Noname manuscript No.
(will be inserted by the editor)

Detecting shadow lobbying

Ivan Slobozhan · Peter Ormosi · Rajesh Sharma

Received: date / Accepted: date

Abstract Lobbying activity is subject to strict disclosure requirements in the US. Failure to comply with these requirements can lead to criminal and civil penalties. It is claimed that these tight lobbying disclosure measures resulted in an increase in 'underground lobbying'. This research proposes a method to discover non-compliance in lobbying disclosure and gauge the magnitude of underground lobbying. We start from the premise that lobbying changes the text of the bills it targets. If these changes happen to some extent systematically, then the texts of lobbied bills should be discernible from non-lobbied bills. We combine the corpus of US legislative bills with a large dataset of lobbying activity to give us a partially labelled dataset, where a positive label indicates a lobbied bill, and the lack of a label indicates either that the bill was lobbied, or was lobbied but not disclosed. To address this partial labelling problem, we first set up a naive classification task, where we assume all unlabelled bills to have a negative label, and train a model on a large corpus of US bills. By finding the best performing model we then design a bagging method and collect out of fold predictions, to predict for each unlabelled bill whether it was lobbied or not. From these predictions we infer that there is a sizable number of bills that are likely to have been lobbied but this lobbying activity was not disclosed. We then investigate how the political affiliation of the sponsoring senators and congressmen relate to these probabilities.

Ivan Slobozhan · Rajesh Sharma
Institute of Computer Science, University of Tartu, Estonia
E-mail: ivan.slobozhan@ut.ee; rajesh.sharma@ut.ee
Peter Ormosi
Centre for Competition Policy, University of East Anglia, UK.
E-mail: p.ormosi@uea.ac.uk

Keywords lobbying · lobbying disclosure · corruption · machine learning

1 Introduction

Lobbying consumes a significant amount of resources. Opensecrets.org website reports that lobbying expenditure reached around $3.5 billion in 2020. The transparency of this intensive lobbying activity is essential in order to protect accountability, the integrity of the process, and to avoid corruption. Driven by this desire to openness, the US Lobbying Disclosure Act, and its modifying legislation, the Honest Leadership and Open Government Act require the registration and disclosure of lobbying activities and funding spent on lobbying.[1] Violations of this requirement carry the possibility of pecuniary (up to $200,000 per violation) and custodial sanctions (up to 5 years in prison). Despite these sanctions, the Secretary of the Senate has referred over 22,000 potential lobbying violations to the U.S. Attorney for the District of Columbia,[2] and we do not know how many lobbying instances go unnoticed and unreported.

Our main contribution is a method to assist in the discovery of lobbying activity by looking at the text of the legislative bills and other features, including bill summary, information about the sponsors and cosponsors of the bill and the topology of the related bills, that may be subject to lobbying activities. For this, we start on the premise that lobbying changes the text of legislation in a way that makes them discernible from non-

[1] [12] offers an overview of 25 years of the Lobbying Disclosure Act.

[2] `https://www.senate.gov/legislative/Public_Disclosure/cumulative_total.htm`

lobbied legislation. This is not a far-fetched assumption. Take rent seeking, which is the economics jargon to describe behaviour which aims at increasing profits through means other than contributing to increasing productivity. Lobbying is a quintessential manifestation of rent seeking behaviour. For example businesses may lobby to block out foreign competition and to thereby cement their domestic market power. If successful, the new legislation will have clauses that limit trade.[3]

A model that can classify bills into lobbied and non-lobbied groups would be useful for multiple reasons. First, it could help the enforcement of lobbying legislation. It has been argued [11] that the increasingly strict disclosure requirements have driven some lobbying activity underground.[4] As long as lobbying effectuates a change in the text of the draft bill, our method should work at flagging the suspicious cases. This has the potential to help gauge the level of undisclosed lobbying activity even when more of the lobbying is now done underground. Second, it could improve the understanding of how lobbying behaviour manifests in legislation and improve legal analysis by discovering classification rules that had been unknown to human analysts. Finally, although the US system is more transparent, the same is not true in jurisdictions where lobbying regulations are relatively new. For example, in the European Union, there is very little information on the laws that are targeted by lobbyists. Using a model trained on US law we could investigate the use of transfer learning together with a much smaller sample of hand-labeled EU data to work on a model fitted to EU laws.

What makes this task more difficult than a straightforward classification problem, is that we have information on the legislative bills that were reported as lobbied (positive labelled), but we have no ground truth for the remaining set of bills (unlabelled). In this latter set, there is a possibility that some were lobbied (and not disclosed), and some were not. To deal with this setup, first, we limit our sample to unlabelled bills and bills that were lobbied at least 50 times. This way we aim to emphasise the difference between the text of unlabelled and lobbied texts ([14] showed that bills lobbied many times are more different in their text from unlabelled bills than bills lobbied only a few times). With this sample we use different sets of features to first draw on a naive model, which assumes that all non-labelled bills were not lobbied (all labelled as negative). We then experiment with different classification models to identify the best performing one. In particular, we used Logistic Regression (LR), Random Forest and Support Vector Machine (SVM). We use the best performing model (LR) as our base estimator in our method to detect suspicious bills. Then we use a bagging method to estimate, for each bill, the probabilities that they were subject to lobbying using out of fold predictions from our base estimators. Through this we show that there is a considerably large number of previously unlabelled US bills where our predictions suggest that some lobbying activity took place. This is more likely to be in certain subject areas, such as energy and healthcare.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 discusses the dataset and Section 4 lays down the proposed methodology. Section 5 describes the results, and Section 6 concludes the paper.

