Fuzzy Spectral Hierarchical Communities in Evolving Political Contribution Networks

Scott Wahl, John Sheppard
Gianforte School of Computing
Montana State University, Bozeman, MT

Abstract
A valuable tool for analyzing social networks is partitioning the complex graphs based on dense sub-networks, usually referred to as communities. This partitioning can discover groups who have similar attributes or behaviors. Using a previously developed method for creating hierarchical fuzzy communities using spectral clustering, these communities are evaluated in a temporal network and tracked through time to evaluate community change. This method is tested on a real world political network based on campaign finance contributions, and it is shown how communities change over time at multiple levels of the hierarchy.

Introduction
There are numerous real world social networks that can be partitioned into dense subnetworks. Referred to as communities, these subnetworks should contain elements that have properties in common with one another. Considerable work has been performed in developing methods for finding communities in social networks. Much of the early work focused on splitting the nodes into distinct and separate communities (Blondel et al. 2008), (Newman 2006), (Newman and Girvan 2004), (Pons and Latapy 2004). Due to its performance on more complex clusters, spectral clustering has proven popular (Ng, Jordan, and Weiss 2001), (Pothen, Simon, and Liou 1990). A limitation with those early approaches is that they do not reflect individuals belonging to multiple communities. An additional issue is that many networks contain hierarchical structure where communities combine to form larger groups at different levels. More recent approaches attempt to handle this by allowing fuzzy clusters as well as creating a hierarchical structure for the communities (Bandyopadhyay 2005), (Devillez, Billaudel, and Lecolier 2002), (Liu 2010), (Palla et al. 2005), (Torra 2005), (Xie, Szymanski, and Liu 2011).

There are a number of real world sources for networks that exhibit community structure. Prior research has yielded such structure in genetics (Zhang and Horvath 2005), neuroscience (Power et al. 2011), and Internet communities (Flake et al. 2002) as a small sample. The focus of this paper is on networks created by political donations. Of particular note, there has been some prior research focusing on political social networks. Much of this research is on social interaction and its effect on political participation (Aldrich et al. 2015), (La Due Lake and Huckfeldt 1998), (Quintelier, Stolle, and Harell 2012). Other work focuses on elitism and the behavior of corporations in politics (Mizruchi 1989). Specifically related to campaign finance, the geography of donations has been shown to be an indicator useful for predicting donations (Gimpel, Lee, and Kaminski 2006).

In order to effectively test the community results of campaign finance data, a useful comparison tool is required. An estimate of political ideology, called the campaign finance score (CFScore), is used within this paper to validate the discovered communities since political ideology has been shown to be a significant factor in political outcomes in a variety of topics. Prior research has studied how candidates fall ideologically when compared to national parties (Stephen Ansolabehere 2001). Ideology also impacts the legislative process (Jenkins 2006). Recent work has been performed in estimating ideology directly from campaign finance datasets (Bonica 2014). This CFScore has the benefit of being applied to a broader spectrum of political entities, including donors. The idea behind this approach is that an entity would prefer to donate to a candidate or group who shares similar ideology. Through that model, CFScores can be calculated and were shown to provide similar ideological measures to prior work, from which accurate voting records for legislators could be predicted. By comparing this metric with community assignments through time, new insight can possibly be gained in the donation patterns within the campaign finance dataset. Such analysis can help identify groups of candidates with similar donors that otherwise may not stand out. Additionally, analyzing such data through time and in different states can hopefully provide new tools for assessing the impact of campaign finance legislation.

Using the prior work for estimating ideologies, we use a hierarchical fuzzy spectral clustering scheme previously developed by us (Anonymous and Anonymous 2015) on the political contribution networks in order to find communities within campaign finance. These communities are validated against community metrics as well as their relationship with ideological scores. New to this work, the communities are analyzed through time with respect to their ideology measures.
Background

There has been a lot of interest in detecting communities in social networks, and a wide variety of techniques have been proposed. In analyzing these partitions, one very popular function for determining the quality of any partitioning of nodes is modularity (Newman and Girvan 2004). The general idea behind this measure is to compare the fraction of links that connect any nodes in a community, $C_i$, to any other community, $C_j$. This ratio of edges is compared against a null model. This null model is a graph where each individual node maintains the same degree, but each edge is reassigned randomly. For a community partitioning to be considered a good partition, the fraction of links within a community should be higher than the fraction of links leaving the community. However, it should be noted that simply putting all nodes into a single community would satisfy this constraint, so more must be done.

