# LightBULB: Report

**Christopher Kang**
Computer Science and Engineering
University of Washington
Seattle, USA
ck32@cs.washington.edu

**Jacob Peplinski**
Electrical and Computer Engineering
University of Washington
Seattle, USA
jpeplins@uw.edu

## Abstract

Today's vitriolic political landscape finds bipartisanship in loathing lobbyists - many citizens regard lobbyists as influence peddlers who pervert democracy in favor of elites. While public watchdogs study the behavior of lobbyists, most work focuses on aggregating data and observing broad trends in lobbyist activity. We propose and demonstrate a new set of tools which can track individual companies while integrating textual data at scale. First, we leverage traditional big data techniques to demonstrate underlying structure in congressional lobbying. Then, we introduce a model which predicts a company's interest in specific bills given the bill text and the company's historical lobbying activity, and show that it generalizes well to unseen bills. Lastly, we showcase a series of use cases which leverage these methods to discover interesting lobbying behaviors. Together, these methods serve as an initial foray for automated tools which empower researchers to identify nuanced relationships between special interests and the legislative process.

## 1 Introduction

While lobbying has always been a protected act under the 1st Amendment, it's significance in American politics and policy has been contentious since the dawn of the Republic. Only recently has the process become more transparent with the Lobbying Disclosure Act of 1995, which required public disclosures by lobbyists and the organizations which employ them. However, the sheer complexity of congressional bills and volume of lobbying activity poses a challenge for ordinary citizens and academics alike interested in understanding lobbying dynamics.

Public watchdog groups like the Center for Responsive Politics and the Public Accountability Initiative already aggregate and report on public lobbying disclosures. These datasets are frequently used to produce easily interpretable trends and first-order statistics (e.g. Amazon and Facebook spent the most money on lobbying of all companies in 2020). While useful for communicating simple trends, these analyses lack the complexity to uncover patterns in lobbying behavior within/across organizations. Moreover, current analysis of lobbying data does not consider the text of a bill itself, which we hypothesize is a core source of information for uncovering insights about lobbying behavior.

In this work, we propose a set of complementary methods for deriving insight from publicly-disclosed Congressional lobbying data. Our analyses span topics such as dimensionality reduction, itemset mining, content-based filtering, and community detection, with the ultimate goal of characterizing the behaviors of special interest groups (SIGs) and identifying anomalous behavior. These methods are based on analysis of the *SIG-Bill Matrix*, a representation of lobbying activity we propose that can be derived from publicly-available data. We report results and discuss the validity of each proposed method, and apply our methods in a case study of a specific SIG. Formally, our contributions include:

1. Introducing the *SIG-Bill Matrix* and describing the data and methods used to build it.

Preprint. Under review.

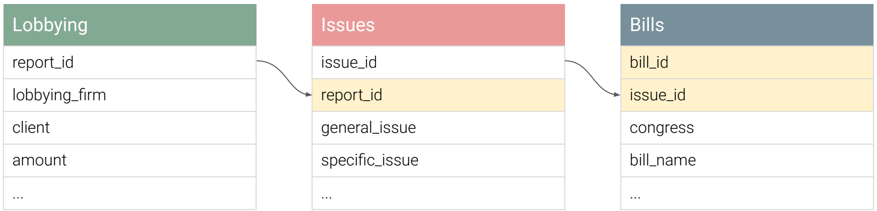

Figure 1: Relational structure of the OpenSecrets lobbying data. Highlighted columns indicate identifiers that are not unique on a table.

2. Applying dimensionality reduction and association rule learning techniques to uncover structure in the *SIG-Bill Matrix*.
3. Showing that the *SIG-Bill Matrix* can be well-approximated using a classifier trained on purpose-built text embeddings of bill titles and that it generalizes to unseen data.
4. Applying a subset of the above methods to uncover non-intuiitive lobbying activity.

## 2 Related Work

### 2.1 Analysis of Lobbying Activity

While political watchdog groups often utilize public lobbying data, work on lobbying behavior in technical research communities is less frequent. Slobozhan et al. [2020] used features created from the text of U.S. congressional bills and machine learning models to classify whether a bill had been lobbied. By creating dataset partitions by lobbying intensity, the authors demonstrated that bills that had been heavily lobbied were easier to differentiate from un-lobbied bills, suggesting that lobbying may effect legislation in a way that can be detected by textual analysis. Meng and Rode [2019] performed a game-theoretic statistical analysis to estimate the social impact of lobbying critical climate legislation in 2010.

### 2.2 Natural Language Processing for Legislative Text

Advances in natural language processing (NLP) have spurred interest in using textual analysis to understand legal texts such as legislation. For example, Yano et al. [2012] used unigram and bigram features derived from the text of bills to increase baseline accuracy of predicting whether a bill will be passed by a congressional committee. Kornilova and Eidelman [2019] explores the use of deep language models for automatically generating summaries of U.S. legislation. Chalkidis et al. [2020] proposes a variant of BERT, a popular deep language model, that is fine-tuned to better model legalese. They demonstrate that their improved model, LegalBERT, outperforms baseline models on multiple domain-specific language modelling and named entity recognition tasks. LegalBERTis trained using the same procedure as BERT, known as masked language modelling, but relies solely on text from EU, UK, and US legislation and court cases.

