Measuring External Stress in Political Committee Networks

John Louie
jwlouie@stanford.edu

Vihan Lakshman
vihan@stanford.edu

1 Introduction

With the increasing abundance of temporal data, recent research in social network analysis has explored the impact of external events or "shocks" on the structure and evolution of the underlying graph. This type of analysis stands in contrast to analysis of network dynamics based on properties intrinsic to the graph itself. An understanding of how social networks are shaped by external events can provide insight into understanding networks as dynamic objects capturing complex relationships and as a new tool for modeling human behavior.

In this project, we seek to measure the impact of external stress in a network of donors to candidates and political action committees (PACs) during the 2016 election cycle. In addition, we propose several null models and associated stress disruption algorithms to develop a theoretical framework for network responses to external effects based on our observed results

The 2016 U.S. Presidential election cycle, notable for several news events that shifted polls and public opinion considerably, struck us as a powerful case study into understanding the interplay between the complex monetary relationships defining political campaigns. Given the recency and historical significance of this election, we ultimately hope that our work can provide an informative analysis of the behavior of donors during the election while, on a more abstract level, providing insight into responses to stress in empirical networks.

2 Related Work

2.1 Changes in Network Topology Under Stress

The inspiration for our work largely comes from the seminal paper of Romero, et al. in 2016, who posed the question of using temporal data to investigate how networks respond to external shocks and pursued this idea through a dataset of instant messages amongst employees at a hedge fund. To determine external stress, the authors utilized historical stock prices and defined an event as an *x-shock* if the magnitude of change in a stock price s on day d exceeded some threshold x while not exceeding the threshold on the previous three days, thereby capturing a notion of unexpectedness. Formally, a stock-day tuple (s, d) is an x-shock if $|\Delta_{s,d}| > x$ and $|\Delta_{s,d'}| \leq x \quad \text{for} \quad d' = d-1, d-2, d-3$ where $\Delta_{s,d}$ is the percent change in the opening and closing price of stock s on day d.

With this precise definition of a network shock, the authors proceeded to investigate changes to the network structure following such major external events and observed substantial changes. In particular, they noticed that, in the midst of a shock, the network would often "turtle-up" meaning that the number of correspondences between hedge-fund insiders would increase while messages between employees and outsiders would decrease. This phenomenon suggests that, in the midst of stress, hedge fund employees would form tighter connections with those in their community while abandoning outsiders. Consequently, the authors reported that the graph exhibited a higher clustering coefficient and a slight increase in the size of the largest connected component.

While Romero, et al. set the foundation for much of the recent interest in studying external influences on network dynamics, these ideas also appeared in earlier work. In 2005, Butts and Petrescu-Prahova studied the network of radio communications amongst first-responders during the attack on the World Trade Center on September 11, 2001. In particular, the authors examined how the radio network formed in response to a catastrophic emergency differed from baseline null models, ultimately concluding that the emergency network was more connected and exhibited a longer tail in its degree distribution, attributes the authors argued contributed to a successful response to the emergency.

2.2 Algorithms for Modeling Graph Disruption

In 2009, Blum et al. also considered the question of understanding network disruptions due to stress, restricting their focus to external events that prompted social networks to splinter. In particular, the authors considered three case studies in the form of real-world networks: a network of members of a karate club, connections between a terrorist group in Southeast Asia, and a graph linking board members across various environmental NGOs. In all three cases, some major event precipitated the members of the network to break up, leading to the loss of numerous edges and, in some cases, disconnecting the graph. The main contribution of the authors' work is proposing algorithms to simulate this effect of network splintering on random graphs. The driving intuition behind these algorithms is the idea that if an edge of a triad is disrupted by some external event, at least one other edge of the triangle is also likely to be removed, a situation akin to a node having to choose between two friends who are fighting. This work represents one of the earliest known approaches to developing generative algorithms to model external network stress, which heavily influenced our modeling work.

2.3 Network Analysis of Political Phenomena

There is also a rich literature of network analysis in politics. Perhaps the most influential paper in this field, igniting work on understanding political polarization in the digital age, is the work of Adamic and Glance in 2004 which analyzed links between political blogs. After identifying the 40 most popular political blogs, the authors constructed a weighted directed graph where the ith node, representing a blog, as an edge pointing into node j if i links to blog j. The weight of the edge is given by the number of links. Adamic and Glance's work is notable for quantitatively showing the political polarization within the blog network since distinct liberal and conservative communities emerged. In addition, they found that conservative blogs tend to link to each other with higher frequency than their liberal counterparts. These ideas also inspired our work, especially in examining polarization within the political donation network. Along similar lines to our work, Grossman and Dominguez in 2009 constructed a network of political interest groups with nodes representing groups and edges linking organizations that either endorsed or donated to the same candidate. Ultimately, the authors found distinct liberal and conservative communities with the liberal network more densely connected. The also sought to connect endorsements to legislative outcomes to identify the most influential nodes in the graph.

3 Data

3.1 Data Overview

For our project, we are using a combination of publicly available datasets from the FEC as well as data on election polls during the 2016 election cycle.

