# Analysis of Political Networks in the US

Neeloy Banerjee, Nolan Skochdopole, Adrien Fallou

October 15, 2017

## 1 Introduction

"It is a truth universally acknowledged that a single person in possession of a good fortune, must be in want of power."

The influence of money has been a prominent issue in American politics since the 1974 'Watergate amendments' to the Federal Election Campaign Act established the Federal Election Commission. The issue took front stage again in 2010 when the *Citizens United vs. Federal Election Commission* Supreme Court ruling removed limits on the amounts corporations and organizations could spend on political advocacy. In that same year, a federal Court of Appeals ruled in *SpeechNow.org v. Federal Election Commission* that there could be no limits in the amounts that corporations, unions, associations and individuals are allowed to give to independent political committees (referred to as "Super PACs") that do not directly support candidates.

We want to understand the links and relationships between money and power in American politics through graphical analysis. We plan to construct multiple graphs representing the links made by money and the links made by power. Our idea is that voting is a proxy for power - when you vote in Congress, you are exercising your power in a certain direction. We expect that campaign finance data may provide an indication as to the direction one might vote.

## 2 Review of relevant prior work

Here we keep relevant sections from our project proposal, but we plan to integrate this section much more within the paper as we write our final draft.

### 2.1 Party Coalitions and Interest Group Networks[1]

This paper utilizes network theory to analyze the structure of political parties and more importantly the structure of interests groups within this network. Interest groups tend to support politicians with two main goals in mind: electoral (just getting a candidate elected) and legislative (influencing a politician to vote on or push legislation in a certain way). After reviewing some relevant literature they ultimately find that interest groups, when choosing to fund primary or general campaigns, tend to not stray too much from party lines, while these party lines become much more blurred when interest groups target funding for legislative purposes directly.

One of the most interesting goals of the paper was to somehow identify internal conflicts within the party coalition networks, whether this came from within the parties themselves or from the choices of funding by the interest groups. The authors only provided some basic qualitative results to this question. However, we see balance theory, which we learned about in class, as a great first tool to tackle this question and we demonstrate this later.

### 2.2 Predicting Positive and Negative Links in Online Social Networks[3] and Low Rank Modeling of Signed Networks[2]

The first paper studies three datasets drawn from online settings: Wikipedia, Slashdot and Epinions. It seeks primarily to predict positive or negative edge signs for these data sets using machine learning on graphical features. However in the process it also enables an empirical study of the social psychological theories of status and balance as applied to online networks and a study of the usefulness of negative edges. It predicts positive or negative signed edges with higher accuracy than a naive predictor; it can learn weights on one network and accurately predict edges on a different one, implying there exists underlying rules behind the formation of positive or negative edges.

The second paper naturally extends from the previous. By considering the same datasets, the authors asked whether a low rank approximation of the adjacency matrix could adequately represent the graph structure and if so, could matrix completion algorithms predict the signs of edges that currently did not exist and enable clustering to take place.

We plan to use the techniques developed in these papers to predict edge signs (and hopefully weights) in the voting graph. This is discussed in the last section.

### 2.3 Extended Structural Balance Theory for Modeling Trust in Social Networks[4]

This paper builds upon the work by Leskovec et al.[3] to improve predictions of edge signs in social networks via balance theory. Previous work in balance theory only involved the use of signed edges with no magnitude; under this regime, all edges (relationships) hold the same weight/significance. However, we may easily imagine a social network in which exist relationships of varying degrees that this simplified model does not capture. Therefore, the authors develop an extended structural balance theory (ESBT) to help model these magnitudes of relationships within a given network. They define five broad categories of relationships: stongly positive ($s+$), weakly positive ($w+$), neutral ($O$), weakly negative ($w-$) and strongly negative ($s-$). Although they do not explicitly assign numerical values to each relationship, they impose an order upon all relationships in the network and given this ordering and develop their extended balance theory from this framework.

