UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Measuring changing levels of political polarization in the U.S. Congress with W-NOMINATE

First examiner:

Javier Garcia Bernardo

Candidate:

Lars Teunissen

Second examiner:

Elena Candellone

In cooperation with:

Dion Custers & Esther Stehouwer

June 27, 2024

Abstract

Polarization is a prominent theme in U.S. politics, with a commonly held view that the level of polarization in political discourse has increased substantially in recent years. Here we want to investigate to what extent this is the case. Using the Poole and Rosenthal W-NOMINATE scores, it is possible to analyze political choices to gain insights into legislators' political beliefs. Placing legislators on an ideological spectrum reveals the distance between their beliefs, which can be translated into a measure of polarization. This can be done through various measures, including the bimodal coefficient, which shows the bimodality of the distribution.

We illustrate how polarization changes in the U.S. political landscape from 1973 to the present day. Dividing our data into multiple political themes we analyze changes in polarization per theme in this temporal frame. Our analysis indicates significant fluctuations between consecutive congresses, influenced by notable historical events. Nonetheless, taking the full temporal aspect into account we can detect a definite increase in our measurements of political polarization over our larger time frame. This nuanced understanding of political polarization dynamics supports views of increasing polarization while challenging the perception of a steady rise in all aspects of legislative behavior.

Contents

Appendix

1. Introduction

Political polarization has in recent times emerged as a salient feature of contemporary discourse, especially within modern democratic societies [1]. The often made claim that polarization is on an inexorable rise has become widely spread in the political narrative [2]. It is frequently cited as either a contributing factor or even the sole reason for legislative gridlock and societal divisions. However, such a disposition towards the subject fails to take into account nuances vital to understand the veracity of this claim fully. For instance, not every political theme necessarily experiences such societal division at the same time. Discourse on a topic as significant as polarization should be informed by more academic works, with measured trends and factual evidence supporting otherwise argumentative statements. Measuring these trends is the main focus of this research paper.

The Poole and Rosenthal W-NOMINATE scores can be used as a framework to divide legislators by ideology based on their voting patterns [3]. It places legislators onto a spatial ideological spectrum, which can function as a basis for assessing the degree of polarization within a congress [4]. Using such a quantitative approach allows us to perform further analysis to determine how political beliefs shift over time, thereby revealing a measure of the underlying dynamics of polarization.

To effectively utilize the W-NOMINATE model, we want to begin with comprehensive and well-prepared data. This involves data gathering, cleaning, and standardization from various sources to ensure the integrity and reliability of our analysis. By addressing these foundational tasks, we can provide robust insights and accurate findings for our research objectives.

All the data in these datasets can then be analyzed to discover the topics the bills belong to. We employ the Latent Dirichlet Allocation (LDA) topic modeling algorithm for this step. LDA facilitates the extraction of topics from large corpora of documents [5]. Dividing our data into topic subsets allows us to identify polarization levels beyond the overall measure, detect issue-specific differences, and observe varying polarization trends. This methodological approach contributes to a more contextual insight into how polarization is affected by changing opinions on dominant political matters over time.

To properly address questions about changes with respect to past levels of political polarization, we must analyze our ideological scores on an extended temporal framework. By examining votes made between 1973 and the present, we are able to measure changes in polarization when they occur and reveal changes in the political landscape of the U.S.

Analyzing the ideological positions of legislators can reveal the extent of polarization within a legislative body. The bimodality coefficient reveals the measure of distinction between the two parties on this ideological spectrum [2]. Incorporating the temporal dimension into this analysis allows us to observe the change in polarization over time and across different topics, providing a clear view of the total political dynamics.

In this paper we present an in-depth analysis of polarization trends using LDA topic modeling, W-NOMINATE scores, and bimodality coefficients within an extended temporal framework, highlighting the changes over time as revealed by the bimodality coefficient to answer a specific question: How can we define polarization and how does it change over time for different topics within the House of Representatives from the 93rd until the 118th congress?

1.1 Related work

The W-NOMINATE model has seen extensive use in similar work, to analyze legislative behavior and ideological positions. Outside of the U.S. Congress, it has also been used for research on the European Parliament and the United Nations General Assembly, to study voting behavior [6] [7]. Furthermore, the model has been used to examine institutional investigators on sustainability practices, revealing regional differences [8]. The W-NOMINATE model has therefore shown its versatility and robustness for use in the field of political science.

2. Data

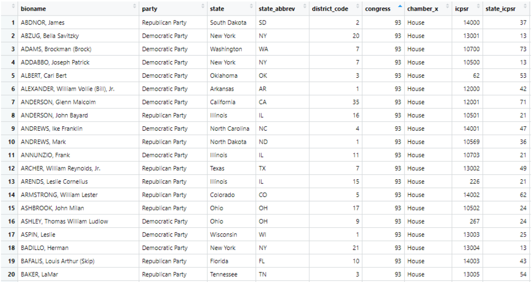
2.1 Description of the data

Our original data was acquired from two different sources. The voteview website [9] provides data on all U.S. Congress members as well as information on all votes taken per Congress. Each dataset is filtered to only contain information belonging to the House of Representatives. It provides us with information on every vote taken by every Congress member. This comprehensive data is suitable for our research because it offers detailed roll call data, which is needed to calculate the ideological positions of legislators. Another reason for this particular dataset's suitability is that its reliability and accuracy have already been proven in earlier research on legislative behavior [4], which further supports its use in this research.

The rest of our data was taken from the ProPublica website [10], which provides detailed information about every bill voted on in Congress. We focus on its full summaries to do robust topic modeling. When using the Latent Dirichlet Allocation algorithm, we will categorize our bills based on their descriptions. ProPublica's [10] open-source data on these bills ensures a solid foundation for any topic modeling algorithm to extract useful information.

Initially, our datasets contained data per Congress and were therefore very numerous, but these were immediately added together to set our starting number of datasets to three, containing information for all Congresses. These datasets are:

- The members dataset, which includes variables such as member names and their party affiliation per Congress. 2.1
- The role call dataset contains all legislators' votes made in Congress on all bills.



	bioname	party	state	state_abbrev	district_code	congress	chamber_x	icpsr	state_icpsr
1	ABDNOR, James	Republican Party	South Dakota	SD	2	93	House	14000	37
2	ABZUG, Bella Savitzky	Democratic Party	New York	NY	20	93	House	13001	13
3	ADAMS, Brockman (Brock)	Democratic Party	Washington	WA	7	93	House	10700	73
4	ADDABBO, Joseph Patrick	Democratic Party	New York	NY	7	93	House	10500	13
5	ALBERT, Carl Bert	Democratic Party	Oklahoma	OK	3	93	House	62	53
6	ALEXANDER, William Vollie (Bill), Jr.	Democratic Party	Arkansas	AR	1	93	House	12000	42
7	ANDERSON, Glenn Malcolm	Democratic Party	California	CA	35	93	House	12001	71
8	ANDERSON, John Bayard	Republican Party	Illinois	IL	16	93	House	10501	21
9	ANDREWS, Ike Franklin	Democratic Party	North Carolina	NC	4	93	House	14001	47
10	ANDREWS, Mark	Republican Party	North Dakota	ND	1	93	House	10569	36
11	ANNUNZIO, Frank	Democratic Party	Illinois	IL	11	93	House	10703	21
12	ARCHER, William Reynolds, Jr.	Republican Party	Texas	TX	7	93	House	13002	49
13	ARENDS, Leslie Cornelius	Republican Party	Illinois	IL	15	93	House	226	21
14	ARMSTRONG, William Lester	Republican Party	Colorado	CO	5	93	House	14002	62
15	ASHBROOK, John Milan	Republican Party	Ohio	OH	17	93	House	10502	24
16	ASHLEY, Thomas William Ludlow	Democratic Party	Ohio	OH	9	93	House	267	24
17	ASPIN, Leslie	Democratic Party	Wisconsin	WI	1	93	House	13003	25
18	BADILLO, Herman	Democratic Party	New York	NY	21	93	House	13004	13
19	BAFALIS, Louis Arthur (Skip)	Republican Party	Florida	FL	10	93	House	14003	43
20	BAKER, LaMar	Republican Party	Tennessee	TN	3	93	House	13005	54

Figure 2.1: Some member data variables. Per Congress, we have the names and party affiliations of every member.

- The bills dataset provided by ProPublica contains extensive information on the content of all bills handled by Congress.

The total member data consists of 11789 rows with information on names, party affiliation, and unique identifiers containing all members between congresses 93 and 118. 2.1 Congress 111 is the largest, with 457 legislators found in the data. The smallest congress is Congress 98 with 440 legislators. The average number of legislators per Congress is 445. Within some Congresses, we find clear majorities for either the democratic or republican parties. For instance, the highest democratic majority is found in Congress 95. This congress counts 295 democratic legislators. The highest number of Republicans exists in Congress 114, where 251 legislators are affiliated with that party.

The three legislators remaining in Congress the longest in our dataset are Donald Edwin Young, sitting from Congress 93 until 117 for Alaska, John Jr Conyers, sitting from Congress 93 until 115 for Michigan and Charles B. Rangel, sitting from Congress 93 until 114 for New York. For further information on the member votes and bills dataset, see appendix A.

2.2 Preparation of the data

As written above, the initial data contains information from Congress 93 until Congress 118. The reason for this is a lack of digitized information on roll calls from before 1973. Considering how important large corpora of data are to retrieve reliable topic modeling outputs, the collective decision was made to set 1973 as the starting point for this research. The bills dataset contains information provided by ProPublica that provides full summaries for bills from this point. Yet, it does not provide comprehensive information on roll calls that are not bills, such as choosing the house speaker or motions to adjourn. Therefore, these were removed to make sure every row that remains in the data can be used for topic modeling. Furthermore, extra information on bills was added, such as 'policy' and 'subjects' to further support topic modeling performance by allowing us to compare our output with the different themes provided in these columns. The resulting dataset contains 28449 bills over the 26 congresses.

The members dataset did have an id system, but these seemed to contain multiple errors when calculating their ideological positions and were therefore not suitable for further use. A new id for unique party and name combinations was created accordingly. This change resulted in 2246 unique combinations for a total of 11580 rows. This id was also added to the votes data, to make the different datasets more integrative. Furthermore, we added a binary variable to the votes data which shows 1 if the vote was 'Yea' and 0 otherwise. The final dataset for the votes contains 12,343,235 votes.

3. Methods

Our methodological approach utilizes two distinct techniques necessary to answer the research question: Latent Dirichlet Allocation (LDA) [5] to do topic modeling and W-NOMINATE [4] to determine legislators' ideological positions on a scale.

3.1 Topic modeling with LDA

Latent Dirichlet Allocation (LDA) is an unsupervised generative probabilistic model designed to retrieve latent topics in large corpora of documents [11]. This makes it a suitable topic modeling method to use on our dataset of bills containing the summaries that we created using the ProPublica [10] dataset. It inherently expects multiple topics to exist within a document and that each word in the document is attributable to one of these topics. By applying this model to the bill data, LDA can determine coherent topics for each bill. The process of using LDA further requires a significant amount of text preprocessing, as well as vectorization using Term Frequency-Inverse Document Frequency (TF-IDF) [12]. The LDA algorithm is used in accordance with the fundamentals provided by the scikit-learn library [13].

LDA models its input according to the following generative process:

1. For each topic k, choose a distribution $\phi_k \sim \mathrm{Dir}(\beta)$.
2. For each document d:
   (a) Choose a distribution $\theta_d \sim \mathrm{Dir}(\alpha)$.
   (b) For each word w_{dn}:
      i. Choose a topic $z_{dn} \sim \mathrm{Mult}(\theta_d)$.
      ii. Choose a word w_{dn} based on z_{dn} and ϕ.

The meaning of the distributions are:

- Dir(α) and Dir(β) are called Dirichlet distributions. They help us decide how likely different topics are in a document and how likely different words are in a topic.
- α is a parameter that controls how topics are mixed in a document. If α is high, each document will talk about many topics. If α is low, each document will talk about fewer topics.
- β is a parameter that controls how words are mixed in a topic. If β is high, each topic will use many different words. If β is low, each topic will use fewer words.
- θ_d is the topic distribution for document d and is chosen using α.
- ϕ_k is the word distribution for topic k and is chosen using β.

These distributions help us model how documents are made up of topics and how topics are made up of words [11].

The implementation begins with the standardization of the input documents, in this research the corpus of bill summaries. During this preprocessing phase all text is converted to lowercase, numbers and punctuation are removed as well as common stopwords in the English language. These common procedures were not adequate to generate usable LDA output, since many words were common in the bill summaries that were not removed but did have a significant impact on the topic modeling algorithm. Therefore, a list of impactful but undescriptive words was created by manually checking LDA output and finding words occurring in most unique topics. These words were then removed from the input corpus in the preprocessing phase. For the list of words, see Appendix B.

This phase is critical to retrieve high-quality LDA output. Common phrases and English stopwords would otherwise take precedence over descriptive words in their topics, since LDA will assume those most often seen words in a topic to contribute the most to its distinction from other topics. This same outcome would happen with bill-specific terms in the case of this research, such as 'election' and 'appropriations'. Without such steps, the topics created by LDA would not nearly be descriptive enough to extrapolate to real-world themes.

The remaining corpus was thereafter vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) [12], to capture the importance of the remaining words in the context of our corpus. It consists of two parts: TF, which measures how frequently a term occurs in a document, and IDF which measures how important a term is across the entire corpus. Together they denote words that are important to specific documents but not across all documents. This further enhances LDA output, since emphasizing the most informative terms in the corpus is beneficial to capturing clear topic distinctions. Furthermore, since bills often have common phrases such as "Requires the Secretary" or "sets forth the rule", using TF-IDF lowers the influence of such phrases so that LDA focuses only on those remaining high-scoring terms that can properly define topics.

In implementing LDA itself, we used the Scikit library [13] due to its optimized nature when dealing with large datasets while also being suitable for preprocessing tasks such as TF-IDF. LDA is then applied to the created TF-IDF vector using the Scikit-learn implementation [13]. The number of topics to be extracted from the corpus is set to seven; for each of these topics, the top 10 most descriptive terms are returned. For more information on our choice for seven topics, see Appendix B.

Each document is assigned to a topic using a probability distribution provided by LDA. This implementation returns the seven most coherent topics that can be found within the bills dataset. 4.1

3.2 Coherence tests

Coherence scores were used to gain a more profound understanding of how well the output topics were generated. This phase of topic modeling output analysis is supported by Gensim [14], providing a robust framework for evaluation. The CoherenceModel class, using the ′c_v′ coherence measure, evaluates semantic similarity on high-scoring words within a topic.

Our approach uses cosine similarity between word vectors. The coherence score for a topic is the average cosine similarity between the word embeddings of the top N words in the topic, where N refers to the number of top words that are being considered for our calculation (in our case this is 10). Coherence is defined through this function:

$$\text{Coherence}(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \text{cosine}(v_{w_i}, v_{w_j})$$

where v_{w_i} and v_{w_j} are the word vectors for w_i and w_j. [15]

Coherence score returns a numerical value between zero and one, providing an easy to interpret metric to ascertain topic consistency. In the context of our research, this helps us to decide how meaningful our final output on the bills within a topic is in the context of the overall political theme.

3.3 The W-NOMINATE model

W-NOMINATE (weighted NOMINATE) is commonly used to estimate the ideological positions of legislators based on their voting behavior. It is an extended version of the original NOMINATE (Nominal Three-Step Estimation). [3] W-NOMINATE also allows for multiple dimensions in the spatial model. [4] We assume that legislators and bills can be represented as points in some dimensional space (in our case one-dimensional). A legislator's ideal point on this space, as well as a bill's position, influences whether the legislator votes 'Yea' or 'Nay'. W-NOMINATE is a probabilistic model that estimates these positions. The basic W-NOMINATE scaling method is:

Let v_{ij} denote the vote of legislator i on bill j, where v_{ij} takes values from a set, in our research {(4, 5, 6), (1, 2, 3), (7, 8, 9)} corresponding to 'Nay', 'Yea', and Missing. The model estimates legislator i's ideal point x_i and bill j's ideal point y_j from the votes using a latent variable model.

$$v_{ij} = \text{sign}(x_i - y_j) \cdot \left(1 - \exp\left(-\frac{|x_i - y_j|}{\gamma}\right)\right)$$

where:

- x_i is the ideal point of legislator i,
- y_j is the ideal point of bill j,
- γ is a scaling parameter that controls the range of the utility function. [4]

The biggest strength of W-NOMINATE as opposed to regular NOMINATE lies in its potential to use more than one dimension, however in running the model we found that democrats and republicans were accurately captured by a single dimension making this step unnecessary. 4.2 The scree plot shows how much variance is explained by each dimension, thereby visualizing the necessary amount of dimensions needed to accurately capture the underlying structure. W-NOMINATE has been used a lot in similar research [6] [7] [8] which makes it a suitable method to use here to find the ideological structures in our data.

3.4 The Bimodality Coefficient

The bimodality coefficient is a function that assesses to what extent a distribution can be seen as bimodal, or having two distinct peaks within the data [16]. The function is defined in the following way:

$$BC = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

where:

$$\begin{aligned} g &: \text{the skewness of the distribution,} \\ k &: \text{the kurtosis of the distribution,} \\ n &: \text{the number of observations.} \end{aligned}$$

- If the output of the bimodality coefficient function exceeds 5/9, or around 0.555, our distribution is likely bimodal. [16]
- If the output of the bimodality coefficient function does not exceed 5/9, our distribution is likely unimodal.

This function is used on the distributions returned by the W-NOMINATE model and translates the ideological positions of all legislators to a single value that denotes if there are two distinct groups visible in this data. Furthermore, a higher value will show the measure to which such groups are distinct.

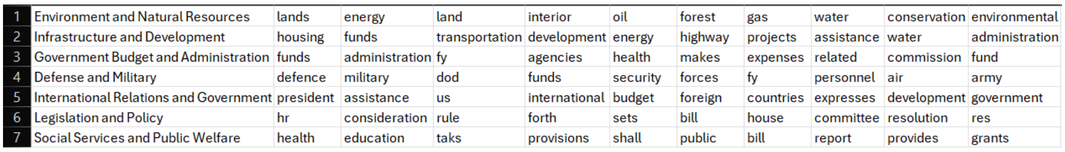
3.5 Bootstrapping

To assess the reliability of the Bimodality coefficient we use bootstrapping. We provide a measure of uncertainty to our method to make outcomes of polarization in Topic-Congress combinations more reliable. Bootstrapping consists of three parts:

- **Resampling:** We generate 1000 bootstrap samples by random sampling with replacement from the original data points.
- **Recalculating:** For each sample, calculate the bimodality coefficient.
- **Estimation:** From the created distribution, derive the standard error of the original bimodality coefficient for the Topic-Congress combination [17].

4. Results

Applying LDA to the preprocessed corpus of text yielded seven topics, each characterized through their 10 most descriptive terms. All topics represent distinct political themes. 4.1



1	Environment and Natural Resources	lands	energy	land	interior	oil	forest	gas	water	conservation	environmental
2	Infrastructure and Development	housing	funds	transportation	development	energy	highway	projects	assistance	water	administration
3	Government Budget and Administration	funds	administration	fy	agencies	health	makes	expenses	related	commission	fund
4	Defense and Military	defence	military	dod	funds	security	forces	fy	personnel	air	army
5	International Relations and Government	president	assistance	us	international	budget	foreign	countries	expresses	development	government
6	Legislation and Policy	hr	consideration	rule	forth	sets	bill	house	committee	resolution	res
7	Social Services and Public Welfare	health	education	taks	provisions	shall	public	bill	report	provides	grants

Figure 4.1: Definitive topics (left) and their 10 most descriptive terms from most to least important.

The coherence score output over our final LDA results is as follows:

Topic	Coherence Score
Environment and Natural Resources	0.4478
Infrastructure and Development	0.4243
Government Budget and Administration	0.4259
Defense and Military	0.5620
International Relations and Government	0.5332
Legislation and Policy	0.5325
Social Services and Public Welfare	0.2867

Table 4.1: Coherence Scores for Each Topic.

The choice of seven topics was empirically determined by taking into account both topic granularity as well as the amount of bills per topic, to support both W-NOMINATE input requirements and output quality. W-NOMINATE will require sufficient bills to work and after dividing them by 26 for Congresses and 7 for topics it meets this requirement for nearly every Congress-Topic combination.

We consider coherence scores higher than 0.5 to be good, scores higher than 0.4 to be decent, and scores lower than that to be poor. [18] [19] This suggests that three of our created topics are highly distinct and coherent, resulting in meaningful further analysis of these topics. Three more have scores above 0.4, making them still coherent and distinct but also containing more noise than can be found in those with scores above 0.5. They are still valuable for research on their distinct topic because of the meaningful patterns captured. The Social Services and Public Welfare topic scores poorly with 0.2867, meaning it contains a lot of noise. 4.1 This makes sense considering the LDA model will find many bills that do not fit with a topic, for instance because the summary is very short (around 10 words in total) or because there are no defining words to be found that have not necessitated removal in preprocessing. Because LDA force-fits all inputs, even the ones that don't fit well in any topic, this usually results in one noisy topic. [20] Therefore the results following further research on this topic will not accurately describe differences in polarization on Social Services and Public Welfare, but rather polarization differences in a more general sense due to the prevalence of noise in this data.

4.1 Inferring legislators' ideology using W-NOMINATE

In our one-dimensional W-NOMINATE method, legislators are positioned on a single ideological spectrum. Here we attempt to visualize if a clear distinction between democrats and republicans is visible in a congress. 4.3 Because we have already divided bills into topics, most polarization is assumed to belong only to between-party differences. Below is a histogram showing two W-NOMINATE outputs of the topic legislation and policy, one belonging to Congress 93 and the other to Congress 117. We can see that in Congress 93, it is impossible to create a clear distinction between Democrats and Republicans. In contrast, in Congress 117 the ideological difference between the two parties is almost total.

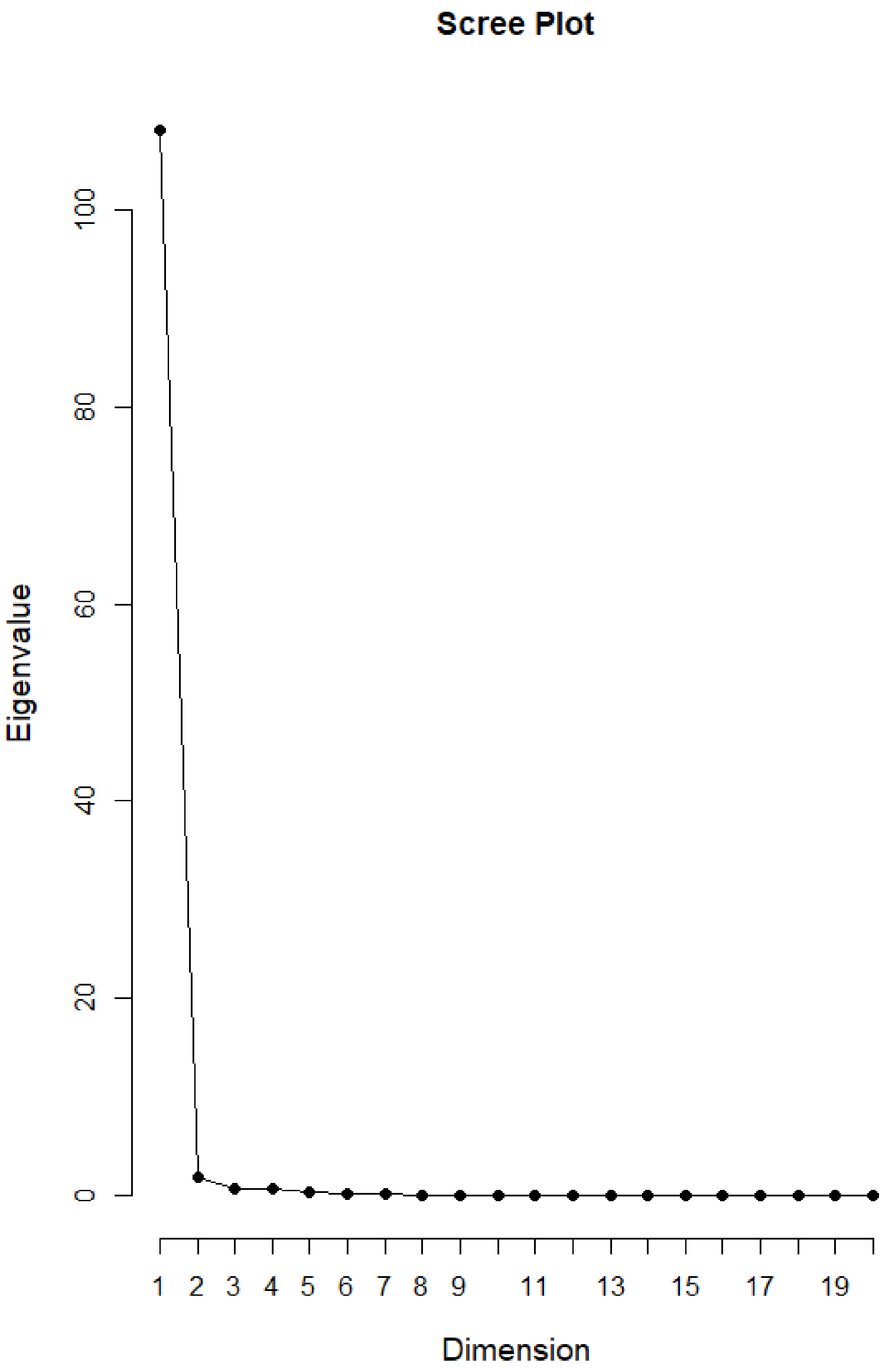
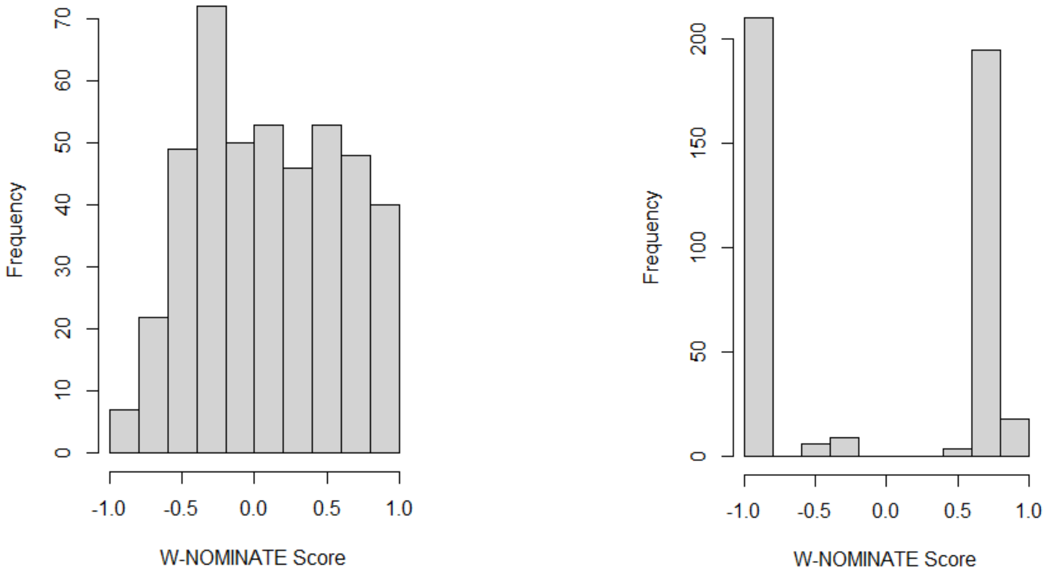


Figure 4.2: W-NOMINATE output can be explained by one dimension. The Scree plot generated for a W-NOMINATE model output (in this case Legislation and Policy, Congress 117 shows an eigenvalue over 100 for a single dimension with higher dimensions having little to no influence on the output.)



(a) W-NOMINATE output congress 93.

(b) W-NOMINATE output congress 117.

Figure 4.3: Output difference between Congress 93 and 117 in topic Legislation and Policy. We see that in Congress 93 there was little difference between Democrats and Republicans visible, with moderates for both parties having only a small ideological divide. In contrast, Congress 117 shows a near-complete disappearance of moderation for both parties, with an almost maximal ideological divide.

4.2 Detecting polarization using the bimodality coefficient with bootstrapping

Since the distribution of our research is always centered around 0, with Democrats between -1 and 0 and Republicans between 0 and 1, we assume bimodality values higher than 0.555 to denote a polarized Congress. We do this because even skewed unimodal distributions show polarized tendencies for one of the parties, therefore resulting in a polarized Topic-Congress combination. However, higher values for the bimodality coefficient do denote more extreme values of polarization as these values increase towards 1. For values lower than 0.555, we consider Topic-Congress combinations to be non-polarized. 4.4 To ensure our findings are robust, we used bootstrapping to create an error margin. If a Topic-Congress combination has a bimodality score of just above 0.555 and its bootstrapped error margin can bring this value below 0.555, we return it as non-polarized.

Our bimodality coefficients show that for the starting point of our research at Congress 93, all topics score below 0.5. 4.5 This implies clear unimodality in the data, which further indicates that at this point it is not simple to divide legislators into Democrats and Republicans based on their ideology. Differences between successive congresses might result in considerable differences in bimodality coefficient output, both in downward and upward directions. This indicates that changes in overall polarization have happened gradually instead of in specific congresses. For the bimodality scores graph of all topics, see Appendix C.

Focusing on specific topics, we determine their trends by fitting a regression line over their outputs. While every topic shows a certain upward trend, the slope varies. Some topics like Defense and Military trend to a coefficient value of around 0.6, which indicates slight polarization 4.7. Topics like Government Budget and Administration trend upwards to almost 0.7, showing more extreme polarization. 4.8 For the regression lines of all topics, please see Appendix C.

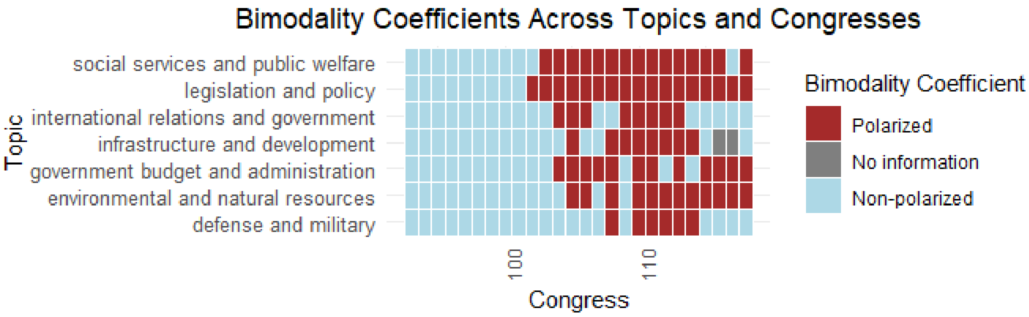


Figure 4.4: bimodality visualization of all topics. We see mostly non-polarized Topic-Congress combinations for Congresses 93 to 103 and mostly polarized combinations for Congresses 104 to 118. This shows the basic change in bimodal nature that occurs in the data. Also notable is that some topics are shown to be less polarized overall, such as defense and military.

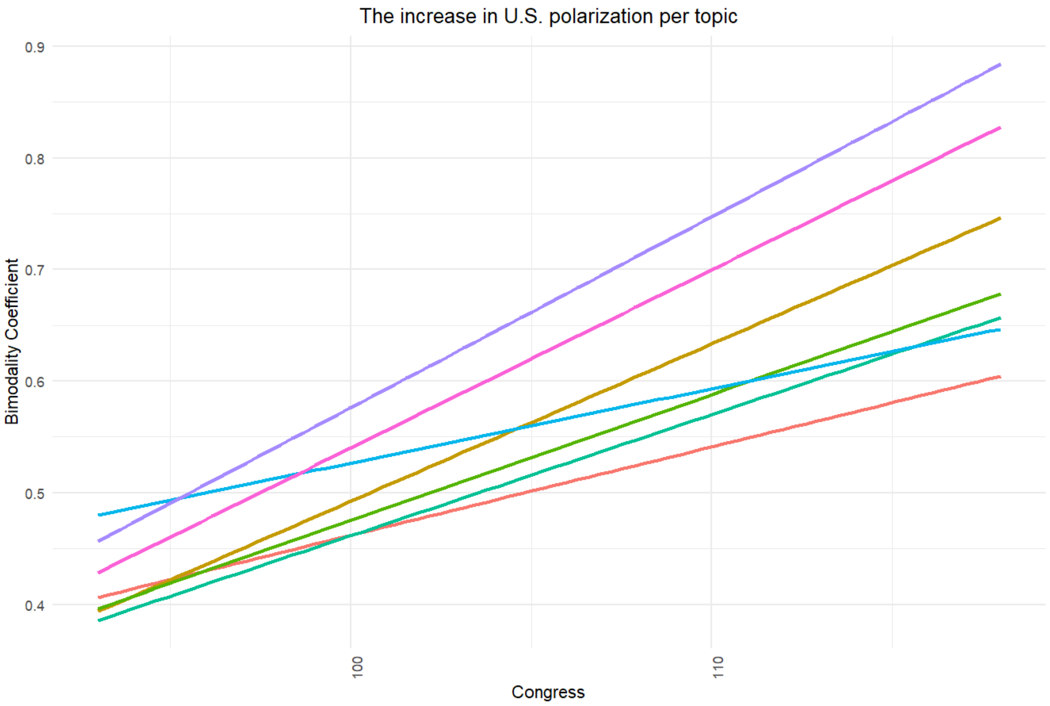


Figure 4.5: bimodality regressions of all topics. We see the changing nature of bimodality for every topic across the time frame. Also notable are the extreme values of bimodality that are reached in later Congresses and the overall positive trend every topic experiences over time. All topics are non-polarized in Congress 93 and are lightly to extremely polarized in congress 118.

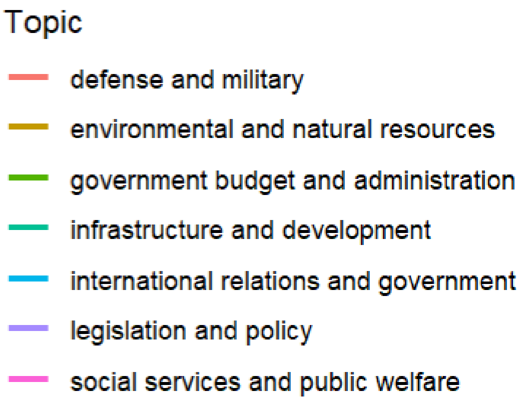

Topic
defense and military
environmental and natural resources
government budget and administration
infrastructure and development
international relations and government
legislation and policy
social services and public welfare

Figure 4.6: labels for the regression lines above

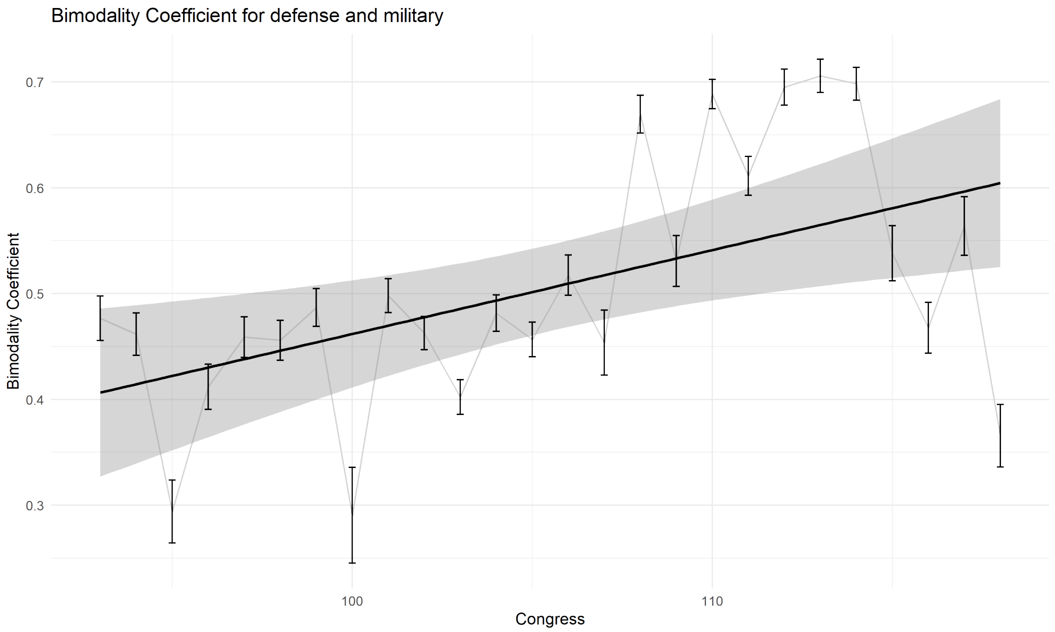


Figure 4.7: The regression over all bimodality coefficients belonging to the Defense and Military topic. We see a slight increase overall, mostly due to the peaks at Congresses 113 to 115. Bimodality reaches about 0.6 at Congress 118, indicating a lightly polarized political theme in the present day.

