Picking Electoral Winners
CS224W Project Report

Ashwin Siripurapu, Audrey Ho, Kai-Yuan Neo

December 2014

1 Introduction

In the political arena, one of the three major pillars of society, a democracy means that the election mechanism is the sole driving force, and yet in America, it is neither well-regulated nor well-understood. As the 2014 elections proved, there are many factors besides plain facts and swayed opinions that affect an election of a politician or the outcome of a bill's passage, to name a few. Many are aware that one of these factors is campaign donations, but few know exactly how donations play that role. Thus, we have set out to answer some of the questions surrounding campaign donations.

This paper assesses how recipient networks affect election outcomes in the United States. Our hypotheses include:

1. Candidates who receive larger amounts of donations are more likely to win
2. Candidates who receive donations from several entities are more likely to win
3. Candidates who receive donations over several cycles are more likely to win

First, we introduce prior work, showing where statistical graph techniques have been lacking in the past. Second, we explain the data collection process and data set, as well the possibilities and limitations therein. Third, we test our hypotheses through data analysis of prior donor networks and report our results. Fourth, we discuss the modeling and analytical process by which our prediction algorithm examines training and test data. Last, we present our findings on the nature of donor networks of top electoral candidates.

2 Literature Review

In the research conducted so far on the implications of campaign finance, a minority of studies include robust quantitative analysis and fewer still venture beyond basic statistical methods. Despite the lack of numerical evidence, however, the authors' typically deep understanding of social science phenomena still result in profound conclusions. Bonica, McCarty, Poole, and Rosenthal conducted a study in which they found campaign donations are increasingly coming from wealthy donors, and that organized labor is diminishing in influence. In addition, small donors were found to be much more ideologically polarized than wealthy individual donors. One exception is that wealthy high-profile donors who staffed the boardroom of a Fortune 500 company were usually more polarized, as it was to the firm's benefit to maintain balanced political access.[1] In answering the question of donor's influential power, Overton concluded using very elementary methods of numerical analysis that the political outcomes are largely influenced by which candidates are funded, which in turn are influenced by the people with monetary power.[2]

[1] Bonica, Adam, et al. "Why Hasn't Democracy Slowed Rising Inequality?" *Journal of Economic Perspectives.* 27:3. 2013

[2] Overton, Spencer A. "The Donor Class: Campaign Finance, Democracy, and Participation" *University of Pennsylvania Law Review.* 152:2004.

The general agreement is that the people with money are influential because they have the power to control the flow of politics, but we are interested in the details. We aim to fill the the enormous lack of quantitative analysis by using more robust techniques, such as graphical analysis, to determine exactly *how* influential a particular donor might be. Because we are reaching beyond basic statistics to drive our quantitative analysis, we hope to answer questions such as how much influence a dollar can buy, and whether a dollar from a higher socioeconomic class will be more influential than a dollar from donor in a lower one.

3 Data Collection

We used a data set provided by Adam Bonica, Assistant Professor of Political Science at Stanford. Professor Bonica's research centers on ideology and campaign finance. To this end, he has prepared a magisterial database, *DIME*[3], which contains campaign donation data from 1980 to 2012. User subscription is required to access the data, but an account is free for all Stanford affiliates and is trivial to set up.

This data illustrates a donor network that can be used to determine macro and micro effects and properties of campaign donations over many election cycles. Some notable facts about this network are that the largest weakly connected component contains over 99% of nodes, and that the fraction of zero in-degree nodes is also over 99%. This means that some people individually donate to multiple people, and many people donate to the same small set of people.

Prof. Bonica's database schema assigns each contributor a contributor ID (or Bonica CID, for short[4]), and each recipient (i.e., candidate) a recipient ID, or Bonica RID[5]. We principally make use of two of his database files:

1. The candidate table. This file contains one row for each candidate for elected office in each of the election cycles from 1980 to the present day. In particular, this includes the RID and (if the candidate also donated money), the CID. It also includes whether or not the candidate won that particular election. The name of this file is `candidate_cfscores_st_fed_1979_2012.csv`.