## 3 Data Description and Collection

### 3.1 OpenSecrets Lobbying Activity Database

OpenSecrets is a public website that aggregates federally mandated lobbying disclosures (called LDAs) into easily searchable formats, enabling large scale analysis of organizations, lobbyists, and specific Congressional bills. These LDAs are required by federal sunshine laws and disclose attributes such as the SIG, lobbying firm, and amount paid for lobbying services (rounded to the nearest $10k).[1]

At the time of writing, the OpenSecrets database contained entries for 1,213,772 LDAs. In our work we focus on three tables in this database: *lobbying*, *issues*, and *bills* (see 1). Each row in the *lobbying*

[1] We discuss caveats associated with this data in section 6.2.1

table represents a unique LDA and each entry in *issues* corresponds to a specific issue listed in the LDA. An excerpt from the *issues* table is shown below (note the identical `report_id` entries).

```
+--------+-----------+-------------+--------------------------------------------+
|issue_id|report_id  |general_issue|specific_issue                              |
+--------+-----------+-------------+--------------------------------------------+
|2433238 |2C518207...|ENG          |Smart grid and solar energy and security issues|
|2433239 |2C518207...|TRA          |Autonomous vehicles, sensors and security      |
+--------+-----------+-------------+--------------------------------------------+
```

Though not required by law, filers will often list the specific bills that were lobbied alongside issues on an LDA. In this case, OpenSecrets stores a link between each issue and referenced bill in the *bills* table. Naturally, several issues (from separate reports) could reference the same bill, or worse, some LDAs may not explicitly link bills and issues, in which case a single issue may map to multiple bills.

### 3.2 Scraping Legislative Text from GovTrack

GovInfo is a website that tracks congressional bills through the legislative process and is run by the US Government Publishing Office (GPO). They provide a bulk dataset which classifies bills by Congress, then Session, and finally bill type (Senate / House resolution, continuing resolution, etc.) Once a list of relevant bills from the OpenSecrets dataset were specified, the GovInfo bulk dataset was queried and the original XML text of the bill at first introduction was downloaded[2]. This XML contains information like the bill's title, formatted bill text (including section/subsection headers), and other metadata. Because the bulk dataset is limited to the 113th - 116th Congress, we also constrain our analysis over this time period. In total, we scraped **22,546** bill XML files from Govinfo.

### 3.3 Dataset Construction

We are interested in the lobbying behavior of SIGs on individual bills. To build our dataset, we inner join the *lobbying*, *issues*, and *bills* tables on the `report_id` and `issue_id` columns respectively, collecting all unique pairs of `report_id` and `bill_id`. Supposing we have a set of SIGs $\mathcal{C}$ and set of bills $\mathcal{B}$, we create a matrix $\mathcal{Y} \in \{0,1\}^{|\mathcal{C}|\times|\mathcal{B}|}$, then assign $\mathcal{Y}_{ij} = 1$ if SIG $i$ lobbied bill $j$ and $\mathcal{Y}_{ij} = 0$ otherwise (example below):

```
         B0 B1 B2 ... BN
SIG_0 [0  1       ... 1 ]
SIG_1 [1  0       ... 0 ]
...   [           ...   ]
SIG_C [1  0       ... 1 ]
```

We call this representation the *SIG-Bill Matrix* - it indicates which SIGs have lobbied which bills as described by the LDAs. In total, **13,507** unique SIGs were found in the OpenSecrets data (looking only at bills for which we could extract XML files). In total $\mathcal{Y}$ is an $[13,507 \times 22,546]$ matrix with a sparsity of 99.945% (or 165,785 non-zero entries). **All of the methods proposed in this work revolve around this matrix and the XML files for each bill.**

## 4 Methodology

We begin by demonstrating the potential utility of $\mathcal{Y}$ as proposed in 3.3 by performing a Principle Components Analysis (PCA) on $\mathcal{Y}$ and visualizing the latent dimensions. Here, we qualitatively search for patterns that correspond to preconceived notions about bills (e.g., SIGs who lobby similar bills should be relatively close together) and discuss possible sources of error.

Next, we search for patterns across SIGs by mining frequent sets of bills and learning association rules. Our objective is to determine the utility of these learned associations between bills and make novel adjustments to more easily find "interesting" rules.

Lastly, we extend our analysis by attempting to approximate a sub-sampled version of $\mathcal{Y}$ using text embeddings extracted from bill titles. We motivate this pursuit by first visualizing a t-SNE plot of

[2] We discuss challenges with this approach in section 6.2.3.

LegalBERTembeddings extracted from bill titles, uncovering structure. We then train a multi-label classifier to approximate rows of $\mathcal{Y}$ using LegalBERTembeddings and show that this classifier can generalize to unseen bills.

