# HIERARCHICAL FUZZY SPECTRAL CLUSTERING IN CAMPAIGN FINANCE SOCIAL NETWORKS

by

Scott Allen Wahl

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2021

# DEDICATION

*To my wife Annie, and my parents, Neil and Kristine, whose support has been immeasurable.*

# ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. John Sheppard. He saw promise in me and encouraged me to pursue an advanced degree. He has at times shown immense patience and encouragement when I struggled, but also unfailing in pushing me to better my research. I would also like to thank the members of my committee, Dr. John Paxton, Dr. Binhai Zhu, and Dr. Elizabeth Shanahan whose advice has aided in navigating my research as well as the difficulties of balancing life. I also thank the rest of the staff in the Gianforte School of Computing, especially Jeannette Radcliffe who was always eager to help anyone in need. I thank all the members of the Numerical Intelligent Systems Laboratory. The research discussions, feedback, and encouragement have been invaluable.

I thank my wife, Annie, for encouraging me along this journey. She supported me when I doubted myself, helped through late nights, and shouldered extra burden while I worked through this research. I also thank my parents, Neil and Kristine Wahl. Their dedication and character are traits to which I aspire. They have always supported me in my adventures in more ways than I can possibly enumerate.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

# LIST OF TABLES

LIST OF TABLES – CONTINUED

LIST OF TABLES – CONTINUED

Table Page

LIST OF TABLES – CONTINUED

Table Page

# LIST OF FIGURES

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

# LIST OF ALGORITHMS

# ABSTRACT

Community detection in networks is an important tool in understanding complex systems. Finding these communities in complex real-world systems is important in many disciplines, such as computer science, sociology, biology, and others. In this research, we develop an algorithm for performing hierarchical fuzzy spectral clustering. The clustering algorithm is applied to small benchmark problems, as well as a large real-world campaign finance network. Afterwards, we extend the hierarchical fuzzy spectral clustering for use in evolving networks. The discovered communities are tracked through the evolving network and their underlying properties analyzed. Third, we apply association rule mining on community-based partitions of the data. A comparison of the results within and between communities show the effectiveness of this method for adding interpretability to the underlying system. Fourth, we examine the ability of hierarchical fuzzy spectral clustering on a graph to predict behavior that is not present in the graph itself. The results are shown to be effective in predicting votes in the United States legislature based on the campaign finance networks. Finally, we develop an orthogonal spectral autoencoder that is used to perform graph embedding. This approximation model avoids the eigenvector decomposition of the full network, as well as allows out-of-sample spectral clustering. The results show the embedding performs comparably to the full spectral clustering.

# CHAPTER ONE

# INTRODUCTION

Within this chapter we present motivation and background for research in community detection and analysis. Community detection is a method of graph analysis used to interpret interactions or relationships in a more complex system. The simplification obtained by the communities can be used to improve interpretability, create predictions, or apply reasoning over the underlying system that created the graph. We develop novel algorithms to find hierarchical and overlapping communities within networks. Our analysis of communities focuses on social networks in political campaign finance. The following sections include a summary of relevant social networks and clustering. Afterwards we summarize the major contributions of this research. The chapter concludes with an overview of the remaining chapters.

## 1.1 Motivation

Complex networks, or graphs, are a large and growing area of important research. Community detection is an important method for simplifying complex systems that can be difficult to analyze as a whole. Large scale graphs such as social networks are one area where this analysis can be applied to provide interpretability, discover patterns in behavior, or find missing relationships. Social networks are a type of graph that are generated from interactions among members of some population. These networks have shown up in many different fields such as genetics [1], neuroscience [2], collaboration networks [3], Internet groups [4], animal social behavior [5], and many

more.

At their core, networks or graphs are made of points or *vertices* that are connected by *edges*. In many real-world networks, these edges are not distributed evenly throughout, but instead there are groups of vertices that tend to have more connections among themselves than connections to the rest of the network [6]. These groups are commonly referred to as communities. Community analysis on these networks can be useful in identifying behaviors or structure within the underlying system that created the graph.

Consider a network made by links between web sites. Web sites that share topics will likely have a higher proportion of links between themselves than to other websites that do not cover similar topics. Community analysis on such a network could aid in categorizing those web sites. As another example, correspondence or associations among friends can be interpreted as a network. Community detection can be used to identify shared interest or be involved in recommendations for pairing individuals who are not directly connected. Clustering the networks into subsets of vertices is a common way to perform community detection. Clustering should result in sets of objects that share properties.

There are many examples where improvements to systems can be made by finding communities. Expanding on the web site example from earlier, clustering can improve the performance of web site access. Latency can be improved by keeping related domains within the same server [7]. Related stores or products can form a network where recommendation systems can increase revenue by showing related items to consumers [8]. It can also provide a way to classify the vertices, which has applications in genetics and metabolic networks [9].

Community detection in social networks has been a focus of considerable research. Finding the communities alone can provide useful information regarding the

vertices within their sets. Vertices that are centrally located in their clusters provide control and stability within the group. Those vertices on the boundary between communities can provide mediation or points of exchange between the different communities [10, 11]. Thus, it is important to not only identify the communities, but also the relationship of each vertex to the community at large.

However, much of the early research focused on communities where vertices could only belong a single set or community. This results in what is called a crisp community assignment [6, 12, 13, 14]. A popular algorithm for community discovery is spectral clustering since it is relatively easy to implement and it can find non-convex clusters [15, 16]. However, there is a limitation with some prior approaches since they do not allow for vertices in the network to belong to multiple communities. This is an issue as the individuals within a social network often belong to more than one community at a time.

Another limitation in some methods is that they do not account for sub-groups within a community. Such sub-groups can consist of smaller groups of individuals within a larger community, thus forming a hierarchy of communities. Military, business, familial, and political hierarchies are all examples of hierarchies where individual smaller groups combine to create a larger group. There are more recent approaches that attempt to improve on the older algorithms by allowing fuzzy clusters as well as creating a hierarchical structure for the communities [17, 18, 19, 20, 21, 22].

## 1.2 Contributions

This thesis details contributions to the development of community detection. These contributions involve the creation of new algorithms for performing clustering, generalization of the communities, and adding interpretability to the clustering. The specific contributions are listed below.

- We introduce an algorithm for performing hierarchical fuzzy spectral clustering [23]. We use outlier detection on the spectral characterization to find the number of communities and hierarchical structure in real-world networks. Results from the experiments show the utility of both the overlap and hierarchy of clusters. We test the hypothesis that overlapping communities in a campaign finance setting can provide insight into the behavior of the individuals within the community. Results show the individuals between communities exhibit different behavior than those entirely within a single community.

- Hierarchical fuzzy spectral clustering is adapted for use in community detection in evolving networks [24]. Clustering is done on snapshots of evolving networks through time. A fuzzy Jaccard similarity is used to identify matching communities for the adjacent snapshots. This method is applied to the campaign finance network. The community assignments are evaluated against known ideological estimates as validation of the communities. Results of tracking the communities through time are shown for a selection of state campaign finance networks.

- We use association rule mining to enhance the interpretability of the communities found using hierarchical fuzzy spectral clustering [25]. We use the community assignments from hierarchical fuzzy spectral clustering to create overlapping partitions of the underlying transactions that created the social network. Association rule mining was applied to the partitions to find the frequent patterns in each of the partitions. We use the intersection of the rule sets among sibling communities to find shared properties of the communities. Those rules that are not in the intersection of the siblings are useful in determining discriminative properties of the communities. We apply this method to the campaign finance data to illustrate its ability to provide

interpretation of the community assignments.

- We show the effectiveness of hierarchical fuzzy spectral clustering in predicting behavior of community members [26]. This deviates from the prior analysis which focused on communities and data that created the social network directly. Instead, we use the community assignments for the individuals by adding them as features to a different set of data that involves those individuals. The results show the generalizability of the community assignments in that it is effective in predicting behavior that was not directly represented in the social network. We applied the hierarchical fuzzy community assignment results obtained from the campaign finance networks and use them to predict the voting behavior of individuals within the United States legislature.

- We developed a novel algorithm for graph embedding referred to as Orthogonal Spectral Autoencoder (OSAE). By using mini-batch sampling and an approximate Laplacian, this model approximates the results of the spectral embedding that is required during hierarchical fuzzy spectral clustering. The model also provides benefit in allowing for simple out-of-sample extensions for clustering new data without performing additional spectral decompositions. The model is validated against the large campaign finance networks, and the results are compared with the clustering of the exact spectral decomposition.

## 1.3 Overview

In this section we provide an overview of the following chapters in this dissertation.

Chapter 2 provides a review of some prior research that is important to the topic of community detection and clustering. First, we provide definitions of graphs

and their components. This is followed by a description of various existing clustering techniques used to find communities within networks. We also discuss metrics for evaluating the clusters. Following these, we provide a summary of the political networks used in the later experiments. We also discuss the reasoning behind the selection of political networks as an application of community detection in graphs.

Chapter 3 includes more review of methods to find overlapping communities in networks. Following this background, we introduce our approach for hierarchical fuzzy spectral clustering. This approach iteratively adds eigenvectors to fuzzy c-means, creating a hierarchy of additional communities. These communities are attached to their parents via a fuzzy Jaccard similarity metric. The effectiveness of this algorithm is shown on two small benchmark datasets in addition to a real-world campaign finance network.

Chapter 4 augments hierarchical fuzzy spectral clustering. This procedure works by first finding communities at each individual time step of the evolving network. After this community discovery, the links between the time steps are added by a similarity metric based on fuzzy set similarity of the adjacent time steps. We analyze the performance of tracking the communities through time on multiple state campaign finance, detailing how individuals and communities change behavior over time.

Chapter 5 adds interpretability to the community detection by applying association rule mining to partitioned data. As discovered in analyzing the communities through time, providing useful semantics to the communities can be difficult. We automate this procedure by partitioning the underlying dataset based on the fuzzy community assignments provided by hierarchical fuzzy spectral clustering. The results on rules found in a state campaign finance network show the automatic rule finding is beneficial in interpreting the community structure.

Chapter 6 considers the generalizability of the communities from HFSC. As-

suming the communities capture patterns of behavior and ideologies of the campaign finance networks, we combine the community assignments with a dataset containing the history of Yea and Nay votes in the United States legislature over a period of 12 different snapshots. The results show the community assignments generalize the behavior of the legislators and is effective at predicting votes.

Chapter 7 introduces a novel graph embedding structure to resolve two issues inherent to the spectral decomposition step necessary for HFSC. The first of these issues is that spectral decomposition can be costly and does not scale well to large datasets. Second, projecting new data points into the spectral domain is not straightforward. The usual method for clustering new data points is to redo the spectral decomposition on the addition data. Instead, the graph embedding introduced uses an approximate Laplacian to limit the size of the matrix used in spectral decomposition. In the process, the auto encoder model used to perform the graph embedding naturally extends to out-of-sample data so that new information can be efficiently clustered without additional spectral decompositions. While the approximate graph embedding does not perform quite as well as the full spectral decomposition, the results are comparable.

Chapter 8 is a summary of the results and contributions. We also give some direction and areas for future work.

# CHAPTER TWO

# BACKGROUND

In this chapter we give a review of terminology and existing methods for community detection in networks. This review provides many definitions necessary for later discussion. The existing methods cover a variety of philosophies in performing community detection. This includes graph partitioning, clustering, and spectral decomposition methods. Following the discussion of those algorithms, we provide motivation for using campaign finance networks as the real-world network. In addition to the practical application, the networks created in political social networks highlight the benefits of community detection, interpreting found communities, and the generalization of the community assignments to a task not strictly defined by the graph.

## 2.1 Networks

Define a network, or graph, as $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ where $\mathbf{V}$ is a set of vertices, or nodes, in the graph, and a vertex in the graph is $v \in \mathbf{V}$. $\mathbf{E}$ is a set of edges connecting the vertices of the graph. Edge $e_{ij}$ defines a connection between vertices $v_i$ and $v_j$ where $e_{ij} \in \mathbf{E}$ and $v_i, v_j \in \mathbf{V}$. For the purposes of this work, edges $e_{ij} \in \mathbf{E}$ are considered to be undirected and some measure of affinity between vertices $v_i$ and $v_j$ is associated with each edge. The degree $d_i$ of vertex $v_i$ is the sum of the edges connected to that node, i.e.

$$d_i = \sum_{j=1}^{n} e_{ij}.$$

Most real-world networks have been found to have useful structural properties that help inform or assist in learning useful knowledge concerning the nodes or entities

within the network. Such real-world graphs usually are not regularly structured like a lattice, i.e., the connections between vertices in the graph are not evenly distributed throughout the graph. Instead, there are substructures within the graph where vertices in that group have a higher proportion of edges with other members of that group than with members of other groups. These substructures are commonly called communities. We define a community $\mathbf{C}_i$ as a subset of vertices, or $\mathbf{C}_i \subset \mathbf{V}$.

Using a simple definition of community structure, nodes within a community should have a high proportion of in-group edges compared with the rest of the network. This value would be at a maximum when there is an edge between every pair of vertices in the community, forming a set of vertices that are completely connected. A completely connected subset of nodes in a network, $e_{ij} \in \mathbf{E}\, \forall v_i, v_j \in \mathbf{C}_i$, is called a clique.

We consider properties of random graphs to determine in what ways real-world graphs may differ. In the random graphs proposed by Erdös and Rényi, the probability of an edge occurring between any pair of vertices is equal for all pairs [27]. With that uniform distribution in the edges, the probability of edge $e_{ij}$ being in $\mathbf{E}$ is the same for all possible edges connecting vertices in $\mathbf{V}$. Any vertex is equally likely to be connected to any other. The degree of any vertex in this graph should be based on the total number of edges in the graph and the number of nodes. The expected degree $d_i$ of vertex $v_i$ is given by

$$E\left[d_i\right] = \frac{|\mathbf{E}|}{|\mathbf{V}|}.$$

The in-group connectivity $\mathcal{C}_{\text{in}}$ of community $\mathbf{C}_i$ is defined as the number of edges between members of that community divided by the number of members in that

community, i.e.,

$$\mathcal{C}_{\text{in}}\left(\mathbf{C}_i\right) = \frac{\left|\{e_{ij} : v_i, v_j \in \mathbf{C}_i\}\right|}{\left|\mathbf{C}_i\right|}.$$

Consider a community $\mathbf{C}_i$ from a random network. In such a community, any edge between vertices $v_j, v_k \in \mathbf{C}_i$ would be equally likely to exist as an edge between vertices $v_j \in \mathbf{C}_i$ and $v_l \notin \mathbf{C}_i$ due to the construction of the network. Therefore, the expected in-group connectivity of the community would be the same as the network at large.

$$E\left[\mathcal{C}_{\text{in}}\left(\mathbf{C}_i\right)\right] = E\left[\mathcal{C}_{\text{in}}\left(\mathbf{V}\right)\right]$$

This would be true for any possible community in the network. However, this is not true for real-world graphs.

Instead, in real-world graphs the distribution of degree for each node often exhibit a tail corresponding to a power law, i.e., most vertices have a small degree while some have much larger degree [28]. As one example, campaign finance networks have been shown to follow a preferential attachment model. In preferential attachment, the probability of an edge being added between nodes is proportional to the existing degree.

In contrast to random networks, real-world networks generally have subnetworks that exhibit community structure. These communities are generally comprised of groups of vertices that have elements in common with each other. Examples of networks that have community structure can be drawn from social [11], biological [2], gene expression [1], and many other types of networks [29]. Since the communities can represent fundamental properties of the network, their discovery is important for understanding the nature of the networks [4, 6].

Community structure in graph is in part shown by a transitive property that is present in real-world networks. In contrast to the random networks, real-world

vertices are more likely to have an edge between them if there is a third vertex to which they are both connected. As an example, assume vertices $v_i$, $v_j$, $v_k$ and edges $e_{ik}$ and $e_{jk}$ exist in a graph $\mathbf{G}$. Then it is more likely that a connection would form between vertices $v_i$ and $v_j$ with edge $e_{ij}$.

One way community information is described is with a local clustering coefficient [30]. The authors defined a clustering coefficient $\mathcal{C}_{\text{cf}}$ as the average of the ratio of edges within a neighborhood over the maximum number of edges the neighborhood could have. In this case, the local neighborhood $\mathbf{N}_i$ of $v_i$ is $\mathbf{N}_i = \{v_j : e_{ij} \in \mathbf{E}\}$. Suppose vertex $v_i$ has $k$ neighbors. If this vertex and its neighbors formed a clique, then that vertex and its neighbors can have at most be $\frac{k(k-1)}{2}$ edges between them. The ratio of edges that actually exist compared to the allowed edges for the neighborhood of $v_i$ is

$$\mathcal{C}_v\left(v_i\right) = \frac{2\left|e_{ij} : v_i, v_j \in \mathbf{N}_i, e_{ij} \in \mathbf{E}\right|}{k\left(k-1\right)}.$$

This defines how close to a clique the neighborhood of $v_i$ forms. The authors defined the global clustering coefficient as the average of $\mathcal{C}_{\text{cf}}\left(v_i\right)$ over all vertices in the graph, i.e., $\mathcal{C}_{\text{g}} = \frac{1}{n}\sum_{v_i \in \mathbf{V}} \mathcal{C}_{\text{cf}}$ where $n$ is the number of vertices in the graph. A network with a high average clustering coefficient indicates high connectivity of members of a neighborhood and that it exhibits community structure.

## 2.2 Clustering

With some background and definitions of networks and communities, the task becomes discovering communities within a graph. There is already a wealth of research on finding communities. Some initial work focused on crisp partition of the network into non-overlapping, non-hierarchical communities [6, 13]. In this domain, a crisp partition is one where a vertex can only be a member of a single

community. Thus, if $v_i \in \mathbf{C}_j$, then $v_i \notin C_k \forall \mathbf{C}_k : k \neq j$. A variety of approaches have been developed for finding communities in networks [12, 14]. A very popular method is spectral clustering [15, 31]. These have proved popular for their ease of implementation and their ability to handle non-convex clusters.

To best represent the communities, a classification of the vertices into clusters should satisfy two important realities of many social networks: cluster overlap and hierarchy. For the first, nodes within the network may belong to multiple communities. Much like in human social groups, an individual may belong to more than one community or have multiple affiliations [32]. Hierarchy is another important component of some social networks wherein smaller communities together make up larger ones. Military, business, and political hierarchies are all examples of hierarchies where individual smaller groups combine to create a larger group.

One early method for finding overlapping clusters comes from computational complexity work done by Karp [33]. Consider a subset of nodes $v_1, \ldots, v_k$ where edge $e_{ij} \in \mathbf{E}$ for all pairs of nodes $v_1, \ldots, v_k$. These nodes are completely connected, or a clique. The minimum clique cover problem is one of Karp's original NP-complete problems. A solution to this problem determines the minimum set of cliques necessary to cover all nodes in graph $\mathbf{G}$. Since a clique can be considered a community, and these sets of cliques may result overlap, such a solution would result in overlapping communities. This solution would not inherently provide a degree of membership to each community as well as have limitations on the kinds of communities it would find. It is even NP-hard to approximate this problem to a specified tolerance [34].

To find fuzzy communities, a variety of approaches have been presented. Palla uses a clique percolation method to find adjacent cliques with overlapping nodes [20]. Other methods use fuzzy modularity and simulated annealing or other techniques to find relevant partitions [17, 19, 22]. Fuzzy c-means is another possibility for

determining fuzzy clusters and has been used to find hierarchies of clusters [18, 21]. The approach presented here differs in its use of spectral clustering and spectral characterization to create a top-down algorithm for finding hierarchical fuzzy clusters.

In addition to the previously mentioned methods, there has also been work in social networks which change over time and methods for tracking and predicting communities [35, 36, 37]. One such method attempts to predict the emergence of future communities using link prediction methods [38]. The following sections provide more detail on some of these methods.

### 2.2.1 Modularity

Modularity is a metric that is commonly used to evaluate the communities discovered by those algorithms. The general idea behind this measure is to compare the fraction of links that connect any nodes in a community, $\mathbf{C}_i$ to any other community, $\mathbf{C}_j$. This ratio of edges is compared against a null model. This null model is a hypothetical graph where each individual node maintains the same degree, but each edge is reassigned randomly.

For this, first define $\mathbf{A}$ as the adjacency matrix for a graph where $a_{i,j} = 1$ if there is an edge between nodes $i$ and $j$. Assume there exists a community partitioning $\mathbf{C}$ that represents the set of communities and $\mathbf{C}_i$ contains the nodes belonging to community $i$. Let $k = |\mathbf{C}|$ and define $\mathbf{E}$ as a $k \times k$ matrix that contains data regarding interconnections between communities. Specifically, for any two communities $\mathbf{C}_i \in \mathbf{C}$ and $\mathbf{C}_j \in \mathbf{C}$, element $e_{ij}$ represents the fraction of edges that connect nodes in $c_i$ to $c_j$. This leaves the diagonal as the ratio of edges that connect nodes within the community to other nodes in that same community. From this representation, determining the value of a partitioning of the graph into communities relies on the trace of $\mathbf{E}$, namely $\mathrm{Tr}\,(\mathbf{E}) = \sum_i \mathbf{e}_{ii}$. Furthermore, we define the value $d_i = \sum_j e_{ij}$,

which gives the ratio of edges within the graph that connect to all the vertices within $C_i$. Then the modularity is then given by

$$Q = \sum_i \left(e_{ii} - d_i^2\right) = \mathrm{Tr}\left(\mathbf{E}\right) - \left|\left|\mathbf{E}^2\right|\right|. \tag{2.1}$$

This can alternatively be written using the adjacency matrix for the social network $\mathbf{A}$ directly as

$$Q = \frac{1}{2m} \sum_{i,j \in V} \left[a_{ij} - \frac{d_i \cdot d_j}{2m}\right] \delta_{C_i,C_j} \tag{2.2}$$

where $m$ is the number of edges in $\mathbf{A}$, $d_i$ is the degree of node $i$, and $\delta_{C_i,C_j}$ is 1 when $i$ and $j$ are in the same community and 0 otherwise. For crisp communities, these values help in determining the quality of the found community assignments.

Using modularity as a tool to perform clustering has a few issues. It has been shown that maximizing modularity is NP-hard [39]. Additionally, there is a resolution limit where it is not possible to evaluate communities with modularity if the community structure is weak and it can be lost in the noise of the network [40]. Other work has extended modularity to create metrics that work with fuzzy communities [41].

### 2.2.2 Divisive Clustering

One of the earliest popular methods for finding communities involved removing edges from a network based on measures of edge centrality [6, 42]. This metric estimates the importance of an edge relative to the graph. The general procedure for this divisive method on graph $\mathbf{G}$:

1. Calculate a measure of betweenness for all edges $e_{ij} \in \mathbf{E}$.

2. Remove the edge with the highest score.

3. Reevaluate the betweenness scores for the remaining edges.

4. Repeat.

This process successively removes edges from the network, creating disconnected components as they are removed. Another relevant method is spectral partitioning. This was a precursor to spectral clustering where we calculate the random walk Laplacian of a graph $\mathbf{L}_{\text{rw}} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is a diagonal matrix where $d_{ii} = \sum_j a_{ij}$. Since all rows and columns of this matrix sum to zero, the eigenvector $\mathbf{z} = [1, 1, \ldots]$ always exists and corresponds to the zero eigenvalue. In the event there are disconnected components of graph $\mathbf{G}$, then the Laplacian matrix will be block diagonal and there will be multiple zero eigenvalues. However, if the network is connected, the second eigenvalue $\lambda_2$ will correspond with how good the split of the network will be in spectral bisection. The bisection itself works by noting the non-degenerate eigenvalues of the Laplacian are orthogonal, and thus have positive and negative components. The sign of the component can be used to assign a community to each of the nodes that correspond to that value in the vector. This method is a hierarchical clustering scheme if the procedure is repeated for the subgraphs containing the communities. Hierarchical clustering schemes are frequently used when there is not a clear number of communities known prior to performing clustering. Using a measure of dissimilarity, sets of nodes are split into smaller communities.

### 2.2.3 Agglomerative Clustering

Agglomerative clustering is one of the general categories of hierarchical clustering. In agglomerative clustering, sets of smaller communities are iteratively joined in a tree-like structure. The tree-like nature of this clustering is often represented by a dendrogram. One early example of this type of network used modularity as a metric for grouping nodes [43]. Using the definition of modularity $Q = \sum_i (e_{ii} - a_i^2)$, the

task becomes identifying the current communities to determine which ones should be merged to into a single community. The authors create a sparse matrix $\mathbf{\Delta Q}_{ij}$ for every pair of communities $\mathbf{C}_i$ and $\mathbf{C}_j$ where there is at least one edge between the communities.

The general procedure works by calculating the initial values of $\mathbf{\Delta Q}_{ij}$ and $a_i = \frac{k_i}{2m}$ for all $i$. A max heap is populated with the largest element of each row of $\mathbf{\Delta Q}$. Then we select the largest value $\mathbf{\Delta Q}_{ij}$ from the heap and join the respective communities $i$ and $j$. The values of the sparse matrix are updated and the process repeats.

### 2.2.4 K-Means Clustering

One early method for performing clustering is $K$-means [44]. $K$-means (Algorithm 2.1) works by identifying $k$ centroids within the data. Each point's proximity to the centroid determines its cluster. This is accomplished by randomly initializing $k$ centroids. Every data point is then classified as belonging to the cluster associated with the nearest centroid. New centroids are then calculated to be the centers of the current clustering. This process continues until convergence or a maximum number of iterations have occurred.

One issue with $K$-means clustering is that it does not handle non-convex clusters well. Figure 2.1 shows an example where clusters discovered via $K$-means do not correspond to the logical clusters in the data [45]. Expectation maximization (EM) is a similar technique used in clustering [46, 47]. Using Gaussian Mixture Models, EM performs an expectation (E) step that calculates the probability of a point $\mathbf{x}_i$ belonging to a cluster $\mathbf{C}_k$ by

$$p\left(\mathbf{x}_i \mid \mathbf{C}_k\right) = \sum_{j}^{N} P\left(\Theta_{j|k} \mid \mathbf{C}_k\right) \mathcal{N}_k\left(\mathbf{x}_i; \Theta_{j|k}\right)$$

**Algorithm 2.1** $K$-means Clustering

**function** $K$-MEANS($\mathbf{X}, k$)

Initialize cluster centroids $c_1, c_2, \ldots, c_k \in \mathbb{R}^n$ randomly

**for** $t \in [1, \ldots, T]$ **do**

**for** $i \in [1, \ldots, k]$ **do**

$u_i = \arg\min_j \|\mathbf{x}_i - \mathbf{c}_j\|^2$

**end for**

**for** $j \in [1, \ldots, k]$ **do**

$\mathbf{c}_j = \frac{\sum_{i=1}^{m} 1\{u_i=j\}\mathbf{x}_i}{\sum_{i=1}^{m} 1\{u_i=j\}}$

**end for**

**end for**

**return u**

**end function**

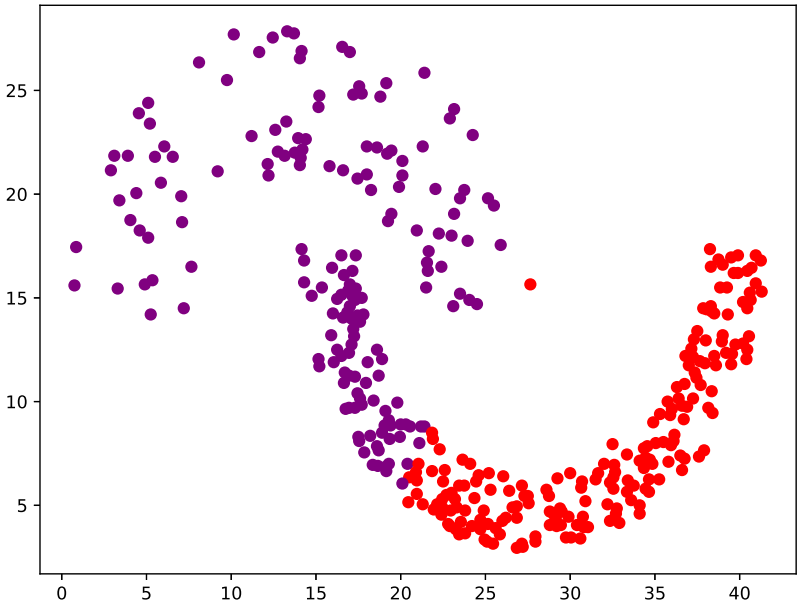



Figure 2.1: $K$-Means Clustering on Non-Convex Data

where $P\left(\Theta_{j|k} \mid \mathbf{C}_k\right)$ is the prior probability of the $j$th component parameter $\Theta_{j|k}$ given cluster $\mathbf{C}_k$ and $\mathcal{N}_k\left(\mathbf{x}_i; \Theta_{j|k}\right)$ is the normal density function on the $j$th component. The maximization step is a maximum likelihood problem over the parameters. New probabilities are calculated based on the likelihood function and the process repeats. EM has been shown to be effective in a wide range of topics.

### 2.2.5 Spectral Clustering

As mentioned in the previous section, $K$-means clustering does not handle non-convex clusters well. The results of $K$-means clustering are biased towards spherical clusters centered on the found centroids. Spectral clustering is a popular alternative that can handle non-convex clusters due to its ease of implementation [15].

Spectral graph partitioning methods already existed that relied on repeatedly cutting a network into smaller partitions [6]. It had been shown that the second eigenvector of a graph's Laplacian could be used to find an approximation of an optimal partition, typically by the sign of the entries in the vector. Each new partition would then be divided by isolating its nodes and performing the split again. Spectral clustering instead uses the top $k$ eigenvectors of the Laplacian instead of doing an iterative split [15]. A disadvantage to this method is that computing the eigenvectors of an $n \times n$ matrix has a time complexity of $\mathrm{O}\left(n^3\right)$. This can be an issue with spectral clustering as it can be costly for large datasets. Approximation algorithms can help as not all eigenvectors are necessary for the calculation [48], but many of these methods are not much faster or suffer from instability depending on the nature of the matrix. In addition to the spectral decomposition, calculating a full affinity matrix can be costly in certain as a full matrix representation requires $\mathrm{O}\left(n^2\right)$ in storage which is not feasible for large datasets. Social networks have an advantage since the adjacency matrix can be used as the affinity matrix, and there are sparse representations that

do not require as much storage.

To perform spectral clustering, consider points within a set of data $\mathbf{S} = \{s_1, \ldots, s_n\} \in \mathbb{R}^l$. Affinity matrix $\mathbf{A}$ is created by

$$a_{ij} = \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right)$$

where $\sigma$ is a parameter to control how quickly affinity drops off as the distance increases. In the case of networks, the adjacency matrix can be used as the affinity matrix since it describes affinity for neighboring points.

Spectral clustering proceeds by performing the following steps as shown in Algorithm 2.2. First, using the affinity matrix $\mathbf{A}$, define $\mathbf{D}$ as a diagonal matrix where $d_{ii} = \sum_j a_{ij}$ and $d_{ij} = 0$ for all $i \neq j$. The random walk Laplacian is defined as $\mathbf{L}_{rw} = \mathbf{D} - \mathbf{A}$. The work described later uses the normalized symmetric Laplacian $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{L}_{rw}\mathbf{D}^{-1/2}$. From there, calculate the $k$ eigenvectors corresponding to the $k$ smallest eigenvalues of $\mathbf{L}$ as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Those eigenvectors form matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$ by stacking the eigenvectors in columns. Each of the rows in matrix $\mathbf{X}$ are normalized to have unit length to create matrix $\mathbf{Y}$. The final step is to use another clustering algorithm on the rows in $\mathbf{Y}$, typically $K$-means. Figure 2.2 shows the results of performing spectral clustering on the same sample problem as before.

## 2.3 Political Networks

Based on data from the National Institute on Money in Politics [1], contributions to candidates and committees in 2016 reached $5.5 billion for federal elections and $3.70 billion for state level politics. For federal House and Senate candidates,

[1] Based on numbers provided by www.followthemoney.org

**Algorithm 2.2** Spectral Clustering

**Require:** $\mathbf{A} \in \mathbb{R}^{n \times n}$

**Require:** $2 \leq k < n$

1: **function** SPECTRAL CLUSTERING($\mathbf{A}, k$)

2: $\mathbf{D}_{ii} = \sum_j a_{ij} \forall i = 1, 2, \ldots, n$ ▷ Diagonal matrix of row sums

3: $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ▷ The normalized symmetric Laplacian

4: $[x_1, x_2, \ldots, x_k] = \text{eigs}(\mathbf{L}, k)$ ▷ Eigenvectors of $k$ smallest eigenvalues

5: $\mathbf{X} = [x_1, x_2, \ldots, x_k]$

6: $\mathbf{Y} = \forall i X_{ij} / \|\mathbf{X}_i\|$ ▷ Normalize the rows of $\mathbf{X}$

7: $\mathbf{U} = K-\text{means}(Y, k)$

8: **return U** ▷ Return the cluster assignments

9: **end function**

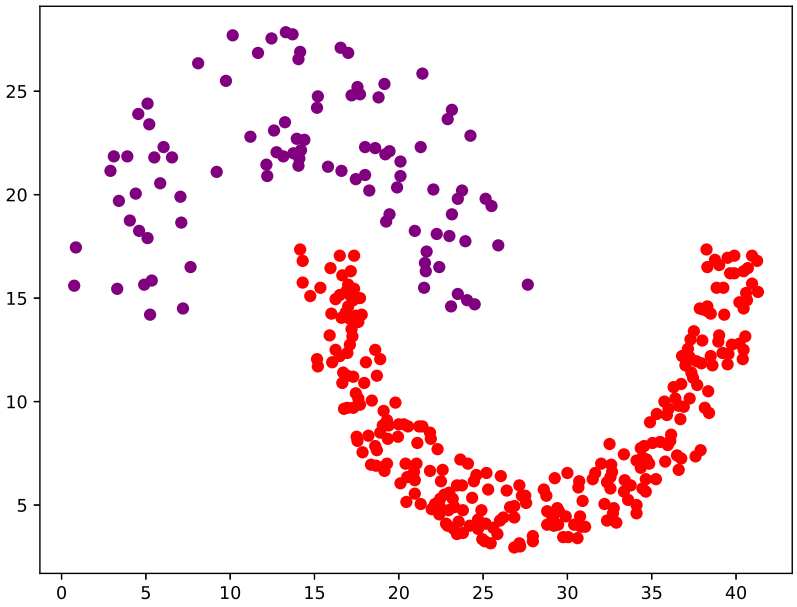



Figure 2.2: Spectral Clustering on Non-Convex Data

donations reached roughly $1.7 billion. The majority of that money comes from a relatively small number of donors. There were only 11,479 donors who gave $10,000

or more. The amount from this group was over $804 million, nearly half the total amount. The remaining half came from roughly 708,000 entities. State level politics tell a similar tale. Only $160 million of that total came from 1.3 million donors who gave less than $1,000 total. The remaining $3 billion came from only 166 thousand donors. Just like in federal elections, most of the money comes from the wealthy or special interest groups. The total amounts have increased since. In 2018, the sum of donations to federal candidates was lower at $4.6 billion, primarily due to that not being a United States presidential election year. However, donations to state candidates increased to $6.0 billion, more than all federal data from the prior presidential election cycle.

With such sums, the question of how such money may impact legislators and legislation is a very important one. It has been shown that a benefit to donating to a politician is that it provides access to that politician. While there is no clear evidence that standard political donations directly influence legislation, those who donate can more easily get their legislators to listen to them [49]. With this in mind, we use campaign finance networks as a real-world dataset where community analysis may provide additional insight, interpretability, and augmentations for additional research in other domains.

We construct the political social networks used later in this dissertation from campaign finance datasets. This data is based on transactions where individuals, businesses, or other organizations make donations to political groups. These groups can be candidates, politicians or political committees and form the vertices in the network. The donations to the political groups can be considered links and create the edges. The following describes areas in where community analysis of political social networks may provide benefit in augmenting other research.