2. The contribution tables. There is one of these tables for each of the election cycles from 1980 to the present day. Each row of this table contains (among other information) the CID of the donor, the RID of the recipient, and the dollar amount of the contribution. These files are named `contribDB_{YYYY}.csv` where {`YYYY`} is an even–numbered year between 1980 and 2006, inclusive.

However, very often, the same individuals who donate money to a political cause or campaign later go on to run for elected office. To this end, we chose to unify the two ID numbers by scanning through Prof. Bonica's database of candidates. Recall that this table contains the RID and the CID of the candidate, with the CID existing only if the candidate has also given money to another candidate at some point in time. That is, if a candidate received money but never gave money, he would have an RID but no CID. In the project milestone, we dealt with this limitation by simply ignoring all nodes who received but did not donate money. The rationale behind this was that, according to Professor Bonica's research, the vast majority of people who run for office previously donated money to another political candidate. Indeed, our research bore out this finding quite well: we noticed that we were ignoring a mere 1.6% of candidates by taking this approach.

However, in our final analysis, we elected to include those candidates who had received but not given donations as well, for the sake of completeness. We did this by assigning all nodes who had an RID but no CID (since they had not contributed) a unique "artificial CID" which would be used to represent them in the graph. These artificial CIDs are included in the `rid_to_cid` mapping in the `generate_graph.py` script.

Every election cycle has its own donation graph. However, many of the same candidates persist from one election cycle to the next. Therefore, we can use each candidate's CID as a persistent ID that tracks them throughout different election cycles.

[3]http://data.stanford.edu/dime

[4]In Prof. Bonica's data files, the column name heading for a candidate's CID is `bonica.cid`

[5]The column name for the Bonica RID in the data files is `bonica.rid`

4 Network Analysis

We first analyze the donor networks across election cycles to determine whether our hypotheses hold. To get a feel of the data, we visualize donor networks in years 1980, 1990, and 2000. For each network, we only visualize a random 3000 edges (1% of the graph) due to efficiency reasons. After visualizing the donor networks, we perform analysis on each network to test our hypotheses. For each network, we find the recipients that have received the largest amount of donations and the recipients with the highest in-degree, and analyze their propensity to win elections. We then find the candidates who have received donations in the largest number of election cycles, and measure their propensity to win.

