think weekend at bernies for one reason
 so he can shuffle in to the oval office sign some paperwork
 then shuffle back out before he soils his depends the reason
 he does not take questions is because he does not have a
 clue as to what he is signing biggest scam in american history
 half do not exist as problems the rest are their own

conservative: not in a political way just as an online friend
 use adblock when you can and do not click on ads and try
 to avoid accepting cookies when you can and after visiting
 website clear cookies for sites you do not trust i wanted
 trump to win maybe you are a democrat who wanted biden to

win that does not matter my advice here will maybe hopefully
 help you not get viruses and maybe add some more civility
 to discourse gold leaf covering a plastic shell for over
 inflated valuation defunding the police is only part of the
 democrat war

conspiracy: trump bottomless fed do you believe this play
 book is from trump because this is mmt in action works k
 that does not mean you cannot talk about it here yeah growing
 up going through catholic schools but having open minded
 parents was the worst question anything within the old or
 new testament and you area sinner but agreed in one text
 he is a big baby and the second he is the saviour of men
 by sacraficing his own son sounds like a egotistical narcissist
 to me lol the stuff that was in warehouses was on the republicans
 watch

pnd: 167670 lockdown inspired suicides on course to dwarf
 covid deaths in australia in time even in us studies lockdown
 health scandal 2 000 extra heart attack deaths 220468 if
 you want to get rid of the jews you got to get rid of the
 central banks that is how the jews fuck over the economy
 collusion with corrupted governance 183661 posts a sys link
 that only mods and admins can access no one else can see
 it pretends not to be the board owner that is absolutely
 fucking hilarious you couldn t be more incompetent if you
 tried google drive

qresearch: any else notice most the q grifters have gone dark
 joe m l lin wood sidney powell where these loudmouth motherfuckers

now 12507953 quartz valley indian reservation searched it
 on duckduckgo and got a page linking to marine corps acronyms
 but qvir was not on the page weird 12444485 weird only post
 not loading the pix for me 11986400 the y head the hebrew
 letter ayin meaning eye 12687227 the absolute state of the
 democratic party 12534696 hot take 11606882 just working
 from what an anon posted yesterday all 14 posts 12563079
 i am not tired of winning yet

climate: conspiracy theories are smoke and mirrors to make
 everyone who talks about 5g surveillance and energy waste
 look like a crazy poor circulation means a lack of nutrients
 moving through ocean currents dead zones and depleted oxygen
 will occur in oceans most spawning fish congregate at such
 points so a loss would be catastrophic based is it still
 us if they are going to die in 10 20 years and not experience
 the consequences sorry but boomers have all the money they
 control most of government they are the group that perpetuated
 climate denial and then believed the denial and

climatechange: i am not sure what it is you just googled but
 gravimetry via satellite accurately measures the mass of
 the planets water you asked for proof and measurement of
 rising seas gravimetric studies have made measurements for
 decades and lo the seas have risen i cannot stop you from
 telling me whatever odd things that you believe if you have
 your heart set on it but if they do not square with observable
 evidence i do not have to take you seriously and so i do
 not climate change is not a belief it is a scientific fact
 at this

immigration: luckily the future vp is one of the major supporters of the bill and fyi u s386andbiden she is half indian this bill may never pass but no matter how much you hate indians do not try to justify it is evident from your comment history one of ours will be in the white house will toast to that nope i was always confused since my parents just said i was canadian growing up guess i should look into getting a us passport spousal open work permits to canada first off where are you in the process did you

conservatives: nothing to see here these idiots think you can prevent a virus from spreading it is a virus it is going to spread and during a pandemic ask Biden about racial jungles i just saw the New York Times article on the freeway parades allegedly shutting down bridges despite the police stating that traffic was never stopped where they called for the immediate arrest of all the drivers involved because it is dangerous and no one has the right to break the law tell that to all the rioters the past several months thanks for that how do you think

democrats: i see people on section 8 that are seriously gaming the system i know of a certain someone who was getting a big ass house for like 280 mo and her and her husband s combined incomes we just a tad over 230k meanwhile an average home in my neighborhood is around 4k i know of more as well but nobody audits this shit and now in my cousin s neighborhood there is multiple homes under section 8 and in this neighborhood all the homes are brand new practically and then you see the section 8 homes all ran the