One method for analysis of the effect of money in politics is to create models

using economic game theory. One such model analyzed the effects of campaign finance policy on legislation [50]. The model showed that by introducing caps, the amount of aggregate contributions is reduced due to the equilibria in the system. This is due to underlying preferences of the legislator. For example, if a lobbyist targets a politician who already had a preference for the favored policy of the lobbyist, then that lobbyist does not need to spend up to the cap since they know that opposition lobbyists would need to spend more to overcome that bias.

Another model shows how legislators may elect to adopt certain policy choices based on how interest groups donate [51]. Notably, even without the expectation of *quid pro quo*, this money can have an impact based on their temporal model. Since incumbents need to raise money for reelection, their knowledge of interest group donations can bias policy choices even in cases where the interest group has donated to the opposition. Some improvement can be made to these models, however, since they do not fully capture the group behavior of donations.

Using a randomized field study, [49] raised multiple questions to showcase how donating to a legislator can improve access. In setting up their experimental design, they not only asked whether or not campaign contributions to a legislator improved the chance to get access to that legislator, but also if donating to different legislators helped gain access to legislators to which they did not donate. Their experiment used a political organization attempting to arrange meetings with legislators and constituents who had previously donated. Whether or not they revealed if the attendees had donated was decided randomly, allowing a better analysis on how donations impact access. From their results, only 2.4% of the offices arranged meetings when only told the attendees were constituents. However, 12.5% arranged meetings for attendees who were donors. Of those meetings, only 5.5% of the constituents met with a senior staffer, compared with 18.8% for donor attendees

As related to the work herein, community analysis may help discover and analyze groups donating to similar candidates, who together, may wield substantially more power in gaining access to lawmakers than others.

As in many social science disciplines, surveys are also useful in gaining insight into behaviors, either as a study or for additional variables when doing regression. One related work concerns the geography of political contributions. Akey notes that most large campaign donations come from areas where political opinion does not necessarily reflect the majority of the population [52]. Using Federal Election Commission data and surveys, it was shown that political opinion from different geographic areas does not match the nation as a whole, especially from areas with fewer donations. Related to that work, though not using survey data, Gimpel, Lee, and Kaminsky found that most money came from areas within major metropolitan areas and not necessarily from traditional bases of support for either party [53].

As noted, survey data is sometimes used to gather independent variables for regression. In one such example, Kirkland [54] hypothesized that legislators in moderate districts would be more likely to be disloyal to their party, whereas legislators from districts with high ideological variance would have higher party cohesion [54]. Survey data was used to determine the ideological variance within states and then used as independent variables in regression to confirm the hypothesis. These studies are exceptionally useful in providing insight into the nature of political contributions, but more can still be done.

Analyzing the laws themselves, prior research questions the impact of stricter campaign finance laws on egalitarian policies [55]. This work shows that states with stricter laws spend larger portions of their budget on public welfare. Although it was noted that individuals from working class backgrounds are underrepresented in politics, it does not appear that stricter campaign finance laws work to improve

that representation. Benz et. al. [56] analyze indirect relationships between policy and political action committee donations. Using principal component analysis on measures of population health in addition to others, they were able to show a correlation between health care lobbying registrations and the number and size of political action committee (PAC) donations. Their claim is that this could be interpreted such that PACs are in effect an extension of lobbying. Other work attempts to look at the effects of campaign contributions directly on foreign policy outcomes [57]. In that work, there is a focus on pro- or anti- embargo votes regarding Cuba along with donations from Cuban American interests. The author was able to show that contributions from that ethnic identity group do have an impact on pro-embargo policies.

A lot of work has been done in analyzing industry or business contributions. Powell and Grimmer [58] used changes in committees after a party loses a majority to analyze how businesses donate to those members. After a change in party control, some members of the committee are removed involuntarily from their assignment. By pairing donation behavior with this information on exiled legislators, they attempt to identify how behavior changes in response to the loss of power of the former committee members. It is shown in their results that the PACs focus on short term access and reduce donations to members who no longer sit on committees that govern their relevant industry and increase donations to new members. This is in addition to the possibility of businesses stopping donations to a political party or members if they feel confident the party will lose their majority, causing these results to be an underestimate.

Other work has shown certain PAC industries have a political tilt where they focus donations to specific parties [59]. Using gross contributions to candidates, Gilbert and Oladi [60] show correlations with industries when analyzing votes on

trade with China. In particular, agricultural interests were very effective per dollar spent. Huber and Kirchler [61] used regression to show that company contributions to winning candidates were correlated with higher returns in the stock market.

There is also substantial work that focuses on election results. Brown [62] was able to use gubernatorial data to show that self-financing is not as effective in getting elected as getting external funds. Benoit and Marsh [63] analyzed the effects of spending in ranked choice voting in Ireland. The hypothesis was that incumbent spending would be less effective. Using variables for spending, incumbency, party strength, and others, they confirmed that incumbent spending is approximately half as effective regarding election success as challengers, although not in the case of same-party challengers.

Holbrook and Weinschenk [64] attempted to ascertain what effect campaign expenditures have on voter turnout in local elections. Using generalized least squares, the experiment analyzed several variables expected to influence turnout in local elections. Based on their results, for all their models, campaign spending had a positive significant effect on turnout levels. Basinger et. al. [65] used models where the voting shares for two parties were modeled independently. Their claim was that just combining the vote shares does not accurately reflect the way other variables cause changes to the voting shares. Their primary contribution was in the way voting shares were modeled, doing comparisons to other methods to reveal additional correlations. Streb and Frederick [66] relied on a Heckman selection model to analyze rolloff for judicial elections. In determining the cause for rolloff, the most important variable was a measure of partisanship and not campaign spending in this case.

More in the area of social networks, Internet presence was also shown to be effective in fund-raising, especially around small-donor contributions [67]. Crespin and Deitz [68] attempted to show the effectiveness of female donor networks in

supporting female candidates. Like many of the other studies, the research used ordinary least squares over several candidate variables. They note that based on their results, women raise more in small dollar donations, indicating the donor networks are effective in that way. The difference between men and women for larger values, however, was not significantly different. Over all their variables, small individual contributions favored Democratic women in particular.

Some similar prior work studied how candidates from parties fall ideologically at local and national [69] This work assumes that under a spatial ideological model, that opposing candidates should move toward the ideological center of a district in order to capture more votes. Their results show that the two parties do not converge, but that there is a correlation regarding policy positions and ideology of candidates with the ideology of the district. Notably, incumbency was associated with more moderate positions of candidates in subsequent elections.

There has also been some work in the intersection of politics and social networks. Some of this prior work looks at social interaction and its effect on political participation [70, 71, 72]. Additional work analyzes elitism and the behavior of corporations in politics [73]. There has also been research on the geography of donations and its usefulness as an indicator for predicting donations [53]. Additional network analysis has been done showing how donations fit other models. Preferential attachment in networks—new connections in a network tend to occur in areas with many preexisting connections—has been shown to help explain how shares of a donor pool can affect the probability of gaining more donors [74]. This, however, does not fully answer the question of whether larger groups exist within the networks that are together having a more significant impact.

While the above methods are useful on their own, the majority of research in this area relies on linear regression. That research gathers a set of variables to determine

the effects of independent variables on the dependent variable. Although these works use similar methods, the variables and questions they are attempting to answer can vary widely. Especially in cases involving campaign finance or social networks, the work presented in this dissertation may be of benefit to this type of research. The fuzzy community values found at each hierarchy can provide additional variables for use in regression. This can help determine if there are specific groups within politics who are having a strong impact.

### 2.3.1 DW-NOMINATE

Identifying ideology of political entities can be a useful tool in predicting or analyzing behavior. Finding numerical estimates of ideology greatly enhances the ability of performing quantitative analysis within politics. Considerable prior work has been performed on analyzing ideological estimates of legislators and the bills upon which they vote. A widely known and used tool is DW-NOMINATE, which stands for dynamic, weighted, nominal three-step estimation [75], [76]. At a high level, DW-NOMINATE is built upon the idea of a random utility model where a legislator $i$'s utility for an outcome (Yea) on a bill $j$ is given by $\mathcal{U}_{ij}^{y} = u_{ij}^{y} + \varepsilon_{ij}^{y}$ where $u_{ij}^{y}$ is a utility function, superscript $y$ represents a Yea vote, and $\varepsilon_{ij}^{y}$ is a random error sampled from an inverse exponential distribution[77]. The initial DW-NOMINATE work was based on a normal distribution utility function. In this work, the utility of a legislator's choice of voting Yea or Nay is centered around an estimated ideal point. The more distant an option is to a legislator's ideology, the less utility is gained by voting for that option.

As examples, the authors refer to the concepts of alienation and indifference. Alienation represents where the set of choices are far removed from the ideal point but on the same side of the ideological space. Indifference is where the choices are far

away but on either side of the space, as in a moderate politician faced with voting for two extremes, one on each side of the political spectrum.

Encapsulating this information requires determining the ideal points of legislation and legislators in order to calculate utility of choices. DW-NOMINATE does so by maximizing a likelihood function based on the probability of a legislator's choices regarding bills. In this model, the utility of a legislator $i$ voting Yea in the $k$th dimension is defined by

$$u_{ij}^{y} = \beta \exp\left(-\frac{1}{2}\sum_{k=1}^{s} w_k (d_{ijk}^{y})^2\right)$$

where $(d_{ijk}^{y})^2$ is the squared distance of the estimated ideal point in dimension $k$ to legislator $i$ to a Yea outcome in dimension $k$, $w_k$ are salience weights, and $\beta$ is an adjustment for overall noise, which is proportional to the variance of the error distribution. From this, the probability of voting Yea in the normal utility model is based upon the relative utilities of voting Yea or Nay. Using those values, this probability can be written as

$$P_{ij}^{y} = P\left(u_{ij}^{y} > u_{ij}^{n}\right) = P\left(\varepsilon_{ij}^{n} - \varepsilon_{ij}^{y} < u_{ij}^{y} - u_{ij}^{n}\right) = \Phi\left[u_{ij}^{y} - u_{ij}^{n}\right]$$

where $\Phi$ is the standard normal cumulative distribution function. In the normal model, this is given by

$$P_{ij}^{y} = \Phi\left[\beta\left\{\exp\left(-\frac{1}{2}\sum_{k=1}^{s} w_k (d_{ijk}^{y})^2\right) - \exp\left(-\frac{1}{2}\sum_{k=1}^{s} w_k (d_{ijk}^{n})^2\right)\right\}\right].$$

Using these equations, the model is estimated by maximizing the log likelihood as

defined by

$$\mathcal{L} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{\tau \in \{y,n\}} C_{ij}^{\tau} \ln(P_{ij}^{\tau})$$

where $\tau$ is an index for choices Yea and Nay, $P_{ij}^{\tau}$ is the probability of voting $\tau$ as defined above, and $C_{ij}^{\tau} = 1$ if the actual choice was $\tau$ and zero otherwise. In learning the model, Rosenthal and Poole estimated a single parameter at a time, holding the others fixed. The authors claim two dimensions worked well in practice.

Application of the model yields estimates of ideal points for both legislator and legislation in two dimensions. In practice, the first of the two dimensions is considered to correspond to a liberal and conservative economic spectrum. The second dimension corresponds to social issues. In later experiments within Chapters 6 and 7, the estimates of legislation are used to show the expressive power of the communities discovered in campaign finance networks. The discovered communities perform well in predicting votes in those two dimensions without explicitly using the legislators' ideal points as determined in the model.

### 2.3.2 CFScore

Recent work by Bonica [78] created an ideological estimate for both donors as well as candidates. In his work, contributions are assumed to represent evaluations of a candidate's ideology. Donors would be more likely to donate to those who share ideology. This common-space campaign finance score (CFscore) has the advantage of applying to both types of entities, where prior research focused solely on legislators or recipients. Calculating the CFScore for federal contribution data begins by creating an $n \times m$ contingency matrix $\mathbf{R}$. The rows of $\mathbf{R}$ map to contributors while the columns map to recipients. Each entry $r_{ij}$ in $\mathbf{R}$ contains a sum of the contributions from $i$ to $j$.

From this matrix, each entry $r_{ij}$ is converted into an integer in the range $[1, 50]$

by dividing $r_{ij}$ by 100 and capping the result to 50. This value is further standardized by dividing each $r_{ij}$ by $\sum_i \sum_j r_{ij}$. From this matrix, singular value decomposition (SVD) is performed to obtain $\mathbf{K} = \mathbf{D_r}^{-\frac{1}{2}}(\mathbf{R} - \mathbf{rc}^\top)\mathbf{D_c}^{-\frac{1}{2}}$ where $\mathbf{r}$ and $\mathbf{c}$ are vectors of row and column sums of $\mathbf{R}$. Additionally, $\mathbf{D_r}$ and $\mathbf{D_c}$ are diagonal matrices where the values of $\mathbf{r}$ and $\mathbf{c}$ are placed on the diagonal. From this SVD calculation, ideal points can then be estimated using $\theta = \mathbf{U}\mathbf{D_c}^{-\frac{1}{2}}$ for contributors and $\delta = \mathbf{V}\mathbf{D_r}^{-\frac{1}{2}}$ for recipients. For these equations, the matrix $\mathbf{U}$ gives the left eigenvectors of $\mathbf{K}\mathbf{K}^\top$, and $\mathbf{D_r}$ is a diagonal matrix of singular values. The matrix $\mathbf{V}$ are the right eigenvectors of $\mathbf{K}^\top\mathbf{K}$. For state data, these federal ideal points are used as bridge observations in an iterative procedure that estimates contributor and recipient CFScores across states.

# CHAPTER THREE

# HIERARCHICAL FUZZY SPECTRAL CLUSTERING

## 3.1 Background

In this section we discuss the clustering algorithms that we utilize in the development of our algorithm for hierarchical fuzzy spectral clustering. We also discuss methods for evaluating fuzzy clusters and finding the number of clusters using spectral characterization.

### 3.1.1 Fuzzy $c$-Means

As discussed earlier, we use fuzzy clustering since nodes in these real-world networks can belong to more than one community. In fact, the nodes that are in multiple communities are of particular interest. Assuming a simple split into two communities based on a left-right ideological spectrum, nodes that are in both communities include those that may be more interested in political access than left-right ideology. One common method for performing fuzzy clustering is fuzzy $c$-means clustering (FCM) [79, 80, 81]

FCM is similar to $K$-means in that it calculates centroids within this data and assigns communities based on the distance from a centroid to a specific data point. Where the algorithm differs is that FCM uses fuzzy sets for the community assignments. Instead of vertex $v_i$ belonging to a single community $\mathbf{C}_j$, each vertex has a fuzzy set $\mathbf{u}_i = \{u_1, u_2, \ldots, u_k\} \in \mathbf{U}$ that defines how much that vertex belongs to each community. For each vertex $v_i$, the sum of its membership to each community is equal to one, i.e.,

$$\sum_{j:u_j \in \mathbf{u}_i} u_j = 1.$$

**Algorithm 3.1** Fuzzy C-Means

**Require:** $\mathbf{X} \in \mathbb{R}^{n \times m}$

**Require:** $2 \leq k < n$

**Require:** $1 \leq m < \inf$

1: **function** FCM($\mathbf{X}, k, m$)

2: $\mathbf{U}^0 \leftarrow dirichlet\,(n,k) : \forall i \in [0,1,\ldots,n) \sum_j u_{ij} = 1$ ▷ Initialize cluster assignments.

3: **for** $t \in [1,\ldots,T]$ **do**

4: **for** $p \in [1,\ldots,k]$ **do**

5: $\mathbf{c}_p = \frac{\sum_i [\mathbf{u}_p(x_i)]^m x_i}{\sum_i [\mathbf{u}_p(x_i)]^m}$ ▷ Update centers using membership matrix and $m$.

6: $\mathbf{u}_p^t\,(x_i) = \frac{\left[\frac{1}{\tilde{d}_{ip}}\right]^{\frac{1}{m-1}}}{\sum_{j=1}\left[\frac{1}{\tilde{d}_{ij}}\right]^{\frac{1}{m-1}}}$ ▷ Update community assignment matrix.

7: **end for**

8: **if** $\max \left\|\mathbf{u}_p^t - \mathbf{u}_p^{t-1}\right\| < \epsilon$ **then**

9: Stop early

10: **end if**

11: **end for**

12: **return** $\mathbf{U^t}$

13: **end function**

Details on FCM are shown in Algorithm 3.1. Like $K$-means, FCM begins by identifying the number of clusters. Initial cluster assignments are assigned randomly to each data point $w_{i,j}$. This initial setting uses a sampling from a Dirichlet distribution such that each row sums to one. Like $K$-means, FCM then proceeds by repeating two steps. The centroid for each cluster is calculated using

$$\mathbf{c}_k = \frac{\sum_x \mathbf{u}_k(x)^m x}{\sum_x \mathbf{u}_k(x)^m}$$

where $m$ is an exponential term that determines the amount of fuzzy overlap between the clusters. A higher value results in fuzzier clusters. The assignments are then updated with the new centroids using

$$\mathbf{u}_{ij} = \frac{1}{\sum_k \left( \frac{\|\mathbf{x}_i - \mathbf{c}_j\|^2}{\|\mathbf{x}_i - \mathbf{c}_k\|^2} \right)^{\frac{1}{m-1}}}$$

until the algorithm converges or the maximum number of iterations is reached. The resulting weights and centroids define the discovered fuzzy clusters. Later work extended fuzzy c-means for hierarchical clustering that uses a different approach [82]. In this Centroid Auto-Fused Hierarchical FCM algorithm, an $\ell_2$ norm penalty is added between cluster centroids to encourage the centroids to fuse. As the algorithm proceeds, cluster centers coalesce, creating a hierarchy as cluster centers agglomerate. The authors begin with a large number of centroids and fuse the centroids until an "optimal" number of clusters is reached. Other methods define a partition function and a similarity measure between two vertices [83]. The community detection becomes a nonlinear constrained optimization problem. This is solved using a gradient-based algorithm and simulated annealing to obtain fuzzy community assignments.

### 3.1.2 Fuzzy Modularity

Like in fuzzy clustering, there are adaptations of clustering metrics for fuzzy communities. A fuzzy modularity $\widetilde{Q}$ was developed to assess the splits created by fuzzy clustering [32]. Its form and principle are similar to that of the original modularity. $\widetilde{Q}$ is defined by

$$\widetilde{Q}\left(\mathbf{U}_k\right) = \sum_{c=1}^{k} \left[ \frac{E\left(\bar{V}_c, \bar{V}_c\right)}{E\left(V, V\right)} - \left( \frac{E\left(\bar{V}_c, V\right)}{E\left(V, V\right)} \right)^2 \right].$$

In the above equation, $\mathbf{U}_k$ is a fuzzy partition of $k$ clusters,

$$E\left(\bar{V}_c, \bar{V}_c\right) = \sum_{i \in \bar{V}_c, j \in \bar{V}_c} a_{ij} \left( \frac{u_{ic} + u_{jc}}{2} \right),$$

$$E\left(\bar{V}_c, V\right) = E\left(\bar{V}_c, \bar{V}_c\right) + \sum_{i \in \bar{V}_c, j \in V \setminus \bar{V}_c} a_{ij} \left( \frac{u_{ic} + u_{jc}}{2} \right),$$

and

$$E\left(V, V\right) = \sum_{i \in V, j \in V} A_{ij}.$$

This gives a method for evaluating different fuzzy community assignments. By using fuzzy modularity, it is also possible to perform similar agglomerative clustering techniques as those used for regular modularity [84], or other approaches like simulated annealing [19].

### 3.1.3 Spectral Characterization

One important consideration when clustering networks is determining how many clusters to use. The problem occurs when determining a stopping point during top-down splitting of communities, when determining the starting number of clusters for bottom-up clustering, or for the optimal clustering number for other methods. Crisp clustering has an advantage that a single value is used to describe the community assignment for each node in a network. Fuzzy clustering requires storing multiple values for the community assignments for every node in the network. This becomes unwieldy with a large number of clusters, especially in hierarchical clustering where there are multiple values for each community. Truncating low values to eliminate some storage does not necessarily solve the issue. Because of these issues, it becomes important to try to determine criteria for determining an appropriate number of clusters $k$. One possible method is to use spectral characterization to limit the number

of communities.

To see how spectral characterization can be used to inform the number of communities, we need to consider the structure of the networks in question. These networks have non-negative and undirected edges between nodes. Thus, these networks are both symmetric and positive. This gives some useful properties when analyzing the eigenvalue decomposition. We start with analyzing the eigenvalues of the adjacency matrix $\mathbf{A}$ of the network. 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 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 [40, 85, 86, 87, 88]. To illustrate this, consider an example random network with 1000 nodes. Suppose 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 its own community since there are no special defining characteristics separating any of the nodes. Figure 3.1 plots the eigenvalues of this directed network. As can be seen, there is a single eigenvalue outside the main cluster of nodes.

As predicted, the largest eigenvalue here is related to the average degree of the nodes in $\mathbf{A}$ [89]. 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, $\lambda_{max} \approx n \times p$, or in this case, $\lambda_{max} \approx 1000 \times 0.04 = 40$. There is additional work on Erdös-Rényi uncorrelated random graphs showing the edge

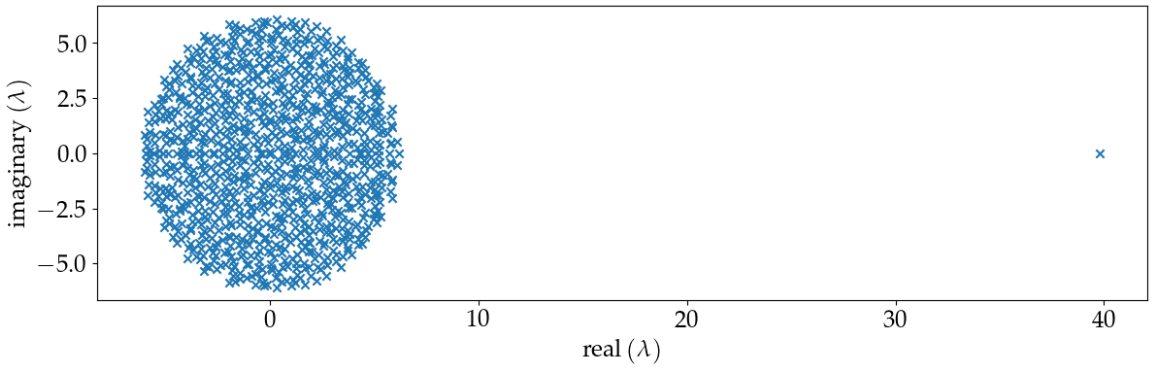



Figure 3.1: Eigenvalues of a random network for $n = 1,000$

of the large cluster of eigenvalues is approximately defined by $\sigma\sqrt{n}$ where $\sigma$ is the standard deviation of the values $a_{ij} \in \mathbf{A}$. However, this does not hold for non-random graphs. Analysis on real-world campaign finance networks shows that using $\sigma\sqrt{n}$ 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. Figure 3.2 also shows eigenvalues, but of a different network of 1000 nodes created with four communities of equal size, $p = 0.1$, and $q = 0.01$. As can be seen in the graph, 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. In another example, Figure 3.3 shows the eigenvalues of a network created with 6 communities. Each community in this

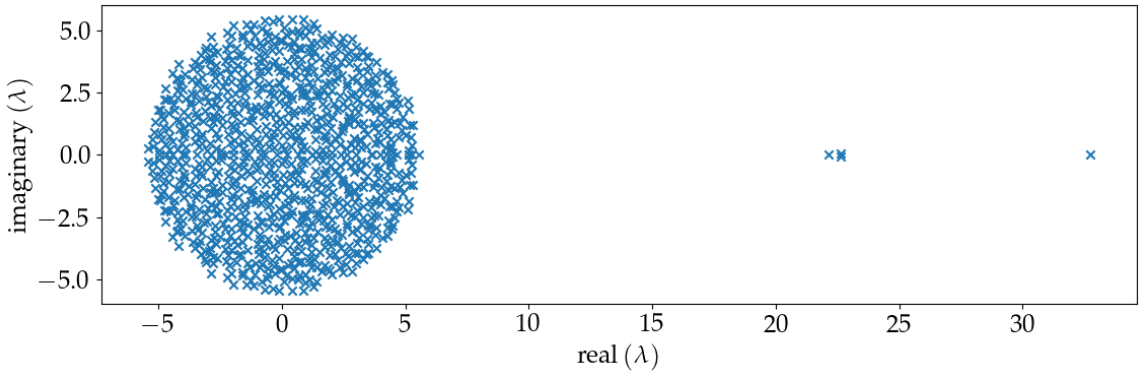



Figure 3.2: Eigenvalues of random network with communities

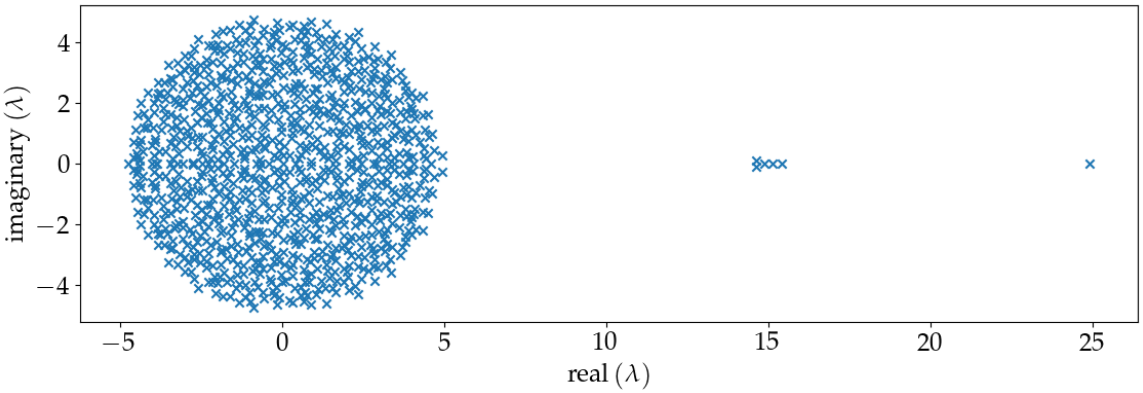



Figure 3.3: Eigenvalues of random network with communities

network contains approximately 166 vertices. This yields the group of eigenvalues at approximately $166 \times (0.1 - 0.01) = 14.94$.

In the case of networks with hierarchical communities, the spectrum of the network shows multiple groups of eigenvalues when the communities are of similar size. The network used to generate Figure 3.4 has a top-level hierarchy of 4 nodes with each having 4 sub-communities, creating 16 total clusters. As can be seen in the graph, 16 eigenvalues are located outside of the main cloud and are split into two separate clusters. The gap between eigenvalues, or eigen-gap, indicates a separation between levels within the hierarchy. This principle is what is used when attempting to

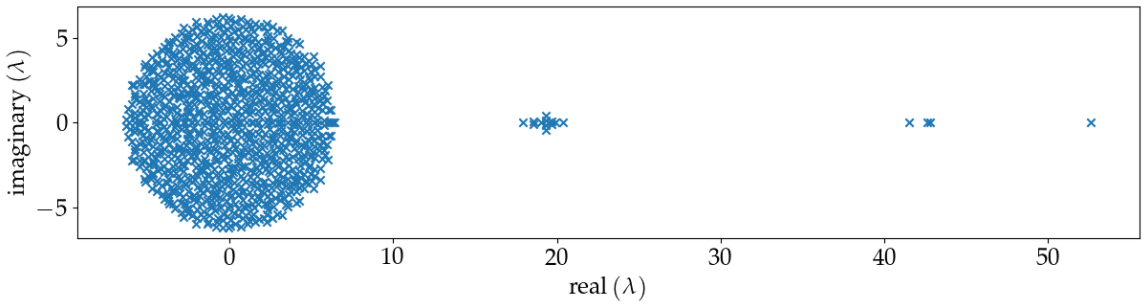



Figure 3.4: Eigenvalues of a random network with hierarchical communities

determine the number of communities in each hierarchical level relevant for spectral clustering.

## 3.2 Hierarchical Fuzzy Spectral Clustering

The approach proposed is primarily based on the spectral clustering work of Ng, Jordan, and Weiss [15] as well as Zhang, Wang, and Zhang [32]. Firstly, the spectral composition of the network must be determined. As described in previous section, we can utilize the spectral characterization in determining the level of hierarchy in the network and the number of clusters at each hierarchical level. This can be done by finding outliers in the gaps between eigenvalues of the spectral decomposition. The method for this outlier detection is given in Algorithm 3.2.

This algorithm proceeds by first finding all the of eigenvalues of matrix $\mathbf{A}$. On the real-world campaign finance networks, limiting the calculation to only eigenvalues $\lambda_i \geq 1$ provided good results in finding $k$. The set of eigenvalues is then sorted in ascending order. We then calculate the gap between eigenvalues as $\delta_i = \lambda_{i+1} - \lambda_i$, creating the set of eigengaps $\delta_i \in \mathbf{\Delta}$. The threshold $\tau = \text{aad}\,(\mathbf{\Delta}) \times 1.482$ is found

**Algorithm 3.2** Spectral Analysis

1: **function** NUMCOMMUNITIES($\mathbf{A}$)
2: $\quad\mathbf{\Lambda} = \text{eigenvalues}\,(\mathbf{A}) : \lambda_i \geq 1$ ▷ Calculate the eigenvalues of matrix $\mathbf{A}$.
3: $\quad\mathbf{\Lambda}' = \text{sort}\,(\mathbf{\Lambda})$ ▷ Sort the eigenvalues in ascending order.
4: $\quad$**for all** $\lambda_i \in \lambda' : i < |\lambda'|$ **do**
5: $\quad\quad\delta_i = \lambda_{i+1} - \lambda_i$ ▷ Calculate the difference between successive eigenvalues.
6: $\quad$**end for**
7: $\quad\tau = \text{aad}\,(\mathbf{\Delta}) \times 1.482$ ▷ Find average absolute deviation of the eigengaps.
8: $\quad ind = \text{first index}\,(\delta_i \in \Delta : \delta_i \geq \tau)$
9: $\quad k = |\Delta| - ind + 1$ ▷ $k$ is the index of $\lambda$ with a gap greater than threshold $\tau$.
10: $\quad$**return** $k$
11: **end function**

using absolute average deviation given by

$$\text{aad}\,(\mathbf{\Delta}) = \frac{1}{n}\sum_{i=1}^{n}\left|\delta_i - \bar{\mathbf{\Delta}}\right|$$

where $\bar{\mathbf{\Delta}}$ is the average eigengap. The constant 1.482 is from an assumption on a Gaussian distribution of the values [90]. The number of clusters $k$ is then calculated by finding the first $\delta_i \geq \tau$.

If the results of spectral analysis are inconclusive, it is possible to fall back on iterative testing of the partitions using fuzzy modularity as an optimization metric. With the number of clusters at the largest level of the hierarchy determined, an initial clustering is performed on the eigenvector decomposition of the normalized symmetric Laplacian $\mathbf{L}_{\text{sym}}$ of adjacency matrix $\mathbf{A}$ as shown in Algorithm 3.3.

The fuzzy spectral clustering algorithm starts by calculating the Laplacian

**Algorithm 3.3** Fuzzy Spectral Clustering

1: **function** FSC($\mathbf{A}, k$)
2: $\mathbf{D} \leftarrow \{d_i = \sum_1^n a_{ij}\}$ ▷ Create the diagonal matrix $\mathbf{D}$ from the row sum of $\mathbf{A}$.
3: $\mathbf{L}_{\text{sym}} = \mathbf{I}^{n \times n} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ▷ Calculate the normalized symmetric Laplacian.
4: $\mathbf{Z} = \text{eigenvectors}(\mathbf{L}, k)$ ▷ Compute the top-$k$ eigenvectors.
5: $\mathbf{X} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]$ ▷ Create matrix $\mathbf{X}$ where column $i$ in $\mathbf{X}$ is eigenvector $\mathbf{z}_i$.
6: **for all** row $\mathbf{x}_i \in \mathbf{X}$ **do**
7: $x_{ij} = x_{ij} / \|\mathbf{x}_i\| \; \forall \, x_{ij} \in \mathbf{x}_i$ ▷ Normalize each row of matrix $\mathbf{X}$.
8: **end for**
9: $\mathbf{U} = \text{FCM}(\mathbf{X})$ ▷ Use fuzzy c-means to get cluster assignments.
10: **return U**
11: **end function**

matrix $\mathbf{L}_{\text{sym}}$. The top-$k$ eigenvectors are calculated from matrix $\mathbf{L}_{\text{sym}}$. The top-$k$ vectors correspond to the $k$ smallest eigenvalues for the normalized symmetric Laplacian. These eigenvectors are oriented as the columns of matrix $\mathbf{X}$. After normalizing each row of $\mathbf{X}$, we use fuzzy c-means to find the cluster assignments $\mathbf{U}$ for each of the vertices of the network.

To obtain hierarchical structure, the process is repeated with a varying $k$ corresponding to the number of clusters in each hierarchical level, shown in pseudo-code in Algorithm 3.4. In practice, calculating the eigenvectors can be performed once with the largest $k$ and reused in subsequent iterations of the clustering. The communities are calculated for each level using FSC to create set of communities on a hierarchical level $i$. Each community in level $i$ is connected to its previous by calculating the fuzzy Jaccard similarity measure $\mathcal{J}_f$ of the communities. $\mathcal{J}_f(\mathbf{C}_i, \mathbf{C_j})$

**Algorithm 3.4** Hierarchical Generation

1: **function** HFSC($\mathbf{A}, k$)
2: **for** $i = 2$ to $k$ **do**
3: $C_i = \mathrm{FSC}(\mathbf{A}, i)$ ▷ Find the clusters for each level in the hierarchy.
4: **end for**
5: **for all** $C_{i,m} \in C_i : 3 \leq i < k$ **do**
6: $P_{i,m,n} = \arg\max \mathcal{J}_f(C_{i,m}, C_{i-1,n})$ ▷ Find parent $P_{i,m,n}$ for each $\mathbf{C}_{i,m}$
7: **end for**
8: **end function**

between two fuzzy communities $\mathbf{C_i}$ and $\mathbf{C_j}$ is given by

$$\mathcal{J}_f(\mathbf{C}_i, \mathbf{C}_j) = \frac{\sum_{u_k \in C_1 \cup C_2} \min(C_{1,i}, C_{2,i})}{\max(C_{1,i}, C_{2,i})}.$$

The results give similarity measures for the smaller clusters that can be used to assign each cluster to its best matching parent. In practice, the possible parent of a child that has the highest similarity is selected as the parent of the community. To test the efficacy of the algorithm, we analyze two real-world networks and present the fuzzy clustering results for those networks. These networks were chosen for their popularity in benchmark testing for community detection as well as the presence of hierarchical communities.

### 3.2.1 Zachary Karate Network

The first real-world example is the Zachary Karate Club network [91], a very common benchmark set used with community detection algorithms. It is popular since it is small and has known clusters. As the background story goes, due to conflict between the club president and the instructor the 34 members split into two separate

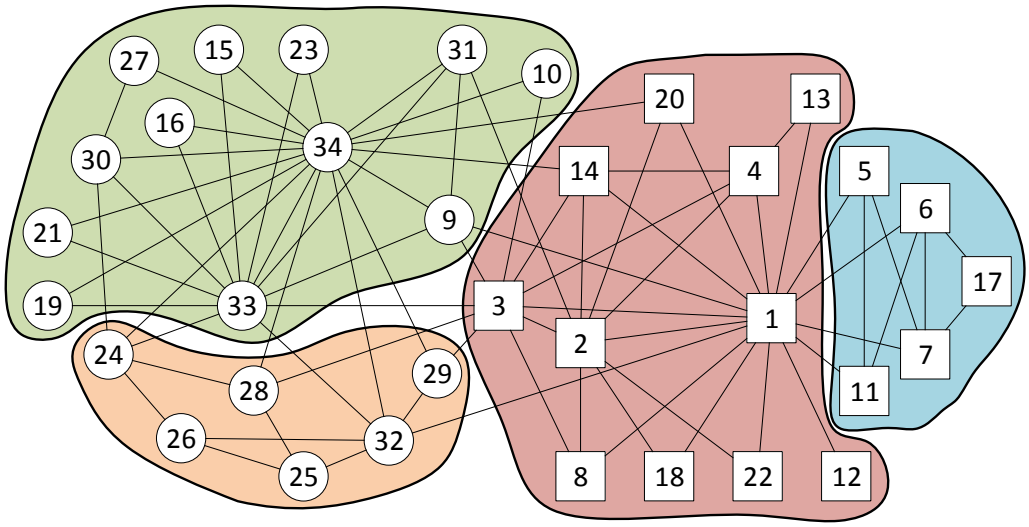



Figure 3.5: Karate Network

groups. This network has another useful property in that there are sub-communities within the two primary groups corresponding to areas of higher relative in-group connections. In fact, the best partition of the network, with respect to modularity, splits the set into four groups [11].

