DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 23-08

# Legislation For Sale? The Influence of PhRMA Campaign Donations on the US House Health Legislation

Brian Flaxman
University of Colorado Boulder

October 26, 2023

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

# Legislation For Sale? The Influence of PhRMA Campaign Donations on US House Health Legislation

Brian Flaxman*

October 26, 2023

**Abstract**

I find that increasing a legislator's previous Congress PhRMA favorability by one standard deviation increases their current Congress PhRMA PAC donations by between 0.107 and 0.147 standard deviations depending on PhRMA favorability measure, and that increasing a legislator's previous congress PhRMA PAC donations by one standard deviation increases their PhRMA favorability in the current Congress by between 0.009 and 0.118 standard deviations depending on PhRMA favorability measure. This is the first paper to analyze empirically the relationship between legislation favorability and special interest group donations. Natural language processing was used to generate similarity scores between mock legislation from the American Legislative Exchange Council, a group associated with Pharmaceutical Research and Manufacturers of America (PhRMA), and legislation introduced in the US House of Representatives to create measures of "PhRMA Favorability" for legislators over the period from 2013 through 2020. These measures were analyzed with campaign donations to House of Representatives members from PhRMA PACs. OLS regression with fixed effects and clustered errors was utilized on unbalanced panel of sitting legislators from the 114th Congress (2013/14) through the 116th Congress (2019/20).

*University of Colorado Boulder

[0]Acknowledgments forthcoming

**JEL Codes:** C80, D72, D78, I18, K16, L38
**Keywords:** campaign finance, natural language processing, House of Representatives, PhRMA, legislation

# 1 Introduction

In order to run a successful campaign for public office, candidates need to have the financial resources for advertising, grassroots level organizing, campaign infrastructure, and a multitude of other expenses. It is no surprise then that the amount of money a campaign is able to raise is a key factor in a candidate's success, both in America and around the world (Abramowitz, 1988; Grier, 1989; Eagles, 1993; Breaux and Gierzynski, 1991; Moon, 2006; Da Silveria and De Mello, 2011; Fourniares and Hall, 2014; Broberg, Pons, and Trocaud, 2022). Given how critical fundraising is for running a successful campaign, it is critical to analyze the sources of candidate funds, and any conflicts of interest that can arise. This is especially true if funds are raised by specific corporate interests, whose sole objective is the profit maximization of firms within the industry, especially since it has been shown that a government and officials that are more favorable for a corporate interest increase their profits (Cooper and Ovtchinnikov, 2010; Huber and Kirchler, 2013). It is therefore quite surprising that nor papers have been published before to empirically analyze the relationship between special interest group campaign donations and the language in legislation crafted by public officials. This paper is the first to do just that. Such analysis is quite feasible utilizing Natural Language Processing (NLP) methods, the computational understanding, analysis, generation, and quantification of natural language. All that is needed is available text that conveys the priorities of a special interest group to compare to introduced legislation. To make this task even more straightforward, some special interest groups produce mock legislation, known as model legislation, that conveys the legislative priorities of the group.

One special interest group that crafts model legislation is known as the American Legislative Exchange Council (abbreviated ALEC), a conservative group that's main focus is influencing policy in the United States at the state level. The group is heavily influenced by corporate America, and has many representatives of corporate America on the "Private Enterprise Board of Directors" wing of their

leadership. For over a decade, the Pharmaceutical Research and Manufacturers of America (abbreviated PhRMA), a prominent pharmaceutical industry trade group, has been on this private enterprise board, thus giving it a hand in crafting the model bills its produces, especially those on healthcare. PhRMA is also very influential in politics at the federal level, raising funds through its corporate PAC for candidates, including many sitting members of the United States House of Representatives. ALEC's model legislation, United States House of Representatives legislation, and PhRMA PAC fundraising data, all of which are publicly available, can thus be utilized to analyze the relationship between legislation favorability and campaign donations behavior.

This paper utilizes an unbalanced panel of sitting United States House of Representative members from the 113th Congress (2013/14) through the 116th Congress (2019/20) that contains PhRMA PAC donations data and measures of the "PhRMA favorability" of the legislation sponsored by House members. These PhRMA favorability measures were created using Natural Language Processing (NLP). Two different Natural Language Processing modeling methods, term frequency and Latent Dirichlet Allocation (LDA), were used to calculate the level of similarity between each piece of US House healthcare legislation sponsored in a Congress and each ALEC model healthcare bill finalized during that same Congress. The highest similarity score between a US House bill and any ALEC model bill from its Congress was assigned as that House bill's "ALEC similarity" score. The ALEC similarity scores from both modeling methods were then used to create two different measures of a legislator's PhRMA favorability each for the term frequency and LDA models . Since the relationship between campaign donations and legislation favorability in a single Congress is likely to be bidirectional, OLS regressions of PhRMA PAC donations lagged by one Congress on PhRMA favorability and OLS regressions of PhRMA favorability lagged by one Congress on PhRMA PAC donations were run, with one of the four PhRMA favorability measures being run in a separate regression. All regressions contained controls for party affiliation, previous election vote percentage, and fixed effects for Congress and location, with robust standard errors clustered on Congress and location being utilized.

Depending on the PhRMA favorability measure used, a one standard deviation increase in a House member's PhRMA favorability in the prior Congress leads to between a 0.908 and a 0.147 standard deviation increase in the PhRMA PAC donations that a House member receives in the current Congress, all statistically

significant at the 1% level. Furthermore, again depending on the PhRMA favorability measure used, a one standard deviation increase in the PhRMA PAC donations received by a House member in the prior Congress leads to a between 0.0096 and a 0.118 (advisors: This is not a typo, the maximum magnitude is in fact over 10x that of the minimum in these regressions) standard deviation increase in their PhRMA favorability in the current Congress, again all statistically significant at the 1% level.

This paper is part of the niche campaign finance literature. Most of the work in the campaign finance literature uses roll call votes, or how legislators vote on certain pieces of legislation, to measure government official behavior with respect to donations. The literature in this area has yielded somewhat mixed results. Most papers show that these roll call votes are affected by donations (Silberman, 1976; Hall and Wayman, 1990; Stratmann 1995; Stratmann, 2002; and Mian et al., 2013), yet some papers show that they do not (Bronars and Lott, 1997; Milyo 1999). Some papers do look at government decisions outside of roll call votes, such as as regulatory decisions (Gordon and Hafer, 2005; De Figuierdo and Edwards, 2007; Stratmann and Monaghan 2017) and appropriated government contracts (Boas, Hidalgo, and Richardson, 2014), with these papers showing a link between donations and these decisions. This paper is wholly different from any of the papers mentioned above in that it looks at how campaign donations from corporations influence the legislation that government officials craft.

This paper is also part of the ever growing economic applications of NLP literature. While NLP has been used in data and computer science for decades, such methods are finally being applied to economics and finance research. NLP has been used in finance to study the relationship between asset prices and news sentiment (Antweiler and Frank, 2006; Tetlock, 2007; Garcia, 2013 Manel and Moreira, 2017), in the economics of risk and uncertainty to study how firms and economies respond to economic volatility (Baker, Bloom, and Davis, 2016), how firms respond to political volatility (Hassan et al., 2017), and in the industrial organization literature to help define product markets (Hoberg and Phillips, 2009). Such methods were also used in the widely circulated Ederer, Goldsmith-Pinkham, and Jensen (2023) investigating the toxicity stemming from the Economics Job Market Rumors Website. In political economy specifically, NLP has been used in some papers to analyze the spoken (Quinn et al., 2010; Gentzkow, Shapiro and Taddy, 2019) and written language of public officials (Sim, Routledge, and Smith, 2015), but

there have been very few applications of analyzing policy. The few existing examples have been utilized NLP to classify policy into certain categories for more standard empirical analysis (Lane, 2022). This would be the first economics paper to utilize NLP to analyze policy directly in any way, let alone being the first to do so in the realm of campaign finance.

This paper is organized as follows. Section 2 contains the relevant institutional information for this paper regarding the US House of Representatives, US elections, campaign finance and relevant law, and ALEC and PhRMA. Section 3 contains a simple model built from first principles to illustrate the relationship between campaign donations and legislation favorability on a fundamental level. This model is unlike other existing models of campaign finance (Gerber, 1996; Pratt, 2002; Coate, 2004a; Coate, 2004b; Ashworth, 2006) as it is built on first economic principles. Section 4 describes in detail the data collection and variable creation processes utilized for the empirical analysis, with special attention paid to the details of term frequency language models, LDA language models, and the creation of the four measures of PhRMA favorability. Section 5 describes the specific empirical strategy utilized to determine the relationship between PhRMA favorability and PhRMA PAC donations, including the specific regression equations utilized, whose results and commentary of them can be found in Section 6. The paper concludes with Section 7.

# 2 Institutional Context

## 2.1 United States Government

The United States has three branches of government: the legislative, executive, and judicial branches. The executive branch is made up of the President, their cabinet, and those working under them. The judicial branch is made of the Supreme Court and all of the federal judiciary. The legislative branch (or legislature) is made of the House of Representatives and the Senate, known collectively as Congress. The legislative branch is responsible for crafting and passing legislation, with the other two branches having checks over the legislature. All three branches have some role over the legislative process. While the legislature is responsible for crafting and passing legislation, the executive branch has the ability to veto this legislation (that

can later be overruled by the legislature with enough support) and the Judiciary has the ability to determine legislation unconstitutional.

The House of Representatives is made of 435 voting members, [1] with each of these 435 Representatives serving a unique portion of a US state known as a district. States that have higher populations have more congressional districts than those that have smaller populations, with each state having at least one Representative. The US Senate is made up of 100 members, with each state having two Senators regardless of state population. Terms for House members last two years and terms for Senators last four. While the House is referred to as the lower chamber and the Senate the upper chamber, the functions of both chambers have very similar,[2], and both chambers are considered to have equal power. Both chambers have committees that are in charge of approving legislation and oversee government and private agencies and organizations. Both play an equal role in creating, drafting, and helping to pass legislation. And legislation in most subject areas can originate in either the House or Senate.[3]

## 2.2 Legislation

There are many categories of legislation, all of which can be introduced in either chamber. These are bills, joint resolutions, simple resolutions, concurrent resolutions, and amendments. While a bit confusing, the term bill can also be used to refer to all types of legislation, as will be done beginning with Section 3. A bill is a piece of legislation dealing with the public that becomes a law if it makes its way fully through the legislation process. A joint resolution is functionally equivalent to a bill. A simple resolution is a piece of legislation that either addresses issues exclusive to one chamber of Congress (such as rule changes) or to express the sentiments of that chamber. They undergo no further action after being approved by that chamber. Legislation that addresses the issues and sentiments of both chambers are known as concurrent resolutions. After being approved by

[1] There are 6 non-voting members, representing the District of Columbia, Puerto Rico, American Samoa, Guam, the U.S. Virgin Islands, and the Northern Mariana Islands.

[2] One exception is that an impeachment indictment against a president occurs in the House and requires a simple majority of the vote, while the trial to remove the President occurs in the Senate and requires a two-thirds majority of the vote

[3] One of the few exceptions is that legislation regarding taxation and other revenue raising must originate in the House.

both chambers, they undergo no further action. Amendments are proposed to alter legislation prior to coming up for a vote.

The life-cycle for all legislation except for Amendments begins with it being introduced into Congress by a legislator, known as the "sponsor." This legislation is usually drafted by staff members, special interest groups, or others, but is on occasionally written by the lawmaker themselves. This original sponsor then seeks to add cosponsors, other legislators that add their names to become initial supporters of the legislation but whose offices did not have a hand in crafting the legislation. That legislation is then sent to its appropriate congressional committee (although simple and concurrent resolutions often skip this step). If the committee approves it, the legislation is then sent to that chamber's floor to be introduced to the floor and debated on. If the chamber passes the legislation with a simple majority, it passes chamber.[4]

If the legislation was not a bill or joint resolution, the process ends. Bills and joint resolutions are then sent to the other chamber where the identical process takes place. Bills and joint resolutions that passed in different forms are sent back to both chambers where an identical "reconciled" form is crafted to then be voted on by both chambers. After this, the bill or joint resolution is sent to the president. If the President signs the bill or join resolution into law, or they don't sign it within 10 days (known as a pocket veto), the bill becomes law. If the President vetoes it, the bill is sent back to both chambers, where it must pass with a two-thirds majority for the veto to be overridden. With how complex this process is, it is no surprise that a majority of the legislation introduced in Congress does not pass, especially bills and joint resolutions. Of the 9,172 bills and joint resolutions introduced in the House in the 116th Congress, 1,156, or 12.6% made it to committees. And of these, 795, or 8.7% were debated and brought up for a House vote. Of these, all but two passed the House in some form (House Speakers oftentimes do not bring up legislation not likely to pass). In total, 344 laws were passed by both chambers of Congress and signed off by President Trump. A higher percentage of resolutions were passed by the House, with 257 of the 1,401 passing, ot 18.3%. These percentages are similar for legislation in the Senate. Part of the reason

[4]One feature of the Senate is that their current rules of operation includes the ability to call a non-standing filibuster on most pieces of legislation. This filibuster can halt proceedings unless 60 or more members call to end the filibuster. Therefore, while it does take a simple majority to pass a piece of legislation in the Senate, it needs the support of 60 or more Senators to be called for a vote in the first place.

the adoption rate of proposed legislation is so low is that much of the legislation drafted is not expected by its original sponsor to pass. It is often drafted with the sole purpose of appealing to voters and potential donors as a signal of their policy priorities.