For this, first define a $k \times k$ matrix $E$ where $k$ represents the number of communities. Within this matrix, the element $E_{ij}$ represents the fraction of the edges that connect any node in community $C_i$ to a node in community $C_j$. Determining the value of a partitioning of the graph into communities relies on the trace of the previously defined symmetric matrix $\text{Tr}(E) = \sum E_{ii}$. This diagonal represents the fraction of all edges that connect any node in community $C_i$ to any other node within itself. Furthermore, we define the value $a_i = \sum_j E_{ij}$, which gives the ratio of edges within the graph that connect to all the vertices within $C_i$. Using all of this, the modularity is then given by

$$Q = \sum_i \left( E_{ii} - a_i^2 \right) = \text{Tr}(E) - \|E^2\|.$$  

This can be written alternatively as

$$Q = \frac{1}{2m} \sum_{i,j \in V} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{c(i),c(j)}$$

where $m$ is the number of edges in $A$ and $\delta_{c(i),c(j)}$ is 1 when $i$ and $j$ are in the same community and 0 otherwise.

In order to validate the communities discovered with fuzzy spectral clustering, a generalization of modularity is used that was created independently by Nepusz et al. (2008) and Shen et al. (2009).

$$Q = \frac{1}{2m} \sum_{i,j \in V} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] s_{ij}$$

where $s_{ij} = \sum_{c \in C} \alpha_{ic} \alpha_{jc}$ and $\alpha_{ic}$ is the fuzzy community assignment of $i$ to community $c$.

Spectral Characterization

One of the limitations of creating a fuzzy hierarchical community assignment for large networks is storage space. As each node has a fuzzy community assignment for each community, as the number of communities grows, the storage space requirements increase exponentially. Due to this, spectral characterization is used to limit the size of the resulting hierarchy. Since the adjacency matrix for the defined social networks have no negative entries, the matrix satisfies the requirements for the Perron-Frobenius theorem, indicating the largest magnitude eigenvalue of the matrix will be real and positive.

As regards social networks and communities, what is important are the properties of the eigenvalues with respect to the number of communities. From other work, it has been shown that community structure in a network has a certain effect on the eigenvalues. More specifically, it has been found that a network with $k$ communities will have $k$ large eigenvalues (Chauhan, Girvan, and Ott 2009), (Sarkar and Dong 2011), (Sarkar, Henderson, and Robinson 2013). To illustrate this, consider a network consisting of 1000 nodes where edges in the network are added randomly between any pair of nodes $n_i$ and $n_j$, with probability $p = 0.04$. With this construction, the network as a whole can be considered to be its own community since there are no special defining characteristics separating any of the nodes. Calculating the eigenvalues of this network reveals a single eigenvalue outside the main cluster of smaller eigenvalues.

As predicted by prior work, the largest eigenvalue here is approximately the average degree of the nodes in $A$ (Farkas et al. 2001). Because of the random construction of the test network, this is approximately the product of the probability of connection between nodes and the number of nodes, $\lambda_{\text{max}} \approx n \times p$, or in this case, $\lambda_{\text{max}} \approx 1000 \times 0.04 = 40$. There is additional work on Erdős-Renyi uncorrelated random graphs showing the edge of the large cluster of eigenvalues is approximately defined by $s \sqrt{n}$ where $s$ is the standard deviation of the values $A_{ij}$. However, some analysis on the real-world campaign finance networks shows that using that as a threshold to determine the $k$ large eigenvalues for those networks, and thus the number of communities, can yield poor results.

Similar principles apply to networks with community structure. Assuming $k$ communities in a random network, two nodes are connected with some probability $p$ if they belong to the same community. Otherwise, the two nodes are connected with probability $q$ where $q < p$. Prior work shows that there are eigenvalues corresponding to $s(p-q)$ where $s$ is the size of the community. Consider another network of 1000 nodes created with four communities of equal size, $p = 0.1$, and $q = 0.01$. For this network, there are four large eigenvalues, three of which are approximately $250 \times (0.1-0.01) = 22.5$. These principles form the basis for estimating the number of communities among the contributors and candidates. Based on these results, the gap in eigenvalues is used to determine an appropriate maximum number of communities.