This network and the known clusters and sub-clusters are shown in Figure 3.5, as shown in [11]. The different node shapes (circles and squares) represent the true clusters with the highlighted regions covering the sub-clusters.

The corresponding spectrum for the network is given in Figure 3.6. This spectrum shows two hierarchical levels, based on the large gaps between eigenvalues located outside the cloud. The two largest eigenvalues correspond to the communities created by the true clusters. Outside of the primary cloud is another cluster of eigenvalues that represent the sub-communities within the primary clusters. Using this information, hierarchical fuzzy spectral clustering is applied to the network.

Using fuzzy spectral clustering as defined earlier, Figure 3.7 shows the overlapping clusters with $k = 2$. Here we set a threshold $\tau_u$ where we consider a vertex $v_i$

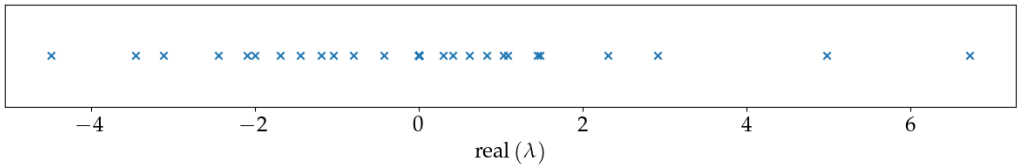



Figure 3.6: Karate Spectral Characteristic

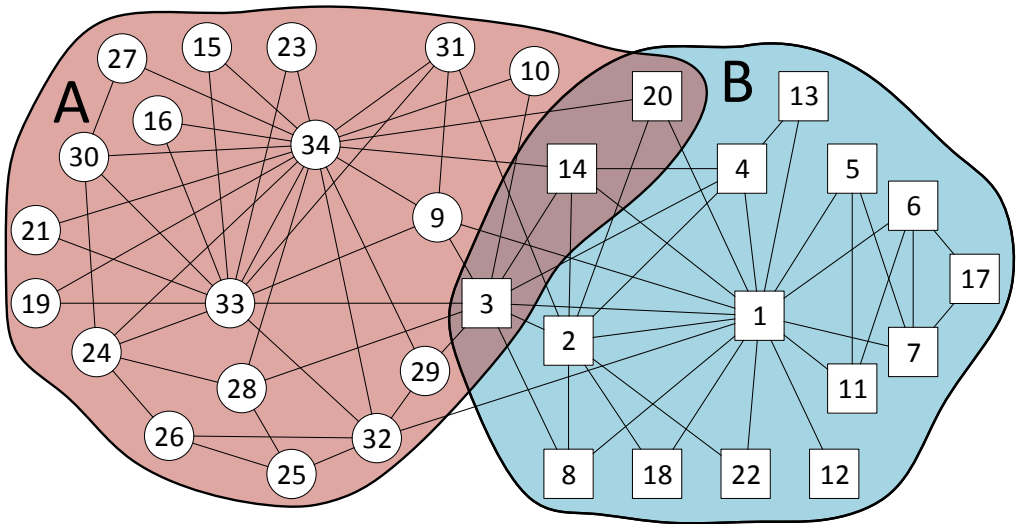



Figure 3.7: Karate network overlapping communities: $k = 2$ and $\tau_u = .25$

to be part of community $\mathbf{C}_j$ if the fuzzy community assignments $u_{ij} \geq \tau_u$. Assigning communities with $\tau_u = 0.25$, nodes 3, 14, and 20 are considered to be overlapping nodes. This appears to make sense as those nodes are connected to the most connected and central nodes of the two different clusters. For cluster **A**, these are nodes 1 and 2, while in **B** these are 33 and 34.

Next, these results are compared with the sub-clusters. Figure 3.8 shows the fuzzy clusters with $k = 4$ and $\tau_u = 0.16$. At this level, 3, 14, and 20 are no longer overlapping nodes due to the dissimilarity of the clusters. These clusters are now less defined by their proximity to the central nodes 1, 2, 33, and 34, and instead more by their local connections. Thus, the set $\{1, 5, 6, 7, 11, 17\}$ becomes its own

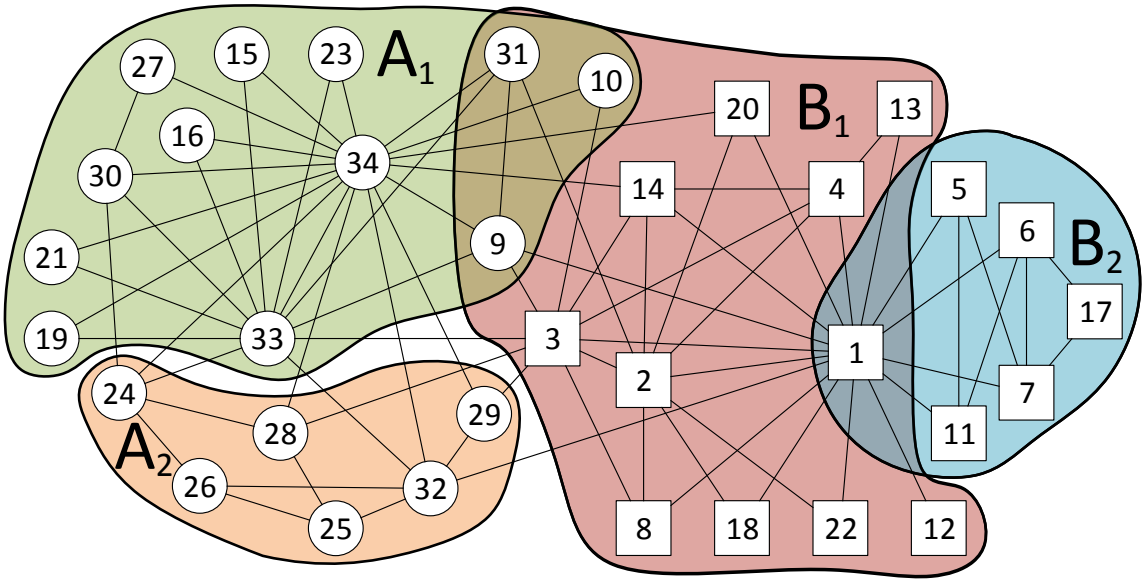



Figure 3.8: Karate network overlapping communities: $k = 4$ and $\tau_u = .15$

cluster since most of these nodes are only connected to each other. The set $\mathbf{A_2} = \{24, 25, 26, 28, 29, 32\}$ IS now its own community, separate from $\mathbf{A_1}$, which are better defined by their proximity to 33 and 34. Additionally, nodes 9, 10, and 31 become assigned to both $\mathbf{A_1}$ and $\mathbf{B_1}$.

It should be noted that increasing the value of $\tau_u$ to $\tau_u = 0.3$ results in an assignment with no overlap where the communities are identical to the sub-communities shown in Figure 3.5.

### 3.2.2 Dolphin Network

The Dolphin network is another well-known example of a social network [5]. This network represents a group of dolphins that were tracked over a period of time. Eventually, the dolphins split into separate groups. From prior work by Lusseau and Newman [5], one of the communities was further broken down into smaller communities. In the later figures, the solid black nodes represent one of the true clusters, and all the grey and white nodes together form the other true cluster. The

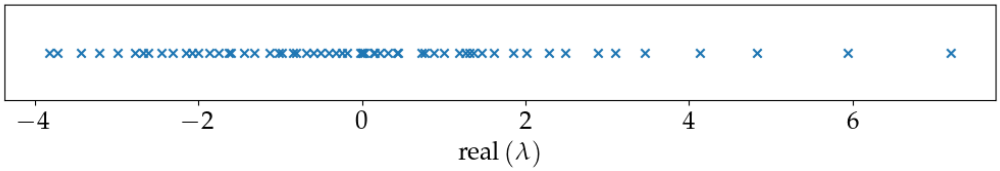



Figure 3.9: Dolphin Spectral Characteristic

different grey-scale colors correspond to the sub-cluster results from Lusseau and Newman.

Viewing the spectral characteristics of this network in Figure 3.9, it is possible to see by the spectral characterization that it does not have as strong of a hierarchical structure when compared to the karate network. There are two hierarchical levels, but the exact number of sub-communities is difficult to determine as it begins to merge with the primary cloud.

As the largest eigen-gap between the values occurs after the second largest eigenvalue, the initial pass clusters into two partitions. The next phase proves more difficult due to the remaining eigenvalues. Since there is a fairly smooth transition from the bulk distribution to the other eigenvalues, we use optimal fuzzy modularity, restricting the search to the approximate number of communities. This procedure gives a best partition using six clusters.

Using this information to get the smaller clusters, the resulting six communities are shown in Figure 3.11. These communities align well with previous results, with the exception of tr88 and tr120, which are added to the community $\mathbf{A_2}$. Unfortunately, these two have considerably different connections in relation to the rest of the members of $\mathbf{B_3}$, weakening their association with those nodes. Since the fuzzy assignment across all nodes must equal 1, this gets distributed across the other nodes, raising the association with $\mathbf{A_2}$ beyond the $\tau_u$ threshold. Likely, it is most strongly tied with

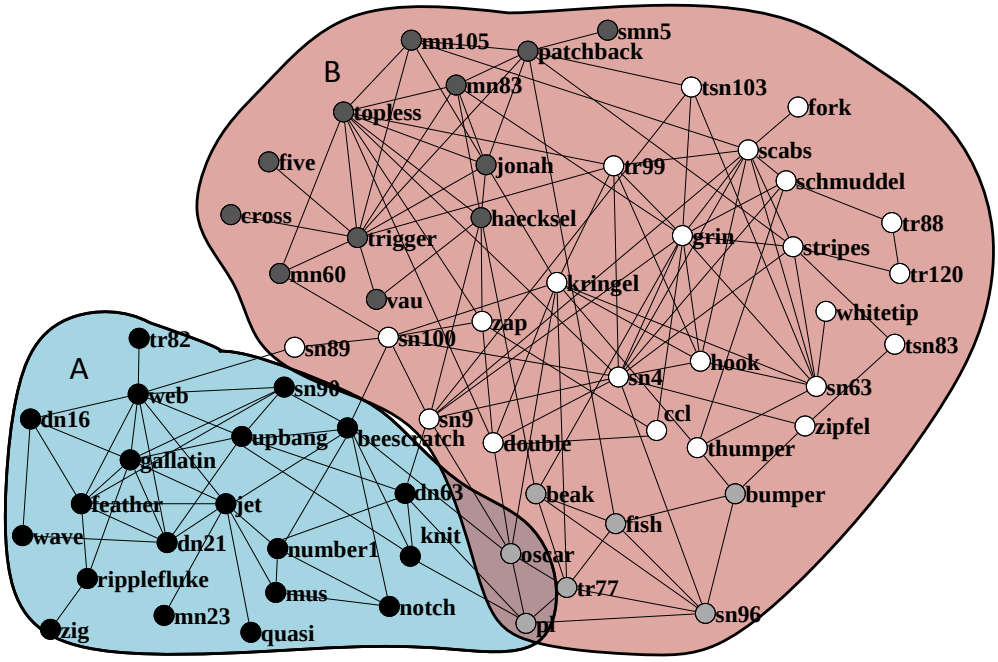



Figure 3.10: Dolphin network overlapping communities: $k = 2$ and $\lambda = .20$

these because $\mathbf{A_2}$ and $\mathbf{B_3}$ share proximity to $\mathbf{B_4}$. Raising $\tau_u$ does place them solely in $\mathbf{B_3}$ but weakens other associations and yields lower modularity.

Still, even with that outlier, communities $\mathbf{B_k}$ closely correspond to one of the true communities. Likewise, communities $\mathbf{A_k}$ match closely with the other true community.

Although there are now more communities than what was determined by Lusseau and Newman, merging $\mathbf{B_2}$ and $\mathbf{B_3}$ into a community and $\mathbf{A_1}$ and $\mathbf{A_2}$ into another community yields very similar results. Attempting to compute $k = 4$ communities directly yields different results as shown in Figure 3.12.

### 3.2.3 Alaska Campaign Finance

Previous work has shown that the primary motivator for donations from individuals is ideology. However, for non-individuals, they may attempt more

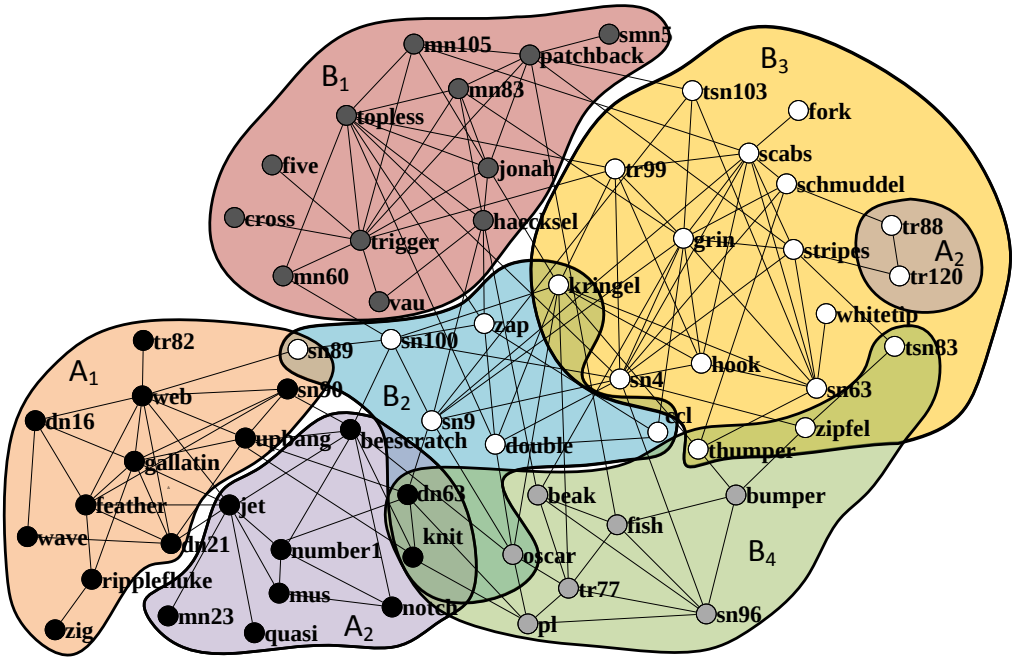



Figure 3.11: Dolphin network overlapping communities: $k = 6$ and $\lambda = .17$

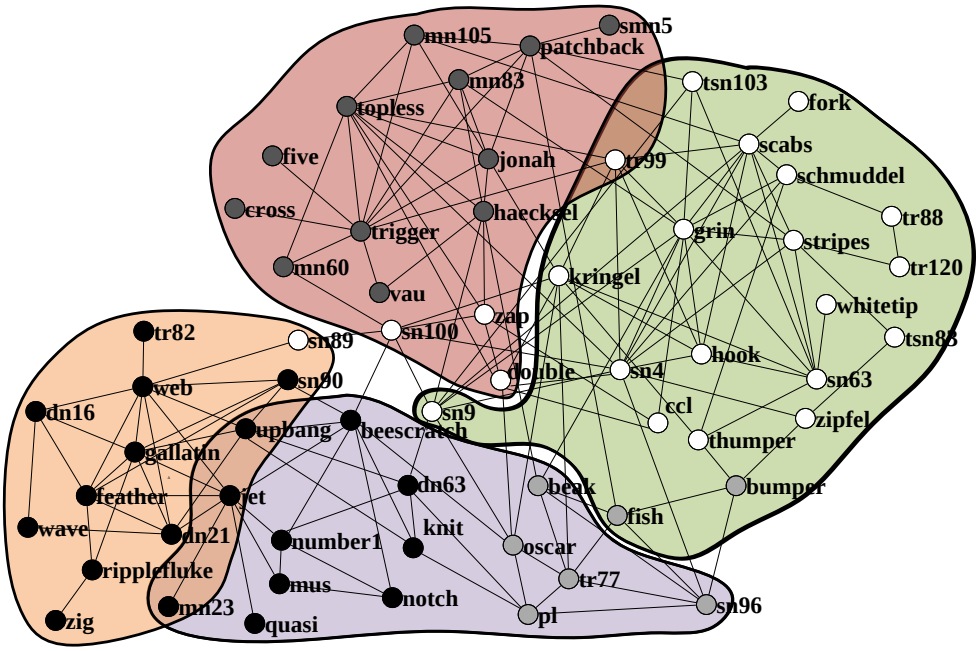



Figure 3.12: Dolphin network overlapping communities: $k = 4$ and $\lambda = .27$

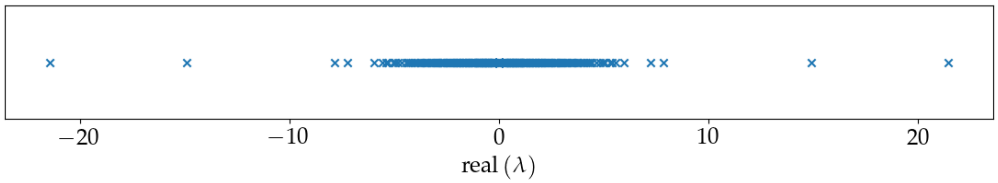



Figure 3.13: Alaska Spectral Characteristic

strategic donations [78]. This premise is tested here with hierarchical fuzzy spectral clustering. This particular set is compiled from donations that were reported in Alaska during 2012 for general elections and represents business and other non-individual donations to candidates, creating a bipartite network where each node is a candidate or donor and an edge is a donation.

In preparing the dataset, vertices were removed if they only contributed once over the course of the election cycle. Similarly, candidates with only one donor were also removed. Multiple donations from a donor to the same candidate were collapsed into a single edge. The remaining data cover 214 nodes and 1426 edges. Despite the simplicity of the graph here, the full scope of information available for creating a fully featured and heterogeneous network from campaign finance is substantial and growing rapidly. Figure 3.13 shows the spectral characteristic of this network. It is rather similar in nature to the Zachary karate network in that it indicates two clusters at the top hierarchy and four clusters at a second hierarchical level. The large negative eigenvalues on the left of the graph are due to the bipartite nature of the graph and can be ignored.

After obtaining the hierarchy from the above method, there are two communities at the top level with each having two child communities. Unsurprisingly, in the parent communities **D** and **R**, the candidates mostly split on party lines, and ideology appears to dominate donations. The overlap between the two groups is especially

Table 3.1: Historical Alaska Donations

| Nodes | Ratio to Winners | Std. Dev. of % |
|---|---|---|
| In Both **D** and **R** | 4.94 | 0.094 |
| Solely in **D** | 1.20 | 0.232 |
| Solely in **R** | 2.80 | 0.182 |

interesting, however, as it includes donors who gave evenly between Democrats and Republicans in Alaska. Moreover, the candidates in this overlap were overwhelmingly winners, with only one candidate losing the election.

To verify these results, the communities were checked against their entire historical data. Based on this data from the National Institute on Money in Politics, these overlap donors have on average given much more to winning candidates with very little variation. The rest of the donors have not done as well at donating solely to winners, though there is far more variation in the percentage of dollars that went to winning candidates, as shown in Table 3.1.

Analyzing the sub-communities at the next hierarchy, there is a clear pattern in the candidates within each community. Analyzing each community separately, we find:

- $\mathbf{D}_1$ comprises exclusively Democratic candidates, 83% of which lost the election. The donors have previously given to Democrats, with only one donation ever to a Republican candidate as well as one to an unaffiliated candidate.

- $\mathbf{D}_2$ comprises mostly Democratic candidates at 88%, 56% of which won their election. The donors have given almost four times as much to Democrats as Republicans.

- $\mathbf{R}_1$ comprises 10 Democratic and 28 Republican candidates. These candidates were almost exclusively winners, with only one losing. Similar to $\mathbf{D}_2$, the donors gave four times as much, in this case favoring Republicans.

- $\mathbf{R}_2$ comprises only Republican candidates as well as a single unaffiliated candidate. Only 55% of these candidates won the election. The donors in this group have, over the years, given over 54 times as much to Republicans as Democrats.

For the children of $\mathbf{D}$, those who gave exclusively to Democrats generally gave to losing candidates while those who gave more evenly donated more to winners. Regarding the children of $\mathbf{R}$, while the donors who gave exclusively to Republicans chose more winners, those who gave to Democrats as well picked almost nothing but winning candidates. This shift may be due to the overall political leanings of Alaska where their legislature has a majority of Republicans.

### 3.2.4 All States

With the success of Alaska as a baseline, we continued by analyzing all the other state communities in 2012. For context, Figure 3.14 shows the proportions of money given to recipients over all the NIMP data. By ratio, donations to Democrats and Republicans were split evenly in the donations, with Republicans slightly favored at a ratio of 1.089 compared to Democrats at 0.918. The difference is much larger when considering winning candidates or incumbents. In the data, winning candidates raised 1.632 times more money than losing candidates. The number is similar for incumbents, where incumbents raised 1.561 times more money than non-incumbents. Deviations from these ratios could indicate different patterns of donations among the communities.

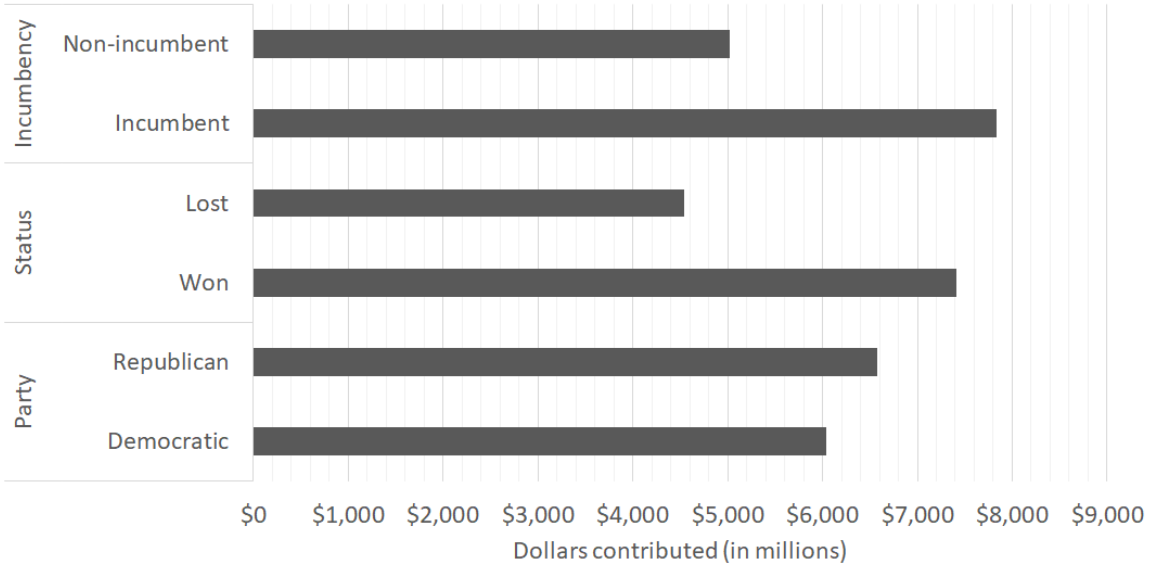



Figure 3.14: Sum of Donations to Recipients by Type

Datasets for each of these states was created in the same manner as the data for Alaska. Each graph is made of non-individual donations in general elections during 2012. Any node $i$ where $d_i = 1$ is removed from the network. Table 3.2 shows the resulting sizes of the graphs for each of the state networks. The following tables give donation amounts and ratios for $k = 2$ and $k = 4$ communities for each of the states. For all communities, a threshold of 0.3 is used for determining if a node is part of a fuzzy community.

Using California as another example, Table 3.3 shows the donation history by members of each community. With the donors in each community at $k = 2$, there are different ratios of their giving based on properties of the recipients. From the results, community $c_{2,0}$ donated more to Democrats than Republicans, but not by a large ratio at 1.342. For both Incumbency and Status, the giving ratios are much higher showing this group of donors has historically given far more to incumbents and winners at 3.277 and 4.606 respectively. These ratios are higher than the overall

Table 3.2: Network Sizes using Non-Individual Donations

| State | Nodes | Edges | State | Nodes | Edges | State | Nodes | Edges |
|---|---|---|---|---|---|---|---|---|
| AK | 214 | 1,426 | LA | 297 | 816 | OH | 1,126 | 9,000 |
| AL | 97 | 204 | MA | 607 | 3,198 | OK | 442 | 4,826 |
| AR | 756 | 6,809 | MD | 523 | 2,365 | OR | 738 | 8,233 |
| AZ | 225 | 1,201 | ME | 249 | 1,067 | PA | 1,397 | 13,532 |
| CA | 1,964 | 17,171 | MI | 901 | 8,494 | RI | 331 | 2,003 |
| CO | 467 | 5,109 | MN | 797 | 4,486 | SC | 884 | 5,670 |
| CT | 68 | 69 | MO | 1,768 | 14,956 | SD | 351 | 2,766 |
| DE | 432 | 2,698 | MS | 69 | 151 | TN | 515 | 6,455 |
| FL | 3,351 | 24,779 | MT | 491 | 2,747 | TX | 2,163 | 24,223 |
| GA | 1,097 | 11,653 | NC | 834 | 7,472 | UT | 469 | 4,579 |
| HI | 288 | 1,890 | ND | 282 | 1,160 | VA | 497 | 4,466 |
| IA | 535 | 5,871 | NE | 217 | 1,481 | VT | 258 | 733 |
| ID | 487 | 4,665 | NH | 303 | 1,640 | WA | 1,369 | 14,260 |
| IL | 2,038 | 18,051 | NJ | 662 | 3,504 | WI | 634 | 3,740 |
| IN | 1,157 | 8,289 | NM | 661 | 6,932 | WV | 387 | 3,466 |
| KS | 843 | 10,327 | NV | 688 | 6,351 | WY | 165 | 1,095 |
| KY | 379 | 2,949 | NY | 2,318 | 14,069 | – | – | – |

underlying data where incumbents raise 1.561 times more than non-incumbents and winning candidates raise 1.632 times more than losing candidates. Community $c_{2,1}$ donated at a higher ratio to Democrats than Republicans at 1.678 when compared to the other community. The ratio of incumbents vs. non-incumbents and winners vs. losers is lower at 1.47 and 2.09 respectively, showing this group overall prioritized party more so than the other.

Table 3.3: Donor History by Community in CA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | \$462,193,658 | 1.342 | I | \$622,936,188 | 3.277 | Lost | \$134,959,301 | 0.217 |
| $c_{2,0}$ | R | \$344,442,457 | 0.745 | N | \$190,093,268 | 0.305 | Won | \$621,631,277 | 4.606 |
| $c_{2,1}$ | D | \$843,949 | 1.678 | I | \$948,145 | 1.470 | Lost | \$515,599 | 0.479 |
| $c_{2,1}$ | R | \$502,919 | 0.596 | N | \$644,964 | 0.680 | Won | \$1,077,510 | 2.090 |
| $c_{2,0} \cap c_{2,1}$ | D | \$185,200 | 2.750 | I | \$181,700 | 2.374 | Lost | \$39,750 | 0.182 |
| $c_{2,0} \cap c_{2,1}$ | R | \$67,350 | 0.364 | N | \$76,550 | 0.421 | Won | \$218,500 | 5.497 |
| $c_{4,0}$ | D | \$17,600 | 0.140 | I | \$109,638 | 0.542 | Lost | \$198,975 | 1.760 |
| $c_{4,0}$ | R | \$125,745 | 7.145 | N | \$202,375 | 1.846 | Won | \$113,038 | 0.568 |
| $c_{4,1}$ | D | \$164,844,588 | 16.600 | I | \$107,950,768 | 1.578 | Lost | \$40,148,123 | 0.308 |
| $c_{4,1}$ | R | \$9,930,533 | 0.060 | N | \$68,428,242 | 0.634 | Won | \$130,421,379 | 3.249 |
| $c_{4,2}$ | D | \$657,249 | 2.023 | I | \$684,475 | 1.760 | Lost | \$287,374 | 0.366 |
| $c_{4,2}$ | R | \$324,824 | 0.494 | N | \$388,839 | 0.568 | Won | \$785,940 | 2.735 |
| $c_{4,3}$ | D | \$377,947,946 | 1.107 | I | \$570,523,492 | 3.691 | Lost | \$112,038,400 | 0.200 |
| $c_{4,3}$ | R | \$341,469,715 | 0.903 | N | \$154,565,697 | 0.271 | Won | \$559,562,532 | 4.994 |

Table 3.4: Individual Preference by Community in CA

| **C** | $\sum$ Party | $\overline{Party}$ | $\sum Inc$ | $\overline{Inc}$ | $\sum Status$ | $\overline{Status}$ |
|---|---|---|---|---|---|---|
| $c_{2,0}$ | $-\$117,751,201$ | $-0.258$ | \$432,842,921 | 0.251 | \$486,671,976 | 0.517 |
| $c_{2,1}$ | $-\$341,030$ | $-0.045$ | \$303,181 | 0.294 | \$561,911 | 0.505 |
| $c_{2,0} \cap c_{2,1}$ | $-\$117,850$ | $-0.580$ | \$105,150 | 0.023 | \$178,750 | 0.696 |
| $c_{4,0}$ | \$108,145 | 0.321 | $-\$92,737$ | 0.232 | -\$85,937 | 0.343 |
| $c_{4,1}$ | $-\$154,914,055$ | $-0.669$ | \$39,522,526 | $-0.013$ | \$90,273,257 | 0.413 |
| $c_{4,2}$ | $-\$332,425$ | $-0.323$ | \$295,636 | 0.368 | \$498,566 | 0.647 |
| $c_{4,3}$ | $-\$36,478,231$ | $-0.175$ | \$415,957,795 | 0.288 | \$447,524,132 | 0.533 |

Moving down the hierarchy to $k = 4$, other behaviors become apparent. Community $c_{4,1}$ heavily favors Democrats at a ratio of 16.6, far higher than any of the other communities listed in CA. In contrast, community $c_{4,4}$ donates almost equally to Democrats and Republicans, but heavily favors both incumbents and

winning candidates, possibly indicating this group is more interested in political access than left-right ideology. Communities $c_{4,0}$ and $c_{4,2}$ are smaller donors that favor Republicans and Democrats respectively.

One important consideration is that these numbers could be dominated by groups who have greater ability to spend large amounts of money. This could cause individual preferences to be lost. Instead of using the raw numbers, we considered a rescaling of the data such that it forms a preference to each type of recipient as a range in $[-1, 1]$. For each feature of a recipient, the data is scaled such that Party $= D \to -1$ and Party $= R \to +1$. A donor who gives exclusively to Democrats would have a preference of $-1$ for party, while a donor who gives exclusively to Republicans would have a preference of 1 for party. Donors who give equally would have a preference of 0. The data is scaled similarly by incumbency (Inc $= N \to -1$, Inc $= I \to +1$) and status (Status $= L \to -1$, Status $= W \to +1$).

With these definitions, Table 3.4 shows the results of scaling individual preference within each community. In the tables listing the scaled preferences, $\sum$Party is the sum of donations where donations to Democratic candidates are multiplied by $-1$. Similarly, $\sum$Inc and $\sum$Status refer to sums of the totals to those features where donations to non-incumbents are multiplied by $-1$ and donations to losers are also multiplied by $-1$. The columns $\overline{\text{Party}}$, $\overline{\text{Inc}}$, and $\overline{\text{Status}}$ are the average scaled preference. Based on this scaling, donors in communities $c_{2,0}$ and $c_{2,1}$ donate to winning candidates at a much higher preference than to either party or incumbency. The donors in $c_{2,1}$ on average have very little preference for either party, whereas $c_{2,0}$ show preference for Democratic candidates. Switching to a lower hierarchical level, communities $c_{4,1}$ and $c_{4,3}$ are the children of $c_{2,0}$. For those, the donors in $c_{4,1}$ show a much stronger party preference than the sibling donors from $c_{4,3}$. Instead, the donors from $c_{4,3}$ have a much higher preference for incumbents, and their candidates more

Table 3.5: Donor History by Community in DE

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | \$49,302,417 | 0.761 | I | \$100,548,273 | 6.991 | Lost | \$14,002,157 | 0.157 |
| $c_{2,0}$ | R | \$64,780,247 | 1.314 | N | \$14,383,165 | 0.143 | Won | \$88,960,053 | 6.353 |
| $c_{2,1}$ | D | \$83,628,947 | 1.331 | I | \$122,664,162 | 4.960 | Lost | \$21,381,708 | 0.192 |
| $c_{2,1}$ | R | \$62,828,631 | 0.751 | N | \$24,732,871 | 0.202 | Won | \$111,164,213 | 5.199 |
| $c_{2,0} \cap c_{2,1}$ | D | \$35,749,106 | 0.903 | I | \$67,306,486 | 7.946 | Lost | \$8,378,331 | 0.140 |
| $c_{2,0} \cap c_{2,1}$ | R | \$39,582,378 | 1.107 | N | \$8,470,122 | 0.126 | Won | \$59,706,953 | 7.126 |
| $c_{4,0}$ | D | \$56,800,432 | 0.925 | I | \$103,544,029 | 6.689 | Lost | \$14,417,455 | 0.156 |
| $c_{4,0}$ | R | \$61,379,789 | 1.081 | N | \$15,480,330 | 0.150 | Won | \$92,164,854 | 6.393 |
| $c_{4,1}$ | D | \$22,841,284 | 14.945 | I | \$16,062,871 | 1.866 | Lost | \$6,761,364 | 0.437 |
| $c_{4,1}$ | R | \$1,528,349 | 0.067 | N | \$8,606,722 | 0.536 | Won | \$15,472,932 | 2.288 |
| $c_{4,2}$ | D | \$41,407,115 | 0.828 | I | \$80,405,994 | 6.950 | Lost | \$10,845,785 | 0.151 |
| $c_{4,2}$ | R | \$50,005,976 | 1.208 | N | \$11,569,824 | 0.144 | Won | \$71,770,034 | 6.617 |
| $c_{4,3}$ | D | \$896,352 | 0.195 | I | \$3,388,763 | 1.555 | Lost | \$1,855,280 | 0.546 |
| $c_{4,3}$ | R | \$4,604,434 | 5.137 | N | \$2,179,473 | 0.643 | Won | \$3,396,076 | 1.830 |

Table 3.6: Individual Preference by Community in DE

| **C** | $\sum$ Party | $\overline{Party}$ | $\sum Inc$ | $\overline{Inc}$ | $\sum Status$ | $\overline{Status}$ |
|---|---|---|---|---|---|---|
| $c_{2,0}$ | \$15,477,830 | 0.244 | \$86,165,108 | 0.239 | \$74,957,897 | 0.378 |
| $c_{2,1}$ | $-\$20,800,316$ | $-0.691$ | \$97,931,292 | 0.351 | \$89,782,505 | 0.501 |
| $c_{2,0} \cap c_{2,1}$ | \$3,833,272 | $-0.339$ | \$58,836,363 | 0.568 | \$51,328,623 | 0.639 |
| $c_{4,0}$ | \$4,579,357 | $-0.462$ | \$88,063,698 | 0.644 | \$77,747,398 | 0.629 |
| $c_{4,1}$ | $-\$21,312,935$ | $-0.911$ | \$7,456,149 | $-0.053$ | \$8,711,568 | 0.277 |
| $c_{4,2}$ | \$8,598,861 | 0.023 | \$68,836,171 | 0.520 | \$60,924,249 | 0.609 |
| $c_{4,3}$ | \$3,708,082 | 0.720 | \$1,209,290 | $-0.257$ | \$1,540,796 | 0.050 |

often win their election. Communities $c_{4,0}$ and $c_{4,2}$ show a similar split where donors in $c_{4,0}$ favor winning Republicans, but $c_{4,2}$ strongly favors winning incumbents with party being the weakest preference.