2 Related works

There is a well-established body of literature on lobbying, and it is beyond the remits of this paper to provide a full-fledged overview of these. In a systematic review of the relevant empirical works, [4] takes account of the main strands of empirical papers and the challenges to empirical research on lobbying. We contribute to this literature along different lines. The closest we can relate our paper to previous literature is in the area looking at the impact of lobbying on the specific bills they are targeting. [6] found a direct association between lobbying activities and bill outcomes, and that public attention reduces the effects of lobbying efforts, suggesting that lobbying is most effective when focused on less salient issues. In another paper, [17] looks at the difference between bills that were lobbied ex post and those lobbied before they were passed. Finally, in [7] the authors look at the determinants of interest group lobbying on particular bills after the bills have been passed, and identifies the areas where lobbying focusing on the implementation (rather than the formation) of legislation is more likely. We draw from this literature in interpreting our results, but fundamentally our work differs in that we are developing a prediction model, rather than make inferences from past data.

The paper also relates to works on lobbying intensity. From the economics and finance literature we know that stakeholders with the largest expected profits from favourable policies and regulations are most likely to lobby most intensively [10]. For this reason we expected more intensive lobbying associated with more discernible (for the algorithm) features when compared to non-lobbied legislation.

[3] [14] looks at the ngrams that are most important indicators of lobbying activity.

[4] The Sunlight Foundation made similar claims: `https://tinyurl.com/ywu2mem6`

In general, our work is also relevant for the literature on NLP applications in the legal domain.[5] Some of these are about automating the process of summarising legal texts, such as court rulings [5] or [9]. A subset of these applied NLP works in law draws on text classification methods. For example, [2] use text classification methods (TF-IDF for feature extraction and SVM for text classification) in order to classify which domain a legal text belongs to. In another paper, [13] propose a semi-supervised learning method to classify legal texts. Finally, a large number of NLP applications in law focus on prediction. [16] set out to predict various aspects of patent litigation, with mixed results. Other works focus on the prediction of court rulings, such as the European Court of Human Rights (ECRH) decisions by [1], or French Supreme Court rulings by [15].

3 Dataset

We use two main sources to assemble our dataset. First, the texts of the bills were from the US Congress' website. Altogether there were 308,125 bills, but we decided to remove old bills (before year 2000) because lobbying information was scarcely available for these earlier years. This left us with a sample of 92,361 legislative bills, which includes all congresses after (and including) the 107th congress.

We downloaded the text of the bills,[6] and the texts of the summary of the bills separately.[7] Metadata on the bills was also collected.[8] This included information on the year the bill passed; the congress number; the type of the bill (House Concurrent Resolution, House Joint Resolution, House of Representatives Bill, House Resolution, Senate Bill, Senate Concurrent Resolution, Senate Joint Resolution, Senate Resolution); the name, the political party, and the state of the sponsor of the Bill; the name and political party of co-sponsors; the main subject of the bill; and the related bills.

We downloaded lobbying information from opensecrets.org, a database constructed and curated by the Center for Responsive Politics (CRP). Lobbying data contained information on each lobbying instance, therefore for each lobbied bill, we had a list of all lobbying instances, which allowed us to derive a measure for lobbying intensity (the number of lobbying instances that were recorded by CRP). From this we could ascertain 58,452 bills that were subject to lobbying, to varying degrees. We considered this set labelled as positive because we had reliable information (they appeared in the records of OpenSecrets.org) that they were lobbied. For other bills, we have no evidence whether they were genuinely not lobbied (negative) or were lobbied but it was not reported (an offence under the Lobbying Disclosure Act). We return to this identification problem in our methodology.

Table 1 summarises the bills in our sample by their primary subject area. Healthcare, taxation, defense, international affairs, and trade are the most frequently submitted bills in our sample. Table 1 also shows the proportion of the bills with Republican sponsors by each subject area. It shows us that in most areas there is a roughly equal split between the bills proposed by Republican and by Democrat sponsors. On the other hand, bills on subjects like economics and public finance are more likely to be proposed by Republican sponsors, and bills on subjects such as education, labour and employment are more likely brought forward by Democrat sponsors. Similar patterns can be observed in our third column, which shows the average percentage of Republican co-sponsors on each bill. Finally, the table reveals the percentage of bills in each subject area for which we had evidence of lobbying. Health, finance, energy, and environmental protection are among the most lobbied subject areas.

Many of the bills in our sample have been subjected to intensive lobbying activity. Opensecrets.org records each lobbying instance as a separate entry, and it is possible (and often the case) that the same bill was the subject of more than one recorded lobbying instances. We use this information as a measure of lobbying intensity. Figure 1 shows the distribution of the bills based on the intensity of lobbying. As expected, the distribution has a long right-tail, most bills recording a small number of lobbying instances, and a few with very high lobbying activity.

Finally, in Table 2 we compare labelled and unlabelled bills regarding some of the features we recorded for each bill. We distinguish between lobbied bills based on the intensity of lobbying (1-49 times and 50 or more times). The main difference seems to appear in the length of the text of the bills (number of words) and the number of cosponsors a bill had. Bills with evidence of lobbying activity are longer and had more cosponsors. This difference is more pronounced for more intensive lobbying activity.

[5] [3] gives an overview of the relevant literature.