## 2.3 Election Cycles

In the United States, there is a general election every year on the Tuesday falling between November 2 and November 8, with primary elections (elections between members of a party to determine the candidate who will run in the general election) taking place months earlier with exact timing varying by state. Unlike Senators who are up for election every 6 years, elections for all House seats take place during every even year general election. Unlike many countries around the world, there is no legally defined campaign window. Because of this, and because House members are up for reelection so frequently, sitting Representatives spend much of their term campaigning and raising money for their upcoming reelections in addition to their actual legislative duties.

## 2.4 Campaign Finance Law

### 2.4.1 Supreme Court Case Law

The first modern day attempts to regulate campaign finance occurred with the passing of the Federal Election Campaign Act of 1971, a measure to ferret out corruption in government in the wake of the Watergate Scandal. This law put into place limits several restrictions on both campaign fundraising and expenditures. The case of Buckley v. Valeo (1976) ruled that while limiting individual contributions to campaigns did not violate the First Amendment,[5] limiting independent expenditures, limiting the amount that candidates can spend of their own personal funds, and limiting the total amount that campaigns can spend in an election were violations of the First Amendment. Colloquially, this decision is referred to as "money is speech," and paved the way for many cases down the

[5]Freedom of speech, religion, and assembly protections, found in the First Amendment to the United States Constitution

road that would declare restrictions on campaign finance violations of the First Amendment.

One of the oldest campaign finance regulations on the books, the Tillman Act, was passed in 1907, barring corporations from participating in federal election campaign efforts. Meanwhile, many states passed their own laws prohibiting such behavior at the state level. One such state was Massachusetts, which prohibited corporations from donating to ballot initiatives, unless that ballot initiative dealt directly with that corporation. The case of First National Bank of Boston v. Bellotti ruled that these laws were violations of a corporation's First Amendment rights. Colloquially, this decision is referred to as "corporations are people," because while the idea of limited"corporate personhood" had been established prior, this was the first to bestow corporations First Amendment rights, ones derived from the Buckley v. Valeo decision.

For the next twenty years after, minor supreme court decisions had been made slowly reducing the ability for state and federal governments to limit campaign donation and expenditure behavior, with regulations being written in response to these cases. The next major relevant Supreme Court case was the well-known decision of Citizens United v. FEC (2010). This case restricts the federal government from limiting independent expenditures in any way, meaning that corporations and individuals can spend an unlimited amount of money on elections as long as it is done through indirect operations. It is worth noting that this case built upon the precedents set forth by both Buckley v. Valeo and Bank of Boston v. Bellotti. While Citizens United v. FEC dealt mostly with outside expenditure groups, the case of McCutcheon v. FEC (2014) impacted more direct channels of fundraising. This decision ruled that it is unconstitutional to limit the amount of donations that an individual can give across all candidates and PACs, declaring the $123,000 limit at the time unconstitutional.

### 2.4.2 Current Campaign Finance Laws

The currrent state of United States campaign finance law have been heavily shaped by the aforementioned Supreme Court decisions. In addition to no legal constraints on campaign timing, there are very few constraints on the ways in which candidates can raise funds. There are strict limits on the amount that candidates can take

from individuals directly ($ 3,300 for the 2023-24 election cycle each for the primary and for the general election). There are also limits on the amount that these PACs can give themselves ($3,300 each for primary and general election and $5,000 each for primary and general election when giving to other PACs). The main way that corporations involve themselves in electioneering efforts is through what are known as corporate PACs. While corporations are not allowed to directly allowed to donate to candidates with their own funds, they can operate corporate PACs with company funds, anybody including company executives can contribute to the money that goes directly to candidates, and the firm can provide incentives for individuals to donate. Furthermore, while it falls outside of the scope of this work, individuals and corporations alike have no restrictions on many types of indirect funding, including Super PACs. America's laws are very different than most of the first world, where candidates have large amounts of public financing at their disposal and have either strict limits or ban entirely the ability for private donations.

As such America has the most expensive elections on Earth, and elections have ballooned in costs over the past several decades. The overall cost in terms of 2020 dollars of all House and Senate elections increasing from $1.32 billion in the 1992 election cycle to $3.97 billion in the 2020 election cycle. Much of this increase is attributed to industry special interest groups such as pharmaceuticals. In terms of 2020 dollars, the overall spending from companies and trade groups associated with pharmaceutical industry increased by over four-fold from $15.0 million in 1992 to $92.0 million in 2020.

## 2.5 The American Legislative Exchange Council (ALEC)

ALEC is a controversial conservative and free market lobbying group dedicated to pushing their ideals in state legislatures across the countries. This organization and its executive structure is made up of state legislators, corporate lobbyists, company executives. The corporate executives on ALEC's executive board are known as its "Private Enterprise Board of Directors." ALEC's main task is to draft model legislation to push priorities in all different policy areas. They then circulate this draft legislation to legislators in state houses, who will many times use this legislation to create their own legislation, eventually becoming state law. The group has been heavily criticized since while the final model legislation is

publicly available, much of the crafting of this legislation is done in secret. One thing that is nearly certain though is that its presence has allowed corporations to greatly impact the legislation passed in the United States on the state level. It is worth emphasizing that ALEC's activities are not aimed at influencing the federal government.

### 2.6 Pharmaceutical Research and Manufacturers of America (PhRMA)

PhRMA is an American trade organization founded in 1958 that represents the pharmaceuticals industry, with current member companies including Bayer, Johnson and Johnson, and Pfizer. They actively push for priorities that benefit the profits of these organizations such as being against letting Medicare negotiate drug prices and increasing drug price transparency. They are heavily involved in activities that influence the government. They donate heavily to many right-wing dark money groups and also are heavily involved in lobbying the US government. While less prolific, they also have a corporate PAC that helps directly support candidates in all types of elections who are in line with their interests, often working to complement their extensive lobbying efforts. PhRMA is also a member of ALEC's Private Enterprise board of representatives, having been a member for over a decade. Therefore, the healthcare model legislation that ALEC produces inherently conveys the interests of PhRMA and its member companies.

## 3 Theory

### 3.1 Setting

Consider a static world with two endogenous actors, a representative legislator seeking reelection and a profit maximizing firm. There is also a an exogenous constituency. The legislator increases the votes they receive $v$ by setting policy more favorable to the constituency's median voter and by receiving more in donations from the firm, denoted as $d$, which it obtains by setting policy more favorable to them. This firm favorability is denoted by $f$, and the following assumptions are

made regarding it:

**Assumption 1a:** $f \in [0, \infty)$.
**Assumption 1b:** The median voter's preference is for $f = 0$

Policies that are more favorable to the firm are assumed to be less favorable to the median voter, and vice versa, but the politician offset their less favorable actual policy choices by presenting themselves in a better light to the public, doing so with money obtained from donations. Firms receive higher profits $\Pi$ from spending more on investment $x$ and from facing more favorable policy, which it obtains by giving more in donations. However, the firm cannot make a profit if they do not spend any money on this investment. The only source of donations for the legislator in this model is those from the firm.

## 3.2 The Legislator's Problem

Formally, the legislator seeks to maximize the number of votes they will receive in an election $v = v(f, d)$ by choosing a function maximizing level of $f$. The function $v(f, d)$ is decreasing and convex in $f$ (trade-off between constituent and firm favorability) and is increasing and concave in $d$ (more favorable policy for the firm yields more in donations from the firm). Firm donations are represented by the function $d(f)$ that is increasing and concave (firms will donate more to the legislator if they give the firm more favorable policy), where $d(0) = 0$ (firms will not donate to the legislator if they set their least favorable policy).

The problem of the legislator can thus be expressed by Equation (1) below:

$$\begin{aligned} &\max_{f \in [0,\infty)} \; v(f, d) \\ &\text{where } d = d(f) \end{aligned} \tag{1}$$

The legislator's first order condition to the problem in Equation (1) is represented by the inequality shown in Equation (L)

$$\frac{\partial v}{\partial d} d'(f) \leq -\frac{\partial v}{\partial f} \tag{L}$$

In the above expression, $\frac{\partial v}{\partial d}d'(f)$ represents the marginal benefit of the legislator increasing favorability to the firm, that benefit being the increase in votes from increasing firm donations. $-\frac{\partial v}{\partial f}$ is the marginal cost, the loss in the value function from setting policy further away from the median voter's preference of $f = 0$.

The behavior of this first order condition can be broken into two cases:

**Case 1:** $\frac{\partial v}{\partial d}d'(f) < -\frac{\partial v}{\partial f}$ for all $f \geq 0$.

**Case 2:** There exist values of $\hat{f}$ such that $\frac{\partial v}{\partial d}d'(\hat{f})|_{\hat{f}} = -\frac{\partial v}{\partial \hat{f}}$.

**Legislator's Solution:** If no $f \geq 0$ solves Equation L (Case 1), the legislator maximizes $v$ by setting $f = f^*0$. Otherwise (Case 2), they set their policy to a level of $f = f^*$ that solves Equation L with strict equality.

## 3.3 Firm's Problem

Formally, the firm seeks to maximize a profit function $\Pi(x, f)$. The firm chooses optimal values of $x$ and $d$ given their exogenous monetary endowment of $m > 0$. The function $\Pi$ is increasing and concave in both $x$ and $f$ (higher investment and more favorable policy lead to higher profits). From the firm's perspective, $f = f(d)$, where $d \geq 0$. Assumption 2 below is made to ensure non-zero values of $x$ :

**Assumption 2:** $\frac{\partial \Pi}{\partial x}|_{(x=0,d=0)} > \frac{\partial \Pi}{\partial f}f'(d)|_{(x=0,d=0)}$

Assumption 2 states that at $(x = 0, d = 0)$, marginal profit of the firm is higher from increasing $x$ rather than from increasing $f$.

The problem of the firm can be expressed by Equation (2) below:

$$\begin{aligned} &\max_{(x,d)} \Pi(x, d) \\ \text{s.t} \quad &m = x + d \\ &d \geq 0 \\ &\text{where } d = d(f) \end{aligned} \tag{2}$$

The first order conditions with respect to $x$ and $d$ can be represented by then (F)

$$\frac{\partial \Pi}{\partial f} f'(d) \leq \frac{\partial \Pi}{\partial x} \tag{F}$$

The behavior of the inequality in F can be broken into two cases:

**Case 1:** $\frac{\partial \Pi}{\partial f} f'(d) < \frac{\partial \Pi}{\partial x}$ for all $d \geq 0$.

**Case 2:** There exist values of $\hat{d}$ such that $\frac{\partial \Pi}{\partial f} f'(\hat{d})|_{(x=m-\hat{d}, d=\hat{d})} = \frac{\partial \Pi}{\partial x}|_{(x=m-\hat{d}, d=\hat{d})}$.

**Firm Solution:** If no $d \geq 0$ solves Equation F (Case 1), the firm maximizes $\Pi$ by setting $d = d^* = 0$ and $x = x^* = m$. Otherwise (Case 2), they set their level of donations $d = d^*$ to the level that solves Equation F with strict equality, with $x = x^* = m - d$.

## 3.4 General Equilibrium

A legislator's policy level $\bar{f}$ and a firm's donation level $\bar{d}$ constitutes a general equilibrium if $\bar{f}$ and $\bar{d}$ satisfy the following conditions:

1. $\bar{f} = f^*$
2. $\bar{d} = d^*$
3. $f(\bar{d}) = \bar{f}$
4. $d(\bar{f}) = \bar{d}$

This model demonstrates that donations and legislation favorability likely have a positive relationship. It also demonstrates a potential pitfall. That is, donations impact language favorability, while simultaneously, language favorability impacts donations. In order to analyze this relationship in the context of regression analysis, both donations and language favorability must be treated as a dependent variable in a set of regressions with the other being run as the main independent variable of interest, but doing so directly is incorrect from an empirical standpoint. This dilemma will be addressed in the Empirical Analysis section.

Note: the functional form, comparative statics, etc analysis is a work in progress. I leave it to the next version

# 4 Data Collection and Variables

As mentioned in the introduction a panel of four Congress cross-sections, starting with the 113th Congress (2013/14) and ending with 116th Congress (2019/20) is utilized.[6] Observations in this panel are the data of a particular US House member for a particular term in the above time-frame that both served their entire term and did not switch parties in the middle of their term.

The most intricate aspect of the data collection and variable creation process involved the variables for PhRMA language favorability of House members. As mentioned previously, this paper utilizes two different NLP modeling techniques to do this. It uses term-frequency (tf) models, a model which is used to compare documents based on their phraseologies, and LDA, a model which is used to compare models based on their subject matters. I will briefly introduce these two methods along with some basic NLP terminology in the next subsection.

