2012 11th International Conference on Machine Learning and Applications

Predicting Congressional Votes Based on Campaign Finance Data

Samuel Smith, Jae Yeon (Claire) Baek, Zhaoyi Kang, Dawn Song, Laurent El Ghaoui, Mario Frank
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA
Email: {samsmith, jaeyeon, kangzy}@eecs.berkeley.edu, dawnsong@cs.berkeley.edu, {elghaoui, mfrank}@berkeley.edu

***Abstract*—The USA is witnessing a heavy debate about the influence of political campaign contributions and votes cast on the floor of the United States Congress. We contribute quantitative arguments to this predominantly qualitative discussion by analyzing a dataset of political campaign contributions. We validate that the campaign donations of politicians are mainly influenced by his or her political power and affiliation to a political party. Approaching the question of whether donations influence votes, we employ supervised learning techniques to classify how a politician will vote based solely upon from whom he or she received donations. The statistical significance of the results are assessed within the context of the debate currently surrounding campaign finance reform. Our experimental findings exhibit a large predictive power of the donations, demonstrating high informativeness of the donations with respect to voting outcomes. However, observing the slightly superior accuracy of the party line as a predictor, a causal relationship between donations and votes cannot be identified.**

***Index Terms*—classification, politics, L1-regularization, behavior prediction**

I. Introduction

In recent years there has been an increased interest in how political campaigns are funded and how those who donate money to members of the United States Congress can influence the outcome of legislation. With the involvement of money in American politics at an all-time high, we attempt to determine the extent of the influence of money on the political process. With the Supreme Court decision of *Citizens United*[1], corporations and rich individuals are now able to inject unlimited amounts of money into election campaigns. Many political scholars and commentators have become greatly concerned that the ability to spend limitless amounts of money on advertising to effectively drown out any opposing candidates or points of view could have grave consequences for democracy.

The tenor of the discussion boils down to if and how much donations influence politics. Many qualitative arguments have been made on this question. However, we feel that the discussion as it is led today, lacks a quantitative basis. We provide the first quantitative analysis of how predictive donations are for political votes. Our hypothesis is that there is a causal relationship between a politician's funding sources and how they vote in Congress. To investigate this hypothesis, we employ supervised learning techniques to build models which predict how a politician will vote on a given bill given only information about his or her funding sources.

Our primary method for predicting votes in this paper is to take a given measure from Congress and train a classification model on some subset of the politicians with knowledge of whom they have received money from and how they voted on the measure. We then test our classification model on the remaining subset of the politicians and assess its accuracy as a measure of the statistical significance of the correlation between a politician's sources of money and how he or she votes. To reason about causal relationships, we compare the results with accuracies obtained by simple baseline methods as well as the party line.

For this research we collaborated with MapLight [1], a nonprofit organization that collects information from publicly available sources about donations from corporations and individuals to politicians, the stated opinions of corporations and other organizations on legislative actions, and the records of how members of Congress voted on these measures. MapLight operates a website that allows users to view bills currently before Congress with a breakdown of the money in support of and in opposition to each bill.

We should note that our model requires information about how some members of Congress vote on a bill in order to train our classifiers to make predictions about the remaining politicians. We cannot simply look at the title and content of a bill and determine how someone will vote as there is no semantic information in our dataset. Thus our methodology only works for predicting votes that have already taken place. This can be a powerful tool in uncovering a possible link between money and votes, but we cannot predict future votes without any training data on those votes.

The main contribution of our paper is a careful analysis of the predictive power of political donations on congressional votes. We show, for the first time, a strong correlation between donations and votes. Moreover, we provide an analysis of the main factors that determine variances in campaign donations.

The remainder of the paper is organized as follows. In Section II, we begin with an overview of the datasets that are used to train our classification models. Before predicting votes, we use dimension reduction techniques in Section III to analyze the main factors dominating the dataset and to explore

[1]Citizens United v. Federal Election Commission, 558 U.S. 50 (2010)

978-0-7695-4913-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMLA.2012.119

the possibility of hidden variables that could influence our models. We provide an overview of the algorithms we used in Section IV and finally present and discuss the experimental findings of our vote prediction test in Section V.