[6] For an example of the text see: https://www.congress.gov/bill/110th-congress/house-bill/2316/text

[7] For an example on the summary of the bill see: https://www.congress.gov/bill/110th-congress/house-bill/2316/summary

[8] https://www.propublica.org/datastore/dataset/congressional-data-bulk-legislation-bills

Table 1 Most common subjects in the sample

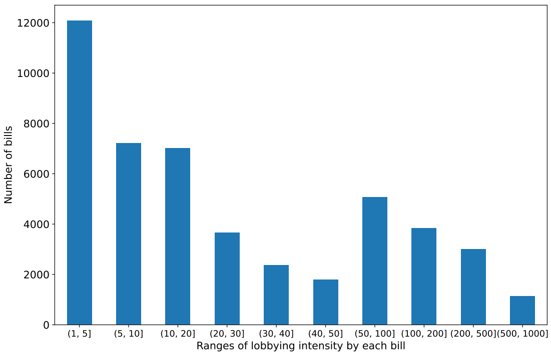
subject	**number of bills**	**ratio of bills with Republican sponsor**	**average ratio of Republican co-sponsors by bill**	**ratio of bills lobbied**
Health	9251	0.406	0.364	0.722
Taxation	8056	0.512	0.489	0.678
Armed forces and national security	6639	0.440	0.408	0.606
Foreign trade and international finance	6575	0.483	0.484	0.534
International affairs	5397	0.491	0.456	0.365
Public lands and natural resources	4974	0.517	0.489	0.574
Government operations and politics	4846	0.430	0.423	0.540
Congress	4105	0.561	0.558	0.203
Crime and law enforcement	4033	0.426	0.400	0.634
Commemorations	3673	0.459	0.442	0.111
Education	3555	0.248	0.249	0.616
Transportation and public works	2941	0.423	0.402	0.683
Finance and financial sector	2569	0.485	0.452	0.775
Energy	2355	0.466	0.451	0.755
Commerce	1975	0.339	0.337	0.694
Labor and employment	1899	0.351	0.339	0.690
Immigration	1899	0.522	0.532	0.644
Environmental protection	1888	0.476	0.456	0.735
Science, technology, communications	1561	0.448	0.427	0.731
Agriculture and food	1508	0.389	0.362	0.707
Economics and public finance	1456	0.662	0.650	0.624
Emergency management	1294	0.442	0.412	0.608
Social welfare	1285	0.422	0.389	0.579
Housing and community development	1113	0.337	0.320	0.707
Native Americans	1069	0.486	0.446	0.610

Table 2 Positive (lobbied) and unlabelled samples compared

	number of words in the bill	**average ratio of rep sponsors**	**ratio of bills under Republican president**	**ratio of bills with Republican sponsor**	**total number of cosponsors**
negative (unlabelled)	1062.892	0.438	0.607	0.456	10.026
	(3961.283)	(0.407)	(0.489)	(0.498)	(21.29)
positive (lobbied 1-49 times)	1991.959	0.416	0.46	0.433	14.229
	(6023.41)	(0.4)	(0.498)	(0.495)	(29.417)
positive (lobbied $\geq$ 50 times)	5739.441	0.459	0.326	0.499	24.007
	(20666.086)	(0.382)	(0.469)	(0.5)	(42.424)

Standard deviation in parentheses.

Fig. 1 Lobbying activity distribution



4 Methodology

In this section, we describe our proposed method to discover non-compliance in lobbying disclosure or, in other words, to detect bills that have been subject to lobbying activity despite not being reported. First we use a naive approach to find the best performing model for a binary text classification problem (lobbied - not lobbied), then we use this model with a bagging out of fold approach to predict lobbying probabilities for each bill.

4.1 Finding the best model for prediction

At the heart of our project is the issue that although we have information on the bills that were lobbied and disclosed (positive label), we do not have the ground truth on whether the legislative bills that were not part of our lobbying dataset were genuinely not lobbied, or were lobbied but the lobbying activity was not disclosed. At the same time, we also cannot validate non-lobbied bills externally (for example, using the support of specifically trained specialists due to lack of resources).

Unfortunately, applying Positive-Unlabelled (PU) learning approaches or metrics is not possible in our case as we do not have a set or reliable negatives for the evaluation. This is simply because those unlabelled bills that had been lobbied are kept secretly by lobbyists in order to avoid being charged with a lobbying disclosure offence. Therefore, to mitigate this issue and obtain negative samples that we can use to make our problem a binary classification one, we do the following two things.

First we reduce our sample to only include cases where the text of lobbied and unlabelled bills is more discernible. For this, we rely on [14], who showed that more intensive lobbying changed the text of the bills more extensively. For this reason, we posit that a sample containing bills that were lobbied at least 50 times, can be confidently assumed to exhibit signs of lobbying. On the other hand, our unlabelled sample may or may not contain instances of lobbying (we are trying to estimate the likelihood of this). However, even if there were undisclosed cases of lobbying in the unlabelled sample, it is unlikely that any of these bills would have seen more than a few instances of lobbying, otherwise the chance of it being disclosed would have become too high. Because of this, this sample (potentially containing bills with low levels of lobbying that were undisclosed) may be to some extent similar to the sample where we have the bills with 1-49 instances of lobbying (containing bills with low levels of lobbying). For this reason, to help our classification, we remove this latter sample of cases.

Second, purely for the model selection stage, in our classification exercise we assume that all unlabelled bills are negatives. We are aware that this assumption may not hold in all the examples, but it should be a reasonable assumption, given that in the first stage we are conducting a simple classification task to distinguish between unlabelled (no recorded lobbying) and intensively lobbied (¿=50 instances of lobbying) bills. Moreover, we believe that this assumption does not jeopardise our results, as we only use this assumption for the model selection, and not the prediction.