The same analysis can be done for other states, such as Delaware. Table 3.5

contains the donation history for each of the donors in the found communities. In this case, the top two communities $c_{2,0}$ and $c_{2,1}$ both have high ratios of donations to winning incumbents, far higher than the database average. They differ in that $c_{2,0}$ favors Republicans while $c_{2,1}$ favors Democrats. However, the overall giving patterns of each of their child communities reveal some differences. Donors who were in both communities had low overall ratios by party, but even higher ratios to incumbents and winners. The child communities of $c_{2,0}$ are $c_{4,0}$ and $c_{4,1}$. In total, the donors in $c_{4,0}$ gave at roughly the database average by party; however, they gave at a considerably higher ratio to incumbents and winning candidates. Donors in $c_{4,1}$ heavily favored Democrats at a ratio of 14.945 over Republican candidates. The ratios to incumbents and winners was quite a bit lower than those of its sibling communities $c_{4,0}$. A similar result is found with the children of $c_{2,1}$: $c_{4,2}$ and $c_{4,3}$. Community $c_{4,3}$ is characterized by a high ratio of party giving to Republicans at 5.137. The ratios to incumbents and winners is closer to the database average. Community $c_{4,2}$ has a party ratio closer to the database average at 1.208 to Republicans, while the ratio to incumbents and winners is far higher than the average at 6.950 and 6.617 respectively.

The results for individual preference reinforce the overall community results but do provide some additional information. Community $c_{2,0}$ shows a high ratio of giving to winning incumbents. However, the individual preferences for incumbency and status are not much different than the party preferences. This can indicate the ratios of giving in community $c_{2,0}$ are high in part due to large donors. Preference was highest to winning candidates at 0.378, with party and incumbency effectively even. Community $c_{2,1}$ had a large preference for Democratic candidates at $-0.691$, with winners also having high preference at 0.501. For those donors in both $c_{2,0}$ and $c_{2,1}$, the party preference favors Democrats, but had higher overall average preference for status and incumbency at 0.639 and 0.568 respectively, indicating these donors did

Table 3.7: Donor History by Community in PA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $149,638,238 | 0.610 | I | $340,136,508 | 5.966 | Lost | $56,734,601 | 0.191 |
| $c_{2,0}$ | R | $245,213,271 | 1.639 | N | $57,010,454 | 0.168 | Won | $297,486,327 | 5.243 |
| $c_{2,1}$ | D | $151,624,395 | 4.616 | I | $118,112,859 | 1.754 | Lost | $53,605,945 | 0.463 |
| $c_{2,1}$ | R | $32,849,300 | 0.217 | N | $67,330,689 | 0.570 | Won | $115,780,389 | 2.160 |
| $c_{2,0} \cap c_{2,1}$ | D | $16,885,302 | 0.919 | I | $27,376,235 | 3.408 | Lost | $7,477,733 | 0.302 |
| $c_{2,0} \cap c_{2,1}$ | R | $18,374,434 | 1.088 | N | $8,033,151 | 0.293 | Won | $24,780,565 | 3.314 |
| $c_{4,0}$ | D | $60,800,441 | 12.011 | I | $37,685,149 | 1.318 | Lost | $23,095,177 | 0.608 |
| $c_{4,0}$ | R | $5,062,144 | 0.083 | N | $28,598,074 | 0.759 | Won | $37,992,954 | 1.645 |
| $c_{4,1}$ | D | $117,777,822 | 6.056 | I | $89,002,626 | 1.816 | Lost | $38,969,172 | 0.452 |
| $c_{4,1}$ | R | $19,447,071 | 0.165 | N | $49,018,365 | 0.551 | Won | $86,240,063 | 2.213 |
| $c_{4,2}$ | D | $137,081,746 | 0.671 | I | $303,275,194 | 7.575 | Lost | $42,692,901 | 0.163 |
| $c_{4,2}$ | R | $204,334,834 | 1.491 | N | $40,037,013 | 0.132 | Won | $262,332,319 | 6.145 |
| $c_{4,3}$ | D | $34,699,562 | 0.396 | I | $98,945,070 | 4.062 | Lost | $23,663,560 | 0.273 |
| $c_{4,3}$ | R | $87,730,404 | 2.528 | N | $24,356,882 | 0.246 | Won | $86,731,143 | 3.665 |

not primarily favor party affiliation. Results for the child communities show similar results to the group dynamic where preference for party in $c_{4,0}$ trends Democrats but has stronger preference for incumbents and winners. Its sibling $c_{4,1}$ has an extreme preference for Democrats at $-0.911$, but far less overall preference for incumbency or status. Communities $c_{4,3}$ and $c_{4,2}$ show similar behavior where $c_{4,3}$ has strong party preference, and $c_{4,2}$ has incumbent and winner preference.

Pennsylvania shows some similar patterns for donations (Table 3.7). The top two communities show some different behavior where $c_{2,0}$ only slightly favors donations to Republicans, but it has high ratio to incumbents and winners. In contrast, community $c_{2,1}$ has a high ratio to Democrats and relatively lower ratios for incumbents or winners. Those donors in both communities gave almost evenly by party, but had higher ratios for incumbents and winners. Moving to the children, only $c_{4,1}$ is a child of $c_{2,1}$, with those donors remaining in this community giving to Democrats at a

Table 3.8: Individual Preference by Community in PA

| **C** | $\sum$ Party | $\overline{Party}$ | $\sum Inc$ | $\overline{Inc}$ | $\sum Status$ | $\overline{Status}$ |
|---|---|---|---|---|---|---|
| $c_{2,0}$ | \$95,575,033 | 0.428 | \$283,126,055 | 0.455 | \$240,751,726 | 0.532 |
| $c_{2,1}$ | $-\$118,775,095$ | $-0.709$ | \$50,782,170 | 0.168 | \$62,174,443 | 0.338 |
| $c_{2,0} \cap c_{2,1}$ | \$1,489,132 | $-0.271$ | \$19,343,083 | 0.463 | \$17,302,832 | 0.474 |
| $c_{4,0}$ | $-\$55,738,297$ | $-0.907$ | \$9,087,075 | $-0.101$ | \$14,897,777 | 0.171 |
| $c_{4,1}$ | $-\$98,330,750$ | $-0.722$ | \$39,984,260 | 0.347 | \$47,270,891 | 0.463 |
| $c_{4,2}$ | \$67,253,089 | 0.253 | \$263,238,181 | 0.669 | \$219,639,418 | 0.695 |
| $c_{4,3}$ | \$53,030,841 | 0.807 | \$74,588,188 | 0.261 | \$63,067,583 | 0.413 |

slightly higher ratio than its parent, but similar ratios for the other two features. In contrast, the other three communities are all children of $c_{2,0}$ and show considerably different behavior than their parent. In $c_{4,0}$, these donors gave to Democrats at a ratio of 12.011, far higher than any other community listed. Communities $c_{4,2}$ and $c_{4,3}$ gave more to Republicans, but $c_{4,2}$ had much higher ratios to incumbents and winners.

Individual preference results for Pennsylvania show the average preferences for donors in community $c_{2,0}$ are strong but close to each other for all of the features. Community $c_{2,1}$ has a very strong preference for Democrat candidates, with far smaller magnitude of preference for incumbents and winners. As expected, its child $c_{4,1}$ shows the same preferences. The children of $c_{2,0}$ have far different preferences from each other. Communities $c_{4,0}$ and $c_{4,3}$ have extreme preference for parties at $-0.907$ and 0.807 respectively. In contrast, community $c_{4,2}$ has incumbents and winners at a much higher preference with 0.669 and 0.695 respectively. The same analysis can be done for each of the remaining states. Appendix A shows the community and preference tables for each of the fifty states.

## 3.3 Conclusion

In this chapter, we introduced a new hierarchical fuzzy spectral clustering algorithm for finding communities in social networks. This HFSC algorithm was shown to be effective in finding meaningful communities within social networks. We used an outlier detection method on the spectral characterization to find an appropriate number of clusters $k$. We showed the vertices in the overlap of clusters found by HFSC showed different behavior than the vertices strictly in a single community. On benchmark datasets, the results of HFSC yielded results consistent with known ground truths for those communities. Applying the algorithm to state campaign finance networks, we obtained fuzzy community structure for each of the fifty states. The results show the community structure is useful in identifying differing behavior among the donors of those communities. It was possible to find information in the overlapping communities where the donors showed different patterns of giving than their siblings within the communities. Additionally, the hierarchical information was also useful to get additional information and further categorize donor behavior. This was shown by the overall group donations, as well as preference by each donor within the communities. The analysis was expanded to all fifty states, highlighting the patterns of donation in communities at various levels in the hierarchy.

# CHAPTER FOUR

# TEMPORAL COMMUNITIES

In this chapter we augment the hierarchical fuzzy spectral algorithm given in Chapter 3 to track communities in evolving social networks. As social networks are representative of the interactions of individuals, as those interactions change the structure of the network should also change. Tracking the communities as the graph changes can be important in understanding how the behavior of communities may change over time. By augmenting HFSC, it is possible to connect communities found at adjacent timesteps. We can then analyze how the behavior of individuals and communities may change.

## 4.1 Background

In the previous section we analyzed communities generated from a single year of campaign finance donations. However, these are networks that change over time. Different candidates are up for election each year. New donors may start donating, some may stop, or they give to different recipients in different ratios. In this section we discuss some prior work performed in trying to discover communities that may change over time.

Much like how early work in community discovery focused on finding crisp communities, early work on networks did not concern tracking communities through time. Early attempts at tracking the communities built upon that work and used static community discovery on crisp communities at different timesteps in order to find communities through time. Some prior work highlighted an issue with static community detection methods in dynamic graphs [37], [92]. Specifically, it was shown

that trying to connect crisp communities discovered at different timesteps can be unstable. Using different crisp community detection algorithms, the authors showed how modularity and communities change when nodes are removed from the graph. As nodes were removed from the network, the modularity of the community assignments found after removal would increase. This would increase up to a certain point when too many nodes were removed, and the modularity sharply fell. Additional analysis on the communities showed that the set of vertices $\mathbf{V}$ in a community $\mathbf{C}_i$ would change dramatically as nodes were removed. This was calculated by the edit distance of the community assignments at each time step.

More recent work attempts to more comprehensively create a model to describe communities through time [93]. The authors developed a dynamic Bayesian model named Dynamic Bayesian Non-negative Matrix Factorization. In their formulation, they define a temporal network as $\mathbf{G} = \{\mathbf{E}_1, \ldots, \mathbf{E}_T\}$ where $T$ is the number of snapshots of the networks. In this model, the number of communities at each time step are determined by an automatic relevance determination. Each snapshot of a graph at $t$ utilizes information from the previous timestep in the community detection.

## 4.2 Approach

One issue with campaign finance networks is that it is not possible to have a smoothly changing network. The datasets available typically are available in two-year increments. The underlying dynamics change dramatically over that time frame as new elections occur. Based on this, we find communities at each of those two-year increments. Then a mutual information criterion based on the fuzzy assignment scores is used to matching communities through time.

The overall approach is similar to the hierarchical fuzzy spectral clustering defined in Chapter 3. Treating each two-year cycle as its own network, first

**Algorithm 4.1** Fuzzy Spectral Clustering

1: **function** FSC($\mathbf{A}, k$)
2: $\mathbf{D} \leftarrow \{d_i = \sum_1^n a_{ij}\}$ ▷ Create the diagonal matrix $\mathbf{D}$ from the row sum of $\mathbf{A}$.
3: $\mathbf{L}_{\text{sym}} = \mathbf{I}^{n \times n} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ▷ Calculate the normalized symmetric Laplacian.
4: $\mathbf{Z} = \text{eigenvectors}(\mathbf{L}, k)$ ▷ Compute the top-$k$ eigenvectors.
5: $\mathbf{X} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]$ ▷ Create matrix $\mathbf{X}$ where column $i$ in $\mathbf{X}$ is eigenvector $\mathbf{z}_i$.
6: **for all** row $\mathbf{x}_i \in \mathbf{X}$ **do**
7: $x_{ij} = x_{ij} / \|\mathbf{x}_i\| \; \forall \, x_{ij} \in \mathbf{x}_i$ ▷ Normalize each row of matrix $\mathbf{X}$.
8: **end for**
9: $\mathbf{U} = \text{FCM}(\mathbf{X})$ ▷ Use fuzzy c-means to get cluster assignments.
10: **return U**
11: **end function**

we determine the eigenvectors of the graph Laplacian $\mathbf{L}_{sym}$ where $\mathbf{L}_{sym} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. Starting with $k = 2$, we use fuzzy c-means clustering on the eigenvectors with the smallest eigenvalues. The hierarchical structure is constructed by repeating the process with increasing $k$. As before, the communities are connected to parents using a fuzzy Jaccard similarity. For convenience, the algorithms are shown again in Algorithms 4.1 and 4.2.

### 4.2.1 Retention Rates

In addition to the static networks, additional work is done in analyzing different networks for each of the election cycles and connecting communities between years. To aid in this, we want to determine the ratio of actors in the network who participate in successive years. This estimate of the level of continuous political participation over

**Algorithm 4.2** Hierarchical Generation

1: **function** HFSC($\mathbf{A}, k$)
2: **for** $i = 2$ to $k$ **do**
3: $C_i = \text{FSC}(\mathbf{A}, i)$ ▷ Find the clusters for each level in the hierarchy.
4: **end for**
5: **for all** $C_{i,m} \in C_i : 3 \leq i < k$ **do**
6: $P_{i,m,n} = \arg\max \mathcal{J}_f(C_{i,m}, C_{i-1,n})$ ▷ Find parent $P_{i,m,n}$ for each $\mathbf{C}_{i,m}$
7: **end for**
8: **end function**

time for nodes in the network we call the retention rate of the network. For any state and year, consider the set of vertices $\mathbf{V}_t$ and $\mathbf{V}_{t+1}$ for election cycles $t$ and $t+1$. We use Jaccard similarity as a metric for the retention rate is defined as the

$$\mathcal{J}(\mathbf{V}_t, \mathbf{V}_{t+1}) = \frac{|\mathbf{V}_t \cap \mathbf{V}_{t+1}|}{|\mathbf{V}_t \cup \mathbf{V}_{t+1}|}.$$

This gives a metric for the number of entities that are in both adjoining years. As mentioned above, any set of vertices $\mathbf{V}$ is first limited by removing those who only have one link. With these datasets, the average retention rate of entities in those networks is 20.90% with a median of 21.66%. Alabama is the only notable outlier. Its retention rate is only 6.52%. Removing this state from consideration, the average retention rate for the other states moves very close to the median.

### 4.2.2 Tracking Communities

The above algorithms aid in understanding static networks, either for all of a state or individual election cycles. However, we are also interested in how communities may change over time. The addition to track communities through time is given in

Algorithm 4.3. For this procedure, HFSC is calculated on each timestep $t$ of the networks to obtain the set of hierarchical communities $\mathbf{U}'_1, \ldots, \mathbf{U}'_T$.

Beginning with $k_i = 2$ for time step $i$, connections are made to time step $i+1$ by searching down the tree for matching communities. Then, for each community in dataset $i$ at $k_o = 2$, we iterate through the hierarchies of dataset $i+1$, also starting with $k_{i+1} = 2$ to find the first hierarchy wherein there exists a community with a fuzzy Jaccard similarity measure $\mathcal{J}_f$ greater than a threshold. In this case, the threshold is determined by the retention rate as it gives an estimate of the number of entities within the community that can possibly match. From this, the fuzzy Jaccard similarity between time steps must be greater than $\mathcal{J}(D_i, D_{i+1})/2$. This process is repeated for the other hierarchy levels with the requirement that a child community cannot be linked to a hierarchy level equal to or above its parent.

## 4.3 Political Donation Networks

The dataset used in this section is different from the data used in Chapter 3. The prior data was obtained from the National Institute on Money in Politics. This was used to determine non-individual donation patterns based on communities found in the campaign network. However, using augmented data from other areas may help in improving the community analysis. The Bonica and the Stanford Social Science Data Collection has an augmented dataset [78, 94]. This dataset combines data from the Federal Election Commission, the Center for Responsive Politics, the National Institute for Money in Politics, the Sunlight Foundation, and other reporting agencies. This combination of data is segmented within each state (and federal) in two-year election cycles spanning 1979 through 2012.[1]

[1]The dataset was updated to include more recent years; however, we continue to use the 1979–2012 subset. Increasing polarization has shown the prediction problem becomes easier over time.

**Algorithm 4.3** Connecting Communities through Time

1: **function** TEMPORALLINK($\mathbf{U}'_1, \ldots, \mathbf{U}'_T$)
2: **for** $t = 1$ to $T - 1$ **do**
3: **for** $\mathbf{C}_{ij} \in \mathbf{U}'_t$ **do** ▷ For each community at time step $t$
4: **for** $\mathbf{C}_{ik} \in \mathbf{U}'_{t+1}$ **do** ▷ For each community at time step $t + 1$
5: **if** $\mathcal{J}_j(\mathbf{C}_{ij}, \mathbf{C}_{ik}) \geq \frac{\mathcal{J}(\mathbf{V}_j \in \mathbf{U}'_t, \mathbf{V}_k \in \mathbf{U}'_{t+1})}{2}$ **then**
6: $\mathbf{M}_t \leftarrow (\mathbf{C}_{ij}, \mathbf{C}_{ik})$ ▷ Add a link between communities if similarity over the threshold.
7: **end if**
8: **end for**
9: **end for**
10: **end for**
11: **return M**
12: **end function**

In performing the analysis for CFScores, Bonica placed certain restrictions upon the dataset. Similar restrictions are applied to the networks here for each combination of state and year. Specifically, an entity is only included as a node in the network if it has at least two connections to other entities. Only direct or in-kind contributions are included. Since loans and similar records do not necessarily indicate support of a candidate, they are removed from consideration. As the data for each state are poorer in earlier years and the starting point is considerably different for each state, the following analysis uses only the data for elections cycles 2004 through 2012 in order to ensure each state has the same number of years of data. It is possible within the data for donors and recipients to be entirely disconnected from the rest of the network. Because the spectral clustering views the whole network, isolated entities

that are not connected to the larger network are also removed from consideration.

Separate files for the list of donors and recipients are also provided. While the donor data is sparse, the recipient data includes information, such as the state, district, seat, party, and other pertinent information regarding a candidate running for office. Further improvements by Bonica include some entity resolution to improve the connections between years and states. The entity resolution assigns unique identifiers to the candidates and donors across states and years.

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 through time. Note that 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 states 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 dataset due to the increase in contributions surrounding recall and gubernatorial races.

### 4.3.1 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, we can check the communities against CFScore for validity.

Table 4.1: Correlation of CFScore and Communities at $k = 2$

| Cycle | AK | WI | NY |
|---|---|---|---|
| 2004 | 0.8984 | 0.9196 | 0.121 |
| 2006 | 0.9162 | 0.9457 | 0.6139 |
| 2008 | 0.9126 | 0.9233 | 0.0173 |
| 2010 | 0.9226 | 0.9427 | 0.1207 |
| 2012 | 0.9057 | 0.9808 | 0.0575 |

CFScores 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 CFScore 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 CFScore 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 CFScore for that entity. Table 4.1 shows the correlations for each of the cycles for all entities. For Alaska, these fuzzy memberships are highly correlated with the CFscore.

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 4.2 shows the average CFScore 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

Table 4.2: CFScore of Alaska Communities

| | $k=2$ | | $k=3$ | | |
|---|---|---|---|---|---|
| Year | $C_1$ | $C_2$ | $C_1$ | $C_2$ | $C_3$ |
| 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 |

of these communities. The results of this correspond to prior political science work indicating an increase in partisanship over the years [78].

Moving down the hierarchy, similar results are obtained for $k = 3$, also shown in Table 4.2. For all but one year, every community at $t_i$ continued into $t_{i+1}$. The community $C_2$ at year 2004 did not have a corresponding community in year 2006 based on the threshold set by the fuzzy Jaccard similarity of the community and the retention rate. At this breakdown, the average CFScore 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.

### 4.3.2 New York

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 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 4.1. 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 who mostly lost were separated from the rest of the network by having similar donor groups, but being considerably different than the rest of the candidates. Comparing the recipients with their candidate information indicates there is no significant correlation between party or district. This indicates that, for New York, ideology is not actually the most dominant factor in determining the pattern of donations.

### 4.3.3 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 CFScore 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 CFScore, shown in Table 4.1. 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 4.3. 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 4.3. 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 CFScore, WI appears to have a considerable, and consistent, set of moderates.

Table 4.3: CFScore of Wisconsin Communities

| | $k=2$ | | $k=3$ | | |
|---|---|---|---|---|---|
| Year | $C_1$ | $C_2$ | $C_1$ | $C_2$ | $C_3$ |
| 2004 | -0.935 | 0.759 | -0.950 | 0.229 | 0.872 |
| 2006 | -0.990 | 0.853 | -1.117 | -0.019 | 0.916 |
| 2008 | -0.986 | 0.728 | -1.084 | -0.134 | 0.780 |
| 2010 | -0.865 | 0.977 | -1.084 | 0.255 | 1.075 |
| 2012 | -1.353 | 1.079 | -1.370 | -0.029 | 1.113 |

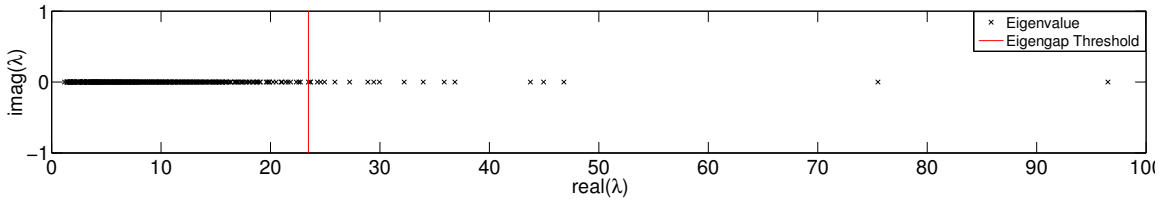



Figure 4.1: Eigenvalues for Alaska Contribution Network, All Years, All Data

## 4.4 Detailed Results

### 4.4.1 Alaska

For the state databases, Alaska showed the highest retention rate of entities from year to year at 29.67% on average. First, communities were 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 resulted in $12,417$ entities and $66,629$ edges. Figure 4.1 shows the eigen-spectrum for this dataset, and the red line indicates the point at which the eigengap first exceeds the average absolute deviation.

Figure 4.2 shows the community assignment values to one of the two found

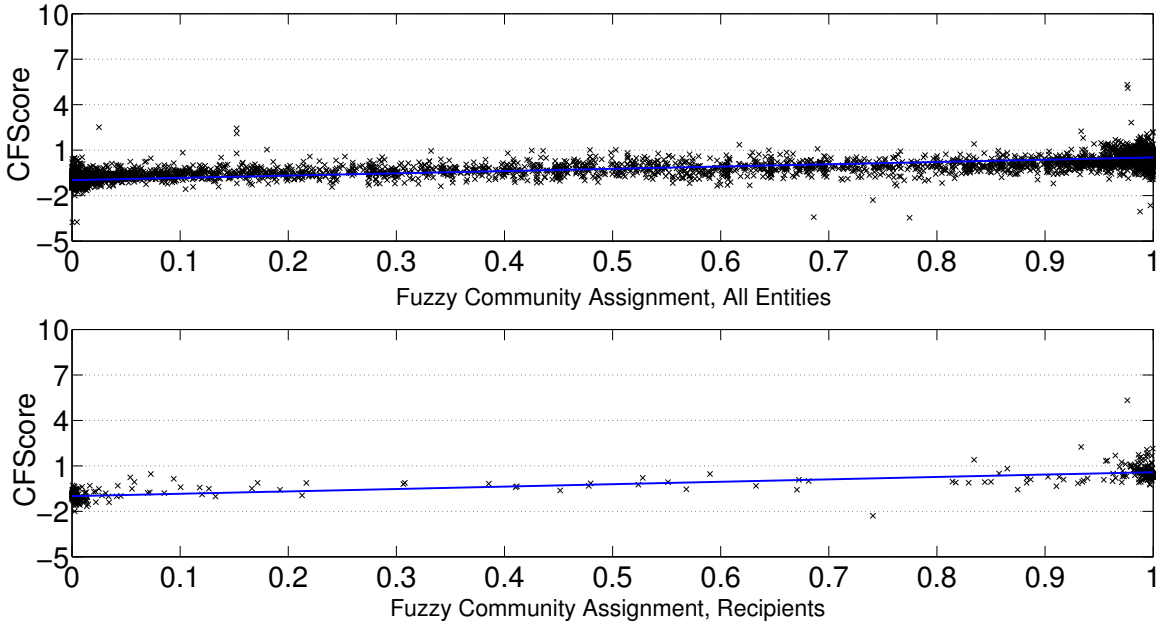



Figure 4.2: Fuzzy Community Assignment for Alaska Contribution Network, All Years, Two Communities

communities and compares them to the CFScore of the entity. This comparison is done twice: once for all the entities in the network, the other for just the recipients. Calculating the Pearson correlation coefficient for all entities gives $\rho = 0.9133$. The correlation coefficient for just the recipient limited set is $\rho = 0.8715$. This indicates that, for Alaska, the CFScore is highly correlated with the community assignments.

Similar graphs can be seen with Figures 4.3 and 4.4 which show the CFScores and the community assignments for each of the four sub-communities, both for all entities and just recipients. Tables 4.4, 4.5, and 4.6 provide context for the 4-community assignments. The results show the intra-community records for those entities who have a fuzzy assignment value of at least 0.5. The resulting analysis shows the money donated strictly within the communities, excluding cross-community records.

In Table 4.4, the column labeled % Dollars shows what percentage of money that community and party comprise of all donations to that party. Thus, the intra-

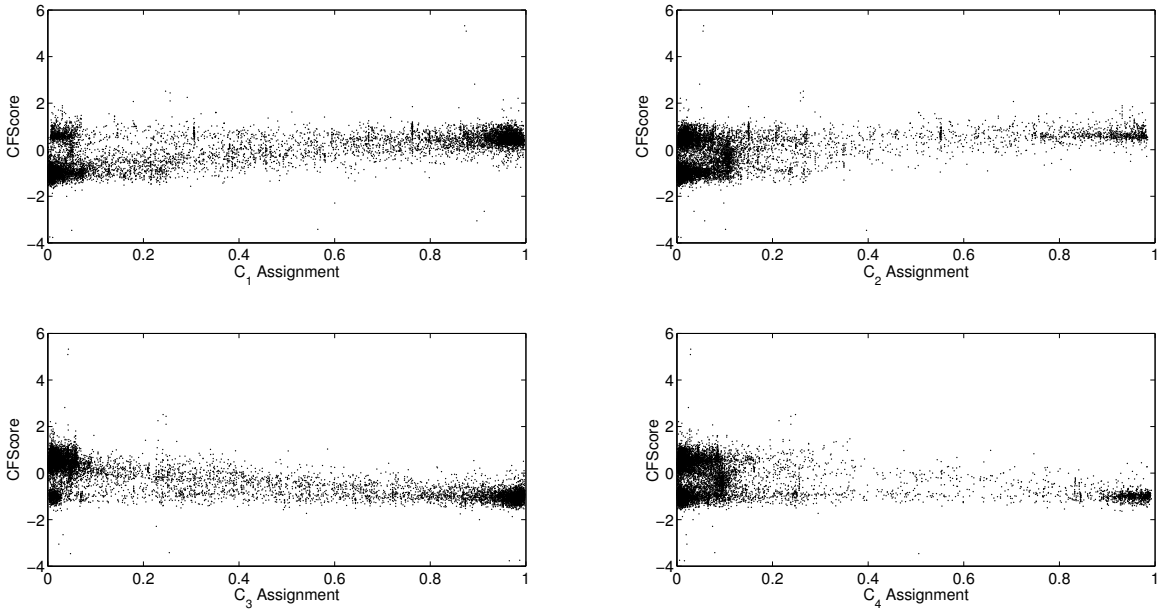



Figure 4.3: Fuzzy Community Assignment for Alaska Contribution Network, All Years, Four Communities

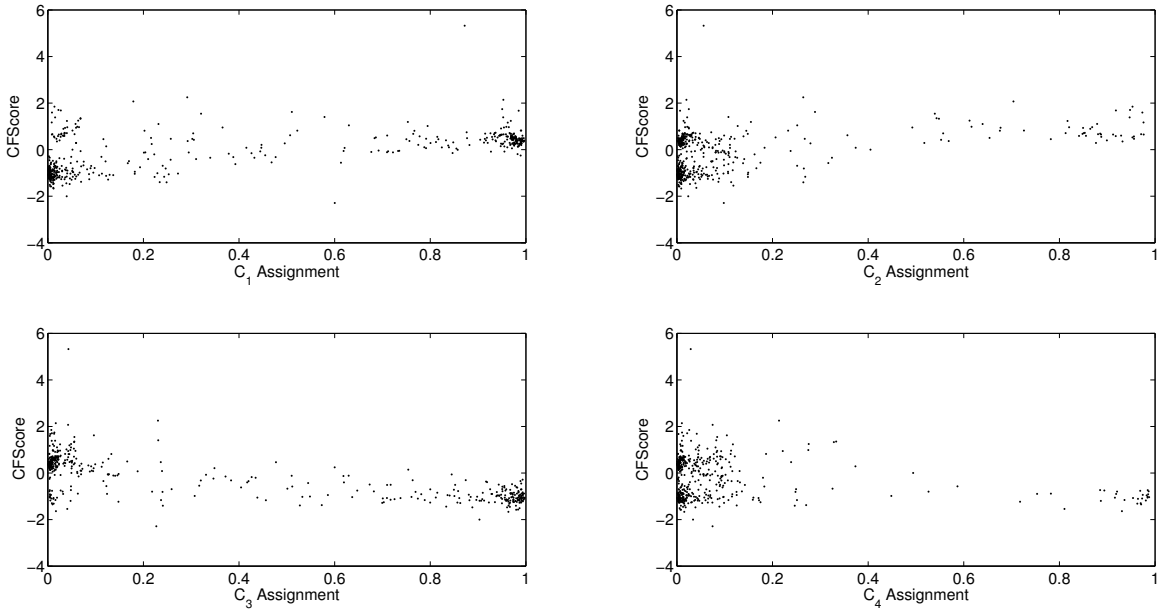



Figure 4.4: Fuzzy Community Assignment for Alaska Contribution Network, Recipients Only, Four Communities

community giving in $C_1$ makes up 59.43% of all donations to Republicans within the Alaska dataset. From this table, it can be seen that communities $C_1$ and $C_2$ identify primarily with Republicans and $C_3$ and $C_4$ with Democrats. More in-depth analysis shows the makeup of the donations is markedly different within the communities. Table 4.5 further breaks down the data by whether or not the recipient won their election. From this, it can be seen that the intra-community giving in $C_1$ contains 67.07% of the donations to Republican winners. It also has a large number of donations to Republican losers as well as Republican affiliated committees or incumbents. $C_2$ donated much more to the losing Republicans than winners, a considerable difference from $C_1$, showing that the losing candidates in this community received money from a different set of people as a whole.

Table 4.7 shows the results of the analysis being expanded to all contributions from the donors in $C_2$, including to those recipients in other communities. From this, 17.95% of the money and 21.18% of the raw number of contributions to Republican losers came from the donors in this community. However, Republican winners only received 9.26% of dollars and 12.06% of number of records from donors in this community, indicating this community consists of, on average, smaller value donations that did not target winning candidates as effectively as those in $C_1$. Based on this, $C_1$ appears to donate perhaps more strategically than $C_2$. This is supported by winning Democratic candidates receiving 18.90% of their money from the donors in $C_1$ despite that group being heavily biased towards Republicans. Communities $C_3$ and $C_4$ mirror this slightly where the ratio of money from donors in $C_4$ favored winning candidates despite losing Democratic candidates raising more money overall. Similar results can be found in Table 4.6, which analyzes incumbency (Incumbent 'I', Open Seat 'O', and Challenger 'C').

Such analysis can be applied to all levels of the hierarchy. At the level with the

Table 4.4: Intra-community Records by Party for Alaska Contribution Networks with Four Communities

| C | Party | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|
| $C_1$ | Republican | $13,296,884.92 | 59.43% | 38,811 | 60.68% |
| $C_1$ | Democratic | $573,678.84 | 2.94% | 1,651 | 2.11% |
| $C_1$ | Independent | $102,579.57 | 3.02% | 368 | 10.16% |
| $C_2$ | Republican | $1,590,281.52 | 7.11% | 5,794 | 9.06% |
| $C_2$ | Independent | $354,543.30 | 10.45% | 207 | 5.72% |
| $C_3$ | Democratic | $10,731,133.80 | 55.05% | 47,798 | 61.11% |
| $C_3$ | Independent | $2,349,740.06 | 69.29% | 1,724 | 47.60% |
| $C_3$ | Republican | $11,420.00 | 0.05% | 28 | 0.04% |
| $C_4$ | Democratic | $1,107,494.10 | 5.68% | 6,924 | 8.85% |
| $C_4$ | Independent | $6,177.11 | 0.18% | 32 | 0.88% |
| $C_4$ | Republican | $1,860.00 | 0.01% | 10 | 0.02% |

highest number of communities, the communities start to become much more focused on the type of individual that strongly identifies with the community. Many of the communities become centered around specific districts that are more isolated in their donations when compared with other districts.

To make sure the resulting communities still represent ideology well after being split into individual 2-year cycles, a similar test was performed on the temporal datasets for Alaska. As before, looking for two communities resulted in splits where the fuzzy community assignment was highly correlated with the CFScore for that entity. Table 4.8 shows the correlations for each of the cycles for both all entities and just the recipients / candidates. For Alaska, these fuzzy memberships are highly

Table 4.5: Intra-community Records by Party and Winner for Alaska Contribution Networks with Four Communities

| C | Party | Winner | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Republican | W | $9,274,341.69 | 67.07% | 27843 | 67.94% |
| $C_1$ | Republican | L | $3,085,176.21 | 45.36% | 8868 | 45.07% |
| $C_1$ | Republican | - | $937,367.02 | 53.72% | 2100 | 63.46% |
| $C_1$ | Democratic | W | $512,676.84 | 5.92% | 1452 | 4.06% |
| $C_1$ | Democratic | L | $61,002.00 | 0.67% | 199 | 0.54% |
| $C_1$ | Independent | L | $59,312.15 | 11.44% | 220 | 11.25% |
| $C_1$ | Independent | - | $42,527.42 | 1.53% | 146 | 9.99% |
| $C_1$ | Independent | W | $740.00 | 0.84% | 2 | 0.98% |
| $C_2$ | Republican | L | $963,487.38 | 14.17% | 3250 | 16.52% |
| $C_2$ | Republican | W | $626,794.14 | 4.53% | 2544 | 6.21% |
| $C_2$ | Independent | - | $344,143.30 | 12.36% | 152 | 10.40% |
| $C_2$ | Independent | L | $10,400.00 | 2.01% | 55 | 2.81% |
| $C_3$ | Democratic | L | $5,426,429.70 | 59.45% | 23968 | 64.61% |
| $C_3$ | Democratic | W | $4,280,814.58 | 49.40% | 20627 | 57.63% |
| $C_3$ | Independent | - | $2,101,182.50 | 75.45% | 663 | 45.38% |
| $C_3$ | Democratic | - | $1,023,889.52 | 60.25% | 3203 | 60.13% |
| $C_3$ | Independent | L | $226,996.56 | 43.77% | 1026 | 52.45% |
| $C_3$ | Independent | W | $21,561.00 | 24.55% | 35 | 17.07% |
| $C_3$ | Republican | W | $7,425.00 | 0.05% | 19 | 0.05% |
| $C_3$ | Republican | L | $3,995.00 | 0.06% | 9 | 0.05% |
| $C_4$ | Democratic | W | $759,435.37 | 8.76% | 4123 | 11.52% |
| $C_4$ | Democratic | L | $348,058.73 | 3.81% | 2801 | 7.55% |
| $C_4$ | Independent | L | $5,987.11 | 1.15% | 30 | 1.53% |
| $C_4$ | Republican | L | $1,860.00 | 0.03% | 10 | 0.05% |
| $C_4$ | Independent | - | $190.00 | 0.01% | 2 | 0.14% |

correlated with the CFscore.