3.1.1 FEC Datasets

The datasets provided by the FEC cover a number of different kinds of contributions and expenditures over different election cycles, the most recent ones being 2013 - 2014, 2015 - 2016 (the cycle we're using) and 2017 - 2018. For the purposes of our project, we were interested in Committee-Committee transactions (movement of money between committees), Committee-Candidate contributions (a committee supporting a candidate financially) and Committee-Candidate links (committees that are specifically support a candidate). Additional data that is of interest are the individual contributions (individuals contributing money to political committees) and committee operational expenditures; however, use of these datasets were beyond the scope of this project. It is worth noting that the definition of a "committee" within the political sphere encapsulates "a PAC, party committee, candidate committee, or other federal committee." The transactions and contributions we are interested in can range from PAC to party, party to state committee, party to national committee, or some other transfer. Each record corresponds to a single transaction and there can be multiple transactions between two committees/candidates.

3.1.2 Candidate Polling

One crucial aspect of our work is rigorously defining an external stress event in the context of our problem. While it is easy to derive examples of events that anecdotally seemed to influence the election, we took inspiration from Romero, et al. and sought to define a shock in a mathematically precise sense. To that end, we elected to consider candidate polling values throughout the given general election of 2016. For the purposes of our analysis, we restricted our ourselves to the ABC/Washington Post general election polling numbers, which received an A+ score of quality by political analysis site 538.com.

3.2 Data Processing

3.2.1 FEC Datasets

The data itself was stored in a SQL variant; based on that format, we elected to use SQLite3 to process our data as the preliminary step to ensure it was well-formed. Because of our choice to use a SQLite3, we had to preprocess the raw text files that we obtained to escape as well as correct transaction dates. Along with the raw text files, the FEC provides the data schema which allowed us to map the data fields to the correct columns. Utilizing SQLite3, we first sought to isolate unique entities (committees, candidates, individuals, etc.) across the various datasets that we were interested in (namely the committee-committee and committee-candidate related datasets). We then mapped these entities to globally unique node IDs that we used to form an edge list to upload to SNAP for analysis. Using this mapping, we generated a preliminary directed network that we used for analysis (some of which is provided below).

Once we verified our network with basic graph metrics (e.g. clustering coefficient, etc.) we moved on to capturing snapshots of our network for analyzing network evolution. We generated a script where, given a specific date in time and a desired interval in days (e.g. ± 15 or ± 30 days), we would capture the edges formed both before and after that specific date. In effect, this gives us the delta of edges formed between nodes between two points in time. Using this, we can see how networks change over specific periods of time (e.g. what new edges are formed? how much additional money is flowing in a particular direction? etc).

3.2.2 Candidate Polling

For processing the polling data, we extracted the records corresponding to the ABC/Washington Post general election poll and isolated the fields of the records that corresponded to either Hillary Clinton or Donald Trump (we elected to omit Gary Johnson). Fields that we were interested in were the date of the poll and the percentage of individuals willing to vote for either candidate.

3.3 Graph Representations of Political Donations

Given that our data contains various different kinds of entities as well as additional features/metadata, there are various ways of interpreting political donation relationships and structure. Some of the approaches that we have used/propose for future analysis are as follows:

3.3.1 Committee-Committee Networks

For a committee-committee network, we are using the committee-committee transactions dataset. For this graph, committees are our nodes (however, note that the definition of "committee" within the political sphere can also encapsulate individuals) while the movement of money between committees form our edges. For example, a committee can give money to another committee (positive weighted edge) or that committee can get a reimbursement from that committee (negative weighted edge). For our purposes, since the transactions dataset contains multiple transactions for a given committee pair, we can consider the net movement of money (a single edge), or we can consider all the different transactions as a multigraph network.

3.3.2 Candidate-Committee Networks

For a committee-candidate network, we also use the committee-committee transactions dataset, but we also supplement it with the committee-candidate linkages (which indicate which committees are specifically in support of a certain candidate). Because the linkages are unweighted (no transaction is associated with it, the natural form of this network is an undirected, unweighted network.

3.3.3 Individual-Committee Networks

Similar to the committee-committee network, this network would be a weighted, directed (potentially multigraph) network. This would represent the movement of money from individuals to committees. Because of the nature of individual contributions (individuals don't contribute to one another, or at least the FEC would have no record of such a transaction), we can almost think of the network as a bipartite graph, if we omit committee-committee edges. Using such a graph could lead to interesting applications in understanding the role committees play in channeling money from individuals towards candidates.

4 Hypotheses

In this work, we test two primary hypotheses regarding the impact of shock events on the structure of the committee network. In the final section of the paper, we also present several null models and generative algorithms to explain the behavior of this network and better understand why our hypotheses succeeded or failed.

1. We will observe statistically significant changes in the network structure following a shock.

2. The underlying evolution process of this network can be modeled by preferential attachment or, more generally, a "rich-get-richer" framework of edge creation based on attaching new edges to nodes with high-degree and rewiring edges to high-degree nodes to simulate the emergence of winning candidates during an election cycle.

   In the subsequent sections, we will formally discuss how we define a network shock as a major change in some candidate's polling numbers. Based on this definition, we will measure important graph metrics that capture notions of network topology such as clustering coefficient, degree distribution, and average degree, and use a paired t-test to measure if these network shocks are in fact statistically significant.