Balance theory suggests that in any triad (group of three nodes), two edges (relationships) will influence the sign of the third relationship, working within an undirected network. If the third edge does create an unbalanced network, stress is introduced into the system. ESBT extends this theory; now, we have more choice for the third edge than just the sign, and we may define a tolerance range within the relationship ordering for which the third edge may take to stay balanced. From these relatively simple ideas of relationship orderings and tolerance ranges, the authors develop a logical framework for their extended balance theory. They argue that social networks tend towards balanced states and develop a convergence model based on the assumption that the networks converge to balanced states in as little relationship change as possible. We will analyze voting data under this framework of balance theory as well, as we may easily assign weights to the signed edges in the network.

## 3 Campaign finance data

### 3.1 Data collection

We collected campaign finance data using the Sunlight Foundation's API, which covers data from the Sunlight Foundation as well as OpenSecrets.org. One major issue with this API is that it does not get people by their unique ID, but instead searches for a person through the database before returning them, meaning the API calls are slow to return. Because of this difficulty, we wrote our data-fetching scripts so as to execute the API calls asynchronously and in parallel, and process the result in a callback. This allows us to get data for the full Congress in under 4 minutes (instead of over half an hour originally).

For each Congress member, we get the top 1000 list of contributions to their campaign for a given cycle split into 27 different contribution types, which include direct contributions to a candidate, as well as spending by outside groups for and against a candidate.

After processing the data, we get three JSON files: one for the people in Congress, one listing the contributors, and one file listing contributions for each candidate:

Listing 1: Sample Congressperson data

```
{
 "congress_numbers": [
  113
 ],
```

```
      "current": true,
      "description": "Representative for New York's 2nd congressional district",
       "firstname": "Peter",
       "id": 400219,
       "lastname": "King",
    }
```

Listing 2: Sample contributor data

```
    {
       "contributor_ext_id": "C00096156",
       "contributor_name": "Honeywell International",
       "contributor_type": "C",
    }
```

Listing 3: Sample contribution data

```
    {
       "recipient_id": 300002,
       "contributors": [
           {
               "amount": "5000.00",
               "transaction_type_description": "Contribution Made to Non-Affiliated",
               "contributor_ext_id": "C00347955"
           }
         ]
    }
```

### 3.2 Visualising the graphs

We wrote a graph visualization script using the d3.js library. Rather than using any off the shelf visualization software, the graphing capabilities of d3 gave us both great control over the specifics of the graph (force-directed allowing us to see clusters and color coding) as well as be interactive in a browser - we could choose nodes and drag them away from the main cluster to see how that affects the cluster movement. It allows interactive visualizations which bring the data analysis to life.

### 3.3 Preliminary results and discussion

We calculated some elementary numbers from our data for the 2012 cycle. The combined amount of donations to candidates for Congress for this cycle is \$1 billion, split into 383000 contributions. A further \$700 million was related to the Obama Presidential Campaign. Here is a short summary for the main contributions type, excluding the Obama data:

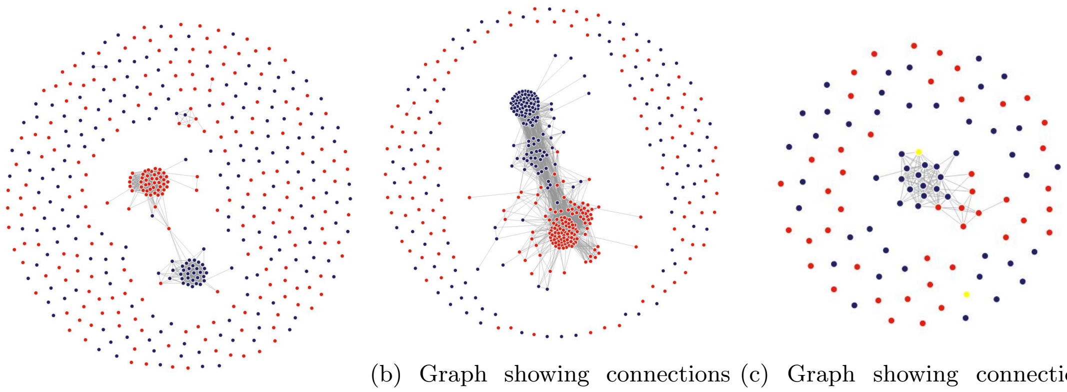
| Contribution type | Contributions count($\times 10^3$) | Total amount($\$ \times 10^6$) |
|---|---|---|
| "Contribution" | 209 | 263 |
| "Contribution Made to Non-Affiliated" | 140 | 281 |
| "Independent Expenditure For" | 2.6 | 46 |
| "Independent Expenditure Against" | 2.5 | 207 |

We notice that individual contributions, while numerous, were almost equalled in amount given by Super PACs ('Independent Expenditure') campaigning against a candidate, and we are looking to understand what impact this has on American politics.

Below we visualize three graphs. Figure 1a shows the graph built from connecting two Representatives who were campaigned against by a common organisation. There are two distinct clusters - representing Democrats (blue) and Republicans (red). This provides a high level initial justification for our method of graph creation - that we see a high level split between the parties. However this graph has many unconnected nodes and we are working on two ways to improve this. The first is to make more complete our data collection process (the text search in the API in particular is prone to errors) and the second is to include more than the 1000 top contributors for each candidate, by thresholding contributions at a certain value (eg take all contributions greater than \$3000).

Figure 1b shows a more complete picture. Here we connect politicians who either share a funding source *or* were campaigned against by a common organisation. This is both a more connected graph and a graph with a less clear boundary between the clusters of Democrats and Republicans. This too makes sense with intuition: funding *for* candidates is more likely to be position-based (eg a candidate will get funding for being pro clean energy rather than just being a Democrat) and funding *against* candidates is more likely to be done by anti-Party organisations. Thus the *against* side will naturally lead to a more distinct clustering.

Figure 1c shows the same thing as figure 1b but for Senators rather than Representatives. We hypothesize that this graph is much less connected because Senators are only voted in every 6 years, so $\frac{2}{3}$ of the Senators in the graph will not have been directly on the campaign trail during the 2012 cycle.



(a) Graph showing connections between Representatives who were campaigned against by the same source

(b) Graph showing connections between Representatives who were funded by the same source or who were campaigned against by the same source

(c) Graph showing connections between Senators who were funded by the same source or who were campaigned against by the same source

# 4 Congressional voting data

## 4.1 Collecting data and building networks

All data for congressional voting (both Senate and House) is available at `www.govtrack.us`, with data since 1989 being readily available in JSON format via API calls. Note also that we used some previously written open source scripts from `https://github.com/unitedstates/congress` to collect this data quickly.

After collecting the data, we wrote some scripts to parse the JSON files and restructure the data into a format from which we could build networks. Currently we are looking at each Congressional house separately. We build a network for either the House of Representatives or the Senate over a given time period (usually two-years corresponding to the Congressional sessions over the same time span); the people in the particular house correspond to the nodes in the network while the weighted, undirected edges represent a measure of voting similarity (or disimilarity) between the two people. Initially, we have modeled this voting similarity as a weight on edge $e = (u, v)$ where $w_e \in [-1, 1]$, $w_e = 1$ if nodes $u$ and $v$ always vote the same (in all shared votes) and $w_e = -1$ if the two nodes vote differently in all shared votesw. Therefore, $w_e$ is simply the fraction of votes on which two members agree over the total number of votes in which they both participate. Note that when we consider only two year Congressional sessions, the graphs will be complete, but if we want to consider a graph spanning many years of Congressional activity, the graphs may become more sparse.

## 4.2 Results

The signed edges of our voting networks allow for easy analysis via balance theory, as we expect, for example, if we have two people who both vote very similarly to a third person, those two people will also vote very similarly to each other. Moreover, because we have exact magnitudes for each edge weight, we have good estimations for each voting “relationship” and we can potentially take advantage of or even expand upon the previously mentioned extended balance theory that takes strengths of relationships into account beyond just the signs of those relationships. We have written functions to calculate the balance coefficients (the fraction of balanced triads over all triads) for these voting graphs. Figure 2 shows the balance coefficients using both classical balance theory and extended balance theory of the Senate for each two-year Congressional session since 1989-1990.

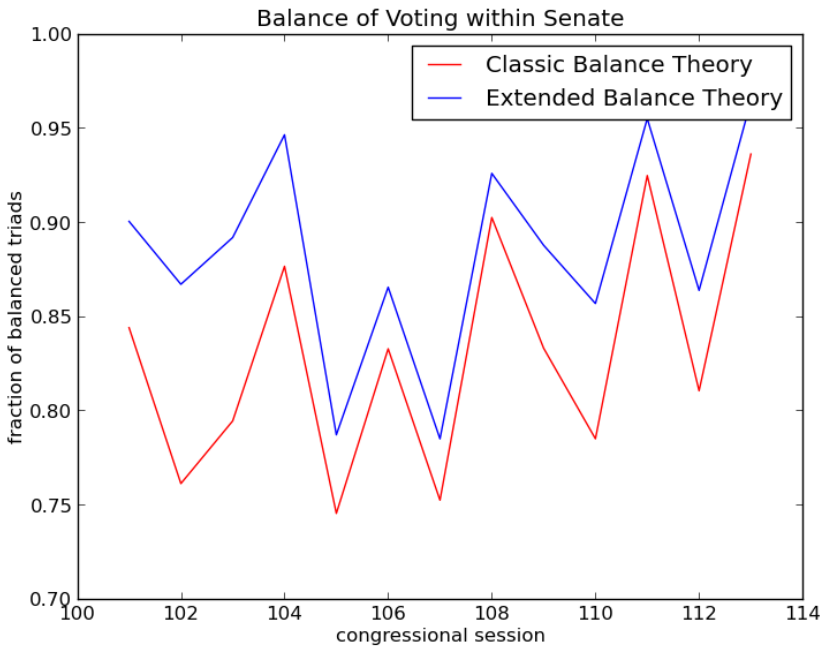
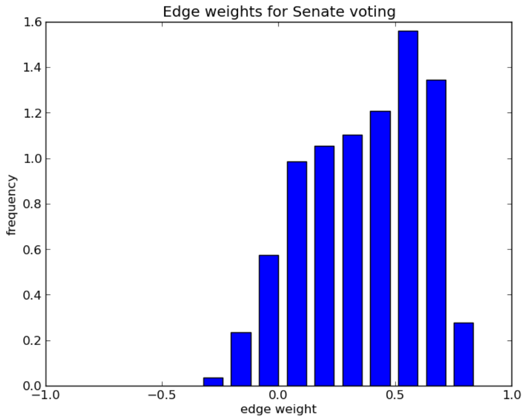


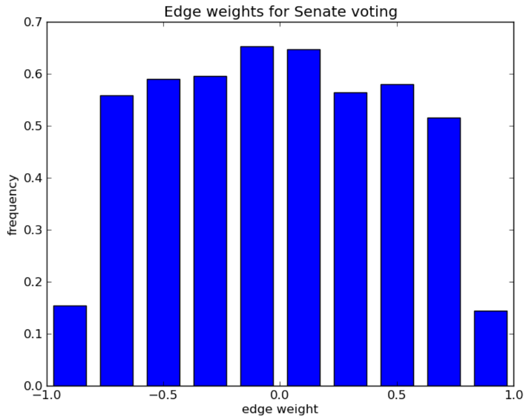

Figure 2: Fraction of balanced triads in senate voting for each congressional session since 1990 (sessions last two years). Note that for the extended balance theory, we used a threshold on strong vs. weak relationships as 0.5.

Note from Figure 2, the Senate voting graphs are very balanced, varying from 75% to 90% total number of triads balanced. However, the balance is very sporadic over time, so we would like to try to give the data some historical context (such as contested midterm elections) to aid in our analysis as well as possibly break up the data into smaller sessions to hopefully observe more meaningful evolution of balance over time. Also, we observe that when we do take the magnitudes of edges into account within extended balance theory, the number of balanced triads increased; we expect this result, because this extended balance theory creates more categories for relationships (strong negative, weak negative, neutral, weak positive and strong positive instead of just negative and positive) and relaxes some previous balance constraints by taking these relationships into account. For example, we allow triads with two weak positive relationships and a weak negative relationship (these weaker relationships have less influence on others, so they are plausible) that would not be considered balanced in classical balance theory. Currently, we have implemented a test of extended balance theory by thresholding the weights separating strong from weak voting relationships, but we still have work to do with optimizing this threshold, especially if we separately optimize the negative and positive thresholds. We also hope to use the exact edge weights to possibly even extend the extended balance theory, for example by adding more categories to the theory or using the distribution of weights directly.

We have yet to analyze the corresponding House of Representatives data under balance theory, which we plan to do as well. However, because these are almost complete graphs, all triads exist, and we must look at them all to calculate balance coefficients, giving an $O(n^3)$ process. The Senate graphs are only 100 nodes, while the House graphs are about 440 nodes, so the calculations will be at least 64 times longer for each. We could not find useful SNAP algorithms to help with the calculation of balance coefficients, just functions that counted triads without enumerating them or checking for balance. If we have time, we would like to possibly find a more efficient and optimize an algorithm for computing these balance coefficients. One idea is to use cut values. We know if the network is balanced, we can partition the nodes so that all negative edges cross these partitions; if we look at the graph with only the positive edges, we will be able to find a min cut of 0 if the network is balanced, so the min cut on this network should give some estimation of the number of unbalanced triads. Similarly, the network on negative edges should have a max cut of all the edges if it is balanced so we may be able to find an approximation for the max cut of this graph on negative edges to estimate the unbalance in the network (although this max cut problem is much harder than the min cut problem). We will explore some of these methods.




(a) Edge weights for all votes in Senate session 101 (1989-1990).




(b) Edge weights for "close" votes in Senate session 101

Figure 3: Edge weight distributions for Senate session 101

We also observed many more positive edges than negative edges within these networks. See figure 3(a) for an example histogram of edge weights for one of the Senate voting networks, which is skewed heavily towards the positive side. One probable reason for this behaviour is that meaningful votes, those that split the voters on ideologies and beliefs, may be rare compared to many procedural votes or universally liked/disliked bills. We have a method of extracting a vote category from our data, such as amendment, procedural or impeachment as a few examples. However, these categories may not be enough to find the more meaningful votes that should provide the best edge weights for the networks. One way we are considering to sift through the data is to consider votes only if they are "close", within a few votes on either side. We show a histogram for these votes in figure 3(b), which gives a much more symmetric distribution around 0 than our previous histogram does. Note that 624 votes were considered in figure 3(a) while 50 were considered for figure 3(b). Although this restriction gives us a nice distribution, we are not sure if this distribution is strictly an artifact of us choosing meaningful votes in this manner; we need to answer whether picking random votes under these conditions would give us identical distributions. We are also considering choosing edge weights slightly differently, such that we may give more weight to votes that are closer but still consider all votes. We will continue to work on these ideas to help build the best possible models for our voting networks.

# 5 Directions for further work

We propose a few different directions for this work to go and we believe we can take more than one of these directions in the coming weeks.

Our current main idea is to use the campaign finance graph to predict edge signs on the voting graph. At present we have a disconnected finance graph, and after we improve our data collection process, we still expect to have disconnected nodes remaining. We propose to have a simple indicator of Republican/Democrat as well as features derived from the finance graph initially to predict edge signs and potentially then edge weights between the politicians. First we will include features from the voting graph (ie remove one edge and predict the sign of the removed edge in the voting graph using voting graph features as well as finance graph features) and then we will attempt to remove the voting graph features. If successful, we would be able to predict macro-voting behaviour amongst politicians just from knowing their campaign finance data.

A second idea is to consider in both graphs, how far away each candidate is from their respective cluster centers. We can then look at the correlations between the two graphs and perhaps give a measure for how intertwined money and politics are, though this would present only correlation and not causation.

Another direction is to package our code into an easy to use front-end to access campaign finance data from any year. Finding and parsing that data was non-trivial (see 'Gathering the Data') and we believe easier access to campaign finance data is something the general public would love to have. This would require cleaning our code and adding a user interface as well as customizable options.

## References


[1] Grossman, D. Party coalitions and interest group networks. *American Politics Research 37*, 5 (September 2009), 767–800.

[2] Hsieh, Chiang, D. Low rank modeling of signed networks. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 507–515.

[3] Leskovec, Huttenlocher, K. Predicting positive and negative links in online social networks. *Proceedings of the 19th international conference on World wide web*, 641–650.

[4] Qian, A. Extended structural balance theory for modeling trust in networks. *Proceedings of Privacy, Security and Trust* (2013).