By using the procedure outlined before, it is possible to track community behavior over time. Looking at one of the communities originating in 2004, analysis shows the candidates within that community are primarily Republicans running for districts in and around Fairbanks. This community continues to show up in

Table 4.6: Intra-community Records by Party and Incumbency for Alaska Contribution Networks with Four Communities

| C | Party | Incumbency | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Republican | I | $5,044,515.85 | 62.01% | 16,068 | 66.60% |
| $C_1$ | Republican | O | $3,724,538.79 | 56.25% | 9,811 | 53.71% |
| $C_1$ | Republican | C | $2,785,329.26 | 56.75% | 8,771 | 56.69% |
| $C_1$ | Republican | - | $932,467.02 | 54.28% | 2,093 | 64.12% |
| $C_1$ | Republican | I,O | $810,034.00 | 81.65% | 2,068 | 72.92% |
| $C_1$ | Democratic | I | $439,957.07 | 6.47% | 1,228 | 4.42% |
| $C_1$ | Democratic | O | $105,846.77 | 1.83% | 324 | 1.46% |
| $C_1$ | Independent | C | $58,512.15 | 16.76% | 216 | 14.18% |
| $C_1$ | Independent | - | $42,527.42 | 1.53% | 146 | 10.02% |
| $C_1$ | Democratic | C | $27,875.00 | 0.53% | 99 | 0.43% |
| $C_1$ | Independent | O | $1,540.00 | 0.72% | 6 | 1.10% |
| $C_2$ | Republican | C | $572,524.29 | 11.66% | 1,944 | 12.56% |
| $C_2$ | Republican | O | $510,127.02 | 7.70% | 1,742 | 9.54% |
| $C_2$ | Republican | I | $507,630.21 | 6.24% | 2,108 | 8.74% |
| $C_2$ | Independent | - | $344,143.30 | 12.37% | 152 | 10.43% |
| $C_2$ | Independent | O | $10,400.00 | 4.84% | 55 | 10.09% |
| $C_3$ | Democratic | C | $3,380,720.36 | 64.29% | 15,597 | 67.78% |
| $C_3$ | Democratic | O | $3,274,571.83 | 56.74% | 13,873 | 62.54% |
| $C_3$ | Democratic | I | $3,054,340.09 | 44.95% | 15,131 | 54.51% |
| $C_3$ | Independent | - | $2,097,732.50 | 75.42% | 659 | 45.23% |
| $C_3$ | Democratic | - | $1,021,501.52 | 61.20% | 3,197 | 60.73% |
| $C_3$ | Independent | C | $177,488.64 | 50.84% | 883 | 57.98% |
| $C_3$ | Independent | O | $56,007.92 | 26.06% | 156 | 28.62% |
| $C_3$ | Independent | I | $18,511.00 | 40.38% | 26 | 26.80% |
| $C_3$ | Republican | I | $5,700.00 | 0.07% | 12 | 0.05% |
| $C_3$ | Republican | C | $4,995.00 | 0.10% | 11 | 0.07% |
| $C_3$ | Republican | O | $725.00 | 0.01% | 5 | 0.03% |
| $C_4$ | Democratic | I | $691,759.31 | 10.18% | 3,543 | 12.76% |
| $C_4$ | Democratic | C | $227,493.42 | 4.33% | 2,067 | 8.98% |
| $C_4$ | Democratic | O | $188,241.37 | 3.26% | 1,314 | 5.92% |
| $C_4$ | Independent | O | $3,120.79 | 1.45% | 14 | 2.57% |
| $C_4$ | Independent | C | $2,866.32 | 0.82% | 16 | 1.05% |
| $C_4$ | Republican | C | $1,860.00 | 0.04% | 10 | 0.06% |
| $C_4$ | Independent | - | $190.00 | 0.01% | 2 | 0.14% |

subsequent years. Figure 4.5 shows the amount of money donors within this group gave to Republicans over the years. As can be seen from the graph, since 2006, the amount of money donated to losing candidates over the years has been much higher than donations to winning candidates. For Republican candidates, the total amount spent to winners and losers was calculated over the years. The percentages given in

Table 4.7: All Records in Alaska for $C_2$ with Four Communities

| | | | | | | |
|---|---|---|---|---|---|---|
| $C_2$ | Republican | W | $1,280,688.01 | 9.26% | 4,943.00 | 12.06% |
| $C_2$ | Republican | L | $1,221,147.61 | 17.95% | 4,167.00 | 21.18% |
| $C_2$ | Independent | - | $354,082.80 | 12.71% | 191.00 | 13.07% |
| $C_2$ | Democratic | W | $127,162.00 | 1.47% | 469.00 | 1.31% |
| $C_2$ | Republican | - | $116,426.04 | 6.67% | 252.00 | 7.62% |
| $C_2$ | Democratic | L | $72,394.86 | 0.79% | 270.00 | 0.73% |
| $C_2$ | Independent | L | $18,560.00 | 3.58% | 92.00 | 4.70% |
| $C_2$ | Independent | W | $7,154.00 | 8.15% | 22.00 | 10.73% |
| $C_2$ | Democratic | - | $4,650.00 | 0.27% | 10.00 | 0.19% |

Table 4.8: Correlation of Fuzzy Community Assignment and CFScore for Alaska Networks by Election Cycle

| Cycle | Correlation for All Entities | Correlation for Recipients |
|---|---|---|
| 2004 | 0.8984 | 0.8384 |
| 2006 | 0.9162 | 0.9158 |
| 2008 | 0.9126 | 0.8480 |
| 2010 | 0.9226 | 0.8923 |
| 2012 | 0.9057 | 0.8238 |

the figure show how much of that money came from this group that focuses more on Fairbanks elections.

### 4.4.2 New York

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

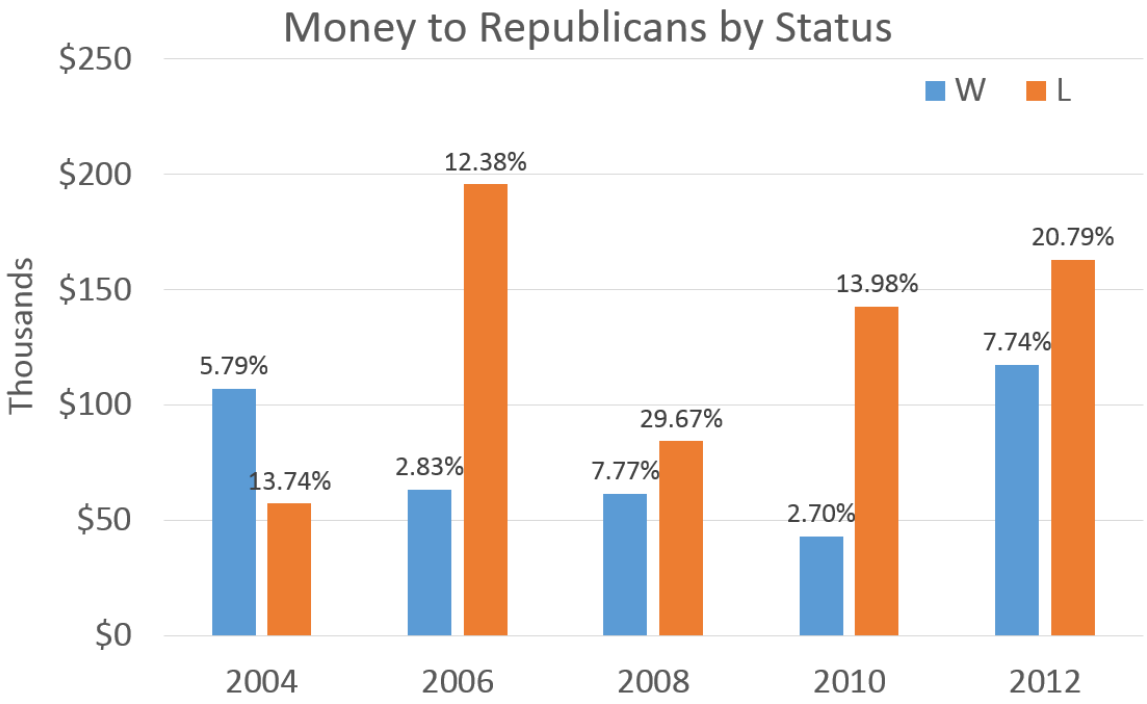



Figure 4.5: Donations from Community Members to Alaskan Republicans

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 to 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. Instead, 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

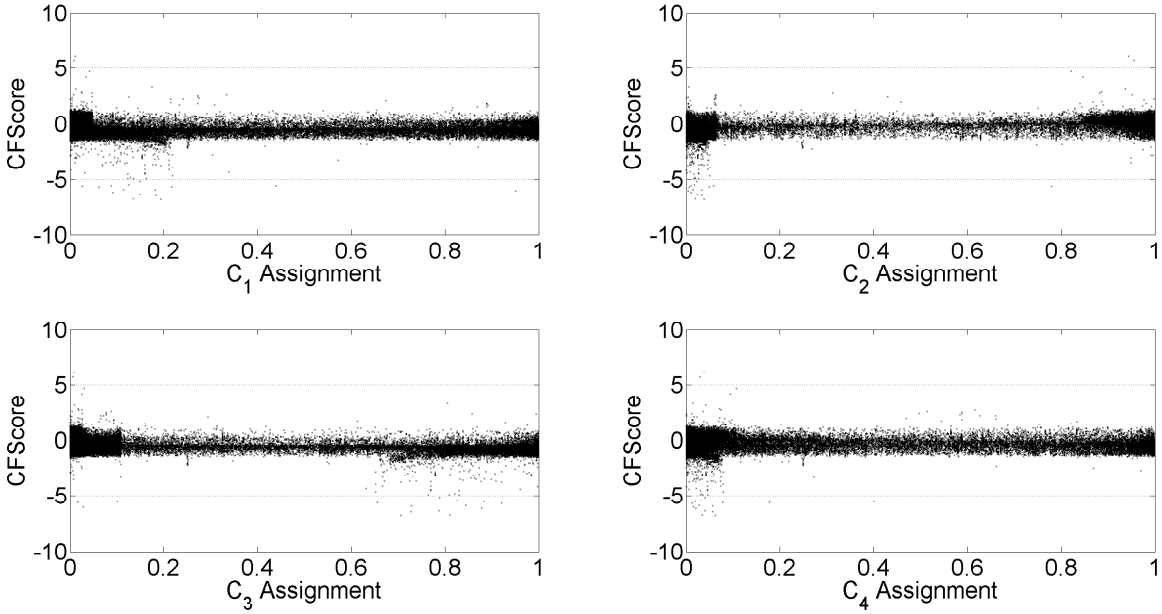



Figure 4.6: Fuzzy Community Assignments for New York Contribution Networks, All Years, Four Communities

information within these communities shows all the candidates for city offices in New York are within $C_2$. While not composed solely of city level candidates, the dominating factor for this breakdown appears to be geography and not ideology.

As a comparison to Alaska, Figures 4.6 and 4.7 show the community assignment scores compared with CFScore for all of New York when split into four sub-communities. At a glance, these graphs are different in that they do not show the tighter groupings of CFScores present in the Alaska data, further indicating that the CFScore ideological estimation is not a good explanation for the patterns of donations at this level.

Within the data, New York City recipients do not have party, incumbency, or status information provided. Due to this, breaking down the data into those categories requires eliminating the city data. After removing that data, Table 4.10 shows the breakdown of money by party and status for each community. Like the analysis of

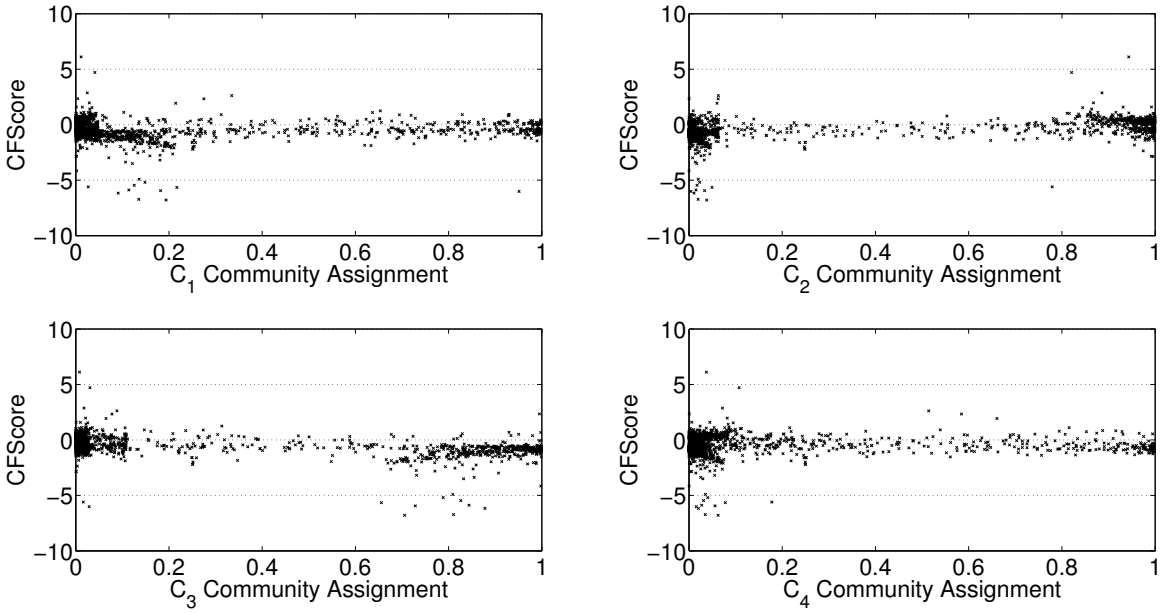



Figure 4.7: Fuzzy Community Assignments for Recipients in New York Contribution Networks, All Years, Four Communities

Alaska before, the data for this table is limited to intra-community donations. From this, it is possible to see that the data has been separated into three communities with primarily Democratic recipients, and only one community for Republican recipients. This indicates that the pattern of donation for Democrats is not as unified as those for Republicans. From the Democratic communities, two of them contain high percentages of the money donated to winning Democratic candidates. The third contains a large percentage of donations to losing Independents. Considering the average and standard deviation of the ideological measure of recipients in $C_3$ is notably different than $C_1$ and $C_4$, this indicates the donation patterns of this group are more highly ideologically motivated, especially since the recipients come from a wide variety of districts. The average and standard deviation for the CFScore is shown in Table 4.9.

To better see the difference between the Democratic communities $C_1$ and $C_4$,

Table 4.9: CFScore Statistics for Four Communities Level in New York Contribution Networks

| Community | Average CFScore | Std. Dev |
|---|---|---|
| $C_1$ | -0.458 | 0.492 |
| $C_2$ | 0.015 | 0.626 |
| $C_3$ | -0.927 | 0.989 |
| $C_4$ | -0.474 | 0.497 |

Table 4.11 shows the money grouped by party and incumbency. Due to the size of this table, any row in the table containing less than $100,000 has been removed. The threshold here is larger than in Wisconsin due to the higher overall contribution totals. These rows only represent a small fraction of the overall data relevant for a community. The grouping in this table reveals that $C_4$ contains most of the Democratic committees and political parties. Based on this information, the candidates in $C_1$ did not receive as much money from those who also donated to political parties directly due to the information gleaned by the community splitting. This demonstrates that despite the CFScore correlation not performing as well as in Alaska, there is still considerable and valuable information to be gained from the community assignments.

As noted in the overall New York results, CFScore does not correlate as well as in other states overall. This poor performance on correlation continues when looking at each individual year, as shown in Table 4.12. Interestingly, the two worst performers—2008 and 2012—have almost no data for New York City candidates. Since this seems counter intuitive given the full state results, we looked at the demographics of the candidates in the 2012 dataset.

Viewing the communities at the top hierarchy, what has happened is that a small set of candidates who nearly isolated from the rest of the network due to their

Table 4.10: Intra-community Records by Party and Winner for New York Contribution Networks with Four Communities

| C | Party | Status | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Democratic | W | $28,288,964 | 23.78% | 37286 | 25.87% |
| $C_1$ | Democratic | L | $10,653,103 | 26.81% | 5029 | 15.04% |
| $C_1$ | Democratic | - | $6,154,370 | 8.44% | 2584 | 12.57% |
| $C_1$ | Republican | W | $2,460,171 | 4.09% | 6012 | 6.25% |
| $C_1$ | Republican | L | $927,944 | 1.78% | 3176 | 9.74% |
| $C_1$ | Independent | L | $442,733 | 10.62% | 714 | 11.68% |
| $C_1$ | Independent | W | $97,959 | 23.68% | 216 | 27.73% |
| $C_2$ | Republican | - | $49,127,303 | 74.26% | 17167 | 80.32% |
| $C_2$ | Republican | W | $42,323,840 | 70.35% | 71341 | 74.12% |
| $C_2$ | Republican | L | $31,621,968 | 60.75% | 19204 | 58.91% |
| $C_2$ | Democratic | W | $11,350,791 | 9.54% | 22187 | 15.40% |
| $C_2$ | Democratic | L | $6,916,774 | 17.41% | 7698 | 23.03% |
| $C_2$ | Independent | L | $1,606,465 | 38.53% | 2387 | 39.05% |
| $C_2$ | Democratic | - | $503,881 | 0.69% | 18 | 0.09% |
| $C_2$ | Independent | - | $386,809 | 39.49% | 52 | 69.33% |
| $C_2$ | Independent | W | $236,507 | 57.17% | 452 | 58.02% |
| $C_3$ | Democratic | W | $3,854,056 | 3.24% | 13763 | 9.55% |
| $C_3$ | Democratic | L | $1,338,103 | 3.37% | 2478 | 7.41% |
| $C_3$ | Independent | L | $1,061,986 | 25.47% | 1553 | 25.41% |
| $C_3$ | Republican | L | $497,078 | 0.95% | 1146 | 3.52% |
| $C_3$ | Republican | W | $83,925 | 0.14% | 55 | 0.06% |
| $C_3$ | Independent | W | $20,275 | 4.90% | 24 | 3.08% |
| $C_4$ | Democratic | - | $42,182,503 | 57.87% | 10171 | 49.46% |
| $C_4$ | Democratic | W | $31,161,011 | 26.20% | 28864 | 20.03% |
| $C_4$ | Democratic | L | $10,341,808 | 26.03% | 11074 | 33.13% |
| $C_4$ | Republican | L | $7,324,020 | 14.07% | 2837 | 8.70% |
| $C_4$ | Republican | W | $4,468,078 | 7.43% | 5295 | 5.50% |
| $C_4$ | Republican | - | $1,685,346 | 2.55% | 312 | 1.46% |
| $C_4$ | Independent | L | $169,938 | 4.08% | 427 | 6.99% |
| $C_4$ | Independent | - | $125,000 | 12.76% | 4 | 5.33% |

donor groups being considerably different than the rest of the candidates' groups. They were not completely disconnected from the rest of the network, however,

Table 4.11: Intra-community Records by Party and Incumbency for New York Contribution Networks with Four Communities

| C | Party | Incumbency | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Democratic | I | $22,488,481 | 24.90% | 32056 | 26.47% |
| $C_1$ | Democratic | O | $18,254,466 | 23.69% | 8701 | 20.89% |
| $C_1$ | Democratic | - | $2,480,785 | 5.31% | 1015 | 7.99% |
| $C_1$ | Republican | I | $2,313,709 | 3.79% | 5689 | 5.97% |
| $C_1$ | Democratic | C | $1,872,705 | 10.69% | 3127 | 13.79% |
| $C_1$ | Republican | C | $600,918 | 2.31% | 2275 | 13.30% |
| $C_1$ | Republican | O | $466,658 | 1.49% | 1193 | 6.50% |
| $C_1$ | Independent | O | $227,495 | 16.87% | 331 | 14.44% |
| $C_2$ | Republican | - | $44,860,514 | 74.69% | 15759 | 81.22% |
| $C_2$ | Republican | I | $39,836,834 | 65.32% | 69082 | 72.44% |
| $C_2$ | Republican | O | $24,121,450 | 77.12% | 13134 | 71.56% |
| $C_2$ | Republican | C | $14,254,312 | 54.75% | 9737 | 56.91% |
| $C_2$ | Democratic | I | $9,624,672 | 10.66% | 19561 | 16.16% |
| $C_2$ | Democratic | C | $4,383,723 | 25.02% | 5517 | 24.33% |
| $C_2$ | Democratic | O | $4,241,036 | 5.50% | 3604 | 8.65% |
| $C_2$ | Independent | I | $794,763 | 46.11% | 1396 | 51.59% |
| $C_2$ | Independent | O | $539,883 | 40.03% | 751 | 32.75% |
| $C_2$ | Democratic | - | $522,015 | 1.12% | 1221 | 9.61% |
| $C_2$ | Independent | C | $514,400 | 33.94% | 624 | 34.27% |
| $C_2$ | Independent | - | $380,734 | 39.07% | 120 | 82.19% |
| $C_3$ | Democratic | I | $2,331,485 | 2.58% | 10742 | 8.87% |
| $C_3$ | Democratic | C | $1,845,159 | 10.53% | 2799 | 12.35% |
| $C_3$ | Democratic | O | $988,890 | 1.28% | 2664 | 6.40% |
| $C_3$ | Independent | C | $683,657 | 45.10% | 670 | 36.79% |
| $C_3$ | Republican | C | $420,248 | 1.61% | 984 | 5.75% |
| $C_3$ | Independent | O | $243,126 | 18.02% | 689 | 30.05% |
| $C_3$ | Independent | I | $155,478 | 9.02% | 218 | 8.06% |
| $C_4$ | Democratic | - | $32,700,882 | 70.05% | 6977 | 54.93% |
| $C_4$ | Democratic | O | $22,967,017 | 29.81% | 12158 | 29.19% |
| $C_4$ | Democratic | I | $22,220,944 | 24.60% | 23758 | 19.62% |
| $C_4$ | Democratic | C | $5,796,479 | 33.09% | 7216 | 31.83% |
| $C_4$ | Republican | C | $5,681,259 | 21.82% | 1449 | 8.47% |
| $C_4$ | Republican | I | $4,747,704 | 7.78% | 5660 | 5.94% |
| $C_4$ | Republican | - | $1,802,519 | 3.00% | 386 | 1.99% |
| $C_4$ | Republican | O | $1,245,962 | 3.98% | 949 | 5.17% |
| $C_4$ | Independent | - | $125,000 | 12.83% | 4 | 2.74% |

and so not removed from the spectral clustering. Looking at the community as a non-overlapping community, it consisted of seven candidates, both Democrats and Republicans, although all their CFScores are less than $-0.5$. All but one of these candidates lost their election. This leaves the vast majority of the candidates in

Table 4.12: Correlation of Fuzzy Community Assignment and CFScore for New York Networks by Election Cycle

| Cycle | Correlation for All Entities | Correlation for Recipients |
|---|---|---|
| 2004 | 0.1215 | 0.1209 |
| 2006 | 0.6139 | 0.5050 |
| 2008 | 0.0173 | 0.0330 |
| 2010 | 0.1207 | 0.1551 |
| 2012 | 0.0575 | 0.0875 |

the other community. Likewise, the number of contributors in this community is small. These 42 donors gave to only 43 different candidates in NY 2012. Only fifteen candidates received money from more than one of these donors. Only nine received money from more than two, seven of those being part of this community.

Moving down the hierarchy, we instead look at the results of splitting the network into 4 sub-communities. Observing the party affiliation of candidates within these groups, even splitting into four communities does not separate well by expected ideology. One of the interesting communities within this group is $C_4$. Within this group are 70 Democrats, 40 Republicans, and a single third-party individual. On average, the candidates had a CFScore of $-0.26$ with a standard deviation of 0.44. This group happens to be comprised of mostly incumbents and includes many politicians in leadership roles, regardless of party. Based on the network of contributions, these individuals have better within community connections than with the rest of their respective parties. Moving backwards in time, connecting this community with its corresponding community in 2010 returns many of the same candidates. However, the group is much more closely aligned ideologically than in 2012.

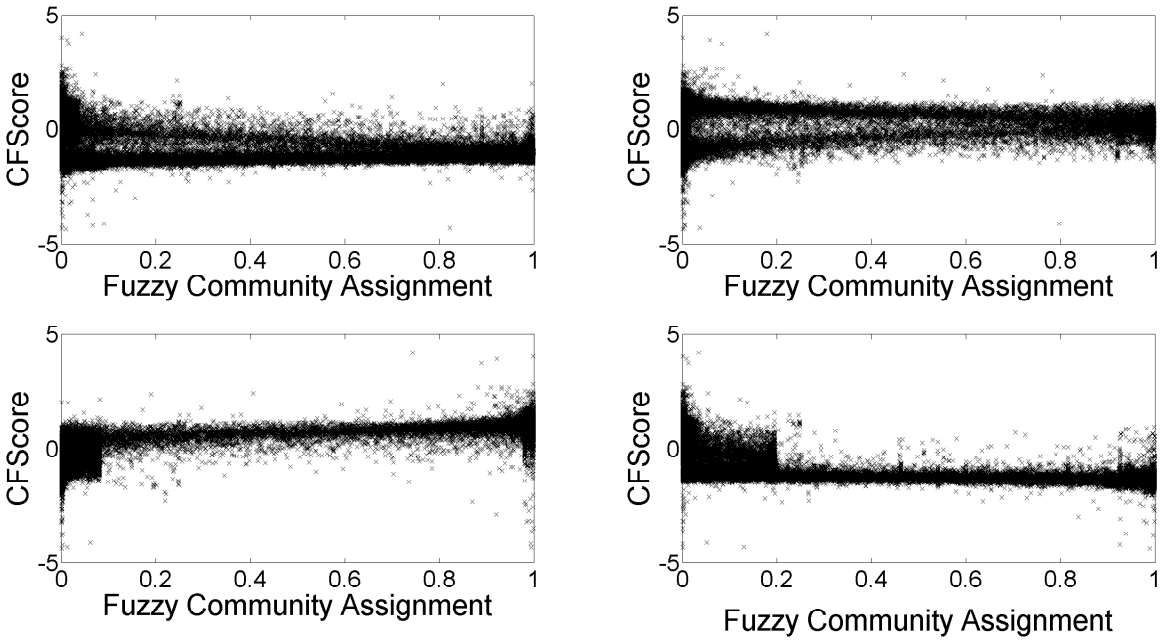



Figure 4.8: Fuzzy Community Assignments for Wisconsin Contribution Networks, All Years, Four Communities

### 4.4.3 Wisconsin

For Wisconsin, creating the network of contributions across all years as before results in one which contained 123,396 nodes and 592,407 edges. Part of the reason for this network being larger are the circumstances surrounding the 2012 recall and regular elections. As in Alaska, the correlation coefficient for CFScore and fuzzy community assignment at the top-level hierarchy is high at $\rho = 0.9745$ for all entities and $\rho = 0.9408$ for recipients. Figures 4.8 and 4.9 plot community assignment and CFScore when splitting the network into 4 sub-communities.

As before, this data is broken into two tables—Tables 4.13 and 4.14—to help make sense of the resulting communities. Due to the size of Table 4.14, rows with less than \$10,000 are removed. Viewing the party and status first, it is apparent that $C_1$ and $C_4$ are made almost entirely of Democratic candidates. These two communities are notably different in that one contains most of the money to winning

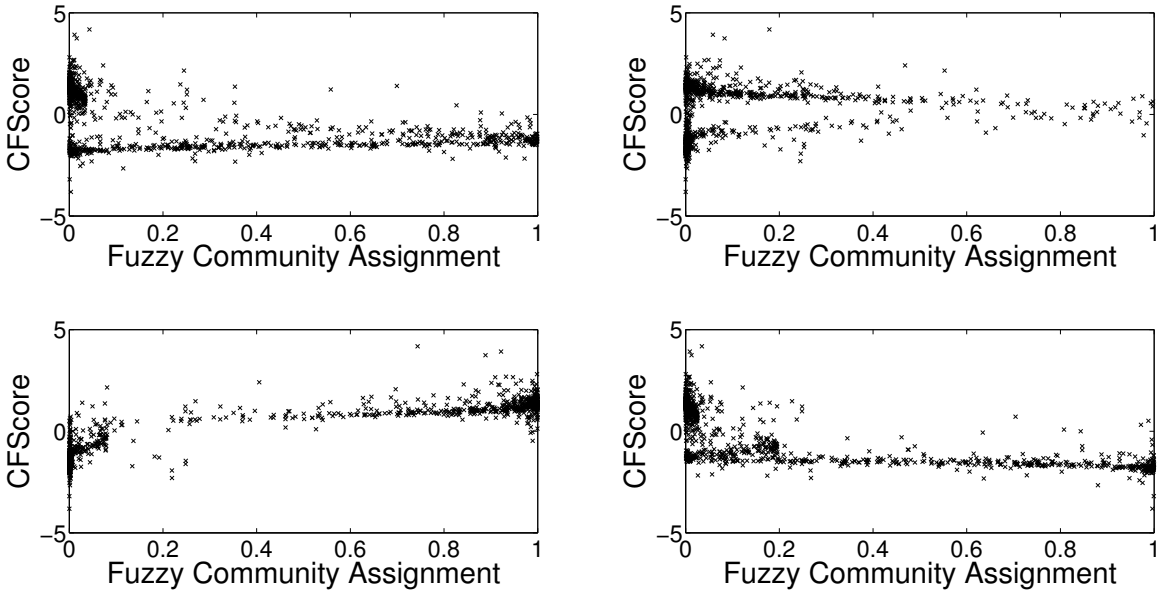



Figure 4.9: Fuzzy Community Assignments for Recipients Wisconsin Contribution Networks, All Years, Four Communities

candidates whereas the other, $C_4$, has much higher proportion of money going to losing candidates. The reason for this is explained better when viewing the incumbency information. With community $C_1$, much of that money went to left-leaning incumbents, where community $C_4$ is focused much more on challengers.

Community $C_3$ is dominated by Republicans. $C_2$ also is mostly Republican, but it is more evenly represented with other parties. Looking at the party and incumbency dataset reveals even more concerning these communities. From there, it becomes apparent that community $C_2$ is focused on incumbency in making donations. While the CFScores of the recipients indicate they are conservative leaning, this group is more ideologically moderate than the others, as shown in Table 4.15. Thus, the community decomposition reveals an ideologically moderate set of incumbent recipients who both share a considerable donor pool, as well as that pool being considerably different than others in their parties.

In individual election years, Wisconsin also shows high correlation at the top

Table 4.13: Intra-community Records by Party and Winner for Wisconsin Contribution Networks with Four Communities

| C | Party | Status | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Democratic | W | $23,874,573 | 45.89% | 168189 | 27.21% |
| $C_1$ | Democratic | - | $21,395,110 | 85.06% | 102666 | 73.40% |
| $C_1$ | Democratic | L | $18,435,910 | 26.56% | 98571 | 14.69% |
| $C_1$ | Independent | L | $5,246,483 | 37.73% | 22032 | 30.09% |
| $C_1$ | Independent | W | $3,546,579 | 31.03% | 17480 | 37.83% |
| $C_1$ | Independent | - | $1,892,033 | 53.42% | 9102 | 57.82% |
| $C_1$ | Republican | L | $2,025 | 0.00% | 12 | 0.01% |
| $C_2$ | Republican | W | $3,531,966 | 3.99% | 29187 | 4.84% |
| $C_2$ | Republican | - | $3,163,367 | 13.48% | 11418 | 7.75% |
| $C_2$ | Democratic | W | $1,624,672 | 4.68% | 12330 | 2.99% |
| $C_2$ | Independent | W | $1,233,344 | 10.79% | 1644 | 3.56% |
| $C_2$ | Democratic | L | $1,196,893 | 1.72% | 6399 | 0.95% |
| $C_2$ | Republican | L | $825,719 | 1.24% | 5346 | 1.80% |
| $C_2$ | Democratic | - | $232,955 | 1.85% | 1626 | 2.33% |
| $C_2$ | Independent | L | $194,537 | 1.87% | 604 | 1.10% |
| $C_2$ | Independent | - | $51,600 | 0.73% | 120 | 0.38% |
| $C_3$ | Republican | W | $113,286,410 | 85.25% | 789213 | 87.29% |
| $C_3$ | Republican | L | $60,748,953 | 91.27% | 265632 | 89.32% |
| $C_3$ | Republican | - | $28,558,583 | 81.16% | 194622 | 88.08% |
| $C_3$ | Independent | W | $8,034,255 | 52.73% | 24848 | 40.33% |
| $C_3$ | Independent | L | $2,023,522 | 19.40% | 10152 | 18.49% |
| $C_3$ | Independent | - | $707,469 | 6.66% | 5079 | 10.76% |
| $C_3$ | Democratic | L | $102,246 | 0.22% | 486 | 0.11% |
| $C_3$ | Democratic | W | $98,922 | 0.57% | 219 | 0.11% |
| $C_4$ | Democratic | L | $25,814,264 | 37.20% | 448410 | 66.84% |
| $C_4$ | Democratic | W | $9,517,339 | 18.29% | 313854 | 50.79% |
| $C_4$ | Independent | L | $4,721,538 | 45.27% | 29636 | 53.97% |
| $C_4$ | Independent | - | $3,199,465 | 45.16% | 20289 | 64.45% |
| $C_4$ | Democratic | - | $2,376,667 | 6.30% | 41916 | 19.98% |
| $C_4$ | Independent | W | $1,225,168 | 16.08% | 8340 | 27.07% |
| $C_4$ | Republican | L | $4,140 | 0.02% | 18 | 0.02% |

Table 4.14: Intra-community Records by Party and Incumbency for Wisconsin Contribution Networks with Four Communities

| C | Party | Incumbency | Total Dollars | % Dollars | Records | % Records |
|---|---|---|---|---|---|---|
| $C_1$ | Democratic | I | $31,734,548 | 49.40% | 182556 | 32.33% |
| $C_1$ | Democratic | C | $12,588,132 | 42.41% | 72657 | 17.60% |
| $C_1$ | Independent | I | $5,561,793 | 66.01% | 28194 | 52.82% |
| $C_1$ | Democratic | O | $4,792,716 | 23.17% | 31058 | 23.25% |
| $C_1$ | Democratic | - | $4,064,613 | 62.97% | 22542 | 59.04% |
| $C_1$ | Independent | O | $1,101,560 | 14.98% | 5786 | 25.69% |
| $C_1$ | Independent | C | $700,486 | 55.75% | 525 | 5.40% |
| $C_1$ | Independent | - | $572,211 | 9.27% | 1338 | 6.27% |
| $C_2$ | Republican | I | $4,169,208 | 3.92% | 33117 | 4.25% |
| $C_2$ | Democratic | I | $2,389,661 | 3.72% | 17973 | 3.18% |
| $C_2$ | Republican | - | $977,313 | 10.79% | 3620 | 6.37% |
| $C_2$ | Independent | I | $754,941 | 13.44% | 552 | 1.55% |
| $C_2$ | Democratic | O | $392,325 | 1.90% | 1426 | 1.07% |
| $C_2$ | Republican | O | $330,243 | 0.79% | 1470 | 0.85% |
| $C_2$ | Independent | O | $226,346 | 3.08% | 738 | 3.28% |
| $C_2$ | Democratic | C | $76,372 | 0.51% | 243 | 0.12% |
| $C_2$ | Republican | C | $59,775 | 0.47% | 336 | 0.47% |
| $C_2$ | Independent | - | $27,525 | 0.30% | 135 | 0.42% |
| $C_3$ | Republican | I | $88,429,229 | 83.17% | 664302 | 85.33% |
| $C_3$ | Republican | O | $56,808,707 | 90.48% | 234558 | 90.09% |
| $C_3$ | Republican | C | $23,430,706 | 91.61% | 131798 | 92.80% |
| $C_3$ | Republican | - | $7,403,317 | 81.76% | 50970 | 89.63% |
| $C_3$ | Independent | O | $3,556,518 | 48.35% | 6848 | 30.41% |
| $C_3$ | Independent | C | $1,741,064 | 46.19% | 10848 | 37.20% |
| $C_3$ | Independent | - | $818,274 | 8.84% | 1182 | 3.69% |
| $C_3$ | Independent | I | $356,687 | 6.35% | 9027 | 25.37% |
| $C_3$ | Democratic | I | $98,922 | 0.46% | 219 | 0.12% |
| $C_3$ | Democratic | C | $89,528 | 0.60% | 429 | 0.21% |
| $C_4$ | Democratic | C | $21,088,075 | 47.37% | 479613 | 77.45% |
| $C_4$ | Democratic | I | $6,323,537 | 9.84% | 222666 | 39.44% |
| $C_4$ | Democratic | O | $5,693,939 | 27.53% | 58194 | 43.57% |
| $C_4$ | Independent | - | $3,174,568 | 51.45% | 19677 | 92.16% |
| $C_4$ | Independent | O | $2,092,836 | 28.45% | 6494 | 28.84% |
| $C_4$ | Independent | C | $1,064,874 | 84.74% | 15189 | 156.25% |
| $C_4$ | Democratic | - | $585,250 | 9.07% | 4870 | 12.76% |
| $C_4$ | Independent | I | $280,799 | 5.00% | 4164 | 11.70% |

hierarchy when comparing community assignments and CFScore (Table 4.16). With the upswing in donations in 2012, the correlation between ideology and communities is exceptionally high. This seems reasonable given the apparent polarizing nature of the elections. Viewing some of the data through time, we isolated the community to

Table 4.15: CFScore Statistics for Four Communities Level in Wisconsin Contribution Networks

| Community | Average CFScore | Std. Dev |
|---|---|---|
| $C_1$ | -1.175 | 0.398 |
| $C_2$ | 0.389 | 0.570 |
| $C_3$ | 1.148 | 0.386 |
| $C_4$ | -1.610 | 0.222 |

Table 4.16: Correlation of Fuzzy Community Assignment and CFScore for Wisconsin Networks by Election Cycle

| Cycle | Correlation for All Entities | Correlation for Recipients |
|---|---|---|
| 2004 | 0.9196 | 0.9186 |
| 2006 | 0.9457 | 0.9452 |
| 2008 | 0.9233 | 0.9378 |
| 2010 | 0.9427 | 0.9357 |
| 2012 | 0.9808 | 0.9561 |

which Scott Walker belonged in 2012. This community is interesting as there was a surge in donations due to a controversial recall election. It is possible to backtrack and look at this community in different years to see how it changed over time. Table 4.17 shows the average CFScores and standard deviation of the recipients and donors within that community. This particular community has rather stable CFScore values, despite the different circumstances.