## 4.1 Natural Language Processing Methods

An n-gram is a sequence of "n" words, where "n" refers the length of the sequence. Unigrams refer to n-grams with $n = 1$ and bigrams refer to when $n = 2$. For illustration, the sentence "bob likes spicy food" converted into unigrams (n=1) would be [bob, likes, spicy, food] and converted into bigrams (n=2) would be [bob-likes, likes-spicy, spicy-food]. A "corpus" is the collection of documents utilized in the creation of a language model. In models utilizing n-grams, a "vocabulary" refers to the set of unique n-grams that can be found in the corpus. For example, in a unigram based model, a corpus consisting only of the phrase "this is bob and this is his dog," the vocabulary would be [this, is, bob, his, dog]. A "vocabulary element" refers to an n-gram found in the vocabulary.

Term frequency and LDA models all belong to what is known as the "bag-of-words" class of model. Bag-of-words models treat documents as an unordered list of n-grams. This is in contrast to models that take into account the ordering of phrases, sentences, and paragraphs. In bag-of-words models, rearranging the order

[6] Since congressional borders were redrawn in the 117th Congress, data from this Congress has been excluded.

of the n-grams of a document will not change how the document is quantified. It is an assumption that in some circumstances can be quite restrictive, but not when analyzing legislation since the order of sections and paragraphs is not relevant when comparing two legislation documents. Furthermore, a bag-of-words model with n>1 still somewhat preserves elements of the ordering of words, as changing the order of two words will end up changing the composition of the n-grams that the document contains. The term frequency models used in this paper utilized bigrams. The LDA model utilized unigrams.

These models also belong to the class of models known as "vector-space" models, in that the modeling process generates a vector-space and documents are represented as vectors in that vector-space. An appealing feature of vector-space language models is that the similarity of two documents can be calculated as the cosine between their vector representations (called cosine-similarity). Term frequence and LDA models are appealing choices for economists as in both models, the dimensions of the vector-spaces and the coordinates of the document vectors have tangible meaning. This is in contrast to methods such as word-embeddings, where the vector-space dimensions are created in a machine-learning process that produces vector-spaces with dimensions uninterpretable to the human reader.

### 4.1.1 Term Frequency

The first model I will detail is the term frequency model. Its origins can be found in Luhn(1958), and is the most commonly used language model found in the NLP found in the economics literature. A term frequency model classifies documents based on counts of the n-grams (in this model, bigrams) they contain, or the "frequency" of the "terms" is contains. Let $\mathbf{V} = \{v_1, \ldots, v_e, \ldots, v_E\}$ represent the vocabulary of a corpus containing $E$ elements. The vector-space generated by a term frequency model has $E$ dimensions, with each dimension representing an n-gram found in the vocabulary. The term frequency vector representation of the document $d_i$ is a $1 \times E$ vector

$$\Theta_i^{tf} = [\theta_{i,1}^{tf}, \ldots, \theta_{i,e}^{tf}, \ldots, \theta_{i,E}^{tf}]$$

where $\theta_{i,e}^{tf}$ is the number of times the vocabulary element $v_e$ occurs in the document $d_i$. NOTE: The use of alcohol in the examples is done in order to facilitate the

LDA example that follows, where a word can be found in both topics. If this is inappropriate, I can try to find something else.

Consider two documents in a corpus that's vocabulary contains two words, "wine" and "beer." Document $d_1$ contains the word wine once and beer four times and document $d_2$ contains the word wine three times and beer twice. The vector representation of $d_1$ and $d_2$ in a unigram term frequency model can be seen in Figure 1. The cosine of $d_1$ and $d_2$ is 0.74, so the cosine similarity of $d_1$ and $d_2$ is 0.74.

### 4.1.2 Latent Dirichlet Allocation (LDA)

To incorporate a modeling method that does take document subject matter into account, "Latent Dirichlet Allocation" language models, first proposed by Blei, Ng, and Jordan (2003), will also be utilized. Rather than being classified by counts of n-grams, documents in an LDA model can be thought of as weighted mixtures of topics. These topics are in turn classified as weighted mixtures of words. The weighted mixtures are more formally categorical distributions (for information about the categorical distribution, refer to D.2 whose probability parameters are obtained from draws of Dirichlet distributions (for information about the Dirichlet distribution, refer to D.4).

LDA assumes that documents were created word by word with a specific word-independent random process:

1. From a Dirichlet distribution of topic probabilities, draw the probabilities for each document's categorical distribution of topics.
2. From a Dirichlet distribution of word probabilities, draw the probabilities for each topic's categorical distribution of words.
3. For each eventual word in each document, pick that word's topic from the document's categorical distribution of topics.
4. For the topic that was drawn in the previous step, draw a word from that topic's categorical distribution of words.

Suppose that documents can pertain to any number of the $S$ topics in $\mathbf{A} = [a_1, \ldots, a_s, \ldots, a_S]$. The vector-space of interest generated by the LDA model has $S$ dimensions, or one dimension for each topic. The coordinates of the vector representation of document $d_i$ is the $1 \times S$ vector

$$\Theta_i^{lda} = [\theta_{i,1}, \ldots, \theta_{i,s}, \ldots, \theta_{i,S}]$$

where $\theta_{i,s}$ is the probability of drawing the topic $a_s$ in step 3 of the document generation process. Consider an LDA model whose documents can belong to one of two topics, alcohol and grapes. The document $d_3$ has the topic distribution $0.5 * \text{alcohol} + 0.5 * \text{grapes}$. In other words, when the topics for the words in $d_3$, there is an 50% chance that the word will come from the topic "alcohol" and a 50% chance that it will come from the topic "grapes." The document $d_4$ has the topic distribution 0.8∗alcohol+0.2∗grapes. In other words, when the topics for the words in $d_3$, there is an 80% chance that the word will come from the topic "alcohol" and a 20% chance that it will come from the topic "grapes." Furthermore, assume that the topic alcohol's distribution is $0.5 * \text{wine} + 0.5 * \text{beer}$, and the topic grape's distribution is $= 0.5 * \text{wine} + 0.5 * \text{jam}$. The vector representations of $d_3$ and $d_4$ are shown in Figure 2. Again, the vector coordinates of $d_3$ and $d_4$ correspond to the probabilities of drawing alcohol and drawing grapes in the process of constructing the document (the distributions of alcohol and grapes themselves are irrelevant). The cosine of $d_3$ and $d_4$ is 0.857, so the cosine similarity of $d_3$ and $d_4$ is 0.857. Obviously, the only observable information prior to training an LDA model are the words in the documents, with the underlying distributions being latent (hence the "latent" in "Latent Dirichlet Allocation"). The process of training an LDA model involves inputting a corpus and some assumed parameters (such as the number of topics and the initial Dirichlet parameters) into an iterative machine learning process to back out the final Dirichlet and categorical distributions. Aside from the fact that the Dirichlet parameters are symmetric and sparse (giving tendency for the categorical distributions to favor certainty), knowledge of this machine learning process is not required for understanding the results of this paper.

## 4.2 PhRMA Favorability Variables

The two Natural Language Processing methods outlined will be utilized to create measures for how favorable a lawmaker's legislation is to PhRMA's interests. Four

different PhRMA favorability measures in total were created, that is two different measures for each of the three different types of models. The process for creating these variables can be summarized in four steps:

1. The text and data of all healthcare US House and ALEC model legislation from 2013-2020 was obtained, with a unigram and bigram corpus being created for each Congress.

2. A term frequency and LDA language model was created for each Congress.

3. The cosine-similarities between each real bill and each model bill in each language model was obtained. The highest such value between a real bill and any model bill in that space were designated as that bill's term frequency and LDA "ALEC Similarities"

4. Two measures of a lawmaker's "PhRMA" favorability were created from each of the three language models:

   (a) The highest ALEC similarity from the term frequency/LDA space of any bill a lawmaker sponsored in a Congress, referred to as "PhRMA maximum favorability".

   (b) The sum of the ALEC similarities from the term frequency/LDA space of all bills the lawmaker sponsored in a Congress "PhRMA sum favorability".

### 4.2.1 Text Data Collection and Preprocessing (Step: 1)

US House of Representatives bills, resolutions, and amendments introduced from the 113th Congress (2013/14) through the 116th Congress (2019/20) were collected. The decision was made early on to utilize all introduced legislation rather than legislation passed, since legislation that doesn't pass can still be used to signal policy priorities to special interest groups. Data on the subject matter of the legislation, the legislation's sponsor, and the date that the legislation was introduced were also collected. This text and data was obtained by web scraping Congress's official website;[7] the Python language, the Selenium package, and the

[7] https://www.congress.gov

Google Chrome web-driver ChromeDriver were used to do this. Pieces of legislation were segregated by the Congress they were introduced in and by their subject matter, with only legislation with the subject "health" being utilized in this paper's textual analysis.

All ALEC model legislation that was finalized from 2013 through 2020 was web scraped. The policy subject and finalization date[8] of each ALEC model bill was also collected. This text and data was all collected by web scraping ALEC's website.[9] Each model bill was assigned to the Congress that coincided with the year it was finalized. So for example, model legislation finalized in the years 2013 and 2014 were assigned to the 113th Congress. The text of both the real and model healthcare bills from a Congress was used to create the two language models for that Congress. The amount of real and model legislation from each Congress can be found in Table B. As you can see, there is far more real than model legislation introduced in each congress.

To prepare for the modeling stage, real and model healthcare legislation from each Congress was then "preprocessed" (converted into the form needed for language model generation); Python's spacy and NLTK packages was used to do this. All of the words in each document were converted to lowercase, and punctuation, Arabic, and Roman numerals were removed. Hyphenated words were maintained by removing the hyphen and joining the words together. So for example, the word "chocolate-covered" would become "chocolatecovered." Each document was then converted into a list of single words, or tokens, in a process known as "tokenization". Stopwords, or words that provide no additional context to the analysis of the texts (such as "the," "a," "of," "is," etc.), were then removed from each document's tokens. In addition to the built-in list of stopwords from the spacy package, additional stopwords in a legislation context such as "article," "section," "act," and "chapter.") were also removed. A full list of stopwords can be found in Appendix F. The preprocessed and tokenized set of real and model documents is that Congress's unigram corpus, and was used directly to create that Congress's LDA model. That Congress's bigram corpus for the term frequency models were then created from the corpus of unigrams.

[8]Some of the model legislation pieces have been updated or revised.

[9]https://alec.org

### 4.2.2 Language Model Creation (Step 2)

Next, each Congress's term frequency model was created from that Congress's bigram corpus and each Congress's LDA model was used to create that Congress's LDA model. Since the models were not used to analyze documents outside of the corpora themselves, the corpora were not divided into train and test sets. The term frequency models were created by hand, and the LDA models utilized the gensim package. The number of topics assumed in a corpus is 50 (default is 100) due to the smaller corpus.[10] The remaining parameters for the LDA model were the package defaults. Because the language models are not being applied to any external documents, the corpora were not split into train and test sets, and all documents from the congress were used to generate the models. I will use the superscript $n$ to refer generically to the language model, that being either term frequency or LDA. Let $r^n_{t,b}$ be the language model $n$ vector representation of the $b$'th real Congressional bill from Congress $t$, and let $m^n_{t,a}$ be the language model $n$ vector representation of the $a$th model bill from Congress $t$.

### 4.2.3 Cosine Calculation and ALEC Similarity (Step 3)

With all language vector spaces created, the cosine between every real bill $r^n_{t,b}$ from Congress and every model bill $m^n_{t,a}$ was calculated for every congress using every model bill. That is, $\cos(r^n_{t,b}, m^n_{t,a}) \forall b, a$. very cosine between every $r^n_{t,b}$ and $m^n_{t,a}$, $\cos(r^n_{t,b}, m^n_{t,a})$. Because ALEC bills have different policy priorities, the similarity of a particular real bill $\bar{b}$ to ALEC's legislative priorities in Congress $t$ using language model $n$ will be defined as the highest cosine between that real bill's vector $r^n_{t,\bar{b}}$ and any model bill vector $m^n_{t,a}$. In other words, the ALEC Similarity of the real bill $\bar{b}$, defined as $s^n_{t,b}$ is calculated as $\max\{\cos(r^n_{t,\bar{b}}, m^n_{t,1}), \ldots, \cos(r^n_{t,\bar{b}}, m^n_{t,A})\}$. For example, suppose in a particular congress $\bar{t}$, there were two real bills and 3 model bills. Language model method $\bar{n}$ yields the vector space shown in 3. As you can see, the ALEC bill vector that has the highest cosine with real bill vector $r^{\bar{n}}_{\bar{t},1}$ (the ALEC bill that creates the smallest angle with $r^{\bar{n}}_{\bar{t},1}$) is $m^{\bar{n}}_{\bar{t},1}$, and the ALEC bill vector that has the highest cosine with real bill vector $r^{\bar{n}}_{\bar{t},2}$ is $m^{\bar{n}}_{\bar{t},2}$. As such, the ALEC similarity of the first real bill, $s^n_{t,1} = \cos(r^n_{t,1}, m^n_{t,1})$ and the ALEC similarity for the second real bill, $s^n_{t,2} = \cos(r^n_{t,2}, m^n_{t,2})$.

