WOMEN IN CONGRESS: AN ANALYSIS OF TOPIC DIFFERENCES BETWEEN GENDERS THROUGH NATURAL LANGUAGE PROCESSING

A Dissertation
Presented to
The Academic Faculty

By

Julia Barnett and Maia Brenner Stainfeld

In Partial Fulfillment
of the Requirements for the Degree of
Master in Data Science

Supervisors: Christian Fons-Rosen and Hannes Mueller

Barcelona Graduate School of Economics

July 2019

"No country can ever truly flourish if it stifles the potential of its women and deprives itself of the contributions of half its citizens."

Michelle Obama

TABLE OF CONTENTS

Abstract

The purpose of this paper is to discover the differences between men and women regarding the topics discussed in the United States Congress. This paper explores both the distribution of topics discussed among men and women in terms of frequency as well as how those topics are discussed from a content perspective. In order to do so, Natural Language Processing techniques have been applied to a unique data set which comprised of the daily records of the United States Congress from the 97th session (1981-1983) to the 114th session (2015-2016). A Structural Topic Model with more than 2,500,000 documents have been estimated identifying that, even when controlling for the party, chamber, year, state and whether the candidate won in a close election, Congresswomen talk significantly more about Health, Budget/Taxes, Infrastructure and Technology, Emergency Relief/Refugees, Foreign Governments, Agriculture/Water, Election Reform, Human Rights, and Crime, while Congressmen talk significantly more about The Constitution, War, The Economy/Natural Disasters, Funding of Departments and National Parks. Furthermore, within topics women are more likely to focus on how the legislation they are discussing will affect citizens and individual lives, whereas men focus on the general discussion and administrative aspect of the legislation. These differences appear to be consistent over the past four decades.

Ultimately this paper confirms the need for a more representative governmental body in terms of the distribution of genders because women and men discuss different topics and discuss them in significantly different ways.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Sexism in the Highest Offices

The Global Gender Gap Index prepared by the World Economic Forum quantifies how great the gender disparity is in 149 countries and the progress made over time with a large focus on the areas of health, education, economics, and politics (World Economic Forum, 2016) [31]. Of these four major categories, they have found that the area in which the gender gap remains the largest is Political Empowerment; as of 2018 only 23 percent of this gap towards equality has been bridged with stagnant growth from the previous year and no country succeeding in fully closing this chasm. The United States of America boasts of being the land of the free, but men are the only ones who can truly live up to this title in the highest office.

Even further, when women make it to the highest offices, they are not able to talk as much as men are. A study done by Victoria L. Brescoll (2011) [5] at Yale University found that though there exists a strong positive correlation between power level in the Senate and amount of words spoken for men, no such relationship exists for women. There have also been many studies done on men interrupting women while they are speaking. Hancock (2014) [12] found that men were 33 per cent more likely to interrupt a woman than a man, and even further Jacobi and Schweers (2017) [13] found that this problem is even more pronounced at the highest level of the government. They found that even though women have only accounted for 24 percent of the Supreme Court seats over the last 12 years, they have been the recipients of over 32 percent of the interruptions, and only the cause of 4 percent. This problem is even getting worse over time: in 1990 the Supreme Court had only one woman (out of nine seats) who received 35.7 percent of the interruptions, in 2002

there were two women (out of nine) who received 45.3 percent of the interruptions, and in 2015 there were three female justices (one third of the Supreme Court) who received 65.9 percent of all interruptions (Jacobi and Schweers 2017). This phenomenon is something that has been a driver of women's behavior in positions of power; women are forced to be more judicious with their words and make their points in a more concise manner because it is likely that they will be interrupted before they are able to finish.

Beyond the lack of equal representation being an issue in and of itself, there are further benefits to be gained when more women are in congress. First of all, studies such as Chattopadhyay R. (2004) [7] have been done across the world that show women are more likely than men to implement policies that are more closely aligned with women's values, yet women are no less sensitive to complaints from their male constituents as they are their female constituents. In addition, Besley et al. (2017) [1] have found that with more women in politics, the competence of their male colleagues rises significantly in both the immediate and long term through both a removal of mediocre male leaders and improvement in future candidates. In addition to these positive impacts of having more women empowered in a political system, women also discuss issues with different frequencies and content than men. We aim to understand how women and men differ in the topics they discuss in the United States Congress, both in terms of the frequency with which they discuss these topics and the content of how they discuss them.

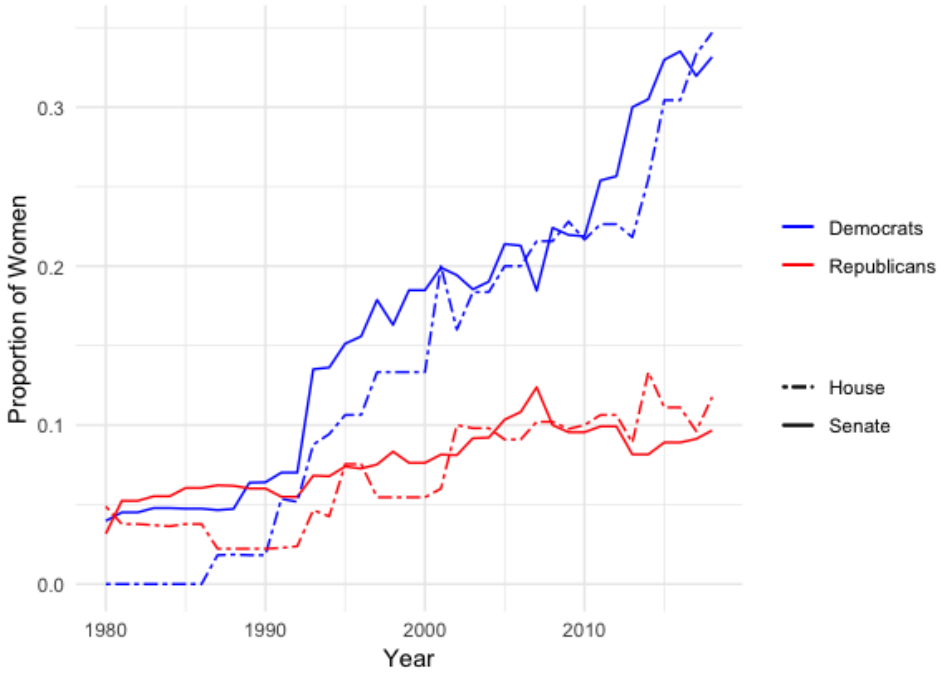
1.2 American Political System

The government in the United States of America is split into three coequal branches: the executive branch, the judicial branch, and the legislative branch. The legislative branch is comprised of the House of Representatives (further referred to as the House) and the Senate, which together form the U.S. Congress. Both halves of congress are designed to be equal and have different powers and responsibilities. The Senate is made up of 100 senators—two from each state. The House has 435 seats and is more representative of the

greater population of the U.S. because the seats are distributed to districts across the U.S. according to population (although each state is guaranteed one representative even if they do not have a large enough population to be eligible for one). All senators serve six-year terms while representatives serve two-year terms.

A defining nature of the United States political system has been the extreme partisanship amongst both constituents and politicians, a phenomenon that gets worse with each passing year. Out of the 535 seats in Congress today, only two seats are held by independents. The rest are either Democratic (the progressive/liberal party) or Republican (the conservative party), and it is important to note that the independents are treated as Democrats under official and practical means (i.e. they participate in Democratic primaries and typically vote as Democrats vote). Even further, one party certainly elects more women; while the Democrats have been steadily growing female representation over the past few decades to 33 percent in 2018, Republicans have consistently hovered around or below 10 percent (Figure 1.1).

Figure 1.1: Proportion of Women in Congress by Party and Chamber



1.3 Structure of This Paper

The remainder of this proceeds as follows: Chapter two explores relavant literature as to the differences in gender in terms of language, political representation and effects of having women in a legislative body. Chapter three describes the methodology of natural language processing and specifically gets into the theory behind a structural topic model. Chapter four explains the data sources utilized in this paper and how we preprocessed the data in order to make it analyzeable. Chapter five discusses how the model was estimated and choices made in fitting the model, followed by chapter six which contains all of our results and analysis. Finally, chapter seven concludes our findings and reiterates the importance of this analysis while chapter eight explores how this field could be studied further.

CHAPTER 2
LITERATURE REVIEW

2.1 Gender Difference in Language

Several papers have examined the gender differences in language by having a look at the actual words men and women use. Some papers claim that women use more intensive adverbs, more conjunctions (e.g. but, and), and more modal auxiliary verbs such as "could", "may", and "might" (Biber, Conrad, Reppen, 1998 [2]; McMillan et al., 1977 [17]; Mehl Pennebaker, 2003 [18]). According to Newman et al. (2008) [21], women use more words related to psychological and social processes whereas men refer more to object properties and impersonal topics.

In terms of the length of the interactions, the findings are more ambiguous. According to (Mulac Lundell, 1994) [20], women come out as the wordier gender both in writing and speaking. However, according to (e.g.,Dovidio, Brown, Heltman, Ellyson, Keating, 1988) [9] men use more words overall and take more turns in conversation. One possible explanation of these ambiguous results is that different contexts can clearly influence the way people talk and, in particular, women may speak differently in male dominated environments such as U.S. Congress.

2.2 Gender Difference in Congress

In particular there are a couple of papers which study specifically the gender differences in the U.S. Congress.

Lenard (2016) [14] studies the gender differences in the political speeches from the 113th U.S Congress. Lenard analyzes 672 speeches from female and 2983 speeches from men and performs one-way ANOVA and two tailed Spearman correlation tests in order

to observe if the differences are statistically significant. According to Lenard, the female politicians were shown to be more formal, critical, and task-oriented. Moreover, the female politicians focused on raising the awareness of different health issues and providing support for patients and their families while the male politicians focused on the consequences and possible solutions to the problems.

Contrarily, Yu (2013) [32] studied the Congressional speeches from the 101st to the 110th Congress (1989-2008). According to Yu, female legislators' speeches demonstrated characteristics of both a feminine language style (e.g. more use of emotional words, fewer articles) and a masculine one (e.g. more nouns and long words, fewer personal pronouns). A trend analysis found that these gender differences have consistently existed in the Congressional speeches over the past 20 years. The findings lend support to the argument that gender differences in language usage persist in professional settings like the floor of Congress.

Even though these papers contribute to understanding how men and female representatives talk in Congress, few research has been found on the different topics men and women address in Congress. In particular, we are interested in understanding the possible effects of having more women in power.

2.3 Effects of Women in Congress

Brollo (2012) [6] exploits a Regression Discontinuity Design (RDD) in close electoral races in Brazil and provides evidence that cities with female mayors have better health outcomes (prenatal visits and percentage of premature births), are awarded more federal discretionary transfers, and are less likely to have administrative irregularities in public procurement practices. Despite these results, male mayors are 20 percentage points more likely to be re-elected than female mayors.

According to the authors, one of the explanations that could potentially rationalize this result is that the extent to which politicians seek the support of special-interest groups

could be gender differentiated (Persson, Tabellini, 2000) [22]. While female mayors seem to put more effort into obtaining better prenatal care outcomes for their constituency, male mayors target special-interest groups of voters by increasing the size of temporary public employees of the municipality (Brollo 2012).

Moreover, Chattopadhyay (2004) [7] provides evidence of a randomized policy experiment in India where one third of Village Council head positions have been randomly reserved for a woman. This study suggests that the reservation of a council seat affects the types of public goods provided. Specifically, leaders invest more in infrastructure that is directly relevant to the needs of their own genders.

These outcomes would suggest that the topics women and men are concerned about are different. All in all, we add to this literature by analyzing a rich data set of U.S. Congressional records which accounts for more than 2,500,000 documents over 35 years and we provide new evidence on the role of gender in politics.

This paper seeks robust answers to the following research questions: (1) Do male and female legislators address different topics?, (2) in what ways do the content of discussions differ within these topics?, and (3) if different patterns are detected, are they consistent in Congressional speeches over the past 35 years?

CHAPTER 3

METHODOLOGY

3.1 Natural Language Processing

An ever-expanding field in data science is natural language processing (NLP), commonly referred to simplistically as ”text mining”. Currently, there are many different techniques available to analyze text and discover patterns in documents via automated procedures that would not have been possible just a couple of years ago. Manning and Schütze [15] stated in their book *Foundations of Statistical Natural Language Processing* that the availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. Phenomena that were previously undetectable or seemed uninteresting in terms of exploration have moved to the central focus of lexical analysis. In particular, NLP has gained considerable attention in the field of Political Science; it is regarded now as a powerful descriptive tool to analyze large bodies of text. Moreover, when combined with steadfast econometric approaches, new inferences can be discovered.

In the following sections of this paper we will introduce the theory for topic modeling approaches, specifically directing our focus to the structural topic model. Afterwards, we will explain how to estimate a structural topic model, analyze the results, and further confirm our estimations with a Fixed Effects Regression Model to observe whether our results are still significant when accounting for other variables. The explanation on why we decided to work with this model and the stm R package can be found in the Discussion section.