5 FEC Network Analysis

5.1 Tools and Methods

For the purposes of graph analysis, we utilized both SNAP and Gephi, a graph visualization tool. While both software packages can compute graph metrics, we elected to use SNAP for the following computations.

5.2 Degree Distribution

Figures 1, 2 and 3 are the degree distribution log-log plots for various committee-committee interaction networks. We consider the degree distribution for the network consisting transactions over all of 2015 and 2016, as well as dividing the data into two halves (all transactions formed only in 2015 and all transactions formed only in 2016) to see how our network evolves over large scales of time.

Visually, it is clear that as the election drew closer, the network evolved and many more edges (transactions/contributions) were formed between committees. Indeed, an approximate comparison of the number of edges shows orders of magnitude difference between the edges in 2015 (over 160k edges) versus those in 2016 (almost 1000k new edges).

Additionally, it is worth noting that the degree distribution suggests a heavy-tailed distribution and that our network follows a power law. This observation, motivates our preferential attachment simulation, explained later.

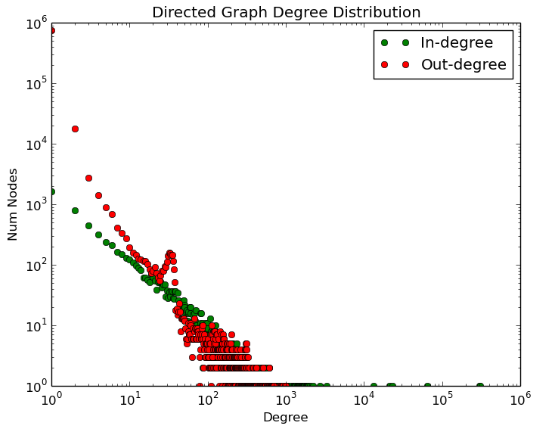

Figure 1: Degree Distribution over years 2015-2016

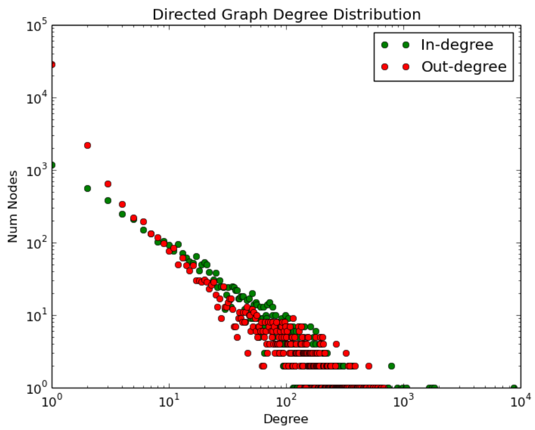

Figure 2: Degree Distribution over 2015

5.3 Clustering Coefficient

Continuing our analysis of the same three networks mentioned above, we computed the clustering coefficient for the directed network of all nodes in the years 2015 and 2016, the year 2015 (the first half) and the year 2016 (the second half).

Table 1: Clustering Coefficient over time

Network	Clustering Coefficient
Entire Network	0.0186477607853
First half of election cycle	0.0321991546445
Second half of election cycle	0.00728842091596

Similar to the visual representation of network evolution, it is clear that the graph evolves from year to year. These findings corroborate our hypothesis that network metrics change over time; in the next section, we will ascertain the significance of these changes a more granular timescale, namely on roughly a month-to-month level in accordance with consecutive polling dates.

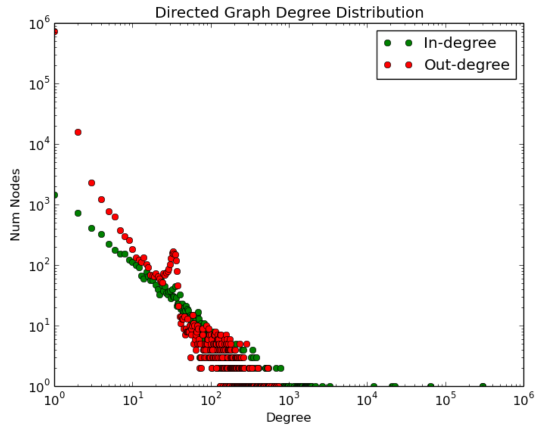

Figure 3: Degree Distribution over 2016

5.4 HITS and PageRank

Within the appendix located at the end of this paper, we list the top nodes based on centrality metrics for our given network of nodes within the years 2015 and 2016. Additionally, we've annotated our findings with their respective committee/candidate names.

5.4.1 Hubs and Authorities

Considering the committees/candidates that we have as part of the hubs and authorities list in tables 2 and 3, it seems reasonable for those particular committees to be a part of those respective lists. For example, the top hub result, InterDigital, Inc., PAC, is a PAC that crosses party lines, both contributing to Democratic and Republican candidates. By contributing to many candidates, it is a "hub" of financial contribution to the overall network. Additionally, we see that Hillary For America is the top authority, which makes sense, seeing as Hillary For America was one of the PACs with the most financial contributions of any candidate in the 2016 election cycle. Interestingly, compared to the results of PageRank (which are discussed below) the authorities list appears to have a larger proportion of Democratic PACs, which might indicate that the Democratic PACs have a greater diversity of unique contributors.