[10] The change from 50 to 100 does not cause substantial changes to the results that follow.

### 4.2.4 Legislator PhRMA Favorability Scores (Step 4)

After all real bills in all sessions have had their ALEC Similarity scores calculated for the two language models, this data is then utilized to determine how friendly legislators are ALEC's (and thus PhRMA's) legislative priorities. Two measures of a legislator's PhRMA favorability were created. The first measure is the PhRMA sum favorability, or the sum of all ALEC similarity scores in a Congress calculated with that modeling method. This measure captures favorability of a legislator from their total legislation output in a Congress. The second measure is the PhRMA maximum favorability, the highest ALEC Similarity of any bill they sponsored in a Congress from each modeling method. This measure captures favorability of a legislator from sponsoring a particular bill that is close in line with a particular PhRMA priority. This means there are four different favorability scores, given that two different language models being used to create two different PhRMA favorabilities each. Legislators who did not sponsor healthcare legislation in a given session are given values of zero for all four of their favorability scores. This is logical, since legislators who did not sponsor any healthcare legislation in a Congress did not benefit PhRMA through their sponsored healthcare legislation in that Congress.

Let $s_{t,b}^{n,i}$ be the the modeling method $n$ ALEC similarity of real bill $b$ from congress $t$ sponsored by legislator $i$, and let $S_t^{n,i}$ be the set of ALEC similarity scores of legislation sponsored by legislator $i$ in Congress $t$ using modeling method $n$. The PhRMA sum Favorability of legislator $\bar{i}$ in Congress $\bar{t}$ from modeling method $\bar{n}$ i calculated as $\sum_{s \in S_{\bar{t}}^{\bar{n},\bar{i}}} s$. The PhRMA maximum Favorability of a legislator $\bar{i}$ in Congress $\bar{t}$ from modeling method $\bar{n}$ is calculated as $\max\left\{S_{\bar{t}}^{\bar{n},\bar{i}}\right\}$. To illustrate how the PhRMA favorability scores of legislators are creatred from the ALEC Similarities of the legislation they sponsor, consider a fictional representative, John Doe, who served during the 113th Congress. During the 113th Congress, he sponsored five pieces of healthcare legislation during this congress. The ALEC similarities for these five bills can be found in Table B. Adding all of the term frequency ALEC Similarities together yields 0.28 and adding all of the LDA ALEC Similarities yields 2.572. As such, John Doe's 113th Congress term-frequency PhRMA sum favorability is 0.28 and the LDA PhRMA sum favorability is 2.572. The maximum term frequency ALEC similarity of the five bills is 0.1 (bill 2) and the maximum and the maximum LDA ALEC Similarity of the five bills is 0.999 (bill 4). As such,

John Doe's 113th Congress term-frequency PhRMA max similarity is 0.1 and his LDA max similarity is 0.999.

PART 1 CUTOFF

## 4.3 Campaign Finance Data

The campaign donations data from the 113th through 116th Congresses was downloaded directly from the non-profit group OpenSecrets[11] who aggregates data from the Federal Election Commision's publicly available donations data. A Congressperson's donations given from PhRMA (and all other corporations and institutions) are classified into three categories, PAC donations, individual donations, and total donations. Donations that candidates or candidate specific PACs receive from PhRMA related PACs are classified as PhRMA PAC donations. Donations given by individuals who list PhRMA as their place of employment to candidates and candidate specific PACs are classified as individual donations. Total PhRMA donations to a candidate are the sum of PhRMA's PAC donations and PhRMA's individual donations. PAC donations were utilized as the campaign finance variable of interest as it is the cleanest variable for PhRMA's direct investment into the candidates.

## 4.4 Control and Fixed Effects Data

The regressions conducted in the empirical analysis were all controlled for the legislator's vote total in the previous election and the legislator's party, and fixed effects were utilized for Congress and for geographic location. Election data from 2010-2018 was obtained from the MIT Election Data and Science Lab.[12] Party and location data was obtained from the [Party-State-District] code from the OpenSecrets data[13] or was input manually in the case of the Congress fixed effects.

[11] opensecrets.org

[12] https://electionlab.mit.edu/data

[13] For example, Democrat Alexandria Ocasio-Cortez from New York's 14th district is referred to as "Alexandria Ocasio-Cortez [D-NY-14]"

The candidate's vote total in the previous election was utilized as a proxy variable for the candidate's electability in the current Congress. The variable utilized for this was the proportion of the vote among the two major candidates. For example, if in the 2012 election, the winning Democrat got 60% of the vote, the Republican Republican got 20% of the vote, and third party candidates and independents received the remaining 20% of the vote, the value of recorded for the Democrat in their 2014 observation is 80% (recorded as 0.8). Electability impacts the Congressperson's fundraising, as incumbents who are in more danger of losing their reelection will likely need to fundraise more. It also impacts impacts their PhRMA favorability scores. Since PhRMA friendly legislation is oftentimes less favorable to the general public, congresspeople who are more likely to win reelection may be more likely to write more PhRMA friendly legislation, all else equal.

## 4.5 Descriptive Statistics

A standard table of descriptive statistics can be found in Table 3. The distributions for all four favorability measures can be seen in the histogram found in Figure 4 and the distributions for these measures removing legislators not sponsoring legislation can be found in Figure 5. The majority of the favorability variables when including only authors have a mode near zero, and demonstrate an exponential distribution-like shape. The exception to this is the LDA maximum favorability, which has modes at zero and one, with relative uniformity in between.

Categorical descriptive statistics were obtained for the language variables and donations data. A table of the PhRMA favorability variables and donations data separated by party can be found in Table 5. Figure 6 shows the mean PhrMA favorabilities for each Congress split by party, and 7 shows the mean donations for each Congress split by party . Given the conservative leaning of ALEC's model legislation and PhRMA's support of conservative candidates, one might expect both the PhRMA favorability and levels of PhRMA donations to be higher for Republicans than Democrats. But while there overall is a difference in PhRMA favorability means between Republicans and Democrats, the picture is murky though when separating the PhRMA favorabilities by Congress. For both term frequency sum and maximum favorabilities, Republicans seem to have higher favorabilities for Congresses 113 through 115, LDA sum favorability is relatively unaffected by

party in these Congresses, and Democrats have higher LDA maximum than Republicans in these three Congresses. All PhRMA favorabilities though are higher on average for Democrats and not Republicans in the 116th Congress. Furthermore, across all Congresses, there is in fact no statistically significant difference in PhRMA donations means between Republicans and Democrats. When looking at the differences between parties, PhRMA donations are higher for Republicans in the 113th through 115th Congresses, with donations being higher for Democrats than Republicans in the 116th Congress. However, it is clear from looking at this that the behavior of both PhRMA donations and PhRMA favorability over time is party dependent.

The difference between PhRMA favorability scores of Representatives receiving donations from PhRMA and those that did not is much more clear cut, as can be seen in Table 6 and Figure 8. The mean of the PhRMA favorability variables among Representatives receiving donations from PhRMA PACs is statistically significantly higher than among Representatives that did not receive any, and the PhRMA favorability variables of all four types is higher in each Congress among Representatives receiving donations verses those that did not. A similar pattern emerges when looking at PhRMA donations when splitting legislators between those who wrote at least one healthcare bill in a Congress (regardless of language favorability) and those that did not, as can be seen in Table 7 and Figure 9. Not only are the mean donations of those sponsoring healthcare legislation higher than those that do not, but the mean among healthcare sponsors is higher in every Congress than the mean among those that did not. All in all, this provides preliminary evidence of a likely relationship between PhRMA PAC donations and PhRMA legislation favorability, a relationship that will be confirmed in the empirical work to follow.

# 5 Empirical Analysis

Two sets of regressions will be run to confirm the positive relationship between PhRMA legislation favorability and PhRMA PAC donations. The first utilizes the PhRMA favorability variables lagged by one Congress as the independent variable with non-lagged PhRMA PAC donations as the dependent variable. The second utilizes PhRMA PAC donations lagged by one Congress as the independent variable with non-lagged PhRMA favorability variables as the dependent variable.

These two sets of regressions are run due to the bidirectional relationship between donations and legislation favorability in the same congress. With the lagged variables, regressions will utilize a three-period unbalanced panel from the 114th through 116th Congresses, with the data from any lagged variables coming from the 113th through 115th congresses.

Let $i$ refers to the specific congressperson, $t$ refer to the Congress, and $k$ refers to either term frequency sum, term frequency maximum, LDA sum, or LDA maximum PhRMA favorability. In both sets of regressions, Congress FEs contains indicator variables for the 114th and 115th Congresses and their interactions with the GOP indicator, location FEs contains US state indicator variables and their interactions with the GOP indicator variable, and $\tilde{\epsilon}_{i,t}$ is the robust standard error clustered on state, party, and congress.

The equation for the regressions of lagged PhRMA PAC donations on non-lagged PhRMA favorability can be found in Regressions B below:

$$\begin{aligned} \text{FAVOR}_{i,t} = \alpha + \beta \cdot \text{DONATIONS}^{k}_{i,t-1} + \gamma^{V} \cdot \text{VOTES}_{i,t-1} + \gamma^{G} \cdot \text{GOP}_{i,t} \\ + \Gamma^{C} \cdot \textbf{CONGRESS FEs} + \Gamma^{L} \cdot \textbf{LOCATION FEs} + \tilde{\epsilon}_{i,t} \end{aligned} \tag{B}$$

The equation for the regressions of lagged PhRMA favorability on non-lagged PhRMA PAC donations can be found in Regressions A below:

$$\begin{aligned} \text{DONATIONS}_{i,t} = \alpha + \beta \cdot \text{FAVOR}^{k}_{i,t-1} + \gamma^{V} \cdot \text{VOTES}_{i,t-1} + \gamma^{G} \cdot \text{GOP}_{i,t} \\ + \Gamma^{C} \cdot \textbf{CONGRESS FEs} + \Gamma^{L} \cdot \textbf{LOCATION FEs} + \tilde{\epsilon}_{i,t} \end{aligned} \tag{A}$$

In addition to the controls mentioned previously, both sets of regression include both Congress and Location fixed effects. The Congress fixed effects are straight forward, an indicator for the 114th and 115th Congresses, with the 116th Congress being the base. Given the relationship between both the PhRMA favorability variables and donations differs over time, interaction terms between Congress and the GOP indicator were also included. The choice of using state fixed effects along with their interactions with the GOP indicator was made in order to best control for location without introducing over-identification. The finest level of fixed effect that could theoretically be utilized for location would be effects for Congressional district. However, given the small number of observations in the

panel, this could lead to issues of over-identification. Furthermore, using just state fixed effects without their interactions does not adequately control for the diversity of congressional districts within a state. The interaction with party was included in order to incorporate the fact that there are often large differences between the Democratic and Republican districts in a states, with the Democratic districts tending to be more ethnically diverse and more urban and Republican districts tending to be more Caucasian and more rural. Robust standard errors clustered on state-party-Congress were also utilized to control both for the heteroskedasticity and for correlation of the standard errors between clusters.

**MAJOR NOTE** The day I'm sending this copy to both of you, I had the idea to add the lagged DV as an independent variable. So regression A would have $\text{DONATIONS}^k_{i,t-1}$ as an independent variable and regression B would have $\text{FAVOR}^k_{i,t-1}$ as an independent variable. For empirical purposes, this REALLY needs to be done, since the donations and favorability are both at least AR(1) (I tested it). When running the regressions, coefficients of interest seem to hold at the 5% level. I'm having issues coding it in the loops I have with Stata, because the only major change will be mentioning the autoregressiveness and the numbers in the results, I'm going to send the version without this now.

# 6 Results

In addition to the regularly reported regression results, additional results are included for both sets of regressions where PhRMA PAC donations, PhRMA favorability, and vote share control data are converted to standard deviations above their four-Congress pooled means to help facilitate interpretation of the results. Table 10 shows the results of Regressions A normally reported. All coefficients of previous Congress PhRMA favorability on current Congress PAC donations in Regressions have positive sign and are statistically significant at the 1% level. Increasing a legislator's previous Congress's term frequency sum, LDA sum, term frequency maximum, and LDA maximum leads to an increase in the PhRMA donations that a legislator receives in the current Congress by \$440.90, \$105.80, \$752.90, and \$239.50 respectively. As seen in Table 11, increasing the previous Congress's term frequency sum, LDA sum, term frequency maximum, and LDA maximum by one standard deviation increases PhRMA donations that a legislator receives in the current Congress by 0.107, 0.147, 0.091, and 0.135.

Table 8 shows the results of Regressions B normally reported. Again, all coefficients of previous Congress PhRMA PAC donations on PhRMA favorability also have positive sign and are also statistically significant at the 1% level. Increasing a legislator's previous Congress PhRMA PAC donations by one dollar leads to an increase of the term frequency sum, LDA sum, term frequency maximum, and LDA maximum PhRMA favorabilities in the current Congress by $3.38 \times 10^5$, $1.35 \times 10^5$, $1.67 \times 10^4$, and $6.29 \times 10^5$ respectively. As can be seen in Table 9, increasing a legislator's previous Congress's PhRMA PAC donations by one standard deviation leads to the current Congress term frequency sum, LDA sum, term frequency maximum, and LDA maximum PhRMA favorabilities increasing by 0.024, 0.010, 0.118, and 0.047 standard deviations respectively.

Some of the coefficients and significance of the controls and fixed effects were intriguing. The main one is the negative and statistically significant coefficients on the GOP indicator variable. Some of this can be explained by the base year in these regressions being the 116th Congress. As mentioned previously, this was the one Congress in which PhRMA donated more to Democrats rather than Republicans. However, when including the Congress fixed effects and the Congress-GOP interactions, the overall impact of being a Republican is negative in these regressions. This is in line though with the descriptive statistics performed earlier in the paper though. Another interesting finding is that the previous election vote share is statistically insignificant with respect to both campaign donations and the PhRMA favorability variables.

# 7 Conclusion

Given that legislators are incentivized to raise as much as possible for reelection efforts, it is easy to conclude that they will act in a manner while in office that will allow them to do this more easily. And given that corporations are profit maximizing entities, it is easy to come to the conclusion that if corporate interests are in fact raising money for candidates, they are doing so with the sole intention of increasing profits. Therefore, it is hard to imagine a world in which corporate interests are raising money for candidates, and there is no relationship between these donations and the actions of government officials. This paper provides evidence that such a relationship exists, and is indeed the first one to do just this.

It does this by first creating a model from first economic principles that demonstrates how the incentives of both politicians and special interest groups lead to a direct relationship between donations and legislation favorability. It then utilizes NLP methods to test this empirically. Using ALEC's model legislation as a text base conveying PhRMA's priorities, I find that PhRMA's PAC donations and a legislator's favorability to PhRMA are in fact linked. There are many other special interest groups that can be analyzed empirically in such a menner, and likely several ways to fine-tune these methods. I leave both for future work.

Bibliography forthcoming

# A Figures

Figure 1: Two Documents in Term Frequency Space

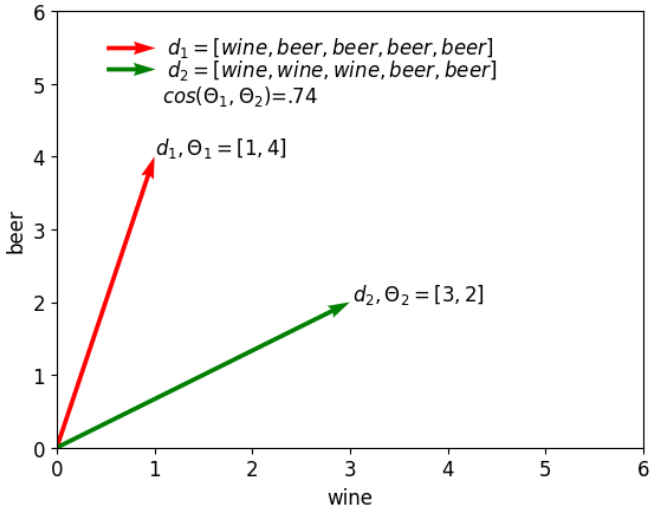


Figure 2: Two Documents in LDA Space

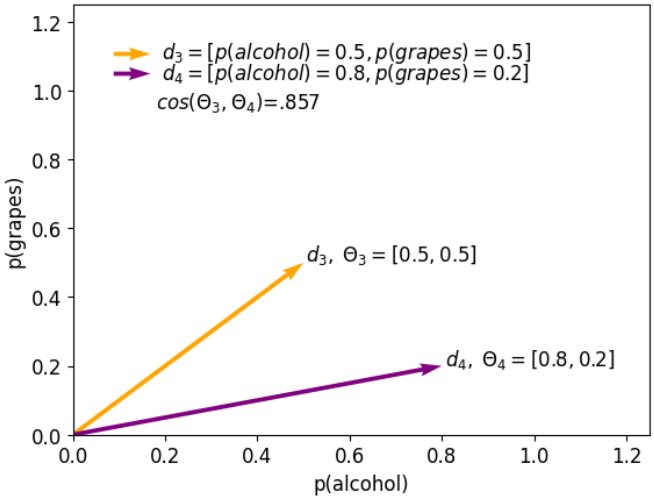



Figure 3: A hypothetical 2D vector representation of two real documents and three model documents.

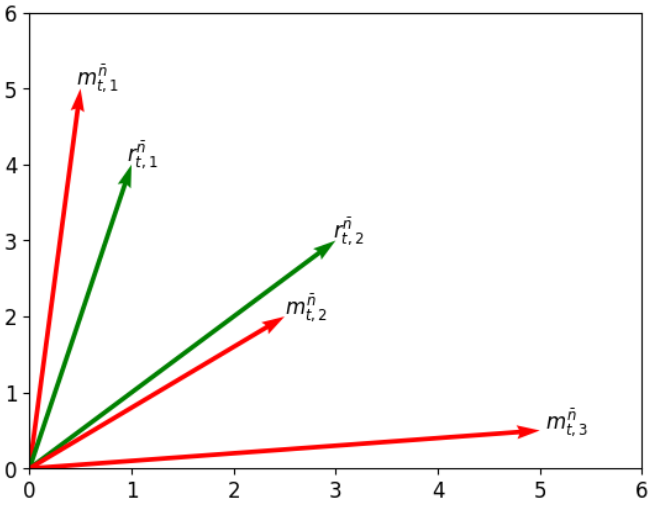


Figure 4: Language Score Distributions, All Observations

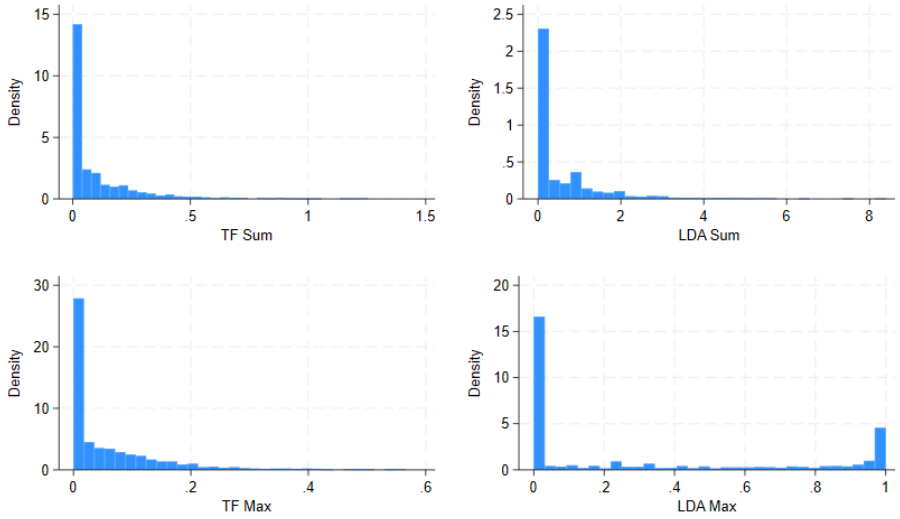



Figure 5: Language Score Distributions, Observations with at least one Healthcare Bill

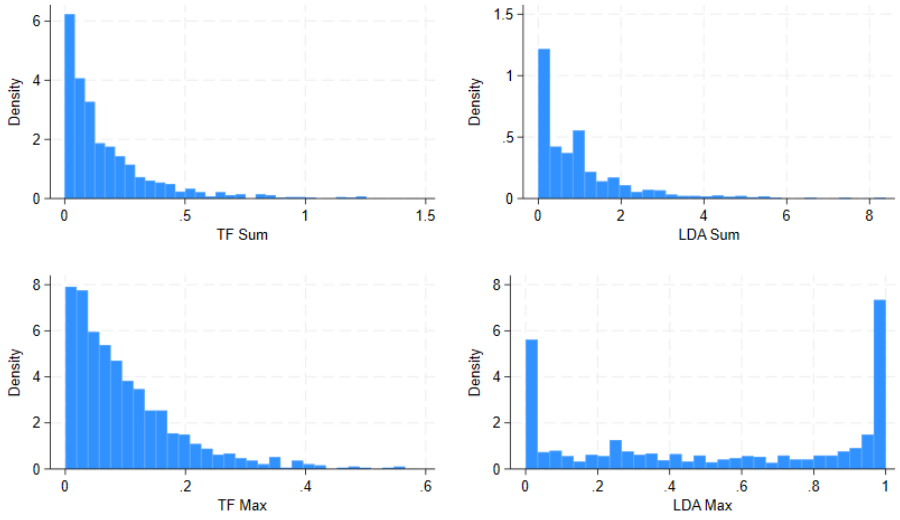


Figure 6: PhRMA Favorabilities, Democrats vs GOP by Congress

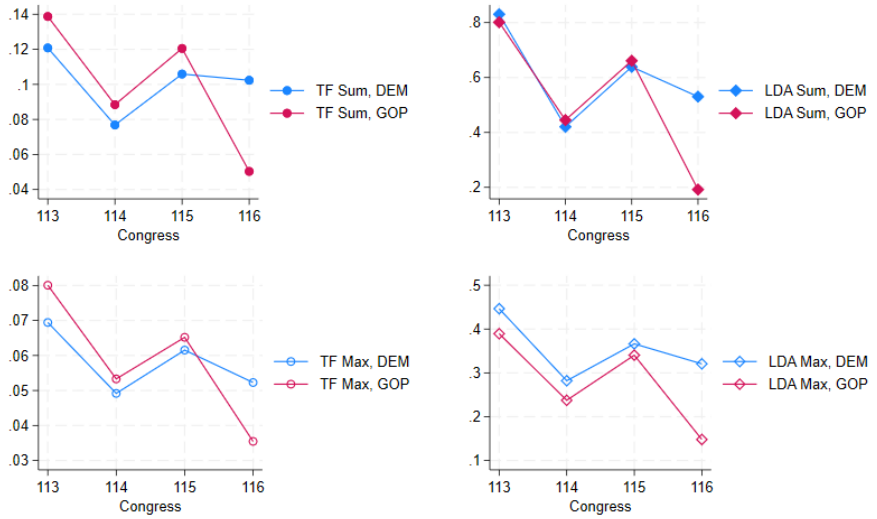



Figure 7: PhRMA Donations, Democrats vs GOP by Congress

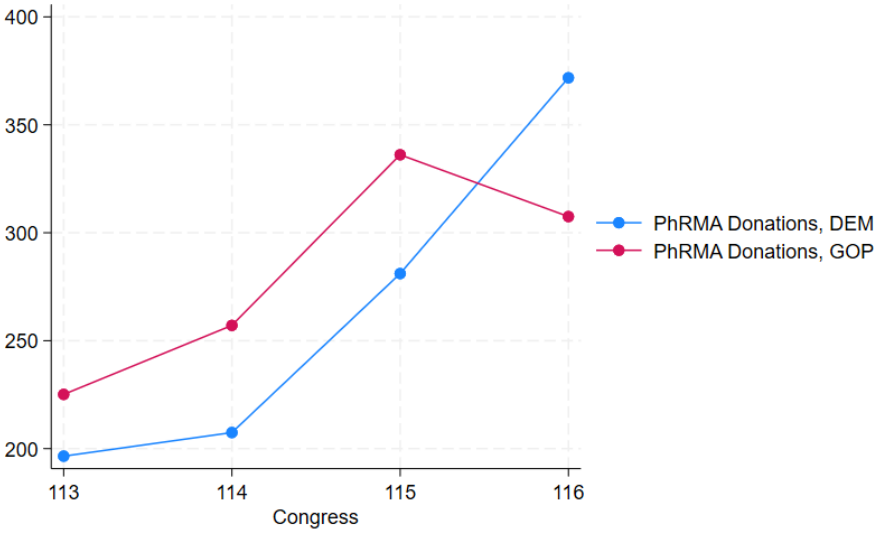


Figure 8: PhRMA Favorabilities, Donations vs. No Donations

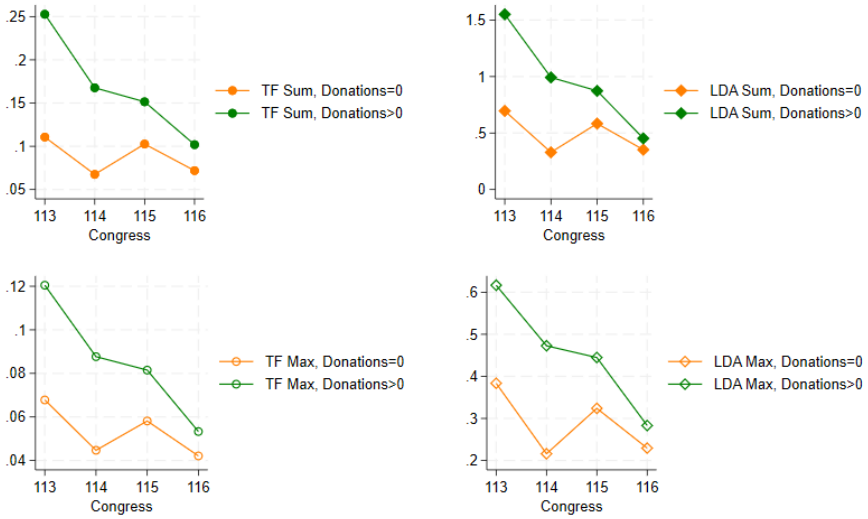
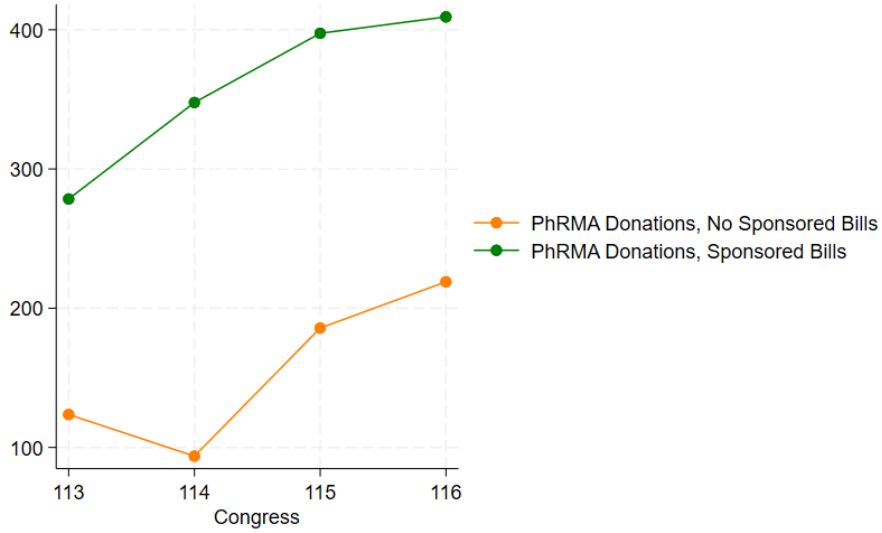



Figure 9: PhRMA Donations, Sponsors verses Non-Sponsors by Congress



# B Tables

Table 1: ALEC and Real Legislation Information

| Congress | Model Bills | Real Bills | Total | Unigram Vocab Size | Bigram Vocab Size |
|---|---|---|---|---|---|
| 113th | 50 | 632 | 682 | 8,317 | 151,609 |
| 114th | 17 | 635 | 652 | 8,244 | 130,341 |
| 115th | 17 | 733 | 750 | 8,532 | 151,602 |
| 116th | 7 | 975 | 982 | 9,806 | 190,293 |

Table 2: John Doe's 112th Congress ALEC Similarity Scores and PhRMA Favorabilities

| Bill | TF | LDA |
|---|---|---|
| 1 | 0.01 | 0.001 |
| 2 | **0.1** | 0.97 |
| 3 | 0.02 | 0.002 |
| 4 | 0.08 | 0.6 |
| 5 | 0.07 | **0.999** |
| Sum | 0.28 | 2.572 |
| | | |
| Sum Favorability | 0.28 | 2.572 |
| Max Favorability | **0.1** | **0.999** |

Table 3: Summary Statistics

| Variable | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| TF (Sum) | 1,732 | 0.375 | 0.571 | 0 | 3.996 |
| LDA (Sum) | 1,732 | 0.436 | 0.809 | 0 | 7.287 |
| TF (Max) | 1,732 | 0.177 | 0.187 | 0 | 0.719 |
| LDA (Max) | 1,732 | 0.271 | 0.385 | 0 | 1 |
| PAC Donations ('000s) | 1,732 | 0.275 | 0.710 | 0 | 7.5 |
| Vote Share(t-1) | 1,732 | 0.697 | 0.149 | 0.5 | 1 |

Table 4: PhRMA Favorability Score Descriptive Statistics

| Variable | N (all) | Mean | SD | Min | Max | N (Bills>0) | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| TF (Sum) | 1,732 | 0.375 | 0.571 | 0 | 3.996 | 1,026 | 0.633 | 0.622 | 0 | 3.996 |
| LDA (Sum) | 1,732 | 0.436 | 0.809 | 0 | 7.287 | 1,026 | 0.736 | 0.94 | 0 | 7.287 |
| TF (Max) | 1,732 | 0.177 | 0.187 | 0 | 0.719 | 1,026 | 0.298 | 0.151 | 0 | 0.719 |
| LDA (Max) | 1,732 | 0.271 | 0.385 | 0 | 1 | 1,026 | 0.457 | 0.406 | 0 | 1 |

Table 5: Language and Donation means, split by political party

| | ALL (N=1732) | GOP (N=816) | DEM (N=916) | Between Groups Mean T-Test T-Statistic | P-value (GOP=DEM) |
|---|---|---|---|---|---|
| TF (Sum) | 0.375 | 0.332 | 0.424 | t=-3.36 | p=0.001 |
| LDA (Sum) | 0.436 | 0.402 | 0.474 | t=-1.84 | p=0.066 |
| TF (Max) | 0.177 | 0.158 | 0.197 | t=-4.32 | p=0.000 |
| LDA (Max) | 0.271 | 0.238 | 0.307 | t=-3.75 | p=0.000 |
| PAC Money ('000s) | 0.275 | 0.281 | 0.269 | t=0.330 | p=0.742 |

Table 6: Language Scores, Split into Positive and Zero donation groups

| FAVOR Measure | Mean, ALL (N=1732) | Mean, $ >0 (N=323) | Mean, $=0 (N=1409) | Between Groups Mean T-Test T-Statistic |
|---|---|---|---|---|
| TF (S) | 0.375 | 0.606 | 0.322 | t=8.203 |
| LDA (S) | 0.436 | 0.693 | 0.377 | t=6.421 |
| TF (M) | 0.177 | 0.235 | 0.163 | t=6.241 |
| LDA (M) | 0.271 | 0.381 | 0.245 | t=5.765 |

Table 7: Donations mean, split by amount of Healthcare Bills Written

| Observations | N | Mean | Std. Dev |
|---|---|---|---|
| All | 1732 | 0.275 | 0.710 |
| Health Bills>0 | 1026 | 0.360 | 0.815 |
| Health Bills=0 | 706 | 0.152 | 0.496 |
| Difference of Means: | t=6.077 | | |

Table 8: State-Party Level OLS Regressions

| VARIABLES | (1) TF Sum | (2) TF Max | (3) LDA Sum | (4) LDA Max |
|---|---|---|---|---|
| Donations (t-1) | 1.60e-05** | 7.51e-06** | 6.39e-05 | 3.74e-05** |
| | (6.90e-06) | (3.07e-06) | (4.01e-05) | (1.83e-05) |
| GOP | -0.0572*** | -0.0154** | -0.332*** | -0.133*** |
| | (0.0152) | (0.00754) | (0.0882) | (0.0437) |
| Vote Share(t-1) | 0.00730 | 0.00366 | 0.0234 | 0.0181 |
| | (0.00525) | (0.00250) | (0.0299) | (0.0134) |
| 114th (2015/16) | -0.0290** | -0.00168 | -0.221*** | -0.0664 |
| | (0.0130) | (0.00710) | (0.0794) | (0.0412) |
| 115th (2017/18) | 0.0260* | 0.0182** | 0.215** | 0.0774* |
| | (0.0149) | (0.00717) | (0.0945) | (0.0422) |
| GOP x 114th | 0.0731*** | 0.0189* | 0.473*** | 0.145** |
| | (0.0201) | (0.0103) | (0.116) | (0.0564) |
| GOP x 115th | 0.0736*** | 0.0215** | 0.369*** | 0.140** |
| | (0.0213) | (0.0108) | (0.128) | (0.0571) |
| Constant | -9.07e-05 | -0.00311 | 0.0292 | 0.0493 |
| | (0.0209) | (0.0139) | (0.125) | (0.0788) |
| Observations | 1,052 | 1,052 | 1,052 | 1,052 |
| R-squared | 0.336 | 0.247 | 0.256 | 0.173 |
| **Fixed Effects** | | | | |
| State | ✓ | ✓ | ✓ | ✓ |
| State-Party | ✓ | ✓ | ✓ | ✓ |

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the state-party-Congress level are in parentheses. This table shows the results of the regressions shown in Equation B from Section 5. Observations are House of Representatives members over a four-Congress unbalanced panel, starting with the 113th Congress (2013/14) and ending with the 116th Congress (2019/20). House member observations are included in a Congress cross-section if they wrote at least one piece of legislation in any subject area during the Congress and if they served the entire two years with the same political party. The dependent variable of interest in each regression, is one of four measures of a Representative's legislation favorability to PhRMA in a Congress. See section 4.2 for a thorough explanation of the creation of these variables. Column (1) shows term frequency sum favorability results, column (2) shows the term frequency maximum favorability results, column (3) shows the LDA sum favorability results, and column (4) shows LDA maximum favorability results. The independent variable in each regression, Donations (t-1), is the amount that a sitting House of Representatives member received from the Pharmaceutical and Research and Manufacturers of America (PhRMA) special interest group Political Action Committee during the previous Congress. Vote Share (t-1) is the share of the major party vote that a legislator received in their last election. GOP is a party indicator variable that takes value of 1 if the legislator is a Republican, and 0 otherwise. 114th (2015/16), and 115th (2017/18) are Congress fixed effects, with 116th (2019/20) being omitted). State fixed effects and interaction terms between the GOP indicator and state fixed effects are utilized in all regression, with results being omitted from the table above.

Table 9: State-Party Level OLS Regressions, Standard Deviation Reporting

| VARIABLES | (1) TF Sum | (2) TF Max. | (3) LDA Sum | (4) LDA Max |
|---|---|---|---|---|
| Donations (t-1) | 0.0659** | 0.0623** | 0.0461 | 0.0662** |
| | (0.0285) | (0.0254) | (0.0289) | (0.0324) |
| GOP | -0.334*** | -0.181** | -0.338*** | -0.332*** |
| | (0.0887) | (0.0882) | (0.0898) | (0.109) |
| Vote Share(t-1) | 0.0426 | 0.0429 | 0.0238 | 0.0452 |
| | (0.0306) | (0.0293) | (0.0304) | (0.0334) |
| 114th (2015/16) | -0.169** | -0.0197 | -0.225*** | -0.166 |
| | (0.0759) | (0.0830) | (0.0808) | (0.103) |
| 115th (2017/18) | 0.152* | 0.213** | 0.219** | 0.193* |
| | (0.0871) | (0.0839) | (0.0961) | (0.105) |
| GOP x 114th | 0.427*** | 0.221* | 0.482*** | 0.363** |
| | (0.117) | (0.121) | (0.118) | (0.141) |
| GOP x 115th | 0.430*** | 0.251** | 0.375*** | 0.350** |
| | (0.125) | (0.126) | (0.130) | (0.143) |
| Constant | -0.300** | -0.479*** | -0.321** | -0.426** |
| | (0.122) | (0.161) | (0.126) | (0.198) |
| Observations | 1,052 | 1,052 | 1,052 | 1,052 |
| R-squared | 0.336 | 0.247 | 0.256 | 0.173 |
| **Fixed Effects** | | | | |
| State | ✓ | ✓ | ✓ | ✓ |
| State-Party | ✓ | ✓ | ✓ | ✓ |

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the state-party-Congress level are in parentheses. This table shows the results of the regressions shown in Equation B from Section 5. All four PhRMA favorability dependent variables, Donations (t-1) and Vote Share (t-1) were transformed into their standard deviations above the panel-wide means. Observations are House of Representatives members over a four-Congress unbalanced panel, starting with the 113th Congress (2013/14) and ending with the 116th Congress (2019/20). House member observations are included in a Congress cross-section if they wrote at least one piece of legislation in any subject area during the Congress and if they served the entire two years with the same political party. The dependent variable of interest in each regression, is one of four measures of a Representative's legislation favorability to PhRMA in a Congress. See section 4.2 for a thorough explanation of the creation of these variables. Column (1) shows term frequency sum favorability results, column (2) shows the term frequency maximum favorability results, column (3) shows the LDA sum favorability results, and column (4) shows LDA maximum favorability results. The independent variable in each regression, Donations (t-1), is the amount that a sitting House of Representatives member received from the Pharmaceutical and Research and Manufacturers of America (PhRMA) special interest group Political Action Committee during the previous Congress. Vote Share (t-1) is the share of the major party vote that a legislator received in their last election. GOP is a party indicator variable that takes value of 1 if the legislator is a Republican, and 0 otherwise. 114th (2015/16), and 115th (2017/18) are Congress fixed effects, with 116th (2019/20) being omitted). State fixed effects and interaction terms between the GOP indicator and state fixed effects are utilized in all regression, with results being omitted from the table above.

Table 10: State-Party Level OLS Regressions

| VARIABLES | (1) Donations | (2) Donations | (3) Donations | (4) Donations |
|---|---|---|---|---|
| Favorability (t-1) | 120.4 | 202.5 | 36.66 | 82.68 |
| | (155.5) | (259.6) | (29.77) | (61.32) |
| GOP | -230.2** | -228.8** | -232.6** | -226.9** |
| | (109.4) | (109.6) | (109.5) | (109.7) |
| Vote Share(t-1) | -17.03 | -16.64 | -17.89 | -15.95 |
| | (22.26) | (22.23) | (22.33) | (22.33) |
| 114th (2015/16) | -227.8*** | -227.0*** | -235.7*** | -233.8*** |
| | (81.01) | (80.92) | (81.86) | (81.22) |
| 115th (2017/18) | -153.7* | -154.5* | -149.9* | -151.3* |
| | (81.74) | (81.61) | (81.54) | (81.52) |
| GOP x 114th | 331.4*** | 328.9*** | 337.2*** | 333.7*** |
| | (114.1) | (114.5) | (114.0) | (114.2) |
| GOP x 115th | 286.3** | 284.5** | 287.5** | 288.0** |
| | (115.3) | (115.6) | (115.3) | (115.2) |
| Constant | 757.4*** | 754.1*** | 757.2*** | 741.5*** |
| | (277.8) | (278.4) | (278.5) | (280.7) |
| Observations | 1,052 | 1,052 | 1,052 | 1,052 |
| R-squared | 0.316 | 0.316 | 0.318 | 0.317 |
| **Language Model** | TF | TF | LDA | LDA |
| **Favorability Type** | Sum | Max | Sum | Max |
| **Fixed Effects** | | | | |
| State | ✓ | ✓ | ✓ | ✓ |
| State-Party | ✓ | ✓ | ✓ | ✓ |

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the state-party-Congress level are in parentheses. This table shows the results of the regressions shown in Equation A from Section 5. Observations are House of Representatives members over a four-Congress unbalanced panel, starting with the 113th Congress (2013/14) and ending with the 116th Congress (2019/20). House member observations are included in a Congress cross-section if they wrote at least one piece of legislation in any subject area during the Congress and if they served the entire two years with the same political party. The dependent variable in each regression, Donations, is the amount that a sitting House of Representatives member received from the Pharmaceutical and Research and Manufacturers of America (PhRMA) special interest group Political Action Committee during a Congress. The independent variable of interest in each regression, Favorability (t-1), is one of four measures of a Representative's legislation favorability to PhRMA in the previous Congress. See section 4.2 for a thorough explanation of the creation of these variables. Column (1) shows term frequency sum favorability results, column (2) shows the term frequency maximum favorability results, column (3) shows the LDA sum favorability results, and column (4) shows LDA maximum favorability results. Vote Share (t-1) is the share of the major party vote that a legislator received in their last election. GOP is a party indicator variable that takes value of 1 if the legislator is a Republican, and 0 otherwise. 114th (2015/16) and 115th (2017/18) are Congress fixed effects, with 116th (2019/20) being omitted). State fixed effects and interaction terms between the GOP indicator and state fixed effects are utilized in all regression, with results being omitted from the table above.

Table 11: State-Party Level OLS Regressions, Standard Deviation Reporting

| VARIABLES | (1) Donations | (2) Donations | (3) Donations | (4) Donations |
|---|---|---|---|---|
| Favorability (t-1) | 0.0291 | 0.0244 | 0.0508 | 0.0467 |
| | (0.0376) | (0.0313) | (0.0413) | (0.0346) |
| GOP | -0.325** | -0.323** | -0.328** | -0.320** |
| | (0.154) | (0.155) | (0.154) | (0.155) |
| Vote Share(t-1) | -0.0240 | -0.0235 | -0.0253 | -0.0225 |
| | (0.0314) | (0.0314) | (0.0315) | (0.0315) |
| 114th (2015/16) | -0.322*** | -0.320*** | -0.333*** | -0.330*** |
| | (0.114) | (0.114) | (0.116) | (0.115) |
| 115th (2017/18) | -0.217* | -0.218* | -0.212* | -0.214* |
| | (0.115) | (0.115) | (0.115) | (0.115) |
| GOP x 114th | 0.468*** | 0.464*** | 0.476*** | 0.471*** |
| | (0.161) | (0.162) | (0.161) | (0.161) |
| GOP x 115th | 0.404** | 0.402** | 0.406** | 0.406** |
| | (0.163) | (0.163) | (0.163) | (0.163) |
| Constant | 0.699* | 0.694* | 0.711* | 0.696* |
| | (0.391) | (0.391) | (0.393) | (0.393) |
| Observations | 1,052 | 1,052 | 1,052 | 1,052 |
| R-squared | 0.316 | 0.316 | 0.318 | 0.317 |
| **Language Model** | TF | TF | LDA | LDA |
| **Favorability Type** | Sum | Max | Sum | Max |
| **Fixed Effects** | | | | |
| State | ✓ | ✓ | ✓ | ✓ |
| State-Party | ✓ | ✓ | ✓ | ✓ |

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered at the state-party-Congress level are in parentheses. This table shows the results of the regressions shown in Equation A from Section 5. Donations, Favorability (t-1), and Vote Share (t-1) were transformed into their standard deviations above the panel-wide means. Observations are House of Representatives members over a three-Congress unbalanced panel, starting with the 114th Congress (2013/14) and ending with the 116th Congress (2019/20), with lagged variables coming from the 113th through 116th Congresses. House member observations are included in a Congress cross-section if they wrote at least one piece of legislation in any subject area during the Congress and if they served the entire two years with the same political party. The dependent variable in each regression, Donations, is the amount that a sitting House of Representatives member received from the Pharmaceutical and Research and Manufacturers of America (PhRMA) special interest group Political Action Committee during a Congress. The independent variable of interest in each regression, Favorability (t-1), is one of four measures of a Representative's legislation favorability to PhRMA in the previous Congress. See section 4.2 for a thorough explanation of the creation of these variables. Column (1) shows term frequency sum favorability results, column (2) shows the term frequency maximum favorability results, column (3) shows the LDA sum favorability results, and column (4) shows LDA maximum favorability results. Vote Share (t-1) is the share of the major party vote that a legislator received in their last election. GOP is a party indicator variable that takes value of 1 if the legislator is a Republican, and 0 otherwise. 114th (2015/16), and 115th (2017/18) are Congress fixed effects, with 116th (2019/20) being omitted). State fixed effects and interaction terms between the GOP indicator and state fixed effects are utilized in all regression, with results being omitted from the table above.

# C Campaign Finance Supreme Court Case History

# D Statistics Foundations

## D.1 Bayesian Methods

Recall that the most basic form of Bayes Theorem is:

$$\mathrm{P}(Y|Z) = \frac{\mathrm{P}(Z|Y)\mathrm{P}(Y)}{\mathrm{P}(Z)} \tag{3}$$

Let $\boldsymbol{\Theta}$ be a vector of parameters whose true values are unknown and $\mathbf{X}$ be a vector of observed values. Given an observed "sampling distribution" of $\mathbf{X}$'s values $\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})$, and assuming an initial "prior distribution" of $\boldsymbol{\Theta}$'s values $\mathrm{P}(\boldsymbol{\Theta})$, Bayes Theorem can be utilized to back out an updated, "posterior distribution" $\mathrm{P}(\boldsymbol{\Theta}|\mathbf{X})$ as seen below:

$$\mathrm{P}(\boldsymbol{\Theta}|\mathbf{X}) = \frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta})}{\mathrm{P}(\mathbf{X})} = \frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta})}{\sum_{\boldsymbol{\Theta}}\left[\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta})\right]} \tag{4}$$

And when the distribution $\mathrm{P}(\mathbf{X})$ is continuous

$$= \frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta})}{\int_{\boldsymbol{\Theta}}\left[\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta})d\boldsymbol{\Theta}\right]}$$

$\mathrm{P}(\boldsymbol{\Theta}|\mathbf{X})$ is known as the posterior distribution, and If Bayes rule is being applied to a partial distribution where parameters $\boldsymbol{\alpha}$,

$$\mathrm{P}(\boldsymbol{\Theta}|\mathbf{X},\boldsymbol{\alpha}) = \frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta},\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{\Theta}|\boldsymbol{\alpha})}{\mathrm{P}(\mathbf{X}|\boldsymbol{\alpha})}\frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta},\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{\Theta}|\boldsymbol{\alpha})}{\sum_{\boldsymbol{\Theta}}\left[\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta},\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{\Theta}|\boldsymbol{\alpha})\right]} \tag{5}$$

And when the distribution $\mathrm{P}(\mathbf{X}|\boldsymbol{\alpha})$ is continuous

$$= \frac{\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta},\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{\Theta}|\boldsymbol{\alpha})}{\int_{\boldsymbol{\Theta}}\left[\mathrm{P}(\mathbf{X}|\boldsymbol{\Theta},|\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{\Theta}|\boldsymbol{\alpha})d\boldsymbol{\Theta}\right]}$$

The functional form of the posterior distribution will not be related to that of the prior and sampling distributions, unless the sampling distribution $D$ has a "conjugate prior" distribution $C$. If the sampling distribution $S(\mathbf{X}|\mathbf{\Theta})$ has the conjugate prior $C(\mathbf{\Theta})$ than the posterior distribution will be of the form $C(\mathbf{\Theta}+\boldsymbol{\epsilon})$.

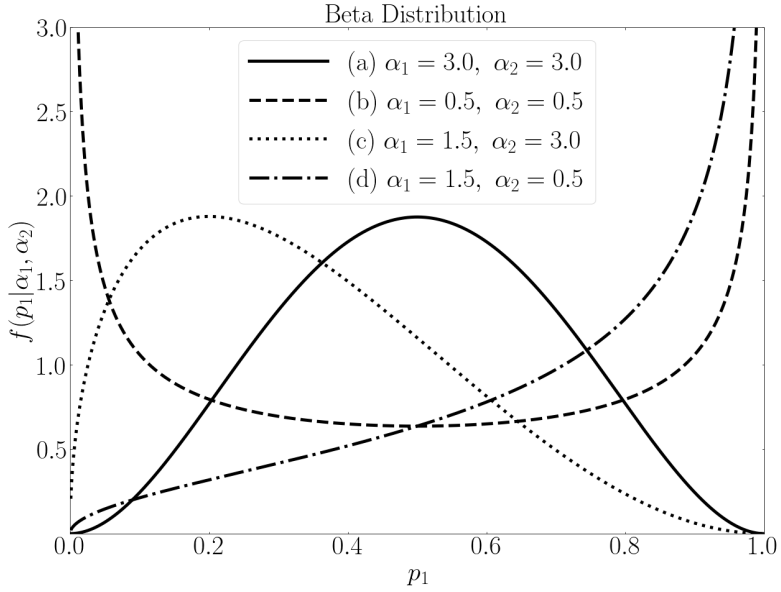
## D.2 Categorical Distribution

The categorical distribution Cat(•) is used to model a random process involving a single event with multiple outcomes. Let $Y = \{y_1, \ldots, y_n, \ldots, y_N\}$, with $y_n$ being the probability that the specific outcome $n$ occurs. For example, if a die has three sides, with the first being twice as likely as the other two, it would be modeled with the distribution $\mathbf{Cat}(1/2, 1/4, 1/4)$. The categorical distribution has analogs to more commonly used distributions in statistics. It is an extension of the Bernoulli distribution that allows for multiple outcomes instead of just one, and a restriction of the multinomial distribution that allows only for a single trial. In other words, the categorical distribution models a single roll of a weighted die, while the Bernoulli distribution models a single flip of a weighted coin and the multinomial distribution models any number rolls of a weighted die.

## D.3 Beta Distribution

The beta distribution Beta(•), models observed probabilities of a binary event with a hidden, latent, true probability. A beta distribution that models the probability of success $p_1$, with $p_2 = 1 - p_1$ being the probability of failure, takes the parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$. The higher that $\alpha_1$ is in comparison to $\alpha_2$, the larger the prior certainty of success and the lower that $\alpha_1$ is in comparison to $\alpha_2$, the higher the prior certainty of failure. The higher that $\alpha_1 + \alpha_2$ is, the more overall uncertainty there is about the events in general (higher mass towards the center of the distribution). Figure 10 shows several beta distributions with different parameters. Consider a beta distribution that models the win-loss probability of a sports team over a particular season. From Figure 10, (a) would designate a historically consistent, average team, (b) a historically streaky team, (c) a historically bad team (d) a historically good team.

Figure 10: Beta distributions with various parameters




The beta distribution's PDF of Beta$(\alpha_1, \alpha_2)$

$$= \frac{1}{B(\alpha_1, \alpha_2)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} = \frac{1}{B(\alpha_1, \alpha_2)} \prod_{k=1}^{2} p_k^{\alpha_k - 1} \tag{6}$$

Where:

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} = \frac{\prod_{k=1}^{2} \Gamma_{k=1}^{2}(\alpha_k)}{\Gamma(\sum_{k=1}^{2} \alpha_k)} \tag{7}$$

[14] Let

$$\tilde{B}(x, \alpha_1, \alpha_2) = \int_0^x t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1} dt$$

The beta distribution's CDF is

$$P(p_1 > x | \alpha_1, \alpha_2) = P(p_2 < x | \alpha_1, \alpha_2) \frac{\tilde{B}(x, \alpha_1, \alpha_2)}{B(\alpha_1, \alpha_2)} \tag{8}$$

[14] Recall that the gamma function is

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

and when $x$ is an integer, $\Gamma(x) = (x-1)!$

Finally, for $p \sim \text{Beta}(\alpha_1, \alpha_2)$, the expected value of $p$ is

$$\mathbf{E}[p] = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{9}$$

The beta distribution is oftentimes used in conjunction with the Bernoulli and binomial distributions. Not only does it make sense that the probability drawn from the beta distribution would follow a Bernoulli or binomial process, but the beta distribution is the conjugate prior distribution of both the Bernoulli and binomial distributions.[15]

## D.4 Dirichlet Distribution

The Dirichlet distribution is the extension of the beta distribution to probabilistic events not limited to two possible outcomes. Like the beta distribution that can be used in conjunction with the Bernoulli and binomial distributions, the Dirichlet distribution can be used in conjunction with the categorical and multinomial distributions. A Dirichlet distribution modeling events $\mathbf{x} = [x_1, ...x_k, ...x_K]$ with the probabilities $\mathbf{p} = [p_1, ...p_k, ...p_K]$ takes the parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K]$. Higher values of $\alpha_k$ increase the observed probability of $x_k$. Holding the relative differences between the parameters constant, the higher $||\boldsymbol{\alpha}||$, the more centered that the distribution is, and thus the more that it favors uncertainty of any particular event occuring. Figure 11 shows several examples of three-dimensional Dirichlet distributions in the two-dimensional simplex.

[15] See Appendix D.1 for the explanation of conjugate prior distributions

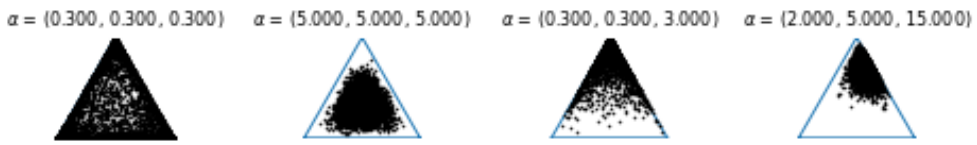



Figure 11: Several three-dimensional Dirichlet distributions.

The Dirichlet distribution's PDF is

$$\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{10}$$

Where $\sum_{k=1}^{K}(p_k) = 1$ and

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma_{k=1}^{K}(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)} \tag{11}$$

The beta distribution is simply a special case of the Dirichlet where $S = 2$. And much how the Beta distribution is used in conjunction with the Bernoulli and binomial distributions, the Dirichlet distribution is often used with the categorical and multinomial distribution, with the Dirichlet distribution being the conjugate prior distribution of both.[16]

[16]See Appendix D.1 for the explanation of conjugate prior distributions

# E Latent Dirichlet Allocation (LDA) Model

As mentioned in the main paper, the LDA model treats documents as categorical distributions of topics, and these topics are treated as categorical distributions of words. The vector of probabilities used as the parameters for the categorical distribution of topics are assumed to be an observation from a Dirichlet distribution[17] of topic probabilities. Likewise, the vector of probabilities used as the parameters for the categorical distribution of words is assumed to be an observation from a Dirichlet distribution of word probabilities. The lengths of documents in an LDA corpus are either assumed to be distributed Poisson, or as I will do, will be taken as given. LDA assumes that documents are generated by the following random process. Prior to generating the documents, a vector of topic probabilities is drawn from each document's Dirichlet distribution of topics, making up the parameters for each document's categorical distribution of topics. Then, a vector of word probabilities is drawn from each topic's Dirichlet distribution of word probabilities, making up the parameters for each document's categorical distribution of topics. The individual documents are then generated word by word. Each word of each document is selected occurs first by drawing a topic from that document's categorical distribution of topics and then a word is drawn from that topic's categorical distribution of words.

Consider a corpus with $P$ documents $\mathbf{D} = [d_1, \ldots, d_p, \ldots, d_P]$ with a vocabulary $\mathbf{V} = \{v_1, ...v_e, ...v_E\}$, that contains the $S$ topics $\boldsymbol{A} = [a_1, \ldots, a_s, \ldots, a_S]$, whose document lengths are $\mathbf{W} = [W_1, \ldots, W_p, \ldots, W_P]$, and $W$ be the maximum number of words in any document in $\mathbf{D}$.

- Let $\Sigma_p = [\sigma_{p,1}, \ldots, \sigma_{p,s}, \ldots, \sigma_{p,S}]$ be the parameters for document $d_p$'s Dirichlet distribution of topics and $\mathbf{\Sigma}$ be the $P \times S$ matrix of these parameters across all documents.
- Let $\Theta_p = [\theta_{p,1}, \ldots, \theta_{p,s}, \ldots, \theta_{p,s}]$ be the drawn probabilities from document $d_p$'s Dirichlet distribution of topics with $\mathbf{\Theta}$ being the $P \times S$ matrix of these probabilities across all documents.
- Let $\Phi_s = [\phi_{s,1}, \ldots, \phi_{s,e}, \ldots, \phi_{s,E}]$ be the parameters for topic $a_s$'s Dirichlet distribution of words and $\mathbf{\Phi}$ be the $S \times E$ matrix of these parameters across all topics.

[17] for information about the Dirichlet distribution, refer to D.4

- Let $\Omega_s = [\omega_{s,1}, \ldots, \omega_{s,e}, \ldots, \omega_{s,E}]$ be the the drawn probabilities from topic $a_s$'s Dirichlet distribution of words and $\mathbf{\Omega}$ be the $S \times E$ matrix of parameters across all topics.

- Let $N_p = [n_{p,1}, \ldots, n_{p,w}, ...n_{p,W_p}]$ be the words found in the document $d_p$, and $\mathbf{N}$ be the $P \times W$ matrix of the words of all documents in a corpus. For $d_p$ where $W_P < W$, row $p$ of $\mathbf{N}$ has indicators null indicators with certainty for the final $W - W_p$ columns to maintain consistent row lengths.

- Let $T_p = [t_{p,1}, \ldots, t_{p,w}, ...t_{p,W_p}]$ be a vector of topics such that $t_{p,w}$ is the topic that word $n_{p,w}$ was drawn from and $\mathbf{T}$ be the $P \times W$ matrix of these topics across all documents. Again, for $d_p$ where $W_P < W$, row $p$ of $\mathbf{T}$ has indicators null indicators with certainty for the final $W - W_p$ columns. [18]

The document generation process can be described as folows:

1. For each document $d_p$, draw the parameters $\Theta_p$ from the Dirichlet distribution $\text{Dir}(\Sigma_p)$.

2. For each topic $a_s$, draw the parameters $\Omega_s$ from the Dirichlet distribution $\text{Dir}(\Phi_s)$

3. For each eventual word in each document $d_p$, pick that word's topic $t_{p,w}$ from the categorical distribution $\text{Cat}(\Theta_p)$

4. Pick the word $n_{p,w}$ from the categorical distribution $\text{Cat}(\Omega_{t_{p,w}})$, where $\Omega_{t_{p,w}}$ is topic $t_{p,w}$'s categorical distribution parameters.

With the distributions being described as follows:

1. $\Theta_p \sim \text{Dir}(\Sigma_p)$, $\forall p$

2. $\Omega_s \sim \text{Dir}(\Phi_s)$, $\forall s$

[18] ADVISOR'S NOTE: This null indicator thing is simply to be able to throw all documents into a matrix regardless of how long it is. If there is a better way of saying "$\mathbf{N}$ is all of the document's words in a matrix and $\mathbf{T}$ all of the latent topics in a matrix, and you can ignore the mismatch dimensions because it doesn't cause any problems in what follows", please let me know. Explanations I've seen either ignore the row length mismatch or assume all documents are the same length.

3. $T_p = [t_{p,1}, \ldots, t_{p,w}, \ldots, t_{p,W_p}]$, where $t_{p,w} \sim \text{Cat}(\Theta_p)$, $\forall p$

4. $N_p = [n_{p,1}, \ldots, n_{p,w}, ...n_{p,W_p}]$, where $n_{p,w} \sim \text{Cat}(\Omega_{t_{p,w}})$, $\forall p$

The probability density function of realizing values of $\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}, \mathbf{N}$ is shown in Equation 12 below: [19]

$$P\Big(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}, \mathbf{N} | \mathbf{\Sigma}, \mathbf{\Phi}\Big) = \underbrace{P\Big(\mathbf{\Theta}|\mathbf{Dir}(\mathbf{\Sigma})\Big)}_{1} \underbrace{P\Big(\mathbf{\Omega}|\mathbf{Dir}(\mathbf{\Phi})\Big)}_{2} \underbrace{P\Big(\mathbf{T}|\mathbf{Cat}(\mathbf{\Theta})\Big)}_{3} \underbrace{P\Big(\mathbf{N}|\mathbf{Cat}(\mathbf{\Omega_T})\Big)}_{4} =$$
$$\prod_{p=1}^{P} \underbrace{P\Big(\Theta_p|\text{Dir}(\Sigma_p)\Big)}_{1} \prod_{s=1}^{S} \underbrace{P\Big(\Omega_s|\text{Dir}(\Phi_s)\Big)}_{2} \prod_{w=1}^{W_p} \underbrace{P\Big(t_{p,w}|\text{Cat}(\Theta_p)\Big)}_{3} \underbrace{P\Big(n_{p,w}|\text{Cat}(\Omega_{t_{p,w}})\Big)}_{4} \tag{12}$$

Where the labeled braces correspond to the steps 1 through 4 from the LDA process.

The process of creating an LDA model is more complex than that of the term frequency and tf-idf models, requiring machine learning algorithms to do so. The LDA training process takes the only known information, the words in the documents $\mathbf{N}$, and uses it to back out $\mathbf{\Sigma}, \mathbf{\Phi}, \mathbf{\Theta}$, and $\mathbf{\Omega}$. The training process takes several inputs, including an assumption on the total number of topics $S$ and initial priors of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ (in practice, all $\sigma_{p,s}$ are assumed to be identical and between 0 and 1, usually less than 0.5, as are all $\phi_{s,e}$). There are a variety of methods used to do this, but all utilize some form of Bayesian methods.

The posterior distribution $P(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}|\mathbf{N}, \mathbf{\Sigma}, \mathbf{\Phi})$ can be represented by Equation 13 below:

$$P(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}|\mathbf{N}, \mathbf{\Sigma}, \mathbf{\Phi}) = \frac{P(\mathbf{N}|\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}, \mathbf{\Sigma}, \mathbf{\Phi}) \tilde{P}(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}|\mathbf{\Sigma}, \mathbf{\Phi})}{P(\mathbf{N})} = \frac{P\Big(\mathbf{\Theta}, \mathbf{\Omega}, \mathbf{T}, \mathbf{N}|\mathbf{\Sigma}, \mathbf{\Phi}\Big)}{P(\mathbf{N})} \tag{13}$$

With the numerator being equal to Equation 12 since parts one through three are statistically independent. Note that because the Dirichlet distribution is the

[19] ADVISOR'S NOTE: The notation I use, for example $P\Big(\mathbf{\Theta}|\mathbf{Dir}(\mathbf{\Sigma})\Big)$, is done to include the specific distributions in the density function statements while avoiding "crimes against probability notation" if you will.

conjugate prior distribution of the categorical distribution, this posterior probability $P\Big(\boldsymbol{\Theta}, \boldsymbol{\Omega}, \mathbf{T}, \mathbf{N} | \boldsymbol{\Sigma}, \boldsymbol{\Phi}\Big) =$ will also be a Dirichlet distribution. The probability density function $P(\mathbf{N})$ in the denominator is obtained by taking Equation 12 and summing across all possible topics $\mathbf{A}$ and all possible values of $\boldsymbol{\Omega}$ and $\boldsymbol{\Theta}$. This density is shown below in Equation 14 below:

$$P(\mathbf{N}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Omega}} P\Big(\boldsymbol{\Theta} | \mathbf{Dir}(\boldsymbol{\Sigma})\Big) P\Big(\boldsymbol{\Omega} | \mathbf{Dir}(\boldsymbol{\Phi})\Big) \sum_{\boldsymbol{\alpha}} P\Big(\mathbf{A} | \mathbf{Cat}(\boldsymbol{\Theta})\Big) P\Big(\mathbf{N} | \mathbf{Cat}(\boldsymbol{\Omega}_{\mathbf{T}})\Big) d\boldsymbol{\Omega} d\boldsymbol{\Theta} =$$

$$\int_{\Theta_p} \int_{\Omega_s} \prod_{p=1}^{P} P\Big(\Theta_p | \text{Dir}(\Sigma_p)\Big) \prod_{s=1}^{S} P\Big(\Omega_s | \text{Dir}(\Phi_s)\Big) \prod_{w=1}^{W_p} \sum_{t_{p,w}=a_1}^{a_S} P\Big(t_{p,w} | \text{Cat}(\Theta_p)\Big) P\Big(n_{p,w} | \text{Cat}(\Omega_{t_{p,w}})\Big) d\Omega_s d\Theta_p \tag{14}$$

The integral in the denominator is not directly computable, and thus numerical methods are utilized. The most common one being variational inference.[20]

# F Computational Details

[20] ADVISOR'S NOTE: I'll likely go through this in more detail in a later draft, but I'm punting for now.