3.1.1 Topic Models

The key innovation that topic models bring to Natural Language Processing is that they allow lexical corpus to be automatically organized into meaningful groups entitled "topics" (Mohr et al. 2013) [19]. In the data science field this would be considered an unsupervised problem since the data (in this case documents) do not have a specified label beforehand of which topic it falls under and the researcher's goal is to be able to categorize these observations into groups with similar characteristics without "telling" the model what the specific groups should be. The researcher can either pre-define the number of topics for the algorithm to uncover or use different criteria for accuracy measures to determine what the "optimal" number of topics is, but this could very well optimize consistency of topics while sacrificing interpretability. Regardless of the algorithm and specific criteria used, most topic models will return the probabilities of words being used in a certain topic and the distribution of those topics across the entire text corpus.

Thus, a topic might be thought of as an assortment of words that tend to come up in a discussion (and thus to co-occur more frequently together) whenever that unobserved and latent topic is being discussed. Note that topic models capture co-occurrences regardless of these words being naturally embedded within other complexities of language, such as syntax, narrative, or location within the text. Instead, each document is treated as if it were a so-called "bag of words". The goals of a topic modeling analysis are then to analyze these various bags of words, identify word co-occurrence patterns across the corpus of text, utilize these to produce a mapping of the distribution of words within the topics, and uncover how the topics themselves are distributed within the greater corpus of text.

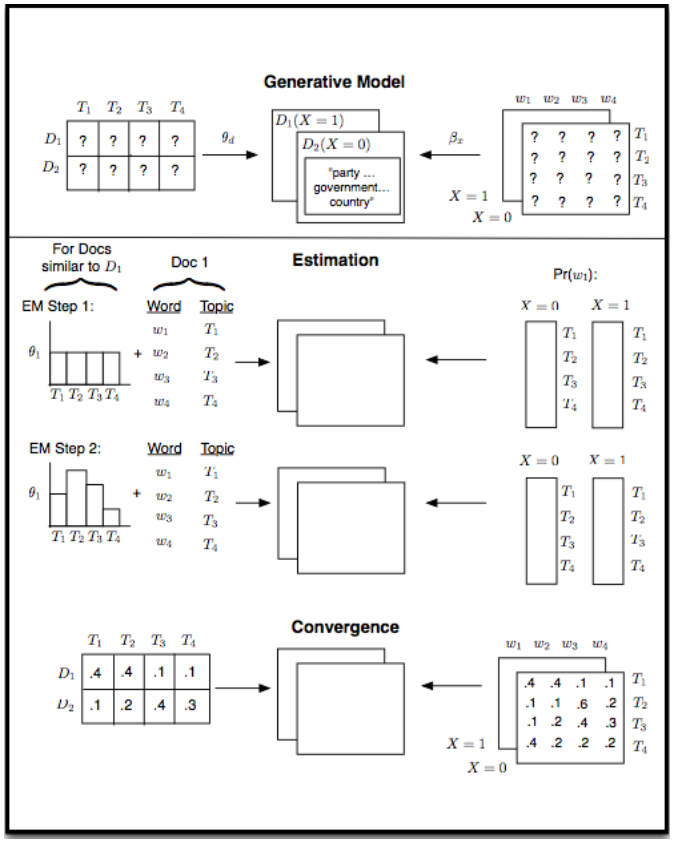
The simplest and most widely used model is Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003) [4]. LDA is a mixed membership topic model, meaning that each document is assumed to be a mixture of topics. After the model is fit to the text corpus of documents, each document will have a percent estimation of how much that document is expected to fall in line with different topics (e.g. document XYZ could be 30% topic

1, 20% topic 2, 0% topic 3, and 50% topic 4). The Structural Topic Model (STM) is a relatively recently introduced variant of LDA that grants the researcher the ability to analyze the lexical corpus via different covariates or metadata associated with each document (Roberts et al. 2013) [25]. In the Discussion section we further explain why we decided to work with the structural topic model and why we estimate it in R.

3.1.2 Structural Topic Model

As mentioned above, the innovation of the structural topic model as it relates to LDA is that is allows users to analyze the topics as they vary with *metadata*, or "data about the data", which is simply additional information about the document itself like the date of the document or who was speaking. The outputs of the model can be used to conduct hypothesis testing about the relationship between different metadata and the topics themselves (Roberts et al. 2013) [25].

Figure 3.1: Heuristic Representation of the Structural Topic Model [24]



Per figure 3.1 above, documents are represented by $(D_1, D_2...)$, topics by $(T_1, T_2...)$, words by $(w_1, w_2...)$, and different levels of metadata (e.g. Gender = F or M) are represented by X. Each word has a probability of belonging to a topic, and each topic has a different distribution of words. Moreover, each document is a mixture of topics. In fact, we estimate the proportion of all topics for a given document which must sum up to one (no change from the general LDA) and similarily, the sum of word probabilities for a given topic must also sum to one.

A key innovation of the STM is the incorporation of the topical prevalence and topical content covariates (Roberts et al. 2016) [24].

- ***Topical prevalence***: how much each topic contributes to a document
- ***Topical content***: how a particular topic is discussed; word rate usage

The model and the R package associated allows for the usage of (a) topical prevalence covariates, (b) topical content covariates, (c) both, or (d) neither. In the case of no covariates, the model reduces to a (fast) implementation of the Correlated Topic Model (Blei and Lafferty 2007) [3]. In our estimation we found that estimating with both topical and content covariates allowed for the most in-depth analysis, even though it was highly costly in terms of the extra computing time necessary.

Each document (D) has a vocabulary defined as the number of unique words present (V). The generative process for a STM model with k topics as described by Roberts (2016) [24] is as follows:

1. Obtain the document-level distribution of each topic within a singular document from a logistic-normal generalized linear model based the chosen document covariates (both content and prevalence) X_d:

$$\theta_d | X_{d\gamma}, \Sigma \text{ LogisticNormal}(\mu = X_{d\gamma}, \Sigma) \tag{3.1}$$

where X_d is a p-by-1 vector (p being the number of levels of covariates), gamma is a p-by-k matrix of coefficients (k being number of topics), and Sigma is a (k-1)-by-(k-1) covariance matrix.

2. Generate the distribution of words representing each topic (k) within each document from the baseline word distribution (m), the topic specific deviation κ_k, the deviation between the covariate groups κ_g and the interaction between the the topic spcific deviation and the covariate deviation κ_i.

$$\beta_{d,k} \propto \exp\left(m + \kappa_k + \kappa_{g_d} + \kappa_{i=(k,g_d)}\right) \tag{3.2}$$

   Each of the above vectors are the length of the vocabulary (V) with one entry per appearance of a word in the vocabulary.

3. Lastly for each word in the document, ($n \in 1, ..., N_d$): obtain the word's mixed-topic assignment from the document-specific distribution over topics:

$$z_d|\theta_d \text{ Multinomial}(\theta_d) \tag{3.3}$$

   and conditional on the topic chosen, draw an observed word from that topic:

$$w_{d,n}|z_{d,n}, \beta_{d,k=z_{d,n}} \text{Multinomial}(\beta_{d,k=z_{d,n})} \tag{3.4}$$

In order to obtain the model parameters, the stm package implemented in R estimates the model with a partially-collapsed variation of the Expectation-Maximization (EM) algorithm. Theoretically, convergence would give the optimal model parameters, however with massive amounts of data, convergence is much too computationally costly to achieve without the assistance of a supercomputer. Regularizing prior distributions are used for γ x, and (optionally) Σ, which help enhance interpretation and prevent overfitting. Further details on the estimation procedure can be found in Roberts et al. (2016) [24].

In a non-technical summarization, after defining the number of topics and the prevalence and content covariates of interest we can run the model and observe the estimated

parameters. Once the model is estimated we obtain a theta matrix (Documents D x Topics k) which corresponds to the proportion of each topic that can be observed in each document. This theta coefficient can be interpreted as a dependent variable and estimate a regression model on all of the covariates.

The stm package in R has an estimateEffects function that allows the user to estimate the effects of the relevant document metadata (covariates) on the topic proportions of the documents. The documentation of the stm package in R stipulates that the regression is run where the documents are the units of analysis, the output is the estimated proportion of each document dedicated to a specific topic estimated via the STM model, and the covariates are the specified document metadata.

In particular, we encourage the reader to read the Online Appendix of Roberts et al. (2016) [27] in order to understand the details of the estimation procedure, in particular how the model accounts for uncertainty through the method of composition. The fitted model estimates effects of the covariates on the topic proportions (thetas). To do so a sample is generated from each document's approximate posterior as $\widetilde{\eta_d} \sim N(\lambda_d, \upsilon_d)$ and then converted to the simplex $\widetilde{\theta}_d = \frac{exp((\widetilde{\eta}_d)}{\sum_k (exp(\widetilde{\eta}_d)}$. Following this, the standard regression model is calculated and simulates trials from the regression's posterior 100 to 1000 times to assure robustness. Then the results are combined to form an approximate posterior which already has present the bounds of uncertainty (confidence intervals). The R package provides the user the ability to both simulate from the approximate posterior over θs and calculate uncertainty under common regression models [26].

CHAPTER 4

DATA

4.1 Data Sources

The primary source of data used in this analysis is the daily records of the United States Congress from the 97th session (1981-1983) to the 114th session (2015-2016). The speeches were made available to the public by HeinOnline and scraped and parsed by Gentzkow, Shapior, and Taddy at Stanford University in 2018 [11]. Parsing consisted of separating the speeches so that the document-level analysis consists of every speech made by one speaker on a given day. This way we can analyze as granularly as each speaker at the daily level, in addition to many separate ways of aggregation. The metadata of the speeches consisted originally of chamber (House or Senate), date, and order of the speeches in a given day. Further cleaning has added important variables for analysis to the metadata such as gender, first name, last name, and state, which allows us to perform the analysis on gender in addition to controlling for other variables that may influence differences among Congressmen and Congresswomen.

We have added an extra source of data in order to control for an additional variable and analyze gender differences while holding another factor constant: whether a Congressperson won with a close margin (as defined by winning with less than a five percent margin; built with data from the MIT Election Data + Science Lab (2019) [29]). This variable and other covariates already present were used for the purpose of affirming that gender is the root cause of the differences in topics, not gender masking itself as a product of other causal variables. In addition to adding this source to the data set we also did our own validation checks on whether the data set was encoded correctly for the metadata covariates.

4.2 Data Preprocessing

In order to make the data analyzable, a series of tasks are performed on the raw data. First, the speeches are converted to lowercase so that capital letters are treated the same as lowercase. Second, punctuation is removed because it typically adds unnecessary noise to the model. Third, *stopwords*, defined as unimportant words that are overly common (e.g. "the", "and", and "is") are removed utilizing a freely available SMART stop word list built by Salton and Buckley and sourced by the online appendix 11 of Lewis et al. (2004) [8]. Fourth, numbers are removed because, similarly to punctuation, they simply add noise to the model. On the words remaining in the corpus, *stemming* is performed using Porter's stemming algorithm developed by Rijsbergen et al. (1980) and Porter (1980) [23]. Stemming is the act of reducing words to their root form (e.g. "voting", "voted", and "vote" all become "vot"). This allows the model to treat these words in the same manner rather than as separate ideas. Finally the corpus is *tokenized* in order to create a vocabulary of unique terms present in the entire body of data. The data as its original input source is simply a mass collection of individual characters; tokenizing the data allows characters separated by white space (i.e. words of a sentence) to be treated as the individual inputs rather than a random string of characters (Vijayarani and Janani 2016) [30].

After this preprocessing is completed, the final result of the data is no longer lexical. Instead, documents (which in this case is everything one Congressperson said on a given day) are vectors of tokenized word counts. Each document is transformed to a vector of word frequencies rather than the raw text; each document has a unique distribution of words because no two speeches are made of the exact same substance. The vectors will be sparse (i.e. mostly zeroes), because the vocabulary is far greater than the words used in a given document so the word counts for most words will be zero.

After the data is completely processed it is stored as three separate variables: documents, vocabulary, and metadata. Documents are vectors of word counts that each Con-

gressperson said on a given day (2,576,458 in total). Vocabulary is the list of unique words used in the corpus (72,203 words). Metadata is the "data about the data" or in other words the additional data we have about the speech (gender, date, state, chamber, etc.).

CHAPTER 5
ESTIMATION OF THE MODEL

We then fit the structural topic model utilizing the documents and vocabulary to uncover the topics represented in the corpus. There is no "right" number of topics, so we tried different numbers until we found one that produced the most interpretable result (40 topics). We then defined the prevalence formula to include as covariates the gender of the Congressperson, chamber of Congress in which he or she resides (House or Senate), state that the congressperson represents, party to which the Congressperson belongs (Democratic or Republican), and date the speech was made. Specifying these covariates in the prevalence formula allows us to observe how these specified metadata affect the frequency a topic is discussed (e.g. we can see if women talk more about health than men).

Additionally, we set the initialization of the model to use a spectral algorithm, as opposed to random initialization or Latent Dirichlet Allocation (LDA) initialization. In order to uncover the model parameters from the data, the spectral algorithm uses a decomposition non-negative matrix factorization (similar to Singular Value Decomposition but restricts all values to be non-negative) to recover the model parameters from the word co-occurrence matrix containing the distribution of words in different topics to generate the parameters (Roberts et al. 2015) [16]. This is especially useful because the initialization is deterministic and can be replicated without randomization influencing differing results.

After fitting this model, we estimate the effect of gender on the distribution of topics. It is important to note that even though we estimated 40 topics, a large proportion of these "topics" are not substantive. Instead, these topics are a catch-all for procedural terms that Congress uses to address the floor, each other, the chairmen, official titles, and many other "pomp and circumstance" matters. We did not thoroughly analyze these topics but instead focused more on the topics that we were able to identify as substantive discussions on key

issues (e.g. health).