### 4.1 Visualizing Latent Structure via the Singular Value Decomposition

In this section, we attempt to identify latent structure within the actual SIG-bill matrix. We are interested in identifying key, approximate properties of $\mathcal{Y}$ through dimensionality reduction. Critically, identifying principal components to express companies (rows of $\mathcal{Y}$) enables visual clustering of companies based on their lobbying behavior; conversely, identifying principal components for columns of $\mathcal{Y}$ enables visual clustering of bills by which companies have lobbied them. We use SVD:

**Theorem 1** (Singular Value Decomposition)**.** *SVD produces some $U, V^T$ orthogonal and $\Sigma$ where:*

$$\mathcal{Y} \approx U \Sigma V^T \tag{1}$$

This decomposition is valuable because $U, V$ serve as a concept space mapping for the individual companies / bills, respectively. (Recall that rows of $U$ represent a SIG's concepts, $\Sigma$ is the strength of a concept, and rows of $V$ map concepts to bills.) After obtaining $U, \Sigma, V$, we plot each row of $U, V$ and observe if there is any notable structure. If there is, this suggests that lobbying behavior has some implicit structure that can be leveraged.

### 4.2 Learning Association Rules from Bill Sets

$\mathcal{Y}$ can be easily thought of as a user-item matrix, as is seen in *Market Basket Analysis*. We begin our exploration of lobbying behavior by attempting to learn association rules from **sets of bills** that clients have lobbied, and investigating rules that defy intuition. In particular, we use the popular *FP-growth* algorithm to mine frequent itemsets and association rules (Han et al. [2000]). FP growth builds frequent itemsets by recognizing that subsets of frequent itemsets must also be frequent, allowing the algorithm to first parse through small sets and incrementally merge them.

One challenge with association rule mining occurs when members submit the same bill over multiple sessions and when identical bills are introduced in both houses. Because our data contains bills from both the House and Senate over multiple congressional sessions, we occasionally capture these "duplicate" bills. We became aware of these patterns after preliminary analysis revealed association rules that frequently linked different versions of the same bill. To mitigate this, we also performed association rule mining on *within-chamber*, *within-session*, *within-chamber-and-session* subsets of the data. Support and Confidence threshold hyperparameters for each grouped analysis are recorded in appendix A.1.

#### 4.2.1 Filtering for Diverse Association Rules

Preliminary attempts at learning association rules reveal bill relationships that are less than surprising. Specifically, we notice numerous association rules where the antecedent and consequent solely comprise of bills with a very similar objective (e.g., defense appropriation). In an attempt to find association rules that defy intuition (and thus unveil potentially useful insight), we integrate categorical data to identify "diverse" rules. Namely, we label an association rule $R$ as diverse if its antecedent set $A$ and consequent set $C$ contain bills that have different *general issue categories* (reference appendix A.2 for further details on creating the general issue categories).

For each association rule, we extract the general issue category for each bill in the antecedent and consequent, and label it as *diverse* if the intersection of their category sets is the null set[3]:

$$d(A, C) = \begin{cases} 1 & |A \cap C| = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

[3] While we elect to use an ensemble based approach, a similar approach could be used via cosine similarity among a normalized vector of categories to classify the diversity of association rules.

### 4.3 Approximating the SIG-Bill Matrix with Bill Titles

#### 4.3.1 Identifying Structure in Bill Title Embeddings

As mentioned in section 2.2, LegalBERT is a variant of BERT that has been fine-tuned on a variety of legal corpora. Although the bill text in our dataset is similar in content to LegalBERT's training corpora, LegalBERT has not been exposed to any US federal legislation, and it is unclear whether the semantics of our bill text can be captured using features from LegalBERT. To motivate our use of LegalBERT, we extract LegalBERT's final hidden layer activations (*embeddings* hereafter) for each token of each bill title, mean-aggregate embeddings across tokens to create a single 768-dimensional vector for each bill title, and perform t-SNE on the set of extracted title embeddings. t-SNE is a dimensionality reduction technique that iteratively minimizes the Kullback–Leibler Divergence between a data distribution and a lower-dimensional representation of the same distribution, and has shown to be a useful tool for visualizing high-dimensional data Van der Maaten and Hinton [2008]. Here, we map LegalBERT title embeddings to $\mathbb{R}^2$, and color points (bill titles) by the bill categories computed in section 4.2.1. If bills of the same category are well clustered in $\mathbb{R}^2$, it is likely that LegalBERT produces embeddings that capture meaningful information about bill titles.

#### 4.3.2 Classifying Lobbying Activity Using Textual Features

Because bill titles describe a bill's purpose, we hypothesize that a SIG's interests can be learned from text features extracted from bill titles. To test this hypothesis, we propose a learning task where the objective is to predict which SIGs lobbied which bills using mean-aggregated LegalBERTembedings over bill titles as input features. Formally, we have a feature matrix $\mathcal{X} \in \mathbb{R}^{n\times 768}$ containing LegalBERT embeddings of bill titles and we want to learn to predict the SIG-Bill Matrix $\mathcal{Y}$. Let $\mathcal{F} : \mathbb{R}^{768} \rightarrow \mathbb{R}^{|\mathcal{C}|}$ be a model with trainable weights $\theta$ adjusted via gradient descent using loss function $\mathcal{L}$. Our objective is then:

$$\min_{\theta}\{\mathcal{L}(\mathcal{F}(\mathcal{X},\theta),\mathcal{Y})\} \tag{3}$$

Because a bill can be lobbied by more than one SIG, this is a multi-label learning task. Therefore, we use binary cross-entropy (*BCE*) to compute loss on predictions for each SIG (i.e. for each output node of $\mathcal{F}$), then average across nodes to produce a single loss value. The sparsity of $\mathcal{Y}$ makes this a very challenging learning task. To make the problem tractable, we produce a subset of $\mathcal{Y}$ created by removing SIGs that have lobbied for less than 500 distinct bills. This reduces the size of our label vector from 13,507 to 159 dimensions (the number of SIGs that have lobbied at least 500 bills), without reducing the number of training examples.