Liberal: i am just surprised it was with a girl tbh countryaboveparty
 it is really not lol it is an important measure but not the
 only measure most are not going to war but the ones that
 are got entirely too close to their goal it is too late for
 want it is already divided the red side chose that division
 picking orange cheeto if one side has divided it and does
 nothing for unifying the country then we have no choice but
 to accept it is divided the federal communications commission
 fcc has listed infotmation regarding these issues on their

math: oh okay i will also look into this thanks a professor
 of mine extended that to three people used excerpts from
 the jules et jim novel to constrain and later validate the
 model and showed that the resulting solution is chaotic they
 also find out that the chaos arises due to the feelings that
 the two men have for each other here is the paper eli5 what
 is haar measure suggestions for linear algebra problems i
 have only heard numerator and denominator rectangular is
 twice as long as it is wide if an are is 144 5 square inches

References

- 8kun (2021). 8kun. <https://web.archive.org/web/20210101064319/https://8kun.top/index.html>. Accessed through the Internet Archive's Wayback Machine on: November 19, 2022.
- 8kun (n.d.a). 8kun. <https://web.archive.org/web/20210302070106/https://8kun.top/index.html>. Accessed through the Internet Archive's Wayback Machine on: July 14, 2022.
- 8kun (n.d.b). Active ISPs. <https://web.archive.org/web/20210205190518/https://8kun.top/activeusers.html>. Accessed through the Internet Archive's Wayback Machine on: July 14, 2022.
- Aggarwal, C. C. and Zhai, C. (2012). A Survey of Text Clustering Algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 77–128, New York, NY. Springer. <https://doi.org/10.1007/978-1-4614-3223-4>.
- Anti-Defamation League (n.d.a). ADL's Mission and History. <https://www.adl.org/about/mission-and-history>, Accessed July 4, 2022.
- Anti-Defamation League (n.d.b). Echo. <https://www.adl.org/resources/hate-symbol/echo>, Accessed: July 4, 2022.
- Apache Software Foundation (2021). Feather File Format. <https://arrow.apache.org/docs/python/feather.html>, Accessed: September 8, 2021.
- archive.today (n.d.). Webpage Capture. <https://archive.ph/>.
- Auxier, B. and Anderson, M. (2021). Social Media Use in 2021. *Pew Research Center*. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>.
- Baayen, R. H. and Shafaei-Bajestan, E. (2019). *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics*. R package version 1.5.0. <https://CRAN.R-project.org/package=languageR>.

- Baele, S. J., Brace, L., and Coan, T. G. (2020). The ‘Tarrant Effect’: What Impact Did Far-Right Attacks Have on the 8chan Forum? *Behavioral Sciences of Terrorism and Political Aggression*, pages 1–23. <https://doi.org/10.1080/19434472.2020.1862274>.
- Baele, S. J., Brace, L., and Coan, T. G. (2021). Variations on a Theme? Comparing 4chan, 8kun, and Other chans’ Far-right “/pol” Boards. *Perspectives on Terrorism*, 15 (1):65–80. <https://www.universiteitleiden.nl/perspectives-on-terrorism/archives/2021#volume-xv-issue-1>.
- Bartlet, J. (2020). Whitespace Characters to Copy and Paste. <https://qwerty.dev/whitespace/>, Accessed: October 7, 2022.
- Baumgartner, J., Lazzarin, E., and Seiler, A. (2019). Pushshift Reddit API Documentation. *Github Repository*. <https://github.com/pushshift/api>.
- Baumgartner, J., Lazzarin, E., and Seiler, A. (2022). Pushshift Dataset. <https://files.pushshift.io/reddit/>, Accessed: July 14, 2022.
- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020a). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839. <https://ojs.aaai.org/index.php/ICWSM/article/view/7347>.
- Baumgartner, J., Zannettou, S., Squire, M., and Blackburn, J. (2020b). The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:840–847. <https://ojs.aaai.org/index.php/ICWSM/article/view/7348>.
- Benkler, Y., Faris, R., and Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, New York, NY.
- Bennett, J. (2022). How War Came Home: From 9/11 to the Storming of the U.S. Capitol. *Reviews in Anthropology*, 51 (3-4):47–67. <https://doi.org/10.1080/00938157.2022.2058662>.
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1987). Occam’s Razor. *Information Processing Letters*, 24 (6):377–380. [https://doi.org/10.1016/0020-0190\(87\)](https://doi.org/10.1016/0020-0190(87))

90114-1.

- Boe, B. (2022). *PRAW: The Python Reddit API Wrapper*. Python package version 7.5.0. <https://praw.readthedocs.io/en/stable/index.html>.
- Bortolon, C. (2022). *In the Eye of the Storm: A Discourse Analysis of Disproval and the Internet's Effect on QAnon*. Major Papers, University of Windsor, Windsor Ontario. <https://scholar.uwindsor.ca/major-papers/213>.
- Bowyer, S. (2021). The Wayback Machine: Notes on a Re-Enchantment. *Archival Science*, 21:43–57. <https://doi.org/10.1007/s10502-020-09345-w>.
- Brath, R. (2018). *Text in Visualization: Extending the Visualization Design Space*. PhD thesis, London South Bank University, London, UK. <https://doi.org/10.18744/PUB.002743>.
- Burch, M., Lohmann, S., Beck, F., Rodriguez, N., Silvestro, L. D., and Weiskopf, D. (2014). RadCloud: Visualizing Multiple Texts with Merged Word Clouds. In *2014 18th International Conference on Information Visualisation*, pages 108–113, Paris, France. IEEE. <https://doi.org/10.1109/IV.2014.72>.
- Butler, U. (2022). Islamophobia in the Digital Age: A Study of Anti-Muslim Tweets. (August 10, 2022). Available at SSRN: <https://ssrn.com/abstract=4227488> or <http://dx.doi.org/10.2139/ssrn.4227488>.
- Carlson, J. and Harris, K. (2022). The Apportionment of Citations: A Scientometric Analysis of Lewontin 1972. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377 (1852):20200409. <https://doi.org/10.1098/rstb.2020.0409>.
- Carlsson, G. and Mémoli, F. (2010). Characterization, Stability and Convergence of Hierarchical Clustering Methods. *Journal of Machine Learning Research*, 11:1425–1470. <http://jmlr.org/papers/v11/carlsson10a.html>.
- Chang, C. and Freeman, J. (2021). Supreme Court Confirmation of Amy Coney Barrett: A “Blatant Act of Bad Faith”? *Documents to the People*, 49 (3-4):30–36. <https://heinonline.org/HOL/P?h=hein.journals/dttp49&i=82>.

- Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). Determining the Number of Clusters using NbClust Package. In *Proceedings of the Meeting on Statistics and Data Mining*, pages 1–7, Djerba, Tunisia. MSDM. https://www.researchgate.net/profile/Mohamed-Limam/publication/323600098.Proceedings_of_MSDM_2014/links/5a9f7cd9aca272d448adb7cb/Proceedings-of-MSDM-2014.pdf#page=6.
- Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2022). *NbClust: Determining the Best Number of Clusters in a Data Set*. R package version 3.0.1. <https://CRAN.R-project.org/package=NbClust>.
- Cleveland, W. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.
- Cosentino, G. (2020). From Pizzagate to the Great Replacement: The Globalization of Conspiracy Theories. In *Social Media and the Post-Truth World Order: The Global Dynamics of Disinformation*, pages 59–86, Cham, Switzerland. Springer International Publishing. https://doi.org/10.1007/978-3-030-43005-4_3.
- CPOST (2022). Understanding Political Violence in America. <https://cpost.uchicago.edu/research/apv/>.
- Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M., and Qu, H. (2010). Context Preserving Dynamic Word Cloud Visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128, Taipei, Taiwan. IEEE. <https://doi.org/10.1109/PACIFICVIS.2010.5429600>.
- De Queiroz, G., Fay, C., Hvitfeldt, E., Keyes, O., Misra, K., Mastny, T., Erickson, J., Robinson, D., and Silge, J. (2022). *tidytext: Text Mining using ‘dplyr’, ‘ggplot2’, and Other Tidy Tools*. R package version 0.3.3. <https://CRAN.R-project.org/package=tidytext>.
- Desgraupes, B. (2018). *clusterCrit: Clustering Indices*. R package version 1.2.8. <https://CRAN.R-project.org/package=clusterCrit>.
- Dinulescu, I. (2021). Religion and Politics in the Context of the 6 January 2021 Assault on the US Congress. *Strategic Impact*, 79 (2):78–92. <https://doi.org/10.53477/1841-5784-21-05>.

- Dowle, M. and Srinivasan, A. (2021). *data.table: Extension of 'data.frame'*. R package version 1.14.2. <https://CRAN.R-project.org/package=data.table>.
- dplyr (n.d.). dplyr. <https://dplyr.tidyverse.org/>, Accessed: July 25, 2022.
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12 (7):2121–2159. <https://www.jmlr.org/papers/volume12/duchilla/duchilla.pdf>.
- Duckett, J. (2011). *HTML and CSS: Design and Build Websites*. John Wiley & Sons, Indianapolis, IN.
- El-Hamdouchi, A. and Willett, P. (1989). Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *The Computer Journal*, 32:220–227. <https://doi.org/10.1093/comjnl/32.3.220>.
- Estivill-Castro, V. (2002). Why So Many Clustering Algorithms: A Position Paper. *ACM SIGKDD Explorations Newsletter*, 4.1:65–75. <https://doi.org/10.1145/568574.568575>.
- Feinerer, I. and Hornik, K. (2020). *tm: Text Mining Package*. R package version 0.7-8. <https://CRAN.R-project.org/package=tm>.
- Formica, T. (2020). A Social (Media) Contract: Reconciling American Freedom and Security in an Age of Online Radicalization and Extremism. *Yale Journal of International Affairs*, 15:131–147. <https://www.yalejournal.org/publications/a-social-media-contract-reconciling-american-freedom-and-security-in-an-age-of-online-radicalization-and-extremism?rq=media%20contract>.
- Galili, T., Benjamini, Y., and Jefferis, G. (2022). *dendextend: Extending 'dendrogram' Functionality in R*. R package version 1.16.0. <https://CRAN.R-project.org/package=dendextend>.
- Gentleman, R. and Temple Lang, D. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16 (1):1–23. <https://doi.org/10.1198/106186007X178663>.
- Goldberg, Y. and Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.’s Negative-Sampling Word-Embedding Method. *arXiv:1402.3722*. <https://doi.org/10>

.48550/arXiv.1402.3722.

- González, R. J. (2015). Seeing Into Hearts and Minds: Part 2. ‘Big Data’, Algorithms, and Computational Counterinsurgency. *Anthropology Today*, 31 (4). <https://doi.org/10.1111/1467-8322.12188>.
- Hannah, M. (2021). QAnon and the Information Dark Age. *First Monday*, 26 (2). <https://dx.doi.org/10.5210/fm.v26i2.10868>.
- Hitkul, Prabhu, A., Guhathakurta, D., Jain, J., Subramanian, M., Reddy, M., Sehgal, S., Karandikar, T., Gulati, A., Arora, U., Shah, R. R., and Kumaraguru, P. (2021). Capitol (Pat) Riots: A Comparative Study of Twitter and Parler. *arXiv:2101.06914*. <https://arxiv.org/abs/2101.06914>.
- Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D., and Watts, D. (2021). Examining the Consumption of Radical Content on YouTube. *Proceedings of the National Academy of Sciences of the United States of America*, 118 (32):e2101967118. <https://doi.org/10.1073/pnas.2101967118>.
- Huang, A. (2008). Similarity Measures for Text Document Clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, volume 4, pages 9–56, Christchurch, New Zealand. NZCSRSC2008. https://www.academia.edu/6003456/SimilarityMeasures_for_Text_Document_Clustering.
- Hussain, M. N., Tokedmir, S., Agarwal, N., and Al-Khateeb, S. (2018). Analyzing Disinformation and Crowd Manipulation Tactics on YouTube. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 1092–1095, Barcelona, Spain. IEEE. <https://doi.org/10.1109/ASONAM.2018.8508766>.
- Internet Archive (2022a). 8kun.top Calendar. https://web.archive.org/web/20210701000000*/8kun.top, Accessed: July 14, 2022.
- Internet Archive (2022b). reddit.com Calendar. https://web.archive.org/web/20210901000000*/reddit.com, Accessed: July 14, 2022.
- Internet Archive (n.d.a). About the Internet Archive. <https://archive.org/about/>, Accessed: September 8, 2021.