The above two assumptions simplify our problem to binary classification and give us the ability to select the best machine learning method using conventional performance metrics.

4.1.1 Feature creation

This section describes the features and how we preprocess them to solve our text classification problem. We used the full texts and summaries of the bills, all additional information available in the bills metadata. More precisely, we utilize four types of features: (1) the text of the bill, (2) a summary of the bill, (3) information about the sponsors and cosponsors, (4) information about related bills for a given bill.

The **text of the bill** is represented by English words. We apply conventional text pre-processing steps to our raw textual documents. Our steps are the following: (1) lowercase the text, (2) delete numbers, (3) delete English stopwords, (4) delete law stopwords, (5) delete HTML tags, (6) delete special characters and punctuation, (7) delete 10% most frequent words and 15 least frequent words, and (8) apply lemmatization. Then we transform the pre-processed text into a set of features that can be fitted into a machine learning model. We use TF-IDF features with a bag of unigrams and bi-grams.

The **summary of the bill** is related to the text of the bill. Summary of the bills is also a piece of textual information, but it is much smaller than the bill's original text in terms of length (number of words). Because of the similarity to the text of the bill, we process the summary of the bill similar to the text of the bill, but without deleting HTML tags, as the summary of the bills are clean texts extracted from metadata rather than scraped from the website. Finally, we also transform cleaned texts from summaries using the TF-IDF technique.

Sponsors and **cosponsors** are represented by the unique identifiers (id) of the senators. To transform this information into numerical features, we perform one-hot encoding. For each of the senators, we firstly assign a unique index from 0 to the number of all unique senators (n). Then, to encode a particular senator, we firstly create a zero vector of length n. After that, we put one on a position that equals the unique index of this senator. Finally, because a particular bill can have sponsors and many cosponsors, we sum up all the vectors for each senator. In the end, we normalize these vectors for each of the bills using min-max scaling (normalization).

Each of the bills has the associated **number of related bills**. More precisely, each bill can have either none of the related bills or many related bills. We generate different network features based on the bill network topology to encode this information. Firstly, we calculate basic network features. For each of the bills, we calculate its (1) centrality, (2) closeness, (3) betweenness, (4) clustering coefficient, and (5) page rank. Because some bills do not have any related bills, we put zero as a value of this feature. In addition, we calculate more sophisticated features, such as node embeddings. For this, we apply the node2vec algorithm [8] and select parameters p, and q equals 1, with the embedding size equal to 16. Then each of the embeddings we concatenate with the rest of the features. Finally, we normalize them across all the bills.

Our final features set is simply a concatenation of the above four features types (All features). These feed directly into our machine learning model.

4.1.2 Metrics

We checked the performance of three algorithms (LR, SVM, and Random Forest) using two main classification metrics: F1-score and area under a receiver operating characteristic curve (AUC ROC), because our dataset has a marginal imbalance problem (there are more negative examples than positives)

1. **F1-score (F1):** is a harmonic mean of precision and recall, where the best possible value is 1 and the worst is 0.
2. **AUC ROC:** is equal to the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative one. AUC ROC is calculated by plotting true positive rate against the false-positive rate at different thresholds. True positive rate is the proportion of actual positives that are identified correctly, and the false-positive rate is the ratio between the false positives and the total actual negative cases. After that the area of this curve is calculated to get AUC ROC. The perfect binary classifier will have AUC ROC equal to 1, and in a random binary classifier ROC AUC equals to 0.5.

4.1.3 Evaluation

To split our dataset into train & test sets we use two approaches:

1. A random stratified one hold split of all the bills (80% train, 20% of test)
2. A chronological split by 2017. All the bills before and including 2017 is in train set, bills after that year is a test set.

Table 3 shows the evaluation metrics for three different methods (LR, SVM, and Random Forest)[9] Each column represents a different type of data. In the first column we only used the text of the bills, in the second column we used the text of the summary of the bills, column 3 shows the results using information on the sponsors/co-sponsors, column 4 shows the results looking at the related bills, and in the last column we present the evaluation metrics where we used all of this information together. One can think of this as a stylised ablation exercise. The figures in Table 3 show that textual features (the text of the bills and the summary of the bills) contribute most to the performance of our models. In fact, the models using only the text of the bills perform almost as well as the models using all features. The identity of the sponsors and co-sponsors also makes a contribution, but related bills seems to offer least in this respect.

Table 3 Model evaluation

	Text	Summary	Sponsors/ cosponsors	Related bills	All features
Logistic regression					
F1	0.7525	0.7451	0.5606	0.4741	**0.7751**
ROC AUC	0.9320	0.9239	0.8017	0.6891	**0.9423**
SVM					
F1	0.7310	0.7186	0.5605	0.3496	0.7745
ROC AUC	0.8325	0.8230	0.7234	0.5939	0.8597
Random forest					
F1	0.6246	0.6180	0.5606	0.4742	0.7750
ROC AUC	0.8754	0.8661	0.8017	0.6879	0.9422

We also looked at how well the our estimated models perform when instead of a random stratified split between train and test data, we trained our models on pre-2017 data, and looked at how well they performed on a test sample of 2017 and 2018 bills. Table 4 reports these results. The performance of our models drops in this case (where our model learns on past bills to predict lobbying in new bills). Particularly, the performance when using text features has dropped most, which would suggest that the way lobbying changes the text of bills changes over time.