For $\mathcal{F}$, we use a simple neural network with a single hidden layer containing 3000 nodes with the ReLU activation. Then, we add an output layer with 159 nodes using the sigmoid activation to produce a probability for each SIG. In total, this network has 2,784,159 trainable parameters. We perform back propagation using the Adam optimizer Kingma and Ba [2014], a batch size of 32, and a learning rate of $10^{-4}$. We split our dataset into training, validation, and evaluation sets:

- **Training / Validation** ($n = 19189, n = 1552$): Of bills from the 113-115th Congress, we randomly sampled 7.5% of them for the validation set, with the others used for training.
- **Evaluation** ($n = 1802$): We select bills from the 116th Congress'. This evaluation strategy was chosen because different versions of the same bill can occur in the same congressional session, which could cause information leakage between the training and evaluation sets.

Despite filtering, our abridged $\mathcal{Y}$ is still a sparse matrix (only 1.36% of entries are non-zero). As a result, our model will naturally under-predict (i.e. due to sparsity, always guessing 0 may result in a low loss). To combat this, we add a regularizaton term to $\mathcal{L}$ which penalizes the squared difference between the euclidean norm of the label vector and model output. This incentivizes $\mathcal{F}$ to output vectors containing probability density proportional to the density of the label vector.

After each training epoch, we monitor training progress by calculating precision & recall (micro and macro-averaged), and total accuracy on the validation set. To combat overfitting, we also propose a custom early stopping strategy. Our application interprets false negatives (predicting a SIG did not lobby a bill when they did) as costlier than false positives. This is because public lobbying behavior can be thought of as *implicit feedback*, where lobbying a bill indicates interest, but not lobbying a bill does not indicate disinterest. Thus, it is critical that our model operates well in a high-recall (low false-negative) regime. Therefore, after each epoch, we calculate the macro-averaged precision (MAP) at a recall of 0.8 (a value we deemed high subjectively). If the MAP falls below 10%, we halt training and retain the model weights from end of the previous epoch.

## 5 Results

### 5.1 Principle Components Analysis

We performed SVD on $\mathcal{L}^{(116)}$ assuming 5 latent dimensions and analyzed the projection of the bills into the latent space. Then, we visualized rows of $V^{(116)}$, where each row is a bill and each element is the coefficient in a concept dimension (full plots are available here: 2D, 3D; top 15 2D, top 15 3D). Studying the interactive plots reveal underlying structure in the bill / concept distribution: in particular, we believe the second dimension characterizes a bill's relation to health issues [4]. This latent structure - obtained by solely analyzing lobbying behavior - demonstrates that there is a significant relationship between bills and lobbying behavior that can be studied with straightforward SVD analysis. This justifies leveraging complex techniques like NLP and itemset mining.

### 5.2 Association Rule Learning

Below are the top three most "interesting" and "diverse" association rules learned for the house and senate across all congresses in the dataset (the second row in Table 1). We use the *Lift* metric as a measure of interest and the method proposed in 4.2.1 as a measure of diversity.

```
Senate
+-----------+-----------+-----------+--------+--------+---------+----------+
|antecedent |consequent |confidence |lift    |support |ant_cat  |con_cat   |
+-----------+-----------+-----------+--------+--------+---------+----------+
|[s3000-114]|[s2943-114]|0.6207     |18.4284 |0.01636 |[BUD]    |[DEF]     |
|[s3036-110]|[s1733-111]|0.6697     |14.2232 |0.01621 |[ENV]    |[ENG]     |
|[s2191-110]|[s1733-111]|0.5773     |12.2607 |0.02070 |[ENV]    |[ENG]     |
+-----------+-----------+-----------+--------+--------+---------+----------+
House
+------------+------------+-----------+--------+--------+--------+---------+
|antecedent  |consequent  |confidence |lift    |support |ant_cat |con_cat  |
+------------+------------+-----------+--------+--------+--------+---------+
|[hr2647-111]|[hr3326-111]|0.6743     |12.7563 |0.02391 |[DEF]   |[BUD]    |
|[hr2-114]   |[hr1628-115]|0.5948     |12.0999 |0.02043 |[MED]   |[HTH]    |
|[hr1585-110]|[hr3222-110]|0.5500     |11.7359 |0.02198 |[DEF]   |[BUD]    |
+------------+------------+-----------+--------+--------+--------+---------+
```

These association rules, while technically having diverse categories, contain bills with relatively similar purposes. For example, `s3036-110` is the *Lieberman-Warner Climate Security Act of 2008*, which establishes a number of requirements around the reporting and limiting of greenhouse gas emissions. This bill implies `s1733-111`, or the *Clean Energy Jobs and American Power Act*, which proposes a very large number of climate-related measures around carbon capturing, water efficiency, and renewable energy use in public transportation. When reading the bill text, it is clear the differences in bill categories (BUD and DEF) are not representative in difference in bill objectives.