II. Dataset Overview

A. Sources

Our source of data is the nonprofit organization MapLight. For our analysis we use the following datasets:

1) Votes: This dataset consists of votes from the United States Senate and House of Representatives on 1262 measures voted on by members of Congress from 2006 to 2012. The data contains a list of entires each with a unique key identifying a politician, another key for a particular Congressional action, and how that politician voted on the measure as either "AYE", "NOE", or "NV" (did not vote).

2) Bill Positions: This data contains publicly stated positions on various bills by various corporations and interest groups. The entries for this dataset include: the name of a particular bill, a description of that bill, a description of the measure[2], the name of a corporation or organization, the opinion of that organization on the measure (support, oppose, indifferent) as decided by MapLight researchers, and finally a full citation for the source of the organization's opinion.

3) Contributions: This is a list of individual and corporate contributions donated to the campaigns of the senators and representatives. Each datum about a corporate contribution contains the name of the organization, a classification of the organization into a particular industrial/political sector and subsector, the amount of the contribution, the politician who received the contribution, and the campaign year for which the donation was counted with the Federal Election Commission (FEC). The data provided was based on MapLight analysis of campaign contributions provided by the Center for Responsive Politics, from candidate filings with the FEC [2]. We would like to emphasize that the particular donation records contain no reference to particular bills. It is illegal in the United States to directly give money to a politician for guaranteeing an outcome on a particular measure. It is however not illegal to publicly indicate the position of your organization with an implicit understanding that your organization may no longer give money to politicians who do not agree with your organization. This makes our task challenging from a technical perspective because our goal is to predict votes on particular bills given the donation data. It should also be noted that these donations are given to political campaigns and not to the politicians themselves as is required under federal law.

4) Politicians: This dataset contains information about each politician. For each politician, it lists their unique key, name, political party, home state, the start of their term, end of their term, and whether they are the currently holding the office listed. For members of the House of Representatives, congressional district is also given [3].

[2]Measures can include voting on actually passing a bill, adding amendments, ending discussion, referring to committee, or various other parliamentary procedures.

[3]Senators represent a state at-large.

5) Sector list: This is a list of 397 different industrial/political subsectors. Each of these subsectors is also grouped into one of 16 more generally defined sectors. For example, the sector "A1300" is defined as "Tobacco and Tobacco Products." The "A" denotes that this industry was part of the general agriculture sector while the "1300" denotes a particular subsector. Our database of contributions also contains various interest groups that donate to politicians. For example, "J7600" represents animal rights groups.

B. Importing and preprocessing

To begin, we take the raw data from MapLight and add up all of the money given to each politician from each of the 16 generally defined sectors as well as the money given to each politician from each of the 397 more precisely defined subsectors. We do not consider the timing of each donation as donations are usually not given with regard to when particular matters come before Congress, but are rather given for each campaign cycle. It would have been possible to construct a time-dependent model, but given the missing temporal relation to bills, we do not think that such a model would have yielded substantially better results.

We choose to use the money from each sector/subsector as the basis for classification instead of the money from each individual corporation or interest group. Considering each organization individually would have made the data highly sparse and difficult to learn from. Looking at the data aggregated by sectors/subsectors may actually yield more useful data as sometimes individual corporations will donate to politicians, but the political opinions of the company may only be made public through an industry-wide advocacy group or political action committee. This is also justifiable for other organizations that are not corporations. For example the National Rifle Association (NRA) is classified as a 501 (c) (3) tax-exempt organization. Organizations in this category are allowed to state positions on particular issues, but they are not allowed to give money to politicians or endorse particular candidates as a provision of their tax-exempt status. There is a legally separate poltical action committee (PAC) called the NRA Political Victory Fund which actively funds candidates. In this case there is a clear relationship between the PAC and the non-profit organization, but this is not always the case.

III. Basic Analysis

We standardize the data and calculate the covariance matrix for the donations. The calculated covariance matrix is visualized in Figure (1). As expected, a high degree of positive correlation occurs near the diagonal, indicating intra-sector correlation. Large blocks can be seen off the diagonal showing sectors that are generally correlated with each other.

We employ principal component analysis (PCA) [3] in an attempt to find a low-dimensional representation of the money given to each politician. PCA involves calculating the dimensions of maximum variance from the covariance matrix. This is done by performing an eigendecomposition of the covariance matrix and sorting the eigenvectors with respect to the largest magnitude eigenvalues.