Datasets and Approach

The political contribution network data used here is provided by Bonica and Stanford’s Social Science Data Collection (Bonica 2013), (Bonica 2014) \footnote{URL: http://data.stanford.edu/dime/}. This dataset combines data from the Federal Election Commission, the Center for Responsive Politics, the National Institute on Money in State Politics, as well as other reporting agencies and the Sunlight...
Data Analysis

For any dataset $D$ and time periods $t$ and $t + 1$, the retention rate of entities is defined as the percentage of entities that are in both sets: $ret(D_t, D_{t+1}) = \frac{|D_t \cap D_{t+1}|}{|D_t \cup D_{t+1}|}$. Using the individual state-year data sets where those who only gave once are removed, the average retention rate of entities in those networks is 20.90% with a median of 21.66%. There is one notable outlier with Alabama where the rate is markedly lower than the other states at 6.52%.

Estimating the Number of Communities

As shown earlier, it is possible to estimate the number of communities in a social network using characteristics of its eigenvalues. To do so requires finding the eigenvalues considered to be large. Unfortunately, directly calculating the predicted edge of the primary cloud of eigenvalues based on the principles of Erdős-Renyı uncorrelated random graphs can yield poor performance in determining the number of clusters. In prior work, a sample of contributors to candidates in Alaska for the 2012 elections showed splitting the network into four communities provided logical groups of connections when compared to donors. Obtaining communities centered around each individual candidate does not provide much in the way of useful information, however.

Algorithm

Our approach used for finding fuzzy clusters is based on the spectral clustering work of Ng, Jordan, and Weiss (Ng, Jordan, and Weiss 2001) and Zhang, Wang, and Zhang (Zhang, Wang, and Zhang 2007). After determining the cutoff for the number of clusters, fuzzy spectral clustering is performed on network $A$ for cluster numbers $K = 2, 3, \ldots, k-1, k$ as follows:

- Let $D$ be a diagonal matrix where $D_{i,i}$ is the sum of the $i$-th row of $A$. This is equivalent to the weighted degree of each node.
- Construct the Laplacian matrix $L = D^{-1/2}AD^{-1/2}$.
- Determine the $k$ largest eigenvectors, $x_1, x_2, \ldots, x_k$ of the Laplacian $L$ and create the matrix $X = (x_1, x_2, \ldots, x_k)$. $X$ is then normalized such that each row has unit length.
- Using $X$, perform fuzzy $c$-means clustering on the data to obtain $U$, a $n \times k$ matrix where $k$ is the number of clusters and $n$ is the number of data points in $A$.

To obtain hierarchical structure, the process is repeated with a varying $k$ corresponding to the number of clusters in each hierarchical level. Each level is connected to its previous level by calculating the fuzzy Jaccard similarity of the communities given by

$$sim(C_1, C_2) = \sum_{i \in C_1 \cap C_2} \min \left( \frac{C_{1,i}}{C_{1,i}}, \frac{C_{2,i}}{C_{2,i}} \right).$$

Connections between two time steps are made in a similar fashion. Beginning with $k_i = 2$ for time step $i$, connections are made to time step $i + 1$ by comparing the communities
at the same level across the two hierarchies. To do so, for each community in data set \( i \) at \( k_0 = 2 \), iterate through the communities of data set \( i+1 \) at \( k_1 = 2 \) to find the best match based on the fuzzy Jaccard similarity.

### Results

The results presented below focus on three different states: Alaska, New York, and Wisconsin. Given Alaska’s high retention rate across years, it provides many opportunities for analyzing how behavior of specific individuals change over the years. While this paper focuses on these three states, all 50 states have been analyzed with the same procedure. For most of those datasets, splitting into two communities at the top level has very high correlation. However, in some state and years, splitting into two communities does not result in a high correlation. This is because there exists a group within the dataset that is more separate from the rest of the network than those with opposing ideologies. New York is one such state. Wisconsin was also selected because of the rapid growth in the data set due to the increase in contributions surrounding recall and gubernatorial races.

### Alaska

For the state databases, Alaska showed the highest retention rate of entities from year to year at 29.67% on average. First, communities are found for the entirety of the state, regardless of the year in which a donation was made. For this dataset, eliminating all single donors and redundant links across all years in the state gives 12,417 entities and 66,629 edges.

At the top level, it is easy to check the communities against CF Score for validity. CF Scores represent a range centered on zero where negative values are associated with liberal ideology and positive with conservative. For Alaska, at the top level, comparing with the CF Score estimation of ideology, the community assignment values show a Pearson correlation coefficient of \( \rho = 0.9133 \). Restricting the comparison to just the recipients in Alaska, the correlation coefficient for this limited set is \( \rho = 0.8715 \). This indicates that, for Alaska, the CF Score ideology estimation is highly correlated with the community assignments.

To make sure the resulting communities still represent ideology well after being split into individual 2-year cycles, a similar test is performed on the temporal datasets for Alaska. As before, checking for two communities results in splits where the fuzzy community assignment is highly correlated with the CF Score for that entity. Table 1 shows the correlations for each of the cycles for all entities. For Alaska, these fuzzy memberships are highly correlated with the CF Score.

Additionally, it is possible to connect the communities in one time step to communities in the next based on the best fuzzy Jaccard similarity. Table 2 shows the average CF Score of entities for each community with a membership value greater than 0.3. As shown, the averages shift fairly consistently away from zero for both of these communities. The results of this correspond to prior political science work indicating an increase in partisanship over the years.

Moving down the hierarchy, similar results are obtained for \( k = 3 \), also shown in Table 2. For all but one year, every community at \( t_i \) continued into \( t_{i+1} \). At this breakdown, the average CF Score of \( C_1 \) and \( C_3 \) does not deviate from zero as in the previous breakdown, despite having similarly high Jaccard similarity measures as the communities in \( k = 2 \). Additionally, the average estimated ideology of \( C_3 \) shifts considerably more than the other two. Viewing additional data about the recipients in this group, community \( C_2 \) corresponds to a specific geographic area, Fairbanks, AK.

### Wisconsin

For Wisconsin, creating the network of contributions across all years as before results in one that contained 123,396 nodes and 592,407 edges. Part of the reason for this network being larger are the circumstances surrounded the 2012 recall and regular elections. As in Alaska, the correlation coefficient for CF Score and fuzzy community assignment at the top level hierarchy is quite high at \( \rho = 0.9745 \) for all entities and \( \rho = 0.9408 \) for recipients.

Wisconsin also shows high correlation at the top hierarchy when comparing community assignments and CF Score, shown in Table 1. As can be seen, when the upswing in donations occurred in 2012, the correlation between ideology and communities is exceptionally high. This seems reasonable given the apparent polarizing nature of the elections.

Analyzing these communities over time yields similar results to that in Alaska. At the top level hierarchy, there are two communities corresponding to left and right ideologies, shown in Table 3. Additionally, as time passes, the overall trend is for both communities to deviate from the center, corresponding with the increase in partisanship.

With \( k = 3 \), the resulting communities look similar again to AK as shown in Table 3. However, community \( C_2 \) in this case does not appear to be isolated to a single geographic area, but has recipients from districts all over the state. Given the overall average CF Score, WI appears to have a considerable, and consistent, set of moderates.

### Table 1: Correlation of CF Score and Communities at \( k = 2 \)

<table>
<thead>
<tr>
<th>Cycle</th>
<th>AK</th>
<th>WI</th>
<th>NY</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>0.898</td>
<td>0.920</td>
<td>0.121</td>
</tr>
<tr>
<td>2006</td>
<td>0.916</td>
<td>0.946</td>
<td>0.614</td>
</tr>
<tr>
<td>2008</td>
<td>0.913</td>
<td>0.923</td>
<td>0.017</td>
</tr>
<tr>
<td>2010</td>
<td>0.923</td>
<td>0.943</td>
<td>0.121</td>
</tr>
<tr>
<td>2012</td>
<td>0.906</td>
<td>0.981</td>
<td>0.058</td>
</tr>
</tbody>
</table>