From the house rules, we can see that `hr2-114` adjusts the methodologies allowed for calculating payments for a physician's services in Medicare (among other medicare-specific changes), and `hr1628-115` removes funding for the Prevention and Public Health Fund, restricts Planned Parenthood from receiving certain types of federal funds for one year, and disallows non-elderly individuals with an income of less than 133% of the poverty level from receiving Medicaid coverage. While these bills were categorized differently, they clearly both cover hotly-debated public health issues.

[4] As suggested, almost all bills in the "Health" plot have positive y values, while the "Taxes" plot seems to have slightly negative values. In addition, the outliers in the "Taxes" plot actually support our argument. Upon further inspection, we identify these bills as H.R. 748, S. 172, and H.R. 1398. H.R. 748 is the "Middle Class Health Benefits Tax Repeal Act of 2019," and both H.R. 1398 / S. 172 are the House / Senate version of the "Health Insurance Tax Relief Act of 2019."

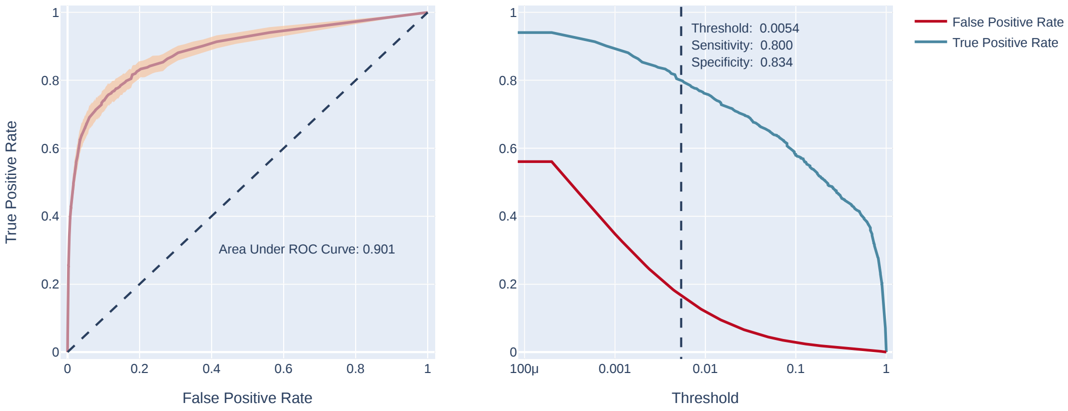

Figure 2: (**Left**) Macro-Averaged Receiver-Operator Curve computed on the evaluation set. The orange area indicates bootstrapped 95% confidence intervals. (**Right**) TPR and FPR as a function of prediction threshold. Lower thresholds are required due to the size and sparsity of the label set $\mathcal{Y}$. The dashed line indicates the threshold, sensitivity, and specificity at which Recall is 80%.

### 5.3 Textual Approximation of the SIG-Bill Matrix

#### 5.3.1 t-SNE & Visualization

We performed a 2-dimensional t-SNE analysis on mean-aggregated LegalBERTembeddings extracted from each bill title. Unfortunately this plot would not fit in our report and has been included in the appendix (See appendix A.3 for more information). This visualization shows that clusters uncovered from t-SNE roughly correspond to bill categories derived in section 4.2.1. From this we can conclude that bills of similar categories exhibit similar textual semantics, motivating the efficacy of a text classifier in the following section.

#### 5.3.2 Text Classification

Our simple neural network trained on LegalBERTembeddings learned to predict which SIGs lobbied with bills relatively well and successfully generalized to the evaluation set without a reduction in performance. Because the label is incredible sparse (i.e., the majority of correct predictions are 0), the final validation accuracy of 0.985 is not very meaningful. Instead we demonstrate the predictive power of the this classifier by creating a receiver-operator characteristic (ROC) curve, which shows the Macro-Averaged true positive rate (TPR) as a function of the false positive rate (FPR). Though the behavior of the classifier will vary based on the chosen threshold, we can assess the overall quality of the model by computing the area under the ROC curve, which is **0.901**. We additionally visualize TPR and FPR as the classifier threshold varies from 0 to 1 (fig. 2). At a sensitivity (i.e., Recall) of **80%** our model exhibits a specificity of **83.4%**, indicating the model is capable of predicting lobbying in the affirmative without under-predicting due to sparsity.

## 6 Discussion

### 6.1 Results

Were omitted from this report because they consistently uncover trivial sets of bills. Examples of trivial relationships include different iterations of the same bill, senate and house versions of the same legislative effort (Note: sometimes the same bill will be introduced in both chambers simultaneously to increase success odds), and legislation that necessarily re-occurs periodically (e.g., federal budget appropriations). Given the uselessness of non-diverse rules, and the marginal intrigue of the above rules, we say this itemset mining and association rule learning fails to uncover interesting or anomalous lobbying behaviors.