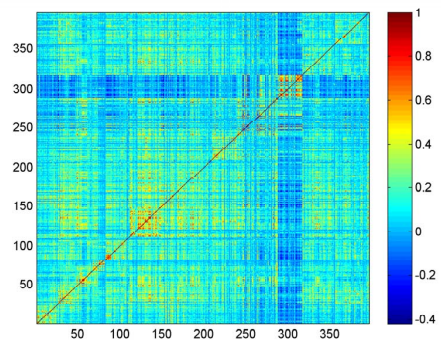

Fig. 1. Correlation matrix for the 397 subsectors

To get a first overview, we project the donations received by each politician onto the two largest principal components. As seen in Figure (2), the politicians are almost perfectly separable by political party along the second principal component. This gives rise to the assumption that the party has a strong causal relationship with the financing sources. As a sanity check of the predictive analysis we will therefore use the party as a baseline predictor.

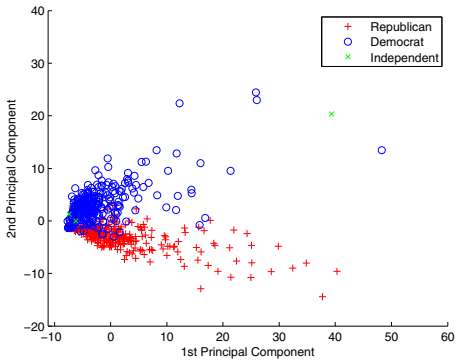

Fig. 2. Projection of each politician and their associated political party onto the first two principal components of the subsector level donation matrix

It is interesting to dissect the principal components of the donation matrix and try to uncover which subsectors were the most influential in maximizing the variance of the contribution money. The first principal component explains 16.48% of the variance in the data as calculated by eigenvalue weight, and projections along this principal component were only slightly dependent on political party. The mean value for Democrats (Republicans) was -1.99 (1.75), with a standard deviation of 7.23 (8.14). A possible interpretation of this principal component is how important the politician is based on how much money they receive from organizations that generally donate to all politicians regardless of ideology. For example, many high ranking members of the Senate were found to have high coefficients for this component. Six of the ten largest elements of this component were part of the financial sector, which is known to give large sums of money to both political parties, as shown in Table I.

The second principal component was highly polarized along party lines. The mean value for Democrats (Republicans) was 3.52 (-3.30) with a standard deviation of 4.85 (2.43). Because of how the results are normalized, sectors with positive loadings have more weight for the Democrats, while negative loadings have more weight for the Republicans. Of the components correlated with positive scores, we found unions, pro-choice advocates, environmentalists, and trial lawyers. For negative scores, the most important elements are pro-gun organizations, builders associations, and small business associations. As noted before, the scores along the second principal component are clearly divisible by political party, as illustrated in Figure (2).

IV. Classification Methods

As motivated before, we want to investigate the relationship between campaign finance donations and the congressional voting of politicians. In this section, we describe different kinds of classifiers to infer voting (AYE or NAY[4]) from campaign money flow. In addition to a naïve baseline predictor, we use two support vector machines (SVM) as parametric predictors, and kNN as a non-parametric predictor.

A. Baseline methods

A naïve prediction method is to toss a coin that outputs either an AYE or NAY vote. If one uses a fair coin, there is an equal probability of getting an AYE or a NAY vote. In addition to that, we construct a predictor that outputs AYE or NAY votes with empirically estimated probabilities given by the proportion of AYE and NAY votes in the training set. Both baseline methods serve as a point of reference for how well our other classifiers predict votes.

B. Party classifier

As mentioned in Section 3, we also chose to use political party as a predicting factor. We constructed a party classifer which takes the majority party vote in the training set and uses the results to predict those in the testing set.

C. k-Nearest neighbors (kNN)

As a non-parametric classifier, we use the k-nearest neighbors (kNN) method. Given a query point X_0, we find the k-nearest neighbors using a distance metric $d(X_i, X_0)$ and assign the class of X_0 by majority vote. kNN requires storing all training observations in memory. However, for the size of our dataset this does not introduce any problem. As k gets smaller, the bias decreases but the variance increases. The reverse holds as k gets larger. We cross-validate on hold-out data to choose k. In most cases, we use $k = 7$ in the evaluation. We also have freedom in choosing the distance function. The candidates include Euclidean distance (L_2-norm) and Manhattan distance (L_1-norm). We find that the L_1-norm tends to give better classification results, probably because the campaign money is not normally distributed.