### Table 2: CF Score of Alaska Communities

\[
\begin{array}{cccccc}
\text{Year} & \text{\( k = 2 \)} & & \text{\( k = 3 \)} & & \\
& C_1 & C_2 & C_1 & C_2 & C_3 \\
2004 & -0.839 & 0.287 & -0.924 & – & 0.387 \\
2006 & -0.851 & 0.346 & -0.847 & -0.790 & 0.363 \\
2008 & -0.885 & 0.334 & -0.899 & -0.318 & 0.293 \\
2010 & -0.895 & 0.373 & -0.913 & -0.492 & 0.391 \\
2012 & -0.872 & 0.391 & -0.897 & -0.223 & 0.357 \\
\end{array}
\]
In order to highlight different behavior of donors in different states, New York was also analyzed in a similar manner. As before, communities were found for the entirety of the state, regardless of the year in which a donation was made. This resulted in a network of 69,369 entities and 264,223 edges. Unlike Alaska, when splitting the network into two communities, the resulting fuzzy assignment values do not have a high Pearson correlation coefficient when compared with the CFScore. This is even true if the same analysis is performed with weighted edges where the weights correspond to the amount of the contributions to an entity. Calculating the correlation coefficient for all entities within New York at the top hierarchy gives a value of $\rho = 0.4451$. For just the recipients within NY, $\rho = 0.2921$. As seen, CFScore is not as well correlated with the communities.

In an attempt to better understand the composition of the communities at the top level, we first look at a strict partitioning of the two top communities where the fuzzy community assignment value must be greater than 0.5. Analyzing the candidate information within these communities, it shows all of the New York city candidates are within $C_2$. While not composed solely of city candidates, the dominating factor for this breakdown appears to be geography and not ideology.

This poor performance on correlation continues when looking at each individual year, as shown in Table 1. Interestingly, the two worst performers 2008 and 2012, have almost no data for New York City candidates. This seems counter intuitive since those years would not have extra data highly centralized in a single geographic location.

Across all years but 2006, viewing the communities at the top hierarchy, what has happened is that a small set of candidates were separated from the rest of the network by having similar donor groups, but being considerably different than the rest of the candidates. In years with local elections reported, the general trend is for this smaller group to include more Republicans and third-party candidates. In the other years, the party distribution is nearly even, very slightly favoring Democrats (58%). For both 2004 and 2012, this smaller community had only a single winning candidate between them and nearly equal numbers of Democrats and Republicans. This indicates that, for New York, ideology is not actually the most dominant factor in determining the pattern of donations. This holds true even when adjusting the weights of the network based on the size of the contributions. This shows community detection methods can provide additional insight into the donation patterns.

### Discussion

Using the described algorithm to separate the network into communities manages to group individuals into logical groups that can be tracked through time and provide additional insight into the area of focus. Communities at the top of the hierarchy arising from campaign finance have ideological estimates that move away from the center as time progresses, matching known partisanship phenomena. However, further breakdowns in the data can yield communities that are not deviating from the center, as shown in Wisconsin. Additional information can be drawn from the results at a smaller view. As an example, for Alaska, analyzing the individuals who change communities at the highest level highlights how their donating behavior changes from year to year. For many of those who have moderate CFScores, it can be seen that the CFScores of the candidates to whom they donate may vary considerably, but of which the average CFScore is moderate. In individual years, the fuzzy clustering scheme highlights how they may donate primarily to a single ideology in a single year. As an example, consider one of the entities present in Alaska 2004 and 2006 (Bonica ID 52297646020). The CFScore for this entity is -0.54. Viewing the target party and CFScore of this entity’s donations shows most of the targets are Democrats and have a lower CFScore. This holds true in 2004 where the donor is solidly in the lower CFScore community. In 2006, however, even though the entity donated primarily to Democrats, only one of those Democrats was not moderate. The others had near zero CFScores. Adding the Republican recipients to that total results in the donor being primarily in the Republican community. Comparing to the full dataset for Alaska, that same entity is mostly within the low CFScore community, but also a small assignment within the high CFScore group, which follows closely with its ideological estimate.

### Conclusion

As shown, fuzzy hierarchical spectral clustering is able to group entities logically within political contribution networks. By splitting networks into two communities, the resulting groupings closely follow previous estimates of ideology. Splitting the network into more communities highlights differing patterns of donations beyond the ideological scores, both within a state and in different election cycles. By using those prior estimates and the communities, it is possible to analyze entities who either shift ideologies over time as well as view groups who differ in their type of donations beyond ideology.

Future work includes analyzing how replacing simple connections within the database with weighted connections based on the donation amount will affect the communities. Additional work will make use of the temporal data sets to attempt to perform predictions in a couple of ways: lawmaker votes, and donations. Much more work can be performed in researching the various communities and their trends in order to gain more insight into the political process. Furthermore, similar analysis for other countries could provide useful comparisons if it is possible to obtain relevant datasets.
References


Bonica, A. 2013. Database on ideology, money in politics, and elections: Public version 1.0 [computer file].