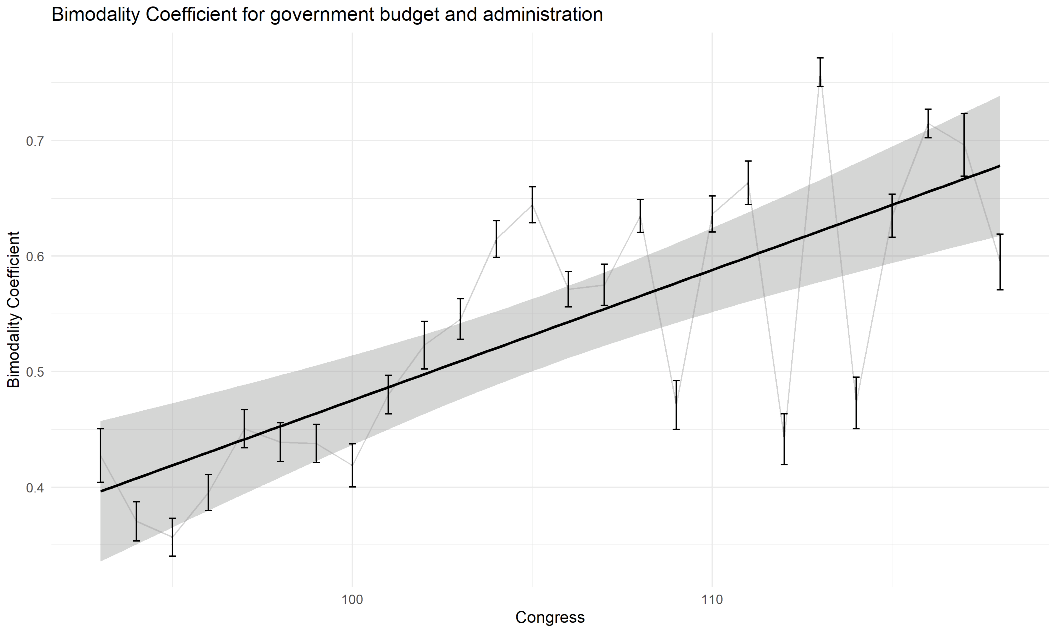
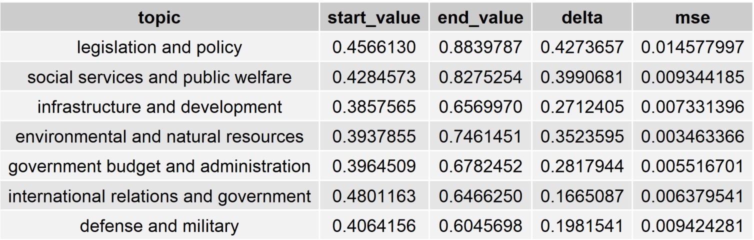


Figure 4.8: The regression over all bimodality coefficients belonging to the Government Budget and Administration topic. We see a higher increase than in the Defense and Military topic, mostly because this topic has more Congresses returning high bimodality coefficient values and experiencing polarization in an earlier stage of our time frame.



topic	start_value	end_value	delta	mse
legislation and policy	0.4566130	0.8839787	0.4273657	0.014577997
social services and public welfare	0.4284573	0.8275254	0.3990681	0.009344185
infrastructure and development	0.3857565	0.6569970	0.2712405	0.007331396
environmental and natural resources	0.3937855	0.7461451	0.3523595	0.003463366
government budget and administration	0.3964509	0.6782452	0.2817944	0.005516701
international relations and government	0.4801163	0.6466250	0.1665087	0.006379541
defense and military	0.4064156	0.6045698	0.1981541	0.009424281

Figure 4.9: linear regression statistics with MSE. Some topics, like legislation and policy, show a big increase in bimodality coefficient values. Here we also see the largest change over our timeframe, denoted in the delta column. International relations and government, as well as defense and military, are both far less polarized and also have the smallest deltas. Every topic is denoted as polarized at the end of our time frame.

5. Discussion

This research project aimed to define political polarization through the bimodality coefficient scores of W-NOMINATE output and examine its changes for different topics within the House of Representatives in a temporal framework ranging from Congress 93 in 1973 until Congress 118 in the present day. We use LDA to define topics, W-NOMINATE to estimate legislators' positions, followed by the bimodal coefficient with bootstrapping to represent this output as a single value that denotes to what extent the output has two distinct peaks. Through this method we were able to retrieve and visualize polarization trends for multiple topics over a period of more than 50 years.

Throughout this research we have defined political polarization in the U.S. House of Representatives as the measure in which the Democrats and Republicans have a visible ideological division on a distinct topic. Through the use of our W-NOMINATE scores we have quantified this division on a one-dimensional ideological spectrum. We then used the bimodality coefficient to simplify this output into a single measure to estimate to what degree the placement of legislators on this spectrum was bimodal and showed an easily distinguishable division between the two parties and used bootstrapping to ensure Topic-Congress combinations were denoted as polarized in a robust manner.

Our analysis allows us to draw some conclusions about the changes in polarization that became apparent over this time frame:

- There is a general increase in political polarization as measured in the time frame between 1973 and 2024.

  We can make this conclusion because we see no trends of decreasing polarization on the topic-specific level while our measured increases all trend towards a level of bimodality much higher than

0.555, allowing us to claim an increase in ideological differences between the parties over time. Furthermore, we have stated that one of our created topics, called Social Services and Public Welfare contains a lot of noise because of the way our LDA model obtains its outputs. Because of this noise, it is unwise to draw strong conclusions from this trend in a topic-specific context, but it can still be looked at in the sense of general polarization with our other topics removed from the overall trend. This trend is positive starting at 0.429 and reaching a bimodality coefficient value of 0.827, meaning we see a clear increase in polarization here.

- Topic-specific trends generally indicate an increase in polarization by varying degrees in the time frame between 1973 and 2024.

  We reach this conclusion because some topics like Defense and Military do not show strong political polarization (having a final bimodality coefficient of 0.605), yet 5 out of 7 topics reach bimodality scores of at least 0.650, and 3 topics go higher than 0.740. The regression lines all show an increase over time. Further considering the unimodal nature of every topic in 1973, we conclude that considerable change has occurred on the topic-specific level throughout our time frame.

- The changes in polarization we measure must be seen as gradual as opposed to sudden.

  Our W-NOMINATE and bimodality outputs show significant shocks between successive Congresses for every topic. This can be because a Congress suddenly has a couple of highly polarizing bills that influence the entire Congress or unpopular bills being brought by the party that has a House majority, causing larger distances between the two. For instance, Congress 114 has the largest Republican majority in our data with 251 legislators, and here 6 out of 7 topics see an increase in their bimodality coefficients. Yet, a following Congress can bring down the bimodality coefficient output just as quickly. We must then conclude that the significant changes measured over the entire temporal scale should be seen as a gradual change instead of a difference created by a large bimodality coefficient change between two specific Congresses.

- Temporal patterns are visible within the data.

  There are multiple instances where bimodality coefficients remain high or low for multiple congresses in a row. For instance, Defense and Military only reached bimodality coefficient values of around 0.7 between Congresses 113 and 115, which coincides with the Bush administration and its response to the 9/11 attacks (at which time the International Relations and Government topic also experienced a sharp increase). This proves that within the data, patterns can be found that reflect certain political dynamics beyond the overall change in polarization.

5.1 Conclusion

The implications of this research are profound. Through our definition of polarization using the bimodality coefficient, we have a strong basis to make the claim that polarization has increased by a sizeable margin in the timeframe used in this research. This growing ideological division between the two ruling parties in the U.S. can possibly lead to a legislative gridlock in one of the most influential governmental bodies in the world. It has the potential to hinder balanced legislation in the House of Representatives, especially in the field of legislation and policymaking. The differences found in this research highlight the varying degrees of polarization on different topics that showcase divisiveness that is growing in the highest legislative body found in the U.S.

From a purely methodological point of view, we can consider the use of LDA topic modeling in conjunction with W-NOMINATE scores and the bimodality coefficient with bootstrapping to be a success. It has proved to be a suitable method to analyze polarization differences within our selected timeframe, suggesting it can also be applied to other timeframes and legislative bodies if required.

That being said, there are also limitations to the method in its current form. For instance, while our W-NOMINATE model accurately places legislators based on their voting behavior, the model does make the assumption that all votes are made without any outside pressure. However, as stated earlier, such pressure does seem to exist quite substantially since we can spot occasions of extra polarization in moments where public opinion played an important role. Any future research could improve here by taking public opinion and lobbying during key moments in U.S. political history into account. This can be done through the addition of full legislative texts outside of just bill summaries as was utilized in this research or through the addition of extra W-NOMINATE dimensions where such a measure is beneficial. Furthermore, extending the timeframe used or checking for factionalism within the Democratic and Republican parties can bring more insight into political behavior over time.

To conclude, this work provides a detailed analysis of the growth of political polarization within the U.S. House of Representatives in the period of 1973 until the present day, or Congress 93 to 118. Defining polarization as the extent to which the two ruling parties are noticeably distinct, we have shown the trend of growing polarization over multiple different topics with differing magnitudes, highlighting a growing ideological division between the two. Addressing or even reversing such trends will require large-scale institutional reforms and societal change. Through such measures, it is doable to counter this growth in polarization and work towards a more effective democratic society.

For all data and code used in this research, please visit `https://github.com/larsteuni/master_thesis`.

A. Appendix A

Voter Name	Total Votes
YOUNG, Donald Edwin	27,217
CONYERS, John, Jr.	25,326
RANGEL, Charles B.	24,705

Table A.1: Top three most active voters.

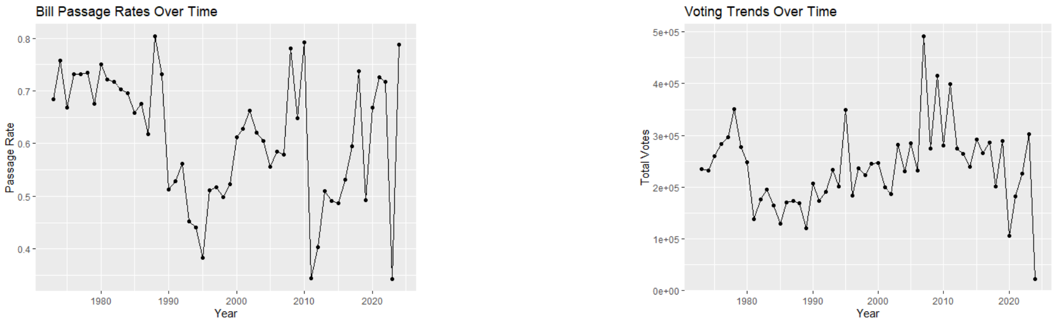
Sponsor Name	Total Bills	Bills Passed
Whitten, Jamie L.	375	268
Sessions, Pete	189	183
Bolling, Richard	173	152

Table A.2: Top three most passed bills sponsors.