4.1 Overall Network Structure

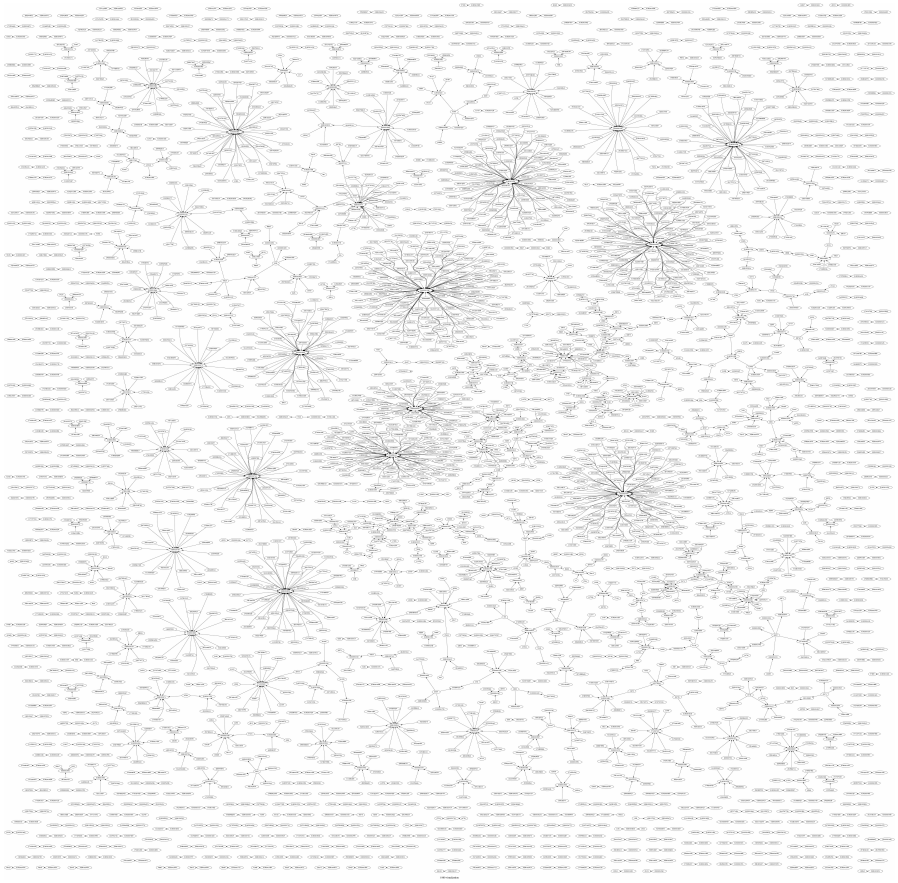

Figure 1: 1980 campaign donation graph

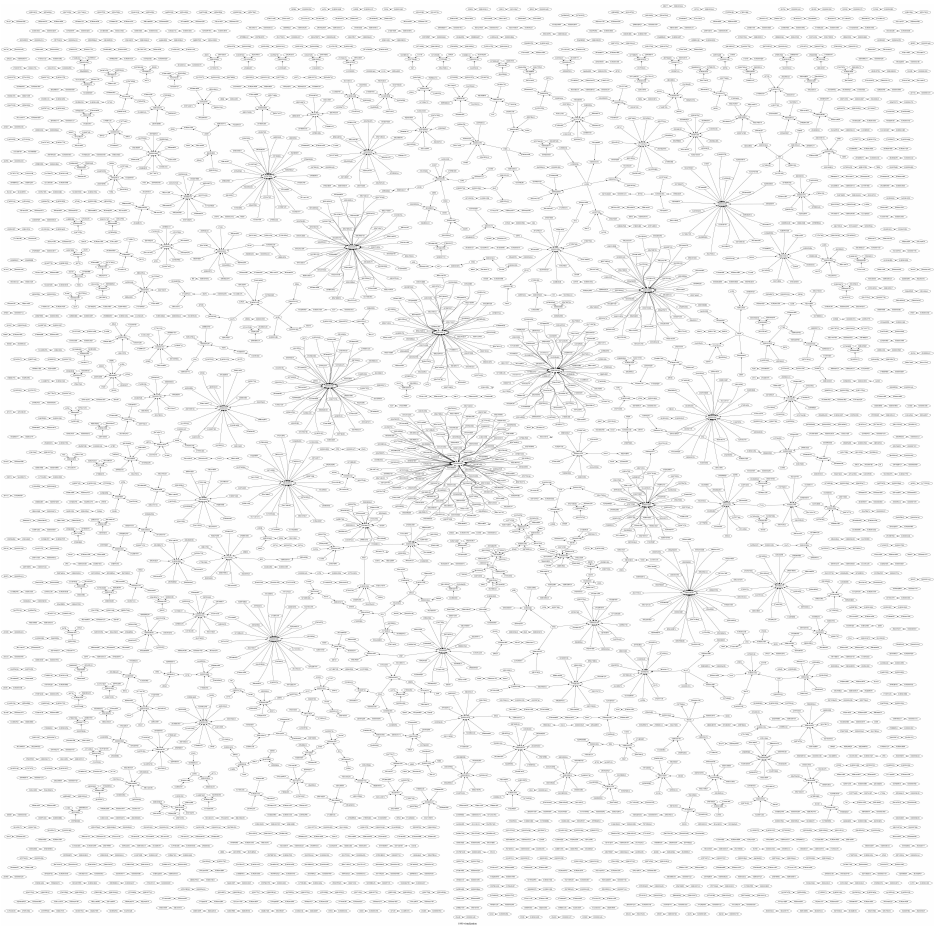

Figure 2: 1990 campaign donation graph

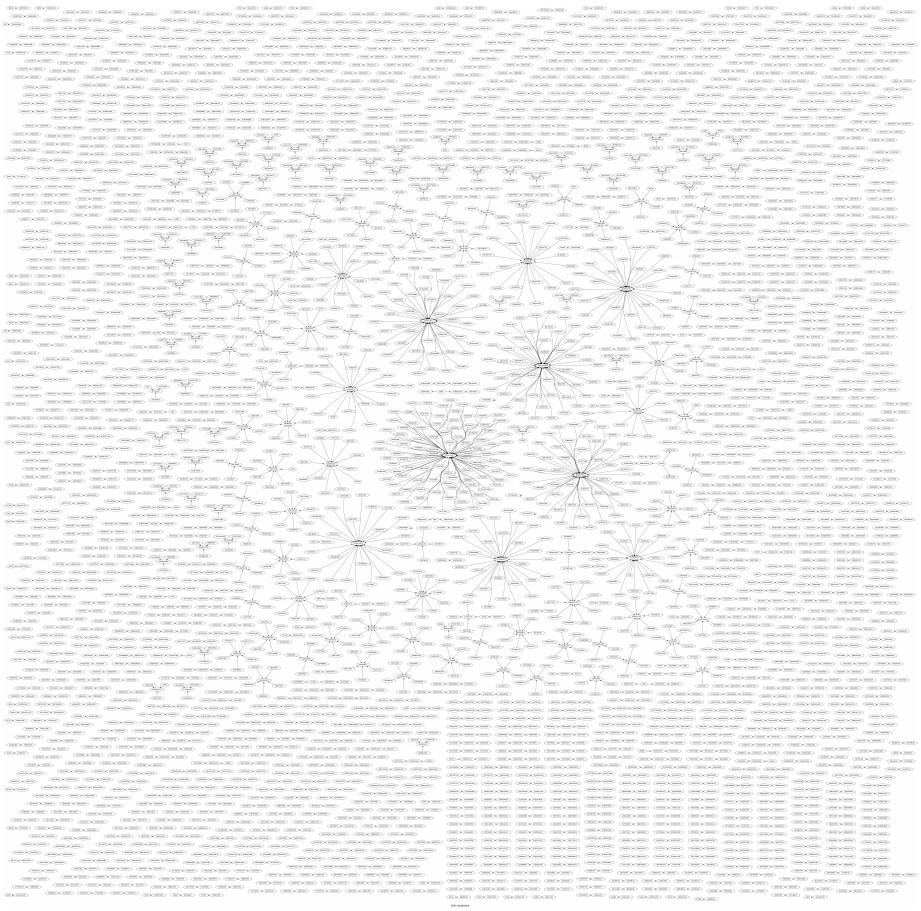

Figure 3: 2000 campaign donation graph

It appears that the donor network in 1980 contains many large local clusters, and as we move to 1990 and 2000 the networks become increasingly sparse and distributed, with many more, smaller clusters and only 1 relatively dense cluster in 2000.

One possible explanation for this phenomenon is the rise of the Internet as a political force. Whereas before, unity under the banner of a political party was essential to ensure a candidate's electoral success (in particular, to martial the resources necessary to spread a candidate's name and platform), the barrier to entry has been significantly reduced. Now, anyone with a laptop and a Twitter account can advertise his positions to the masses. This could account for the fracturing of the donation graph among several smaller recipients.

4.2 Validation of Hypotheses

1. Candidates who receive larger amounts of donations are more likely to win.
   False. Of the top 30 (.01%) of candidates ranked by amount of donations received in 1980, 1990, and 2000, 0, 1, and 2 candidates won elections respectively. This is a 6% success rate.

2. Candidates who receive donations from more entities are more likely to win.
   False. Of the top 30 (.01%) candidates ranked by number of unique contributors in 1980, 1990, and 2000, 0, 0, and 3 candidates won elections respectively. This is a 6% success rate.

3. Candidates who receive donations over more election cycles are more likely to win.
   False. Of the top 30 (.01%) candidates ranked by number of election cycles in which they received donations, 2 candidates won elections respectively. This is a 7% success rate.