[4]We ignore abstentions in our analysis as there are usually not enough for each bill to gather significant data.

Ranking	Sector ID	Sector Description	Loadings
1	F0000	Finance, insurance & real estate	0.0999
2	F4100	Real estate developers & subdividers	0.0979
3	T9100	Hotels & motels	0.0945
4	G2900	Restaurants & drinking establishments	0.0944
5	F5100	Accountants	0.0942
6	F3100	Insurance companies, brokers & agents	0.0934
7	F4000	Real estate	0.0933
8	M2300	Industrial/commercial equipment & materials	0.0921
9	F2100	Security brokers & investment companies	0.0920
10	B1500	Construction, unclassified	0.0913

TABLE I
LARGEST TEN ELEMENTS OF THE FIRST PRINCIPAL COMPONENT

Ranking	Sector ID	Sector Description	Loadings
1	J1200	Democratic/liberal	0.1444
2	L1300	Teachers unions	0.1298
3	J7150	Abortion policy/pro-choice	0.1249
4	JE300	Environmental policy	0.1242
5	K1100	Trial lawyers & law firms	0.1239
27	J6200	Pro-guns	-0.1027
32	B0500	Builders associations	-0.0993
33	G1200	Small business associations	-0.0975
48	J2200	Republican leadership PAC	-0.0894
49	J2400	Republican officials, candidates & former members	-0.0840

TABLE II
FIVE LARGEST POSITIVE AND FIVE LARGEST NEGATIVE ELEMENTS OF THE SECOND PRINCIPAL COMPONENT

D. Linear support vector machine

Support vector machines [4] are popular and powerful binary classifiers. SVMs divide the feature space by a hyperplane such that the margin between the two classes is maximized, i.e. SVMs squeeze a maximally thick hyper-brick between the boundary observations of both classes, the so-called support vectors. In contrast to k-nearest neighbors, SVM generalizes from the observed data, i.e. it forgets the individual observations after training and only saves the decision hyperplane in a parameterized way. For more robustness against outliers, a small number of boundary observations are tolerated within the margin. A parameter C controls the trade-off between maximizing the margin and minimizing the number of such exceptions.

For feature x_i with p dimensions and response variable $y_i = \{-1, 1\}$ in which $i = 1, \ldots, N$, where N is the training size, we can construct a hyperplane $\{x : w^T x + b = 0\}$ in which $\frac{w}{\|w\|}$ is the unit vector normal to the hyperplane, and $\frac{w^T x + b}{\|w\|}$ is the signed distance from some vector x to the hyperplane. For data that are not fully linearly separable, we introduce slack variables $\xi_i, i = 1, \ldots, N$, such that

$$w^T x_i + b \geq 1, \qquad y_i = 1 - \xi_i \tag{1}$$

$$w^T x_i + b \leq -1, \qquad y_i = -1 + \xi_i, \qquad \xi_i \geq 0 \ \ \forall i. \tag{2}$$

The above problem can be formalized into a convex optimization problem as below:

$$\min_{w,b,\xi} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{N} \xi_i \tag{3}$$

$$s.t. \ \ y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \ \ \forall i$$

The primal problem is a convex quadratic program with linear inequality constraints. Strong duality also holds. Finally, the classification rule can be written as

$$\hat{G}(x) = \text{sign}(\hat{w}^T x + \hat{b}).$$

SVM is actually itself L_2 regularized. Thus, the regularization coefficient C needs to be determined in advance. With cross-validation, $C = 0.5$ is found to be reasonable value. For this work, we use the MATLAB Bioinformatics Toolbox. In our work, we use a linear kernel.

E. L_1 regularized SVM

SVM has good performance in classification, regression and novelty detection, compared to traditional methods, especially for high dimensional datasets. However, the interpretability of SVM is problematic when a *sparse* result is preferred. There are sparse methods for linear models such as LASSO [5] and L_1-Logistic Regression [6]. In SVM, we can also add an L_1 penalty term to the loss function to yield a sparse result. With an L_1-regularization term, the target function of the optimization problem before becomes

$$\min_{\omega} ||\omega||_1 + C\sum_{i=1}^{l} (\max(1 - y_i \omega^T x_i))^2. \tag{4}$$

where $||\cdot||_1$ denotes the L_1-norm. We use the LIBLINEAR toolbox for MATLAB as an implementation of this method. LIBLINEAR solves the above equation by a subgradient descent method. Due to the sparsity of the optimal solution, some coefficients become zeros. Thus we can shrink our variable set during the calculation. More details about L_1-regularized SVM can be found in the LIBLINEAR paper [7].

In L_1 SVM, we also have the freedom to choose C, the regularization factor. This is done by cross-validation, and $C = 0.5$ is found to be a good choice. Also, we observe that the the outcome is robust with respect to this choice.

V. Results

In this section we report on our experimental findings and provide a discussion. We run the classifiers given above on the donations that each politician received from each subsector. Three different sets of training data are used. The first one, hereon called the subsector-level, consists of aggregated contributions from all subsectors (e.g., A1300 is the subsector label for Tobacco and Tobacco Products) regardless of whether they have expressed an opinion about the bill or not. The second dataset consists of subsector-level donations with only the subsectors that give an opinion (support or oppose) on the bill under consideration. We will refer to it as the subsectors w/ opinion dataset in all plots. Lastly, the third dataset, labeled as subsectors alpha-grouped, is similar to the second dataset, except that this set includes the subsectors that give an opinion and as well as the other subsectors within the same general sector as those subsectors. For each dataset, we choose only bills for which support or opposition was not unanimous. We also filter out the bills that had almost no listed opinions in our dataset for the second and third datasets. For each bill, we choose 70% of the politicians as the training set and the other 30% as testing set. The accuracy of the classifiers is recorded for each bill, and the comparison between different classifiers and between different training sets is depicted below. All finance datasets have been normalized as a preprocessing step. Our votes dataset contains 1103 bills with public information available about which organizations supported or opposed them. After eliminating bills on which votes are unanimous, there are a total of 669 bills left for analysis.

Figure 3 shows the performance of all the classifiers on the mentioned three different datasets that we have introduced before. There are several interesting findings from the results and we discuss them separately.

a) Significance and comparison of classifiers: To begin with, the accuracy for kNN, SVM, and L_1-SVM are significantly higher than those of the fair and empirical coin tosses for all three sets, which indicates that the campaign finance and Congressional voting have a significantly strong relationship. Out of all different classifiers used, kNN exhibits the highest accuracy on average, and L_1-SVM has a slightly higher accuracy than standard SVM. L_1-SVM filters out the critical features (subsectors) in the money matrix, and is more robust in prediction.

b) Incompleteness of dataset: Regarding the three different sets, the variance for the subsectors w/ opinion only seems to be the highest of the three. This is to some extent caused by the incompleteness in the data we have. The opinions from the subsectors are obtained from different sources, such as the letters, newspapers, speeches, etc. It is likely that some expressions of opinion can be missed during the collection of the dataset. Thus, the training set is incomplete and accordingly will lead to large variance.

c) Money based prediction vs. party line: We see that the party classifier achieves the highest classification accuracy out of all the classifiers, indicating that the political party is generally a much better factor for determining how a politician will vote than any of the classification schemes based on money. We also carried out two other studies, one, to determine how good of a predictor money is on bills with a large number of people voting across party lines and two, to determine the performance of the classifiers within the two parties, thus eliminating the party factor in the model construction process.

In our first study on this issue, we find that for most bills where political party is a poor predictor, money is an equally poor predictor. There are of course a few bills where money is a far better predictor than political party, but to attribute money as the cause of these outliers would exemplify a confirmation bias. For our second study, we isolated each party and ran the classification only on bills where at least 25% of people voted against the majority of their party. The classifiers had the following prediction accuracy mean(standard deviation) for each party: for the Democrats, biased coin 0.5355(0.0973) kNN 0.6208(0.1147), SVM 0.5987(0.1005); for the Republicans, biased coin 0.5236(0.1015), kNN 0.6144(0.1082), SVM 0.5910(0.1153). After eliminating bills for which support or opposition within a party is unanimous, there were 66 bills for Democrats and 170 bills for Republicans considered in this analysis. These results indicate a significant decrease in prediction accuracy of the money-based classifiers compared to the two-party case, indicating that the relativly high accuracies of the classifiers in the two-party case were in fact dependent on political party. In all cases, the classifiers give accuracies that while not high, are better than the biased coin by a statistically significant amount.

As mentioned before, PCA showed that the political party and financial contributions had very high correlation with each other. Based on the above classification results, we think that the original problem of uncovering the relationship between how a politician votes and his/her contributions is now a problem of exploring the intertwined relationships among the former two factors but also the politician's party. In addition, we discuss some reasons why the classifiers using financial contributions have limited prediction power compared to that of the party classifier.

d) Missing direct link between donations and bills: Another reason it is very difficult to use money as a predictor for votes is that there is no direct way to infer how much money should actually be counted as influencing a bill. Generally most corporations and other organizations hold positions on a number of issues. Most of these positions are not made public. Elections are generally complicated and revolve around more than a single piece of legislation the politician must make a decision on in the future. These factors make establishing any sort of causal relationship extremely difficult. It does not logically follow that because a company supports a bill and a politician also supports that bill that the company has somehow corrupted that politician with money. In fact, many politicians actively solicit donations. It might be interesting to look at the problem from the other side and see if a politician's votes are

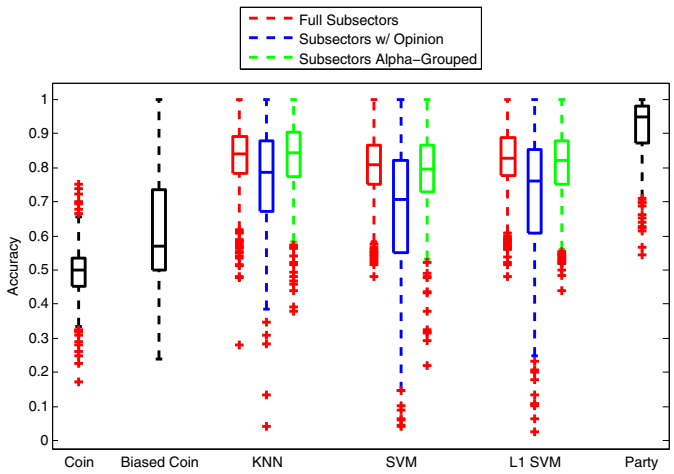

Fig. 3. Performance on the data set for various classifiers and all three datasets. The money-based classifiers significantly outperform the baseline methods of random guessing and guessing according to the empirical probabilities. The party line is more predictive than the donations. See discussion in text.

an indicator of how successful he or she is at fund raising.

e) Influence of aggregation by sector: The money from all of the different sectors/subsectors was aggregated together in our analysis to resolve the issues the sparsity of the donation matrix and our lack of substantial information about individual donors' positions on particular measures. While we feel that the grouping of the money is generally a valid strategy, it can lead to an obscuring of the varying views within a particular sector. Many large firms in some sectors actively engage in rent-seeking behavior, such as seeking new regulations that create a substantial barrier to entry into a market. Obviously, smaller firms do not generally support such measures from which they cannot benefit. Monopoly status, illegal price collusion agreements, intellectual property disputes, and other such complications within a sector can lead to problems in our analysis and cannot be captured without a higher level of granularity.

VI. Conclusions

Politician votes were predicted for a given bill based on a contribution matrix comprised of contributions from each subsector/sector to a politician. k-Nearest neighbors, standard SVM, and L_1-regularized SVM were used on testing sets to predict politician votes based on three levels of complexity: all subsectors, subsectors w/ opinion, and subsectors alpha-grouped. Using PCA, we were able to show a strong correlation between political party and financial contributions. Additional classification analysis revealed that predicting votes along party lines rather than using a classifier constructed through donation data had significantly higher accuracy. Out of the classifers using donation data, the k-nearest neighbors method had the highest classification accuracy. This is not surprising as contributions from different sectors of the industry would relate to political party, and a certain politician is likely to vote similarly to another especially along party lines. From these results, we must conclude that there is no strong evidence politicians vote solely based on the financial contributions they receive from certain industries. Rather, there is a strong correlation between money flow and political party that gets reflected in the voting process where an individual politician is very likely to vote along his/her party line.

Acknowledgments

We would like to acknowledge MapLight for providing us with their data and helping us understand it. We would also like to thank Henry Brady for discussing the politics of campaign finance and lobbyist contributions with us.

References

[1] "Maplight: Revealing money's influence on politics." [Online]. Available: http://maplight.org/

[2] "Center for responsive politics." [Online]. Available: http://www.opensecrets.org/

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, 10.1007/BF00994018. [Online]. Available: http://dx.doi.org/10.1007/BF00994018

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[6] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *ICML'04: Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 78.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.