Bill ID	Yea Votes	Nay Votes
b_583	162	162
b_613	204	204
b_1722	200	200

Table A.3: Three of the most controversial bills.

Some detailed examples we found in the bills and votes datasets, showing the activity of its most active legislators and some of its most divisive bills. Following are plots that show the percentage of bills that pass per year and the number of votes per year. Interesting is that some years feature exceptionally low passage rates, such as 1995, 2011, and 2023, even though the number of votes in those years remains stable.



(a) Bill passage rates over time.

(b) Total votes over time

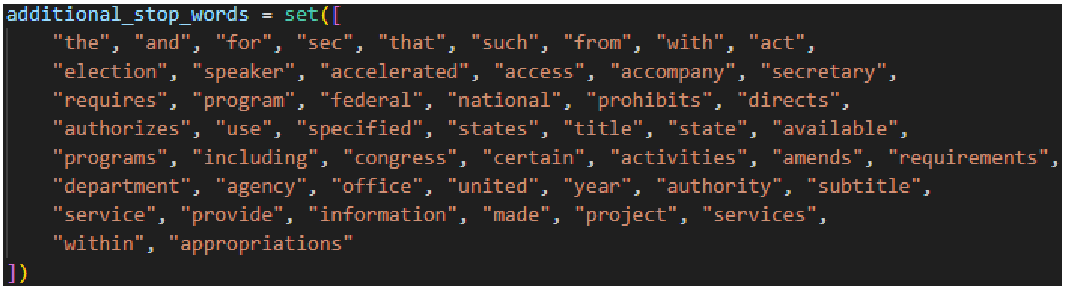
Figure A.1: bills and voting trends between Congresses 93 and 118. Bill passage rates show large-scale change between Congresses, with notable low points visible in the graph. Voting trends are stable throughout our time frame, with the exception of the current Congress which has not yet ended, explaining the lower amount of votes.

B. Appendix B

There are multiple reasons to decide on seven topics for our LDA output. Firstly, as is visible in the graph B.2, seven topics represent a local optimum based on cosine similarity after running LDA on a subset of three congresses with differing amounts of topic allocations. While having few topics can be seen as an oversimplification of the data, this is partly dealt with by having broad themes for our topics and partly opposed by the acceptable and good coherence score outputs in the full LDA model.

Although a five-topic structure does represent a higher average cosine similarity, we chose for the seven-topic structure because it returns results for more political themes and because this amount of topics still shows itself to be robust in its performance. Furthermore, we wanted to find some balance between topic interpretability while also making them more granular. The seven-topic model does well in both regards.

Lastly, we have to take into account the requirements for running the W-NOMINATE model. Before the model runs, it removes any bills that do not meet its minimum voting requirements. This means that most W-NOMINATE inputs will become smaller when the model is called. When we combine that with running the model for every Congress and further divide it by our topics, we risk not having enough bills to do W-NOMINATE. We see this problem twice with our seven-topic structure, so having more topics would likely have more negative effects than positive ones.



```
additional_stop_words = set([
    "the", "and", "for", "sec", "that", "such", "from", "with", "act",
    "election", "speaker", "accelerated", "access", "accompany", "secretary",
    "requires", "program", "federal", "national", "prohibits", "directs",
    "authorizes", "use", "specified", "states", "title", "state", "available",
    "programs", "including", "congress", "certain", "activities", "amends", "requirements",
    "department", "agency", "office", "united", "year", "authority", "subtitle",
    "service", "provide", "information", "made", "project", "services",
    "within", "appropriations"
])
```

Figure B.1: influential but undescriptive words removed in preprocessing phase.

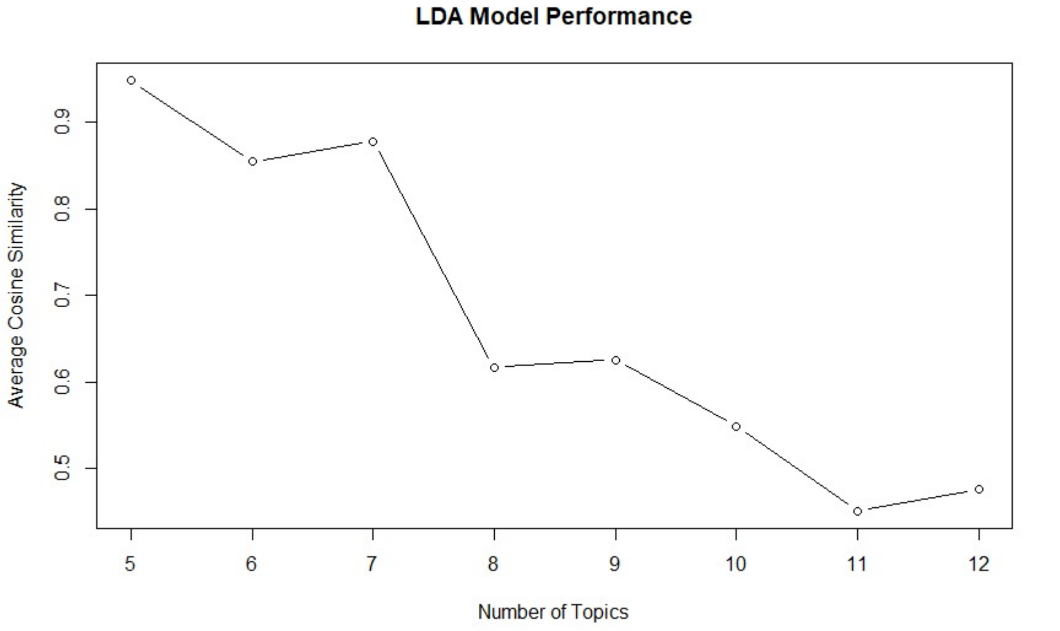


Figure B.2: 7 topics as a local optimum. We see a notable decrease in cosine similarity after seven topics. This, in combination with the lower cosine similarity for six topics, shows seven topics as a local optimum.

C. Appendix C

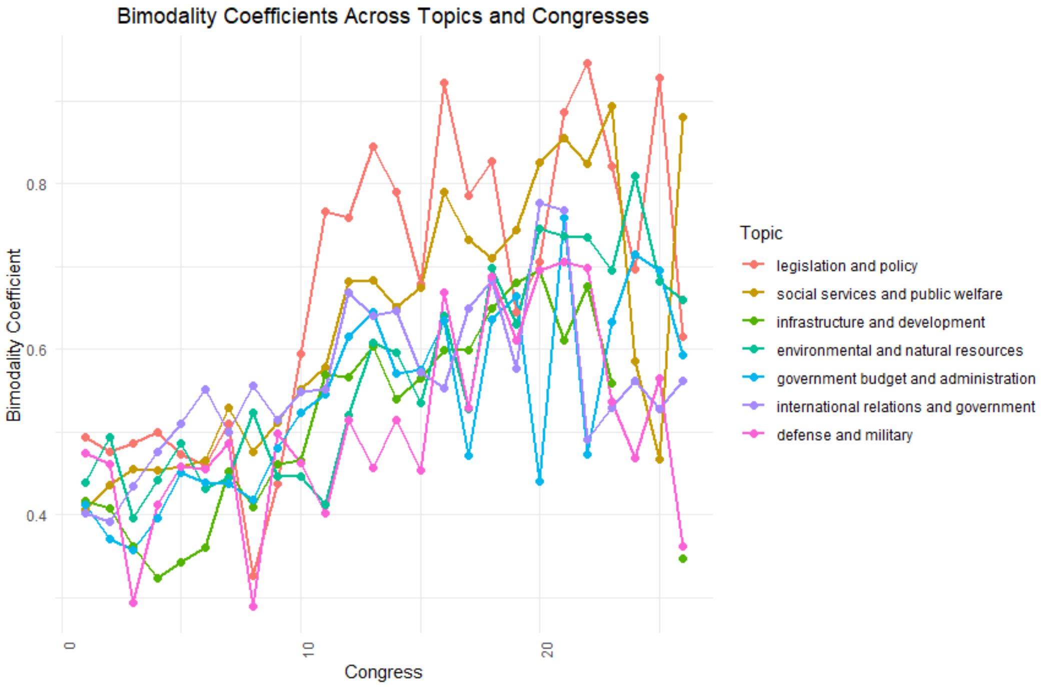


Figure C.1: Bimodality scores across all topics. This graph is without bootstrapped error margins for clarity. We see that all topics were non-polarized in 1973 and that their bimodality coefficients are between 0.4 and 0.5. In later Congresses topics show very different values from each other and are generally higher.