5.4.2 PageRank

For the list of top PageRank nodes in table 4, Hillary For America is again the top node; however, the following ranking nodes are almost entirely Republican, save for the DNC and DCCC committees. This result might hint at the importance, "weight" or financial power of the edges that are flowing into the listed Republican committees. In order to ascertain a more nuanced understanding of the PageRank rankings, we hope to perform weighted PageRank with the associated financial amounts contributed as part of our future work.

5.5 Other Network Properties

5.5.1 Weakly and Strongly Connected Components

Below are relative sizes of the weakly and strongly connected components:

- Relative size of WCC in Directed Graph: 0.99807436315
- Relative size of SCC in Directed Graph: 0.00613820768435

These values match our overall understanding of the structure of our network where most of the edges indicate money flowing "forward" in the network, from committees to the candidates. Having edges that point back from committees to candidates should be relatively rare and if so, most likely indicates some sort of financial reimbursement. Additionally, the fact that the SCC is non-zero also matches our understanding because committees can have cyclical financial transactions (e.g. Hillary For America and Hillary Victory Fund).

5.6 Network Visualizations

Leveraging Gephi, we were able to obtain the network visuals shown in Figures 4, 5, 6. Figure 4 corroborates our intuition of the network where the overall structure consists of "spokes" that become conjoined over time. This intuition motivates our last model described below (Algorithm 4).

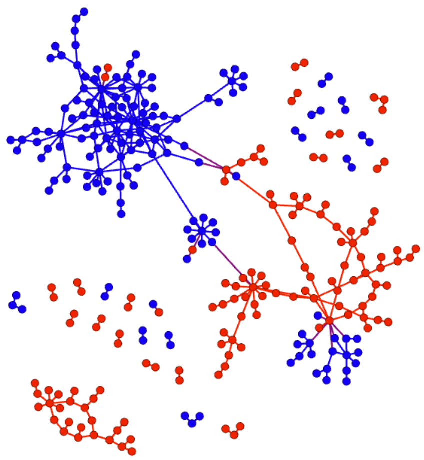

Figure 4: Close up visual of a subset of nodes

In addition to intuition about the structure of our network, we also used Gephi to understand partisanship and it's evolution in face of controversial events. Figure 5 shows the inter and intra-party committee transactions from before the release of the Access Hollywood tapes (visualization is based off of the closest poll before the event, 9/22/2016) on 10/7/2016. Figure 6 represents the network after the tape release (we elected to use the second closest poll to the release in order to allow information/media coverage on the event to propagate). The visuals show that partisan delineation evolves and even blurs around significant events. While the before visual can easily be cut according to party, the after image shows entanglement with nodes that interact with both parties.

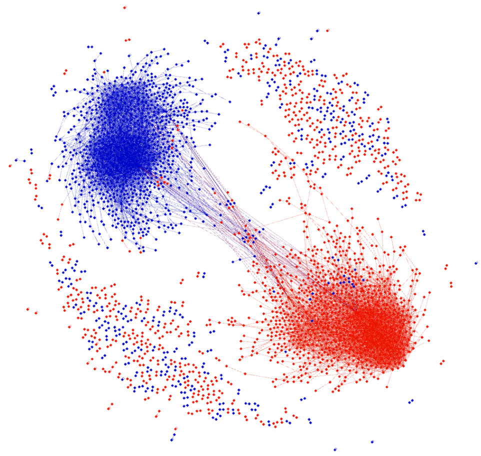

Figure 5: Democratic (Blue) and Republican (Red) committees before release of Access Hollywood Tapes

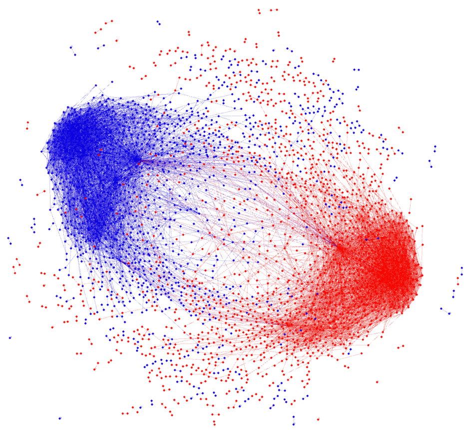

Figure 6: Democratic (Blue) and Republican (Red) committees after release of Access Hollywood Tapes

6 Measuring Shocks

As mentioned above, we examined candidate polling numbers over the course of the election cycle, from December 2015 to October 2016. In total, we constructed a dataset of 24 general election polls comparing major-party candidates Hillary Clinton and Donald Trump. We in turn defined a "shock" for our purposes as any poll which saw the percentage change in either candidate's polling numbers shift in absolute value by more than the threshold $t = 4\%$ between any two consecutive polls. Once we identified a "shock", we then constructed cumulative donor graphs for January 1st 2015 (the start of our dataset) to the end date of the poll. In order to compare the different polls, we calculated the clustering coefficient for each of our constructed cumulative donor graphs. In table 2, we provide the dates of particularly drastic poll changes for both Hillary and Trump.