Table 4 Model evaluation (split around 2017)

	Text	Summary	Sponsors\ cosponsors	Related bills	All features
Logistic regression					
F1	0.5889	0.5929	0.4940	0.4888	**0.6078**
ROC AUC	0.7849	0.7943	0.6719	0.6320	**0.8099**
SVM					
F1	0.5800	0.5762	0.4854	0.4743	0.5753
ROC AUC	0.7068	0.7033	0.6238	0.6090	0.7026
Random forest					
F1	0.5179	0.5348	0.4982	0.4799	0.5527
ROC AUC	0.7050	0.7834	0.6940	0.6380	0.7469

4.2 Identifying suspicious bills

Finally, we present our method to identify suspicious bills or unlobbied bills that are likely to be lobbied. In order to do so we first select the best performing model from Section 4.1.3, which is the logistic regression (LR) on a full set of features. Then we select two samples that are most likely to be different (lobbied vs unlobbied) in terms of lobbying intensity. Finally, we use a bagging method to estimate probabilities for each unlabelled bill that it was lobbied using out of fold (oof) predictions and then averaging them.

[9] In [14] we evaluated other models on a similar dataset, such as CNN and LSTM, with various text embedding representations (BoW, TF-IDF, GloVe, Law2Vec) and the logistic regression always performed at the top.

4.2.1 A bagging out of fold approach

To estimate a model that predicts lobbying in a bill, we took our 14,103 bills that were lobbied at least 50 times, took a sample of 14,103 bills from the unlabelled bills and labelled them as non-lobbied. Then we estimated our model (using a logistic model given its relatively good performance and speed) and deployed it on the remaining 43,934 - 14,103 = 29,831 ‘unlabelled’ sample to predict the probability that a given unlabelled bill was lobbied. We then moved on to the next iteration, where we used the same lobbied sample, but another 14,103 unlabelled bills were selected from the unlabelled sample of 43,934 bills and labelled as non-lobbied. Then we estimated our model for this new set of labelled bills, and deployed it on the remaining sample, and so on. For each unlabelled bill and for each iteration, we stored the estimated probabilities that it was lobbied. We ran 5 iterations, and for each bill we took the average of the predictions as a probability that an unlabelled bill was directly or indirectly affected by lobbying activity. This cross-validation iteration process is shown on Figure 2.

Figure 3 plots the annual average of these predicted probabilities over time. As this is an average figure over a full sample of unlabelled bills, this figure is likely to be directly proportional to the annual probability of a lobbying disclosure offence. Figure 3 shows an increasing trend of unlabelled bills being affected by lobbying. It also shows that since 2016, 10% of the unlabelled bills had over 90% probability that it was lobbied but not reported.

In the following section we look at how this predicted probability of lobbying offence is correlated with the bill metadata.

5 Results

In this section we evaluate the probabilities derived in Section 4. Our interpretation of these predictions is that they represent the probability that a bill was lobbied, despite the fact that it does not feature in the list of bills registered in the lobbying database. Put differently, this is evidence that some lobbying activity was not disclosed as required by the Lobbying Disclosure Act. The higher the predicted probability, the more likely that the bill was lobbied (and not reported), i.e. the more likely that a lobbying offence took place.

Figure 4 plots the kernel density curves of this probability for three groups of observations in our sample, depending on the percentage of Republican sponsors and co-sponsors that were associated with the bill. The figure implies that bills, where the share of Democrat sponsors is higher (i.e. where the share of Republican sponsors is lower), are more likely to have been lobbied and not reported. This would suggest that lobbyists targeting bills put forward by Democrat sponsors may be less prudent in disclosing their lobbying activity.

On the other hand, when looking at the sponsors that are associated with bills with the highest predicted likelihood of a lobbying disclosure offence, we can see that there is a dominance of Republican sponsors, with 15 out of the 20 sponsors with the highest average probability of a lobbying disclosure offence are Republican. These two descriptive findings suggest that the political affiliation of the sponsors and co-sponsors is important in terms of the probability of committing a disclosure offence, but the relationship requires a more detailed look.

Table 5 Average probability of unreported lobbying - by sponsor

sponsor’s name	probability	sponsor’s state	sponsor’s party
Costello, Ryan A.	0.867	PA	Republican
Strange, Luther	0.859	AL	Republican
Smith, Tina	0.816	MN	Democratic
Hassan, Margaret Wood	0.797	NH	Democratic
Bergman, Jack	0.791	MI	Republican
Cortez Masto, Catherine	0.778	NV	Democratic
Rutherford, John H.	0.769	FL	Republican
Johnson, Mike	0.755	LA	Republican
Massie, Thomas	0.753	KY	Republican
Pocan, Mark	0.750	WI	Democratic
Bost, Mike	0.743	IL	Republican
Rounds, Mike	0.742	SD	Republican
Carter, Earl L. ”Buddy”	0.742	GA	Republican
Walters, Mimi	0.741	CA	Republican
Smith, Jason	0.738	MO	Republican
Kustoff, David	0.731	TN	Republican
Katko, John	0.721	NY	Republican
Comer, James	0.718	KY	Republican
Clark, Katherine M.	0.711	MA	Democratic
Sasse, Ben	0.711	NE	Republican