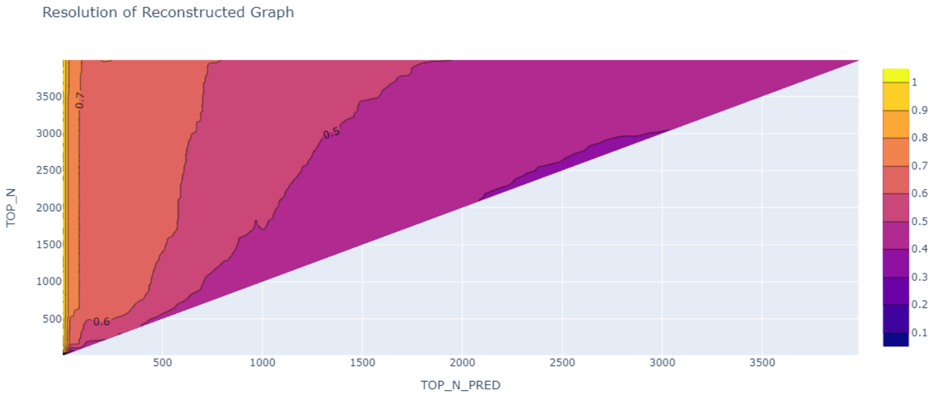

Figure 3: Contour plot of the reconstructed graph's resolution. The X axis is $N_R$, the Y axis is $N_O$, and the contour values are $PR(N_R, N_O)$.

However, the text classifier performed better than initially expected, generalizing to a set of unseen bills with both high sensitivity and specificity. An approximation of our SIG-bill matrix $\mathcal{Y}$ can be thought of as a content-based recommendation system, where model output probabilities are estimated measures of interest in bills for which there was no lobbying activity. Thus, we can use the text classifier in subsequent sections to detect lobbying communities and anomalous bill interest.

### 6.1.1 Case study: Peer Synthesis & Generalizeability

With simple modifications, the data described above can be used to create networks describing similarity between companies. We demonstrate how an artificial peer graph can be constructed, and, paired with modularity analysis (e.g. the Louvain method), further insights can be detected. In particular, a similarity network better contextualizes an individual SIG's behavior given other SIGs, especially as the content of bills evolves over time.

To begin, consider our bill / SIG matrix $\mathcal{Y}^{n \times |\mathcal{C}|}$. We seek to create some adjacency matrix $\mathcal{A} \in [0,1]^{n \times n}$ that represents the similarity between companies. So, we leverage cosine similarity and set:

$$\mathcal{A}_{ij} = \frac{\langle \mathcal{Y}_i \mathcal{Y}_j \rangle}{\|\mathcal{Y}_i\| \, \|\mathcal{Y}_j\|} \tag{4}$$

Where $\mathcal{Y}_i$ is the $i$th column of $\mathcal{Y}$ (or the lobbying profile for a specific SIG). When performed on the ground truth data, we can leverage the Louvain method to classify different companies.

We extend the above graph construction method to demonstrate the generalizability of our classification model. Instead of using $\mathcal{Y}$, we directly feed in probabilities produced by our LegalBERT model. If the synthetic graph produced from our classification model's outputs has high similarity with the ground truth graph $\mathcal{Y}$, we can argue that our model is maintaining the relationships *between* companies. To assess this similarity, we create the “Percent Reconstructed" metric (PR). Namely, suppose $\mathcal{E}_N(G)$ is the set of the top $N$ weighted edges in the graph $G$[5]. Then, we compute:

$$PR(N_R, N_O) = \frac{\mathcal{E}_{N_R}(G_R) \cap \mathcal{E}_{N_O}(G_O)}{N_R} \tag{5}$$

Where $N_R, N_O$ are the parameters describing the number of top edges to collect for the reconstructed/original graph and $G_R, G_O$ are the reconstructed/original graphs.

To test this analysis, we plot $PR(N_R, N_O)$ in fig. 3 over the 116th Congress. The plot implies that we can reconstruct high resolution graphs *even on unseen data*! Thus, this validates the output of our classification model, as it implies that we can use our probabilities to reconstruct a moderate fidelity peer graph (e.g. when constraining $N_R < 500$, our $PR \in [60\%, 70\%)$).

[5] We ignore self-edges and only check if edge endpoints are the same, not edge weights

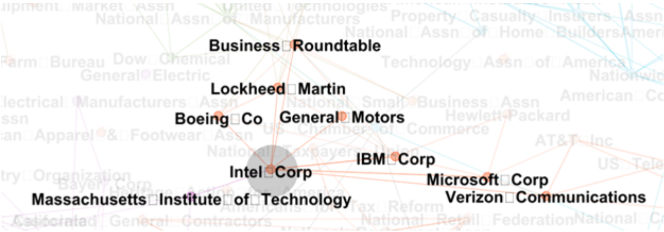

Figure 4: Intel's nearest neighbors, shown in Gephi

#### 6.1.2 Case study: General Motors, IBM, and the Early Push for 5G

We again construct a graph from our model's probability outputs, but now we focus on leveraging this as a tool for finding anomalous behavior. In particular, we can study diverse companies with "close" relationships to find anomalies. For example, GM is Intel's 8th strongest connection (implying a high cosine similarity between the model outputs) and were categorized in the same community by the Louvain method (shown in fig. 4). However, this initially seems strange because Intel and GM are not in similar industries and thus should have different lobbying priority.

So, we craft a simple heuristic to extract interesting bills that both have lobbied. In particular, we can create a general surprise heuristic $S$ where:

$$S(c_1, c_2) = \frac{\mathbb{I}_{c_1} \cdot \mathbb{I}_{c_2}}{\sum_{c_i \in \mathcal{C}} \mathbb{I}_{c_i}}$$

Assuming the denominator is nonzero and where $\mathbb{I}_{c_i} = 1$ only if the model probability $p(c_i) \geq T$ where $T$ is our probability threshold; otherwise, $\mathbb{I}_{c_i} = 0$. This simplistic heuristic is a good metric of "surprise" because it identifies bills that both $c_1, c_2$ are thought to have lobied but few other SIGs are predicted to be interested in. When we perform this analysis, an immediate anomaly appears: S. 2505 from 2014, the Wi-Fi Innovation Act, which broadens 5G access. While Intel has an established 5G division, GM's interest has only recently become apparent - GM is interested in selling 5G cars and this year announced their first 5G car would be sold in China in 2022 Wayland [2020]! Thus, even with a rudimentary heuristic, our methods can help identify anomalous behavior and empower researchers to identify deeper relationships between SIGs which lobby.

### 6.2 Limitations & Challenges

#### 6.2.1 LDA Incompleteness

While the LDA requires lobbyists to disclose their activities, sources attest to the growing unenforcability of the statute Fang [2015] GAO [2020]. In particular, LDAs are aggregated by client, so the exact dollar amount is unknown per bill. In addition, some lobbyists consider themselves "public relations consultants" and thus exclude themselves outside the scope of lobbying.

This assumption is a fundamental challenge for political and computer scientists alike as it means the proxy for measuring lobbying behavior is incomplete. In particular, we believe that greater administrative data (and enforcement mechanisms) are necessary to truly gain a more accurate picture of the extent of federal lobbying.

Last, we acknowledge that using OpenSecret's database introduces issues of data purity - for example, the "Issues" field seems to be human-generated and exhibit a many-to-many relationship with individual bills. This adds further complexity to the dataset which was omitted for simplicity.

#### 6.2.2 Nature of Political Influence

In our conversation with Professor Thorpe of the Political Science department, we learned that the nature of lobbying dynamics also differs from what is depicted in popular culture (e.g. *Scandal* and *House of Cards*). Namely, lobbyists focus less on changing minds or politically pressuring representatives; rather, they advise clients on where to donate money so as to keep friendly Congresspeople in power. Thus, studying the influence of money in politics should not be constrained just to the floor of the Capitol, but also to the earliest stages of the campaign process.

#### 6.2.3 Challenges in the Stupidity of Congress

We have also assumed a simplified bill lifecycle that assumes that the semantics of a bill at introduction are similar to the bill at engrossment and passage. However, this is not always true; for example HR636-114 began as “America’s Small Business Tax Relief Act of 2015," but then was amended to be “Federal Aviation Administration Reauthorization Act of 2016." While this is humorous and shows how quirky Congress can be, it is certainly frustrating from a scientific perspective, as this muddles the semantics and metadata of the Congressional bills in our dataset.

### 6.3 Extension

The tools we’ve developed scrape the surface of machine learning for automated lobbying analyses. In particular, there is a host of metadata on campaign contributions (provided by the Federal Election Commission), lobbyist personal relationships (LittleSis.org), and full bill history/amendments, to name a few. These sources could provide invaluable metadata to improve model learning. Combined with usability improvements - like improved visualizations or a GUI for our techniques - these tools could also include non-technical yet still be flexible and powerful.

## 7 Conclusion

While the impact of lobbyists on Capitol Hill is often opaque, the tools and methods we have developed are one step towards uncovering the behavior of America’s most powerful interests. By leveraging existing big data techniques and pairing them with novel heuristics, we have demonstrated an expressive set of tools for quantifying lobbyist impact and identifying anomalous activity.

First, we outlined the use of traditional big data / dimensionality reduction techniques like PCA and itemset mining for a simplistic analysis of the macro behavior of SIGs. In particular, we noted that these text-naive approaches often fail because there is insufficient metadata to truly expose interesting relationships. This constraint led us to transfer the LegalBERT model to the domain of congressional bills and verify the validity of text embeddings with T-SNE. Then, we leveraged LegalBERT to tackle the multilabel classification problem; in particular, using bill titles as a proxy for bill content and predicting interested companies. We showed that our model generalizes well to unseen bills.

Finally, we described two use cases which build upon proposed methods and demonstrate the utility of work for characterizing lobbying behavior. In particular, we showcased a hypothetical data analytics pipeline that can leverage graph-based or graph-agnostic methods to identify anomalous lobbying activity. This work serves a critical foray into better tooling political scientists. While the host of methods we’ve developed are broad, they are also a solid foundation to innovate additional heuristics and techniques upon, thereby empowering further study.

## 8 Acknowledgements & Team Contributions

We are enormously grateful for the many people who have helped this project succeed. We are generously supported by no grant but will gladly accept payment in the form of high grades. For our bulk data, we thank the Center for Responsive Politics, which runs the OpenSecrets website, and the Government Publishing Office, which runs GovInfo. We also thank the lobbyists we’ve met with, Sarah Castro of UW Federal and Dale Learn of GTH Fed, who have provided invaluable context on the nature of political lobbying and bills in Congress. We are also grateful for Rebecca Thorpe, an Associate Professor of Political Science at the University of Washington, who contributed invaluable background knowledge into the dynamics of lobbying and theory of political science.

### 8.1 Team Member Contributions

Chris: Extracted bill texts from GovInfo, produced PCA/T-SNE analysis, reached out to subject matter experts, created peer synthesis analysis. Jacob: Mined itemset association rules, implemented LegalBERT pipeline, trained LegalBERT and implemented metrics, parsed OpenSecrets bulk data into PySpark dataset. Both: Contributed to research ideation, participated in conversations with subject matter experts, wrote project proposal/milestone report, contributed to project presentation.

## References

I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

L. Fang. Where have all the lobbyists gone?, Jun 2015. URL https://www.thenation.com/article/archive/shadow-lobbying-complex/.

GAO. 2019 lobbying disclosure: Observations on lobbyists' compliance with disclosure requirements [reissued with revisions on jun. 9, 2020.], Mar 2020.

J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. Kornilova and V. Eidelman. Billsum: a corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*, 2019.

K. C. Meng and A. Rode. The social cost of lobbying over climate policy. *Nature Climate Change*, 9 (6):472–476, 2019.

I. Slobozhan, P. Ormosi, and R. Sharma. Which bills are lobbied? predicting and interpreting lobbying activity in the us. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 285–300. Springer, 2020.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

M. Wayland. Gm to launch 5g-connected vehicles in china starting in 2022, Aug 2020. URL https://www.cnbc.com/2020/08/19/gm-to-launch-5g-connected-vehicles-in-china-starting-in-2022.html.

T. Yano, N. A. Smith, and J. Wilkerson. Textual predictors of bill survival in congressional committees. In *proceedings of the 2012 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 793–802, 2012.

# A Itemset Mining

## A.1 Hyperparameters

Table 1: Itemset Mining Data Partitions and FP-Growth Parameters.

| By Session | By Chamber | `minSupport` | `minConfidence` |
|---|---|---|---|
| No | No | 0.03 | 0.6 |
| No | Yes | 0.02 | 0.5 |
| Yes | No | 0.02 | 0.6 |
| Yes | Yes | 0.04 | 0.5 |

## A.2 Merging Issues with Rules

Unfortunately, bills are not labelled by the general issues they address (by the federal government or by OpenSecrets in post). We devised a method of assigning a *general issue category* to each bill by analyzing all issues linked to a bill. Recall that the *issues* table contains a `general_issue` field, which is a simple three-letter code (e.g., *DEF* for defense, *BUD* for budget). Though a bill can represent multiple issues, we propose categorizing a bill by finding the most frequent `general_issue` associated with the bill. For each bill, find all `issue_ids`, select `general_issue`, then find the most frequent `general_issue`. We verify that for most bills there is a clear `general_issue` category that is more frequent than others.

## A.3 T-SNE

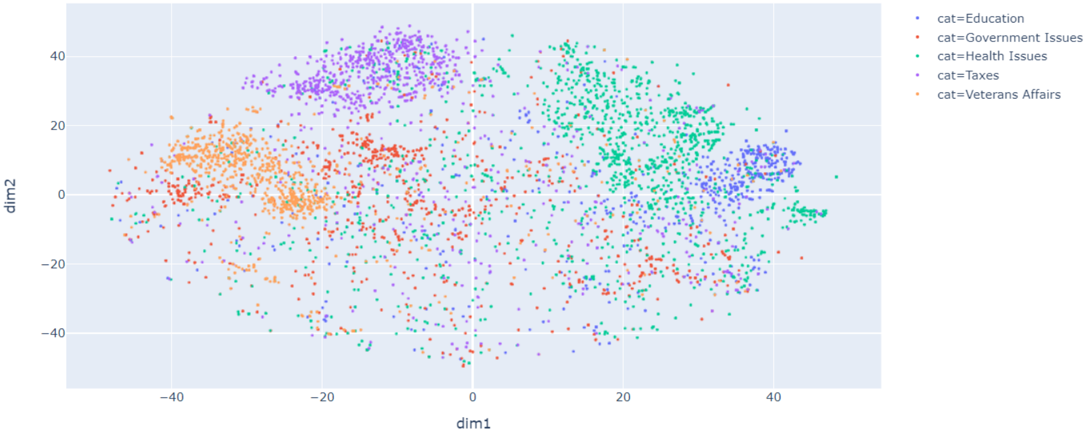

Figure 5: T-SNE Plot with clusters highlighted. The top five categories have clear clusters, suggesting that LegalBERT is appropriately embedding bill titles in a latent space.