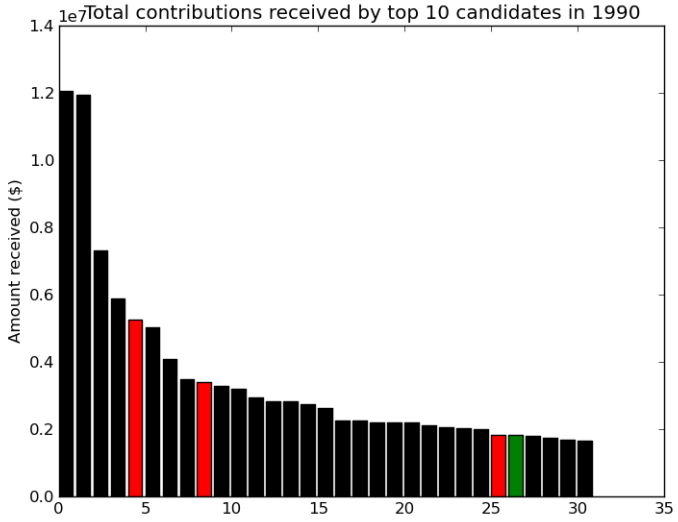

Figure 4: Top 30 total amounts received in 1990 (green = won, red = lost, black = did not run)

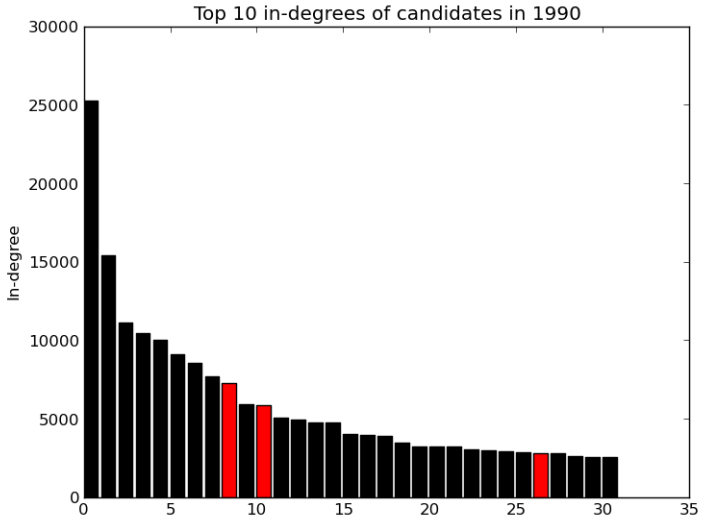

Figure 5: Top 30 indegrees in 1990 (green = won, red = lost, black = did not run)

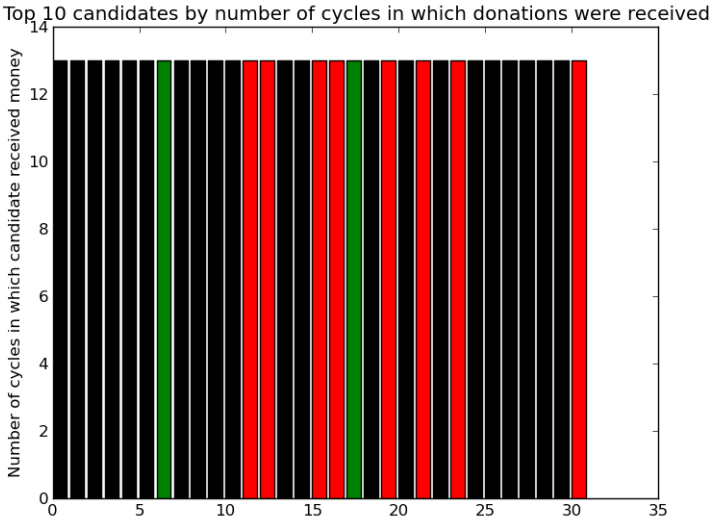

Figure 6: Top 30 values of cycles in which donations were received

5 Algorithmic Model

We determine connectivity of people by the quantity and value of their given and received donations. Based on previous work, we assume there is a correlation between amount donated and the wealth of a person, and we use wealth to be a measure of well-connectedness in the donor graph.