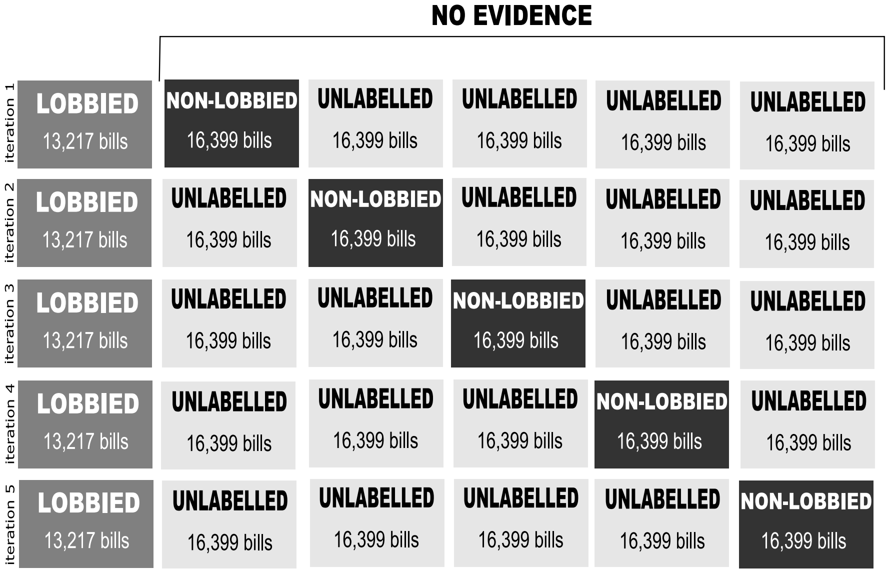
To get some further insight, we estimated the following linear model:

$$\begin{aligned} prob_i = {} & \beta_1 ratio_i + \beta_2 pres_i + \beta_3 spon_i + pres \\ & \times (\beta_4 ratio_i + \beta_5 spon_i) + \beta_6 ratio_i \times spon_i \\ & + \beta_7 ratio_i \times pres_i \times spon_i + \vec{\gamma}\vec{X} + \varepsilon_i \end{aligned} \tag{1}$$

Where $prob_i$ is the predicted probability for bill i that a bill was lobbied but the lobbying was not disclosed, $pres_i$ is the president in power at the time of passing bill i, $ratio_i$ is the ratio of Republicans co-sponsoring bill i, $spon_i$ is whether the main sponsor of the bill is republican, and $\mathbf{X}$ is a vector of other features, such as a time trend, fixed effects for sponsor’s home state, and the length of the bill. We are interacting our main variables of interest to investigate non-linear relationships between them and our dependent variable ($prob$).

The estimates from this regression are presented in Table 6. As the coefficients of interaction and quadratic

Fig. 2 Extracting information from non-labelled data



NO EVIDENCE

iteration 1	LOBBIED 13,217 bills	NON-LOBBIED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills
iteration 2	LOBBIED 13,217 bills	UNLABELLED 16,399 bills	NON-LOBBIED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills
iteration 3	LOBBIED 13,217 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	NON-LOBBIED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills
iteration 4	LOBBIED 13,217 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	NON-LOBBIED 16,399 bills	UNLABELLED 16,399 bills
iteration 5	LOBBIED 13,217 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	UNLABELLED 16,399 bills	NON-LOBBIED 16,399 bills

Fig. 3 Average annual probability of lobbying offence

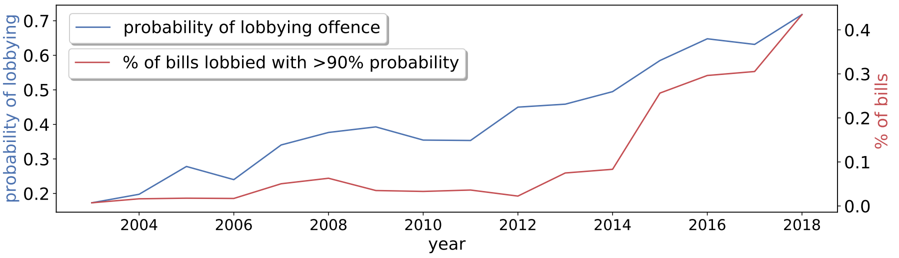
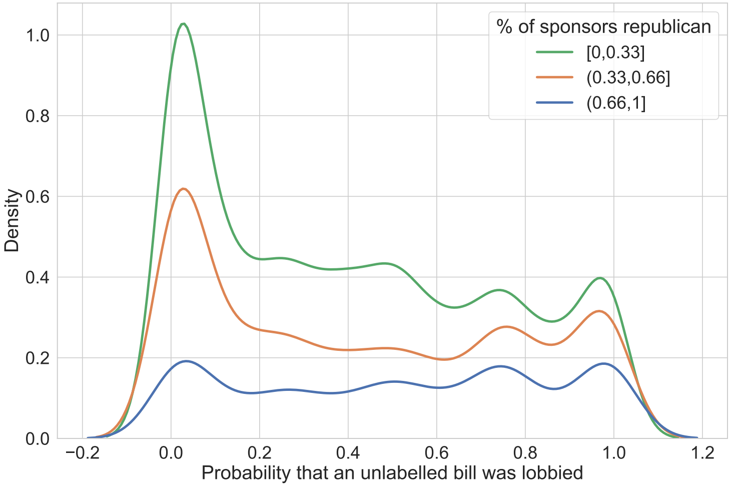
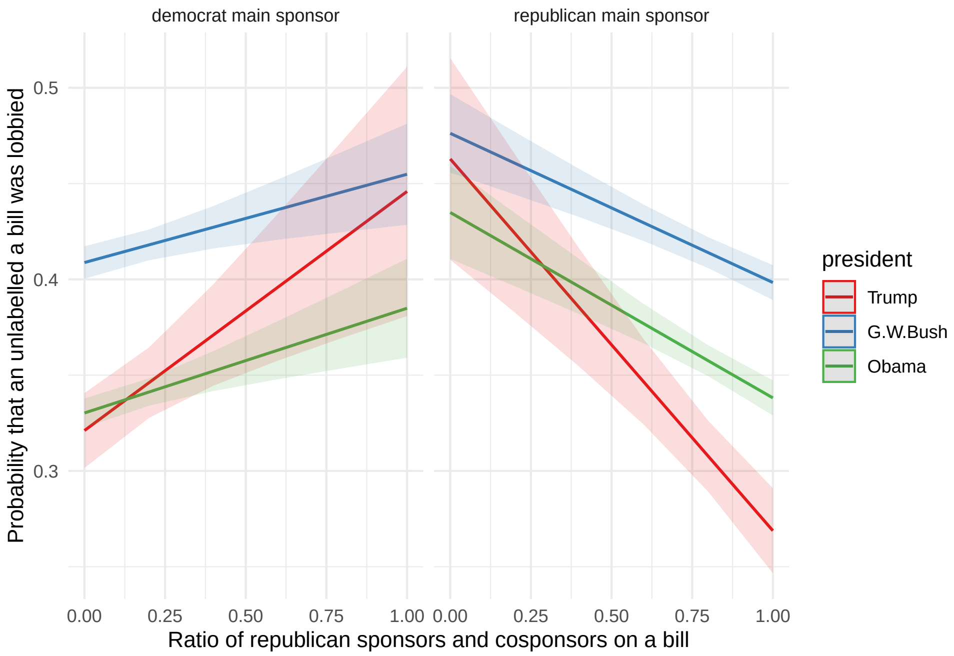