Table 2: Clustering Coefficient and Poll %

Poll End Date	Clustering Coefficient	Clinton Poll	Trump Poll
3/6/2016	0.02895	47.52	42.94
5/19/2016	0.02350	43.42	48.35
6/23/2016	0.01772	49.04	42.09
9/22/2016	0.01507	46.45	45.67
10/23/2016	0.01742	48.31	39.00
10/27/2016	0.01769	45.39	45.25
11/6/2016	0.01806	45.22	41.70

With our formal definition of a shock and having identified a set of dates corresponding to such poll swings, we are in a position to test our hypothesis that the change in clustering coefficient after a shock is statistically significant. The final major technical hurdle we need to address is how to control for the natural evolution of the network over time; both new and existing donors constantly create new edges, meaning that we cannot assume as a null hypothesis that the network should remain static following a shock. Our proposed solution to this problem was to train a learning algorithm $\mathcal{A}$ that sought to predict clustering coefficient using the number of nodes and edges in the network at any given time as features. We trained this model on examples from our dataset.

This predictive model then served as our baseline for how we would expect the network to involve following a shock, which then allows us to measure whether the change we actually observed was statistically significant. After some experimentation, we chose a least-squares regression model defined by $\mathcal{A}(|V|, |E|) = \alpha_0 + \alpha_1|V| + \alpha_2|E|$ to predict the clustering coefficient of a committee network $G = (V, E)$ at any given time. We chose this model for its simplicity, ease of training, and good performance on a validation data set. Given this predictive algorithm, we constructed two arrays of data: one representing the observed change in clustering coefficient between identified shock events in the true network while the other represented the change in *predicted* clustering coefficient between the same identified dates where the predicted values were computed via our least-squares regression algorithm. This mechanism allowed us to control for natural network evolution.

Given this data, we then conducted a paired t-test for difference of means with our null hypothesis being that the means between the actual and predicted differences would be the same. If the the network changes were substantially different than predicted, then we would expect to see statistically significant changes in the mean of the sets of differences. Ultimately, we obtained a p-value of 0.64 when running this test on clustering coefficient and a p-value of 0.67 when running this test over the average degree metric. At a significance level of $\alpha = 0.05$, we therefore cannot reject our null hypothesis and conclude that we cannot reject our null hypothesis. In the last section of this paper, we will propose potential reasons why certain assumptions of our model and limitations of our dataset may have prevented us from being able to detect statistically significant changes due to shocks, providing a direction for future work.

7 Null Models and Algorithms

7.1 Algorithm Overview

In addition to analyzing the effect of external events on the political donation network, we sought to conduct some mathematical analysis and propose a theoretical model for the network evolution over time that we observed. As a baseline, we implemented perhaps the simplest network disruption model due to stress: Blum, et al.'s 'triadic disruption' algorithm on Erdös-Renyi and Preferential Attachment random graphs. In our work, we considered both randomized and deterministic versions of the disruption algorithm.

Algorithm 1 Triadic Disruption (Randomized)

1: Choose a triad (i, j, k) uniformly at random. Delete edge (i, j)
2: Flip a fair coin. If "HEADS" delete edge (j, k). If "TAILS", delete (i, k)
3: If the edge deleted in Step 2 is part of other triads, repeat Step 2 on each of these triads

Algorithm 2 Triadic Disruption (Deterministic)

1: Choose a triad (i, j, k) uniformly at random. Delete edge (i, j)
2: **if** degree(i) $<$ degree(j) **then**
   Delete (i, k)
3: **else**
   Delete (j, k)
4: **end if**
5: If the edge deleted in Step 2 is part of other triads, repeat Step 2 on each of these triads

In addition to our triadic disruption simulation, the degree distribution plots (Figures 1, 2, 3) provided us with the intuition that our network obeys a power law. Given this observation and the intuition that nodes form edges in a preferential manner (i.e. committees will most likely support/provide funds to the best-performing candidate that they agree with), we wanted to determine whether our network obeys a preferential attachment model. Algorithm 3 provides our adaption of the preferential attachment model for our given network.

Algorithm 3 Preferential Attachment (Randomized given real network start)

```
1: Define G as the network at poll time t and define G' as the network at poll time t + 1
2: for n ∈ V_{G'}, n ∉ V_G do
3:     With probability p: Select a node m ∈ V_G uniformly at random to form an edge with n
4:     With probability 1 − p: Select a node m ∈ V_G with probability proportional to m's in-degree
5:     With m, form edge (n, m)
6: end for
7: for i in range |E_{G'}| − |E_G| do
8:     Repeat steps 3 to 5
9: end for
```

The final generative model we considered was what we call the "Planted Spoke Model." The intuition for this model largely comes from our exploratory visualizations of the network. As shown in 4, the committee network is largely organized in a hub-and-spoke-like structure where several central nodes, representing candidates or influential political operatives, have in-coming edges from several other nodes. These ancillary nodes organized around a central hub only occasionally have edges amongst themselves, but the existence of these edges prevents the graph from being bipartite. Given this driving intuition that the graph is organized by new nodes attaching themselves to one of several hubs, we defined the evaluation objective of this model to replicate the change in clustering coefficient we saw over time in our network as shown in 7.1 where we see the clustering coefficient first decrease over time and then begin increase roughly three quarters of the way through the election cycle. We aimed to see if we could design a generative algorithm to mimic this process, which led us to the planted spoke model.