Table 4.17: CFScore Statistics for Walker Community in Wisconsin across Election Cycles

| Cycle | Ave. Recipient CFScore | Std. D. | Ave. Donor CFScore | Std. D. |
|---|---|---|---|---|
| 2004 | 1.036 | 0.475 | 0.882 | 0.373 |
| 2006 | 1.161 | 0.575 | 1.069 | 0.283 |
| 2008 | 1.112 | 0.389 | 0.994 | 0.340 |
| 2010 | 1.249 | 0.463 | 1.115 | 0.224 |
| 2012 | 1.077 | 0.570 | 1.088 | 0.331 |

## 4.5 Conclusion

In this section we augmented hierarchical fuzzy spectral clustering to track communities in social networks through time. Despite the volatile nature of the campaign finance datasets, most communities analyzed were able to be tracked through the years by use of a fuzzy similarity metric. Analyzing the communities showed how behavior of the group, and individuals, changed over time.

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, 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.[2]. 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

[2]This entity is identified by the entity resolution on the dataset by Bonica ID 52297646020

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.

Despite the top-level results not having high correlation, the communities in New York provide a great amount of information on donation habits. Analyzing the communities in depth helps to reveal more information about the overall donation patterns within New York politics. From there it was possible to find a community that appeared to have more ideological focus in the donations. It also showed that the donor network of Republicans appears to be tighter than that of Democrats.

Despite the anomalous events regarding the recall election in Wisconsin, the results of the temporal hierarchical fuzzy spectral clustering showed stability in the discovered communities. It was possible to track the communities at the top hierarchies through all the individual years. The resulting communities' assignments had high correlation with existing ideological metrics for the recipients and candidates.

Hierarchical fuzzy spectral clustering is able to group entities logically within political contribution networks. At two communities, the groups closely follow previous estimates of ideology. Splitting networks into more communities highlights differing patterns of donations beyond ideological scores within a state and in different election cycles. The community assignment values could be useful in other regression analyses in order to better identify correlations. Using ideological estimates and communities, it was possible to analyze entities who shift ideologies over time, as well as view groups who differ in their type of donations beyond ideology.

# CHAPTER FIVE

## ASSOCIATION RULE MINING

Based on the results of the previous chapters, one issue that occurs is determining useful semantics for each of the communities. Automated tools for analyzing the types of data within and between the different clusters would aid in understanding the actual utility of community detection. Using additional features of the data used in creating the campaign finance social networks, it is possible to create a transaction database describing the donation of money to political actors. These transactions include additional information about the donor and recipient. As examples, the recipient information includes party, incumbency, status of the election, district, as well as other information. Donor information varies by state according to what is required to be filed, but includes address, employer, occupation, and industry codes. Then these transactions can be used to establish relationships between nodes to generate a social network.

The previous chapters (3 and 4) have shown the effectiveness of finding hierarchical fuzzy communities within these networks in relation to ideological estimates. Even still, a better analysis of the patterns within each of the communities is desired. To that end, we apply association rule mining to the information contained within the transactions. The rule mining is conditioned on the fuzzy community assignment of the entities within those transactions. Motivating this approach is that donations can be made for a variety of reasons. Some of this is captured in the descriptive information regarding both the donor and the recipient. One individual may focus donations on incumbents within a set of districts, for example. By treating each of the components of an individual transaction as elements in a market basket,

the rules found highlight associations in the pattern of donations. This allows us to obtain rules characterizing these communities. Given the hierarchy of communities, more importantly it is possible to analyze the differences in rules arising between communities, especially in those that share a parent in the hierarchy.

## 5.1 Background

The results in Chapters 3 and 4 the hierarchical fuzzy spectral clustering found communities are correlated with ideological estimates. Further, in cases where the communities do not reflect ideology as well as in other areas, the resulting communities still contain useful information on patterns of donations. Part of the issue with these communities, however, is their "understandability." Analyzing the resulting communities, especially in how they are different from each other, can be difficult and sometimes only the obvious is discovered. Rule mining is proposed in this chapter as a possible method for assisting in that area. The following sections introduce some background on association rule mining.

### 5.1.1 Frequent Itemsets

In classical frequent itemset mining, we are dealing with a set of binary attributes called items. Additionally, there exists a database $\mathcal{D}$ of transactions $\mathbf{T}$. Each transaction $\mathbf{t}$ is a set of items that are a part of that transaction. The support of an item in $x_i$ is the percentage of transactions in the database that contain $x_i$, i.e.,

$$s\left(x_i\right) = \frac{\left|\mathbf{t} : x_i \in \mathbf{t}\right|}{\left|\mathbf{T}\right|}.$$

A frequent itemset $f_i = \{x_1, \ldots, x_k\}$ is combination of items in the database where together they have a support $s\left(f_i\right) = \frac{\left|\mathbf{t}:x_i \in \mathbf{t}\right|}{\left|\mathbf{T}\right|} \geq s_{min}$. Such frequent items are useful in

that they provide insight into trends within the database. Calculating all the frequent items in the data due to the exponential combinations of possible itemsets. Apriori was an early algorithm that aided in finding these frequent itemsets.

### 5.1.2 A Priori

An early rule mining algorithm uses the concept that a subset of a frequent itemset must also be large by definition [95]. The Apriori algorithm performs multiple passes, starting with finding the large 1-itemsets. These are the set of frequent items where the items $\mathbf{f}$ are found that have $s(f_i) \geq s_{min}$. The next passes generate new candidate itemsets based on the superset of previous itemsets. Then any non-frequent items are removed, and the algorithm continues. Algorithm 5.1 details the method for generating the large frequent itemsets from the database.

Once we obtain frequent itemsets, there is still more that can be done to gain insight into the relationships of the items. There are a couple of important statistics that assist in revealing additional relationships in the data. The confidence of a set of items is how likely an item is part of a transaction if the other items are part of the transaction. This is measured by the confidence

$$conf(x_i \rightarrow x_j) = \frac{s(x_i, x_j)}{s(x_i)}.$$

Lift is another important measure as it describes how likely it is that item $x_j$ is part of the transactions with $x_i$ while controlling for the frequency of $x_i$. This measure is defined as

$$lift(x_i \rightarrow x_j) = \frac{s(x_i, x_j)}{s(x_i) \times s(x_j)}.$$

When the lift for a pair of items $x_i \rightarrow x_j$ is greater than one, then $x_j$ is more likely to be a part of a transaction if $x_i$ is part of that transaction. These measure form

**Algorithm 5.1** Apriori Algorithm for Large Itemsets

```
function LargeItemsets(database)
    L = ∅
    F = ∅
    while F ≠ ∅ do                                   ▷ Make a pass over the dataset
        Candidate set C = ∅
        for database tuple t do
            for itemset f ∈ F do
                if c_f ∈ C then
                    c_f.count = c_f.count + 1        ▷ Add to the count if it is part of the candidate set
                else
                    c_f.count = 0                    ▷ If not part of the candidate, reset
                    C = C + c_f
                end if
            end for
        end for                                      ▷ Consolidate
        F = ∅
        for itemsets c ∈ C do
            if count(c) / |D| > s_min then
                L = L + c
            end if
            if c should be used as a frontier then
                F = F + c
            end if
        end for
    end while
end function
```

the basis of evaluating the association rules discovered later and highlighting patterns that may be of use in providing interpretability to the community assignments.

A similar process can be used to create rules from the data based on these itemsets. One of the augmentations to association rule mining was to integrate classification with the rule mining [96]. As noted in their work, the framework they developed was intended to help solve an "understandability problem" where rules produced by classification are difficult to understand. The algorithm Classification Based on Associations, or CBA, contains a rule generator and a classifier builder. First, a set of frequent *ruleitems* is found from within a transactional database. One example listed is $\langle \{(V_1, 1), (V_2, 1)\}, (class, 1)\rangle$ where $V_1$ and $V_2$ are attributes. From these set of frequent rules, multiple passes are performed over the data to generate candidate rules. This list is refined to create the final rules for use in a classifier, the results of which gave an improvement on C4.5.

## 5.2 Association Rules Across Communities

Within this chapter, the network used is based on political donations among candidates, committees, and donors. Each edge in the network represents a transaction between two entities. These transactions form both the information used in creating the community, as well as the transactions used in performing association rule mining afterwards.

We begin in the same manner as described in Chapter 3. A network is created from the campaign finance social network and we use hierarchical fuzzy spectral clustering to find communities within the network. As described before, once we obtain these communities, it can be difficult to analyze all the communities to determine behavior. In order to provide interpretability to the results, these discovered communities are used to create partitions in the transaction data to use in association rule mining. For the association rule mining, each of these transactions are tagged with the relevant metadata fields to create lists of features. Table 5.1 has

Table 5.1: Example transactions for political donations

| District | Status | Party | Office | Incumb. | Donor Type | Industry | Zip |
|---|---|---|---|---|---|---|---|
| Assembly 042 | Lost | D | House | C | Individual | Uncoded | 92220 |
| Assembly 031 | Won | D | House | I | Non-Individual | Health | 95814 |
| Senate 029 | Won | D | Senate | O | Non-Individual | Party | 94518 |

example rows from the data.

Using the fuzzy community values from the clustering procedure, the transactions within the full dataset are separated based on the fuzzy community assignments of the donors within a community. We allow membership for node $i$ in a community $j$ if $u_{i,j} \geq 0.3$. This creates overlapping communities and overlapping partitions of the underlying transactions. Using the Apriori Association Rule Mining package within R, association rules are found within the data based on the membership values of the donors. This procedure is performed for each community at each level of the hierarchy.

Since we are interested in discriminatory rules between communities, the focus of the analysis are on communities who share a parent. Any pair of communities could be analyzed in the same manner, however. Consider two sibling communities, $\mathbf{C}_{i,m}$ and $\mathbf{C}_{i,n}$. First, association rules are discovered for the transactions belonging to each of those communities, generating rule sets $R_{i,m}$ and $R_{i,n}$. Next, the two rule sets are compared against each other to generate categories of rules. First, the rules in common can be found by taking the intersection of the two sets. Such rules help identify overall trends in the data but are not as useful as discriminatory information.

More interesting is the set of conflicting rules between the sets. This conflict set can be determined from the rule sets by using the intersection of the antecedents as $\mathcal{A} = \text{ant}\,(R_{i,m}) \cap \text{ant}\,(R_{i,n})$. For each of the antecedents in common, we determine

Table 5.2: Fields used for rule association

| Field | Description | Examples |
|---|---|---|
| Donor Type | An individual or non-individual | Individual, Non-Individual, Other |
| Industry | Category for donor industry | Health, Labor, Agriculture, etc. |
| Zip Code | ZIP Code reported by the donor | 94131, 94028, etc. |
| District | The area a candidate represents | Assembly 027, Senate 029, etc. |
| Office Type | Legislative body for the office | House, Senate, Gubernatorial, etc. |
| Party | Party affiliation of the candidate | Democratic, Republican, etc. |
| Status | Candidate won or lost their election | Won, Lost, Withdrew, etc. |
| Incumbency | Incumbency status of candidate | Open, Incumbent, Challenger |

if there is a conflict in the consequents between the rules in $R_{i,m}$ and $R_{i,n}$. The resulting conflict sets provide information on discriminative donation patterns within that community. Additional information can be gleaned from the rules where the consequent is the same, but the antecedent is different.

The data used in this chapter is taken from the National Institute on Money in Politics. This particular dataset uses data from California in 2016. As before, the data is used to generate a social network where edges in the network represent donations between vertices in the graph. In addition, we use additional features of the transactions that includes additional information such as party of recipient, industry categorization of donors, etc. For the analysis below, a variety of fields were selected for mining association rules for both donors and recipients (Table 5.2. While there are more fields within the data, many of them are generalizations or specializations of other fields and including them would create rules defining those relationships instead of finding more interesting relationships.

For this research, data for each state is analyzed separately. We do not co-mingle

the state-level datasets since much of the analysis is focused on state elections. In addition, communities found in combined state data tend to cluster around each state anyway, instead of across map boundaries. Following prior work in political science, the network of relationships is subject to filtering [78]. The primary filter is that any entity who gave or received money only once is removed from consideration. This eliminates any nodes connected to the network by a single edge. As a second filter, any isolated subnetwork which is not connected to the largest set of connected nodes is also removed since any such group becomes its own community by default.

## 5.3 Results

For state elections, California sees the most money donated to candidates and committees within that state[1]. For this reason, we focus on that state for the analysis presented here. For just candidates and committees involved in 2016 elections, California had nearly one billion dollars in contributions. Out of that money, non-individuals outspent individuals at a rate of almost five to one. Using the 2016 data for initial analysis, the network of relationships for California in that year total 5372 nodes (of which 232 are recipients) and 32,309 edges.

As a baseline, association rule mining based on Apriori is performed over the transactions without partitioning the data by community. Many of the resulting rules discovered here are not especially insightful. Most are things already well known, such as incumbents have a much higher chance of winning their elections, winning candidates raise more money, non-individuals spend more widely than individuals, etc. In all, 120 different rules were discovered for the entire set of transactions. A rendering of the groups of rules discovered is shown in Figure 5.1 where the rules are

[1]Based on 2016 data from https://www.followthemoney.org

grouped by common antecedents. In these figures, the size of the circle is relative to the support of the rules and darker shades indicate higher lift. Additionally, due to the length of the labels, Incumbency and Donor Type are abbreviated as ICO and DType respectively.

It is possible to get more interesting results after splitting the data by community. After performing fuzzy spectral clustering over the California data, the resulting communities are fairly evenly split. Using the donor membership values, the transactional data is split into two overlapping datasets. Using 0.3 as a threshold for membership, there are 3154 entities (141 candidates) in $C_{2,1}$ and 3382 entities (180) candidates in $C_{2,2}$. Rule groups found are shown in Figures 5.2 and 5.3.

The results for mining rules over two communities provides additional information. For $C_{2,1}$, it becomes immediately clear that the members of this community donate to Democrats since the rule $\{\} \Rightarrow Party = Democratic$ has 0.91 support. Another rule discovered within this community is that of $\{Incumbency = I\} \Rightarrow Status = Won$ at a support of 0.35 and confidence 0.94. As expected based on prior knowledge is that the incumbent is more likely to win their election.

What is more interesting is when these rules are compared with those for $C_{2,2}$. One expectation is that for this evenly split data where one community reflects Democrats is that the other community should be comprised mostly of Republicans. However, this is not the case. Instead, this group consists of mostly candidates who won their election, regardless of party. Many of these donations in $C_{2_2}$ come from non-individuals, as shown by the rules

$$\{\text{Party} = \text{Democrat}\} \Rightarrow \text{Donor Type} = \text{Non-Individual}$$

$$\{\text{Party} = \text{Republican}\} \Rightarrow \text{Donor Type} = \text{Non-Individual}$$

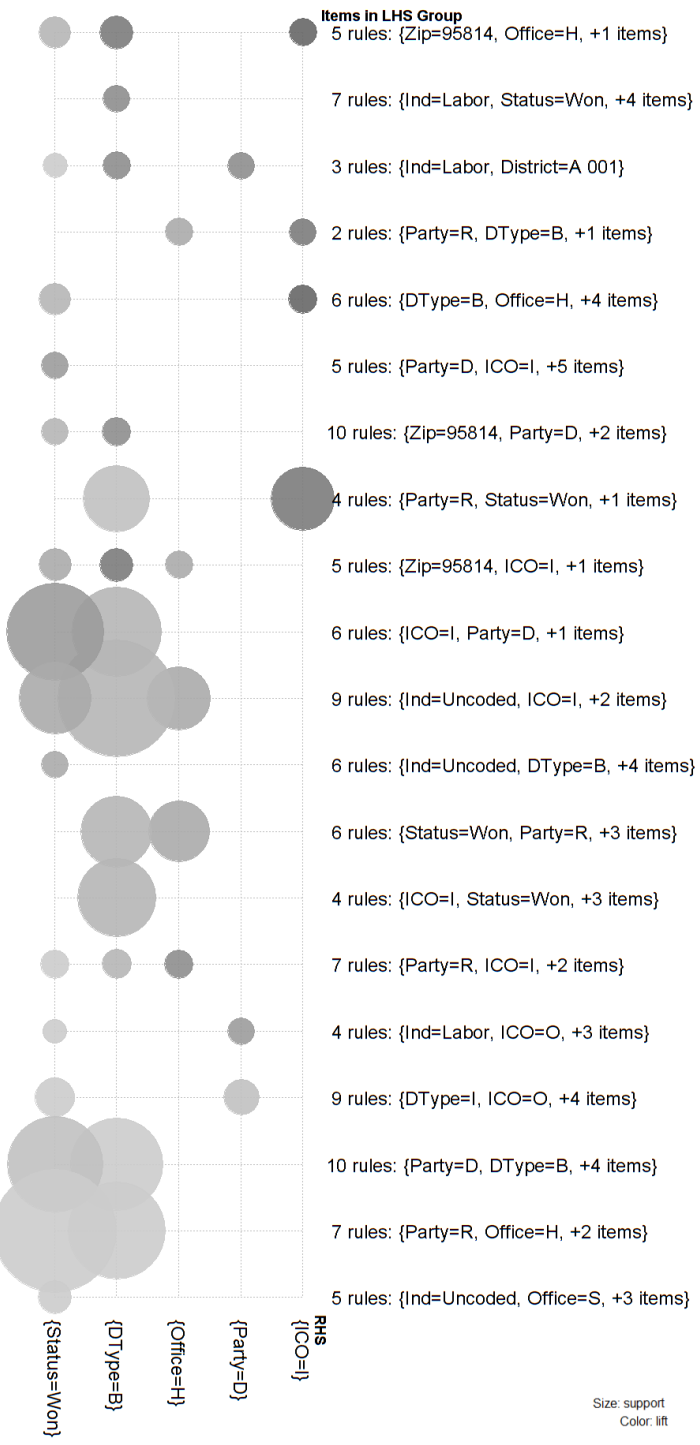



Figure 5.1: Rules visualization for all California 2016 transactions

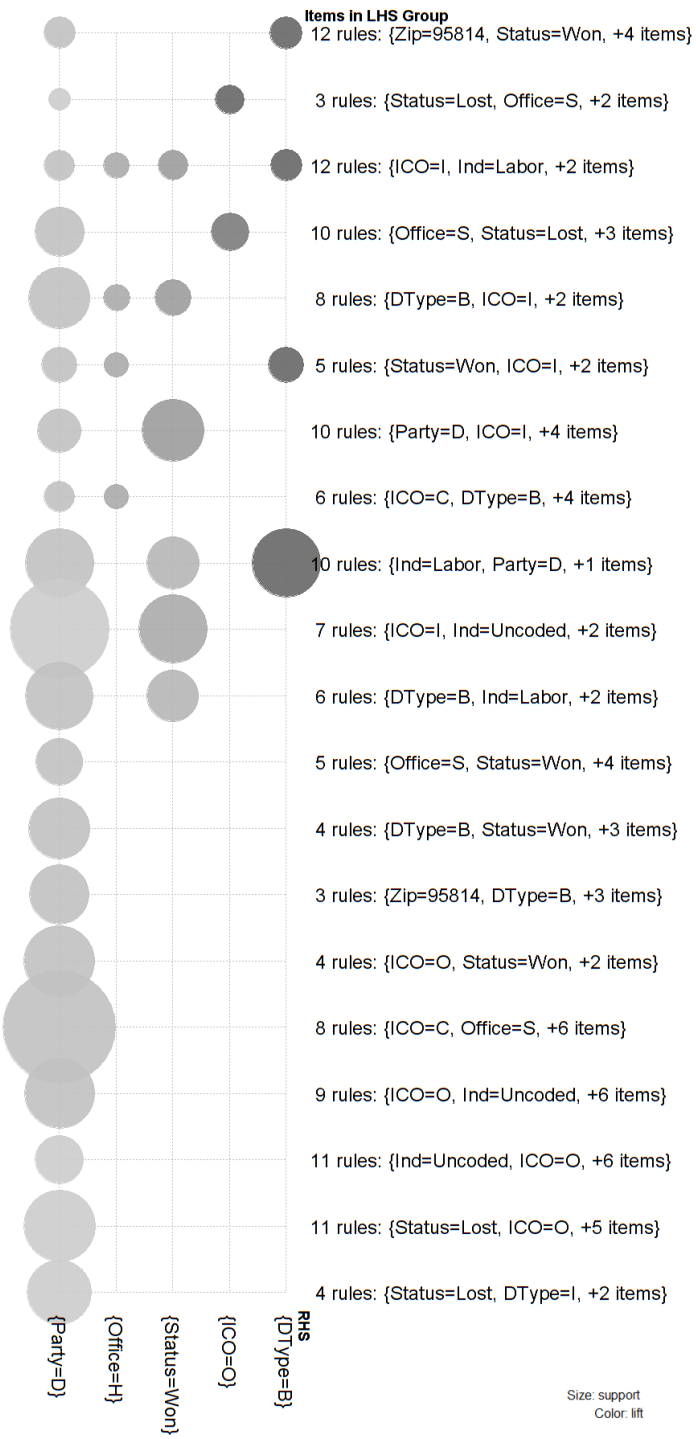



Figure 5.2: Rules visualization for California 2016 community $C_{2,1}$

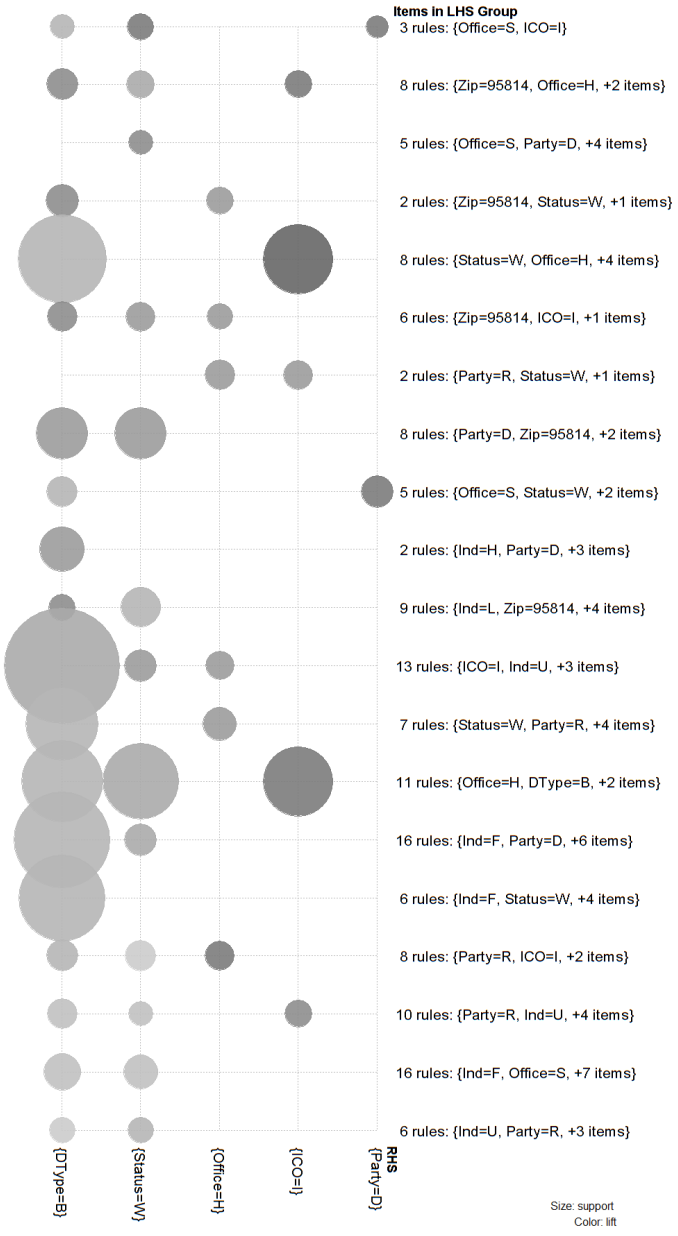



Figure 5.3: Rules visualization for California 2016 community $C_{2,2}$

While both rules were extracted, the support for the former is 0.63 and the latter is 0.24. Community $C_{2,1}$ has the only rule referencing individuals of

$$\{\text{Donor Type} = \text{Individual}\} \Rightarrow \text{Party} = \text{Democratic}$$

with 0.39 support and 0.87 confidence. This helps show the difference between the two communities where $C_{2,1}$ contained more individuals donating to Democrats whereas $C_{2,2}$ is made of non-individuals donating to winning candidates independent of party, highlighting the differing implicit strategies of the types of donors.

Moving down the hierarchy provides even more information. Communities $C_{3,2}$ and $C_{3,3}$ both are composed of many Democrat candidates. To compare these communities, we calculate the intersection of the rules that were discovered in each community individually. The intersection of the rules are shown in Table 5.3. In particular, we look at the rules with the highest lift amongst the two communities. The top rules help show that both communities contain useful patterns and rules showing how non-individuals, especially in the labor industry, gave to winning candidates. Unsurprisingly, in both datasets, incumbents tended to win. The important information here is that it confirms this is an integral part of the parent community.

Also important are the ways in which the communities $C_{3,2}$ and $C_{3,3}$ differ. Tables 5.4 and 5.5 contain the highest lift rules that were not shared by the two communities. The high lift rules in $C_{3,2}$ immediately highlight behaviors related to donations to elections where there is no incumbent. Notably, these rules also refer mostly to individuals. There is no similar rule in the other community. This helps highlight that, although both communities contain primarily Democratic candidates, there is considerable difference in the donating habits between individuals and non-

individuals. Confirming this are the high lift rules for $C_{3,3}$. These rules focus on non-individuals and highlight that they are frequently associated with winning candidates. This is especially true for two different types of donors: those from the labor industry, and those from ZIP Code 95814. ZIP code 95814 corresponds the state capital building and area. The rules indicate nearly all donations from that area were from non-individuals (0.997 confidence) and that this money went to winning candidates (0.853 confidence). Much of this came from party committees, in support of candidates likely to win. While this is mostly incumbents as indicated by the rule $\{Incumbency = I\} \Rightarrow Status = Won$ with 0.53 support, the rule $\{Incumbency = O\} \Rightarrow Party = Democrat$ at support 0.33 indicates that money was funneled to those going for open seats as well, but it seems they were less likely to win. All of this helps to demonstrate the ability for rule finding to improve analysis of the communities.

## 5.4 Conclusion

In this chapter we presented a method for trying to add understandability to community detection. This was done using hierarchical fuzzy spectral clustering on a campaign finance social network. The communities were hierarchical overlapping clusters of entities in California. Since the addition of the hierarchy adds more complexity in performing analysis, it becomes important to automatically find shared and discriminatory data between the large number of communities discovered. Using the additional data provided with the transactions forming the links between individuals in the network, association rule mining found additional discriminatory information for the communities. The resulting rules aid in providing insight into the donation patterns of groups within the data beyond what was readily apparent from the rules found using the dataset in its entirety. While this work relied on Apriori for

Table 5.3: Common Rules for $C_{3,2}$ and $C_{3,3}$

| Antecedent | Consequent | $C_{3,2}$ Lift | $C_{3,3}$ Lift |
|---|---|---|---|
| Status=Won<br>Industry=Labor | Donor=Non-Individual | 2.039 | 1.421 |
| Status=Won<br>Party=D<br>Industry=Labor | Donor=Non-Individual | 2.039 | 1.421 |
| Office=H<br>Industry=Labor | Donor=Non-Individual | 2.036 | 1.421 |
| Party=D<br>Office=H<br>Industry=Labor | Donor=Non-Individual | 2.036 | 1.421 |
| Industry=Labor | Donor=Non-Individual | 2.025 | 1.421 |
| Party=D<br>Industry=Labor | Donor=Non-Individual | 2.025 | 1.421 |
| Party=D<br>Incumbency=I | Status=Won | 1.574 | 1.226 |
| Party=D<br>Office=H<br>Incumbency=I | Status=Won | 1.557 | 1.221 |
| Incumbency=I | Status=Won | 1.520 | 1.188 |
| Incumbency=I<br>Donor=Non-Individual | Status=Won | 1.515 | 1.204 |

Table 5.4: Rules only in $C_{3,2}$

| Antecedent | Consequent | Lift |
|---|---|---|
| Status=Lost, Party=D, Office=S | Incumbency=O | 1.736 |
| Status=Lost, Office=S<br>Donor=Individual | Incumbency=O | 1.719 |
| Status=Lost, Office=S | Incumbency=O | 1.709 |
| Party=D, Office=S, Industry=Uncoded<br>Donor=Individual | Incumbency=O | 1.617 |
| Office=S, Industry=Uncoded<br>Donor=Individual | Incumbency=O | 1.607 |
| Party=D, Office=S<br>Donor=Individual | Incumbency=O | 1.596 |
| Party=D, Incumbency=C | Office=H | 1.585 |
| Office=S, Donor=Individual | Incumbency=O | 1.585 |
| Status=Lost, Party=D, Industry=Uncoded<br>Donor=Individual | Incumbency=O | 1.575 |
| Party=D, Office=S, Industry=Uncoded | Incumbency=O | 1.573 |

rule mining as a proof of concept, future work can use more sophisticated algorithms for determining rules as well.

Table 5.5: Rules only in $C_{3,3}$

| Antecedent | Consequent | Lift |
|---|---|---|
| Status=Won, Party=D<br>Office=H, Industry=Labor | Incumbency=I | 1.459 |
| Status=Won, Party=D, Office=H<br>Donor=Non-Individual, Industry=Labor | Incumbency=I | 1.459 |
| Status=Won, Office=H, Industry=Labor | Incumbency=I | 1.451 |
| Status=Won, Office=H, Industry=Labor<br>Donor=Non-Individual | Incumbency=I | 1.451 |
| Industry=Labor, ZipCode=95814 | Donor=Non-Individual | 1.423 |
| Party=D, Industry=Labor, ZipCode=95814 | Donor=Non-Individual | 1.423 |
| Status=Won, Office=H, Industry=Labor | Donor=Non-Individual | 1.420 |
| Status=Won, Party=D<br>Office=H, Industry=Labor | Donor=Non-Individual | 1.420 |
| Incumbency=I, ZipCode=95814 | Donor=Non-Individual | 1.420 |
| Status=Won, ZipCode=95814 | Donor=Non-Individual | 1.420 |

# CHAPTER SIX

# VOTE PREDICTION

Much of the work described in the preceding chapters focuses on providing context to the communities and analyzing the behavior directly within donation patterns. These patterns are drawn directly from HFSC and the transactions underlying the construction of the social networks. However, we want to determine if the discovered communities are generalizable to predicting behavior in other areas. Thanks to the work by Bonica and Voteview, there is information tying the donations to federal legislators with their voting history beginning in 1980 [94, 97].

Using the communities discovered from the campaign finance, we analyze if it is possible to generalize the community assignments to predict voting behavior in the legislature based on estimated ideological of the bills themselves. To do so, we merge the community assignment features with the voting data and implement random forest classifiers to predict Yea or Nay votes. As one way to see if donation amounts affect voting, multiple weighting schemes are tested for the connections between donors and recipients. In addition to just adjacency, we find communities based on weighted networks using a logarithmic scale as well as the raw donation sums.

## 6.1 Background

In this section we discuss some of the classification models used to evaluate the generalization of the HFSC community assignments.

### 6.1.1 Decision Trees

For the purposes of this research, once the data is separated into clusters, or communities, that information is used to predict behavior of the members of the

network. By combining the cluster data with voting history, the goal is to predict voting in the future. By training a classification model on current data, it is possible to predict behavior based on the provided features.

Decision trees perform classification by generating a tree from repeatedly splitting training data based on tests of the features [98]. Each internal node of the tree represents a test of a value on some feature or set of features that partitions the data as a result of that test. The leaf nodes of the tree represent classes based on the partition of the data at that node. New samples can be classified by applying the tests, or rules, beginning at the root of tree and proceeding to a leaf node to determine the class. The appeal of most decision trees are that they are simple to create and are also simple to interpret due to the rules that can be inferred from the tree.

As a simple example, consider a set of data with party and policy information for a bill in the legislature. Positive and negative classes correspond to voting yea or nay on the bill. Figure 6.1 shows a hypothetical tree from that data. The first test concerns the general policy position of the bill, progressive or conservative. The next layer of the tree checks the party position of the legislator and classifies the most likely vote.

The criteria for choosing a feature and value to partition the data can vary greatly. There are numerous possible ways to perform the splitting. Common techniques utilizes information theory to determine what feature and values create the best partitions. The Iterative Dichotomizer 3 (ID3) algorithm is one well known decision tree algorithm that uses entropy and information gain [99]. Consider a set of classes $c \in C$. The entropy of any partition of the data $D$ is defined as

$$H(D) = \sum_{c \in C} -p(c) \log_2 p(c)$$

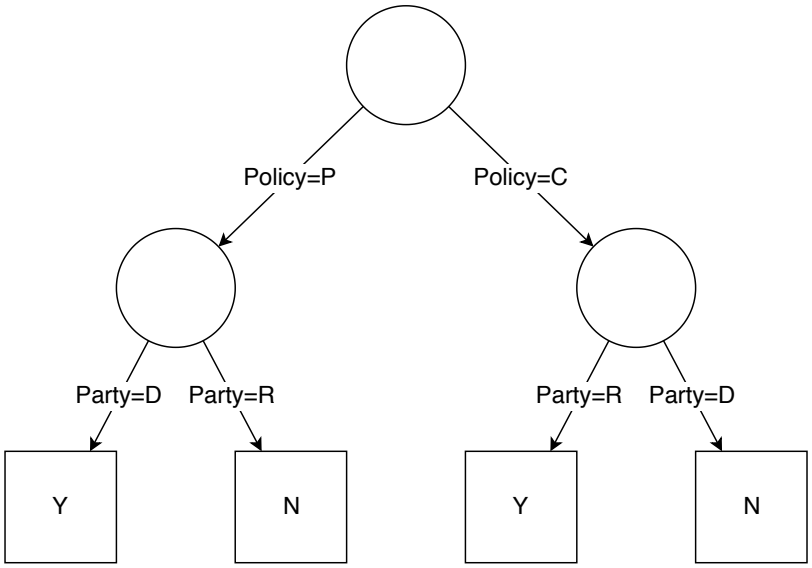



Figure 6.1: Example Decision Tree

where $p(c)$ is a proportion of the data with class $c$, or

$$p(c) = \frac{|d_c \in D|}{|D|}.$$

ID3 works by selecting the test $t$ that maximizes the information gained when partitioning the data into sets $D_t$ using

$$H(D) - \sum_{D_t} p(D_t) H(D_t).$$

Since the entropy of $D$ is fixed for dataset $D$, this is equivalent to minimizing the entropy of partitions $D_t$.

ID3 builds trees in an iterative fashion. Beginning with the root of the tree, the attribute that maximizes the information gain is selected to partition the data. The same procedure is applied to the resulting partitions until a stopping criterion is met.

Another method for determining how to split data at each point is Gini impurity,

which is used in the Classification and Regression Tree (CART) algorithm [98]. The principle behind this metric is to minimize the impurity at each split. The impurity is defined by the probability of a data point being associated with the wrong class when randomly assigning a class to each member in the partition based on the class distribution [100]. If the partition contains only one class, then the impurity would be zero as the class distribution would allow only that class. The measure is worst when there is an even split of classes in the partition. More formally, the probability of misclassification is defined as

$$\phi = 1 - \sum_{j=1}^{n} \left(p\left(c_j|t\right)\right)^2$$

where $p\left(c_j \mid t\right)$ is the probability of assigning the incorrect class based on the class distribution in partition $t$.

### 6.1.2 Random Forests

One issue decision trees can have is that they overfit the data unless pruned. Random forests were originally created as an ensemble method in an attempt to avoid overfitting and increase generalization [101, 102]. Random forests work by creating an ensemble of decision trees that each vote on the predicted class. Each of the decision trees uses a bootstrap aggregated (bagged) sample of the dataset. The bagged method samples with replacement from the original data to create an equal sized dataset for training. This results in approximately one-third of the data being left out for each tree. In the method used within this dissertation, each tree randomly selects a subset of the features for split at each point. Performance of training a random forest can be estimated by the out-of-bag error rate. This rate $OOB$ is defined by the prediction error of each training sample $x_i$ using those trees

that did not include $x_i$ in its bootstrap dataset.

## 6.2 Dataset Preparation

The dataset of political contributions used in this chapter is the one provided by Bonica and Stanford's Social Science Data Collection [78, 94]. As provided, the Stanford dataset uses two separate identifiers for candidates and donors. Using them separately would cause duplication of nodes in the network as an individual in the network may both donate and receive money. The candidate (or recipient) data include the donor id so that it is possible to bridge the two separate ids and create a single unique identifier for an entity. This new identifier is used to populate the edges in the induced social network.

In performing the analysis for CFScores, Bonica placed certain restrictions upon the dataset. Similar restrictions are applied to the networks here. Specifically, since loans and similar records do not necessarily indicate support of a candidate, they are removed from consideration. The included list of transaction types are as follows: 10, 11, 12, 13, 15, 15C, 15E, 15F, 15I, 15J, 15L, 15PD, 15S, 15T, 15Z, 18G, 18H, 18J, 18K, 18S, and 18U. A small description of these types are shown in Table 6.1 as reported by the US Federal Election Commission and Bonica. From this data, an initial network is created out of all the donations where the edge is weighted as $a_{ij} = amount$ where $amount$ is the sum of all donations or receipts between entities $i$ and $j$. As done in the previous chapters, any node $i$ with degree $d_i = 1$ is removed from the network. The largest connected component of the remaining network is then used as the network of contributions. Below is a summary of the procedure to generate each network.

1. Load the contributions for each two-year cycle into a database.

Table 6.1: List of Transaction Types used in the Analysis

| Type | Description |
|---|---|
| 10 | Contribution to Independent Expenditure-Only Committees |
| 11 | Native American Tribe contribution |
| 12 | Non-federal other receipt |
| 13 | Inaugural Donation Accepted |
| 15 | Contribution to political committees |
| 15C | Contribution from Candidate |
| 15E | Earmarked contributions to political committees |
| 15F | Loans forgiven by candidate |
| 15I | Earmarked contribution passed on to committee |
| 15J | Recipient committee's percentage of contribution from an individual |
| 15T | Earmarked contribution entered into intermediary's treasury |
| 15S | Contributions to state elections |
| 15L | Contributions to local elections |
| 15PD | Contribution made as payroll deduction |
| 15Z | In-kind contribution received from registered filer |
| 18G | Transfer in from affiliated committee |
| 18H | Honorarium received |
| 18J | Recipient committee's percentage of contribution to joint fundraising committee |
| 18K | Contribution received from registered filer |
| 18S | Receipts from Secretary of State |
| 18U | Contribution received from unregistered committee |

2. Attach the contributions to the recipient and donor information.

3. Create a unique identifier by merging donor and recipient ids.

4. Build the initial edges in the network, limited to relevant transaction types.

5. Determine the degree $d$ of each node $i$ and remove any with $d_i = 1$.

6. Using breadth first search, find the largest connected component of the remaining network.

7. Extract the list of edges of the largest connected component as the final graph.

Using all years from 1979 to 2012, the resulting network consists of over 5.26 million nodes and 29.85 million edges. Each node in the network is attached to the voting data by use of the Inter-university Consortium for Political and Social Research identifier (ICPSR id). The voting data includes 9.2 million observations of 'Yea' or 'Nay' votes on various bills from the years 1979 through 2012. Each bill also contains two features that are the spatial estimates of the socioeconomic measure: DW-NOMINATE midpoints $mid1$ and $mid2$. These are the ideological estimates generated from the voting behavior of legislatures as given by Voteview. The final voting dataset is created by combining the discovered communities' assignments, the DW-NOMINATE ideological scores, and the predictive class variable of 'Yea' or 'Nay'.

## 6.3 Experimental Design

The entire process of performing the prediction requires two primary steps. The first step is to perform hierarchical fuzzy spectral clustering on the campaign contributions network to find the community assignments for the legislatures [23]. In this experiment, communities are found for $k \in [2, 12]$. These values are then attached to the corresponding entries in roll call data.

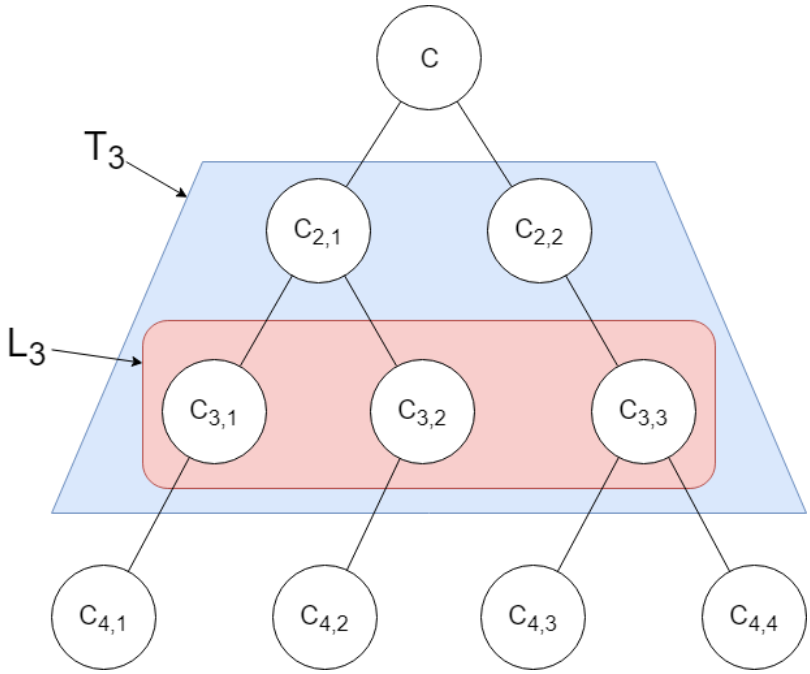



Figure 6.2: Example Hierarchy

An example hierarchy is shown in Figure 6.2. At the top level is the entire network. Level two splits the network into two overlapping communities. At each increasing level one more community is added to the network. For the purposes of the research herein, there are two important types of community assignment values within the tree. $T_i$ includes all assignments up to level $i$ in the tree. $L_i$ is defined only by those assignments on level $i$ itself. As shown in the example, the portion labeled by $T_3$ is the tree hierarchy that includes $L_3$ and all ancestors of that hierarchical level. All of the communities encapsulated by $T_3$ are used in vote prediction instead of just the lowest level of that sub-tree. This is in contrast to the portion labeled by $L_3$ that contains only the communities at level three.

In this paper, the communities are used to predict new behavior of the actors within the network. The behavior in question pertains to voting behavior within the legislature. Performing this classification relies on additional data provided by

Voteview in the form of socio-economic estimates of bills and voting records of United States legislators [76]. The resulting dataset contains a record for each recorded vote in the legislature. Each record also contains the fuzzy assignment values for that legislator, as calculated by $\mathbf{U}$ with varying $k$, the two DW-NOMINATE dimensions for the bill being voted upon. With each record is the class to be predicted: a 'Yea' or 'Nay' vote.

The federal datasets used in this chapter are considerably larger than the state data used in previous chapters. In 2012 there are over 1.36 million nodes and 6.15 million edges. Due to the large nature of the datasets, an approximate eigen-solver was used to find the eigenvectors of the network to obtain only the vectors and values of interest. The following chapter describes a graph embedding that avoids calculating eigenvectors for the entire matrix at once. A value of $m = 10$ for fuzzy $c$-means was used to find the clusters. This was set to ensure highly fuzzy communities instead of closer to crisp clusters.

Four different primary experiments were performed to investigate properties of the classifier and data. These include 1) different weighting of the network edges, 2) using all data combined and individual years separately, 3) varying community types and numbers, and 4) the performance of random forest when compared to a single decision tree. For the experiments, each combination of experimental parameters (3,564 trials) was tested using 10-fold cross validation.

### 6.3.1 Experiment 1: All vs Yearly Data

Along with the edge weighting, a test is performed to compare the results of testing the entire dataset against individual election cycles. The full network utilized the entirety of the data from all available years. Separate networks were also created for each individual 2-year cycle from 1979 through 2012 as provided in the data. This

created 17 different networks. The sizes of the networks greatly increase over time due to the increase in the amount of money in politics. After performing the same processing step for each cycle, the number of nodes ranged from approximately 34 thousand in 1980 to 1.4 million in 2012. Similarly, the number of edges ranged from 169 thousand in 1980 to 6.2 million in 2012. Table 6.2 shows the growth of both nodes and edges over time.

### 6.3.2 Experiment 2: Classifier

In addition, a test was done to compare the relative performance of a single decision tree to a random forest. Each classifier used implementations provided by the *sklearn* package of Python: DecisionTreeClassifier and RandomForestClassifier. We do not use early stopping or pruning in either classifier. Early experiments showed that the community assignment features were not equally expressive when trying to predict votes. Using the common method of limiting the number of features available during a split to $\sqrt{(n_{features})}$, however, resulted in poor performance. Thus, the algorithms were allowed to use any feature when determining how to perform a split. Both classifiers use the Gini impurity metric when performing splits.

Another issue in the data is that the class distribution for 'Yea' and 'Nay' is skewed. There are approximately 6 million 'Yea' votes and 3.2 million 'Nay' votes for the relevant years. This can be a problem as the skewed data may result in the classifier defaulting to the dominant class and reducing performance. To prevent this issue, both the decision tree and random forest weight the classes. The class weight is inversely proportional to the number of instances of that class. The final parameter to note is that the random forest used 50 trees. Both the random forest and decision trees were evaluated using 10-fold cross-validation.

Table 6.2: Size of Networks by Year, in thousands

| Cycle | Num of Nodes | Num of Edges |
|---|---|---|
| 1980 | 34 | 169 |
| 1982 | 21 | 142 |
| 1984 | 31 | 192 |
| 1986 | 32 | 208 |
| 1988 | 51 | 271 |
| 1990 | 70 | 335 |
| 1992 | 115 | 510 |
| 1994 | 116 | 514 |
| 1996 | 196 | 892 |
| 1998 | 359 | 1,570 |
| 2000 | 435 | 1,901 |
| 2002 | 604 | 2,494 |
| 2004 | 728 | 2,957 |
| 2006 | 779 | 3,249 |
| 2008 | 1,026 | 3,992 |
| 2010 | 1,174 | 4,944 |
| 2012 | 1,362 | 6,154 |

### 6.3.3 Experiment 3: Communities

For each network type, three different applications of the spectral clustering were applied. The first of these methods used a single level of the hierarchical tree (shown in the figures as FCM-L). This represents the process of using the best single performing clustering as would be typical in many hierarchical clustering techniques.

The second included all parents of the level in question, effectively pruning anything further down the tree (shown in figures as FCM-A). Finally, as a point of comparison, the eigenvectors themselves were used directly in performing prediction (shown in figures as EV). This was chosen as a baseline to show any loss in predictive power by using fuzzy $c$-means.

### 6.3.4 Experiment 4: Edge Weighting

Three different weighting schemes were analyzed with the network in order to analyze different effects on the results. The first weighting scheme just used 1 to indicate an edge between two entities in the network, regardless of the amount of the contribution. The second used the base-10 logarithm of the total amount of contributions between two entities as $a_{ij} = \log_{10}(amount)$. Prior to the calculation, any amount less than \$10 was raised to \$10 to ensure a minimum value of 1 on the edges. Finally, the raw amount $a_{ij} = amount$ of total donations between two entities was used, again with a minimum value of 1. By analyzing these different measures, we can infer how connectivity versus donation amount impacts the prediction of voting behavior.

## 6.4 Results and Discussion

Figure 6.3 shows the results of predicting votes over the matrix formed from all years of data. In nearly all cases, the random forest ensemble outperformed the single decision tree, which was as expected. The more interesting result comes from the differing behavior of the edge weights on the network. For just the adjacency matrix, performance began by middling but improved until about $k = 10$ communities. At this point, the performance of the single level (FCM-L) began to drop. However, the eigenvector (EV) and full tree (FCM-T) held steady due to not losing the information

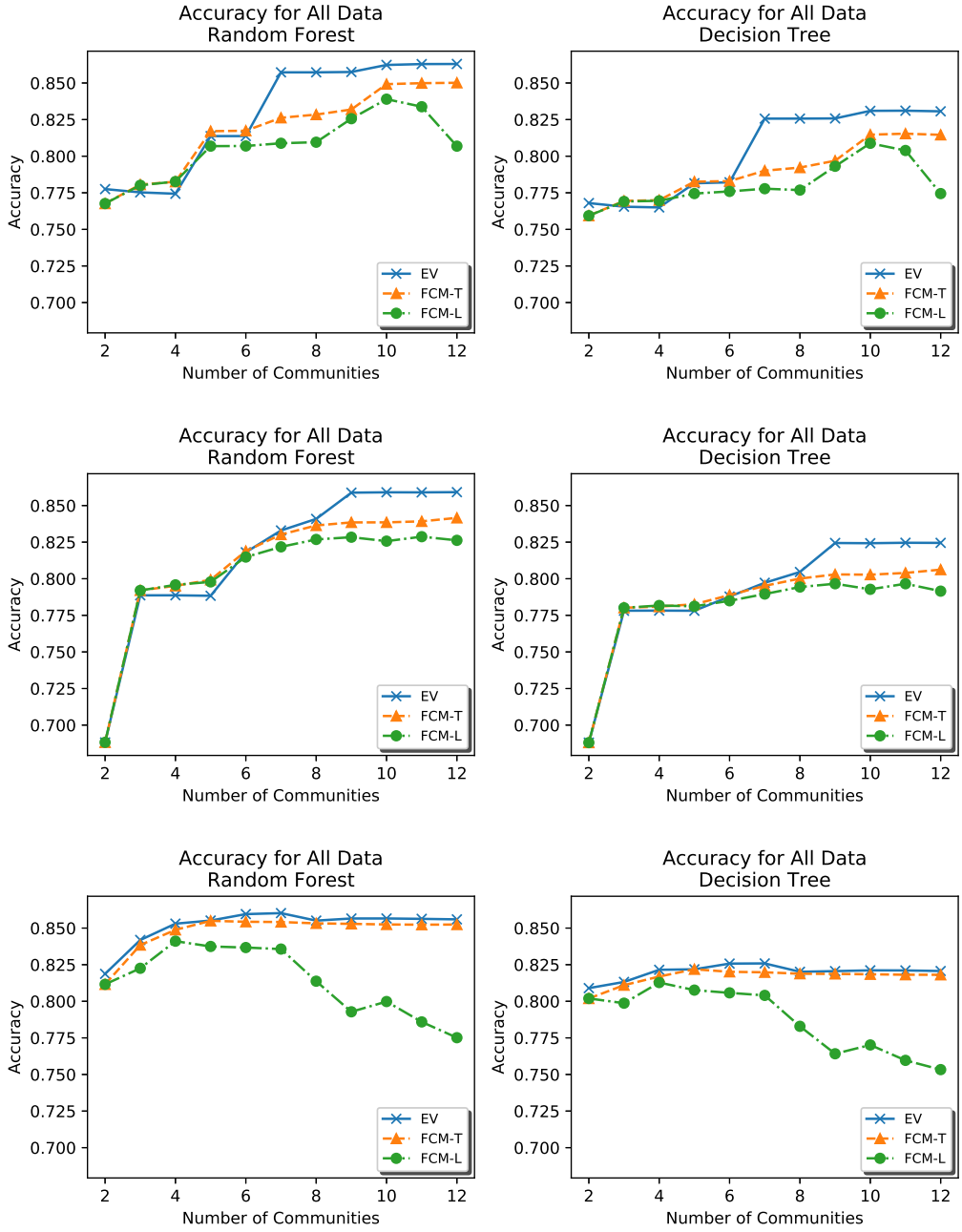



Figure 6.3: Results from using all years of data.

provided by fewer communities.

The raw contributions show almost all the predictive power of these communities

is held within the first few levels of communities. The large donations appear to dominate the community analysis at that stage, and additional communities do not provide much in additional predictive ability. This can be shown in that after $k = 7$, the performance of the single level drops rapidly.

Unlike the other two, the scaled contributions performed poorly at first for this particular set of data. As shown by EV, this was not due to the performance of FCM, but integral to the structure of the network. As the tree grew, however, the performance improved and matched that of the other two weightings.

Breaking out the votes by year showed additional evidence of polarization within the legislature. Beginning with 1980, the prediction accuracy was lower than that of subsequent years. In general, the accuracy increased as time passed, hitting a peak in 2010 with votes being predicted at roughly 94% for each of the weighting methods. This fell back to approximately 90% to 91% in 2012, which was more in line with 2008. The performance of the differing number of communities also flattened over time, implying that fewer communities are necessary to define splits between members of Congress.

### 6.4.1 All vs Yearly Data

A little more information was necessary to compare the models learned from all the data to those trained on two-year cycles. When calculating the performance of the all-data models, the accuracy of each individual year was calculated as well as that of the entire dataset. Using the prediction of the entire dataset, we can determine if individual years predict votes better for that year than the larger dataset that covers many years. The notable result from this experiment is that using all the data at once did not typically do better than using data in a specific year. Instead, out of the different comparisons, the models built from individual cycle data were statistically

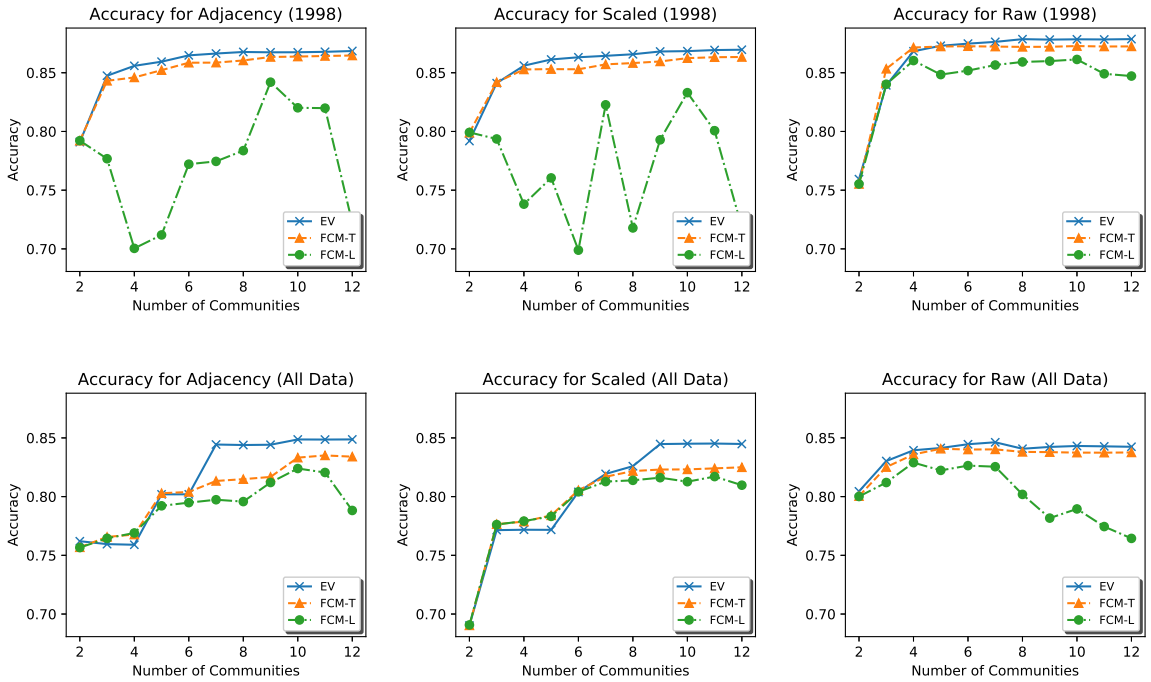



Figure 6.4: Comparing models trained with 1998 data vs. all data on votes in 1998

better in 91.7% cases. Figures 6.4 and 6.5 show this trend from examples of these experiments.

The results of the experiments show another issue related to polarization of the legislature. This can be seen from two different aspects of the results from individual cycles. First, the overall accuracy of prediction increased over time. This hit a peak in year 2010, though it was still quite high in the following 2012 cycle. Additionally, the impact of the number of communities is less pronounced in those later years. More community information here is not helpful as most of the important information is captured by the small numbers of communities. This is true even for FCM-L, which mostly stops behaving as in prior years where there are peaks and valleys in performance based on the number of communities.

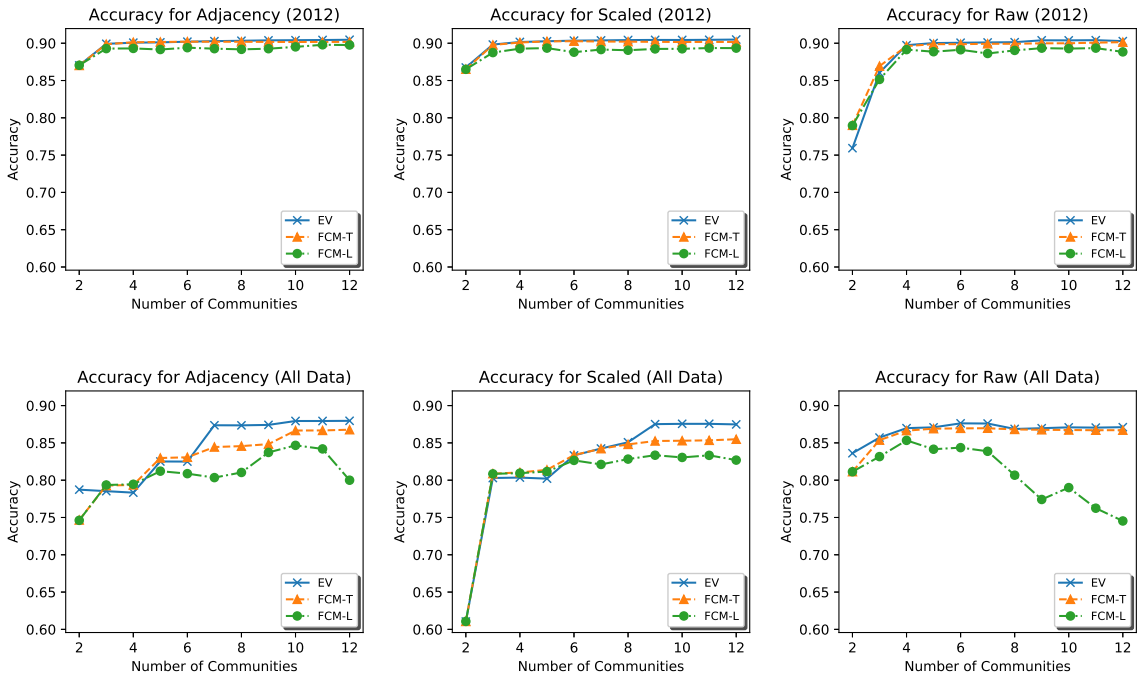



Figure 6.5: Comparing models trained with 2012 data vs. all data on votes in 2012

### 6.4.2 Classification

As expected from the results, the random forest outperformed the single decision tree fairly consistently. It should be noted, however, that the random forest takes considerably longer to build. In this case, it was building $n = 50$ individual decision trees. This was especially noticeable on the larger datasets. Viewing the side-by-side results as shown in Figure 6.3, the average performance across the folds for random forest was considerably higher than that of the decision tree. Analyzing the folds with the Student's t-test shows that in all but three cases the random forest had significantly better results at $\alpha = 0.05$. All of those occur in the scaled weighting scheme at $k = 2$ communities. Table 6.3 shows the performance of statistical tests for all weightings and types of communities at $k = 2$. As mentioned in the prior section, this is likely due to the structure of the network and not the performance of FCM as

Table 6.3: Accuracy with $k = 2$ Communities using All Data

| Weight | Type | Random Forest | Decision Tree |
|---|---|---|---|
| Adj | EV | $\mathbf{0.778 \pm 4.49E-04}$ | $0.768 \pm 5.74E-04$ |
| Adj | FCM-T | $\mathbf{0.768 \pm 4.21E-04}$ | $0.759 \pm 8.26E-04$ |
| Adj | FCM-L | $\mathbf{0.768 \pm 4.09E-04}$ | $0.759 \pm 8.31E-04$ |
| Scaled | EV | $0.688 \pm 3.88E-04$ | $0.688 \pm 4.10E-04$ |
| Scaled | FCM-T | $0.688 \pm 3.97E-04$ | $0.688 \pm 4.10E-04$ |
| Scaled | FCM-L | $0.688 \pm 3.41E-04$ | $0.688 \pm 4.10E-04$ |
| Raw | EV | $\mathbf{0.819 \pm 3.64E-04}$ | $0.809 \pm 4.96E-04$ |
| Raw | FCM-T | $\mathbf{0.812 \pm 3.68E-04}$ | $0.802 \pm 3.31E-04$ |
| Raw | FCM-L | $\mathbf{0.812 \pm 3.84E-04}$ | $0.802 \pm 3.55E-04$ |

all the different types of clustering and vectors perform similarly.

Out of the 1,782 different comparisons between random forest and decision trees, the decision tree outperformed random forest in only 33 instances to a significant degree at $\alpha = 0.05$. These occurred in varying years, with all but 5 of those examples at $k = 2$ communities. In all these cases, despite the statistically significant difference, the relative performance difference is not large. The largest difference in average accuracy between random forest and the decision tree is only 1.01%. Based on these results, the remaining experiments consider only random forests.

### 6.4.3 Communities

As shown in Figure 6.3, increasing the number of communities (or vectors used) improved the classification accuracy of the vector (EV) and whole tree models (FCM-T). Both models showed consistent performance where adding more features helped the performance, with only small deviations from this pattern. However, this was

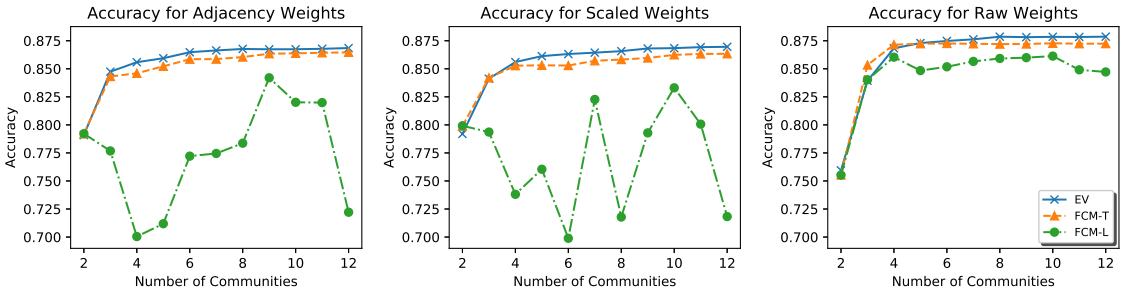



Figure 6.6: Results for 1998

not true when using single levels of the hierarchy (FCM-L). In two of the weighting schemes, the accuracy increased with the number of communities until a certain point at which it fell sharply. It is expected that as the number of communities grow in the scaled weighting, a similar pattern will hold, and the accuracy will begin to drop. The reason for this is that if the number of communities discovered exceeds that of the actual strong community structure in the network, then the resulting assignments will be not be of use in classification.

As a more extreme example, consider the results from the experiments using only data in 1998 from Figure 6.6. While the performance of FCM-L was consistent in the case of raw contributions, the performance in both the adjacency matrix and the scaled values were highly erratic. This can occur when the number of communities used does not match the proper number of communities at that resolution. This behavior can explain the performance of FCM-L in these cases. As can be seen, for certain numbers of communities, the accuracy improved to nearly match that of EV and FCM-T. However, since both EV and FCM-T have access to the entirety of the data up to that point, those two methods are far more robust to those changes. Due to these issues, the selection of $k$ for FCM-L is far more important than in the other methods and should be done carefully.

Another notable difference between FCM-L and FCM-T can be seen in the

behavior in the low number of communities. Moving down the hierarchy, the performance of FCM-L decreased for $k = 3$ and $k = 4$ for both the adjacency and scaled weights. However, despite using those same values from the hierarchy of community assignments, the performance of FCM-T increased during that period. This shows that FCM-T is more robust to individually poor community assignments and that the combination of data from the levels is more useful.

In general, the best performance was obtained using the eigenvectors directly. This was not always the case, however, especially for small numbers $k$ of communities. FCM-T was typically comparable in performance and did not suffer the same issues as FCM-L. EV was more easily computable due to the number of features required. With EV, $k$ indicates the number of eigenvectors to use. For FCM-T, the entire tree structure is used up to the level with $k$ communities. This results in $k\,(k-1)\,/2$ features. This increase in the number of features, along with the large amount of voting data, resulted in FCM-T taking longer to compute the models than EV.

Note, however, that interpretability is an important factor in the selection and use of models, and FCM-T has the benefit of being easier to interpret. Two vectors representing two different entities does not have meaning out of the larger context. However, two lists of community assignments for those same entities can be more easily interpreted for anyone attempting to use these data or methods.

### 6.4.4 Edge Weighting

Edge weighting plays a significant role in the accuracy of prediction across the different experiments. When looking at the entire dataset, the raw contribution amounts yielded the best performance when using few communities. Using the analysis from the prior section, however, it is possible to see that this performance was dominated by those few communities. Increasing past $k = 7$, the performance of

the raw weights dropped rapidly in FCM-L, indicating those communities are not well representative of voting behavior. This drop happened much later for the adjacency matrix at $k = 11$. Despite the poor performance of the scaled data at the start, it improved quickly, matched the other weighting schemes, and had not fallen in performance by $k = 12$.

Overall, the raw values outperformed the other two measures, though not by much. Out of the possible comparisons, raw weighting was better statistically than the adjacency matrix 53.1% of the time and better than scaled 50.8% of the time. Adjacency was better than raw 34.3% of the time and scaled was better than raw 36.9% of the time. From these results, the best communities for predicting votes were more often those generated by large, raw dollar amounts. This impact is small, however. Even though the different weights had an impact in different tests, in aggregate the results from edge weighting were quite similar in their performance.

## 6.5 Conclusion

As shown, the clusters discovered by hierarchical fuzzy spectral clustering were generalizable to behavior not present in the underlying system that created the graphs. This generalizability was demonstrated by predicting voting behavior of legislators based on campaign finance records. In addition, the expressive power of the full hierarchy was demonstrated by its superior performance to the predictions given by a single level in the hierarchy. This full hierarchy performance nearly equals using eigen-vectors directly, while providing interpretability for those who use the data. While different weightings of the edges had an impact on the results of the predictions, overall the best performing weighting was based on raw dollar amounts. However, both the adjacency matrix and scaled weights were promising. The results also again highlight the growing issue of partisanship within the legislature.

# CHAPTER SEVEN

## ORTHOGONAL SPECTRAL AUTOENCODER FOR GRAPH EMBEDDING

In this chapter we examine approximate methods for finding encodings for the social networks. There are a couple of notable limitations in the previous methods for calculating encodings. One of the issues is that these methods for eigenvector embedding do not easily generalize to new data. Another issue is in the cost of calculating the eigenvalues and eigenvectors themselves.

Here we develop a mini-batch spectral embedding method using an orthogonal spectral autoencoder (OSAE). Two different versions of the autoencoder is tested, one with an orthogonal constraint and loss on the encoding layer (in addition to the reconstruction error). The second version adds an approximate spectral decomposition by creating a smaller sampled Laplacian and performing eigenvector decomposition on the smaller Laplacian. This adds a loss term for the difference between the orthonormal encoding and the top-$k$ eigenvectors of the approximate Laplacian.

The results of these experiments show that the network embeddings are effective in predicting behavior of actors within the graph. In the testing, we cluster the embeddings using fuzzy c-means. The resulting fuzzy clusters are used as community assignments for the actors within the graph. Once again we use the federal campaign finance networks and voting history for the legislators within the graph. Each legislator is assigned to their communities based on the results from fuzzy c-means. Those community assignments are used as an ideological approximation and combined with the voting data to generate a classifier for voting behavior. The results are comparable to the hierarchical spectral decomposition from prior chapters.

## 7.1 Background

Some of the limitations of spectral embedding was discussed in Chapter 2. Considerable work has been done by other researchers to mitigate some of the issues of complexity and out-of-sample embedding. The following section discusses some related work done by these other researchers in related areas of dimensionality reduction, approximate spectral clustering, and out of sample augmentations.

### 7.1.1 Deep Learning and Autoencoders

A considerable amount of research has been performed recently in the area of neural networks and autoencoders. These networks have shown to be very effective in a variety of tasks. Notably for the work in this dissertation, these tasks include dimensionality reduction, function approximation, and graph embedding. Early work on multilayer feedforward networks showed that such a network can be a universal approximator of a function, provided it has sufficient parameterization [103].

Training these networks can be efficient thanks to the use of graphical processing units (GPU) which are able to parallelize the matrix computations. This is important as passing a mini-batch of size $b$ to the network requires matrix multiplication at each layer, making the computation $\mathrm{O}\left(b \times n \times m\right)$. Training multiple batches requires multiple runs, resulting in $\mathrm{O}\left(b \times n \times m \times e\right)$. In practice, the performance of training the network over a GPU is quite effective, however.

