Duke University
Samuel DuBois Cook Center on Social Equity at Duke University
FALL 2023

Seminar: Global Inequality Research
Lecturers: Ph.D Elizabeth Degefe
Ph.D. Quran Karriem

Term Paper Draft

Debating Redistribution in Times of Economic Crisis

An Examining of U.S. Budgetary Debates using Structural Topic Modeling

Authors:

Xinyuan Li
xinyuan.li@duke.edu
Liberal Arts (Master)
1st Year

Philipp Weisenburger
philipp.weisenburger@duke.edu
Political Science (Master)
Exchange Student

Word Count: 5,822 | Page Count: 20
Submission Date: December 05th, 2023

Contents

Abstract:

Using Structured Topic Modeling we analyze speeches in debates on the federal budget in the U.S. House of Representatives from 2002 to 2011 to investigate how economic crises affect redistribution during budgetary debates. The difference-in-difference design allows us to examine theoretical ambiguities in existing interdisciplinary literature about what shapes the role of redistribution in fiscal policy. We find redistribution related topics to be crowded-out by economic crisis. At the same time Democratic Representatives appear to be more engaged with redistribution, while their Republican colleagues neglect the topic. The responsiveness of Representatives to the economic situation of their constituency could not be tested due to the limited scope of this research. However, we showed that the applied machine learning algorithm provides a powerful tool to explore this topic.

1. Introduction

"The public finances are one of the best starting points for an investigation of society, especially though not exclusively of its political life." (Schumpeter 1991, 101)

Public budgets reallocate resources within an economy. Both the spending and revenue side provide powerful but arbitrary tools, that can either increase or lower existing inequalities among societal groups. This is especially the case during economic crisis. With exacerbating inequalities and a declining economic output, the need for redistribution becomes more dire while the redistributive potential declines. Fiscal debates then more than ever take the monetary form of moral and societal debates on *who gets what*.

Our research focuses on the spending side and tries to identify what shapes demand for redistribution in the budgeting process. Redistribution in fiscal policy is subject of multiple social sciences and subfields. Hence, various competing arguments have emerged on the driving forces of changes in the demand for redistribution in public budgets. Conflicting economic theories describe redistribution as a deduction of economic efficiency (Okun 2015), while others see it as a tool to stimulate growth during *times of crisis* (Keynes 1939). The political preferences of elected officials similarly influence their redistributive preferences which can be shaped by *partisanship* (Buchanan 1949). Likewise, does the *responsiveness* to their constituencies (Gilens 2005; Elsässer/Haffert 2022) and therefore the economic situation of their district play a role in the budgeting process. To the best of our knowledge, these competing explanations have not yet been subject to a comparative empirical test.

We apply a difference-in-difference research design to answer the following question: *How do economic crises affect redistribution in budgetary debates?* Given the existing theories, we expect to see speakers addressing redistribution related topics differently given the macro-economic circumstances (H_1), their party affiliation (H_2), and the average

wealth of their constituency (H_3). Our findings will allow us to evaluate the theoretical arguments and examine their empirical validity.

By introducing an innovative methodology to this problem, we hope to close the gap that currently exists in the literature. Structural Topic Modeling will be applied to utilize transcripts of debates on the federal budget in the United States House of Representatives. We focus on a ten-year time span around the 2008 financial crisis. Using a text-as-data approach we deviate from the mainstream of public economics (Weber 2023), allowing us to investigate the nexus between actual policy choices and their justification to the public.

The article will first provide a clustered overview of the existing literature on redistribution in fiscal policies and derive the hypotheses thereof. A second section will give an overview of the research design and offers a conceptional understanding of the applied methodology – structured topic modeling. Benefits and potential shortcomings of the chosen approach will be discussed. The following section justifies the case selection and depicts the data source. The findings of our analysis are discussed in the analysis section, followed by a conclusion and critical assessment of remaining uncertainties.

2. Theory

Many social sciences deal with public budgets. Accordingly, plenty of theories have emerged around the process of budgeting, its impact on the economy as well as its redistributive potential. The following section will outline the relevant theoretical foundation and therefrom derive the hypotheses to be tested in the conducted research project. The main objective is to situate this research in existing literature and an ongoing scientific debate about redistribution and public budgets in the context of economic crisis.

The reviewed literature included articles and chapters from Economics, Political Science, and Political Economy dealing with redistribution and public budgets. They have been clustered into three categories. Each category in itself displays a more or less heterogeneous subfield. From economic literature we focused on competing theories on

the effect of redistribution on economic output and efficiency. Closely related, Political Economy deals with the impact of institutions and ideologies. For feasibility constraints we focused on the latter. Lastly, responsiveness theory was identified in Political Science addressing how representatives are responding to the redistribution demands of their constituency.

2.1. Macroeconomic implications of Redistribution and the Role of Crisis

Redistribution through fiscal policy and public budgeting in particular has implications for the overall output and efficiency of an economy. Different economic schools of thought assess its impact differently, especially during economic recessions. This subsection will give an overview on how the two most prominent schools of economic thought – Neoclassic and Keynesianism – assess the role of redistribution differently and derive our first hypothesis thereof.

Economic literature is mainly concerned with the impact of public budgets on the overall performance of the economy in a country or currency zone. Two schools of thought provide diametrically opposing views on the role of redistributive public spending on the economic performance. Macroeconomic considerations inevitably shape the budgeting process, as they are considered a key indicator for the performance of a society and it is seen as a necessary condition for a prosperous society (Yu et al. 2019) – even though not sufficient in many cases as we will see.

The neo-classical doctrine assumes a trade-off between redistributive spending and efficiency. This thought is infamously captured in Arthur M. Okun's metaphor of redistribution as a leaky bucket (Okun 2015). He states that redistribution inevitably comes with a loss in efficiency. Hence, a society will only be able to redistribute as much until the rate of redistribution is matched or even surpassed by the rate of efficiency loss.

While Okun's metaphor could be banished into the realm of fables, supportive research has come out in the wake of the 2008 financial crisis that has provided empirical evidence

for the claim that redistributive spending would hamper or thwart economic recovery and growth.

Alesina/Ardanga 2010 examine the impact of tax cuts as opposed to increased spending on economic recovery. In investigating a cross-sectional dataset of OECD countries, they argue that tax cuts yield a higher multiplying effect on the economic output and thus are the more efficient fiscal stimulus. The case they make is grounded in the argument of supply-side economics, a concept developed mainly by the economist Milton Friedman (Stiglitz 2016). A crucial role is attributed to the future expectations of consumers. They echo the argument of Olivier Jean Blanchard who suggested that current fiscal tightening could "eliminate the need for larger, maybe much more disruptive adjustments in the future" (Blanchard 1990, 111). Consumers would see currently enacted measures as a soft alternative compared to more draconian measures in the future. "Consumers anticipate a permanent increase in their lifetime disposable income, and this may induce an increase in current private consumption and in aggregate demand" (Alesina/Ardanga 2010). Giavazzi/Pagano (1990) have found a similar effect in their case studies of Ireland and Denmark.

Similarly in the context of the Great Recession Reinhart and Rogoff (2010) find an overall negative economic impact of increased public spending in times of crisis and an already inflated national debt. In their cross-sectional time-series analysis they make the point that saving and thus a reduction in social spending is required to overcome economic crisis. Joseph Stiglitz refers to these approaches as "expansionary austerity" (Stiglitz 2016, 45).

This notion is challenged by Keynesian or Neo-Keynesian economists. Herndon et al. criticize the findings by Reinhart/Rogoff stating, "selective exclusion of available data, coding errors and unconventional weighting of summary statistics" (Herdon et al. 2013, 2). Leading Dean Baker to speak of "The Myth of Expansionary Fiscal Austerity" (Baker

2010). Reifschneider et al. find a negative impact of austerity not only on current but also on future economic performance (Reifschneider et al. 2013).

In contrast to the neo-classical school of thought, Keynesian economists see public spending and especially redistributive spending as a powerful tool for state intervention in times of economic crisis. Keynes, being introduced as the eponym, can be quoted saying: "The boom, not the slump, is the right time for austerity at the Treasury." Keynes 1937, 390)

Accordingly, Stiglitz points to inequality as a root cause of the Great Recession. As "those at the top spend a smaller percentage of their income than the rest" (Stiglitz 2016, 44) the aggregate demand can be increased by redistribution from the top to the bottom.

Besides all differences the two economic schools of thought focus on the extension or the reduction of redistributive spending as a tool to overcome economic crises. Other objectives, such as the improvement of equality or the increase of the quality of life are subjected to the restauration of economic growth. We therefore expect:

> H_1: Redistribution related topics are mentioned more often during times of economic crisis as opposed to times of economic prosperity.

2.2. Elected Officials' Preferences

The following section will review academic literature on the political partisanship around redistribution in public budgeting. It will show the implications of redistribution for the political arena. This review is necessary to derive a hypothesis on how the party affiliation of a speaker interacts with their mentioning of redistribution related topics.

Alesina and Ardagna state that the two competing economic schools of thought have aligned with partisan politics in the United States, with Republicans favoring tax cuts and cuts to spending as opposed to democrats favoring progressive taxation and expansive fiscal policies (Alesina/Ardagna 2010, 36).

Relying on Paul Peterson's typology of public spending (Peterson 1981), Yu et al. analyze the budget trade-offs and the role of partisanship (Yu et al. 2019). Peterson distinguishes between spending in four categories: Development, Allocation, Redistribution and Education (Peterson 1981). Yu et al. examine state budgets over time and find partisan labels together with institutional limitations to play an essential role in shaping the spending choices (Yu et al. 2019, 254). More detailed, they find the Democratic Party to be in general more supportive of redistributive spending than the Republican Party. During budgetary debates, we thus hypothesize:

H_2: Democratic Politicians mention redistribution related topics more often, while Republicans avoid the topic.

2.3. Responsiveness

The ability of politicians to make choices first and foremost depends on them being elected and gaining a majority as a party. The democratic paradigm therefore assumes accountability of elected officials towards their electorate. Gilens' (2005) seminal work examined the differences in responsiveness to class-interests. He finds responsiveness to be highly dependent on class affiliation. Where policy preferences deviated for various income groups, more affluent citizens were more likely to see their preferences turned into policies. Gilens contrasts his findings with the democratic paradigm and is left with the paradox of a *de jure* democratically constituted society that *de facto* only represents the preferences of its top income members.

The latter part of Gilens paper tries to examine the causal mechanism of this correlation. He establishes three potential pathways: (1) Public preferences shape policy choices of elected officials; (2) Policy choices shape public preferences; (3) Both public preferences and policy choices are confounded by "real-world events". He rules out the second pathway arguing that many constituents do not even know their congress representatives. The third pathway remains under-investigated. Examining potential

differences during times of economic crisis vs. prosperity we may be able to test this third pathway to a limited degree. His findings are replicated by Elsässer and Haffert (2022) who find a similar phenomenon for Germany examining opinion polling on fiscal policy proposals from 1980 until 2016. We therefor hypothesize:

> H3: Representatives from poorer districts speak more about distributions. Representatives of more affluent districts tend to avoid the topic.

Finally, it is important to state that the explanatory approaches laid out above are not necessarily mutually exclusive. It can easily be thought of as an intersection of the varying theoretical assumptions in one case as opposed to another.

3. Methods

In the following section we will layout the chosen research design and applied methodology. We will justify why we relied on Structured Topic Modeling for the data analysis, give a brief overview of its operating principles, and discuss potential shortcomings and pitfalls as well as how we planned to deal with them. In a separate subsection we will explain the research design we used to investigate potential causal mechanisms. Again, we will justify why we chose Difference in Difference as a research design and provide a conceptual framework of DiD. Necessary assumptions and potential sources of errors and biases will be made transparent accordingly. We will discuss corresponding robustness checks.

3.1. Structured Topic Modeling

In order to analyze the nexus of technocratic policy making and the democratic paradigm, we turn to parliamentary debates on budgets. The analysis of large qualitative text datasets deviates from mainstream economics and political economy (Weber/Wasner 2023). Text data analysis is costly if it relies on human coding. Thus, researchers often times opt for approximations using quantifiable data (Roberts et al. 2014).

Roberts et al. present a statistical topic modelling approach that allows us to achieve multiple goals at the same time. First, by automating the supervision and coding of text data it dramatically reduces the implied cost of a qualitative approach. Larger datasets can be explored and the larger sample size in turn can lead to more robust findings. Second, their STM approach is considered "unsupervised" (Roberts et al. 2014, 3). This means that their method can infer content from the text data in form of clusters with minimal prior input by the researcher.[1] The potential sources of biases introduced by the researchers are thereby reduced, as they can never be eliminated totally.

Before briefly examining Structural Topic Modeling, a disclaimer is necessary. Being Sociology and Political Science students, none of us is a trained statistician or computer scientist. The understanding we have established in the course of this seminar remains conceptual and heavily reliant on peer-reviewed literature that we analyzed and tried to understand to the best of our knowledge and capabilities. Nonetheless, it remains a basic understanding.

Structural Topic Modeling reduces the complexity of text data by assigning a number to every word. The text is then treated as a vector. In order to find common themes, Structural Topic Modeling estimates the proximity of different words to form clusters. The analysis therefore needs to have a set number of kernels around which the density distribution can be explored by the algorithm. Roberts et al. distinguish between "single-membership models and mixed-membership models" (2014, 3). This can be understood as sampling with backfilling as opposed to without. The latter seems to be more applicable to our research interest since we do not rely on a sharp discrimination between different words but rather an accurate clustering. For example, if redistribution is a dominant word that forms many clusters with multiple variations of other words, this could be particularly of

[1] Setting the number of kernels that are to be found by the algorithm is one of the few inputs by the researchers.

interest. Context matters and by opting for a mixed-membership model we allow more ambiguity to enter the potential findings. The increased complexity of the findings may be seen as a trade-off. As the approach is after all a qualitative one and much of the complexity of the original text data has already been reduced using STM, this appears justified.

The STM algorithm starts with a "global prior distribution" (Roberts et al. 2014, 5). This enables the algorithm to infer proximity of words from a global training set. We can think of this as a basic familiarity with the language. Due to limitations in time and scope we relied on the model provided in the lecture. The *stm* package in R operates using a Dirichlet prior (Roberts et al. 2013).

The algorithm provides a number of word clusters that appear in greatest relative proximity and requires only a few assumptions. It thus is a "fast, transparent, replicable analysis" (Roberts et al. 2014, 4) that allows causal inference using text-as-data.

In a following step, the clusters will be reviewed by the researchers. Words that have no interpretational value will be hand coded as stop words, i.e. excluded from the analysis. This iterative process is repeated until no substantive improvement of the clusters is observed. The clusters are then interpreted by the researchers according to the words they contain.

3.2. Difference-in-Difference Design

The observed clusters will be analyzed using a Difference-in-Difference design, which Keele identifies as a "straightforward approach" (2020, 823) to estimate causal effects. In essence, the idea of this design is to compare two groups – one having received a treatment as opposed to the other which serves as a control. The two groups are observed before and after treatment. Hence, the trend of the control group can be interpreted as a counterfactual state of how the treatment group would have behaved had it not been treated. (Angrist/Pischke 2009).

In general, the average treatment effect on the treated is calculated as the net expected outcome of the treated subtracting the net treatment effect of the control group:

$$\mathbf{ATE = \delta = (E[Y^T|post] - E[Y^T|pre]) - (E[Y^C|post] - E[Y^C|pre])}$$

The difference estimation can be visualized by Hill et al. Here the deviation in slopes is considered the treatment effect. The counterfactual state (dashed line in the figure) can have a different intercept. In this way the design allows to compare differences in differences and does not rely on identical cases as opposed to twin studies or the synthetical construction of a control (Abadie et al. 2015).

APPENDIX 1 – FIGURE 1 (Hill et al. 2011)

The average treatment effect can also be estimated using a regression equation. It will estimate δ as the estimator of an interaction term of an indicator variable for the treatment group (1 = treat; 0 = control) and the time (1 = post treatment; 0 = prior to treatment):

$$\mathbf{\delta = \alpha + \beta_1 * TREAT + \beta_2 * POST + \gamma (TREAT*POST)}$$

The regression form will allow to add variables as controls and thereby enable the testing of the robustness of any causal effect found. The regression equation is plugged into the *estimateEffect()* function of the *stm* package in R.

In case no significant impact being found, the research project could also contribute to the debate on responsiveness theory. This outcome could indicate a neglect of redistribution in times of multiple crises.

4. Data

Investigating the impact of economic crisis, we pick a ten-year period around the 2008 financial crisis from 2002 to 2011. The Great Recession has been chosen as it triggered the biggest fiscal response in peace time (Alesina/Ardanga 2010). Since we wanted to focus on an economic crisis, we excluded the Covid-19 pandemic. Although the economic impact had been significant, the pandemic came as a multi-dimensional event that would have made a clear attribution of effects to a treatment less straightforward.

Net GDP growth was chosen as an indicator for economic crises. We are aware of the fact that economic crises can be multifaceted and need not manifest themselves in a decline in the Gross Domestic Product. It is after all only an approximation. As data source the GDP growth data by the IMF was used. Figure 2 shows the annual GDP growth rates:

APPENDIX 2 – FIGURE 2 (IMF 2021)

The main problem the outlined research project had to overcome was in fact a data problem. This may appear irritating at first glance since records of the United States House of Representatives and the Senate are publicly available and transcripts of debates are provided for almost all sessions. However, the problem arises when it comes to the cleanness of the data. In order to utilize structured topic modeling, the text data must have a specific form.

In the case of the data from Congress.gov the transcripts had not been separated by speaker. To distinguish them manually by copying and pasting surpassed the capacities of the research project. Using scraping and other techniques to clean the data into separate observations turned out to be more complex and equally time consuming. A first idea, to introduce cutting points before the mentioning of the name of the speaker was rendered impracticable due to the random naming of representatives in the speeches themselves.

For a preliminary examination, we focused on presidential speeches on passed budgets. The fallback came at a cost. As there is no ‘opposition president’ the DiD design was not fully applicable. Also, a sitting president does not represent a single state or district. H_3 therefore remained untestable.

The data problem was overcome by utilizing the Social Science Data Collection provided by Stanford Libraries[2]. It provides a clean version of all speeches from the 43rd to the 114th congress. We identified the relevant sessions using the digitalized archive of the Congressional Records[3]. The relevant speeches were then integrated into one database including the name of the speaker, party affiliation, gender, state and congressional district, year, and an indicator variable for economic crisis as metadata. By being able to use the Stanford data set we increased the number of observed speeches to 846.

5. Analysis

The following section will first review our analysis process and discuss the findings of our preliminary analysis using the speeches by U.S. presidents on the submitted presidential budget proposal as text-data. This had been our fallback option and will be revisited here briefly to assess the validity of the approach as well as the limitations of the preliminary data source critically. After that, the test for each hypothesis will be conducted in a separate subsection using speeches in the U.S. House of Representatives on the federal annual budget text-data source, as originally intended.

In the first step, all text-data has been subject to a data cleaning process. To ensure comparability all letters were converted to lower case, punctuation and paragraph signs had been removed and words have been stemmed as is best practice. Common stop words and short words with two or fewer letters have been removed since they do not hold informational value in most cases.

[2] Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. (https://data.stanford.edu/congress_text)

[3] Congressional Record by GovInfo.gov (https://www.govinfo.gov/app/collection/crec/)

As discussed above, the number of kernels (k) has to be set by the researchers. Thus, after a testing phase k had been set to 20, rendering 20 clustered topics that are displayed in Figure 3 (preliminary data set) and Figure 4 (final data set) ranked by frequency of appearance:

APPENDIX 3 – FIGURE 3 | APPENDIX 4 – FIGURE 4

The next step had been a closer look at the top ten categories that had been established by the structural topic modeling algorithm:

APPENDIX 5 – FIGURE 5[4] | APPENDIX 6 – FIGURE 6

In an inductive process the categories have been labeled by the researchers accordingly to increase the usability for the following comparisons. When comparing the findings of the preliminary data set to the final data set, a reduction of stop words can be observed. While the first one still contained stop words like “can”, “will”, “must”, “new”, the number of words that have no informational value for the research question has been drastically reduced. However, some stop words remain, especially in less frequent topic clusters (e.g., “day”, “year”, “people”). After several iterations of hand coding additional stop words, no substantive improvement was observed. Thus, a point of saturation had been reached.

Comparing the two Top Topics (Fig. 3 and Fig. 4) we further can see that the frequency between the topics shows greater variability. While the preliminary data shows most of the Topics clustered around 0.05, the final data shows a bigger range of frequencies covered. From topic 17 and 20 which are barely distinct from zero all the way up to Topics 5, 8, 19 and 18 closer to 0.10.

Contextwise, it must be noted that the number of technical terms that have entered the most frequent clusters has increased. This is unsurprising given that the presidential speeches from the preliminary data set are much shorter and are focused on the policy

[4] For an interpretation of the preliminary Top Topics, also see Appendix 5

lines. In the congressional data, funding bills are discussed in greater detail, which is why we should see an increase in terms such as “additional”, “appropriation”, “fund”, “earmark” (all words taken from Topic 10).

For the interpretative analysis of the topics, we focused on the ten most frequent topics using the frequency of 0.05 (median) as a cutoff point. The reduction appeared necessary given the requirements of the paper. It is to be noted however, that this restriction ruled out Topic 16, which contained the words “medicaid”, “care”, and “health” and Topic 11 and 17 both mentioning “job”. These clusters can be seen as redistribution related. Excluding them will make our research design more conservative and thereby increase the robustness of our findings.

The most frequent cluster (Topic 5) is focused on „Fiscal Policy” containing mainly words that focus on the legislative process. Topic 8 deals with “Technology regulations”. It mentions among other things the FCC – the Federal Communications Commission. Given the time period under study, the frequent mentioning of internet regulation appears plausible. For the interpretation, this category holds less value since regulatory policies often have only limited monetary impact and thus are less relevant for the public budget. Topic 19 again relates to more technical aspects of the budgeting process. It focuses on “Balanced Budget”, touching on the question of deficit spending and public debt. Topic 18 is more policy related. It centers around “Defense”. The topic is not as clear since it still contains plenty of words that are not unambiguously definable. “Budget technicalities” are the subject of Topic 10 as mentioned above, while Topics 2 and 14 have a redistributive focus. Topic 2 is focused on “Social Security and Medicare” and Topic 14 on “Education and Veterans”. The revenue side is discussed in Topic 13 focusing on “Corporate Taxation”. “Government shutdown” is the ninth most frequent cluster (Topic 4). Government shutdowns have not occurred during the period under study. However, the frequent mentioning of this topic can be seen as a precursor to the shutdown of 2013, marking a

break in the long period since 1996 without a technical illiquidity of the government. We tested the precursor hypothesis by plotting the distribution of Topic 4 over time. The finding very much approves our reasoning.

APPENDIX 7 – FIGURE 7

The last interpreted topic (Topic 7) is concerned with the "Energy sector".

The main part of the analysis is focused on the comparison of the identified topics using the assigned meta data, i.e. time of the speech, the party affiliation and congressional district of the speaker[5]. Before examining the individual hypotheses in detail, let us discuss two general differences between the findings for preliminary data set as opposed to the final data set.

First, we found the most frequent topics to be different. This is unsurprising, given that the data is completely different. Nonetheless, some similarities still stand out, such as the mentioning of Social Security in both data sets. On the other hand, the more technical focus of the Congressional data is expected as discussed above. Further research could utilize the findings and conduct a more extensive exclusion of stop words related to budget technicalities. Unfortunately, this remained beyond scope for this research project.

The second difference refers to the comparative analysis[6]. When regressing the topics identified in the Congressional data set for the metadata of the text-data, the number of statistically significant findings increases. This reflects the insufficient power of the preliminary data set using only ten presidential speeches (10 speeches in the preliminary vs. 846 in the final data set).

[5] For an interpretation of the preliminary findings in a difference-in-difference setup see Appendix 7.

[6] See for comparison Appendix 8 (preliminary data set) and Appendix 9 and 10 (final data set)

5.1. Testing Hypothesis 1: The Role of Economic Crisis

To test how economic crisis affects redistribution related topics in budgetary debates, we compared the ten most frequent topics in the years coded as crisis to those defined non-crisis. The findings are summarized in the following plot:

Appendix 9 – Figure 10

As discussed above, the spreads have widely increased compared to the model using the preliminary data set. In addition to the increased data set, we now have Republican and Democratic speakers for the same year, as opposed to only the incumbent President for each year. The reflective nature of a debate will therefore lead to an increase in the differences between the time periods. An example might help to illustrate this point: If a topic gets mentioned more often during times of crisis, a Presidential speech might not be mentioned during economic prosperity, while it is referred to once in a speech given during an economic crisis. If we transfer this into a debate setting, the topic mentioned once would be picked up by the opposing side and thereby counted at least twice.

"Technology regulations" and "Corporate taxation" remain insignificant. "Defense" is clearly identified as a non-crisis topic. In this case economic crisis appears to have a crowding out effect on defense related topics. In this particular time period, it could also be attributed to the 2001 "war on terror" which might have drawn the mentioning more into the pre-crisis period in our data set. Further research observing other time periods could help to close this ambiguity. Similarly, "Government shutdown" is mentioned less frequently during times of crisis. It has been identified as a precursor for the 2013 government shutdown above. This would point to the conclusion that the shutdown was less related to an economic necessity and more likely to be a phenomenon of partisan politics. Again, more research would be needed to investigate this specific issue. However, it is outside the scope of this research project. The "Energy sector" is clearly mentioned

more frequently during crisis times. This may be a result of the important role of the energy sector for an economy.

The identified redistribution related topics "Social Security and Medicare" and "Education and Veterans" are mentioned less frequently during times of crisis. This rejects our first hypothesis. We could conceptualize this as a crowding out effect. In our preliminary analysis the findings for redistribution related topics were insignificant. With the increased data set we could overcome this shortcoming. The less frequent mention can be interpreted as a sign of a more dominant neoclassical paradigm in U.S. fiscal politics.

In full disclosure we also want to mention the findings for "Balanced Budget" which had been identified as non-crisis topic and "Budget technicalities" which is more prevalent during economic downturns. Both findings are counterintuitive since we would expect technicalities to be overshadowed by policy terminologies of urgent crisis related topics. To the contrast, "Balanced Budgets" and the question of deficit spending should have been more likely to appear during crises. These inconclusive findings will have to be subject to further research.

5.2. Testing Hypothesis 2: Partisanship:

The bigger spreads observed with the test for the effect of economic crisis have narrowed down when regressing the topic frequency on party affiliation. We attribute this finding to the dialogical character of a debate in Congress. As mentioned before, topics mentioned by one side are more likely to be picked up by the other within the same session to respond and engage in a discussion. This being said, the findings whose confidence bands not crossing "party lines", i.e., the zero-line, can be more solidly interpreted as statistically significant findings.

Appendix 10 – Figure 11

Unsurprisingly, half of the identified clusters are rendered insignificant. Party affiliation does not seem to have an impact on the mentioning of "Balanced Budgets",

"Budget technicalities", "Technology regulations", the "Energy Sector" and "Education and Veterans". The latter we identified as a redistribution related topic. However, the redistribution related topic "Social Security and Medicate" is significantly more frequently mentioned by Democratic members of the U.S. House of Representatives, than their Republican colleagues. This is in line with Yu et al. (2019) and confirms our second hypothesis. The finding further appears to be robust as it can be found in our preliminary findings as well.