- Internet Archive (n.d.b). Using the Wayback Machine. <https://help.archive.org/help/using-the-wayback-machine/>, Accessed: July 14, 2022.
- Internet Archive (n.d.c). Wayback Mahine General Information. <https://help.archive.org/help/wayback-machine-general-information/>, Accessed: July 14, 2022.
- Kean, T. H. and Hamilton, L. H. (2004). *The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States (9/11 Report)*. U.S. Government Printing Office, Washington DC. <https://www.govinfo.gov/app/details/GPO-911REPORT>.
- Kessler, J. (2017). Scattertext: A Broser-Based Tool for Visualizing How Corpora Differ. In *Proceedings of ACL 2017, System Demonstrations*, pages 85–90, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.1703.00565>.
- Khalil, S. and Fakir, M. (2017). RCrawler: An R Package for Parallel Web Crawling and Scraping. *SoftwareX*, 6:98–106. <https://doi.org/10.1016/j.softx.2017.04.004>.
- Kleinberg, J. (2002). An Impossibility Theorem for Clustering. In *Advances in Neural Information Processing Systems*, volume 15, pages 446–453, Boston, MA. MIT Press. <https://papers.nips.cc/paper/2002/file/43e4e6a6f341e00671e123714de019a8-Paper.pdf>.
- Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 957–966, Lille, France. Proceedings of Machine Learning Research. <http://proceedings.mlr.press/v37/kusnerb15.html>.
- Landry, A. P., Ihm, E., Kwit, S., and Schooler, J. W. (2021). Metadehumanization Erodes Democratic Norms During the 2020 Presidential Election. *Analyses of Social Issues and Public Policy*, 21 (1):51–63. <https://doi.org/10.1111/asap.12253>.
- Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of Machine Learning Research*, pages 1188–1196, Beijing, China. <https://>

proceedings.mlr.press/v32/le14.html.

- Ledwich, M. and Zaitsev, A. (2020). Algorithmic Extremism: Examining YouTube’s Rabbit Hole of Radicalization. *First Monday*, 25:(3). <https://doi.org/10.5210/fm.v25i3.10419>.
- Li, D. and Zhou, X. (2016). “Leave Your Footprints in My Words”-A Georeferenced Word-Cloud Approach. *Environment and Planning A: Economy and Space*, 49 (3):489–492. <https://doi.org/10.1177/0308518X16662273>.
- Lohmann, S., Heimerl, F., Bopp, F., Burch, M., and Ertl, T. (2015). Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. In *2015 19th International Conference on Information Visualisation*, pages 114–120, Barcelona, Spain. IEEE. <https://doi.org/10.1109/iv.2015.30>.
- lxml Dev Team (2022). *lxml - XML and HTML with Python*. Python package version 4.9.1. <https://lxml.de/>.
- Maass, H. (2020). (October 14). 10 Things You Need to Know Today: October 24, 2020. *The Week*. <https://theweek.com/10things/943617/10-things-need-know-today-october-14-2020>.
- MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Berkeley Symposium on Mathematical Statistics and Probability*, 5:281–297. <https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and-chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992>.
- Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M. (2022). *cluster: “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al.* R package version 2.1.4. <https://CRAN.R-project.org/package=cluster>.
- Manning, C., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

- Martin, A. E. and Fournillier, J. B. (2022). It's Me(me), Revolution Elizabeth: Social Media and a Practice of Critical Social Commentary. *Cultural Studies ↔ Critical Methodologies*, 22 (5):466–476. <https://doi.org/10.1177/15327086221097631>.
- Marx, D. (2018). *PSAW*. Python package version 0.1.0. <https://psaw.readthedocs.io/en/latest/>.
- McQuitty, L. (1957). Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educational and Psychological Measurement*, 17:207–222. <https://doi.org/10.1177/001316445701700204>.
- Michalke, M. (2021). *koRpus: Text Analysis with Emphasis on POS Tagging, Readability, and Lexical Diversity*. R package version 0.13-8. <https://CRAN.R-project.org/package=koRpus>.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546*. <https://doi.org/10.48550/arXiv.1310.4546>.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*. <https://doi.org/10.48550/arXiv.1301.3781>.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2021). Computing Numeric Representations of Words in a High-Dimensional Space (U.S. Patent No. 10,922,488). <https://patents.google.com/patent/US10922488B1/en>.
- Ng, L. H. X., Cruickshank, I., and Carley, K. (2021). Coordinating Narratives and the Capitol Riots on Parler. *arXiv:2109.00945*. <https://arxiv.org/abs/2109.00945>.
- Norris, N. (1997). Error, Bias and Validity in Qualitative Research. *Educational Action Research*, 5 (1):172–176. <https://doi.org/10.1080/09650799700200020>.
- Obaidi, M., Kunst, J., Ozer, S., and Kimel, S. Y. (2022). The “Great Replacement” Conspiracy: How the Perceived Ousting of Whites Can Evoke Violent Extremism and Islamophobia. *Group Processes & Intergroup Relations*, 25 (7):1675–1695. <https://doi.org/10.1177/13684302211028293>.

- O'Donnell, T. (2020). (October 25). 10 Things You Need to Know Today: October 24, 2020. *The Week*. <https://theweek.com/10things/945802/10-things-need-know-today-october-24-2020>.
- Pape, R. (2022a). American Face of Insurrection: Analysis of Individuals Charged for Storming the US Capitol on January 6, 2021. *Chicago Project on Security and Threats*, University of Chicago, Chicago, IL. <https://cpost.uchicago.edu/publications/american-face-of-insurrection/>.
- Pape, R. (2022b). Deep, Divisive, Disturbing and Continuing: New Survey Shows Maintream Support for Violence To Restore Trump Remains Strong. *Chicago Project on Security and Threats*, University of Chicago, Chicago, IL. <https://cpost.uchicago.edu/publications/deep-divisive-disturbing-and-continuing-new-survey-shows-maintream-support-for-violence-to-restore-trump-remains-strong/>.
- Pennycook, G. and Rand, D. G. (2021). Examining False Beliefs about Voter Fraud in the Wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review*, 2 (1). <https://doi.org/10.37016/mr-2020-51>.
- Perma.cc (n.d.). About Perma.cc. <https://perma.cc/about>.
- Pion-Berlin, D., Bruneau, T., and Jr., R. B. G. (2022). The Trump Self-Coup Attempt: Comparisons and Civil–Military Relations. *Government and Opposition: An International Journal of Comparative Politics*, pages 1–18. <https://doi.org/10.1017/gov.2022.13>.
- Podolak, M. (2022). *pmaw*. Python package version 3.0.0. <https://pypi.org/project/pmaw/>.
- Price, D. H. (2011). How the CIA and Pentagon Harnessed Anthropological Research During the Second World War and Cold War With Little Critical Notice. *Journal of Anthropological Research*, 67 (3):333–356. <https://doi.org/10.3998/jar.0521004.0067.302>.
- Proferes, N., Jones, N., Gilbert, S., Fiesler, C., and Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7 (2). <https://doi.org/10.1177/20563051211019004>.

- Queirós, A., Faria, D., and Almeida, F. (2017). Strengths and Limitations of Qualitative and Quantitative Research Methods. *European Journal of Education Studies*, 3 (9):369–387. <https://doi.org/10.5281/zenodo.887089>.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- R Core Team and Contributors Worldwide (2022a). *The R Stats Package*. R package version 4.3.0. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html>.
- R Core Team and Contributors Worldwide (2022b). *The R Utils Package*. R package version 3.6.2. <https://rdocumentation.org/packages/utils/versions/3.6.2>.
- Ravn, S., Barnwell, A., and Neves, B. B. (2020). What Is “Publicly Available Data”? Exploring Blurred Public–Private Boundaries and Ethical Practices Through a Case Study on Instagram. *Journal of Empirical Research on Human Research Ethics*, 15 (1-2):40–45. <https://doi.org/10.1177/1556264619850736>.
- Reddit (2021). Reddit by the Numbers. <https://www.redditinc.com/press>, Accessed: November 19, 2022.
- Reddit (2022a). Eyebleshoot. <https://www.reddit.com/r/Eyebleshoot/>, Accessed: November 10, 2022.
- Reddit (2022b). Reddit API Documentation. <https://www.reddit.com/dev/api/>, Accessed: August 3, 2022.
- Reid, J. C. and Craig, M. O. (2021). Is it a Rally or a Riot? Racialized Media Framing of 2020 Protests in the United States. *Journal of Ethnicity in Criminal Justice*, 19 (3-4):291–310. <https://doi.org/10.1080/15377938.2021.1973639>.
- Richardson, L. (2022). *Beautiful Soup Documentation*. Python package version 4.8.1. <https://beautiful-soup-4.readthedocs.io/en/latest/>.
- Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. *Journal of Classification*, 35:345–366. <https://doi.org/10.1007/s00357-018-9259-9>.

- Salzmann, Z., Stanlaw, J. M., and Adachi, N. (2012). *Language, Culture, and Society: An Introduction to Linguistic Anthropology*. Westview Press, Boulder, CO.
- Schwartz, A., Eichstaedt, J., Blanco, E., Dziurzynski, L., Kern, M., Ramones, S., Seligman, M., and Ungar, L. (2013). Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach. In *Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 296–305, Atlanta, GA. Second Joint Conference on Lexical and Computational Semantics, Volume 1.
- Scrapy (n.d.). Meet the Scrapy Pros. <https://scrapy.org/companies/>, Accessed October 25, 2022.
- Scrapy Developers (2022a). *Scrapy Documentation*. Python package version 2.6.1. <https://docs.scrapy.org/en/latest/pdf/>.
- Scrapy Developers (2022b). Scrapy Tutorial. <https://docs.scrapy.org/en/latest/intro/tutorial.html>, Accessed August 1, 2022.
- Setty, S. (2021). The January 6, 2021, Capitol Riots: Resisting Calls for More Terrorism Laws. *Journal of National Security Law & Policy*, Special Online Issue: Capitol Insurrection 2021. <https://jnslp.com/2021/01/25/the-january-6-2021-capitol-riots-resisting-calls-for-more-terrorism-laws/>.
- Singh, A. K. and Shashi, M. (2019). Vectorization of Text Documents for Identifying Unifiable News Articles. *International Journal of Advanced Computer Science and Applications*, 10:7. <https://doi.org/10.14569/IJACSA.2019.0100742>.
- Sneath, P. (1957). The Application of Computers to Taxonomy. *Microbiology*, 17:201–226. <https://doi.org/10.1099/00221287-17-1-201>.
- Sokal, R. R. and Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Science Bulletin*, 38, pt. 2:1409–1438. https://ia800703.us.archive.org/5/items/cbarchive_33927_astatisticalmethodforevaluatin1902/astatisticalmethodforevaluatin1902.pdf.
- stringr (n.d.). stringr. <https://stringr.tidyverse.org/>, Accessed: July 25, 2022.

- Temple Lang, D. (2022). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.98-1.8. <https://CRAN.R-project.org/package=RCurl>.
- Tidyverse (n.d.). R Packages for Data Science. <https://www.tidyverse.org/>, Accessed: September 19, 2022.
- Townsend, L. and Wallace, C. (2018). The Ethics of Using Social Media Data in Research: A New Framework. In Woodfield, K., editor, *The Ethics of Online Research*, chapter 8, pages 189–207. Emerald Publishing Limited, Bradford, UK. <https://doi.org/10.1108/S2398-601820180000002008>.
- Tran, T. and Ostendorf, M. (2016). Characterizing the Language of Online Communities and its Relation to Community Reception. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.1609.04779>.
- Tsou, A. (2016). How Does the Front Page of the Internet Behave? Readability, Emoticon, and Links on Reddit. *First Monday*, 21 (11). <http://dx.doi.org/10.5210/fm.v21i11.7013>.
- United States Department of Justice (2021). One Year Since the Jan. 6 Attack on the Capitol. <https://www.justice.gov/usao-dc/one-year-jan-6-attack-capitol>, Accessed: November 3, 2022.
- US Senate (2021). Examining the US Capitol Attack: A Review of the Security, Planning, and Response Failures on January 6. Committee on Homeland Security and Governmental Affairs, Committee on Rules and Administration. <https://www.rules.senate.gov/imo/media/doc/Jan%206%20HSGAC%20Rules%20Report.pdf>.
- Van Dijcke, D. and Wright, A. (2022). Profiling Insurrection: Characterizing Collective Action Using Mobile Device Data. *Becker Friedman Institute for Economics at the University of Chicago*. <https://bfi.uchicago.edu/working-paper/profiling-insurrection-characterizing-collective-action-using-mobile-device-data/>.
- van Rossum, G. and the Python Core Development Team (2020). *Python 3.7.12 documentation*. Scotts Valley, California. <https://docs.python.org/3.7/>.

- Wang, E., Cook, D., and Hyndman, R. J. (2020a). Calendar-Based Graphics for Visualizing People's Daily Schedules. *Journal of Computational and Graphical Statistics*, 29 (3):490–502. <https://doi.org/10.1080/10618600.2020.1715226>.
- Wang, S., Tang, J., Aggarwal, C., and Liu, H. (2016). Linked Document Embedding for Classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 115–124, Indianapolis, IN. CIKM. <https://doi.org/10.1145/2983323.2983755>.
- Wang, Z., Zhou, D., Yang, M., Zhang, Y., Rao, C., and Wu, H. (2020b). Robust Document Distance with Wasserstein-Fisher-Rao Metric. In *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129, pages 721–736, Bangkok, Thailand. Proceedings of Machine Learning Research.
- Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Educational and Psychological Measurement*, 58:236–244. <https://doi.org/10.2307/2282967>.
- Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>.
- Wickham, H. (2022). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.3. <https://CRAN.R-project.org/package=rvest>.
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., and Dunnington, D. (2022a). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.6. <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, H., François, R., Henry, L., and Müller, K. (2022b). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, H., RStudio, Feather developers (Bundled feather library), Google (Bundled flatbuffers code), and LevelDB Authors (2019). *feather: R Bindings to the Feather 'API'*. R package version 0.3.5. <https://CRAN.R-project.org/package=feather>.
- Wijffels, J. (2021). *word2vec: Distributed Representations of Words*. R package version 0.3.4. <https://CRAN.R-project.org/package=word2vec>.

- Wilke, C. (2020). *ggtext: Improved Text Rendering Support for 'ggplot2'*. R package version 0.1.1. <https://CRAN.R-project.org/package=ggtext>.
- Wilkinson, L. and Friendly, M. (2009). The History of the Cluster Heat Map. *The American Statistician*, 63 (2):179–184. <https://doi.org/10.1198/tas.2009.0033>.
- Wodak, R. (2011). Critical Linguistics and Critical Discourse Analysis. In Zienkowski, J., Verschueren, J., and Ola Östman, J., editors, *Discursive Pragmatics*, pages 50–70, Philadelphia, PA. John Benjamins Publishing Company.
- Wu, L., Yen, I. E.-H., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., and Witbrock, M. J. (2018). Word Mover’s Embedding: From word2vec to Document Embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4524–4534, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1482>.
- Yang, Y., Akers, L., Klose, T., and Yang, C. (2008). Text Mining and Visualization Tools – Impressions of Emerging Capabilities. *World Patent Information*, 30 (4):280–293. <https://doi.org/10.1016/j.wpi.2008.01.007>.
- Yardi, S., Romero, D., Schoenebeck, G., , and danah boyd (2009). Detecting Spam in a Twitter Network. *First Monday*, 15:1–4. <https://firstmonday.org/ojs/index.php/fm/article/download/2793/2431>.
- Zamani, M., Rabbani, F., Horicsányi, A., Zafeiris, A., and Vicsek, T. (2019). Differences in Structure and Dynamics of Networks Retrieved from Dark and Public Web Forums. *Physica A: Statistical Mechanics and its Applications*, 525:326–336. <https://doi.org/10.1016/j.physa.2019.03.048>.
- Zeng, J. and Schäfer, M. S. (2021). Conceptualizing “Dark Platforms”. Covid-19-Related Conspiracy Theories on 8kun and Gab. *Digital Journalism*, 9 (9):1321–1343. <https://doi.org/10.1080/21670811.2021.1938165>.
- Zobel, J. and Moffat, A. (1998). Exploring the Similarity Space. *ACM SIGIR Forum*, 32 (1):18–34. <https://doi.org/10.1145/281250.281256>.