We initialize our model with n spoke centers and simulate the network evolution process by adding new nodes and connecting them to spoke centers based on a probability distribution given to the algorithm as input (this distribution intuitively represents the popularity of candidates). In addition, with a very small probability ϵ, these ancillary nodes will connect to each other. This comprises the first phase of our algorithm. We rigorously prove that, under this model, with very high probability the clustering coefficient will go to zero as the number of nodes approaches infinity, which allows us to capture the first portion of the graph in 7.1. The second phase of our algorithm then accounts for the subsequent increase in clustering coefficient. This phase intuitively corresponds to a select few candidates emerging as "winners" and thereby funding resources will get redirected to them. We model this process by only attaching new nodes to these winners and furthermore, with some probability, wiring existing edges to these winning hubs. For a sufficiently large rewiring probability, this results in an increase of clustering coefficient as we see in 7.1. We conclude this section by proving that the first phase of the spokes model does result in a decrease in clustering coefficient.

Proposition: As $n \to \infty$, the average clustering coefficient of the graph produced by the Planted Spokes Model goes to zero with probability $1 - o(1)$.

Proof. The clustering coefficient of the ith node is given by

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

Since edges are added independently, the degree of a node is independent of the degree of its neighbors so

$$E[C_i] = \frac{2E[e_i]}{E[k_i(k_i-1)]} = \frac{2E[e_i]}{E[k_i^2 - k_i]}$$

$$= \frac{2E[e_i]}{Var[k_i] + E[k_i]^2 - E[k_i]}$$

If the node under consideration is the central hub node, then the expected degree is np_i. Moreover, we expect the neighbors of this node to have have $\epsilon\binom{np_i}{2} = \frac{\epsilon np_i(np_i - 1)}{2}$ edges. Since the edge creation process follows a binomial distribution, the variance of the degree will be given by $np_i(1 - p_i)$ Plugging this into our definition of clustering coefficient. we find that

$$E[C_i] = \frac{\epsilon(np_i - 1)}{p_i(n-1)}$$

Moreover, we note that the expected clustering coefficient of non-central nodes will be $o(1)$ for a sufficiently small choice of ϵ. So the sum of the clustering coefficients of the non-central nodes will be $o(n)$ Therefore the average expected clustering coefficient will be

$$\frac{1}{n+1}\left(\frac{\epsilon(np_i - 1)}{p_i(n-1)} + o(n)\right)$$

which goes to zero as $n \to \infty$. Since this result holds for one spoke component, it will hold for all components by independence.

To show that this results holds with high probability, we will leverage the Azuma-Hoeffding concentration inequality defined on the *edge-exposure martingale*. Let X be a random variable denoting the clustering coefficient of the graph generated by the spoke model and let X_i be a Bernoulli random variable that takes on a value of 1 if the ith edge is present in the graph. Define the *edge-exposure* martingale $\{Z_i\}$ as $Z_i = E[X \mid X_1, \ldots, X_i]$. We note that since the martingale differences are bounded by 1 due to the fact that the addition of a single edge cannot increase the clustering coefficient by more than 1, we can apply the Azuma-Hoeffding tail bound to note that the probability that the clustering coefficient will deviate from its expectation is $o(1)$ which completes the proof. □

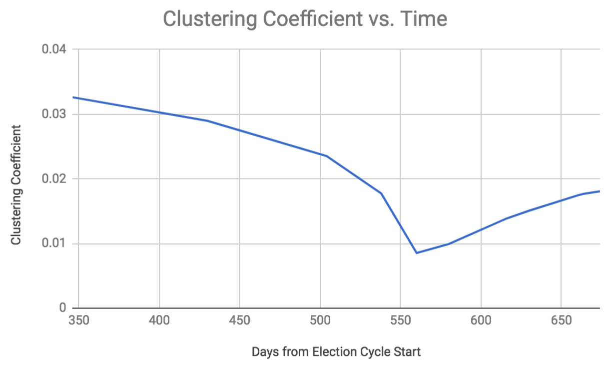

Figure 7: Clustering Coefficient over Time in Committee Network

Algorithm 4 Planted Spoke Model (Randomized)

```
Input: "spoke centers" v_1, ..., v_n, probability distribution D over these centers, probability parameter ε
Phase I:
for j = 1, 2, ... do
    Create new node u_j
    Sample spoke center v_i from distribution D. Add edge (u_j, v_i)
    For each neighbor of w_k of v_i such that w_k ≠ u_j, add edge (u_j, w_k) w.p. ε
end for
Phase II:
Update ε to ε'
for j = 1, 2, ..., do
    Create new node u_j
    Add edge from u_j to one of the top k highest degree spoke nodes v_i with probability proportional to degree of these k nodes
    Select spoke center v_k uniformly at random. Attach each neighbor of v_k to v_i with probability ε'
end for
```

7.2 Evaluation

In addition to running our triadic disruption algorithms on random graph models to measure their effects on null models, we also tested these baseline disruption simulations on our donor network. In particular, we aimed to see if the disruption cascade on the network just prior to a shock would converge to the same network structure following the shock. Such convergence would indicate that our theoretical model accurately captures the dynamics of the network's responses to external effects. To measure the convergence of our algorithms, we have so far considered the following metrics clustering coefficient and average degree

For the purposes of verifying our adapted preferential attachment model, we compared the clustering coefficients generated from our proposed model to the real clustering coefficients observed anecdotally.