Finally, after fitting the STM we obtain the theta matrix (D x k; documents x topics) matrix. This is to say that for each document we have the estimated proportion for each topic. After analyzing the observed results, we then use the estimated thetas as the dependent variable in our fixed effects linear regressions. The goal of this estimation is to see whether the gender (β_1) is still significant when controlling for other variables of interest. It is important to highlight that we are not going to state any causal inference nor we will analyze the absolute magnitude of the β_1 coefficient since the magnitude of the topic proportion depends on the amount of topics we define.

We estimate for each given topic the following linear regression:

$$\begin{aligned} Y_{ist} = \alpha &+ \beta_1 Gender + \beta_2 Party + \beta_3 CloseElection \\ &+ \beta_4 Chamber + \beta_5 Gender * Party \\ &+ \lambda_t + \delta_s + \mu_{st} \end{aligned} \tag{5.1}$$

Where Y_{ist} is equal to the topic proportion of document i in time period t from speaker from state s; $Gender$ (β_1) is a binary variable that takes the value one for the documents generated by male speaker while takes value zero if the document corresponds to a female speaker; $Party$ (β_2) is a binary variable taking the value one when the speaker is Republican and zero otherwise; $CloseElection$ (β_3) is a binary variable that takes value one if the speaker was elected that given year by a close margin (less than 5% difference in the number of votes); $Chamber$ (β_4) is a binary variable that takes value one if the speech corresponds to the Senate and zero if in the House; $Gender * Party$ (β_5) is an interaction term; λ_t denotes time (year) fixed-effects, δ_s denotes state fixed effects and μ_{ist} denotes the error term.

CHAPTER 6

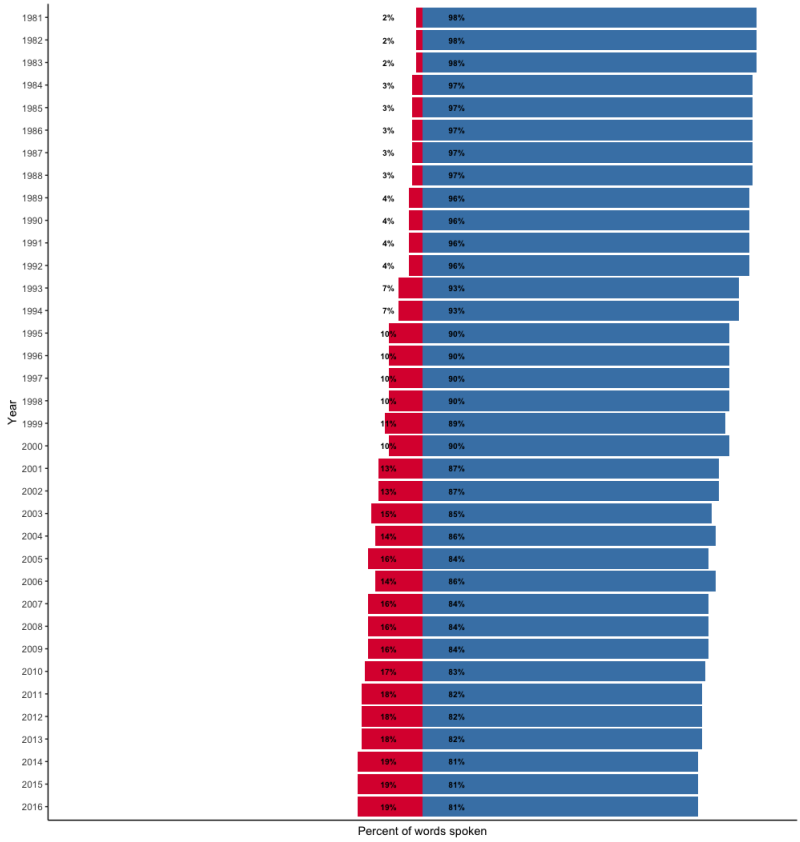
RESULTS

6.1 Introduction

First we will look at the general descriptive statistics: the relative word count of men and women in Congress over time and the frequency with which men and women discuss different topics over the entirety of the period for which we have data (1981-2016). Then we will explore more in depth how often men and women discuss topics, both in absolute terms and over time. Then we will focus on the actual content differences of men and women within a topic. Finally we estimate a Fixed Effects Regression Model to confirm that our findings are consistent.

6.2 Descriptive Statistics

Figure 6.1: Percent of Words Spoken by Gender Over Time

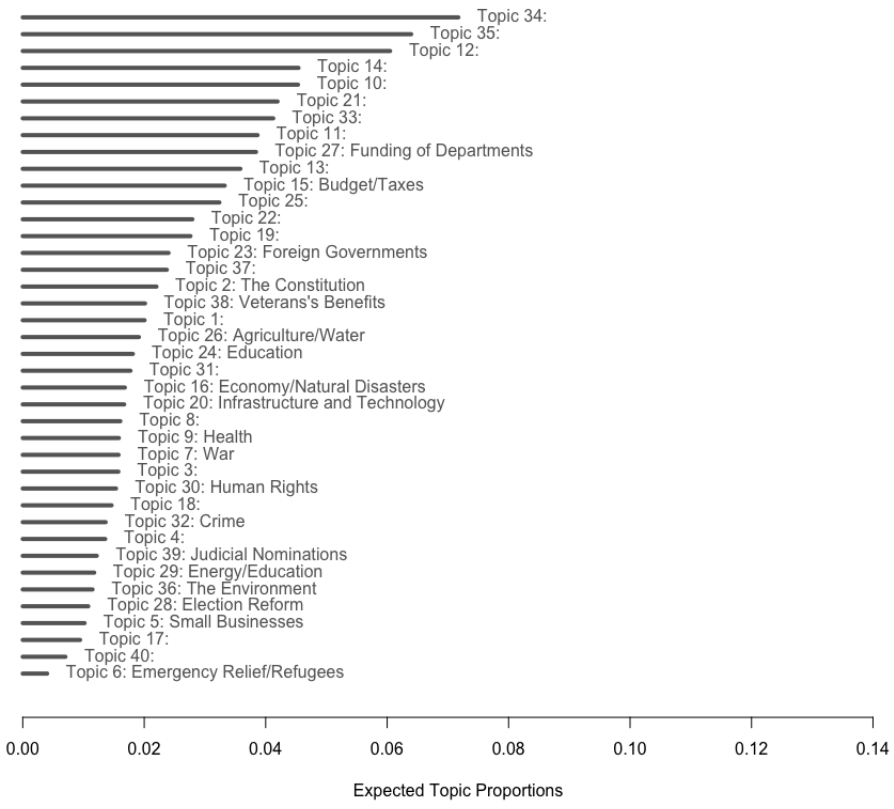


Though the amount of women in Congress has been steadily increasing (still to drastically lower than equal representation with men), the amount of words spoken in congress by different genders has faced an even slower decline towards equality.

As mentioned above, there is a positive correlation between position of power in the Senate and words spoken for men that simply does not exist for women. This adds a new understanding to the discrepancy between the proportion of words spoken by gender and membership in congress by gender: when men gain power they talk even more but women are not afforded the same luxury. Furthermore, as mentioned above studies have shown that women are incredibly more likely to be interrupted than men (Jacobi and Schweers 2017) and (Hancock 2014). This phenomenon could partially explain the disproportionate amount of words spoken in congress by gender because women are not able to say as much as men are before they get cut off.

6.3 Topic Frequency

Figure 6.2: Mean Proportion of Topics Discussed

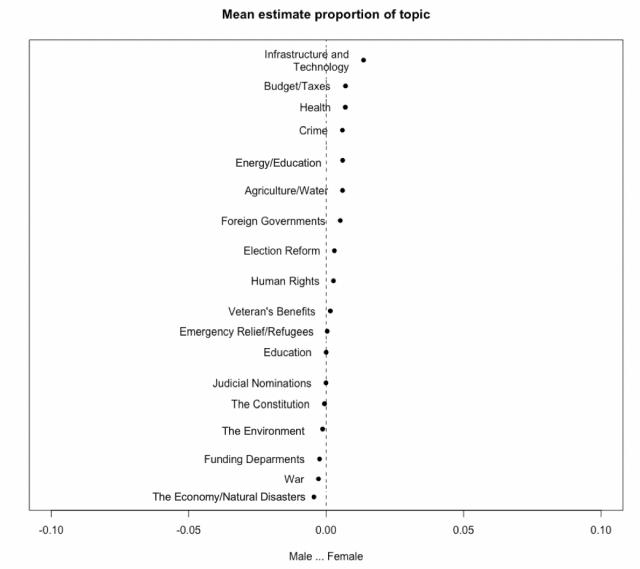


The topics are manually labeled by examining a combination of the words that had the highest probability of appearing in a topic and the *FREX words* of a topic. FREX words are those that are not only highly prevalent in a topic but also exclusive to the topic of interest; they are calculated with a formula created by Roberts et al. (2016) [24] and utilized in the stm R package. The words with the highest probability and the FREX words of every topic estimated can be found in the appendix in table A.1.

Of the forty topics our model estimated, we found 18 to be interpretable and analyzable. The remaining 22 unlabeled topics were either (a) procedural Congressional syntax (e.g. Topic 13: voting on bills with specific syntax yea and nay), (b) a topic that covered too many ideas to be easily analyzable for a specific subject (e.g. Topic 37 that discusses Democrats and Republicans and also home foreclosures), (c) filler words that did not constitute any substantive topic but rather were just used to fill time in a sentence (e.g. Topic 34), or (d) topics that were not interpretable (e.g. Topic 1 mostly grouping names of states). A key insight here is that the majority of the topics discussed in Congress (in terms of word counts devoted to this topic) is actually devoted to topics that are not substantive or dedicated to productive topics but rather to filler or procedural words. The topics we analyzed can be seen in figure 6.2. Of the analyzable topics, the three that had the highest prevalence in the corpus were Funding of Departments, Budget/Taxes, and Foreign Governments.

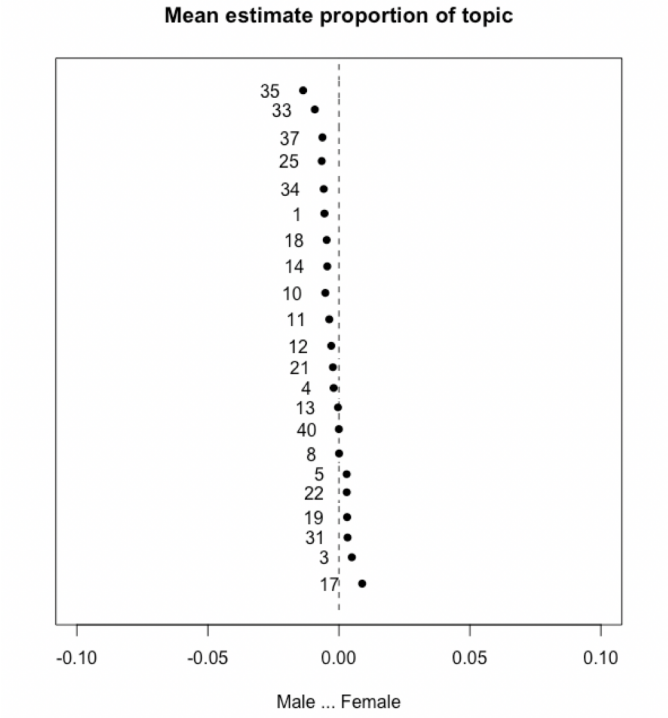
As can be seen in figure 6.3, of all the topics estimated, women talk significantly more than men about "Infrastructure and Technology", "Budget/Taxes", "Health", "Crime", "Energy/Education", "Agriculture/Water", "Foreign Governments", "Election Reform", "Human Rights", "Veterans Benefits", and "Emergency Relief/Refugees". Men on the other hand talk more than women about "The Economy/Natural Disasters", "War", "Funding Departments", "The Environment", and "The Constitution". "Education" and "Judicial Nominations" did not have either gender talk significantly more one way or the other.

Figure 6.3: Gender Differences of Mean Proportion of Interpretable Topics



As can be seen in the following figure 6.4, men talk about non-substantive topics at much greater rates than women do. The fact that women understand the phenomenon that they get interrupted more than men (Jacobii and Schweers 2017) could also drive the way they formulate their points; women spend their time using words that are directly related to the topic and content at hand rather than wasting the time they are allotted to speak by using filler words and non-substantive topics. This could be one of the reasons that we see men speaking significantly more often than women on the vast majority of the topics deemed non-substantive. When analyzing even further topic 17 (the non-substantive topic that women speak more about than men with the greatest margin), the words that comprise this topic tend to be “note”, “discuss”, “suggest”, “might”, “perhaps” (See table A.1 in the Appendix). These words are used in English to soften a belief or statement someone may have; women have been conditioned to phrase their sentences in a less aggressive or direct way than men. This also supports the findings mentioned above that women use more modal auxiliary verbs in language (McMillan et al., 1977) [17].