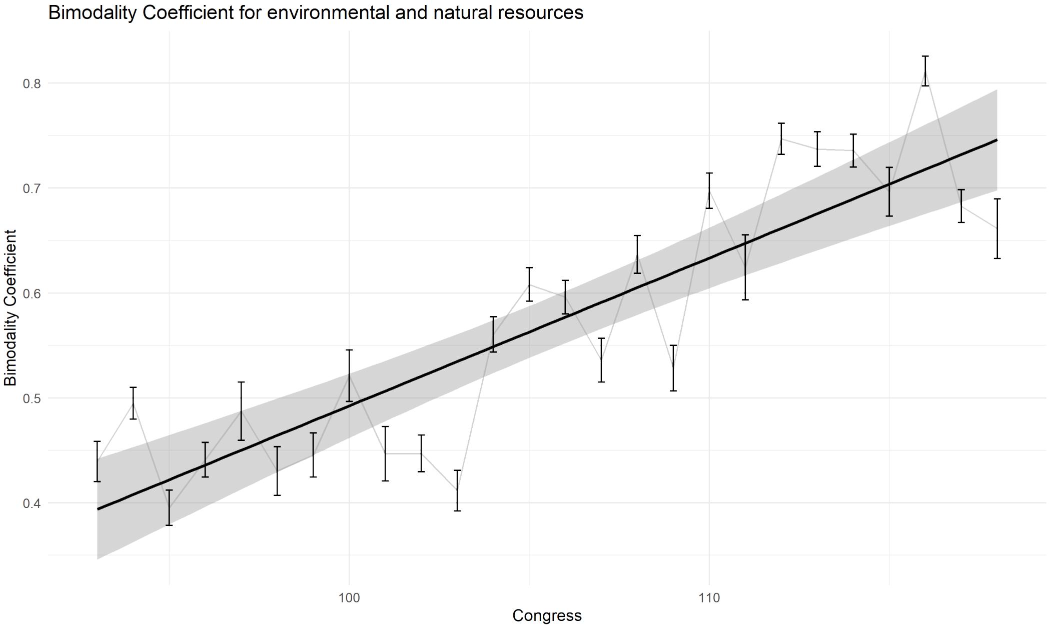


Figure C.2: The regression over all bimodality coefficients belonging to the Environmental and Natural Resources topic. The slope on this regression is steep, starting below 0.4 and ending just below 0.75. Bimodality coefficient scores align quite well with the regression.

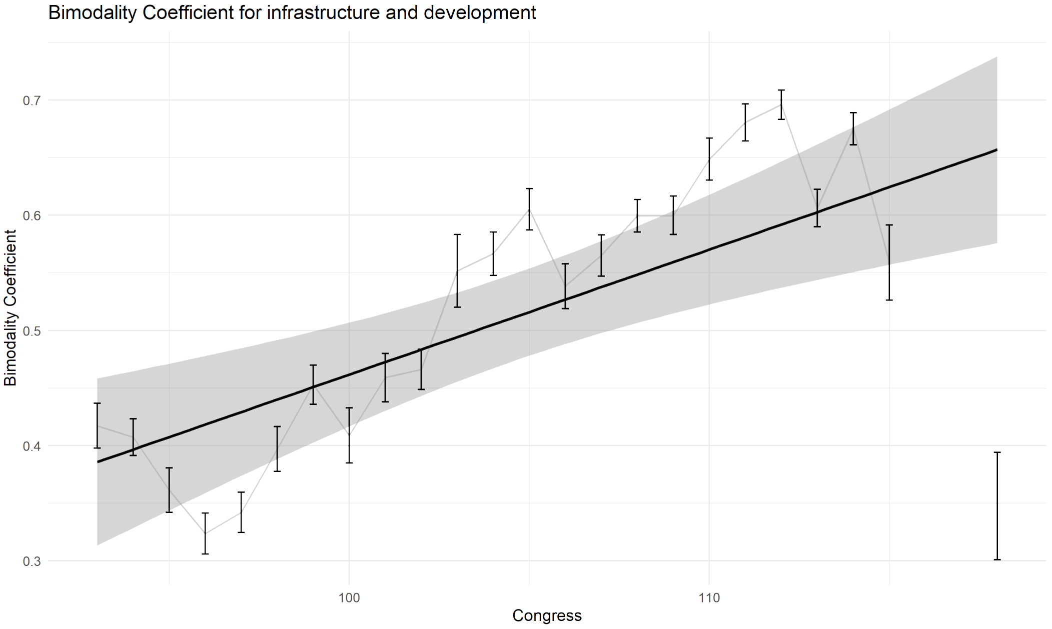


Figure C.3: The regression over all bimodality coefficients belonging to the Infrastructure and Development topic. Two values are missing since W-NOMINATE could not be run with the number of bills and votes for these Topic-Congress combinations. However, the regression line aligns well with the trend between Congress 93 and 115.

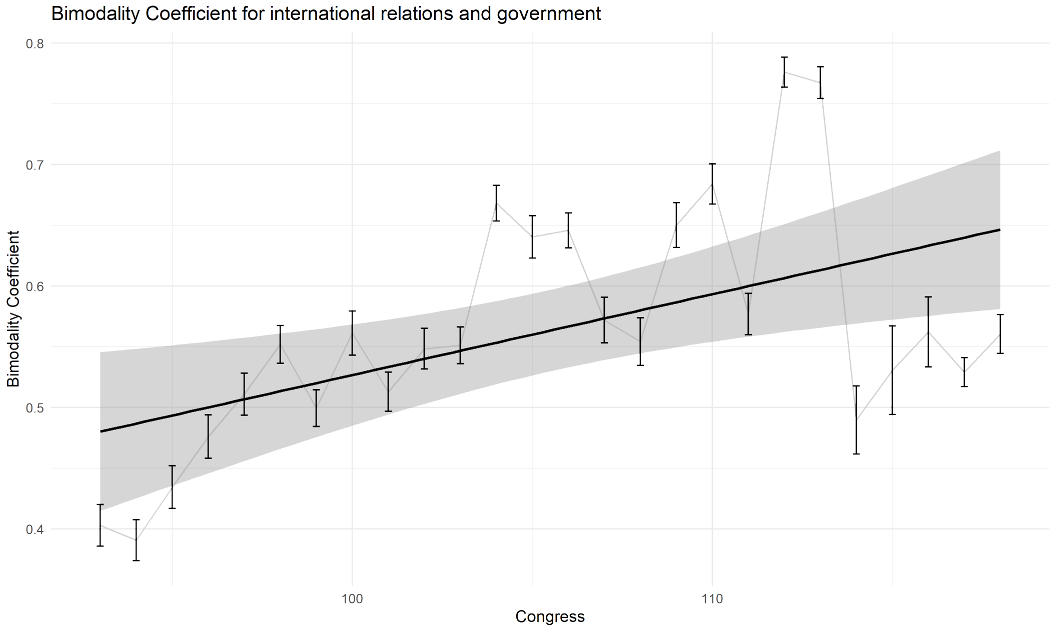


Figure C.4: The regression over all bimodality coefficients belonging to the International Relations and Government topic. The difference in bimodality coefficients between Congress 93 and Congress 118 shown on the regression line is small. The regression for this topic, like the other topics, is non-polarized at Congress 93 and polarized at Congress 118.

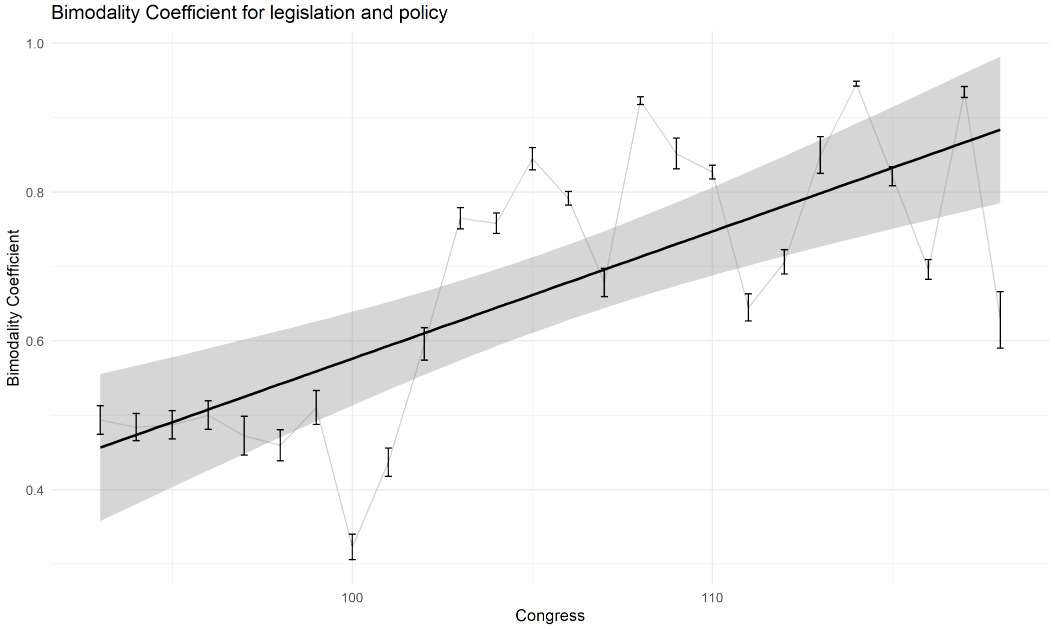


Figure C.5: The regression over all bimodality coefficients belonging to the Legislation and Policy topic. The slope of the regression line is steep, almost reaching a score of 0.9 at Congress 118. This shows extreme polarization.

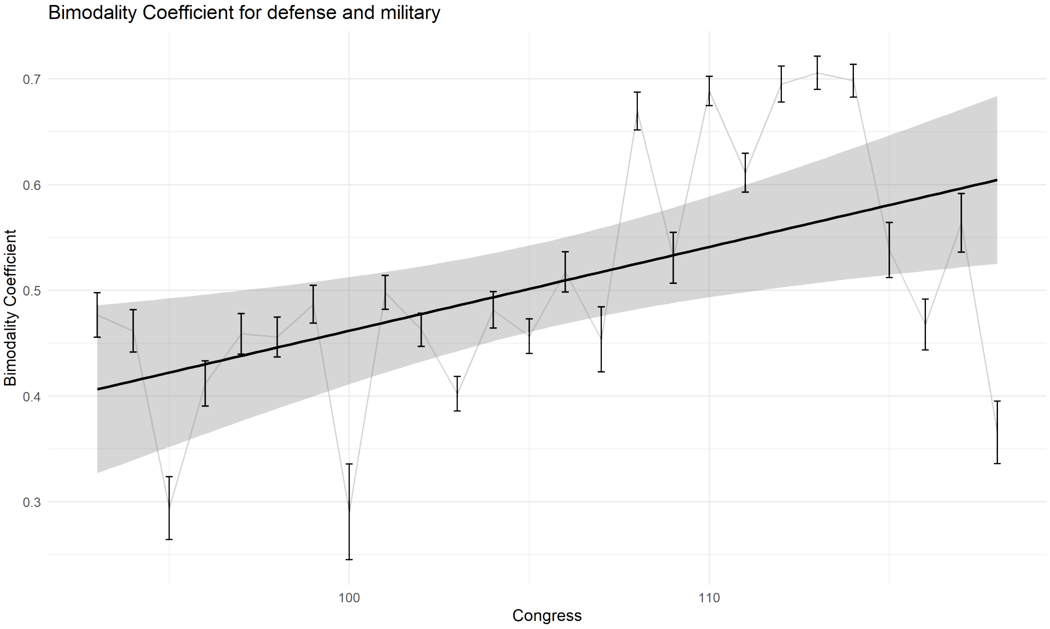


Figure C.6: The regression over all bimodality coefficients belonging to the Defense and Military topic. We see a slight increase overall, mostly due to the peaks at Congresses 113 to 115. Bimodality reaches about 0.6 at Congress 118, indicating a lightly polarized in the present day.

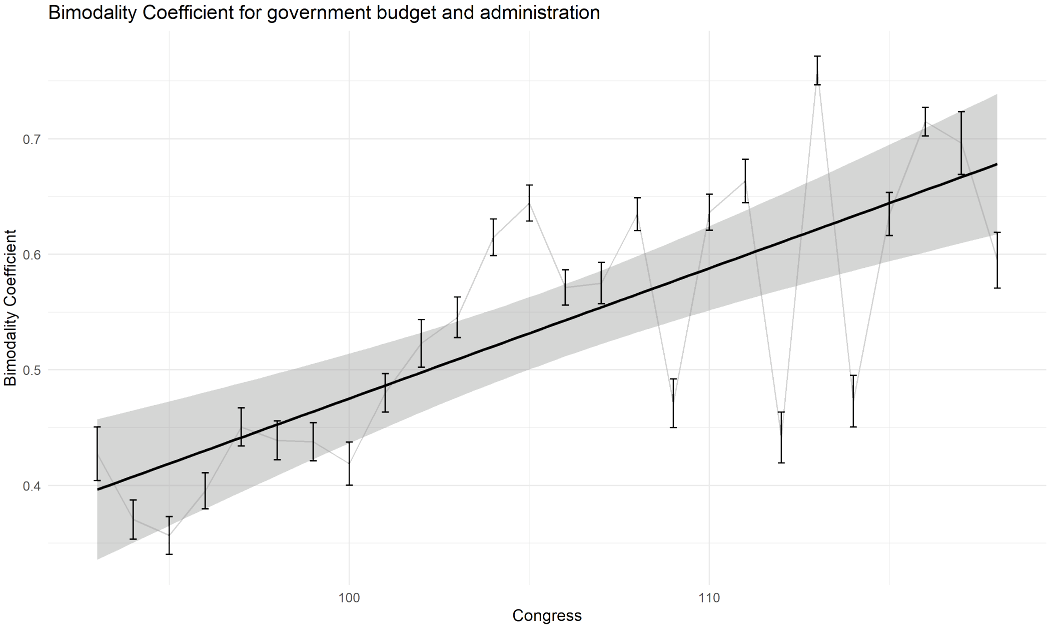


Figure C.7: The regression over all bimodality coefficients belonging to the Government Budget and Administration topic. We see a higher increase than in the Defense and Military topic, mostly because this topic has more Congresses returning high bimodality coefficient values and experiencing polarization in an earlier stage of our time frame.

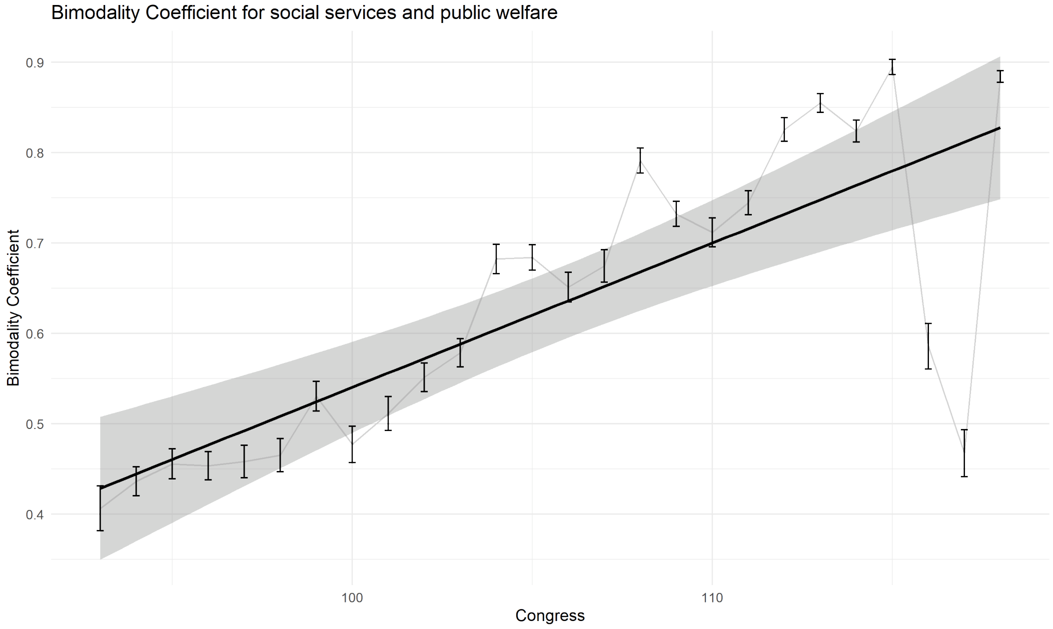


Figure C.8: The regression over all bimodality coefficients belonging to the Social Services and Public Welfare topic. The regression aligns well with the actual bimodality values. Ending at around 0.825, we see high polarization for this topic in the present day.

Bibliography

[1] A. Schedler, “Rethinking political polarization,” *Political Science Quarterly*, vol. 138, no. 3, pp. 335–359, 2023, ISSN: 1538-165X. DOI: 10.1093/psquar/qqad038. [Online]. Available: http://dx.doi.org/10.1093/psquar/qqad038.

[2] N. McCarty, K. T. Poole, and H. Rosenthal, *Polarized America: The dance of ideology and unequal riches*. mit Press, 2016.

[3] K. T. Poole and H. Rosenthal, “A spatial model for legislative roll call analysis,” *American journal of political science*, pp. 357–384, 1985.

[4] K. T. Poole, J. B. Lewis, J. Lo, and R. Carroll, “Scaling roll call votes with w-nominate in r,” *Available at SSRN 1276082*, 2008.

[5] P. Kherwa and P. Bansal, “Topic modeling: A comprehensive review,” *EAI Endorsed transactions on scalable information systems*, vol. 7, no. 24, 2019.

[6] E. Voeten, “Clashes in the assembly,” *International Organization*, vol. 54, no. 2, pp. 185–215, 2000.

[7] S. Hix, A. Noury, and G. Roland, “Dimensions of politics in the european parliament,” *American Journal of Political Science*, vol. 49, no. 2, pp. 494–511, 2005.

[8] E. Dimson, O. Karakaş, and X. Li, “Active ownership,” *The Review of Financial Studies*, vol. 28, no. 12, pp. 3225–3268, 2015.

[9] J. B. Lewis, K. T. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet, *Voteview: Congressional roll-call votes database*, 2023. [Online]. Available: https://voteview.com/.

[10] ProPublica, *Propublica: Congress api*, Accessed: 2024-06-20, 2023. [Online]. Available: https://projects.propublica.org/api-docs/congress-api/.

[11] H. Jelodar, Y. Wang, C. Yuan, *et al.*, “Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey,” *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, Nov. 2018, ISSN: 1573-7721. DOI: 10.1007/s11042-018-6894-4. [Online]. Available: http://dx.doi.org/10.1007/s11042-018-6894-4.

[12] S.-W. Kim and J.-M. Gil, “Research paper classification systems based on TF-IDF and LDA schemes,” en, *Hum.-centric Comput. Inf. Sci.*, vol. 9, no. 1, Dec. 2019.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[14] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora*, In Proceedings of the LREC 2010 Workshop on New

Challenges for NLP Frameworks, https://radimrehurek.com/gensim, Valletta, Malta, May 2010.

[15] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[16] R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. Freeman, "Good things peak in pairs: A note on the bimodality coefficient," *Frontiers in psychology*, vol. 4, p. 67 498, 2013.

[17] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: CRC Press, 1994, ISBN: 978-0412042317. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9780429246593/introduction-bootstrap-bradley-efron-robert-tibshirani.

[18] S. Syed and M. Spruit, "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation," pp. 165–174, 2017.

[19] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," pp. 952–961, 2012.

[20] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of topic models," *Foundations and Trends® in Information Retrieval*, vol. 11, no. 2-3, pp. 143–296, 2017. DOI: 10.1561/1500000030.