Conversely, Republican representatives appear to talk more about "Fiscal Policy" and "Corporate Taxation". The combination could be reflective of the supply-side economics ideology that is prevalent in the Republican Party (Bartlett 2016, Stiglitz 2016).

5.3. Testing Hypothesis 3: Responsiveness Theory

Despite having solved the data problem, our third hypothesis remained untestable. We had planned to approximate the affluency of a representative's congressional district by using the median income of the district. The US census could be used as a source here. For applicability reasons a transformation into a bivariate variable would have been necessary. Therefore, we would compare the median income to the national average.

However, this procedure would have requested to hand code this variable for the >800 speakers in our data set. It was beyond the scope of this research project but offers the ability for future research.

6. Conclusion

In the first section the presented paper examined and clustered the existing literature on redistribution in public budgeting. It therefore combined insights from different subfields of Economics and Political Science and derived three testable hypotheses thereof. The following section discussed the research design. A conceptual understanding of Structural Topic Modeling had been given as well as a brief description of a Difference-in-Difference design. Fortunately, the data problem was resolved allowing us to present and

compare our preliminary to our final finding and thereby subjecting them an additional robustness check.

The research project consistently rejected H_1 and thereby showed how redistribution related topics can be affected by a crowding-out effect during times of economic distress. This finding is particularly alarming since economic crises do not affect all members of society equally but rather tend to enforce preexisting inequalities.

At the same time redistribution turned out to be a partisan issue. The investigation of differences between the speeches of a Democratic and Republican Representatives revealed that the former are more likely to bring up redistributive topics. This finding confirmed similar differences in our preliminary examination of Presidential speeches.

The responsiveness of members of the U.S. House of Representatives to their constituency could not be tested due to limited time and resources. However, we were able to show the potential of applying unsupervised machine learning algorithms to the topic under study. Future research can build on our findings, for example subject them to a more qualitative examination. The data set we utilized provides further metadata that we have thus far not included in the analysis. Again, this holds potential for subsequent research.

Most importantly, a regional expansion of the research project is necessary. We focused on the United States due to our own language limitations and limited timely resources that did not allow to study several countries in the necessary depth. However, the application of this research design to world regions that are systematically understudied seems of enormous importance We call on future research to consider the potential of the design and methodology outlined and applied in this paper to broaden our understanding of what shapes demand for redistribution in fiscal policy during times of crises.

7. Literature

Abadie, Alberto; Diamond, Alexis; Hainmueller, Jens (2015): Comparative Politics and the Synthetic Control Method. In: *American Journal of Political Science* 59 (2), p. 495-510.

Angrist, Joshua D.; Pischke, Jörn-Steffen (2009): *Mostly Harmless Econometrics. An Empiricist´s Companion*. Princeton (USA): Princeton University Press.

Baker, Dean (2010): The Myth of Expansionary Fiscal Austerity. Center for Eocnomic and Policy Research Issue Brief. Online at: https://www.cepr.net/documents/publications/austerity-myth-2010-10.pdf

Bartlett, Bruce (2016): The Rise and Fall of Supply-Side Economics. In: *The New American Economy.* Online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2840495

Broadbent B.; Daly, K. (2010) Limiting the Fall-out From Fiscal Adjustment. In: *Goldman Sachs, Global Economics Paper* 195, April 2010.

Buchanan, James M. (1949): The Pure Theory of Government Finance: A Suggested Approach. In: *Journal of Political Economy* 57 (6), p. 496-505.

Elsässer, Lea; Haffert, Lukas (2022): Does fiscal pressure constrain policy responsiveness? Evidence from Germany. In: *European Journal of Political Research* 61, p. 374-397.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. Online at: https://data.stanford.edu/congress_text

Herndon, Thomas; Ash, Michael; Pollin, Robert (2014): Does High Public Debt Consistently Stifle Economic Growth. A Critique of Reinhart and Rogoff. In: *Cambridge Journal of Economics* 38 (2), p. 257-279.

International Monetary Fund (IMF) Datamapper (2021): World Economic Outlook. Real GDP growth. Annual percent change. Online at: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/WEOWORLD

Keele, Lukas (2020): Difference-in-Difference: Neither Natural nor an Experiment. In: Curini, Luigi; Franzese, Robert (Eds.): *The SAGE Handbook of Research Methods in Political Science and International Relations.* p. 822-834.

Keynes, John M. (1936): *The General Theory of Employment, Interest and Money.* London (UK): Macmillan.

Gilens, Martin (2005): Inequality and Democratic Responsiveness. In: *Public Opinion Quaterly* 69 (5), p. 778-796.

Hill, Carter R.; Griffiths, William E.; Lim, Guay C. (2011): *Principles of Econometrics*. 4th Edition. Hoboken (USA): Wiley.

Okun, Arthur M. (2015): *Equality and Efficiency. The Big Tradeoff.* Washington (USA): The Brookings Institution.

Peterson, Paul E. (1981): *City Limits.* Chicago: The University of Chicago Press.

Reifschneider, Dave; Wascher, William; Wilcox, David (2014): Aggregate Supply in the United States: Recent Developments and Implications for the Conduct of Monetary Policy. IMF-Working Paper. Online at: https://www.imf.org/external/np/res/seminars/2013/arc/pdf/wilcox.pdf

Reinhart, Carmen M.; Rogoff, Kenneth S. (2010): Growth in a Time of Debt. In: *American Economic Review* 100, p. 573-578.

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

Schumpeter, Joseph A. (1991): The Crisis of the Tax State. In: Swedberg, Richard (Ed.): *The Economics and Sociology of Capitalism*. Princeton (United States): Princeton University Press, p. 99 – 140.

Stiglitz, Joseph E. (2016): How to Restore Equitable and Sustainable Economic Growth in the United States. In: *American Economic Review* 106 (5), p. 43-47.

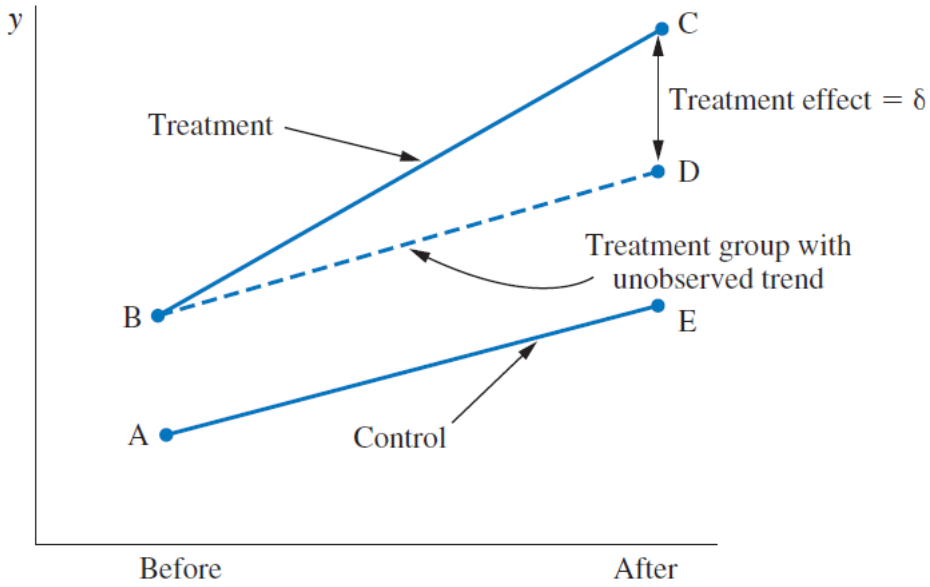
Tilley, Cameron (2023): The Miraculous Old-Time Fiscal Religion: How a Political Norm Discouraged Deficit Spending. *Working Paper Duke University* (Forthcoming).

Weber, Isabella; Wasner, Evan (2023): Sellers´ Inflation, Profits and Conflict: Why can Large Firms Hike Prices in an Emergency? In: *Economics Department Working Paper Series*. Online: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1348&context=econ_workingpaper

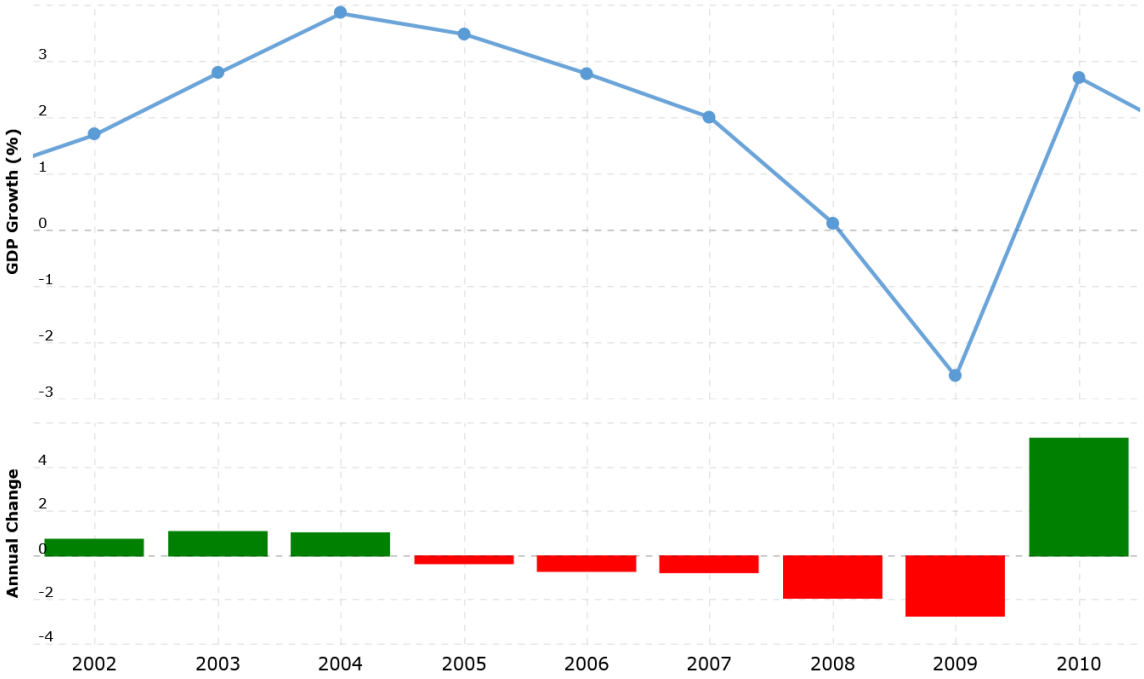
Yu, Jinhai; Jennings Jr., Edward T.; Butler, J.S. (2019): Dividing the Pie. In: *State Politicis and Policy Quaterly* 19 (2), p. 236-258.

8. Appendix

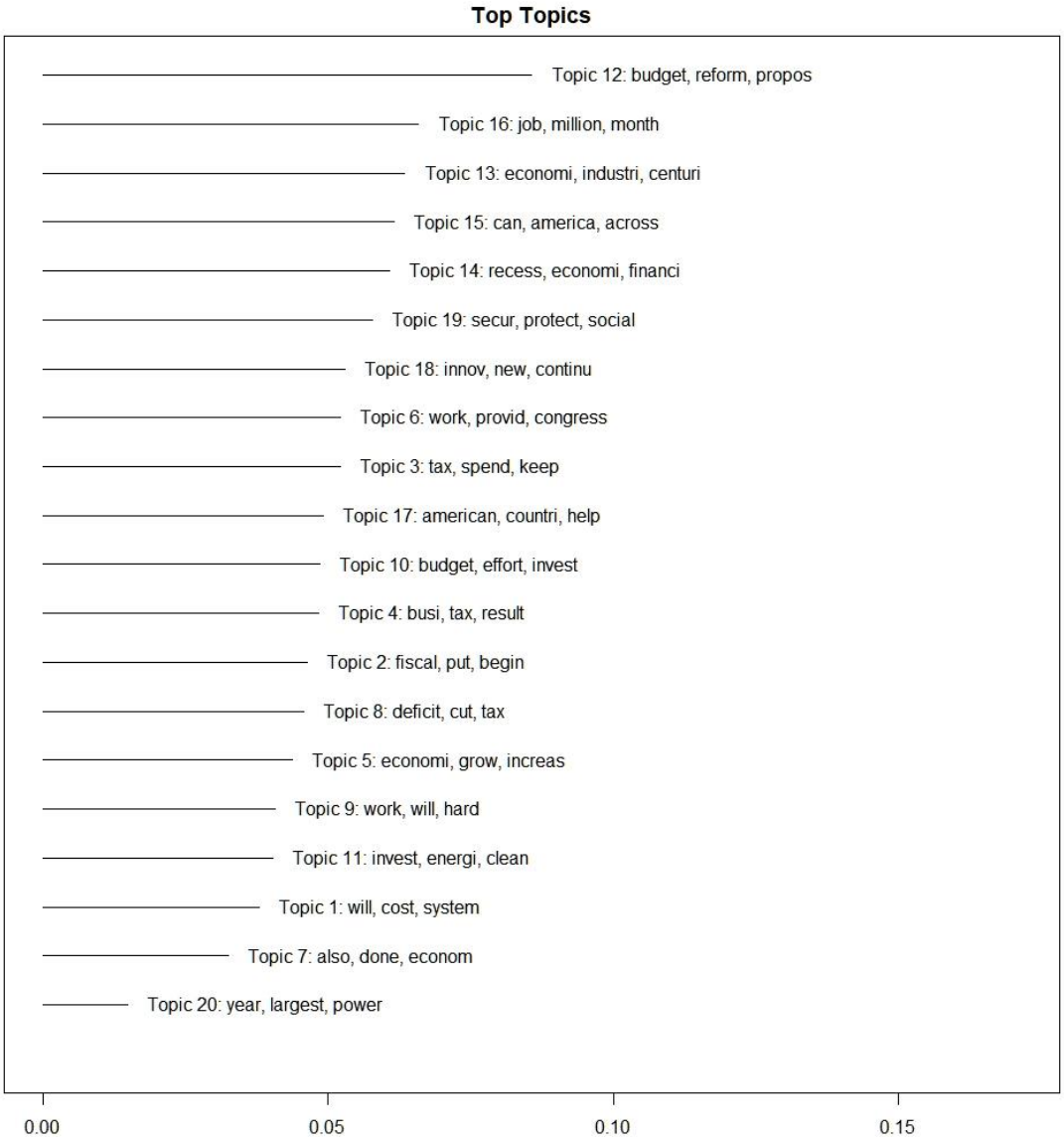
Appendix 1 – Figure 1: Difference-in-Difference Design



Appendix 2 – Figure 2: Annual GDP Growth USA



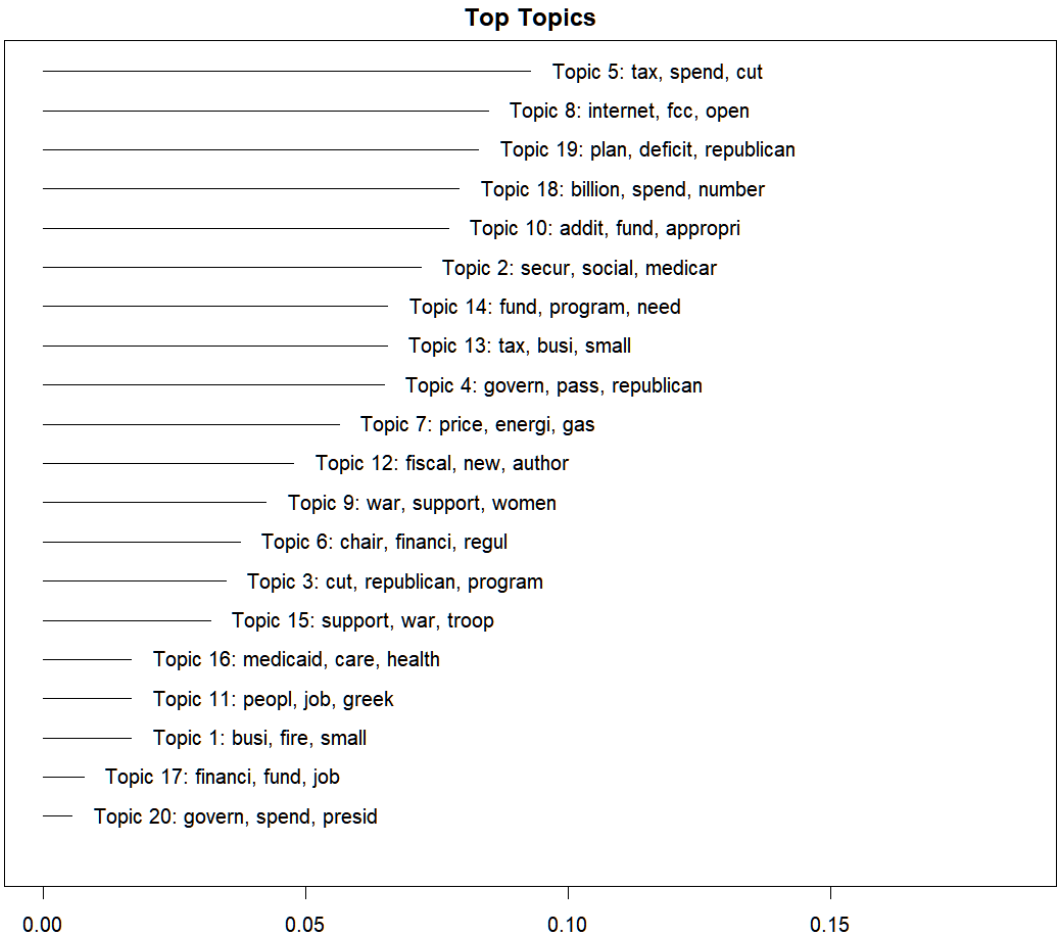
Appendix 3 – Figure 3: Top Topics in U.S. Presidential Speeches Data Set



Topic 12 focuses on the reform process itself and has therefore been named "Fiscal Policy and Reform". Topic 16 appears to focus on "Employment and Unemployment". Topic 13, which is the third most frequent kernel in the overall text dataset deals with "Economic and Industrial Development" It matches the Development category established by Peterson (1981). The next Topic appears to have a stop word still included. 'Can' referring to the verb does not hold a specifically informative value and will be removed in a next iteration. Topic 15 has been named "International Relations and Global Challenges", due to the presence of the words 'world' and 'difficult'. Topic 14 focuses on the "Economic Crises and Financial Trouble" marked by the word great and the stem 'recess' most likely

resembling the ‘Great Recession’. The last Topic has been labeled “Social Security and Future Planning”. This last Topic could be identified as the category of most interest since it comes closest to the demand for redistribution.

Appendix 4 – Figure 4: Top Topics in U.S. House of Representatives Data Set



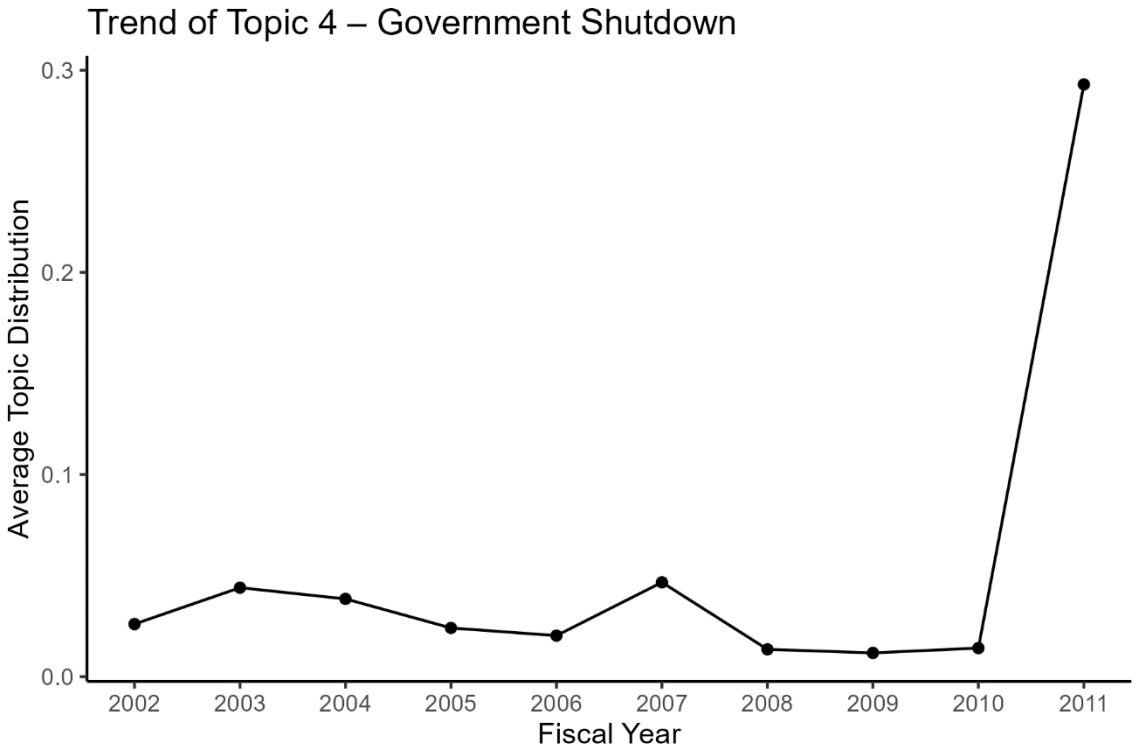
Appendix 5 – Figure 5: Top 6 Topics detailed in U.S. Presidential Speeches Data Set

```
Topic 12 Top Words:
         Highest Prob: budget, reform, propos, program, billion
         FREX: propos, billion, congress, program, reduct
         Lift: branc, common, hes, regul, unjustifi
         Score: regul, propos, billion, congress, reform
Topic 16 Top Words:
         Highest Prob: job, million, month, will, last
         FREX: million, month, job, time, last
         Lift: lose, mortgag, lost, experienc, million
         Score: lose, million, month, job, lost
Topic 13 Top Words:
         Highest Prob: economi, industri, centuri, new, must
         FREX: industri, centuri, economi, strong, must
         Lift: envi, industri, cycl, centuri, strong
         Score: envi, industri, strong, economi, centuri
Topic 15 Top Words:
         Highest Prob: can, america, across, world, difficult
         FREX: across, difficult, america, can, win
         Lift: 've, news, factori, reach, anywher
         Score: 've, across, america, difficult, can
Topic 14 Top Words:
         Highest Prob: recess, economi, financi, help, great
         FREX: street, disast, inc, natur, financi
         Lift: attac, conceal, devast, disast, inc
         Score: post, street, inc, disast, natur
Topic 19 Top Words:
         Highest Prob: secur, protect, social, generat, futur
         FREX: social, protect, secur, command, highest
         Lift: chief, exampl, retire, overcom, command
         Score: chief, protect, social, secur, command
```

Appendix 6 – Figure 6: Top 10 Topics detailed in U.S. House of Representatives Data Set

```
> Topic 5:
    Marginal Highest Prob: tax, spend, cut, balanc, blue, govern, percent, growth, increas, dog
  Marginal FREX: blue, dog, wast, growth, balanc, relief, tax, spend, budget, dig
  Marginal Lift: dilig, toomey, tricar, rsc, phasein, dog, blue, undertax, ammunit, runaway
  Marginal Score: dilig, hid, kennard, nonmedicar, tax, dog, blue, toomey, spend, rsc


  Topic 8:
    Marginal Highest Prob: internet, fcc, open, regul, consum, rule, innov, provid, compani, broadband
  Marginal FREX: fcc, internet, broadband, fccs, innov, parliamentari, disapprov, communic, content, open
  Marginal Lift: comcast, deregulatori, dna, ebay, googl, netflix, walden, wireless, amazon, ancillari
  Marginal Score: hid, kennard, nonmedicar, parliamentari, internet, fcc, broadband, fccs, regul, cra


  Topic 19:
    Marginal Highest Prob: plan, deficit, repyublican, veteran, cut, tax, democrat, billion, care, spend
  Marginal FREX: plan, texa, deficit, veteran, south, carolina, show, trillion, substitut, surplus
  Marginal Lift: dumb, hoyer, raw, sidelin, swing, principi, fascist, icit, undemocrat, revert
  Marginal Score: hid, hoyer, kennard, nonmedicar, veteran, deficit, surplus, substitut, tax, spratt


  Topic 18:
    Marginal Highest Prob: billion, spend, number, money, defens, year, peopl, confer, pay, need
  Marginal FREX: wisconsin, confer, defens, iowa, shell, number, correct, game, billion, discretionari
  Marginal Lift: bypass, flatlin, longrang, macroeconom, mug, placehold, riverboat, tradeoff, wisconsin, committeereport
  Marginal Score: hid, kennard, nonmedicar, wisconsin, shell, scare, conceal, omb, chairman, baselin


  Topic 10:
    Marginal Highest Prob: addit, fund, appropri, million, earmark, legisl, major, includ, provid, billion
  Marginal FREX: addit, earmark, omnibus, obey, appropri, staff, contain, hous, subcommitte, research
  Marginal Lift: headquart, meth, withhold, airdrop, explanatori, nabor, addit, archiv, chastis, darfur
  Marginal Score: addit, hid, kennard, nonmedicar, earmark, obey, omnibus, subcommitte, formula, fund


  Topic 2:
    Marginal Highest Prob: secur, social, medicar, fund, republican, tax, cut, year, drug, surplus
  Marginal FREX: medicar, social, prescript, drug, trust, surplus, senior, secur, raid, cbo
  Marginal Lift: fiveyear, dav, doughnut, echo, fantasi, freefal, giveback, hast, hid, horizon
  Marginal Score: hast, hid, nonmedicar, medicar, prescript, drug, social, raid, trust, surplus


  Topic 14:
    Marginal Highest Prob: fund, program, need, educ, provid, increas, veteran, nation, support, billion
  Marginal FREX: frank, child, educ, highway, program, grant, fund, behind, cbc, black
  Marginal Lift: postur, frank, elev, inmat, pertain, selfsuffici, socioeconom, americorp, reintegr, dismay
  Marginal Score: frank, hid, kennard, nonmedicar, cbc, fund, educ, highway, teacher, veteran


  Topic 13:
    Marginal Highest Prob: tax, busi, small, ask, day, thing, pay, peopl, like, person
  Marginal FREX: ask, code, person, owner, sell, small, sometim, busi, death, ration
  Marginal Lift: dispens, lehigh, patienc, comedian, imposit, layer, lobbyist, perkiomen, dread, outhous
  Marginal Score: dispens, hid, kennard, nonmedicar, tax, suspens, quorum, restaur, farm, sell


  Topic 4:
    Marginal Highest Prob: govern, pass, republican, peopl, spend, shut, today, that, shutdown, last
  Marginal FREX: shutdown, shut, that, cant, yesterday, pass, adjourn, govern, parti, reid
  Marginal Lift: adjourn, zeppelin, unveil, reid, scout, freshmen, circul, hemorrhag, shes, steni
  Marginal Score: adjourn, hid, kennard, nonmedicar, shutdown, shut, reid, cant, didnt, what


  Topic 7:
    Marginal Highest Prob: price, energi, gas, oil, bay, democrat, chesapeak, countri, new, today
  Marginal FREX: chesapeak, gas, ill, bay, oil, price, drill, gasolin, watertrail, pump
  Marginal Lift: continent, gag, americanmad, cellulos, gallup, ill, illconsid, offlimit, pellet, ret
  Marginal Score: hid, ill, kennard, nonmedicar, chesapeak, gas, drill, oil, watertrail, bay
```

Appendix 7 – Figure 7: Frequency Distribution Topic 4 over time

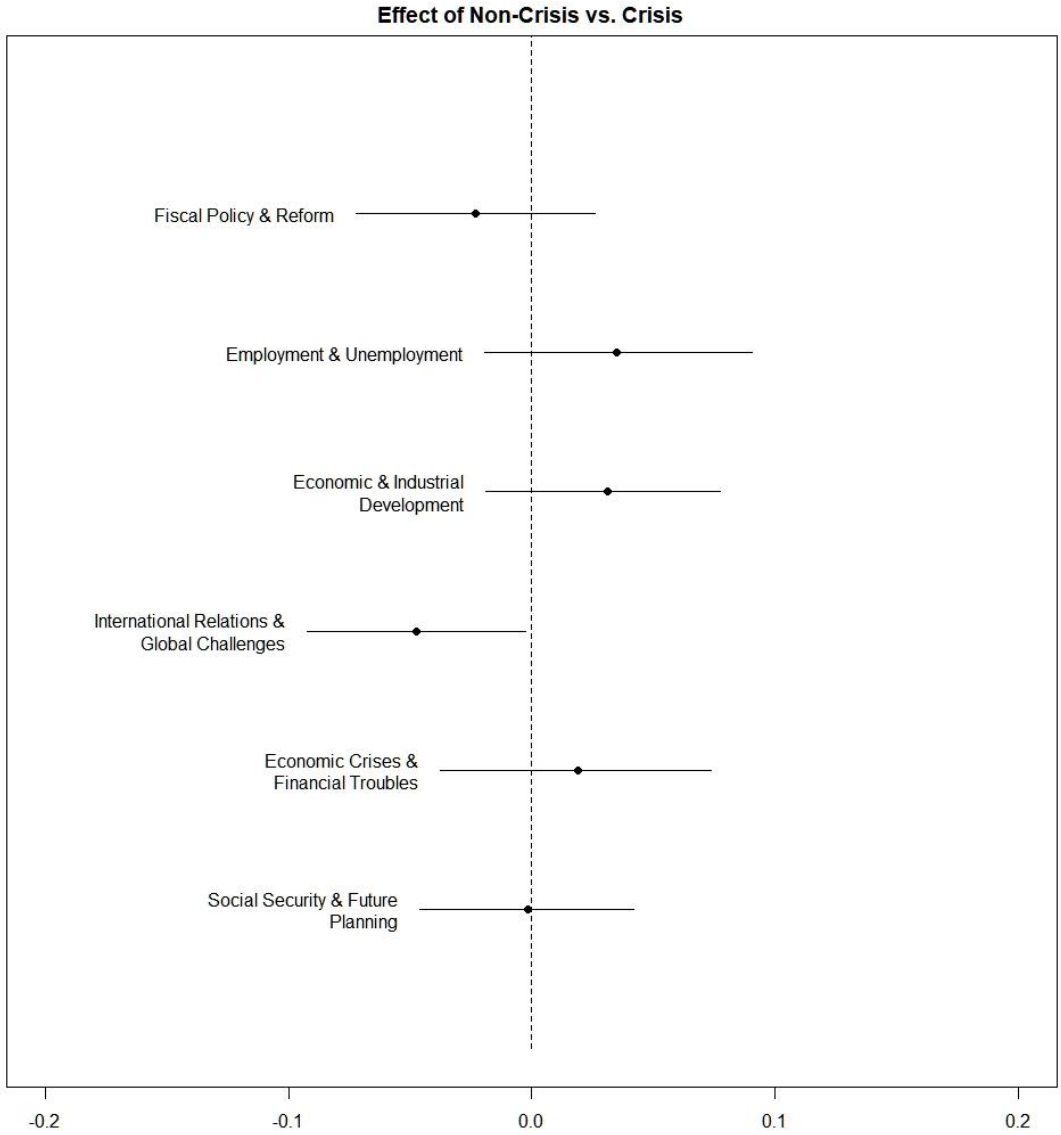


Appendix 8 – Interpretation of preliminary findings

The following plot shows the distribution of the identified topics for the years that had been assigned crisis as opposed to those that have not. The first impression may dampen expectations since all but the "International Relations and Global Conflicts" category are insignificant. This category however appears significantly more often during times of prosperity. The impact of redistribution related topic "Social Security and Future Planning" to the contrast appears insignificant and the estimate is closest to zero. For Hypothesis 1 first implications can be drawn from this. The potential treatment (times of crisis vs. prosperity) appears to have no effect of the mentioning of redistribution related issues.

It is important to state that the findings are thus far preliminary since we aim to exploit a greater dataset in order to increase the validity of the findings and to achieve other objectives of this paper as laid out in the sections on Method and Data.

Figure 8 – Preliminary analysis of the Effect of Economic Crisis:



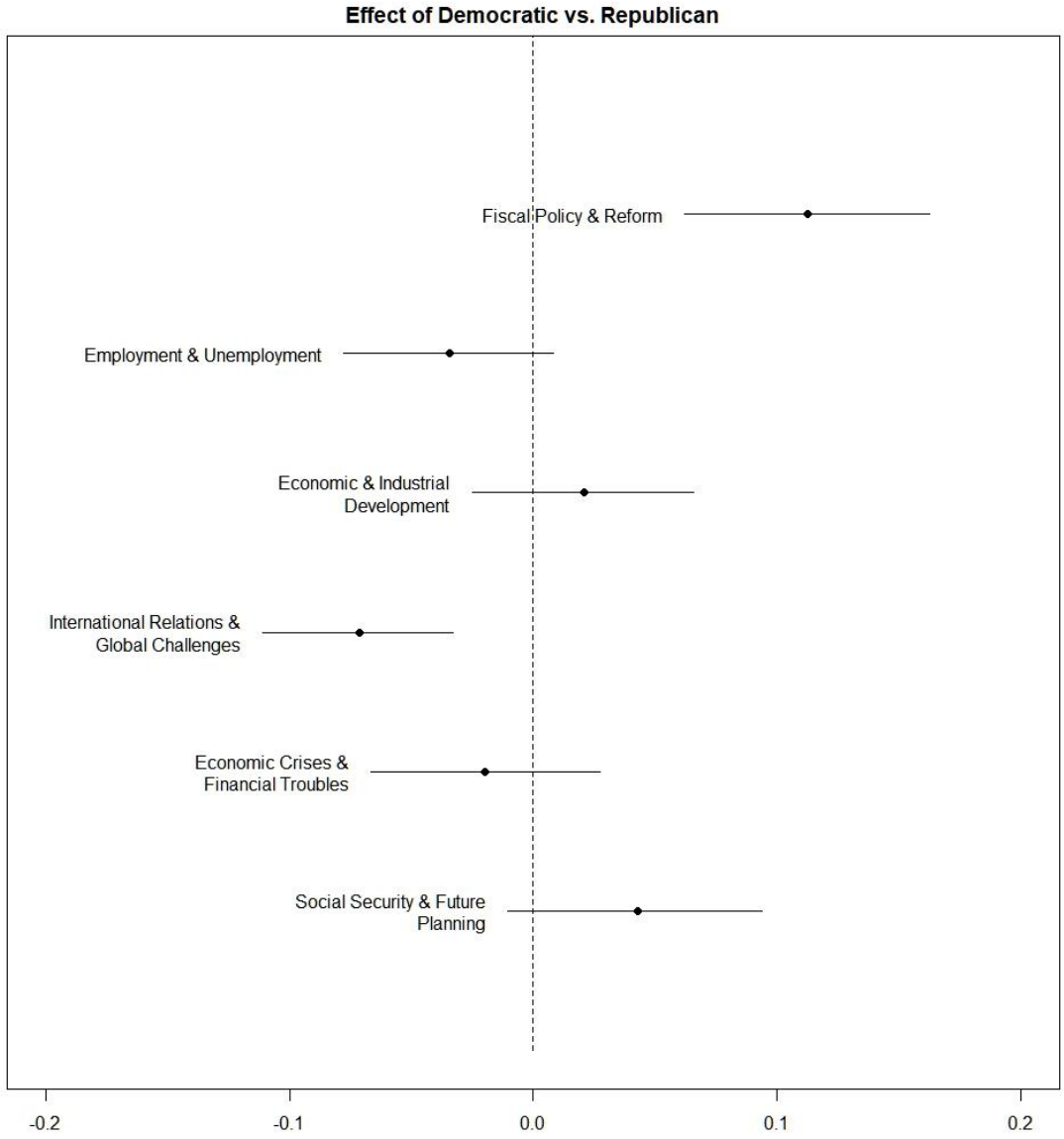
In an effort to establish the causal mechanism by which redistributive topics appear or disappear in budgetary debates, we plotted the different frequencies in mentioning of the established topics conditioned on the party affiliation of the speaker.

Here the results appear significant in two out of six categories and the redistribution related category, although still insignificant, has moved in the direction of a positive effect. "Fiscal Policy and Reform" appears to be a topic mentioned more frequently by President Obama of the Democratic Party. To the contrary the topic "International Relations and Global Challenges" is more frequently used by his Republican predecessor Georg W. Bush. This could eventually be the result of a cohort effect. Given that the Bush presidency was

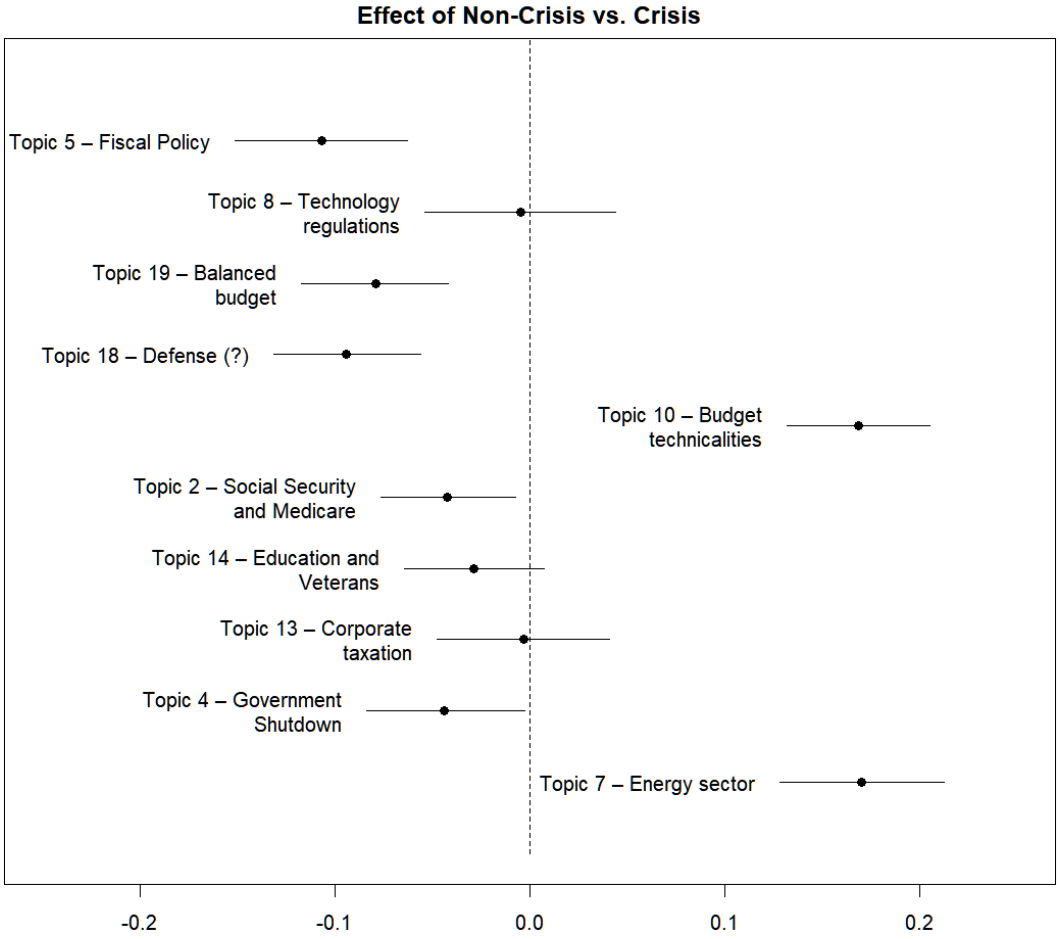
shaped by the terrorist attack on September 11th in 2001 and the subsequent "war on terror" declared by his administration. On the other hand, besides having inherited ongoing wars and military operations around the world, the Obama presidency and especially the time frame under study has predominantly shaped by the Great Recession and its aftermath. This the difference may not be entirely causal. The findings are, after all, preliminary and an investigation of an equal number of speakers from each party during the whole time period could render this finding insignificant.

"Social Security and Future Planning" as previously mentioned remains insignificant. However, it appears to be more common among Democratic speakers. This is in line with the second hypothesis, should the investigation of additional data further this trend.

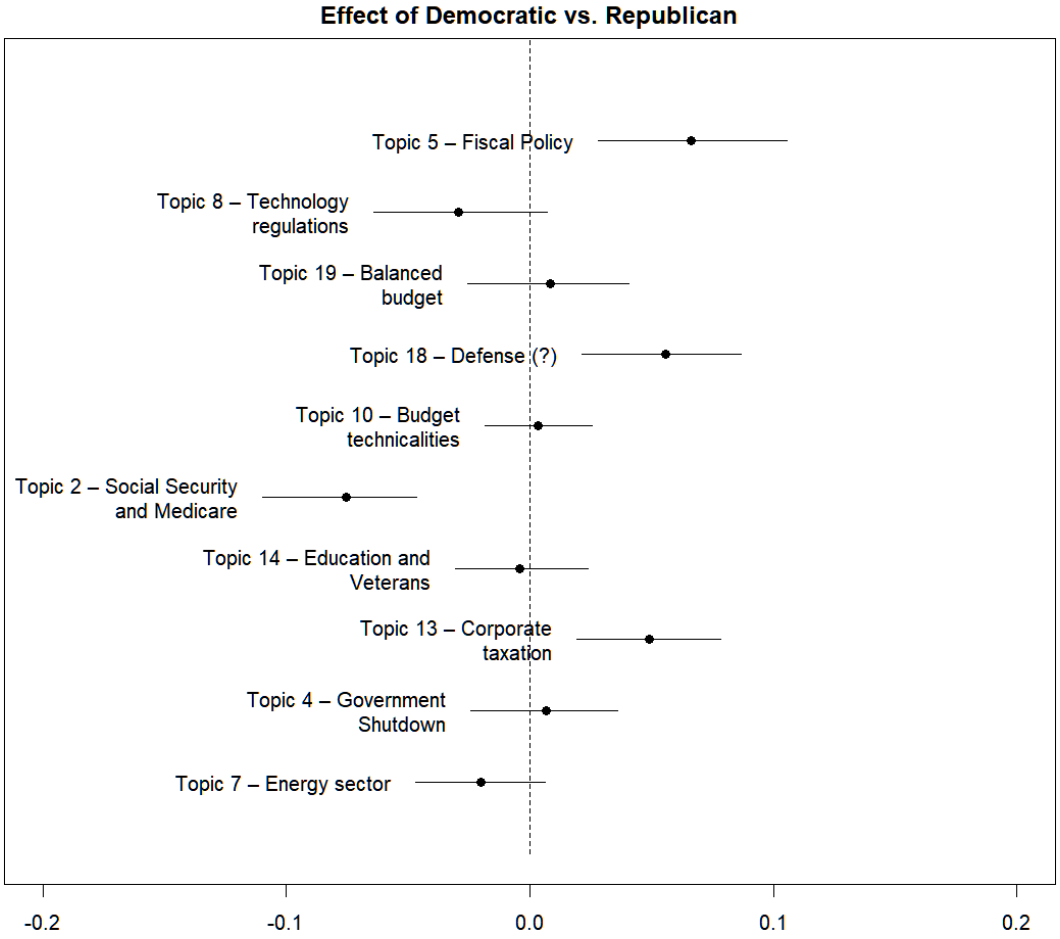
Figure 9 – Preliminary findings of the Effect of Party Affiliation



Appendix 9 – Figure 10: Final findings of the Effect of Economic Crisis



Appendix 10 – Figure11: Final findings of the Effect of Party Affiliation



Appendix 11: R Code

Xinyuan Li & Philipp Weisenburger

2023-12-05

Data Cleaning

```
#House's Debate Analysis
rm(list = ls())

#Install the stm package
#install.packages("stm")

#Load the package
library(stm)

## stm v1.3.6.1 successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com

#Load the data
setwd("C:/Users/scene/Documents/R/GIRI/project")

# read text and basic cleaning
speeches_paths <- c("data/hein-daily/speeches_107.txt",
                    "data/hein-daily/speeches_107.txt",
                    "data/hein-daily/speeches_108.txt",
                    "data/hein-daily/speeches_108.txt",
                    "data/hein-daily/speeches_109.txt",
                    "data/hein-daily/speeches_110.txt",
                    "data/hein-daily/speeches_110.txt",
                    "data/hein-daily/speeches_110.txt",
                    "data/hein-daily/speeches_111.txt",
                    "data/hein-daily/speeches_112.txt")

maps_paths <- c("data/hein-daily/107_SpeakerMap.txt",
                "data/hein-daily/107_SpeakerMap.txt",
                "data/hein-daily/108_SpeakerMap.txt",
                "data/hein-daily/108_SpeakerMap.txt",
                "data/hein-daily/109_SpeakerMap.txt",
                "data/hein-daily/110_SpeakerMap.txt",
                "data/hein-daily/110_SpeakerMap.txt",
                "data/hein-daily/110_SpeakerMap.txt",
                "data/hein-daily/111_SpeakerMap.txt",
                "data/hein-daily/112_SpeakerMap.txt")
speech_list <- list()
map_list <- list()
data_list <- list()
speech_line <- list(c(25349,89),
                    c(113464,203),
```

```
                c(22078,231),
                c(131692,266),
                c(14950,182),
                c(9917,115),
                c(127081,32),
                c(170707,65),
                c(69547,64),
                c(22559,231)
                )


# House speeches for ten fiscal years
for (i in 1:3){
  column_names <- read.table(speeches_paths[i], sep = "|", nrows = 1, header
= FALSE)
  speech_list[[i]] <- read.table(speeches_paths[i], sep = "|",
                                 skip = speech_line[[i]][1], nrows =
speech_line[[i]][2], header = FALSE)
  colnames(speech_list[[i]]) <- as.character(unlist(column_names))
  map_list[[i]] <- read.delim(maps_paths[i], sep = "|", header = TRUE)
  data_list[[i]] <- merge(speech_list[[i]], map_list[[i]], by = "speech_id",
all.x = TRUE)
}

column_names <- read.table(speeches_paths[3], sep = "|", nrows = 1, header =
FALSE)
speech3a <- read.table(speeches_paths[3], sep = "|",
                               skip = 20215, nrows = 220, header = FALSE)
colnames(speech3a) <- as.character(unlist(column_names))
map3a <- read.delim(maps_paths[3], sep = "|", header = TRUE)
data3a <- merge(speech3a, map3a, by = "speech_id", all.x = TRUE)

for (i in 4:10){
  column_names <- read.table(speeches_paths[i], sep = "|", nrows = 1, header
= FALSE)
  speech_list[[i]] <- read.table(speeches_paths[i], sep = "|",
                                 skip = speech_line[[i]][1], nrows =
speech_line[[i]][2], header = FALSE)
  colnames(speech_list[[i]]) <- as.character(unlist(column_names))
  map_list[[i]] <- read.delim(maps_paths[i], sep = "|", header = TRUE)
  data_list[[i]] <- merge(speech_list[[i]], map_list[[i]], by = "speech_id",
all.x = TRUE)
}

#merge

data <- rbind(
  cbind(data_list[[1]], year = "2002", crisis='non-crisis'),
  cbind(data_list[[2]], year = "2003", crisis='non-crisis'),
  cbind(data_list[[3]], year = "2004", crisis='non-crisis'),
  cbind(data3a, year = "2004", crisis='non-crisis'),
```

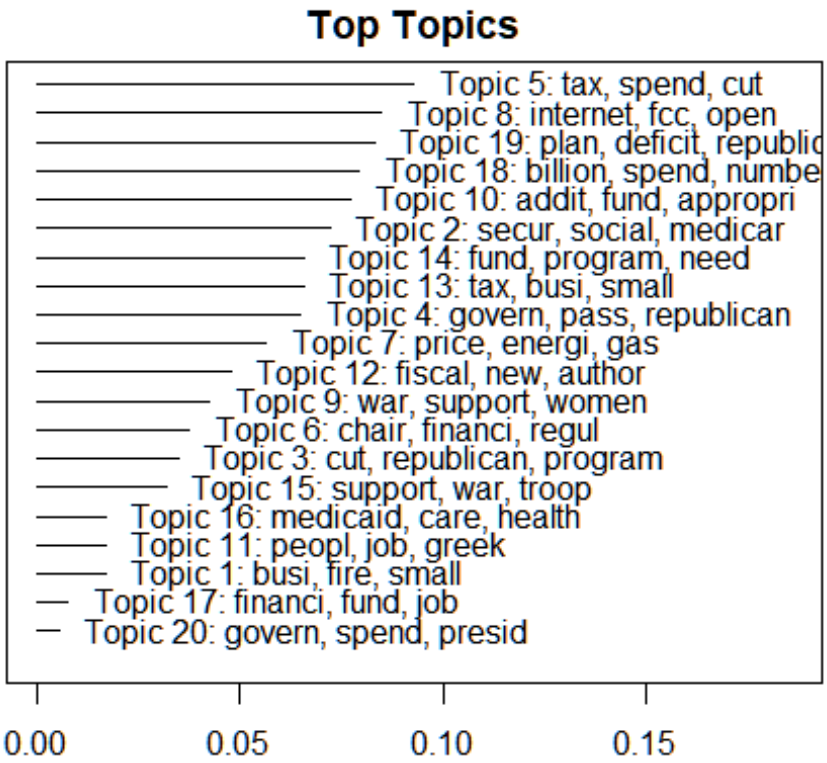
```
  cbind(data_list[[4]], year = "2005", crisis='non-crisis'),
  cbind(data_list[[5]], year = "2006", crisis='crisis'),
  cbind(data_list[[6]], year = "2007", crisis='crisis'),
  cbind(data_list[[7]], year = "2008", crisis='crisis'),
  cbind(data_list[[8]], year = "2009", crisis='crisis'),
  cbind(data_list[[9]], year = "2010", crisis='crisis'),
  cbind(data_list[[10]], year = "2011", crisis='non-crisis'))

data<- na.omit(data)

saveRDS(data, file = "data/data_house.rds")


#Clean text data
data <- readRDS(file = "data/data_house.rds")
processed<- textProcessor(data$speech, #the column that has the text
                          metadata = data, #the name of data set
                          lowercase=TRUE, removestopwords=TRUE,
removenumbers=TRUE,
                          removepunctuation=TRUE, stem=TRUE,
                          wordLengths=c(3,Inf),  #remove words shorter than 3
letters
                          sparselevel=1, language="en",  verbose=TRUE,
                          onlycharacter= FALSE, striphtml=FALSE,

customstopwords=c("economic","economics","economical","economies","economy",

"america","america's","american","americans",
                                          "speaker","can","minute",
"minutes","gentleman","gentlewoman",

"just","now","time","madam","miss",

"will","that","ladies","gentlemen",
                                          "take","what","there","across",
                                          "one","year","want","back","say",

"yea","nay","get","rollcall","aye","noe",

"yield","chairman","committee","committees",

"amendment","amendments","amend","amends","amending",
                                          "budget","member","members",
"vote","votes",
                                          "voter","voters", "house",
"rule",
                                          "come","let","resolution",
"debate", "debates",
                                          "thank", "ask", "asks", "asking",
                                          "distinguish", "distinguished",
"may",
```

```
                                                  "question","questions", "work",
"works", "working",
                                                  "bill","bills","congress",
"congressional",
                                                  "process",
"senate","senator","senators",
                                                  "side","sides",
"friend","friends",
                                                  "know","known","knowing","knows",
                                                  "colleague","colleagues",
"think","thinks","thinking",
                                                  "unanimous","unanimously",
                                                  "caucus","caucuses",
"make","makes")) #custom stopwords

## Building corpus...
## Converting to Lower Case...
## Removing punctuation...
## Removing stopwords...
## Remove Custom Stopwords...
## Removing numbers...
## Stemming...
## Creating Output...

saveRDS(processed, file = "data/processed_house.rds")
```

STM

```
#House's Debate Analysis
rm(list = ls())

#Install the stm package
#install.packages("stm")

#Load the package
library(stm)

#Load the data
setwd("C:/Users/scene/Documents/R/GIRI/project")
processed <- readRDS(file = "data/processed_house.rds")

#Now we separate our data into neat bits for our analysis
out <- prepDocuments(processed$documents, processed$vocab, processed$meta)

## Removing 5328 of 10375 terms (5328 of 81833 tokens) due to frequency
## Removing 2 Documents with No Words
## Your corpus now has 846 documents, 5047 terms and 76505 tokens.

docs <- out$documents
vocab <- out$vocab
meta <-out$meta
```

```r
# Find the best K value
set.seed(2008)
findingk <- searchK(out$documents, out$vocab,
                    K = c(19:21),
                    prevalence =~ crisis * party, #topic prevalence
                    data = meta, verbose=FALSE)


#Fit a model with our given k and topic prevalence and content equations

set.seed(2008)
First_STM <- stm(documents = out$documents, vocab = out$vocab,
                 K = 20,
                 prevalence =~ crisis * party,
                 max.em.its =75, data = out$meta,init.type = "Spectral",
                 verbose = FALSE)

#We built the model!
saveRDS(First_STM, file = "data/First_STM_house.rds")
#Now we analyze it
First_STM <- readRDS(file = "data/First_STM_house.rds")

#First, we identify topics and interpret them


#Plot the most prevalent topics in this model
par(mar=c(2,2,2,2))
plot(First_STM)
```




```
sageLabels(First_STM,n=10)

## Topic 1:
##       Marginal Highest Prob: busi, fire, small, servic, nation, peopl,
state, also, plan, legisl
##       Marginal FREX: fire, forest, speci, interior, colorado, endang,
cleveland, steve, island, manag
##       Marginal Lift: buildup, roadless, wildemess, chairmani, gradual,
mall, prestigi, mid, ski, speci
##       Marginal Score: chairmani, hid, kennard, forest, madam, cleveland,
speci, celebr, ski, interior
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 2:
##       Marginal Highest Prob: secur, social, medicar, fund, republican,
tax, cut, year, drug, surplus
##       Marginal FREX: medicar, social, prescript, drug, trust, surplus,
senior, secur, raid, cbo
##       Marginal Lift: fiveyear, dav, doughnut, echo, fantasi, freefal,
giveback, hast, hid, horizon
##       Marginal Score: hast, hid, nonmedicar, medicar, prescript, drug,
social, raid, trust, surplus
##
##       Topic Kappa:
##       Kappa with Baseline:
##
```

```
## Topic 3:
##       Marginal Highest Prob: cut, republican, program, billion, veteran,
million, nation, tax, year, care
##       Marginal FREX: slash, david, spratt, blueprint, veteran, wealthiest,
cut, valu, child, back
##       Marginal Lift: clever, hay, heath, oregonian, prioritiesa, scrip,
illadvis, toughen, sbas, skew
##       Marginal Score: hay, hid, nonmedicar, medicaid, david, spratt,
slash, veteran, cut, blueprint
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 4:
##       Marginal Highest Prob: govern, pass, republican, peopl, spend, shut,
today, that, shutdown, last
##       Marginal FREX: shutdown, shut, that, cant, yesterday, pass, adjourn,
govern, parti, reid
##       Marginal Lift: adjourn, zeppelin, unveil, reid, scout, freshmen,
circul, hemorrhag, shes, steni
##       Marginal Score: adjourn, hid, kennard, nonmedicar, shutdown, shut,
reid, cant, didnt, what
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 5:
##       Marginal Highest Prob: tax, spend, cut, balanc, blue, govern,
percent, growth, increas, dog
##       Marginal FREX: blue, dog, wast, growth, balanc, relief, tax, spend,
budget, dig
##       Marginal Lift: dilig, toomey, tricar, rsc, phasein, dog, blue,
undertax, ammunit, runaway
##       Marginal Score: dilig, hid, kennard, nonmedicar, tax, dog, blue,
toomey, spend, rsc
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 6:
##       Marginal Highest Prob: chair, financi, regul, market, requir, end,
swap, madam, risk, deriv
##       Marginal FREX: swap, deriv, chair, user, risk, exchang, madam,
market, exempt, financi
##       Marginal Lift: adler, cras, neglig, columbia, euronext, issuer,
kanjorski, liabl, mccarthi, nrsros
##       Marginal Score: columbia, kennard, swap, madam, deriv, bailout,
regul, collater, overthecount, cfpa
##
##       Topic Kappa:
##       Kappa with Baseline:
```

```
##
## Topic 7:
##       Marginal Highest Prob: price, energi, gas, oil, bay, democrat,
chesapeak, countri, new, today
##       Marginal FREX: chesapeak, gas, ill, bay, oil, price, drill, gasolin,
watertrail, pump
##       Marginal Lift: continent, gag, americanmad, cellulos, gallup, ill,
illconsid, offlimit, pellet, ret
##       Marginal Score: hid, ill, kennard, nonmedicar, chesapeak, gas,
drill, oil, watertrail, bay
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 8:
##       Marginal Highest Prob: internet, fcc, open, regul, consum, rule,
innov, provid, compani, broadband
##       Marginal FREX: fcc, internet, broadband, fccs, innov, parliamentari,
disapprov, communic, content, open
##       Marginal Lift: comcast, deregulatori, dna, ebay, googl, netflix,
walden, wireless, amazon, ancillari
##       Marginal Score: hid, kennard, nonmedicar, parliamentari, internet,
fcc, broadband, fccs, regul, cra
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 9:
##       Marginal Highest Prob: war, support, women, men, troop, peopl,
countri, world, unit, nation
##       Marginal FREX: men, inquiri, love, soldier, women, mission, uniform,
hussein, gulf, son
##       Marginal Lift: chronicl, demean, gear, guidanc, inquiri, matthew,
proverb, speedi, teeth, energet
##       Marginal Score: hid, inquiri, kennard, nonmedicar, saddam, hussein,
men, war, love, women
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 10:
##       Marginal Highest Prob: addit, fund, appropri, million, earmark,
legisl, major, includ, provid, billion
##       Marginal FREX: addit, earmark, omnibus, obey, appropri, staff,
contain, hous, subcommitte, research
##       Marginal Lift: headquart, meth, withhold, airdrop, explanatori,
nabor, addit, archiv, chastis, darfur
##       Marginal Score: addit, hid, kennard, nonmedicar, earmark, obey,
omnibus, subcommitte, formula, fund
##
##       Topic Kappa:
```

```
##       Kappa with Baseline:
##
## Topic 11:
##       Marginal Highest Prob: peopl, job, greek, state, mani, greec,
countri, independ, year, nation
##       Marginal FREX: greek, greec, nay, independ, ireland, ancient, polic,
celebr, nuclear, northern
##       Marginal Lift: gothic, ancestor, britain, coffer, colonist, cyprus,
fledgl, greek, greekamerican, hellen
##       Marginal Score: hid, kennard, nay, nonmedicar, greek, greec,
ireland, turkish, celebr, ancient
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 12:
##       Marginal Highest Prob: fiscal, new, author, outlay, michigan, vote,
res, concurr, consider, year
##       Marginal FREX: michigan, outlay, author, fiscal, new, concurr,
consider, roll, detain, vote
##       Marginal Lift: michigan, rolical, nos, poster, outlay, detain,
appro, herebi, author, propriat
##       Marginal Score: hid, kennard, michigan, nonmedicar, outlay, fiscal,
author, detain, rolical, new
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 13:
##       Marginal Highest Prob: tax, busi, small, ask, day, thing, pay,
peopl, like, person
##       Marginal FREX: ask, code, person, owner, sell, small, sometim, busi,
death, ration
##       Marginal Lift: dispens, lehigh, patienc, comedian, imposit, layer,
lobbyist, perkiomen, dread, outhous
##       Marginal Score: dispens, hid, kennard, nonmedicar, tax, suspens,
quorum, restaur, farm, sell
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 14:
##       Marginal Highest Prob: fund, program, need, educ, provid, increas,
veteran, nation, support, billion
##       Marginal FREX: frank, child, educ, highway, program, grant, fund,
behind, cbc, black
##       Marginal Lift: postur, frank, elev, inmat, pertain, selfsuffici,
socioeconom, americorp, reintegr, dismay
##       Marginal Score: frank, hid, kennard, nonmedicar, cbc, fund, educ,
highway, teacher, veteran
##
```

```
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 15:
##       Marginal Highest Prob: support, war, troop, famili, nation, iraq,
men, women, presid, forc
##       Marginal FREX: brave, tonight, saddam, god, prayer, command, arm,
bless, troop, hussein
##       Marginal Lift: commenc, disrespect, routin, sympathi, tyrant,
unanim, hampshir, tile, aleadi, badger
##       Marginal Score: hampshir, hid, kennard, nonmedicar, saddam, hussein,
prayer, iraq, tonight, gratitud
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 16:
##       Marginal Highest Prob: medicaid, care, health, cut, state, program,
peopl, billion, million, cost
##       Marginal FREX: medicaid, illinoi, coverag, uninsur, nurs, care,
health, hospit, insur, vulner
##       Marginal Lift: awak, bottl, illinoi, payer, rangel, renaiss, tort,
uncompens, threequart, gold
##       Marginal Score: hid, illinoi, kennard, nonmedicar, medicaid,
uninsur, coverag, medicar, rangel, prescript
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 17:
##       Marginal Highest Prob: financi, fund, job, peopl, billion, spend,
million, democrat, republican, includ
##       Marginal FREX: dealer, financi, yea, bailout, tarp, loan, system,
arbitr, manufactur, fed
##       Marginal Lift: franchis, statutorili, yea, arbitr, chrysler, takeov,
fanni, freddi, yemen, hook
##       Marginal Score: hid, kennard, nonmedicar, yea, dealer, bailout,
tarp, arbitr, regul, madam
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 18:
##       Marginal Highest Prob: billion, spend, number, money, defens, year,
peopl, confer, pay, need
##       Marginal FREX: wisconsin, confer, defens, iowa, shell, number,
correct, game, billion, discretionari
##       Marginal Lift: bypass, flatlin, longrang, macroeconom, mug,
placehold, riverboat, tradeoff, wisconsin, committeereport
##       Marginal Score: hid, kennard, nonmedicar, wisconsin, shell, scare,
conceal, omb, chairman, baselin
```

```
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 19:
##       Marginal Highest Prob: plan, deficit, republican, veteran, cut, tax,
democrat, billion, care, spend
##       Marginal FREX: plan, texa, deficit, veteran, south, carolina, show,
trillion, substitut, surplus
##       Marginal Lift: dumb, hoyer, raw, sidelin, swing, principi, fascist,
icit, undemocrat, revert
##       Marginal Score: hid, hoyer, kennard, nonmedicar, veteran, deficit,
surplus, substitut, tax, spratt
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 20:
##       Marginal Highest Prob: govern, spend, presid, cut, peopl,
republican, countri, need, shut, fund
##       Marginal FREX: parenthood, shut, shutdown, agreement, abort, tea,
cancer, rider, boehner, awar
##       Marginal Lift: cervic, addon, awar, contracept, defund, libya,
distract, boehner, parenthood, bewar
##       Marginal Score: awar, hid, kennard, nonmedicar, shutdown, internet,
parenthood, shut, boehner, libya
##
##       Topic Kappa:
##       Kappa with Baseline:
##
```

```
sink("output/sageLabels-selected.txt", append=FALSE, split=TRUE)
print(sageLabels(First_STM,n=10))
```

```
## Topic 1:
##       Marginal Highest Prob: busi, fire, small, servic, nation, peopl,
state, also, plan, legisl
##       Marginal FREX: fire, forest, speci, interior, colorado, endang,
cleveland, steve, island, manag
##       Marginal Lift: buildup, roadless, wildemess, chairmani, gradual,
mall, prestigi, mid, ski, speci
##       Marginal Score: chairmani, hid, kennard, forest, madam, cleveland,
speci, celebr, ski, interior
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 2:
##       Marginal Highest Prob: secur, social, medicar, fund, republican,
tax, cut, year, drug, surplus
##       Marginal FREX: medicar, social, prescript, drug, trust, surplus,
senior, secur, raid, cbo
```

```
##       Marginal Lift: fiveyear, dav, doughnut, echo, fantasi, freefal,
giveback, hast, hid, horizon
##       Marginal Score: hast, hid, nonmedicar, medicar, prescript, drug,
social, raid, trust, surplus
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 3:
##       Marginal Highest Prob: cut, republican, program, billion, veteran,
million, nation, tax, year, care
##       Marginal FREX: slash, david, spratt, blueprint, veteran, wealthiest,
cut, valu, child, back
##       Marginal Lift: clever, hay, heath, oregonian, prioritiesa, scrip,
illadvis, toughen, sbas, skew
##       Marginal Score: hay, hid, nonmedicar, medicaid, david, spratt,
slash, veteran, cut, blueprint
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 4:
##       Marginal Highest Prob: govern, pass, republican, peopl, spend, shut,
today, that, shutdown, last
##       Marginal FREX: shutdown, shut, that, cant, yesterday, pass, adjourn,
govern, parti, reid
##       Marginal Lift: adjourn, zeppelin, unveil, reid, scout, freshmen,
circul, hemorrhag, shes, steni
##       Marginal Score: adjourn, hid, kennard, nonmedicar, shutdown, shut,
reid, cant, didnt, what
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 5:
##       Marginal Highest Prob: tax, spend, cut, balanc, blue, govern,
percent, growth, increas, dog
##       Marginal FREX: blue, dog, wast, growth, balanc, relief, tax, spend,
budget, dig
##       Marginal Lift: dilig, toomey, tricar, rsc, phasein, dog, blue,
undertax, ammunit, runaway
##       Marginal Score: dilig, hid, kennard, nonmedicar, tax, dog, blue,
toomey, spend, rsc
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 6:
##       Marginal Highest Prob: chair, financi, regul, market, requir, end,
swap, madam, risk, deriv
##       Marginal FREX: swap, deriv, chair, user, risk, exchang, madam,
```

```
market, exempt, financi
##       Marginal Lift: adler, cras, neglig, columbia, euronext, issuer,
kanjorski, liabl, mccarthi, nrsros
##       Marginal Score: columbia, kennard, swap, madam, deriv, bailout,
regul, collater, overthecount, cfpa
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 7:
##       Marginal Highest Prob: price, energi, gas, oil, bay, democrat,
chesapeak, countri, new, today
##       Marginal FREX: chesapeak, gas, ill, bay, oil, price, drill, gasolin,
watertrail, pump
##       Marginal Lift: continent, gag, americanmad, cellulos, gallup, ill,
illconsid, offlimit, pellet, ret
##       Marginal Score: hid, ill, kennard, nonmedicar, chesapeak, gas,
drill, oil, watertrail, bay
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 8:
##       Marginal Highest Prob: internet, fcc, open, regul, consum, rule,
innov, provid, compani, broadband
##       Marginal FREX: fcc, internet, broadband, fccs, innov, parliamentari,
disapprov, communic, content, open
##       Marginal Lift: comcast, deregulatori, dna, ebay, googl, netflix,
walden, wireless, amazon, ancillari
##       Marginal Score: hid, kennard, nonmedicar, parliamentari, internet,
fcc, broadband, fccs, regul, cra
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 9:
##       Marginal Highest Prob: war, support, women, men, troop, peopl,
countri, world, unit, nation
##       Marginal FREX: men, inquiri, love, soldier, women, mission, uniform,
hussein, gulf, son
##       Marginal Lift: chronicl, demean, gear, guidanc, inquiri, matthew,
proverb, speedi, teeth, energet
##       Marginal Score: hid, inquiri, kennard, nonmedicar, saddam, hussein,
men, war, love, women
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 10:
##       Marginal Highest Prob: addit, fund, appropri, million, earmark,
legisl, major, includ, provid, billion
```

```
##       Marginal FREX: addit, earmark, omnibus, obey, appropri, staff,
contain, hous, subcommitte, research
##       Marginal Lift: headquart, meth, withhold, airdrop, explanatori,
nabor, addit, archiv, chastis, darfur
##       Marginal Score: addit, hid, kennard, nonmedicar, earmark, obey,
omnibus, subcommitte, formula, fund
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 11:
##       Marginal Highest Prob: peopl, job, greek, state, mani, greec,
countri, independ, year, nation
##       Marginal FREX: greek, greec, nay, independ, ireland, ancient, polic,
celebr, nuclear, northern
##       Marginal Lift: gothic, ancestor, britain, coffer, colonist, cyprus,
fledgl, greek, greekamerican, hellen
##       Marginal Score: hid, kennard, nay, nonmedicar, greek, greec,
ireland, turkish, celebr, ancient
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 12:
##       Marginal Highest Prob: fiscal, new, author, outlay, michigan, vote,
res, concurr, consider, year
##       Marginal FREX: michigan, outlay, author, fiscal, new, concurr,
consider, roll, detain, vote
##       Marginal Lift: michigan, rolical, nos, poster, outlay, detain,
appro, herebi, author, propriat
##       Marginal Score: hid, kennard, michigan, nonmedicar, outlay, fiscal,
author, detain, rolical, new
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 13:
##       Marginal Highest Prob: tax, busi, small, ask, day, thing, pay,
peopl, like, person
##       Marginal FREX: ask, code, person, owner, sell, small, sometim, busi,
death, ration
##       Marginal Lift: dispens, lehigh, patienc, comedian, imposit, layer,
lobbyist, perkiomen, dread, outhous
##       Marginal Score: dispens, hid, kennard, nonmedicar, tax, suspens,
quorum, restaur, farm, sell
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 14:
##       Marginal Highest Prob: fund, program, need, educ, provid, increas,
```

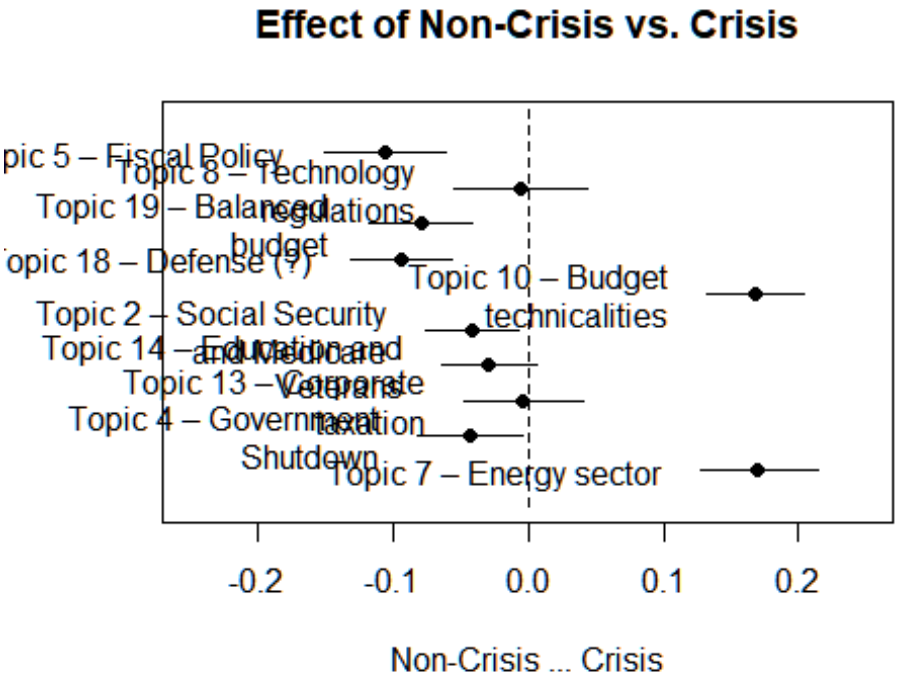
```
veteran, nation, support, billion
##       Marginal FREX: frank, child, educ, highway, program, grant, fund,
behind, cbc, black
##       Marginal Lift: postur, frank, elev, inmat, pertain, selfsuffici,
socioeconom, americorp, reintegr, dismay
##       Marginal Score: frank, hid, kennard, nonmedicar, cbc, fund, educ,
highway, teacher, veteran
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 15:
##       Marginal Highest Prob: support, war, troop, famili, nation, iraq,
men, women, presid, forc
##       Marginal FREX: brave, tonight, saddam, god, prayer, command, arm,
bless, troop, hussein
##       Marginal Lift: commenc, disrespect, routin, sympathi, tyrant,
unanim, hampshir, tile, aleadi, badger
##       Marginal Score: hampshir, hid, kennard, nonmedicar, saddam, hussein,
prayer, iraq, tonight, gratitud
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 16:
##       Marginal Highest Prob: medicaid, care, health, cut, state, program,
peopl, billion, million, cost
##       Marginal FREX: medicaid, illinoi, coverag, uninsur, nurs, care,
health, hospit, insur, vulner
##       Marginal Lift: awak, bottl, illinoi, payer, rangel, renaiss, tort,
uncompens, threequart, gold
##       Marginal Score: hid, illinoi, kennard, nonmedicar, medicaid,
uninsur, coverag, medicar, rangel, prescript
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 17:
##       Marginal Highest Prob: financi, fund, job, peopl, billion, spend,
million, democrat, republican, includ
##       Marginal FREX: dealer, financi, yea, bailout, tarp, loan, system,
arbitr, manufactur, fed
##       Marginal Lift: franchis, statutorili, yea, arbitr, chrysler, takeov,
fanni, freddi, yemen, hook
##       Marginal Score: hid, kennard, nonmedicar, yea, dealer, bailout,
tarp, arbitr, regul, madam
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 18:
```

```
##       Marginal Highest Prob: billion, spend, number, money, defens, year,
peopl, confer, pay, need
##       Marginal FREX: wisconsin, confer, defens, iowa, shell, number,
correct, game, billion, discretionari
##       Marginal Lift: bypass, flatlin, longrang, macroeconom, mug,
placehold, riverboat, tradeoff, wisconsin, committeereport
##       Marginal Score: hid, kennard, nonmedicar, wisconsin, shell, scare,
conceal, omb, chairman, baselin
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 19:
##       Marginal Highest Prob: plan, deficit, republican, veteran, cut, tax,
democrat, billion, care, spend
##       Marginal FREX: plan, texa, deficit, veteran, south, carolina, show,
trillion, substitut, surplus
##       Marginal Lift: dumb, hoyer, raw, sidelin, swing, principi, fascist,
icit, undemocrat, revert
##       Marginal Score: hid, hoyer, kennard, nonmedicar, veteran, deficit,
surplus, substitut, tax, spratt
##
##       Topic Kappa:
##       Kappa with Baseline:
##
## Topic 20:
##       Marginal Highest Prob: govern, spend, presid, cut, peopl,
republican, countri, need, shut, fund
##       Marginal FREX: parenthood, shut, shutdown, agreement, abort, tea,
cancer, rider, boehner, awar
##       Marginal Lift: cervic, addon, awar, contracept, defund, libya,
distract, boehner, parenthood, bewar
##       Marginal Score: awar, hid, kennard, nonmedicar, shutdown, internet,
parenthood, shut, boehner, libya
##
##       Topic Kappa:
##       Kappa with Baseline:
##

sink()

#Now that we understand topics, we can see if their prevalence differs based
on certain factors


#Specify a model that assesses the prevalence of topics
prep <- estimateEffect(c(1:20) ~ crisis * party,
                       First_STM,meta = meta, uncertainty = "Global")
saveRDS(prep, file = "data/prep_house.rds")
```

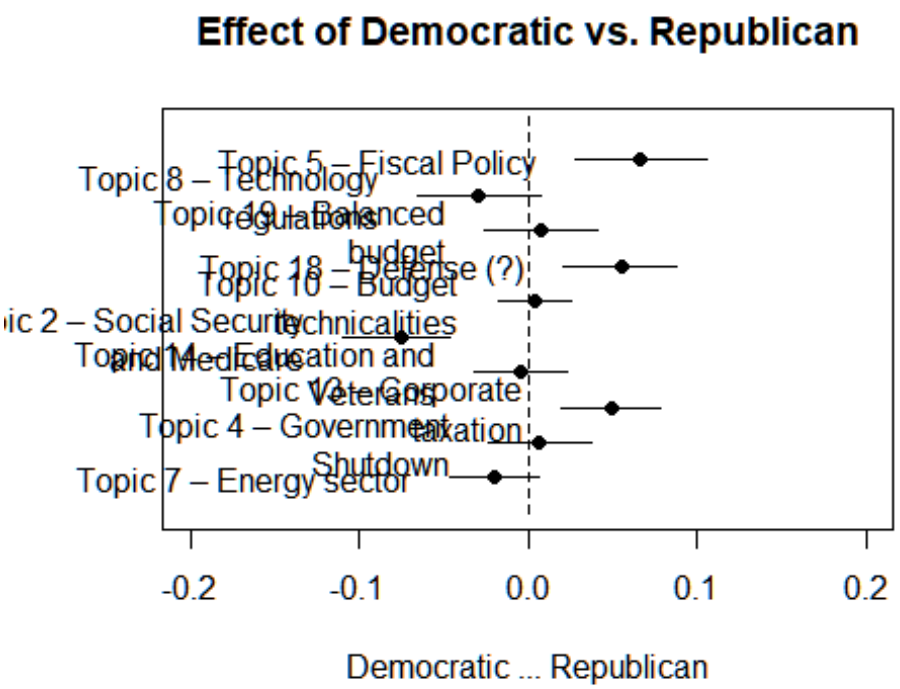
Analysis

```
#House's Debate Analysis
rm(list = ls())

#Install the stm package
#install.packages("stm")
#install.packages("igraph")

#Load the package
library(stm)
library(igraph)

library(ggplot2)
library(dplyr)

set.seed(2008)
#Load the data
setwd("C:/Users/scene/Documents/R/GIRI/project")
First_STM <- readRDS(file = "data/First_STM_house.rds")
prep <- readRDS(file = "data/prep_house.rds")
processed <- readRDS(file = "data/processed_house.rds")

#Now we separate our data into neat bits for our analysis
out <- prepDocuments(processed$documents, processed$vocab, processed$meta)

## Removing 5328 of 10375 terms (5328 of 81833 tokens) due to frequency
## Removing 2 Documents with No Words
## Your corpus now has 846 documents, 5047 terms and 76505 tokens.

docs <- out$documents
vocab <- out$vocab
meta <-out$meta
meta$year <- as.integer(meta$year)

topic_number = c(5,8,19,18,10,2,14,13,4,7)
topic_labels = c("Topic 5 – Fiscal Policy",
                 "Topic 8 – Technology regulations",
                 "Topic 19 – Balanced budget",
                 "Topic 18 – Defense (?)",
                 "Topic 10 – Budget technicalities",
                 "Topic 2 – Social Security and Medicare",
                 "Topic 14 – Education and Veterans",
                 "Topic 13 – Corporate taxation",
                 "Topic 4 – Government Shutdown",
                 "Topic 7 – Energy sector" )

#Now we plot the effect of CRISIS
Sys.setlocale(locale="en_US.UTF-8")

## [1] "LC_COLLATE=en_US.UTF-8;LC_CTYPE=en_US.UTF-8;LC_MONETARY=en_US.UTF-
8;LC_NUMERIC=C;LC_TIME=en_US.UTF-8"
```

```
plot(prep,
     covariate = "crisis",
     topics = topic_number,
     model = First_STM, method = "difference",
     cov.value1 = 'crisis',
     cov.value2 = 'non-crisis',
     xlab = "Non-Crisis ... Crisis",
     main = "Effect of Non-Crisis vs. Crisis",
     xlim = c(-.25, .25),
     labeltype = "custom",
     custom.labels = topic_labels
     )
```




```
#Now we plot the effect of PARTY
plot(prep,
     covariate = "party",
     topics = topic_number,
     model = First_STM, method = "difference",
     cov.value1 = 'R',
     cov.value2 = 'D',
     xlab = "Democratic ... Republican",
     main = "Effect of Democratic vs. Republican",
     xlim = c(-.2, .2),
     labeltype = "custom",
     custom.labels = topic_labels
     )
```




```r
#Trend of top topics
topic_distributions <- First_STM$theta
topic_distribution_with_year <- cbind(meta$year, topic_distributions)
topic_distribution_df <- as.data.frame(topic_distributions)
topic_distribution_df$year = meta$year

for (i in seq_along(topic_number)) {
  topic_index <- topic_number[i]
  topic_label <- topic_labels[i]
  yearly_distribution <- topic_distribution_df %>%
    group_by(year) %>%
    summarise(Average = mean(get(paste0("V", topic_index))))
  p <- ggplot(yearly_distribution, aes(x = year, y = Average)) +
    geom_line() +
    geom_point() +
    scale_x_continuous(breaks = yearly_distribution$year) +
    labs(x = "Fiscal Year", y = "Average Topic Distribution",
         title = paste("Trend of", topic_label)) +
    theme_classic()
  ggsave(filename = paste0("t", topic_index, ".png"), plot = p, width = 6,
height = 4)
}

#Speech Count by States
meta_count <- meta %>% count(state) %>% filter(n > 15) %>% arrange(desc(n))
ggplot(meta_count, aes(x = reorder(state, n), y = n)) +
  geom_bar(stat = "identity") +
  labs(x = "State", y = "Speech Count (>15)") +
```

```
  coord_flip() +
  theme_classic()
```