Fig. 4 Kernel density of the predicted probabilities that a bill was lobbied



terms are not straightforward to interpret, on Figure 5 we plot the predicted values of the probability of failing to disclose information on a lobbied bill using these regression coefficients. The figure suggests that in cases where the party composition of co-sponsors is similar to the party affiliation of the main sponsor, the probabil-

Fig. 5 Sponsors' political affiliation and probability that a bill was lobbied



ity of a lobbying disclosure offence is the smallest. For example, where the main sponsors is republican (right hand side panel in Figure 5), having mainly democrat co-sponsors made it significantly more likely to be associated with a bill where lobbying was not disclosed. Moreover, although more recent instances of lobbying have a higher probability of not being reported, once we control for this time effect, the estimates from Eq.(1) suggest the highest probability of a lobbying disclosure offence under the G.W.Bush presidency. The estimated models have a high R^2, but we are less concerned about the fit of this model, as our main interest at this stage is in making inferences on the relationship between the political affiliation of the sponsors rather than to find the best possible model for recall and accuracy.

We report the coefficients for our main variables in Table 6 using 4 different model specifications. The first model only controls for the party affiliation of co-sponsors and the presidency, model (2) controls for the length of the bill, and adds a time trend, model (3) adds a fixed effect for the state of the main sponsor, and model (4) adds sponsor fixed effects. This shows that adding more and more granular information that could explain variation in the predicted probability of a lobbying disclosure offence (for example controlling for sponsor fixed effects) increases the fit of the model. On the other hand, our main findings remain robust across these different models. Other factors also matter, for example longer bills, or more recent bills are also associated with higher probability of a lobbying disclosure offence.

Finally, we also looked at how our predicted probabilities of a lobbying disclosure offence varied across different bill subjects. Table 7 shows the average probability that a bill was lobbied (and not disclosed) by subject area. The subjects that most likely to attract failure to disclose lobbying are energy, finance, and environmental protection.

6 Conclusion

This paper presented a method to estimate the probability of lobbying disclosure offences. This is the first evidence on the potential magnitude of shadow lobbying, i.e. where lobbying activities are not disclosed to the public. This is particularly timely, as a number of commentators have warned of the possibility that lobbying disclosure legislation would drive lobbying activities underground. Our proposal helps in gauging the level of unreported lobbying even if much of this lobbying now happens underground.

We offer a bagging out of fold approach to address the problem of not knowing the ground truth about any of the unlabelled bills (whether they were not lobbied, or were lobbied but not disclosed), which is inherent in our data. This allowed us to estimate the probability that a given legislative bill was lobbied. Looking

Table 6 Regression results from Eq.(1)

	(1)	**(2)**	**(3)**	**(4)**
ratio_of_rep_cosponsors	0.054***	0.046***	-0.139***	0.033**
	(0.015)	(0.015)	(0.014)	(0.015)
rep_sponsor	0.043***	0.067***	-0.066***	0.369***
	(0.011)	(0.010)	(0.010)	(0.101)
length		0.020***	0.023***	0.019***
		(0.001)	(0.001)	(0.001)
year		0.040***	0.023***	0.032***
		(0.001)	(0.001)	(0.001)
G.W.Bush	0.255***	-79.015***	-45.511***	-63.655***
	(0.003)	(1.398)	(1.294)	(1.504)
Obama	0.447***	-79.094***	-45.555***	-63.709***
	(0.003)	(1.402)	(1.298)	(1.509)
Trump	0.633***	-79.103***	-45.571***	-63.724***
	(0.009)	(1.406)	(1.301)	(1.512)
ratio_of_rep_cosponsors x Obama	-0.008	0.009	-0.116***	-0.003
	(0.022)	(0.021)	(0.019)	(0.020)
ratio_of_rep_cosponsors x Trump	0.058	0.079**	-0.049	0.031
	(0.041)	(0.039)	(0.036)	(0.039)
ratio_of_rep_cosponsors x rep_sponsor	-0.114***	-0.124***	0.201***	-0.116***
	(0.020)	(0.019)	(0.018)	(0.019)
Obama x rep_sponsor	0.091***	0.037**	-0.125***	0.002
	(0.017)	(0.016)	(0.015)	(0.017)
Trump x rep_sponsor	0.093***	0.074**	-0.065**	0.041
	(0.031)	(0.029)	(0.027)	(0.030)
ratio_of_rep_cosponsors x Obama x rep_sponsor	-0.039	-0.027	0.258***	-0.0003
	(0.029)	(0.028)	(0.026)	(0.027)
ratio_of_rep_cosponsors x Trump x rep_sponsor	-0.192***	-0.195***	0.087*	-0.138***
	(0.054)	(0.052)	(0.048)	(0.052)
Observations	39,137	39,137	39,137	39,137
R2	0.648	0.683	0.604	0.721

Table 7 Average probability of unreported lobbying - subjects with highest probability

top subject	**probability**
Environmental protection	0.621
Finance and financial sector	0.620
Energy	0.605
Commerce	0.571
Labor and employment	0.559
Science, technology, communications	0.554
Economics and public finance	0.536
Health	0.527
Immigration	0.526
Agriculture and food	0.524

at these probabilities reveal us how political affiliation, and central administration are correlated with the probability of lobbying disclosure offence.

For the future, we would like to build on this paper to help develop a method for estimating lobbying activity in jurisdictions with lower levels of lobbying transparency, such as the European Union.

7 Declaration

7.1 Funding

This work has received funding from the EU H2020 programm under the SoBigData++ project (grant agreement No. 871042).

7.2 Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest.

7.3 Availability of data and material

All the data we used in our experiments are in open access on the website govinfo.gov[10] and can be extracted with the help of the repository[11] on GitHub. We also provide the already aggregated dataset from different sources [12].

7.4 Code availability

For code snippets, please contact the corresponding author, that is, ivan.slobozhan@ut.ee.

7.5 Authors' contributions

Ivan Slobozhan: Data curation, Formal analysis, Investigation, Validation, Visualization, and Writing – original draft.
Peter Ormosi: Data curation, Formal analysis, Problem formulation, Investigation, and Writing – original draft.

[10] https://www.govinfo.gov/

[11] https://github.com/unitedstates/congress

[12] https://css.cs.ut.ee/data.html

Rajesh Sharma: Supervision, and Writing – Review and Editing.

References

1. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ Computer Science **2**, e93 (2016)
2. Boella, G., Di Caro, L., Humphreys, L.: Using classification to support legal knowledge engineers in the eunomos legal document management system. In: Fifth international workshop on Juris-informatics (JURISIN) (2011)
3. Dale, R.: Law and word order: Nlp in legal tech. Natural Language Engineering **25**(1), 211–217 (2019)
4. De Figueiredo, J.M., Richter, B.K.: Advancing the empirical research on lobbying. Annual review of political science **17**, 163–185 (2014)
5. Farzindar, A., Lapalme, G.: Legal text summarization by exploration of the thematic structure and argumentative roles. In: Text Summarization Branches Out, pp. 27–34 (2004)
6. Grasse, N., Heidbreder, B.: The influence of lobbying activityin state legislatures: Evidence from wisconsin. Legislative Studies Quarterly **36**(4), 567–589 (2011)
7. Grossmann, M., Pyle, K.: Lobbying and congressional bill advancement. Interest Groups & Advocacy **2**(1), 91–111 (2013)
8. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855–864 (2016)
9. Hachey, B., Grover, C.: Extractive summarisation of legal texts. Artificial Intelligence and Law **14**(4), 305–345 (2006)
10. Hill, M.D., Kelly, G.W., Lockhart, G.B., Van Ness, R.A.: Determinants and effects of corporate lobbying. Financial Management **42**(4), 931–957 (2013)
11. LaPira, T.: Lobbying in the shadows: How private interests hide from public scrutiny, and why that matters. Cigler, Allan J, Burdett A. Loomis, and Anthony J. Nownes (2015)
12. LaPira, T.M., Thomas, H.F.: The lobbying disclosure act at 25: Challenges and opportunities for analysis. Interest Groups & Advocacy **9**(3), 257–271 (2020)
13. Li, P., Zhao, F., Li, Y., Zhu, Z.: Law text classification using semi-supervised convolutional neural networks. In: 2018 Chinese Control and Decision Conference (CCDC), pp. 309–313. IEEE (2018)
14. Slobozhan, I., Ormosi, P., Sharma, R.: Which bills are lobbied? predicting and interpreting lobbying activity in the us. In: International Conference on Big Data Analytics and Knowledge Discovery, pp. 285–300. Springer (2020)
15. Sulea, O.M., Zampieri, M., Vela, M., Van Genabith, J.: Predicting the law area and decisions of french supreme court cases. arXiv preprint arXiv:1708.01681 (2017)
16. Wongchaisuwat, P., Klabjan, D., McGinnis, J.O.: Predicting litigation likelihood and time to litigation for patents. In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, pp. 257–260. ACM (2017)
17. You, H.Y.: Ex post lobbying. The Journal of Politics **79**(4), 1162–1176 (2017)