For our spokes model, we again used clustering coefficient to understand how the network evolves over Phases I and II. Our main evaluation goal in this model was centered on producing a graph of clustering coefficient over time that replicated the reality we observed with an initial decreasing phase and then an increasing phase.

7.3 Experiments

7.3.1 Triadic Disruption

We tested our disruption algorithms on two random graph models: Erdös-Renyi and Preferential attachment. In addition, we ran the algorithms on the network of all campaign contributions in the 2016 election cycle just prior to March 1st (the day of the Super Tuesday primaries). Our results are summarized in the table below.

Table 3: Simulation Results

Graph	Time	Clust. Cf.	Avg. Deg.
Erdos-Renyi (Dense)	Before	0.4015458	80.0
	After	0.1751956	33.19
Pref. Attachment (Dense)	Before	0.3203727	37.9
	After	0.256756	33.81
Erdos-Renyi (Sparse)	Before	0.0001902	5.13227
	After	0.0001902	5.13227
Pref. Attachment (Sparse)	Before	0.0898319	3.97
	After	0.0898319	3.97
Committee Network	Before	0.011002704	5.132272
	After	0.011002704	5.132272

The most striking observation from the table was the role of sparsity in determining the length of cascades. We constructed both dense random graphs with relatively large average degrees as well as sparse networks with roughly the same number of nodes and edges as in the committee network. We observed cascading triadic disruption effects in the dense random graph models but not in the sparse networks since there are few triads who share edges and thus the cascade process dies out quickly. Since the real-world committee network is sparse, we can make the preliminary conclusion that the triadic disruption model is not a good choice to explain the network disruptions in our political donation graph.

7.3.2 Adapted Preferential Attachment

In Table 4, we see the performance, quantified with clustering coefficient as mentioned before, of using $p = 0.5$ and $p = 0.25$. The results show that. We see that in some cases, we are able to obtain simulated values that are quite comparable to the ground truth clustering coefficient (values are in bold). On observation, there is no discernible trend or particular characteristic about the network that dictates whether the preferential attachment algorithm will generate a reasonable approximation. This suggests that perhaps more values of p need to be tested and perhaps the value of p should evolve with increasing proximity to the ending election/event (e.g. perhaps p should decrease closer to event/election date). Additionally, other graph metrics (e.g. bipartite clustering coefficient, see "Future Work") could potentially be leveraged to understand network evolution.

Table 4: Clustering Coefficient (CC) for Real and Generated Networks using PA

Poll End Date	Ground Truth CC	Simulated ($p = 0.5$)	Simulated ($p = 0.25$)
12/13/2015	0.033	-	-
3/6/2016	0.029	0.037	0.046
5/19/2016	0.024	0.028	0.033
6/23/2016	0.018	0.024	0.025
7/14/2016	**0.0085**	0.0075	**0.0087**
8/4/2016	0.0099	0.016	0.013
9/8/2016	0.014	0.025	0.020
9/22/2016	0.015	0.022	0.020
10/23/2016	0.017	0.029	0.025
10/27/2016	**0.018**	**0.019**	**0.019**
11/6/16	0.018	0.021	0.020

7.3.3 Planted Spokes Model

To make the spokes model computationally tractable, we ran an instance of the algorithm with four central nodes (motivated by the four major primary contenders in the general election: Hillary Clinton, Bernie Sanders, Donald Trump, and Ted Cruz). We chose our probability distribution of nodes attaching to a central hub based on the polling numbers of these candidates relative to each other. We chose the probability of an ancillary spoke edge to be $\epsilon = 0.01$. Then, in the second phase of the algorithm, we picked the two highest degree nodes as our winners (intuitively Clinton and Trump) and added edges according with $\epsilon' = 0.2$ representing an aggressive re-directing of resources. The change in clustering coefficient over time is shown below.

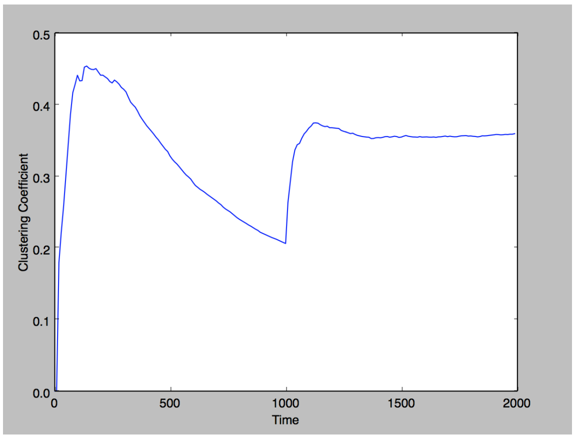

Figure 8: Clustering Coefficient over Time in Simulated Spokes Model

8 Future Work

1. **Party-specific network analysis:** While a large portion of our analysis was over the entire network and differing temporal snapshots, observing how the party-specific networks evolve might expose a more nuanced picture and more significant results. For example, the clustering coefficient metrics we used for our network shock were computed over a network containing both the Democratic and Republican parties; however, if we computed metrics for just the Democratic party for events like the DNC leak, we might see significant shifts in public opinion that might be masked by an increase in Republican clustering.