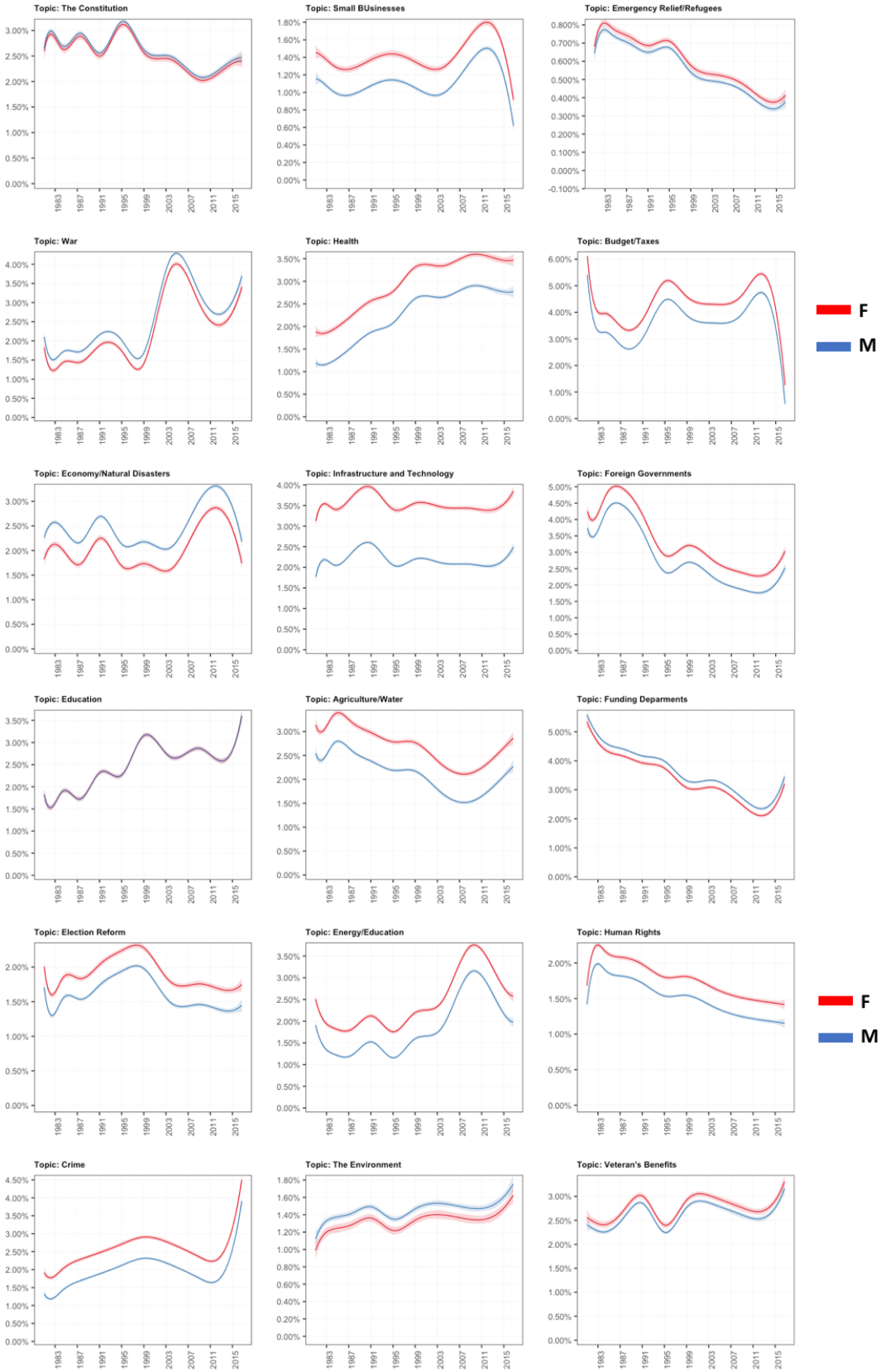
Figure 6.4: Gender Differences of Mean Proportion of Non-Substantive Topics



In the appendix in table A.5 the mean proportion of topics discussed over time by gender are listed for each topic for the total corpus, exclusively men, and exclusively women. The final column of this table is the most useful because it contains the percent difference of how much more (or less) likely a woman is to talk about a certain topic than men are. When analyzing at this level, we can see that women are 83.5% more likely to talk about health than men, 82.5% more likely to talk about infrastructure and technology, and 67.8% more likely to talk about crime than men. Men, on the other hand, are incredibly more likely to talk about the topics deemed non-substantive (the greatest differentials of these categories are 54.2% (topic 14), 37.3% (topic 8), and 34.5% (topic 12)). Within the labeled topics, men are 20.5% more likely to discuss the funding of departments, 14% more likely to discuss judicial nominations, and 9.5% more likely to talk about the Constitution.

This information is incredibly useful for voters. For example, a recent poll by an analytics firm with a focus on political events (Gallup 2018)[10] found that 80% of voters

Figure 6.5: Evolution of Topic Discussion over Time by Gender



consider healthcare to be either “extremely important” or “very important”; since women are 83.5% more likely to talk about healthcare than men, these voters would benefit from having more women in Congress.

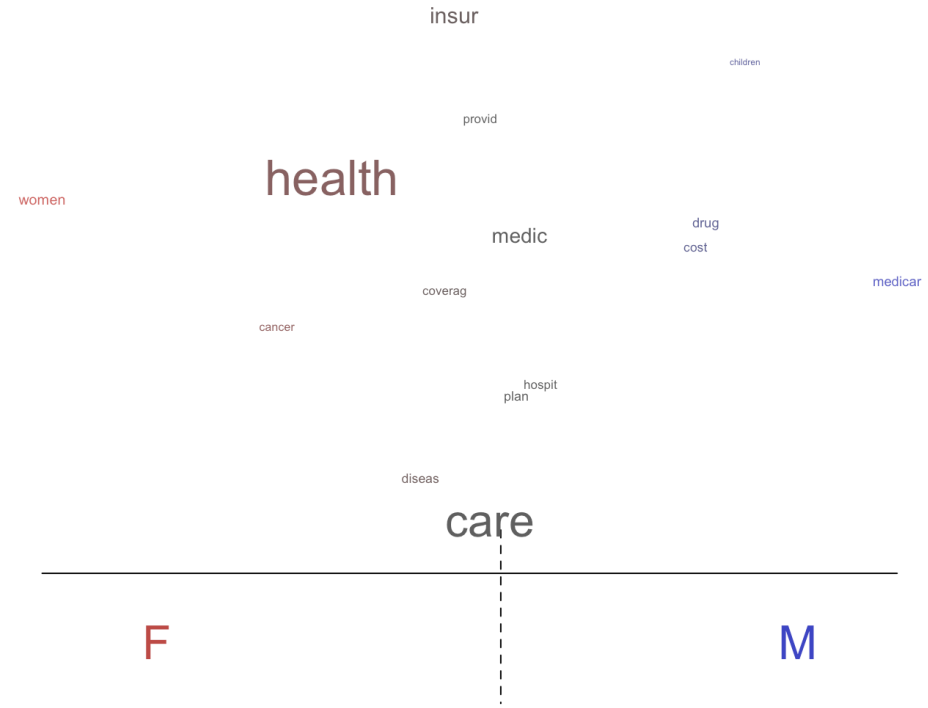
Even further, we can observe that these differences between the mean proportion dedicated to a topic by men and women remain relatively consistent over time [Figure 6.5]. The percentages are of the whole words spoken by both genders, so the differences observed are relative (i.e. they already control for the disparity between the number of women and men in congress). The most fascinating insight from plotting the relevant topics over time by expected proportion of discussion by Congressmen and Congresswomen is that the differences between expected proportion of topics does not face much variability over time. This means that for any given topic, regardless of external events that cause a topic to spike in discussion, the differences in how often that topic is brought up by either gender remains mostly constant throughout time. For example, a spike occurs in “Agriculture/Water” in 1985, but women still spend the same amount more of their time on discussing the topic than men do, the same as when the topic fell to lower proportions in 2007.

6.4 Differences in Content of Topics

Finally, we go deeper into the data to see exactly how men and women discuss different topics in terms of content [all of these content plots can be found in Figure A.1 in the appendix]. The figures are plotted from Female to Male with the dashed line in the center representing when a word is not significantly more prevalent within either gender. The size of the word represents the frequency the word appears in the corpus, and the color/distance from center represents how specific to one gender that word is. We will now evaluate key topics individually.

6.4.1 Health

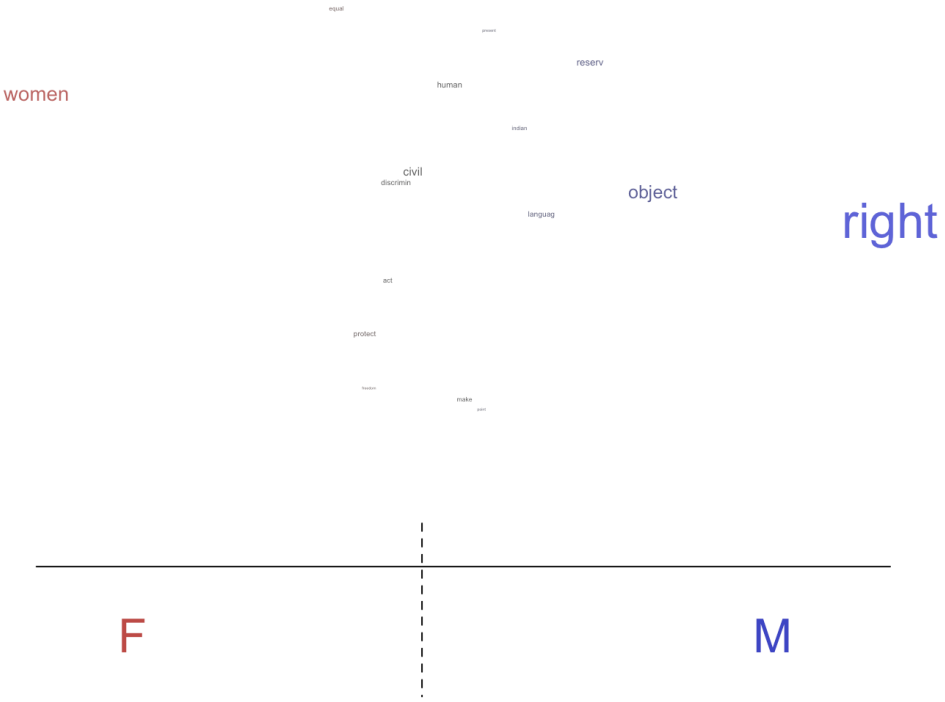
Figure 6.6: Content Differences: Health



When health is discussed in Congress, it is almost always to discuss healthcare. This is why it is no surprise that care is not significantly skewed to either gender—it is the primary focus of the discussion (similar to insurance, providers, coverage, [insurance] plans, and diseases). While men are more prone to discussing drugs and the costs associated with healthcare, women are more focused on the individuals affected by the legislation and approach the topic more from a health perspective rather than a monetary perspective. The biggest insight here is how far skewed to the left the word "women" is. Unsurprisingly, women talk about women's healthcare substantially more than men do. This is an important insight because women's healthcare is of the utmost proportion to over half of the country (over 50.8% of the U.S. population is female according to the latest census).

6.4.2 Human Rights

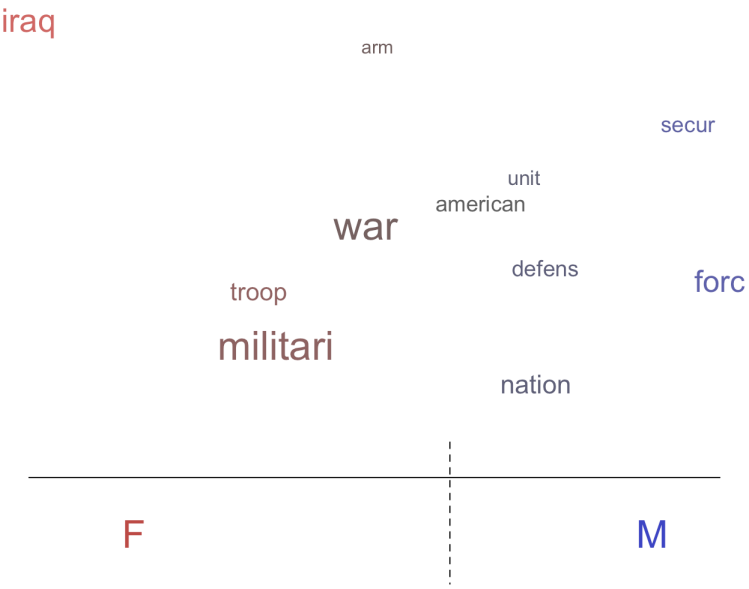
Figure 6.7: Content Differences: Human Rights



Similarly to health, when the topic is human rights, women are much more likely to discuss women's rights as its own unique and necessary subject whereas men are more likely to focus on rights as a more general concept as they apply to all people. Though it is necessary to focus on the rights of all of the citizens that a government represents, when only the minority of the Congressional body is focusing on a group that has historically been oppressed, the Congress may not be fairly representing its people. This is highlighted by a recent poll of American voters in 2018 that found that 74% of voters believed that the "way women are treated in U.S. society" is either an "extremely or very important" issue to them (Gallup 2018) [10].

6.4.3 War

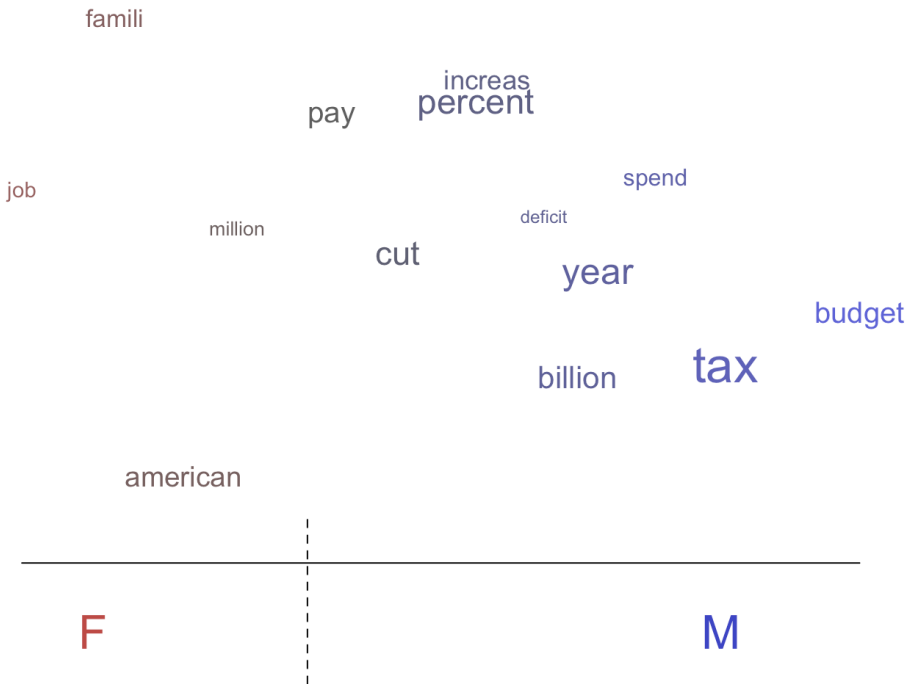
Figure 6.8: Content Differences: War



Within the subject of war, men tend to focus on the overall security of the American nation whereas women tend to focus on the individual troops that comprise the army. Notably, women were much more likely to discuss the Iraq war than men, which was war that lasted for almost a decade, or over 25% of the entire lexical corpus.

6.4.4 Budget/Taxes

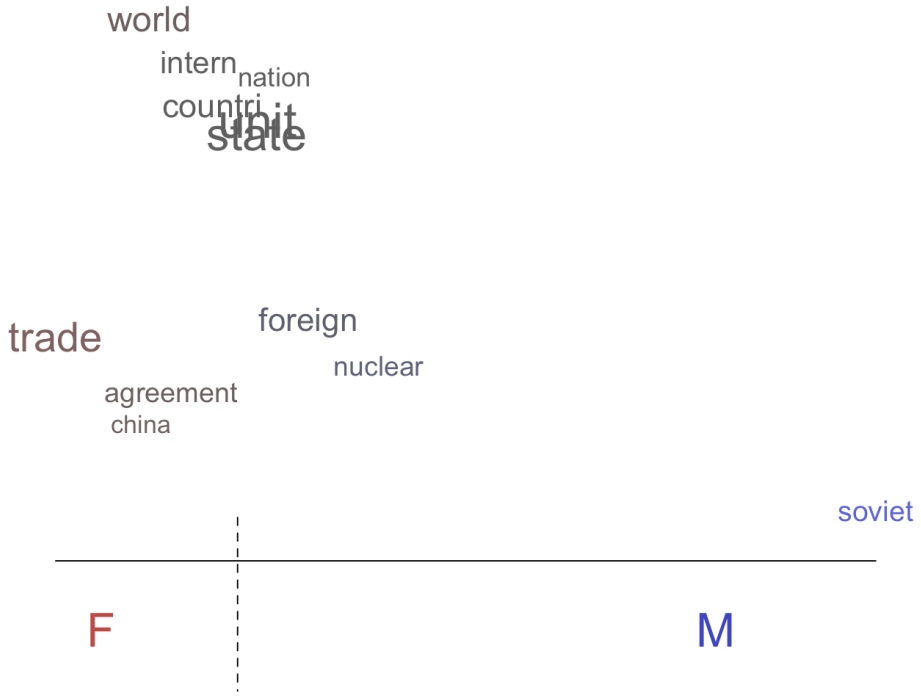
Figure 6.9: Content Differences: Budget/Taxes



Similarly to health, when it comes to the budget and taxes, women tend to focus on the people affected by the decisions Congress makes regarding the topic (e.g. “american”, “family”, “job”), whereas men spend more of their time focusing on the monetary aspect (“increase”, “spend”, “deficit”). Given that this is one of the most commonly discussed (interpretable) topics in Congress, the fact that women are spending their time talking about the budget and taxes in regards to how it actually affects American citizens more than men are is noteable.

6.4.5 Foreign Governments

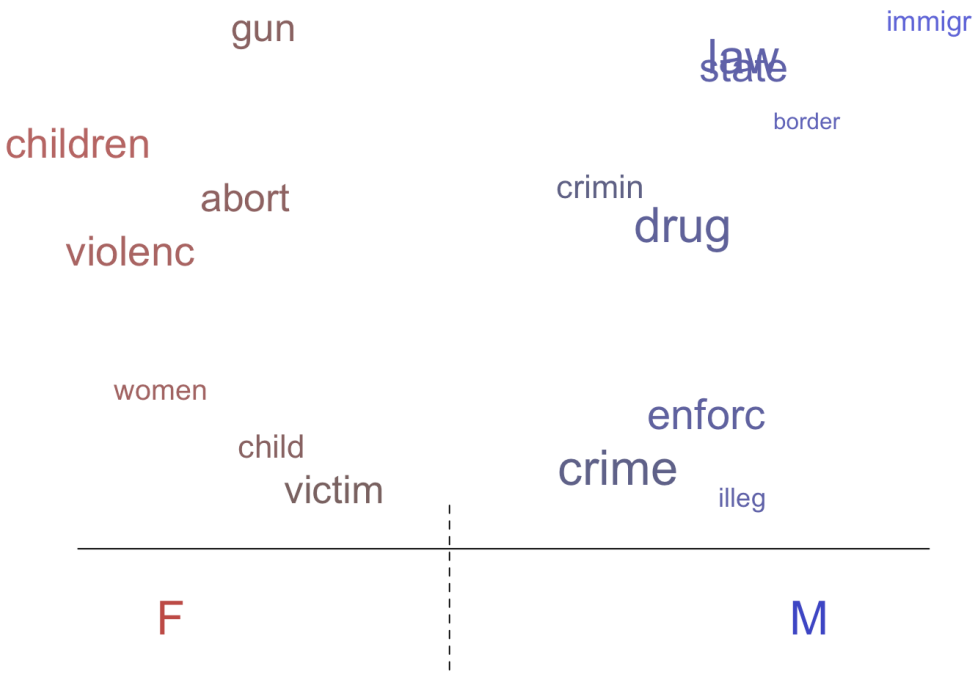
Figure 6.10: Content Differences: Foreign Governments



This topic of Foreign Governments can be seen as a clear indicator of how staggeringly disproportionate the gender balance in the late 20th century used to be. The fact that that "soviet" is so strongly skewed male is most likely indicative of the fact that the Cold War (which ended in 1991) was one of the most prominent aspects of foreign policy up until that point in time; concurrently the largest proportion of words spoken in Congress by women during this time was 4% [see figure 6.1]. Since this time represents over quarter of the lexical corpus (1981-1991), it is not too surprising that the extreme gender imbalance during this time is still made obvious in this topic overall. Another insight worth noting here is that women spend more time discussing trade between nations than men do.

6.4.6 Crime

Figure 6.11: Content Differences: Crime



In line with the previous topics mentioned, the topic of crime unveils a similar pattern between men and women's linguistic dedication in Congress. Men focus on the topic in general and the administrative aspect (e.g. "crime", "enforce", "law", "state"), whereas women focus on the people involved and affected and how the topic specifically relates to women (e.g. "children", "victim", "women"). Within the topic of crime, men are also much more likely to focus on immigrants and the border with Mexico whereas women direct much more of their time to gun violence than men do. There are not nearly as many grey central words in this plot which indicates that women and men discuss this topic incredibly differently. This pattern of the way men and women vary in topic content—general versus the people involved—is consistent in more topics analyzed in the corpus which can all be found in the appendix in figure A.1.

6.5 Fixed Effects Regression Model

Finally we estimate for each topic a Fixed Effects Regression Model to determine whether the gender variable is still significant when controlling for other covariates. As mentioned in the methodology sector we are not analyzing the magnitude of the β_1 (gender) coefficient but rather we are analyzing the sign of the coefficient and whether it is significant or not. In the appendix the summary results for the 18 estimated models can be found (table A.6).

All in all, the $Gender$ variable seems to be significant in 17 out of the 18 models we estimated for the substantive topics of interest. Veterans' Benefits is the only topic which is not addressed in significantly different proportions by men and women during the 1981-2016 period. Regarding the sign of the coefficient, all the estimated coefficient go in line with our previous results. Even when controlling for the $Party$, $Chamber$, $Year$, $State$ and $CloseElection$ covariates, women seem to talk significantly more about "Emergency Relief/Refugees", "Health", "Budget/Taxes", "Infrastructure and Technology", "Foreign Governments", "Agriculture and Water", "Election Reform", "Human Rights", "Crime" and "Judicial Nominations". On the other hand, men talk significantly more about "The Constitution", "War", "The Economy/Natural Disasters", "Funding of Departments", and "The Environment".

Regarding the non-substantive topics, table A.9 shows that Congressmen address topic 34 and 35 in greater proportion than Congresswomen (the two topics with the highest word count out of the entire lexical corpus [Figure 6.2], While Congresswomen address in greater proportion topic 12 which was the non-substantive topic which had the "soft" words or "modal auxiliary verbs" as described earlier.

The topics "Education" and "Energy/Education" provide ambiguous results. While the first one seems to be significantly more addressed by men the second one seems to be significantly more addressed by women. These results are not surprising given that the composition of these topics seems to be different by gender and neither of them should

be really interpreted as a homogeneous Education topic. In particular this is one clear limitation of this model which we discuss in the next subsection.

6.6 Limitations of the Model

After several iterations we decided to work with 40 topics since it was a reasonable number for our interpretation and analysis goals. However, as we mentioned before this arbitrary choice (K = 40) has some limitations. For example, with our choice of 40 topics we are unable to capture a homogeneous topic about Education. In particular topic 24 and 29 which we labeled as "Education" and "Energy/Education" have a strange composition by gender and are not clearly related to Education. The first topic (Topic 24) seems to capture the male speeches related to education with the words with higher probability of appearing being: educ, school, program, student, children, communiti, nation [see table A.1 in the appendix]. However, in the case of women the topic does not seem to cluster the education related topics since the words with highest probability of appearing are: communiti, nation, year, citi, univers, counti, san [table A.1].

Moreover, something similar can be observed when considering Topic 29 which we labeled "Energy/Education". Female speeches related to education seem to be clustered in this group (words with highest probability: school, educ, children, student, program, child, colleg [table A.1]) however the male speeches seem more related to energy than to education (energi, oil, price, gas, fuel, product, produc [table A.1]).

If we are particularly interested in observing the effect of gender on the discussion of education during the whole period, we should estimate an STM with a greater number of topics or with more iterations in order to find a homogeneous and clear topic about Education. As it stands in our current estimation of the model, we are unable to derive any meaningful results about this topic. A similar situation happened with topic 16 with "The Economy" and "Natural Disasters".

Given that running the model with more than 2,500,000 documents is incredibly costly

in terms of time, we perform several robustness checks by estimating the STM models with yearly data. That is to say that apart form estimating the model with the 2,500,000 documents containing all 35 years we additionally estimated STM models with 20 topics for the documents generated in 1981, 1991, 2001 and 2015 independently.

These results are not perfectly comparable because not all of the topics observed in the 1980-2016 period are found in the yearly models. However, most of the main topics were present and results were robust.

In particular we can observe from the yearly models (e.g 1981 and 2015) that an independent "Education" topic appears (not mixed with "Energy"). The most probable words and the FREX words differ by year but in all of the cases Congresswomen seem to address this topic in greater proportions than men.

Table 6.1: Example of Education related topics in yearly models

	Highest Probability	**FREX**
1981	program, educ, school, food, children student, will, pay, famili, percent	stamp, student, child, lunch, nutrit, educ, school, food, children, infant
2015	budget, tax, year, educ, program, student, school, will, need, spend	debt, tax, budget, student, educ, spend, trillion, cut, teacher, incom

We invite our readers to continue playing interactively with our different models and results through the following ShinyApp which is an implementation of Schwemmer's (2018) [28] stminsights package. More details in how to access the data, code, and Graphical User Interface (GUI), can be found in the end of Appendix.

CHAPTER 7
CONCLUSION

Though political representation of women has been growing for the past few decades, the gender representation in the world (and, specifically, the United States government) is still abysmally skewed towards men. This misrepresentation is even further accentuated by the disproportionate imbalance in the amount of words spoken in Congress by men and women. Though United States politics has been a popular subject for many different types of natural language processing, most of the focus has been on partisan differences (Democrats and Republicans) or specific subject matters. The exploration of how men and women differ politically is largely an untapped field of study in relation to NLP techniques. By estimating a structural topic model to the text corpus of U.S. Congressional daily speeches for the last four decades, it becomes possible to analyze how men and women differ in terms of topics discussed both in terms of frequency and content.

Women are significantly more likely to discuss infrastructure and technology, the budget/taxes, health, crime, energy/education, agriculture/water, foreign governments, election reform, human rights, and veterans' benefits. Contrastingly, men are more likely than women to discuss the economy/natural disasters, war, funding of departments, and the environment, though the sentiment of all subjects mentioned has not been anlayzed in this paper. Furthermore, these findings illustrate that men dedicate a greater proportion of their time discussing "topics" that do not offer any content (e.g. filler words/sentences or procedural jargon). These are incredibly important findings for the American voter because knowing who would be more likely to discuss issues they care about will influence voting behavior. For instance, a recent poll found that 80 percent of voters believe that healthcare is either an "extremely important" or "very important" issue influencing how they will vote (Gallup 2018) [10] and our findings indicate that women are 83.5% more likely to discuss

health in Congress than men are.

Our results also indicate that men and women have distinct patterns with how they approach topics: men talk about a topic from a general and administrative point of view, whereas women focus on the individuals and lives affected by the legislation. Furthermore, women make a point to discuss how women are specifically affected by legislation whereas men opt for a more general discussion. The men and women of the United States Congress are elected to represent the American people, and this paper shines an important spotlight on how men and women differently take on that responsibility.

This paper offers an insight into how Congressional representatives utilize their time. It confirms the need for a more representative governmental body in terms of the distribution of genders because women and men indeed discuss different topics and discuss them in significantly different ways. Ultimately, if people's views align more strongly with what women discuss than men, then they should vote for more women.

CHAPTER 8
DISCUSSION

8.1 Why A Structural Topic Model

There are many data science techniques to utilize when exploring textual sources of data; the resulting question becomes why do we choose to use a structural topic model and in particular the stm R package (for example as opposed to an LDA and Python, respectively). First of all, the key innovation with structural topic models as we already mentioned is that users are able to incorporate *metadata*, defined as information about each document, into the topic model. Researchers are able to discover topics and estimate their relationship to document metadata. The outputs of the model can be used to conduct hypothesis testing about these relationships, and the user becomes able to analyze the data from more perspectives than just solitary textual content. In fact, this is the type of analysis that social scientists perform with other types of data, where the goal is to discover relationships between variables and test hypotheses. In particular, we believe this is an excellent example of how machine learning and data science can be applied to social sciences.

In our case the documents correspond to the speeches given in congress from 1980-2016. In particular, we are interested in estimating the effect of the gender variable in the topic distribution. The structural topic model is a powerful tool that allows us to understand not only what men and women are discussing in congress and how often, but also how they discuss the topics from a content perspective.

Secondly, other packages have been built on top of the stm R package such as the stminsight package (Schwemmer, 2018) [28] which can be used by people without prior coding knowledge. These R packages and shiny app democratize access to these types of complex machine learning models by enabling people without prior coding knowledge to

explore the results to the same enjoyment as the author of the code. Moreover, all the code for the GUI and the model estimation is open source and can be easily accessed and modify

.

8.2 Why US Congressional Data

We have previously acknowledged that political empowerment is the area globally in which the gender gap remains the largest (only 23% of the gap is bridged as of 2018 (World Economic Forum, 2016) [31]). The question becomes why we chose to focus on United States political speeches rather than those of other countries. We acknowledge that though that we cannot assert with certainty that our findings can be generalized to more countries around the world, we are of the firm belief that a significant difference in topic prevalence and content at the highest level of governmental office can be found across genders in most countries. We chose to analyze the United State Congressional daily speeches because there was a rich data source of almost four decades available to us free of charge, and additionally there have been enough women in political positions that allowed us to study across genders rather than for a few individual women. Furthermore, the United States government is on a global stage, and many citizens of other countries will at least have some baseline knowledge of the political system of this country so it will be easily understood.

CHAPTER 9
FURTHER STUDY

Though we extensively examined how U.S. Congressional topics discussed varied across gender in terms of both frequency and content, this is still a field rich with possibility of further study. First of all, there are many demographic variables which, though we held some constant in the linear regression to solidify our natural language processing findings, can be examined at a closer level. For example, it would be interesting to examine how Democratic and Republican women differ in topical frequency and content, as well how region of the United States affects how different genders discuss topics. As for variables we did not examine in this paper, further study could go into how race affects content and frequency of topic discussion (both within and apart from gender identities), as well as other demographic variables in terms of age and experience level. Another topic that could be examined further is how the content of topics varied over time. Though we explored this in our Shiny app (which we again encourage the reader to explore, more information in the appendix), this subject matter was ultimately beyond the scope of this paper. Further exploration could also be done into the sentiment associated with topics discussed in Congress; we evaluated the content without the lens of if men or women were talking positively or negatively about different topics. Finally, any type of actions associated with topics discussed was beyond the scope of this paper. Further study could be done into how bills were put in motion varying by gender, as well as votes taken in Congress regarding the subjects discussed. This field of study is rife with possibilities.

APPENDIX A

Table A.1: Substantive Topics Highest Probability Words and FREX Words

	Topic		Female	Male
2	Constitution	Highest Prob	law, court, case, constitut, state, decis, protect	law, court, constitut, case, state, govern, decis
		FREX	lawsuit, fisa, alleg, plaintiff, unconstitut, statut, gay	plaintiff, lawsuit, statut, constitut, juri, unconstitut, litig
5	Small Business	Highest Prob	busi, small, compani, job, bank, loan, creat	will, busi, small, yes, creat, compani, job
		FREX	busi, sba, bank, entrepreneur, womenown, small, lend	sba, entrepreneur, small, parliamentari, busi, enterpris, ventur
6	Emergency Relief/Refugees	Highest Prob	situat, procedur, status, will, haiti, refuge, relief	situat, procedur, status, emerg, refuge, will, inquiri
		FREX	caribbean, haiti, haitian, asylum, refuge, status, procedur	haiti, haitian, inquiri, refuge, procedur, status, situat
7	War	Highest Prob	militari, war, iraq, forc, troop, nation, american	war, militari, forc, iraq, nation, secur, defens
		FREX	iraqi, iraq, troop, afghan, soldier, detaine, qaeda	iraqi, saddam, hussein, iraq, troop, alqaida, baghdad
9	Health	Highest Prob	health, care, insur, women, medic, cancer, diseas	health, care, insur, medic, medicar, drug, cost
		FREX	cancer, breast, diabet, diagnos, alzheim, diseas, uninsur	diabet, transplant, alzheim, diseas, embryo, hmo, hmos
15	Taxes/ Budget	Highest Prob	tax, american, cut, percent, job, famili, year	tax, year, budget, percent, billion, cut, spend
		FREX	unemploy, deduct, stimulus, wealthiest, wage, snap	deduct, tax, debt, trillion, bracket, incom, middleincom
16	Economy/ Natural Disasters	Highest Prob	safeti, need, transport, disast, emerg, air, fire	job, industri, bank, market, loan, compani, worker
		FREX	disast, fema, airlin, airport, hurrican, flood, aviat	bank, loan, lender, unemploy, cabl, foreclosur, mortgag
20	Infrastructure and Technology	Highest Prob	fund, program, million, develop, nation, year, project	technolog, research, nation, transport, develop, air, system
		FREX	hivaid, amtrak, nasa, broadband, space, rail, nist	amtrak, aviat, nasa, faa, cargo, noaa, congest
23	Foreign Governments	Highest Prob	unit, state, trade, world, countri, intern, foreign	unit, state, trade, soviet, world, foreign, countri
		FREX	cyprus, china, colombia, iran, palestinian, castro, darfur	sandinista, abm, taiwan, soviet, cuban, cyprus, baltic
24	Education	Highest Prob	communiti, nation, year, citi, univers, counti, san	educ, school, program, student, children, communiti, nation
		FREX	francisco, art, artist, music, scout, smithsonian, patsi	classroom, teacher, nea, student, faculti, elementari, scholar
26	Agriculture/ Water	Highest Prob	water, state, bill, protect, land, agricultur, farmer	water, agricultur, farmer, bill, farm, state, product
		FREX	fisheri, epa, fish, contamin, fishermen, mercuri, mtbe	dairi, crop, wetland, usda, farmer, speci, pesticid
27	Funding Departments	Highest Prob	secur, depart, report, requir, agenc, feder, bill	fund, program, year, author, bill, appropri, report
		FREX	cyber, homeland, dhs, gao, tsa, audit, inspector	gao, procur, dod, agenc, outlay, audit, fiscal
28	Election Reform	Highest Prob	elect, district, state, immigr, american, citizen, peopl	elect, polit, campaign, public, state, peopl, reform
		FREX	census, alien, deport, ballot, voter, citizenship, immigr	broadcast, ballot, voter, pac, puerto, elector, campaign

Table A.2: Substantive Topics Highest Probability Words and FREX Words (Cont.)

	Topic		Female	Male
29	Energy/ Education	Highest Prob	school, educ, children, student, program, child, colleg	energi, oil, price, gas, fuel, product, produc
		FREX	classroom, teacher, student, educ, elementari, afterschool	ethanol, gas, drill, oil, gallon, petroleum, gasolin
30	Human Rights	Highest Prob	right, women, civil, equal, protect, discrimin, act	right, object, reserv, languag, indian, civil, human
		FREX	discrimin, religi, religion, gender, civil, equal	object, withdraw, indian, tribe, tribal, right, bia
32	Crime	Highest Prob	children, violenc, crime, abort, drug, gun, victim	law, drug, crime, enforc, state, immigr, crimin
		FREX	abort, rape, gun, traffick, assault, fetus, firearm	firearm, cocain, traffick, crime, narcot, handgun, heroin
36	Environment	Highest Prob	energi, oil, gas, price, state, alaska, fuel	land, state, south, nation, area, park, forest
		FREX	drill, gallon, arctic, ethanol, energi, anwr, alaskan	miner, wilder, forest, blm, alaska, canyon, timber
38	Veterans' Benefits	Highest Prob	servic, veteran, legisl, provid, employe, worker, bill	legisl, bill, act, servic, provid, veteran, benefit
		FREX	veteran, spous, employe, compens, disabl, osha, pension	copyright, veteran, disabl, internet, pension, osha, employe
39	Judicial Nominations	Highest Prob	judg, court, nomin, quorum, attorney, justic, general	judg, quorum, nomin, hear, court, confirm, committe
		FREX	circuit, bench, ashcroft, judg, nomine, alito, appoint	nomine, nomin, bench, estrada, vacanc, judgeship, alito

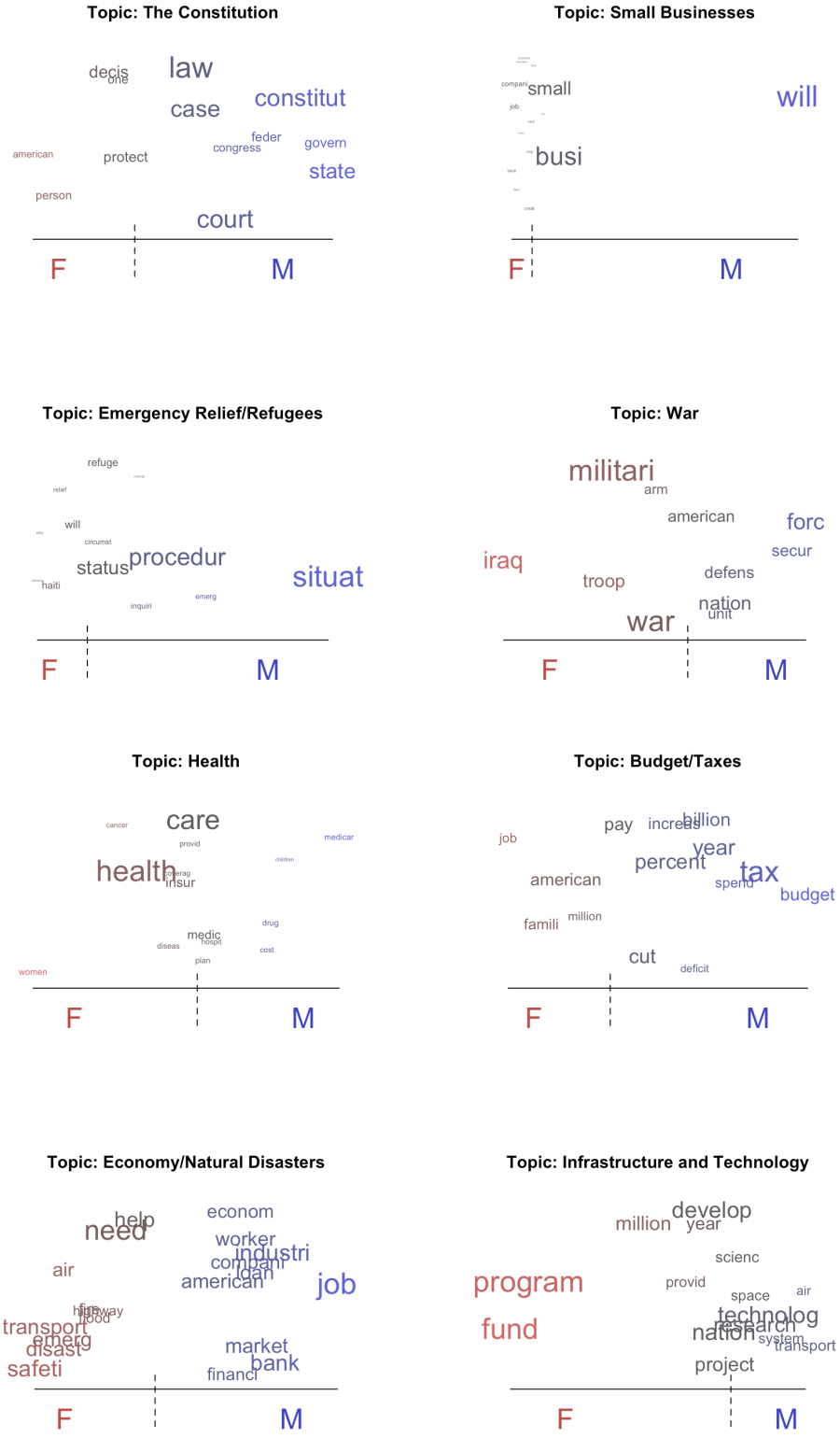
Table A.3: Non-Substantive Topics Highest Probability Words and FREX Words

Topic		Female	Male
1	Highest Prob	california, texa, york, florida, ohio, island, illinoi	california, york, texa, florida, virginia, illinoi, ohio
	FREX	california, virgin, florida, hawaii, island, texa, colorado	california, florida, georgia, virginia, rhode, connecticut, indiana
3	Highest Prob	presid, budget, spend, administr, bush, year, deficit	presid, administr, congress, bush, budget, polici, reagan
	FREX	budget, bush, clinton, obama, deficit, veto, reagan	veto, absenc, reagan, clinton, bush, lineitem, ronald
4	Highest Prob	new, pleas, north, carolina, happi, hampshir, colleagu	new, distinguish, pleas, north, mexico, carolina, happi
	FREX	carolina, yorker, hampshir, north, manhattan, delight, manchest	carolina, hampshir, thle, delight, happi, mexico, distinguish
8	Highest Prob	agre, resolut, motion, tabl, action, upon, lay	agre, move, motion, tabl, lay, upon, laid
	FREX	bosnia, kosovo, yugoslavia, balkan, bosnian, albanian, serb	motion, bahai, tabl, lay, agre, laid, tile
10	Highest Prob	ask, consent, unanim, order, call, madam, presid	ask, consent, unanim, order, call, print, madam
	FREX	consent, rescind, unanim, inquir, ask, order, call	consent, rescind, unanim, ask, print, vitiat, order
11	Highest Prob	speaker, madam, resolut, member, legisl, day, rule	speaker, resolut, member, madam, rule, legisl, may
	FREX	res, speaker, extran, nay, revis, concurr, con	speaker, nay, extran, concurr, revis, resolut, suspend
12	Highest Prob	presid, senat, follow, record, committe, author, read	presid, senat, consider, follow, proceed, read, record
	FREX	desk, dispens, dirksen, text, supra, designe, paragraph	desk, dispens, wednesday, calendar, designe, seconddegre, thursday
13	Highest Prob	bill, vote, pass, move, confer, floor, present	vote, bill, major, leader, debat, side, floor
	FREX	rollcal, aye, vote, reconsid, yes, bill, detain	clotur, rollcal, reconsid, vote, oclock, tomorrow, aye
14	Highest Prob	senat, will, state, louisiana, floor, can, wish	senat, will, question, can, matter, hope, understand
	FREX	arkansa, mikulski, byrd, louisiana, hutchison, inouy, reid	domenici, moynihan, hatfield, packwood, exon
17	Highest Prob	fact, believ, discuss, note, concern, suggest, might	suggest, might, discuss, fact, perhap, seem, inde
	FREX	note, discuss, suggest, might, inde, indic, perhap	suggest, perhap, might, inde, seem, somehow, discuss
18	Highest Prob	offic, team, marin, yea, armi, post, divis	offic, record, articl, post, washington, servic, name
	FREX	battalion, regiment, championship, messr, infantri, coach, sergeant	yea, postmast, bryant, taylor, gari, steve, kelli

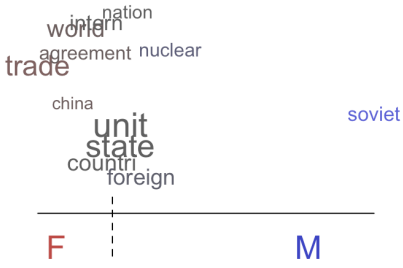
Table A.4: Non-Substantive Topics Highest Probability Words and FREX Words (Cont.)

Topic		Female	Male
19	Highest Prob	honor, famili, nation, american, today, live, life	year, american, nation, peopl, famili, honor, life
	FREX	armenian, condol, mourn, reverend, luther, commemor, boko	armenian, commemor, chaplain, condol, reverend, belov, remembr
21	Highest Prob	work, thank, colleagu, import, want, issu, member	work, colleagu, issu, thank, member, import, want
	FREX	congressman, togeth, caucus, congresswoman, chair, forward	chair, compliment, appreci, togeth, dilig, forward, work
22	Highest Prob	amend, committe, rule, chairman, appropri, member, section	amend, committe, rule, section, consid, member, act
	FREX	amend, section, markup, languag, waiv, titl, rule	amend, xxi, section, titl, jurisdict, xvi, claus
25	Highest Prob	time, back, balanc, consum, may, reserv, much	time, back, balanc, may, consum, much, reserv
	FREX	consum, balanc, reserv, back, cfpb, comp, overtim	balanc, consum, back, reclaim, remaind, time, inquir
31	Highest Prob	will, support, rise, urg, offer, oppos, propos	will, support, offer, propos, urg, rise, adopt
	FREX	opposit, adopt, oppos, urg, substitut, under, rise	adopt, substitut, offer, opposit, oppos, accept, propos
33	Highest Prob	gentleman, minut, gentlewoman, distinguish, demand, jersey	gentleman, minut, gentlewoman, demand, maryland, recommit
	FREX	gentlewoman, gentleman, customari, xviii, guam, recommit	gentleman, gentlewoman, gentleladi, guam, minut, recommit
34	Highest Prob	peopl, say, want, get, think, know, can	say, think, peopl, get, want, can, just
	FREX	dont, got, talk, everybodi, mayb, guess, bit	mayb, somebodi, everybodi, anybodi, got, guess, talk
35	Highest Prob	yield, chairman, thank, michigan, like, colleagu, subcommitte	yield, chairman, thank, will, correct, pennsylvania, wisconsin
	FREX	colloquy, yield, michigan, chairman, kansa, wisconsin, reclaim	yield, wisconsin, pennsylvania, minnesota, chairman, kentucki
37	Highest Prob	hous, republican, congress, democrat, american, govern, repres	hous, bill, republican, congress, pass, committe, confer
	FREX	shutdown, foreclosur, hud, mortgag, gingrich, homeownership	hous, confere, hud, shutdown, coin, confer, reconcili
40	Highest Prob	send, messag, sent, state, much, clear, can	send, messag, sent, clear, believ, state, signal
	FREX	messag, send, sent, signal, clear, much, behalf	send, messag, sent, signal, clear, believ, letter height

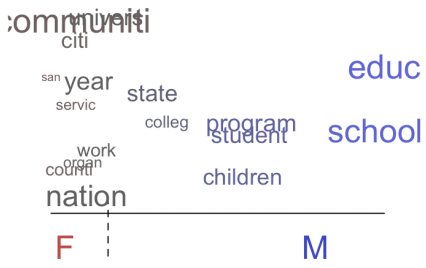
Figure A.1: Content Plots of Topics by Gender



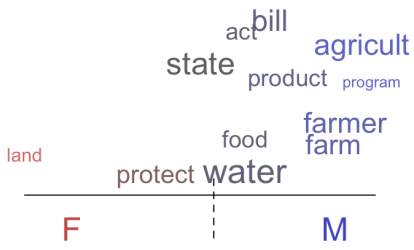
Topic: Foreign Governments



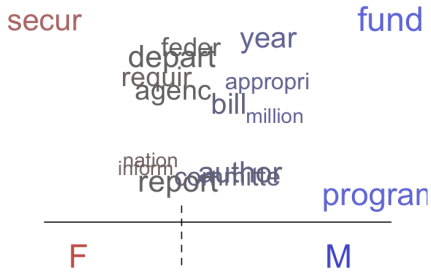
Topic: Education



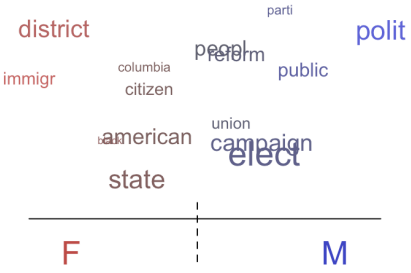
Topic: Agriculture/Water



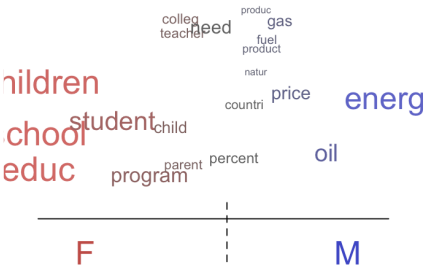
Topic: Funding of Departments



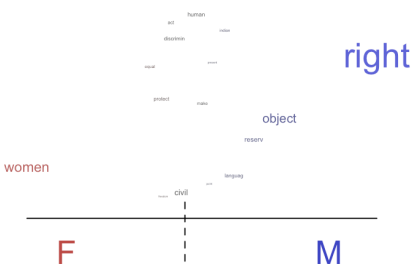
Topic: Election Reform



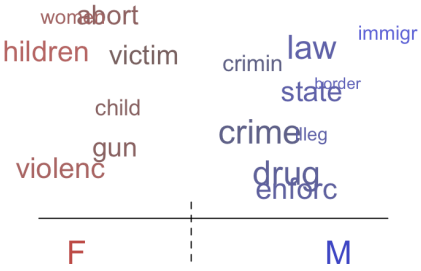
Topic: Energy/Education



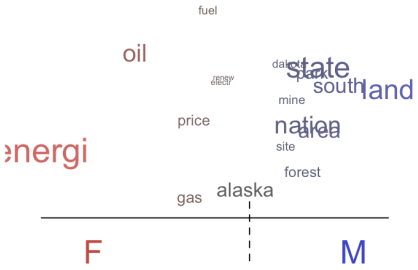
Topic: Human Rights



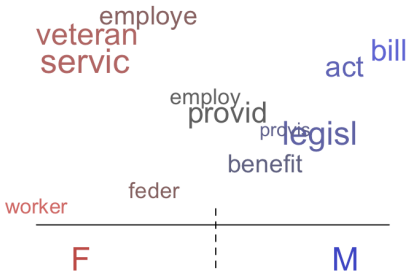
Topic: Crime



Topic: The Environment



Topic: Veterans's Benefits



Topic: Judicial Nominations

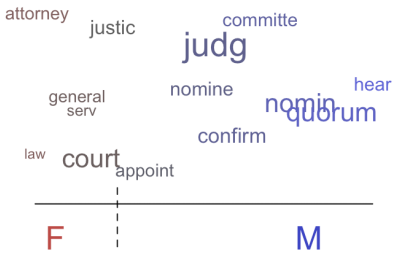


Table A.5: Mean Proportion of Topic Discussed by Gender

Topic Label	Topic Number	$\overline{\theta}$	$\overline{\theta_f}$	$\overline{\theta_m}$	$\frac{\overline{\theta_f}-\overline{\theta_m}}{\overline{\theta_m}}$
Non-Substantive	1	0.0201	0.0174	0.0203	-14.3
Constitution	2	0.0220	0.0201	0.0222	-9.5
Non-Substantive	3	0.0157	0.0173	0.0156	10.6
Non-Substantive	4	0.0136	0.0103	0.0138	-25.3
Small Business	5	0.0102	0.0136	0.0099	37.7
Emergency Relief/Refugees	6	0.0040	0.0037	0.0040	-7.5
War	7	0.0158	0.0191	0.0155	23.3
Non-Substantive	8	0.0161	0.0104	0.0165	-37.3
Health	9	0.0158	0.0273	0.0149	83.5
Non-Substantive	10	0.0453	0.0359	0.0461	-22.1
Non-Substantive	11	0.0387	0.0478	0.0379	26.0
Non-Substantive	12	0.0605	0.0407	0.0621	-34.5
Non-Substantive	13	0.0358	0.0321	0.0361	-11.1
Non-Substantive	14	0.0454	0.0217	0.0474	-54.2
Taxes/Budget	15	0.0332	0.0448	0.0323	38.7
Economy/Natural Disasters	16	0.0168	0.0151	0.0170	-11.0
Non-Substantive	17	0.0094	0.0116	0.0093	25.0
Non-Substantive	18	0.0146	0.0135	0.0147	-8.1
Non-Substantive	19	0.0276	0.0358	0.0269	32.8
Infrastructure and Technology	20	0.0167	0.0286	0.0157	82.5
Non-Substantive	21	0.0420	0.0446	0.0417	7.0
Non-Substantive	22	0.0279	0.0282	0.0279	1.2
Foreign Governments	23	0.0240	0.0251	0.0239	5.0
Education	24	0.0181	0.0220	0.0178	23.5
Non-Substantive	25	0.0324	0.0314	0.0325	-3.3
Agriculture/Water	26	0.0191	0.0235	0.0188	25.3
Funding Departments	27	0.0384	0.0310	0.0390	-20.5
Election Reform	28	0.0108	0.0141	0.0105	34.1
Energy/Education	29	0.0118	0.0202	0.0110	83.1
Human Rights	30	0.0154	0.0163	0.0153	6.3
Non-Substantive	31	0.0177	0.0215	0.0174	23.8
Crime	32	0.0136	0.0218	0.0130	67.8
Non-Substantive	33	0.0412	0.0387	0.0415	-6.8
Non-Substantive	34	0.0717	0.0693	0.0719	-3.7
Non-Substantive	35	0.0640	0.0518	0.0650	-20.3
Environment	36	0.0115	0.0114	0.0115	-0.5
Non-Substantive	37	0.0237	0.0227	0.0238	-4.8
Veterans' Benefits	38	0.0201	0.0227	0.0199	13.8
Judicial Nominations	39	0.0122	0.0106	0.0123	-14.0
Non-Substantive	40	0.0070	0.0065	0.0071	-8.0

Table A.6: Results: Fixed Effects Regression Models

	Dependent variable:					
	The Constitution	Emergency	War	Health	Budget Taxes	Economy/Natural
	(1)	Relief Refugees (2)	(3)	(4)	(5)	Disasters (6)
Gender: Male	0.001^{***}	-0.0004^{***}	0.003^{***}	-0.008^{***}	-0.012^{***}	0.004^{***}
	(0.0002)	(0.00003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
Party: Republican	0.001^{***}	-0.0001^{***}	0.002^{***}	-0.004^{***}	-0.005^{***}	-0.002^{***}
	(0.0003)	(0.00005)	(0.0003)	(0.0003)	(0.0004)	(0.0003)
Close_Election	-0.002^{**}	-0.0004^{***}	0.003^{***}	0.001^{**}	0.003^{***}	0.003^{***}
	(0.001)	(0.0001)	(0.001)	(0.001)	(0.001)	(0.001)
Chamber : Senate	0.003^{***}	0.0005^{***}	-0.002^{***}	0.001^{***}	-0.009^{***}	-0.0004^{***}
	(0.0001)	(0.00001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Gender*Party	0.0003	0.0002^{***}	-0.002^{***}	0.002^{***}	0.006^{***}	-0.002^{***}
	(0.0003)	(0.00005)	(0.0003)	(0.0003)	(0.0004)	(0.0003)
Year Fixed Effects	✓	✓	✓	✓	✓	✓
State Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458
R^2	0.009	0.012	0.028	0.019	0.021	0.012
Adjusted R^2	0.009	0.012	0.028	0.019	0.021	0.012
Residual Std. Error	0.060	0.009	0.054	0.058	0.082	0.049
F Statistic	236.496^{***}	316.829^{***}	775.129^{***}	515.937^{***}	573.615^{***}	337.626^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.7: Results: Fixed Effects Regression Model (Cont.)

	Dependent variable:					
	Infrastructure and Technology (1)	Foreign Governments (2)	Education (3)	Agriculture Water (4)	Funding of Departments (5)	Election Reform (6)
Gender: Male	-0.012^{***}	-0.002^{***}	0.002^{***}	-0.006^{***}	0.004^{***}	-0.003^{***}
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
Party: Republican	-0.001^{***}	0.009^{***}	0.003^{***}	0.004^{***}	0.007^{***}	-0.001^{***}
	(0.0003)	(0.0004)	(0.0003)	(0.0003)	(0.0004)	(0.0002)
Close_Election	0.001^{**}	0.0005	0.013^{***}	0.004^{***}	0.002^{*}	0.001^{**}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)
Chamber: Senate	-0.002^{***}	-0.0003^{**}	0.003^{***}	-0.001^{***}	0.002^{***}	-0.003^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00005)
Gender*Party	-0.002^{***}	-0.010^{***}	-0.006^{***}	-0.004^{***}	-0.009^{***}	0.002^{***}
	(0.0003)	(0.0004)	(0.0003)	(0.0003)	(0.0004)	(0.0002)
Year Fixed Effects	✓	✓	✓	✓	✓	✓
State Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458
R^2	0.010	0.022	0.012	0.012	0.017	0.012
Adjusted R^2	0.010	0.022	0.012	0.012	0.017	0.012
Residual Std. Error	0.051	0.077	0.061	0.060	0.078	0.034
F Statistic	265.397^{***}	598.139^{***}	341.269^{***}	340.537^{***}	471.962^{***}	333.742^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.8: Results: Fixed Effects Regression Model (Cont.)

	Dependent variable:					
	Energy/Education	Human Rights	Crime	The Environment	Veterans Benefits	Judicial
	(1)	(2)	(3)	(4)	(5)	Nominations (6)
Gender: Male	-0.006^{***}	-0.004^{***}	-0.007^{***}	0.001^{***}	0.0001	-0.001^{***}
	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
Party: Republican	-0.004^{***}	-0.0005^{***}	-0.001^{**}	0.003^{***}	0.003^{***}	-0.001^{***}
	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0003)	(0.0002)
Close Election	0.001^{***}	-0.002^{***}	0.003^{***}	0.003^{***}	0.007^{***}	-0.0001
	(0.001)	(0.0004)	(0.001)	(0.0004)	(0.001)	(0.0004)
Chamber: Senate	-0.002^{***}	-0.003^{***}	-0.001^{***}	0.001^{***}	0.002^{***}	0.015^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Gender*Party	0.003^{***}	0.002^{***}	0.002^{***}	-0.004^{***}	-0.005^{***}	0.001^{***}
	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0003)	(0.0002)
Year Fixed Effects	✓	✓	✓	✓	✓	✓
State Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458	2,576,458
R^2	0.024	0.010	0.014	0.021	0.008	0.054
Adjusted R^2	0.024	0.009	0.014	0.021	0.008	0.054
Residual Std. Error	0.046	0.035	0.052	0.038	0.052	0.036
F Statistic	657.512^{***}	260.936^{***}	377.408^{***}	575.520^{***}	211.313^{***}	$1{,}546.708^{***}$

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.9: Results: Fixed Effects Regression Model for Select Non-Substantive Topics

	Dependent variable:		
	'34'	'35'	'17'
	(1)	(2)	(3)
Gender: Male	0.001^{***}	0.017^{***}	-0.0002
	(0.0003)	(0.0003)	(0.0003)
Party: Republican	-0.008^{***}	0.003^{***}	-0.009^{***}
	(0.001)	(0.0005)	(0.0005)
Chamber: Senate	-0.017^{***}	-0.076^{***}	0.092^{***}
	(0.0001)	(0.0001)	(0.0001)
Gender*Party	0.013^{***}	-0.002^{***}	0.012^{***}
	(0.001)	(0.0005)	(0.001)
Constant	0.050^{***}	0.102^{***}	0.039^{***}
	(0.001)	(0.001)	(0.001)
Observations	2,576,458	2,576,458	2,576,458
R^2	0.027	0.161	0.234
Adjusted R^2	0.027	0.161	0.234
Residual Std. Error (df = 2576363)	0.099	0.091	0.096
F Statistic (df = 94; 2576363)	765.223^{***}	$5{,}255.388^{***}$	$8{,}354.541^{***}$
Note:			*p<0.1; **p<0.05; ***p<0.01

Shiny App

We estimated a number of different models in order to be sure that our results are robust and anyone with Internet access can play around with these results. In fact, we made publicly available part of our results in the following Google Drive folder[1].

The steps to visually follow and interactively play around with the results are:

- Access the Google Drive folder through your preferred browser to download the data
- Access the following ShinyApp[2] through your preferred browser. Or through your RStudio Session install the stm_insights package and run the following command run_stminsights()[3]
- Upload in the ShinyApp one of the yearly results RData files (each of this RData files contains the estimated model, the estimated effect and the out file with the metadata for a given year)
- Play around with the GUI. For example, on the first tab one can find the words with highest probability of appearing in each topic that enable the user to label the topics according to her own criteria. Secondly, one could click on the "Plots" tab then choose type of plot "continuous" and plot variable: date_numeric. By clicking the Interaction Term and choosing Gender one can observe the plots presented in this paper.

[1] https://drive.google.com/drive/folders/1cUjwCHheXf4L3YB2X6BmzOBuzEbxoEEq?usp=sharing

[2] https://maiabrenner.shinyapps.io/text_analysis_app/

[3] The App has been deployed with a free account of shinyapps.io for academic purpose only and in order for people with no rstudio knowledge to be able to use it. Accessing the App through this link may not have optimal speed. Accessing it from the original stm_insights package is much faster

REFERENCES

Besley, T. et al. (2004). "Gender Quotas and the Crisis of the Mediocre Man: Theory and Evidence from Sweden". In: *American Economic Review* 107 (8), pp. 2204–42.

Biber, D., Conrad S., and Reppen R. (1998). "Corpus linguistics: Investigating language structure and use." In: *Cambridge, England: Cambridge University Press.* 84, pp. 857–870.

Blei, D. M and Lafferty J. D. (2007). "A correlated topic model of Science". In: *Ann. Appl. Stat.* 1 (1), pp. 17–35.

Blei, D.M., Ng A. Y, and Jordan M. I. (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research 3*, pp. 993–1022.

Brescolli, V. L. (2011). "Who Takes the Floor and Why: Gender, Power, and Volubility in Organizations". In: *Administrative Science Quarterly* 56 (4), pp. 622–641.

Brollo, F. and Troiano U. (2013). "What Happens When a Woman Wins an Election? Evidence from Close Races in Brazil". In: *Competitive Advantage in the Global Economy (CAGE)* (161).

Chattopadhyay, R. and Duflo E. (2004). "Women as policy makers: evidence from a randomized policy experiment in India". In: *Econometrica* 72 (5), pp. 1409–1443.

D. Lewis, et al. (2004). "Rcv1: A new benchmark collection for text categorization research". In: *Journal of machine learning research 5*, pp. 361–397.

Dovidio, J.F. et al. (1988). "The relationship of social power to visual displays of dominance between men and women." In: *J Pers Soc Psychol* 54 (2), pp. 233–42.

Gallup (2018). "Gallup's Midterm Election Benchmark poll". In:

Gentzkow, M., Shapiro J. M., and Taddy M. "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts". In: *Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. https://data.stanford.edu/congress*$_text$.

Hancock, A. and Rubin B. A. (2014). "Influence of Communication Partner's Gender on Language". In: *Journal of Language and Social Psychology* 34 (1), pp. 46–64.

Jacobi, T. and Schweers D. (2017). "Justice, Interrupted: The Effect of Gender, Ideology and Seniority at Supreme Court Oral Arguments". In: *Northwestern Law Econ Research Paper No. 17-03. Available at SSRN: https://ssrn.com/abstract=2933016*.

Lenard, D.B. (2016). "Gender Differences in the Political Speeches from the 113th United States Congress". In: *Doctoral Thesis, Josip Juraj Strossmayer University of Osijek, Faculty of Humanities and Social Sciences.*

Manning, C. D. and Schtze H. (1999). *Foundations of Statistical Natural Language Processing*. Prentice-Hall, Inc.

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley (2015). "Navigating the Local Modes of Big Data: The Case of Topic Models". In:

McMillan, J. R et al. (1977). "Womens language: Uncertainty or interpersonal sensitivity and emotionality?" In: *Sex Roles* 3, pp. 545–559.

Mehl M. R., Pennebaker J. W. (2003). "The sounds of social life: A psychometric analysis of students daily social environments and natural conversations." In: *Journal of Personality Social Psychology* 84, pp. 857–870.

Mohr, J. W. and Bogdanov P. (2013). "Introduction-Topic Models: What They Are and Why They Matter". In: *Poetics* 41 (6), pp. 545–569.

Mulac, A. and Lundell T.L. (1994). "Effects of gender-linked language differences in adults written discourse: Multivariate tests of language effects." In: *Language Communication* 14, pp. 299–309.

Newman, M.L. et al. (2008). "Gender Differences in Language Use: An Analysis of 14000 Text Samples". In: *Taylor Francis Group, LLC*. 45.

Persson, T. and Tabellini (2000). "Language and gender in Congressional Speech." In: *Political Economics, Cambridge, MA: MIT Press.*

Porter, M. (1980). "An algorithm for suffix stripping". In: *Journal of machine learning research 5*, pp. 130–137.

Roberts, M. E., Stewart B. M., and Tingley D. (2016). "stm: R Package for Structural Topic Models". In: *Journal of Statistical Software*.

Roberts, M. E. et al. (2013). "The Structural Topic Model and Applied Social Science." In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Roberts, M.E. et al. (2014a). “Structural Topic Models for Open-Ended Survey Responses”. In: *American Journal of Political Science* 58 (4), pp. 1064–1082.

— (2014b). “Structural Topic Models for Open-Ended Survey Responses: Online Appendix”. In: *https://scholar.princeton.edu/sites/default/files/bstewart/files/ajpsappendix.pdf*.

Schwemmer, Carsten (2018). *stminsights: A Shiny Application for Inspecting Structural Topic Models*. R package version 0.3.1.

Technology Election Data + Science Lab, Massachusetts Institute of (2019). “U.S. Senate 19762018 and U.S. House 19762018”. In:

Vijayarani1, S. and Janani R. (2016). “Text Mining: Open Source Tokenization Tools An Analysis.” In: *Advanced Computational Intelligence: An International Journal* 3 (1), pp. 37–47.

World-Economic-Forum (2016). *The Global Gender Gap Report*.

Yu, B. (2013). “Language and gender in Congressional Speech.” In: *Literary and Linguistic Computing* 29 (1).