We are model the log–odds function of winning as being a linear function of various features of the candidate. That is to say, we are fitting an ℓ_1–regularized logistic regression model to the data. Where our analysis differs from a more classical social sciences approach is that some of the features (inputs) in our model are network analysis–style attributes of the candidates' nodes in the donation graph. Specifically, we consider for each candidate:

1. Total donations received in the current election cycle
2. Total donations given
3. PageRank score in the donation graph for the current election cycle
4. HITS score in the donation graph for the current election cycle

The motivation for using PageRank and HITS scores is that they encode information about how important a given node is. Recall that the PageRank score of a node represents the steady–state probability of being at that node during a random walk of the graph. HITS scores, too, measure a node's importance as a hub (in our case, source of money) or authority (sink of money). Degree centrality and clustering coefficient measure the node's position among its neighbors; neither of these is directly related to election prospects, nceessarily, but we hope to tease out some relationship between a node's relationship to its neighbors in the contribution graph and its probability of winning an election.

After receiving feedback on the milestone, we also decided not to evaluate candidates in a vacuum but to consider the impact of opponents as well. In particular, for each candidate, we consider the maximum of the donations received, donations given, PageRank and HITS scores over all of his opponents, as well as the sums of each of those four features over all of his opponents. Lastly, we consider the ratio of the candidate's feature values to the maximum of his opponents' feature values for each of the features enumerated above. That is, we have three classes of interaction features between a candidate and his opponents:

1. Sum of opponents' feature values (PageRank, Hubs score, Authorities score, total incoming donations, total outgoing donations)

2. Maximum of opponents' feature values

3. Ratio of candidate's feature values to maximum of opponents' feature values

The intuition behind using the sums of the opponents' feature values is that they represent the total amount of money or influence stacked against the candidate. The intuition behind using the maximum of the opponents' feature values, as well as the ratio of that maximum to the candidate's feature values, is that most of the money in a race will go to the top one or two candidates, depending on how close the race is. Therefore, it makes to examine whether the candidate is one of the front runners, and taking the maximum over all his opponents helps the model determine the candidate's relative status.

Candidates who ran uncontested would have an invalid value for the ratio of their incoming or outgoing donations to that of the best of their opponents (since they would have no opponents). We simply removed these rows from our data matrix; in practice, if a candidate is running unopposed, they will obviously win, so they are of no interest to a classification problem such as ours.

We use Python's `scikit-learn` machine learning package to fit the logistic regression model. We also considered using the `statsmodels` package, but ended up going with `scikit-learn` because of the latter's superior regularization capabilities. In particular, it was difficult to implement ℓ_1–regularized logistic regression using `statsmodels`, and even more difficult to have the regularization parameter selected via cross validation. The `scikit-learn` package, being more machine learning–oriented, made these tasks very easy. The downside is that `scikit-learn` is not as well–suited to interactive analysis and displays fewer statistical results (t–scores and p–values for fitted coefficients, for example). However, since we were not concerned so much with the classical statistical results of our model, this was not much of a problem.

6 Results

6.1 Metric

Before evaluating any model, it is necessary to define a criterion or metric for evaluation. Since this is a classification problem, it is conventional in statistics, data mining/information retrieval, and machine learning to use the F_1 score, i.e.,

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where P and R are, respectively, precision and recall, i.e.,

$$P = \frac{\#\text{truly positive test points classified as positive}}{\#\text{test points classified positive}}$$

$$R = \frac{\#\text{truly positive test points classified as positive}}{\#\text{truly positive test points}}.$$

This approach yields a score between 0 and 1 that trades of symmetrically between precision and recall. More generally, one can use the F_β score, defined as

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$

in order to get a metric that favors either precision or recall. For our purposes, however, we decided that precision and recall should be equally important, because our overarching goal is to figure out how to use network analysis concepts to influence candidates optimally, given a constrained budget. To this end, it is equally important to know which candidates will win as it is to know which candidates will lose (so that we do not spend money on losers unnecessarily).

Using this evaluation metric, the model outperformed the naive baseline (predict everyone to be a winner) by a significant margin: the model got an F_1 score of 0.65 compared to the naive baseline's score of 0.59. These are the results on an unseen test set, which consisted of candidates in the 2002 election cycle. The training set consisted of candidates and whether or not they won from each biennial election cycle from 1986 through 2000, inclusive.

6.2 Key Features

One of the nice properties of ℓ_1–regularization is that it automatically performs feature selection, setting features weights to zero when they are sufficiently small. The final feature weights output by our model (again, trained on the contributions and election results data from 1986 through 2000) is as follows:

		Feature Value
1	(Intercept)	-0.13
2	PageRank	0.0
3	Hub Score	0.0
4	Authority Score	42.9
5	Total Incoming	-3.77×10^{-7}
6	Total Outgoing	-1.75×10^{-5}
7	Max. Opp. PageRank	0.0
8	Max. Opp. Hubs	0.0
9	Max. Opp. Auths	0.0
10	Max. Opp. Incoming	-4.8×10^{-7}
11	Max. Opp. Outgoing	1.87×10^{-5}
12	PageRank Ratio	-0.13
13	Hubs Ratio	0.0
14	Auths Ratio	-0.31
15	Incoming Ratio	1.5
16	Outgoing Ratio	8.5×10^{-3}
17	Sum Opp. PageRank	0.0
18	Sum Opp. Hubs	0.0
19	Sum Opp. Auths	-0.76
20	Sum Opp. Incoming	1.82×10^{-7}
21	Sum Opp. Outgoing	-1.84×10^{-5}

As noted above, the "PageRank Ratio" refers to the ratio of the candidate's PageRank score to the greatest PageRank score of his opponents. The other "Ratio" features are defined similarly.

In analyzing these weights, it is critical to understand what optimization problem was solved to yield them; in particular, depending on the formulation of the objective function, a positive feature weight could mean that higher values of the feature are associated with higher probability of the candidate winning or it could mean that higher feature values are associated with lower candidate victory probabilities. Examining the LIBLINEAR paper [6] (whose associated software is used internally by `scikit-learn`) shows that the primal problem solved for ℓ_1–regularized logistic regression is such that higher feature weights imply higher probability of victory.

Some interesting results are apparent here. First of all, it is apparent that the two features with the highest weights are authority score and the ratio of a candidate's inflow of money to the highest of his opponents' inflows, i.e., features (4) and (15) in the above table. Intuitively, this makes sense: a node's authority score is a measure of how much it is "linked to" (donated to) by other nodes in the graph, and of course it should help to have a high quantity of donations relative to your opponents. What is more surprising

[6] 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 9(2008), 1871-1874. Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`

is that, in the presence of these features, the weight associated with the total incoming donation flow is in fact *negative*. Moreover, the features given by the maxima of the opponents' graph–related features (namely, PageRank, Hubs score, and Authorities score, corresponding to rows (7) through (9) in the above table) all were given weight 0, indicating that in the presence of the other features, these particular features were not informative to the predictive model. The intuition here is that, in the presence of the opponent–ratio features, the actual numerical value of the opponents' values of these scores is unimportant. What matters is how much *better* a given candidate is, relative to his opponents. Again, this is a fairly common–sense observation: an election is a zero–sum game, and having a high relative (rather than absolute) status, in terms of connectedness and money, is what counts.

7 Further Work

Though our hypotheses were not validated, they have also not been proven wrong. We would like to perform further studies to formally prove or disprove them. In particular, there were some issues with the data set, such that candidates such as Ronald Reagan were reported to have lost in the 1980 presidential election. After correcting for such data set errors, we would like to re-evaluate if donation amount, recipient in-degree, and duration of donations affect a candidate's election outcome.

We would also like to go deeper in our analysis of how received amount and candidate in-degree affect election outcome. Though it seems that the candidates that received the highest amount of donations from the highest number of sources did not win elections, that does not imply that candidates with the opposite properties won. We would like to find the received amount and in-degree sweet-spots, if any, at which candidates win elections.

We have yet to fully examine the effects of campaign donations from one election cycle to the next. In particular, we have not yet examined the effect of a candidate's (e.g.) PageRank or HITS score in previous years on his election prospects in the current year.

We also have yet to consider a slightly more complex weighted logistic regression model, in which training data from years closer to the present year are given more weight than training data further in the past. We would want to do this to capture the intuition that data from further in the past is less relevant to classification of data closer to the present day.

In addition, we could consider other tools from the field of time series analysis (commonly used in political science and economics) such as ARMA (autoregressive and moving average) models of graph properties and seasonality transforms.

As a motivating example, we could explore whether the campaign contribution graph becomes more densely connected or more separated into connected components during presidential election years versus during midterm election years. This would be an example of seasonality.