2. **Develop different network metrics:** A large portion of our analysis was motivated by use of the standard clustering coefficient metric. However, while our analysis suggests that clustering coefficient can be useful in understanding our network and potential null models, the sparsity of our network and low density of triads (as was seen with the triadic disruption simulation) means that all our calculated values are very small in magnitude. This motivates the need for a different metric, for example, the bipartite clustering coefficient, which is defined intuitively as follows: $C = \frac{\text{closed 4-paths}}{\text{all 4-paths}}$

   In addition, we hope to refine our statistical significance tests by developing more sophisticated methods for controlling for natural network evolution. One potential avenue for future work is learning how this particular network evolves based on historical election data and applying that for hypothesis testing.

   As suggested by the degree distribution plots, it would be interesting to compute the value of α as the network evolves over time; this could also be used as a potential network metric to quantify network change.

3. **Further Investigation of Node Centrality:** Another major goal of our project is to identify which committees are the most influential forces during the presidential campaigns. In this milestone, we considered basic Pagerank and HITS calculations. For future work, we hope to calculate a weighted Pagerank score using the transaction amounts as edge weights and explore other measures of centrality that could shed light on the most significant actors. We ultimately hope that this analysis can shed light on understanding the outcomes of the 2016 Presidential election.

9 Appendix

Rank	Hubs Score	Name
1	0.0025073706	INTERDIGITAL, INC., PAC
2	0.0025042011	BAKER, DONELSON, BEARMAN, CALDWELL & BERKOWITZ, PC PAC
3	0.0024857571	MCGUIREWOODS LLP
4	0.0024732734	POET PAC
5	0.0024587963	HILES, NANCY
6	0.0024549180	SCHROEDER, BRIAN
7	0.0024549180	RICHARDSON, STEVE
8	0.002454918028	AMERICAN SOCIETY FOR METABOLIC AND BARIATRIC SURGERY POLITICAL ACTION COMMITTEE, INC.
9	0.00189915028	TEXANS FOR HENRY CUELLAR CONGRESSIONAL CAMPAIGN
10	0.001886693	RICHARD E NEAL FOR CONGRESS COMMITTEE

Table 5: Top 10 Hubs (HITS) Scores for Final FEC Network

Rank	Authority Score	Name
1	0.8782566387	HILLARY FOR AMERICA
2	0.47703173686	DONALD J. TRUMP FOR PRESIDENT, INC.
3	0.02574272509	DNC SERVICES CORP./DEM. NAT'L COMMITTEE
4	0.003993509835	MINNESOTA DEMOCRATIC-FARMER-LABOR PARTY
5	0.003967880303	GEORGIA FEDERAL ELECTIONS COMMITTEE
6	0.003919832181	MAINE DEMOCRATIC PARTY
7	0.00391181508	WY DEMOCRATIC STATE CENTRAL COMMITTEE
8	0.003851354562	TEXAS DEMOCRATIC PARTY
9	0.003839277290	DEMOCRATIC PARTY OF OREGON
10	0.00379194257	MISSOURI DEMOCRATIC STATE COMMITTEE

Table 6: Top 10 Authorities (HITS) Scores for Final FEC Network

Rank	PageRank Score	Name
1	0.05979465194198752	HILLARY FOR AMERICA
2	0.05403736501253616	DONALD J. TRUMP FOR PRESIDENT, INC.
3	0.014955650898969108	REPUBLICAN NATIONAL COMMITTEE
4	0.008329295545095027	RYAN FOR CONGRESS, INC.
5	0.007550920272046351	NRCC
6	0.0056419738168701685	DNC SERVICES CORP./DEM. NAT'L COMMITTEE
7	0.005315949507569311	TEAM RYAN
8	0.004865050767399944	NRSC
9	0.004395055316056207	SCALISE FOR CONGRESS
10	0.0039604086992859455	DCCC

Table 7: Top 10 PageRank Scores for Final FEC Network

Individual Contributions:

- Vihan: Analysis tooling, simulations and network analysis, visualizations, mathematical/graph theory/proofs

- John: Data preprocessing/network creation tools, analysis tooling, simulations and network analysis, visualizations

References

Adamic, Lada A., and Natalie Glance. 2005. *The political blogosphere and the 2004 US election: divided they blog*. Proceedings of the 3rd international workshop on Link discovery. ACM.

Blum, D., Nipun Dave, and Raymond Hsu. 2009. *Mechanisms of Network Disruptions Due to Stress*. 78th Military Operations Research Society Symposium, WG. Vol. 7

Grossman, Matt, and Casey B. K. Dominguez. 2009. *Party Coalitions and Interest Group Networks. American Politics Research*. American Politics Research.

Petrescu-Prahova, Miruna, and Carter T. Butts. 2005. *Emergent coordination in the World Trade Center disaster*. Institute for mathematical behavioral sciences (2005): 1-23.

Romero, Daniel M., Brian Uzzi, and Jon Kleinberg. 2016. *Social networks under stress*. Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee.