Title

Discourse Networks: Dynamic network modeling of the Brexit negotiations

Author

Kappos, Cybele

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Discourse Networks:

Dynamic network modeling of the Brexit negotiations

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Statistics

by

Cybele Kappos

2024

ABSTRACT OF THE THESIS

Discourse Networks:

Dynamic network modeling of the Brexit negotiations

by

Cybele Kappos

Master of Science in Statistics

University of California, Los Angeles, 2024

Professor Chad J. Hazlett, Chair

Political discourse is constantly in flux: the key issues, the actors that define the discourse are changing from moment to moment. Quantitative approaches to discourse are an important methodological tool that can help researchers measure the structure of and changes in discourse. However, existing approaches are often time-consuming and not scalable to large datasets. In this thesis, I apply a fully automated approach to discourse analysis that combines Structural Topic Modeling and Network Analysis. I apply the model to a novel dataset of the Brexit negotiations in British parliament. I find that the method captures the state of relations between actors and the progress in the negotiation process.

This thesis of Cybele Kappos is approved.

Frederic Paik Schoenberg

Mark Stephen Handcock

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2024

Contents

List of Figures

List of Tables

1 Introduction

Topic Modeling is a powerful tool in text analysis that uncovers latent patterns of words or phrases (topics) in a set of documents. Structural Topic Modeling (STM) is a topic modeling framework that allows the researcher to include covariate information in the detection of topics. STM is a flexible tool that can be used in various contexts such as studying open-ended survey responses (Roberts et al., 2014), understanding problems that leaders face, (Tonidandel et al., 2022) and better understanding how topics and coalitions change during historical events (Ahmed et al., 2020). STM allows researchers to discover how documents talk about the same topics, organize a set of texts, and study the prevalence and content of different topics.

The topics detected by the STM model are not unlike the patterns of discourse. Michel Foucault defines discourse "the totality of all effective statements", which together form a "conceptual network on the basis of the intrinsic regularities of discourse" (Foucault, 1972). A quantitative approach to examining discourse would be an important development in an era where discourse – discourse on foreign policy, reproductive rights, political campaigns – are constantly in flux. Studying discourse would entail being able to study the relations between topics and actors as well as the changes in these relations over time.

Bridging STM and Network Analysis can help us achieve this task. These two approaches may seem incompatible because Network Analysis is methodologically distinct from topic modeling and is not a tool in the toolkit of Natural Language Processing. In the social sciences, Social Network Analysis quantitatively measures the structure of groups or individuals. This method is often used to understand the position and importance of different actors, as well as the connections between them. It is also used to investigate how actors group together into communities or clusters.

A closer study of the similarities of the two methods shows that both STM and network analysis have a structural focus, with the objective of understanding how units are organized.

Additionally, as STM can be used to study how topics are shared among actors, network analysis studies the relations between individuals. The first method that was developed to combine STM and network analysis is called Discourse Network Analysis (DNA). (Leifeld, 2013) While this methodological approach was the first of its kind and powerful, the method is not easily applicable to projects using larger corpora of texts because it requires the researcher to conduct qualitative content analysis. How then, can we make DNA more scalable to cases where the dataset consists of thousands of documents?

In this thesis, I show how discourse network analysis can be automated. I use STM to detect topics in documents, replacing the step of qualitative content analysis of the DNA method. I then use network analysis to transform the output of the STM model into a network. Using these networks, I can then capture network-level changes in structure and the flow of information, and detect communities of speakers. I do a comparative analysis of these networks over time.

I apply the method to the context of the Brexit negotiations in the British House of Commons. Using a novel dataset of speeches from members of parliament (MPs), I am able to capture the most important topics of the negotiation period. I also measure how communities of MPs are formed and change over the course of the time period.

The thesis proceeds as follows. I review the fundamentals of STM and DNA. I then describe the historical context of the data and the data itself. I provide a technical explanation of my methodological approach. I then describe the results, which are followed by a comparative analysis of the networks in the discussion.

2 STM

There are several methods to estimate a topic model. A popular approach is the Latent Dirichlet Allocation (LDA), which uses a Bayesian probabilistic model. Structural Topic Modeling (STM) is an extension of LDA modeling but allows for more information in the production of the model as the user can incorporate covariate information. The covariates should be theoretically informed and associated with the prevalence of topics in a given document. In this paper, I use the metadata of the date of the document because I believe the date of the discussion is likely to be associated with the importance of a topic in a given time period.

Under STM, the generative process for each document d with k topics can be captured with the following notation.

1. Each document consists of a mix of the k topics, drawn from a logistic-normal generalized linear model with mean μ based on the document's covariates X_d. The covariates are specified by the user.

$$\overrightarrow{\theta}_d | X_d\gamma, \Sigma \sim LogisticNormal(\mu = X_d\gamma, \Sigma)$$

Where X_d is a 1-by-p vector, γ is a matrix of the document's coefficients and Σ is a covariance matrix.

2. Given the document-level attention to topics, the document-specific distribution over words representing each topic (k) using the baseline word distribution (m)

$$\beta_{d,k} \propto exp(m + \kappa_k^{(t)} + \kappa_{yd}^{(c)} + \kappa_{yd,k}^{(i)})$$

This captures the topic-specific deviation from m ($\kappa_k^{(t)}$), the covariate-group deviation ($\kappa_{yd}^{(c)}$) and the interaction between the two ($\kappa_{yd,k}^{(i)}$))

3. Finally, for each word in the document $n \in 1, \ldots, N_d$

- The topic assignment of each word is drawn based on the document's specific distribution over topics

$$x_{d,n} | \overrightarrow{\theta}_d \sim Multinomial(\overrightarrow{\theta}_d)$$

- Conditional on a given topic, draw an observed word from that topic

$$w_{d,n}|z_{d,n}, \beta_{d,k=z_{d,n}} \sim Multinomial(\beta_{d,k=z_{d_n}})$$

3 DNA

Discourse Network Analysis, or DNA, is a method developed by Philip Leifeld (2017). It is a combination of content analysis and dynamic network analysis. First, the researcher qualitatively reads through documents, annotating (a) the statements that are present in each and (b) the valence of the document, i.e. whether the author is in support of or against the statement. Once each document has been annotated, the data can be converted into a matrix form –documents are rows and every unique concept present in the data is a column– where each cell denotes whether the statement was mentioned in the document and, if it was mentioned, whether the valence was positive or negative. The matrix can then be converted into a network object, where ties between actors are formed when the actors are either both in support of or in opposition to the issue. In bimodal networks, we can examine the relation between actors, between actors and concepts, and between concepts.

This approach was developed in the context of the Advocacy Coalition Framework, which conceptualizes policymaking as the outcome of competing coalitions of policy actors who share policy beliefs and preferences.

While this is a powerful method that examines political texts through a network framework, it is a tedious process to manually code each document and is not feasible for large numbers of texts. Therefore, I demonstrate a modification of this method to expedite the process and extend its use to larger corpora of text.

4 Comparing Quantitative Approaches to the DNA method

Vaughan (2020) developed an automated approach to the annotation of concepts in documents. Instead of manually coding each document, the author uses LDA to detect the presence of a topic-concept in a given document. The author then uses the resulting discourse networks to examine the discursive similarity between two advocacy organizations. Later, this approach was extended by Harper, Kappos, and Neumeier (2023) to examine how focusing events disrupt discourse networks. The authors use STM, as opposed to Vaughan's LDA approach, to annotate documents. In both methods, valence is not calculated. This is significantly different to Leifeld's DNA method.

Using a dynamic network framework, they examine network changes over time. By selecting a sample of focusing events (sudden events that draw the attention of the public, the media, and policymakers) related to immigration, they examine the discourse networks 90 days before and 90 days after the focusing event. They measure changes in the discourse network that are associated with the event.

In this paper, I take a similar approach to Harper et al. The primary difference is that the context I use is much more specific. The authors examine immigration discourse in the British House of Commons over a 40-year period, whereas I examine only the Brexit negotiations period which began in 2016 and ended in 2020 (the Brexit withdrawal agreement was finalized in October 2019 but was put into effect in February 2020). This narrower timeframe generates more granular, delineated topics. Additionally, many actors in the discourse networks remain in the dataset across the years (the same MPs participate in the same committees) and there are fewer changes in MPs in the House of Commons since there was only one election in the time period. This makes the networks more directly comparable over time.

5 Context

The data used in this paper covers the period of Brexit negotiations, following the referendum in 2016, when UK voters voted to leave the European Union. The negotiations involved many important points of discussion including new legislation on immigration from inside the EU, the negotiation of trade deals, the implications of the end of funding from the EU in areas such as the arts and agriculture, and the devolvement of EU oversight in certain legal areas.

I use this case study in the rest of my research to study how to measure anxiety in speech and audio data.

6 Data

The dataset is an amalgamation of existing and novel data. Initially, the Hansard data seemed appropriate for this project. The Hansard dataset is a compilation of all British House of Commons debates dating from the 1988-89 parliamentary session to the present day. The dataset is often used in text analysis and scholarly work on discourse analysis (Archer, 2017, 2018; Foxlee, 2018; Kruger, van Rooy, and Smith, 2019; Mair, 2019; Slembrouck, 1992). This is a large, comprehensive dataset and therefore, it is useful for computationally-expensive text analysis methods. The data is relatively clean, meaning it is more or less consistent in format and only requires basic pre-processing to be able to analyze the data. The dataset has been criticized for limitations such as the accuracy of the transcript (Mollin, 2007). The format of the data is each row is a speech by a Member of Parliament (MP). Each speech includes metadata such as speech class (division, procedural, speech, table, or upper division), speaker name, a speaker identification number, constituency, the title of the debate where the speech occurred and the URL to the debate.

This data is interesting but is not a comprehensive look at the Brexit negotiations. It only includes the debates in the main chamber (the chamber where members of the government and opposition are seen sitting opposite each other). However, the work of the House of Commons extends to Special Committees. I scraped the data from oral evidence sessions of Select Committees to supplement the Hansard data. The resulting data is referred to as the Brexit dataset.

Select Committees are formed in the main chamber and the members of the committees consist of backbencher MPS. Some committees have interdepartmental roles and other examine specific topics, such as women's rights. In the case of Brexit, many select committees were formed specifically to address Brexit-related issues, e.g. the Committee on the Future Relationship with the European Union and the Exiting the European Union Committee.

Committees scrutinize and analyze government decision-making. They gather evidence from stakeholders and experts to form reports which are then submitted to the main chamber.

The resulting Brexit dataset consists of main chamber debates about Brexit and Select Committee oral evidence sessions about Brexit. The resulting dataset consists of $n = 47,116$ speeches.

The full methodological approach for starting with a corpus to performing DNA is detailed in Algorithm 1.

Algorithm 1: Quantitative Discourse Network Analysis

For a corpus of document D where each document is denoted as d_n where $n \in 1, \ldots N$, number of topics K and matrix θ where each cell p_{nk} is the proportion of topic k in document d_n where $k \in 1, \ldots, K$

1. Pre-process documents d_n

   (a) Shift all words to lowercase

   (b) Remove punctuation, special characters and numerical characters

(c) Remove common stopwords as well as procedural stopwords e.g. minister, hon, member

(d) Stem all tokens

2. Prep data for STM and choose K

3. Apply STM to data and extract matrix θ for all documents

4. Choose threshold $\lambda \in [0, 1]$ for cutoff value to identify most prominent topics in each document

   (a) All p_{nk} below threshold λ are converted to 0, all values above λ are converted to 1

5. Convert into network object

6. Get subgraphs depending on variables of interest (e.g. date)

7. Apply walktrap algorithm to network object to identify communities

---

The data is pre-processed by converting all lemmas to lowercase. Punctuation, special and numerical characters are removed. Common stopwords are removed from the corpus as well as a set of procedural words. All words are tokenized.

The data is then prepped for the stm package. Included in the metadata for the model is the speech's date. The number of topics is chosen according to a usual set of diagnostics (Held-Out Likelihood, Residuals, Semantic Coherence and Lower Bound).

For this analysis, I included only the date as a covariate in the STM model. I do not expect the party of the speaker to determine which topics they are most focused on. This is because Brexit was not an explicitly partisan issue, despite the fact that the event is associated with the Conservative party (Tolvanen, Tremewan, and Wagner, 2021). Euroskepticism existed in both major parties, albeit for different reasons. Politicians from all different political backgrounds joined both campaigns (the Leave campaign and the Remain campaign).

Rather, I hypothesize that only the date is associated with the prevalence of topics over time. I expect that at different time periods, different topics are discussed more heavily. This is dependent on the general context which includes the priorities of the House of Commons at the time and any relevant important political events, such as when a Prime Minister steps down.

One of the outputs of the *stm()* function is a matrix θ. As mentioned in section 2, θ is the documents' distribution over topics. The size of the matrix is N number of documents by K number of topics. Each cell p_{nk} is the proportion of topic k in document d_n. Each row, therefore, sums to 1. This matrix is converted to a dataframe. I append the speech's speaker, the data, the committee, the party of the speaker (if available), and the topic of discussion as columns. Then, to focus on the most important topics discussed in each speech, I determine a threshold λ as a cutoff point. In this case, I found $\lambda = 0.2$ to be an appropriate threshold. Using this threshold, any topic proportion p_{nk} above the threshold is converted to a binary 1 and converted to 0 if below the threshold. I experimented with other cutoff points but found this one to be accurate at capturing whether the speech clearly discussed topics that exceed $\lambda = 0.2$. The resulting dataframe is a matrix of 1s and 0s.

6.0.1 Walktrap Algorithm

A frequently-used tool in network analysis is community detection. Community detection is used to uncover how clusters of nodes, connected by edges, are grouped together. There are many algorithms that use different metrics, such as modularity or density of edges. In this paper, I use the walktrap algorithm to detect communities.

I chose the walktrap algorithm because research shows that it is well-suited to network analysis studying interpersonal relations, communication, and information exchange (Smith et al., 2020).

The walktrap algorithm uncovers communities based on random walks in the network. Random walks are used to determine the distance between nodes. Using hierarchical clustering, each node is then assigned a community based on small intra-community distance and large inter-community distance (Bitten, 2019). Compared to other community-detection algorithms, this algorithm assigns nodes to only belong to one community. A parameter determined by the user is the length of the random walk. In my analyses I choose $steps = 5$, meaning the random walk takes 5 steps.

Translating this to discourse network analysis results in communities that represent discursive proximity. Actors belonging to a certain community share inter-communal similarity (disussing many of the same topics), while also sharing discursive distance from another community that has its own set of central topics. Since networks in some years contain many nodes (i.e. many speakers), I also color the nodes by party affiliation. This is to help visually disentangle the properties of the network and its communities.

7 Results

7.1 Selecting K

The choice of K is an important decision for researchers. There is no rule of thumb for this number and it is highly dependent on the data. The STM package includes a function $searchK()$ where researchers can search a grid of values of K to determine the best fit for the data (Roberts, Stewart, and Tingley, 2019). I searched all values of K between 20 and 40 by increments of 2. The diagnostics (Held-Out Likelihood, Residuals, Semantic Coherence and Lower Bound) are shown in figure 1.

Similarly, plotting semantic coherence against exclusivity helps decide the best choice for K. This is captured in figure 2, once again confirming K = 30 is a good choice. Semantic coherence measures how often the most probable terms in a topic co-occur. This is highest when there are few topics. Exclusivity measures how unique top terms are to a given topic.

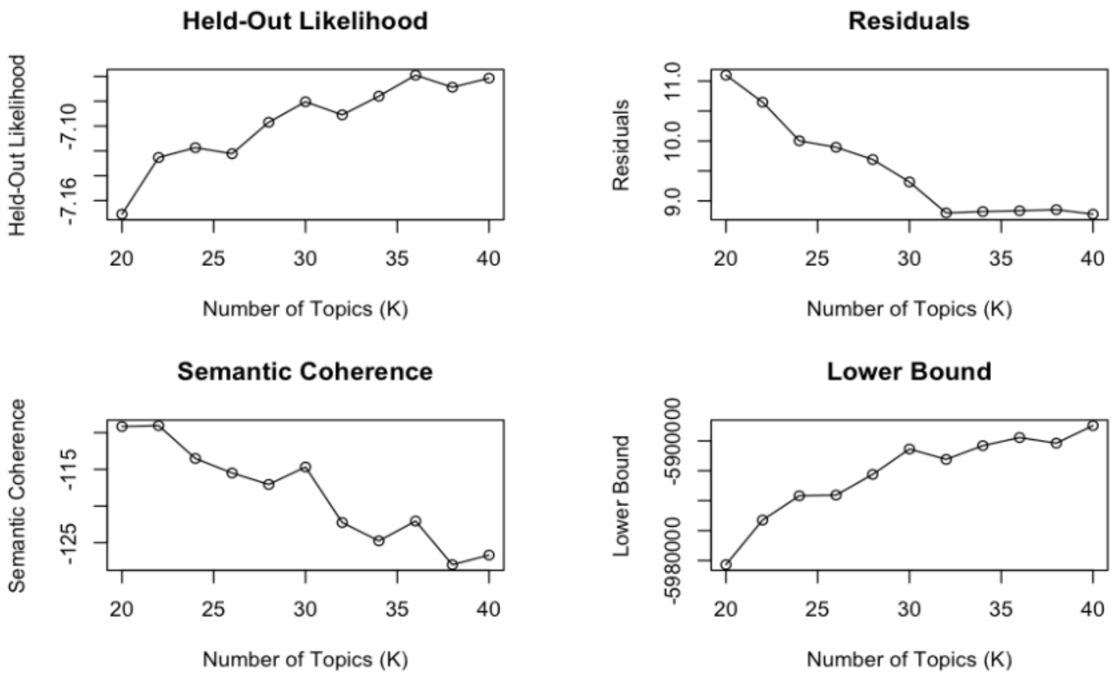

Figure 1: Diagnostic Values by Number of Topics

This is highest when K is large. Therefore, there is always a tradeoff between the two measures.

The figures suggest K = 30 is a good choice for the parameter.

7.2 Topics

The topics produced by the model were very coherent from a qualitative perspective. This is likely due to the nature of the dataset, which is already a very focused area of discourse in British parliament, i.e. there is less noise. The stm package in R provides the top probability words for each topic, as well as the top FREX, Lift and score words for each topic. In table 1 I show the 30 topics, as well as the highest probability words and the score words. Score divides the log frequency of the word in the topic by the log frequency of the word in other topics. I find that using both the highest probability words and score words helps the researcher label the topics. Using the function FindThoughts() from the stm package shows

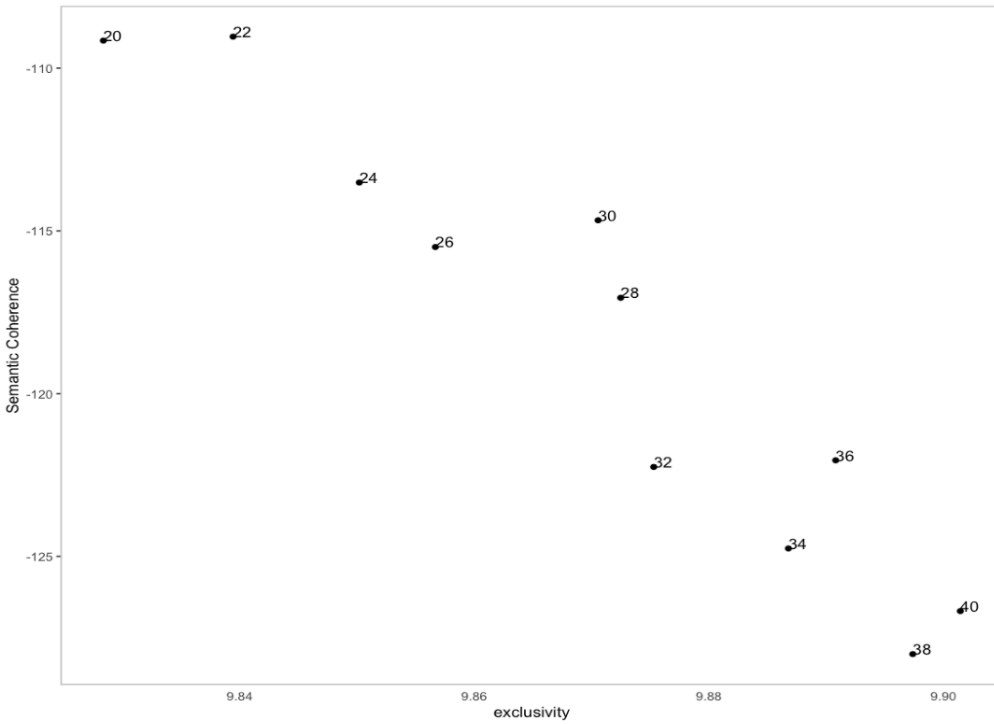

Figure 2: Exclusivity vs Semantic Coherence

key documents for each topic. These results help provide a meaningful label and reduces guesswork.

Almost all topics (except perhaps topic 28) are distinct (little to no semantic overlap with other topics) and meaningful. All of these topics were important parts of the Brexit negotiations. For example, the validity of trade deals (topics 4, 8, 10, 23, 27 and 29), immigration (topic 7) and future relations to the EU (topics 3, and 11) were all key aspects of the Brexit deal. The model also, however, captures more abstract topics quite well, such as topic 2, which captures possibilities and uncertainty/certainty. Topic 15, which captures complications in negotiations. The negotiations were famously difficult and lengthy. Topic 25 captures possibilities after Brexit, what *might* happen.

In summary, topic modeling on this unique dataset provides some topics that are rather clear and meaningful.

Description	Highest Prob	Score
1: regulation of pharmaceuticals	differ, mani, move, direct, use, medicin, ensur	medicin, differ, mani, move, direct, contract, regul
2: possible outcomes, certainty	may, concern, certain, impact, sort, reason, potenti	impact, concern, sort, risk, leav, may, certain
3: funding health research after Brexit	health, public, research, patient, nhs, univ, world	research, patient, student, univ, health, nhs, clinic
4: changes in trade agreements after Brexit	trade, countri, agreement, deal, world, tariff, free	trade, tariff, agreement, wto, free, deal, countri
5: negotiations with the EU	european, negoti, agreement, union, state, futur, relationship	negoti, agreement, european, union, relationship, state, norway
6: to-do list	need, year, now, put, work, thing, done	need, year, now, put, work, done, next
7: migrant workers	skill, work, employ, worker, immigr, labour, job	skill, worker, employ, labour, migrat, recruit, immigr
8: fisheries	local, fund, fish, manag, region, share, project	local, fund, fish, fisheri, mayor, region, fishermen
9: policing data, INTERPOL after Brexit	system, offic, oper, inform, home, polic, agenc	offic, polic, arrest, extradit, inform, home, europol
10: import/export of meat and dairy	sector, export, import, industri, market, tariff, europ	export, lamb, tariff, sector, market, industri, meat

11: EU law after Brexit	law, legisl, legal, court, parliament, withdraw, right	law, court, legisl, legal, justic, withdraw, parliament
12: creative industries in Europe	industri, opportun, well, unit, sir, great, kingdom	sir, industri, unit, creativ, kingdom, opportun, tourism
13: asylum seeking	countri, case, appli, right, exampl, individu, nation	countri, citizen, case, applic, resid, appli, status
14: transition period, implementation period	deal, process, plan, end, arrang, place, possibl	deal, plan, transit, period, arrang, cont, prepar
15: complications in negotiations	tri, way, deci, data, challeng, polit, difficult	data, polit, deci, adequaci, tri, commiss, mutual
16: devolvement of powers	polici, area, approach, devolv, framework, common, within	devolv, framework, polici, power, common, welsh, administr
17: supply chain	busi, cost, compani, suppli, small, chain, manufactur	busi, suppli, compani, cost, chain, manufactur, small
18: communication across parliament re: negotiations	issu, clear, work, discuss, depart, made, posit	discuss, depart, clear, engag, meet, respon, minist
19: welcoming guests to panels	question, ask, answer, thank, evid, much, help	thank, question, ask, answer, evid, morn, session
20: Northern Ireland border	ireland, northern, good, irish, protocol, republ, great	ireland, northern, protocol, republ, irish, unfett, vat
21: short- vs long-term consequences of Brexit	term, back, come, mean, long, get, short	back, term, come, long, get, short, mean

22: construction targets and energy	detail, target, transport, okay, turn, use, certif	target, heat, electr, transport, detail, certif, gas
23: the single market	market, servic, access, singl, free, movement, financ	market, singl, servic, custom, movement, free, regulatori
24: timeline, deadline	time, last, chang, given, day, month, week	time, month, week, last, vote, referendum, day
25: possibilities after Brexit (positive and negative)	point, like, might, someth, happen, whether, brexit	point, might, someth, like, brexit, view, whether
26: understanding, clarity in communication	talk, quit, sure, understand, bit, don, earlier	talk, bit, understand, quit, don, sure, littl
27: border checks of goods	border, check, custom, port, capac, good, requir	border, port, check, custom, freight, dover, holyhead
28: approximations	just, yes, much, realli, give, say, seem	yes, just, give, say, much, realli, figur
29: quality of produce and food	product, food, standard, produc, consum, anim, price	food, product, standard, sugar, produc, anim, price
30: agriculture in Scotland	agricultur, scotland, support, scottish, farm, farmer, environ	agricultur, farm, farmer, scotland, scottish, payment, support

Table 1: Topic labels, highest probability words and score words

7.2.1 Topics by committee

Another way to check the success of the method is to check documents grouped by committee. Within each committee, we can calculate the top topics by proportion, and verify that these topics align with the general agenda of the committee. By following this process and examining the top 5 topics, I found promising results.

For example, the Business, Energy and Industrial Strategy Committee's top topics included topics 17 (supply chain) and 10 (imports and exports of meat and dairy). The Culture, Media and Sport Committee's top topics included 12 (creative industries in Europe). The Energy and Climate Change Committee's top topic was topic 22 (construction targets and energy). Northern Ireland Affairs Committee's top topic was topic 20 (the Northern Ireland border).

8 Actor-topic networks

Discourse networks consists of topics and actors. Edges between actors (blue circular nodes) denote a shared topic. Edges between topics (purple square nodes) denote that an actor discusses both topics. Edges between topics and actors convey that the actor speaks about the given topic.

Figures 3 and 4 show the actor-topic networks for the corresponding years. This is a sample of the networks from the full time period, selected to show how differences in discourse networks can be visualized. I will not cover these types of networks extensively because although they can be visually interesting, there are too many nodes to be able to qualitatively analyze the position and importance of each node. Nevertheless, there are a few aspects of the networks that are visually apparent.

The 2018 network has more speakers, while the 2019 network is more sparse. The density of the 2018 network is apparent, and this is confirmed in a later section on network measures. In both networks, topic 21 (V21 on the networks - a square, purple node) are more removed

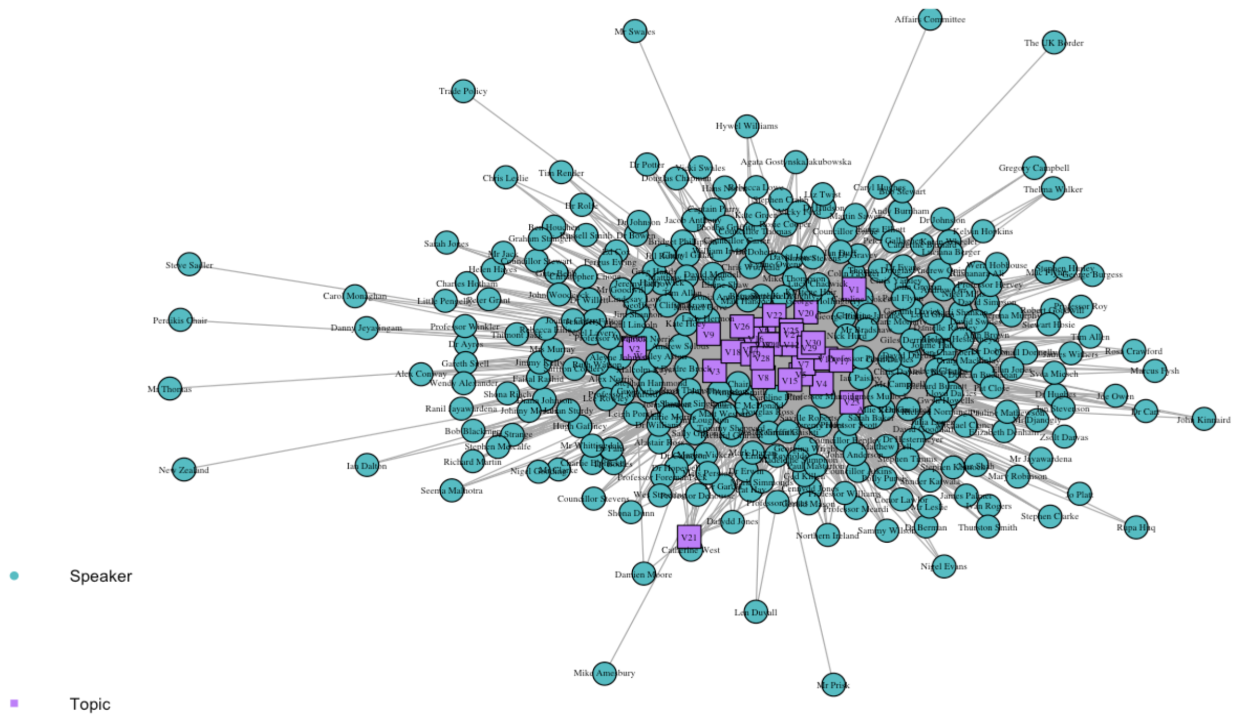

Figure 3: 2018 actor-topic network

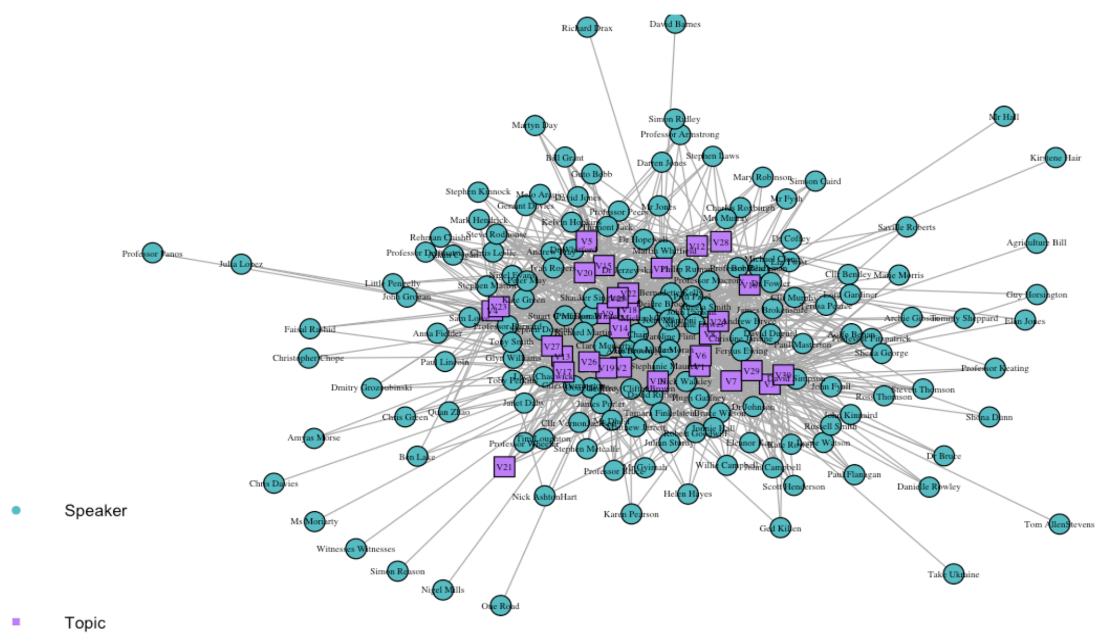

Figure 4: 2019 actor-topic network

from the central cluster of topics. This topic is about short- and long-term consequences of Brexit. This may denote that only a few actors were considering the lasting consequences of the referendum.

These figures can offer a first pass at understanding how discourse changes over time. However, in order to quantify changes, I examine network-level measures.

9 Network measures across time

There are several measures that quantify the structure of a network. These measures can be separated into node-level measures, meaning metrics that measure the position or importance of a given node, and network-level measures, meaning metrics that capture the structure of the network as a whole. In this paper, I compare the results between three network-level measures, namely edge density, average path, and transitivity. These measures, as will be apparent, are highly correlated and thus convey similar information about the networks. I chose to include all three for comprehensiveness. I briefly explain what each measure is.

Edge density is a measure that is a ratio of the number of edges that exist to the number of possible edges in the network. It captures the number of actual connections compared to the number of possible connections. As edge density increases, it means the overall density of the network increases.

Transitivity is a type of clustering coefficient. The one I use in this paper is the global clustering coefficient, which is the "ratio of triangles to connected triples" (Shizuka, 2019). That is, the measure captures how many nodes that share a connection are connected to each other. As transitivity increases, more nodes are connected to one another.

Finally, *average path* is a simple measure that calculates the average number of nodes you have to cross to get from one node to another, using the shortest possible path. As average path decreases, nodes are more closely connected to one another.

Figure 5 shows edge density, transitivity, and average path for each year in the dataset.

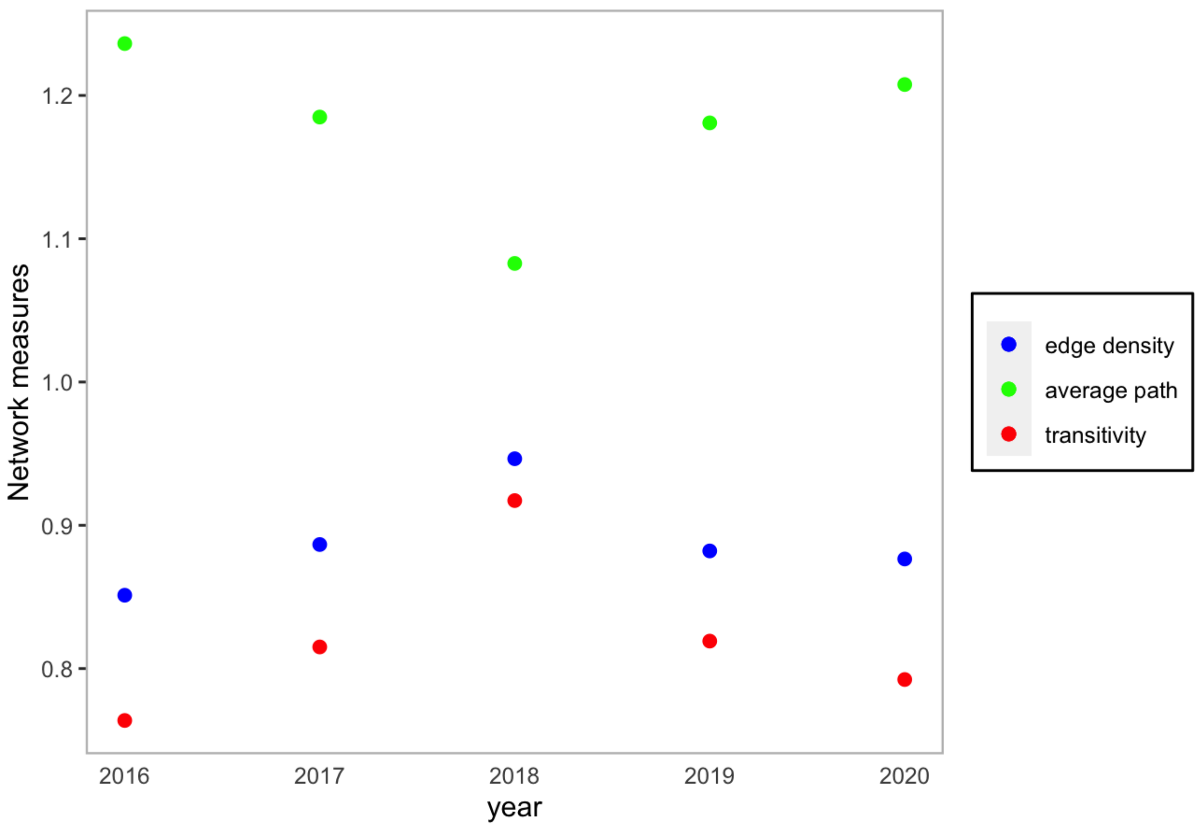

Figure 5: Edge density, average path and transitivity by year

All measure move similarly, with edge density and transitivity increasing from 2016 to '17 and 2017 to '18. Then there is a slight decrease in the subsequent years, 2019 and 2020. Average path decreases from 2016 to '17, and from 2017 to '18. All three show the same pattern, namely that 2018 exhibits a "peak" by all measures. The network becomes denser, leading up to 2018, information flows more freely and actors partake in more similar topics of discussion – a convergence in discourse.

10 Community Analysis

Community analysis was conducted using the walktrap method as discussed in the Data and Methods section of the paper. Networks only include actors for this step. Nodes have been colored by party affiliation. Since there are several parties in British parliament (10 in the dataset), it would not be easy to visually identify the party composition of different communities. To clarify this, I classify nodes into four categories; Labour, Conservative, other (MPs who belong to the other parties in parliament) and witness. This classification was chosen because the Labour party and the Conservative party are the two largest parties in parliament. Other parties with fewer seats in parliament are grouped together. Witnesses are included in the analysis because they are an important component of Committee meetings and thus represent a large proportion of the speakers in the data.

Results are shown in figures 6, 7, 8, 9 and 10. Community structure, meaning the composition of communities and count of communities, varies noticeably from year to year.

In 2016 (figure 6), two main communities emerge with a minor difference in the size of each. The community members consist mostly of Conservative MPs, Labour MPs and witnesses. Only two nodes from other parties are present. The distribution of the different speakers across the communities is rather even, with the red community consisting mostly of witnesses.

In 2017 (figure 7), interestingly, only a single community is detected. The community consists of members from all four categories of speakers. The detection of a single community means that the distribution of topics by speaker across the set of speakers was not different enough to result in the separation of speakers into separate communities.

Community analysis of the 2018 subset of the data (8) shows four communities, with two dominant communities (red and green) and two smaller communities (blue and purple). The red community is the largest of the four and consists heavily of witnesses. There is a central cluster of witnesses. The second largest community has a more even distribution of different types of speakers.

Community analysis of the 2019 data (9) shows two communities, with the red community being significantly larger than the blue community. There is no visible clustering of any type of speaker (Conservative, Labour, other, or witness), and witnesses and Conservative MPs seem to constitute most of the nodes. The blue community consists mostly of witnesses.

	2018 - Community 1	2018 - Community 2	2018 - Community 3
2017 - Community 1	46	31	1

Table 2: Communities in 2017 compared to 2018. Cells show number of members who shared a community in 2017 (rows) and 2018 (columns)

	2019 - Community 1	2019 - Community 2
2018 - Community 1	35	7
2018 - Community 2	23	2
2018 - Community 3	0	0
2018 - Community 4	0	0

Table 3: Communities in 2018 compared to 2019. Cells show number of members who shared a community in 2017 (rows) and 2018 (columns)

Finally, the 2020 community analysis (10) shows two communities of almost equal size. The blue community consists largely of witnesses. The red community has a more equal distribution of speakers.

Members are not always the same from year to year. However, for those who do maintain a presence from one year to the next, we can compare the membership across years.

Table 2 tabulates the community membership for speakers who were present in both the 2017 and the 2018 negotiations. The single large community in 2017 split into communities 1 (red) and 2 (green) in 2018.

For members who were present in the 2018 and 2019 negotiations, table 3 shows the tabulation of the membership. The largest community in 2018 largely remained in the same community in 2019. The second community in 2018 remained also more or less in the same community (2019 - Community 1). A few members switched community to community 2.

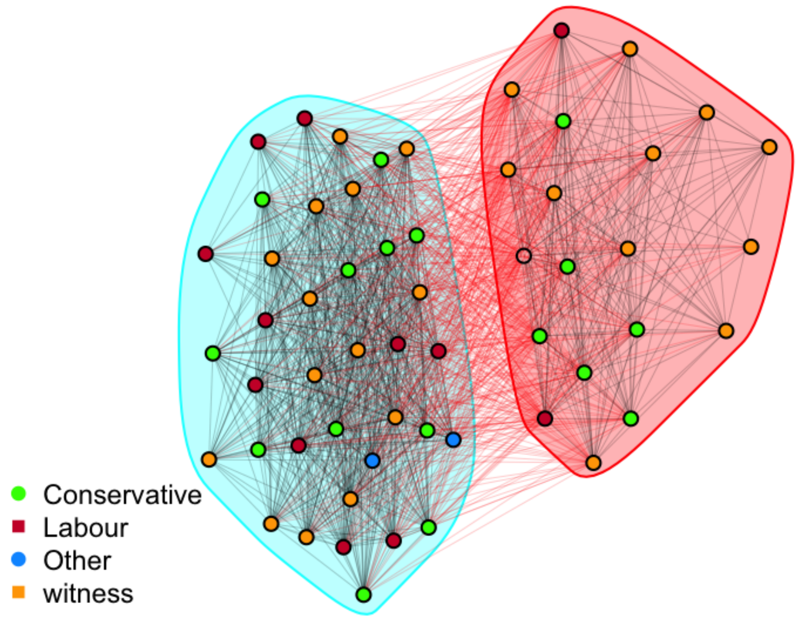

Figure 6: 2016 Community analysis

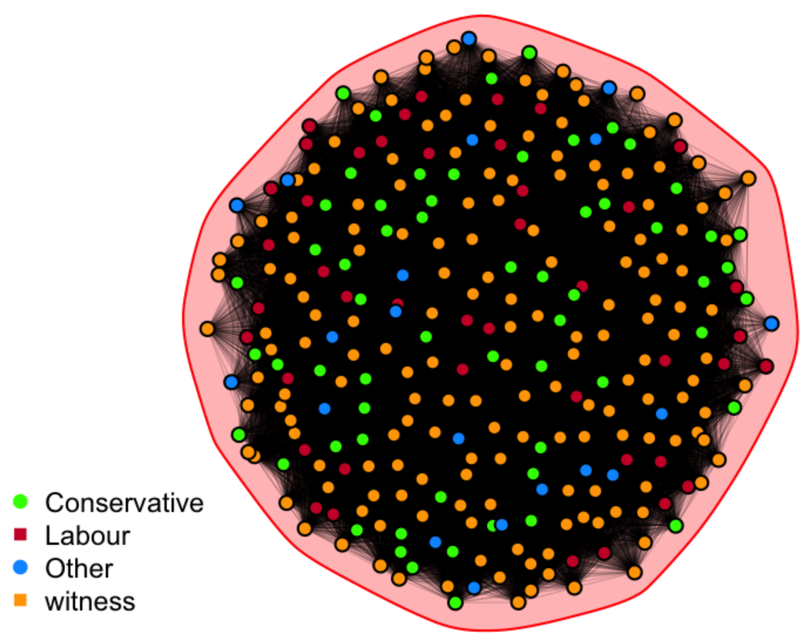

Figure 7: 2017 Community Analysis

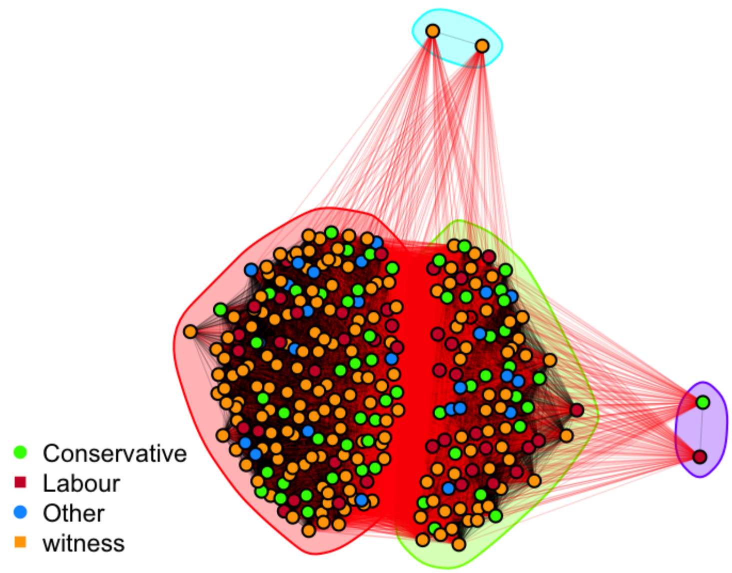

Figure 8: 2018 Community Analysis

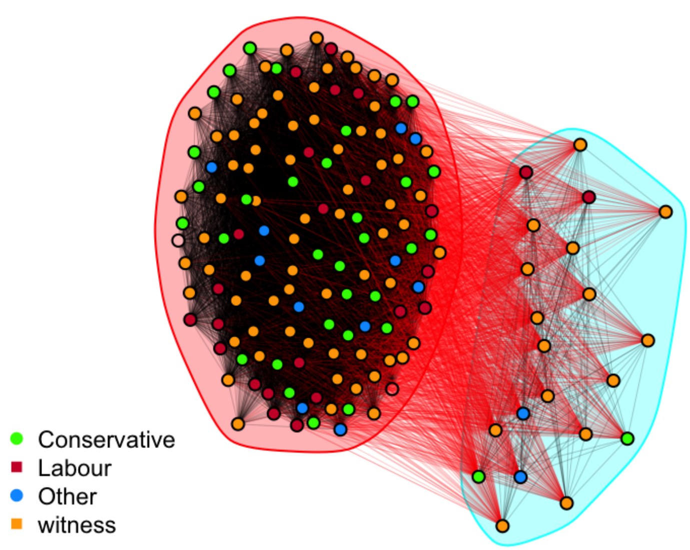

Figure 9: 2019 Community Analysis

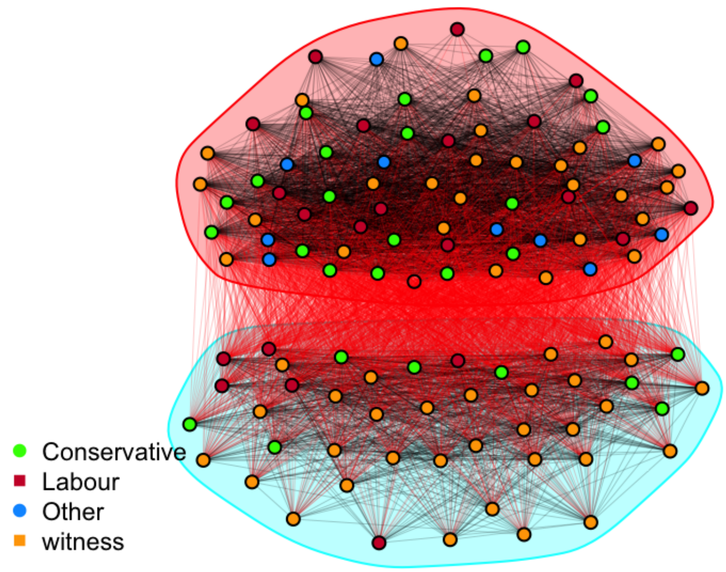

Figure 10: 2020 Community Analysis

11 Discussion

The discourse networks measures over the years (figure 5) show a pattern of growing density, culminating in a peak in the flow of information in 2018, followed by a relative decrease in the density and flow of information.

To interpret the metrics, one needs to consider the timeline of the negotiations. 2016 was the year of the referendum when the British public voted to leave the European Union (June 2016). The Prime Minister at the time, David Cameron, resigned and Theresa May took over the leader of the Conservative party, i.e., May became the new Prime Minister. The outline of the negotiation process (the issues that would need to be negotiated) were outlined in December of that year. Edge density and transitivity are the lowest of the five-year period. Average path is the highest. This may represent the lack of coherent discourse around Brexit. Key issues had not been identified nor were negotiations underway.

We find support for this in figure 6, the community analysis of 2016. The size of the network is rather small, showing that not many actors had become involved. MPs from the Conservative and the Labour party are the main political nodes on the graph. Witnesses are present in the network but are mostly present in the red community.

2017 marked the start of negotiations. In March, article 50 was triggered, formally initiating the process of the UK leaving the EU (Council, Council). In June negotiations began with the EU (meetings at the parliament-level were held before this date). Simultaneously, the government announced snap elections. The Conservative party remained the largest party in parliament but lost 13 seats. The slight increase in edge density and transitivity, and decrease in average path, shows that the negotiations were slightly underway but not moving at a fast pace because MPs were preoccupied with the snap election.

Figure 7 (community analysis in the 2017 network) shows a drastic change in structure compared to the 2016 community network (figure 6). The lack of separation of nodes into multiple categories marks a lack of organization of actors into communities. This may reflect the general disarray of this year of negotiations. There was not much progress in negotiations.

This unique finding of a single community may suggest that there need to be multiple communities to spur negotiations. The presence of different communities shows the presence of different focal topics in each distinct community. The grouping of all actors into a single community may be due to a common discursive focus, where different opinions are sure to exist. This may inhibit progress on the community's discursive focus, because the community is simply overwhelmed by political positions.

2018 was a particularly fraught year for Brexit negotiations. Theresa May did not ultimately manage to pass a deal through the House of Commons. Discussions of a no-deal Brexit emerged in June 2017 (Wikipedia, 2024) despite the government reporting great progress on negotiations, claiming "95% of the Brexit deal was done" (BBC, 2018). The draft deal finally failed in parliament in December 2018. Therefore, the high level of transitivity, high density of the network may be associated with an increase in the urgency of the discussions happening in Committees and chamber. The disparate topics of Brexit all became salient simultaneously to the actors involved (therefore increasing the number of edges) because there was a sense of urgency in needing to finalize the deal. A draft document was produced and debated in parliament in December. It was not voted on in 2018 as discussions were delayed to early 2019.

Figure 8 (community analysis for 2018) supports this interpretation as there are two main communities (red and green), each with a large number of actors. Table 2 confirms the splintering of one homogeneous community into multiple communities. Members who were present in both years of negotiations split almost evenly into the main 2018 communities. In the context of community analysis, it is surprising to have such few communities given the number of actors is so high (n = 295). These two distinct communities may reflect the progress of the negotiations, as each community is sizeable and consists of actors of all types. Importantly, witnesses show an important and rather central presence in the two main communities. Witnesses' job is to bring opinions and expertise to the meetings, helping inform the political decision-making process. Witnesses could be politicians in the EU, or interest groups from the UK (e.g., farmers or fishermen). Their input into the negotiation process was crucial for determining the decisions made.

In 2019, another five-day debate began, the draft exit document failed three votes in January, early March and late March. The deadline to exit the EU was extended and Theresa May resigned. Negotiations continued and the deadline was extended twice more in April and October. Finally, a draft was passed on December 20th. The slight downward trend (upward for average path) may mark the diffusion of the discourse, as final details were sorted.

Figure 9 (community analysis for 2019) shows one large community (red) and one smaller community (blue). Witnesses are less present in this network, reflecting that the bulk of the negotiations with the EU were complete and the remaining deliberations were at the level of the British parliament. There are, however, a large proportion of witnesses in the blue community. This may reflect that witness input was a distinct contribution to the discussions. The tabulation of members from 2018 to 2019 (table 3) shows that the communities were more stable from 2018 to 2019, as members who shared a community in 2018 often shared a community in 2019.

In January 2020, the UK finally exited the European Union. Although negotiations with the EU concluded, there were and continue to be, parliamentary meetings to discuss the transition period and issues that arise in the areas impacted by Brexit. These measures are only slightly different than those from 2019, possibly capturing the slowing down of discourse. Information is no longer flowing as intensely.

Figure 10 still includes witnesses because although the UK had officially exited the EU, discussions about the technicalities and practice of the new EU-UK relations were still being discussed. Similar to the 2019 community analysis, the blue community consists mostly of witnesses, which may reflect that the witnesses were involved to give expertise on the new international relations. The conclusion of negotiations is reflected in the relative sparsity of the network.

12 Conclusion and future directions

Discourse is a dynamic phenomenon. The key issues, the key actors and the different groups in discourse are constantly in flux. In this thesis, I have shown an approach to studying the structure of discourse over time that eliminates the need for tedious, manual annotation of texts. Using the Brexit negotiations as a case study, I showed how STM and network analysis can be combined to study how different discourse was structured over the nearly five years of negotiations. I showed how the flow of information changed over time using network-level measures and how different political actors grouped together in communities.

I mainly addressed network- and community- level variation over time. Future research is needed to examine the actor-level and topic-level variation across time. This requires an in-depth examination of topics and actors. Research could answer how key actors or communities may causally change the structure of discourse. Studying the political implications of changes in discourse such as change in policy is another potential direction for future studies. How do changes in discourse result in a changes of policy or the formation of coalitions in government. Lastly, future research should compare how topic modeling compares to Leifeld's manual annotation. Sentiment analysis could be used to replace Leifeld's coding of actors' attitudes towards a given issue (Leifeld, 2013).

13 Bibliography

Ahmed, Wasim, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. 2020, May. "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data". *Journal of Medical Internet Research* 22 (5): e19458.

Archer, Dawn. 2017, December. "Mapping Hansard Impression Management Strategies through Time and Space". *Studia Neophilologica* 89 (sup1): 5–20. Publisher: Routledge _eprint: https://doi.org/10.1080/00393274.2017.1370981.

Archer, Dawn. 2018, July. "Negotiating difference in political contexts: An exploration of Hansard". *Language Sciences* 68 : 22–41.

BBC. 2018, October. "Theresa May says 95% of Brexit deal is done". `https://www.bbc.com/news/uk-politics-45938444`.

Bitten, Harry. 2019, July. "Detecting communities in a language co-occurrence network". `https://towardsdatascience.com/detecting-communities-in-a-language-co-occurrence-network-f6d9dfc70bab`.

Council, European. . "Timeline - The EU-UK withdrawal agreement". `https://www.consilium.europa.eu/en/policies/eu-relations-with-the-united-kingdom/the-eu-uk-withdrawal-agreement/timeline-eu-uk-withdrawal-agreement/`.

Foucault, Michel. 1972. *The Archaeology of Knowledge.* Pantheon Books.

Foxlee, Neil. 2018, June. "Pivots and Levers: Political Rhetoric around Capitalism in Britain from the 1970s to the Present". *Contributions to the History of Concepts* 13 (1): 75–99. Publisher: Berghahn Journals Section: Contributions to the History of Concepts.

Harper, Heather, Cybele Kappos, and Stefanie Neumeier. 2023. "The Disrupting Effect of

Focusing Events on Immigration Discourse in the UK: an Automated Discourse Network Analysis". *Working Paper*.

Kruger, Haidee, Bertus van Rooy, and Adam Smith. 2019, September. "Register Change in the British and Australian Hansard (1901-2015)". *Journal of English Linguistics* 47 (3): 183–220. Publisher: SAGE Publications Inc.

Leifeld, Philip. 2013. "Reconceptualizing Major Policy Change in the Advocacy Coalition Framework: A Discourse Network Analysis of German Pension Politics". *Policy Studies Journal* 41 (1): 169–198. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psj.12007.

Leifeld, Philip. 2017, September. "Discourse Network Analysis". In J. N. Victor, M. N. Lubell, and A. H. Montgomery (Eds.), *The Oxford Handbook of Political Networks*. Oxford University Press.

Mair, Christian. 2019, August. "Brexitiness: The Ebbs and Flows of British Eurosceptic Rhetoric since 1945". *Open Library of Humanities* 5 (1). Number: 1 Publisher: Open Library of Humanities.

Mollin, Sandra. 2007, November. "The Hansard hazard: gauging the accuracy of British parliamentary transcripts". *Corpora* 2 (2): 187–210. Publisher: Edinburgh University Press.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2019. "stm: R Package for Structural Topic Models". *Journal of Statistical Software*.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014, October. "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58 (4): 1064–1082.

Shizuka, Dai. 2019, June. "Measuring Networks Part 1: centrality and Global Measures". `https://dshizuka.github.io/networkanalysis/04_measuring.html`.

Slembrouck, Stef. 1992, May. "The parliamentary Hansard 'verbatim' report: the written construction of spoken discourse". *Language and Literature* 1 (2): 101–119. Publisher: SAGE Publications Ltd.

Smith, Natalie R., Paul N. Zivich, Leah M. Frerichs, James Moody, and Allison E. Aiello. 2020, October. "A Guide for Choosing Community Detection Algorithms in Social Network Studies: The Question Alignment Approach". *American Journal of Preventive Medicine* 59 (4): 597–605.

Tolvanen, Juha, James Tremewan, and Alexander K. Wagner. 2021, October. "Ambiguous Platforms and Correlated Preferences: Experimental Evidence". *American Political Science Review*: 1–17. Publisher: Cambridge University Press.

Tonidandel, Scott, Karoline M. Summerville, William A. Gentry, and Stephen F. Young. 2022, October. "Using structural topic modeling to gain insight into challenges faced by leaders". *The Leadership Quarterly* 33 (5): 101576.

Vaughan, Michael. 2020, April. "Talking about tax: the discursive distance between 38 Degrees and GetUp". *Journal of Information Technology & Politics* 17 (2): 114–129. Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2019.1705220.

Wikipedia. 2024, January. "Brexit negotiations". `https://en.wikipedia.org/w/index.php?title=Brexit_negotiations&oldid=1198532094`. Page Version ID: 1198532094.