Autoencoders are one example of using neural networks for dimensionality reduction. These are typically neural networks used in unsupervised machine learning where the autoencoder learns a representation of the data by encoding the information in a smaller feature space. Bengio introduced a method for training deep autoencoders by training individual layers and stacking the individual autoencoders [104]. The

results showed improved performance than when trying to train entire deep networks at once. Other research has shown the relationship between autoencoders and principal component analysis [105]. Using a bottleneck layer in the autoencoder, the network shown was able to eliminate nonlinear correlations in the data and reduce the dimensionality of the embedded.

Many other methods have been examined for using neural networks in machine learning. Other researchers have tried to develop functional embedding by using classification loss and pairwise point similarity [106].

### 7.1.2 Approximate Spectral Clustering

As mentioned, since spectral clustering on large datasets is difficult due to space and time complexity, there have been several approximation methods for spectral clustering developed. Nyström approximation is one such algorithm used to calculate the eigenvector decomposition [107]. This has been adapted for use in spectral clustering [108].

In the fast spectral clustering algorithm proposed by Choromanska et al. (Algorithm 7.1), the process begins by selecting $l$ columns sampled uniformly without replacement from the affinity matrix. This creates matrix $\hat{\mathbf{A}}$. Two diagonal matrices $\mathbf{D}$ and $\mathbf{\Delta}$ are created based on the row sums of $\hat{\mathbf{A}}$. An approximate Laplacian $\hat{\mathbf{C}}$ is calculated using the two diagonal matrices and the sampled columns. Then the best r-rank approximation is taken of W (usually singular value decomposition) to obtain approximate eigenvectors given by $\hat{\mathbf{U}}$. The remainder of the procedure is the same as standard spectral clustering where the column matrix containing the eigenvectors is row normalized followed by clustering the resulting matrix.

Similar work regarding Nyström approximations is used in other work [109]. In this work the authors treated a sample of the data as *landmarks*. These landmarks

**Algorithm 7.1** Fast Spectral Clustering

**function** FastSpectralClustering($\mathbf{A}, k, l, r$)
  $\mathbf{L} \leftarrow$indices of $l$ sampled columns ▷ Sample columns from $\mathbf{A}$ uniformly without replacement
  $\hat{\mathbf{A}} \leftarrow \mathbf{A}(:, \mathbf{L})$
  $\mathbf{D} \in \mathbb{R}^{n \times n} : \mathbf{D}_{ij} = \delta[i = j]\, 1/\sqrt{\sum_{j=1}^{l} \hat{\mathbf{A}}_{ij}}$ ▷ Diagonal $\mathbf{D}$ as row sums of $\hat{\mathbf{A}}$
  $\boldsymbol{\Delta} \in \mathbb{R}^{l \times l} : \mathbf{D}_{ij} = \delta[i = j]\, 1/\sqrt{\sum_{j=1}^{l} \hat{\mathbf{A}}_{ij}}$ ▷ Diagonal $\boldsymbol{\Delta}$ as row sums of $\hat{\mathbf{A}}$
  $\hat{\mathbf{C}} \leftarrow \hat{\mathbf{I}} - \sqrt{\frac{l}{n}}\mathbf{D} \times \hat{\mathbf{A}} \times \Delta$
  $\mathbf{W} \leftarrow \hat{\mathbf{C}}(\mathbf{L}, :)$
  $\mathbf{W}_r \leftarrow$ best r-rank approximation to W
  $\tilde{\Sigma} = \frac{n}{l}\Sigma_{W_r}$ and $\tilde{\mathbf{U}} = \sqrt{\frac{l}{n}}\hat{\mathbf{C}}$
  $\mathbf{Y} = \forall i X_{ij} / \|\mathbf{X}_i\|$ ▷ Normalize the rows of $\mathbf{X}$.
  **return** $\mathbf{U} \leftarrow K - \text{means}(\mathbf{Y})$ ▷ Return clusters.
**end function**

were drawn from the exact eigenvector decomposition as calculated from the data. A linear transformation from the landmark set to the full set is applied to approximate the spectral embedding of the original data.

One relevant example for work with autoencoders is that of mini-batch spectral clustering [110]. As the authors note, calculating the Laplacian of a dataset as well as the spectrum can be $O(n^2)$ in storage for the Laplacian, and $O(n^3)$ time complexity to calculate the spectral decomposition. The primary motivation is to calculate the spectrum of the Laplacian by finding the principal eigenvectors $\mathbf{Z}$ without the direct calculation. They note this can be reworked to optimize the following trace problem

$$\arg\min \left\{ \text{Tr}\left( -\frac{1}{2}\mathbf{Z}^\top \mathbf{L}\mathbf{Z} \right) \right\} : \mathbf{Z}^\top \mathbf{Z} = \mathbf{I}.$$

**Algorithm 7.2** Stochastic Riemannian Gradient with Mini-Batches

**function** HTILDE($\mathbf{L}, p, N_r, \mathbf{W}$)
  Initialize $\tilde{\mathbf{G}} \in \mathbb{R}^{n \times k}$ with all elements zero
  **for** $i = 1$ to $N_r$ **do**
    Sample the components of $\mathbf{r}_i$ ▷ Randomly select columns
    $\tilde{\mathbf{G}} += \frac{1}{N_r} \mathbf{L}\mathbf{r}_i\mathbf{r}_i^\top\mathbf{W}$ ▷ Calculate the gradient using sampled Laplacian
  **end for**
  **return** $\tilde{\mathbf{H}} = \left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)\tilde{\mathbf{G}}$
**end function**

Specifically, they note that the orthonormality constraint causes $\mathbf{W}$ to lie on a Stiefel manifold. The Riemannian gradient on this manifold is $H = \left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)\mathbf{G}$. This allows some theoretical guarantees that a stochastic gradient optimization will converge. In the case of this work, the authors use the normalized Laplacian $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}A\mathbf{D}^{-\frac{1}{2}}$. In this form, the top-$k$ eigenvectors are those from the largest $k$ eigenvalues. Algorithms 7.2 and 7.3 describe their algorithm for calculating the approximate eigenvectors. In Algorithm 7.2, they calculate the stochastic gradient using samples of the Laplacian. The $\mathbf{r}_i$ vectors introduce the stochasticity by only selecting certain columns of the Laplacian. The gradient is used in Algorithm 7.3 to update the estimate the embedding and then project the space back onto the Stiefel manifold by way of the QR decomposition. They use this formulation to appeal to results in Bonnabel [111] such that convergence of the method can be proved in the limit of iterations.

One possible solution to the out-of-sample issue for spectral embedding a graph is to use neural network to find a function $f(x)$ mapping a data point to the corresponding row of an eigenvector decomposition. Using the top-$k$ eigenvectors,

**Algorithm 7.3** Mini-Batch Spectral Clustering

**function** MINI-BATCH SPECTRAL CLUSTERING($\mathbf{L}, k, \epsilon$)
  Initialize $\mathbf{W}^{(0)} \in \mathbb{R}^{n \times k}$ as a random orthonormal matrix
  Initialize $\mathbf{M}^{(0)} \in \mathbb{R}^{n \times k}$ with all elements zero
  **for** $t = 1$ to $T$ **do**
    $\tilde{\mathbf{H}}^{(t)} = \text{Htilde}\left(\mathbf{L}, p, N_r, \mathbf{W}^{(t-1)}\right)$
    $\mathbf{M}_{ij}^{(t)} = \mathbf{M}_{ij}^{(t-1)} + \left|\tilde{\mathbf{H}}_{ij}^{(t)}\right|^2$
    $\hat{\mathbf{H}}_{ij}^{(t)} = \frac{\tilde{\mathbf{H}}_{ij}^{(t)}}{\epsilon + \sqrt{\mathbf{M}_{ij}^{(t)}}}$
    $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \lambda \hat{\mathbf{H}}_{ij}^{(t)}$
    $\mathbf{W}^{(t)} = QR_Q\left(\mathbf{W}^{(t)}\right)$
  **end for**
  Apply $K$-means on $\mathbf{W}^{(T)}$
**end function**

the goal is to find $f(x) = \mathbf{z}_i$ where $\mathbf{z} \in \mathbb{R}^k$. Prior work has developed this idea with a stacked autoencoder [112]. This method is a two-step procedure wherein the initial step is to train the network to encode the graphs using a stacked autoencoder. Once the autoencoder is trained over the input data, the decoder portion is removed and the encoder is fine-tuned on learning the mapping from the input to the eigenvectors.

Deep embedded clustering attempts to simultaneously learn the set of cluster centers in the feature space along with the parameters that maps the data points into the reduced feature space [113]. Their algorithm alternates between updating the network parameters and cluster centers. Using a kernel derived from the Student's $t$-distribution to calculate similarity between the embedded points and the cluster centers, the authors use a Kullback-Leibler divergence loss function for a selected target distribution. These descriptions only cover a small fraction of the work being

performed in this field. There are many different techniques for attempting to use neural networks for clustering [114, 115], or improve the out of sample clustering performance using embedding [116, 117, 118, 119].

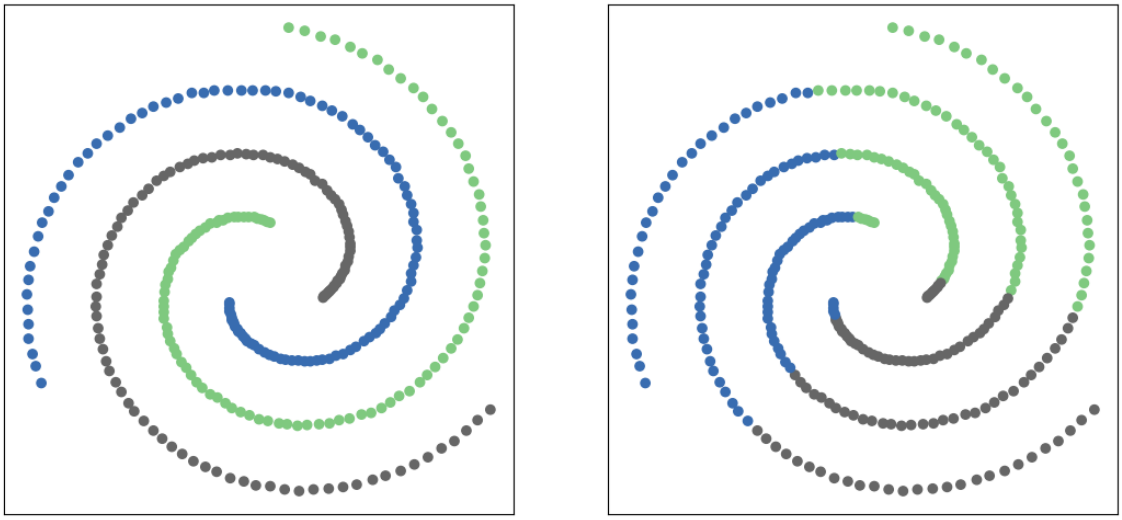
## 7.2 Kernels and Spectral Clustering

One of the benefits of using the adjacency matrix for spectral clustering is that it does not require use of an additional kernel function. This is useful as many kernel functions require tuning a bandwidth parameter to improve the clustering performance. However, in the Orthogonal Spectral Autoencoder, an approximate Laplacian is calculated by use of one of these kernel functions. Optimizing the selection of bandwidth for a kernel can be difficult. When there are known class labels for the data, it can be possible to tune the bandwidth by using multiple trials and selecting the best option. With unlabeled data, other metrics such as within-cluster similarity may need to be used to estimate the clustering.

$$\mathcal{K}_{rbf}\left(v_1, v_2\right) = \exp\left(\frac{-\left\|v_1 - v_2\right\|^2}{2\sigma^2}\right)$$

To highlight the effect of the bandwidth hyperparameter on clustering, consider a spiral dataset as shown in Figure 7.1a. This graph contains non-convex clusters and is not suited for clustering with algorithms that assume convex clusters such as $K$-means. Using $K$-means clustering directly over the dataset yields the poor clustering shown in Figure 7.1b.

Instead of performing $K$-means on the data points directly, spectral clustering uses a kernel function to find pairwise affinity among the individual data points. Using an optimal bandwidth $\sigma$ for this dataset yields the spectral clustering shown in Figures 7.2a and 7.2b. However, a poor selection of $\sigma$ has no better performance



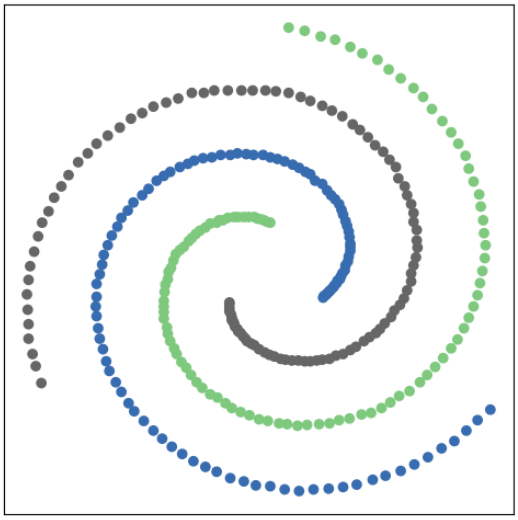
(a) Spiral Dataset (b) $K$-means on Spiral Dataset

Figure 7.1: Spiral Dataset and Convex Clustering

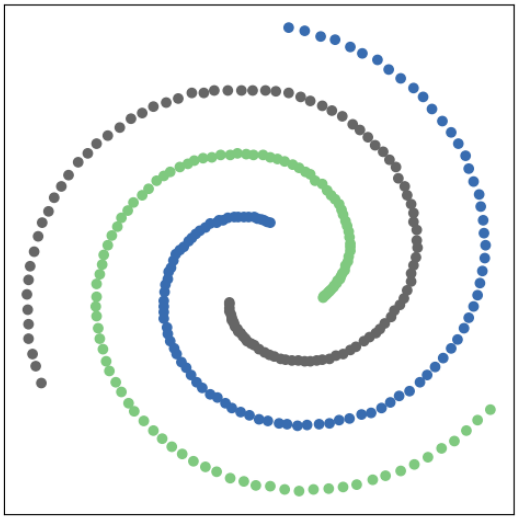
than $K$-means, as shown in Figures 7.2c and 7.2d.

Unfortunately, the results of the clustering here is worse than it appears as it regards fuzzy clustering. To better understand the effect of the bandwidth $\sigma$, consider the graphs of the eigenvectors of the spectral decomposition in Figure 7.3. This decomposition uses the normalized symmetric Laplacian $\mathbf{L}_{\text{sym}}$ defined in previous chapters. Despite spectral clustering finding the exact clusters for $\sigma = 0.5$ (Figure 7.3a) and $\sigma = 1$ (Figure 7.3b), the eigenvector decomposition is considerably different between the two. The small value of $\sigma = 0.5$ separates the three communities entirely, compressing all the data into the same point for each cluster. In contrast, there is some variation in the values of the eigenvectors for $\sigma = 1$. Increasing $\sigma$ further would have an effect of increasing the overlap between the communities, but at the expense of the overall clustering accuracy.

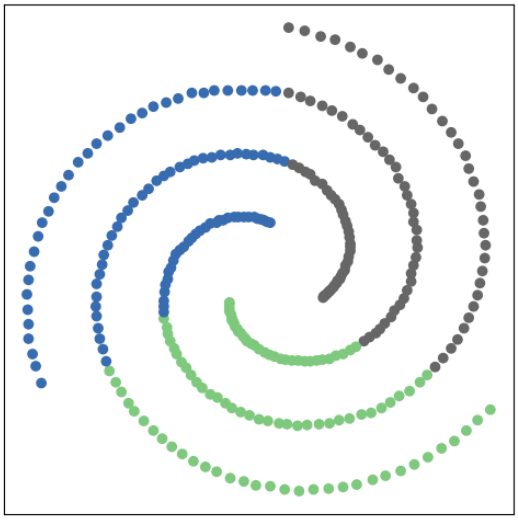
We attempt to mitigate the issue of tuning $\sigma$ in an unsupervised setting by use of methods for selecting an appropriate bandwidth prior to clustering. There are a



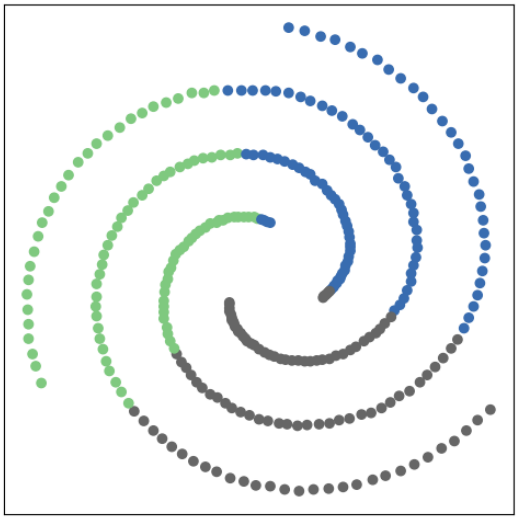
(a) Spectral Clustering on Spiral, $\sigma = 0.5$



(b) Spectral Clustering on Spiral, $\sigma = 1$



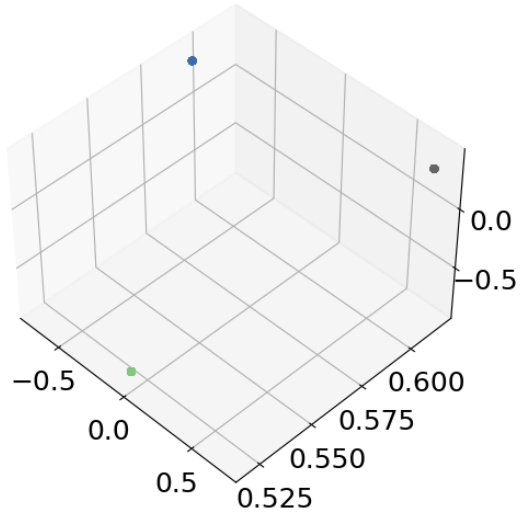
(c) Spectral Clustering on Spiral, $\sigma = 2$



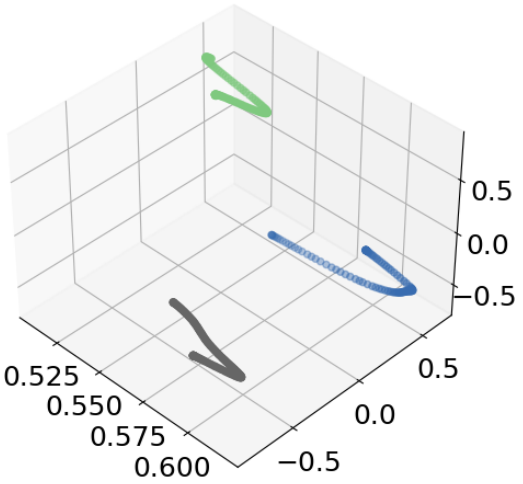
(d) Spectral Clustering on Spiral, $\sigma = 4$

Figure 7.2: Spiral Dataset and Spectral Clustering for Varying $\sigma$

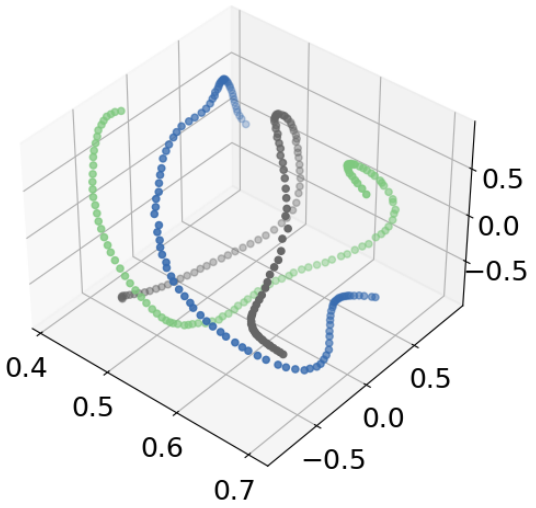
few methods that have been developed for selecting an appropriate bandwidth prior to performing the clustering, especially regarding the radial basis function kernel [120, 121]. In this dissertation, we make use of a self-tuning $\sigma$ as described by Zelnik-Manor and Perona[122].



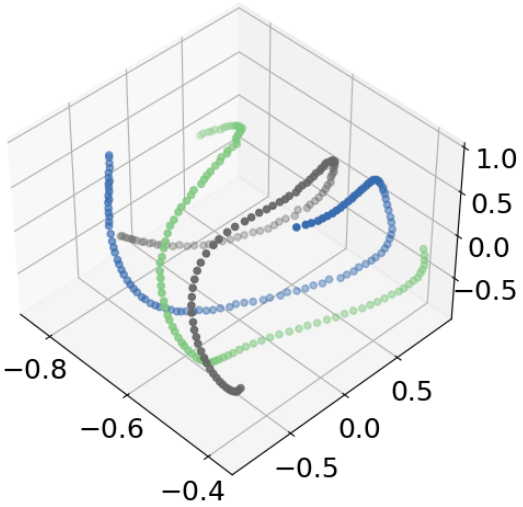

(a) Eigenvectors with $\mathcal{K}_{\text{rbf}}$ and $\sigma = 0.5$




(b) Eigenvectors with $\mathcal{K}_{\text{rbf}}$ and $\sigma = 1$




(c) Eigenvectors with $\mathcal{K}_{\text{rbf}}$ and $\sigma = 2$




(d) Eigenvectors with $\mathcal{K}_{\text{rbf}}$ and $\sigma = 4$

Figure 7.3: Spiral Dataset and Eigenvectors of $\mathbf{L}_{\text{sym}}$ for Varying $\sigma$

Local density adaptive similarity calculates a $\sigma_i$ for each vertex $v_i$ based on a distance to its local neighborhood. Consider the neighborhood of vertex $v_i$ as the nearest points based on Euclidean distance $\|v_i - v_j\| \, \forall \, v_j \in \mathbf{V}$ The $\text{n}^{\text{th}}$ nearest neighbor $v_n = NN\,(v_i)$ provides an approximation of the density of the neighborhood

around $v_i$. Using this principle, the scaling sigma is calculated by $\sigma_{ij} = \sigma_i \sigma_j$ where $\sigma_i = \|v_i - NN(v_i)\|$. The numerator in the RBF kernel remains the same, yielding a scaled kernel function

$$\mathcal{K}_{rbfs} = \frac{-\|v_1 - v_2\|^2}{\sigma_i \sigma_j}$$

The authors of that work claim that the seventh nearest neighbor of $v_i$ consistently provided good clustering performance.

## 7.3 Orthogonal Spectral Autoencoder

Our approach allows for out-of-sample clustering as well as limiting the size of the Laplacian needed in the calculations of the eigenvectors such that $l \ll n$. The primary framework for the orthogonal spectral autoencoder is a deep undercomplete autoencoder. In this architecture, there are multiple hidden layers between the input and output. The encoding layer is termed undercomplete since the dimension of this layer is smaller than the input layer. This structure can be useful in performing feature extraction and dimensionality reduction on the data. In general, this type of autoencoder uses a reconstruction loss based on the distance from values of the input and output, commonly mean squared error.

Figure 7.4 describes the general structure of the orthogonal spectral autoencoder. Our additions to this framework include another layer between the encoder and decoder that orthogonalizes the output of the encoding layer $\mathbf{E}$ by use of QR decomposition where $\mathbf{E} = \mathbf{QR}$. There are two additional components to the loss function that penalize encoding layers that are not orthogonal, as well as approximate the eigenvectors of a sampled Laplacian.

Each of the hidden layers within the networks are linear layers which apply the function $y = wH^\top + \beta$ where $w$ is the weight vector, $H$ is the input to the layer,

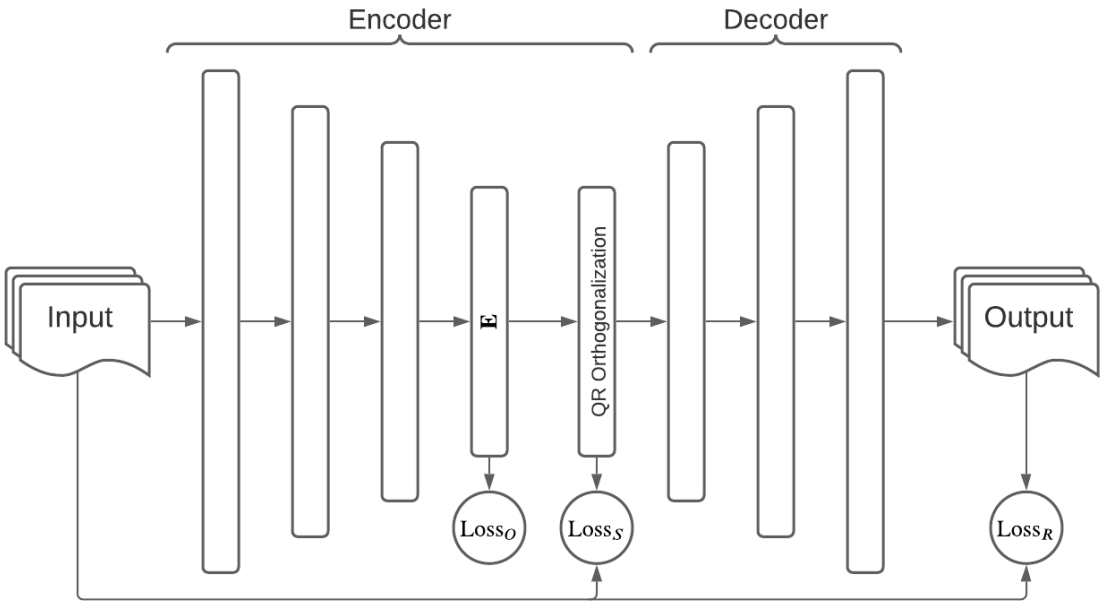



Figure 7.4: Structure of the Orthogonal Spectral Autoencoder

and $\beta$ is an additive bias. The output of each hidden layer uses the SELU activation function defined as

$$SELU\left(y\right) = s \times \left(\max\left(0, y\right) + \min\left(0, \alpha \times \left(\exp\left(y\right) - 1\right)\right)\right).$$

For these autoencoders, we use $k < b$, where $k$ is the maximum number of clusters and the dimension of the encoding layer. Scalar $b$ is the size of the batch. Therefore, the matrix generated by the autoencoder at the layer $\mathbf{E}$ is $\mathbf{E} \in \mathbb{R}^{b \times k}$. Matrix $\mathbf{Q}$ is an orthogonal matrix such that the columns are orthogonal. Since $\mathbf{AR}^{-1} = \mathbf{Q}$, the inverse $\mathbf{R}^{-1}$ is saved within the QR Orthogonalization layer during each training step. The output of the QR Orthogonalization layer becomes $\mathbf{Z} = \mathbf{ER}^{-1}$. During evaluation, the last calculated $\mathbf{R}^{-1}$ is used to project the results of the encoding layer.

The loss function for OSAE is comprised of three parts. The first is the

reconstruction error on the input data. In addition, there is a separate orthogonality loss, and a spectral loss. Total loss is then given by

$$\text{Loss} = \text{Loss}_R + \text{Loss}_O + \text{Loss}_S.$$

Reconstruction loss $\text{Loss}_R$ is the mean squared error of the batch input $\mathbf{A}_b$ and the output of decoder portion of the autoencoder $\mathbf{A}'_b$. This is calculated by

$$\text{Loss}_R = \text{MSE}\left(\mathbf{A}_b, \mathbf{A}'_b\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{ij} - a'_{ij}\right)^2$$

where $\mathbf{A}_b = \mathbf{A}\left[\mathcal{I}, :\right]$ and $\mathcal{I}$ are the indices of the minibatch input.

$\text{Loss}_O$ is the orthogonalization loss and is defined as the mean absolute error of the identity matrix $\mathbf{I}^k$ and the product of the transpose of $\mathbf{E}$ with itself. By definition, if $\mathbf{E}$ is column orthogonal, then $\mathbf{E}^\top \times \mathbf{E} = \mathbf{I}^k$

$$\text{Loss}_o = \frac{1}{n^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\left|I^k - \hat{e_{ij}}\right|$$

where $\hat{\mathbf{E}} = \mathbf{E}^\top\mathbf{E}$.

The spectral loss is calculated by using the orthogonal output of the QR decomposition layer and an estimate of the Laplacian $\mathbf{L}$. In the spectral decomposition defined in earlier chapters, the normalized symmetric Laplacian $\mathbf{L}_{sym}$ was defined as

$$\mathbf{L}_{sym} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}.$$

Since the input of the graph is a sparse adjacency matrix, slicing the matrix such that it contains only the rows and columns of the indices $\mathcal{I}$ of minibatch $\mathbf{A}_b$ would

result in a mostly empty matrix. Instead, we use the Laplacian kernel

$$\mathcal{K}_L\left(\mathbf{A}_b, \mathbf{A}_b\right) = \exp\left(\frac{-\left\|a_{b,x} - a_{b,y}\right\|}{\sigma}\right)$$

where $\sigma$ is the kernel bandwidth for each pair of rows $x$ and $y$ in $\mathbf{A}_b$. In the experiments given later we use $\sigma = 2$. The Laplacian kernel $\mathcal{K}_L \in \mathbb{R}^{k \times k}$ yields a new affinity matrix $\mathbf{A}_b^L$ for minibatch $\mathbf{A}_b$. Using the pairwise Laplacian similarity, we calculate the normalized symmetric Laplacian of $\mathbf{A}_b^L$ as

$$\mathbf{L}_{sym} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_b^L\mathbf{D}^{-\frac{1}{2}}$$

where $\mathbf{D}$ is a diagonal matrix where $d_{ii} = \sum_{j=1}^{k} \hat{a}_{ij}$ where $\hat{a}_{ij}$ is element $i, j$ in $\mathbf{A}_b^L$. The next step calculates the $k$ eigenvectors corresponding to the smallest eigenvalues of $\mathbf{L}_{sym}$ for $\mathbf{A}_b^L$. Matrix $\mathbf{Z}'$ is constructed by the top-$k$ eigenvectors as given by the eigen decomposition $\text{eigs}\left(\mathbf{L}_{sym}, k\right)$. With this matrix, the spectral loss can be calculated as the mean absolute error of the approximate eigenvectors and the orthogonalized encoding.

$$\text{Loss}_S = \frac{1}{n}\sum_{i=1}^{b}\sum_{j=1}^{k}\left|\mathbf{Z}_{ij} - \mathbf{Z}'_{ij}\right|$$

Those three losses ($\text{Loss}_R$, $\text{Loss}_O$, and $\text{Loss}_S$) summed together make the full loss calculation for each batch during the training phase.

For every epoch during training, we sample $b$ rows from the adjacency matrix $\mathbf{A}$. For a single batch $b$, the rows of the adjacency matrix $\mathbf{A}$ are sample uniformly without replacement to create minibatch $\mathbf{A}_b$. Matrix $\mathbf{A}_b \in \mathbb{R}^{b \times n}$ will have $b$ rows from $\mathbf{A}$. This sample is the training data for a single epoch. At the next epoch, a new batch is sampled in the same manner from $\mathbf{A}$. With this sampling method, some rows may be selected more than once during the full training session. Other rows

may not be sampled at all. We repeat this procedure until reaching the maximum number of training epochs.

The final trained network is then used for performing clustering over the entire set of data. Using the orthogonal encoder portion of the network, the output of the QR decomposition layer ($\mathbf{Z}$) is used as the $n \times k$ embedding for the graph. Much like in spectral clustering, we normalize each row of $\mathbf{Z}$ as

$$\mathbf{z}_i^{\ell} = \frac{z_i}{\sqrt{\sum_{j=1}^{k} z_{ij}^2}}.$$

The normalized rows of $\mathbf{Z}$ are used as the data points for fuzzy c-means clustering. The resulting $\mathbf{U}$ from FCM are the fuzzy cluster assignments for the embedding. As before, it is possible to use the embedding for hierarchical spectral clustering as well. Since the spectral loss imposes an ordering on the columns of the output based on the ordering of the approximate spectral decomposition, we can similarly use the ordered columns of $\mathbf{Z}$ as the top-$k'$ for clusters $k' < k$. The construction of the hierarchy is the same as that detailed in prior chapters.

## 7.4 Vote Prediction using OSAE Graph Embedding

Using the same federal data from the prior chapters, we run several experiments testing the efficacy of using orthogonal spectral autoencoders for vote prediction. The general procedure for performing the clustering is as follows. A weighted graph adjacency matrix is used as input to the neural network. Each element $a_{ij}$ in $\mathbf{A}$ is scaled logarithmically by $a_{ij} = \log\left(\sum \text{amt}_{ij}\right)$ where $\text{amt}_{ij}$ is the value of a donation between nodes $i$ and $j$ in the graph.

For each different year of data, we keep the structure and procedure of training the autoencoder the same. The only exception to this is that the input and

output dimension must change for each cycle to fit the different data. No other hyperparameters are changed between runs. All runs were done with a mini-batch size of $b = 128$. Based on the results from the clustering on the previous results, we set $k = 8$ as the maximum number of clusters, and thus the dimension of the encoding layer. Each network is trained over 500 epochs.

In the following experiments we test the effectiveness of the graph embedding for predicting future behavior for each 2-year cycle separately. The classification method is the same as described in Chapter 6. In each cycle, a random forest is used to predict if a legislator will vote Yea or Nay based on their community assignments and the ideological estimate of the bill in question. For these experiments, we work with a smaller dataset than the full vote data used in Chapter 6. A sample of $200,000$ votes is drawn from the relevant set of Yea and Nay votes for that two-year cycle. The random forest is constructed with an ensemble of 50 trees. As we are more interested in the relative performance, the random forests are not optimized or pruned.

The first experiment listed in the results is a baseline for comparison. Multiple trials were performed using the spectral decomposition in Chapter 6. In each trial, the spectral decomposition with $k = 8$ clusters calculated for each of the networks in years $1980, 1982, \ldots, 2012$. For each of those years, we use three separate feature sets to predict behavior of the community members. The three separate feature sets are the row normalized top-$k$ eigenvectors of the normalized symmetric Laplacian $\mathbf{L}_{sym}$, the hierarchical fuzzy c-means community assignments for each $k = 2, \ldots, 8$. In addition to testing the performance of spectral decomposition on the entire adjacency matrix, we add a comparison to the vote prediction when using the CFScore ideological estimates. This is done to add more context to the relative performance of the algorithms. Each of these features are assigned to the individual nodes within the graph. These features are then attached to the voting history that pairs legislators

within the graph to bills. Each bill has its own ideological estimate from DW-NOMINATE that gives an estimate of the policy position of each bill.

The next experiments test the efficacy of the embeddings for an orthogonal autoencoder and OSAE. Each network is trained for 500 epochs, where every epoch samples a minibatch from the adjacency matrix and uses stochastic gradient descent to update the weights. The row normalized graph embedding $\mathbf{Z}$ is then used in the calculation of the features that are assigned to each legislator in the campaign finance network. As before, we analyze the results of different treatments of the graph embedding: embedding itself, fuzzy c-means clustering, and hierarchical fuzzy spectral clustering. For all the following experiments, trials were repeated 10 times to get the average performance of each experiment.

First we test OSAE on clustering of a dataset of points on a 2-dimensional plane. In this dataset, the 2-dimensional data is fed into the same orthogonal spectral autoencoder architecture as that used by the later experiments on graph embedding. The Laplacian of the entire data set is approximated using the $\mathcal{K}_{\text{rbfs}}$ described in Section 7.2. The network is trained over mini-batches of size 128 until the loss function fails to improve by $\epsilon = 0.001$ using a patience window of 100 epochs. Results of consecutive $K$-means clustering on the graph using the true number of clusters is shown in Figures 7.5 and 7.6. As seen in the graphs, the results do not perfectly capture the true clusters in the data, although the results still show promise in providing an embedding that can be useful for other networks.

We analyze the results of training the encoders for each cycle of the data. As baseline points of comparison, the following figures show results for predictive voting behavior of legislators within the networks for each cycle. Included in these results are the results for using the CFScore ideological measure to predict voting behavior instead of the spectral decomposition and clustering. Table 7.1 shows the results of

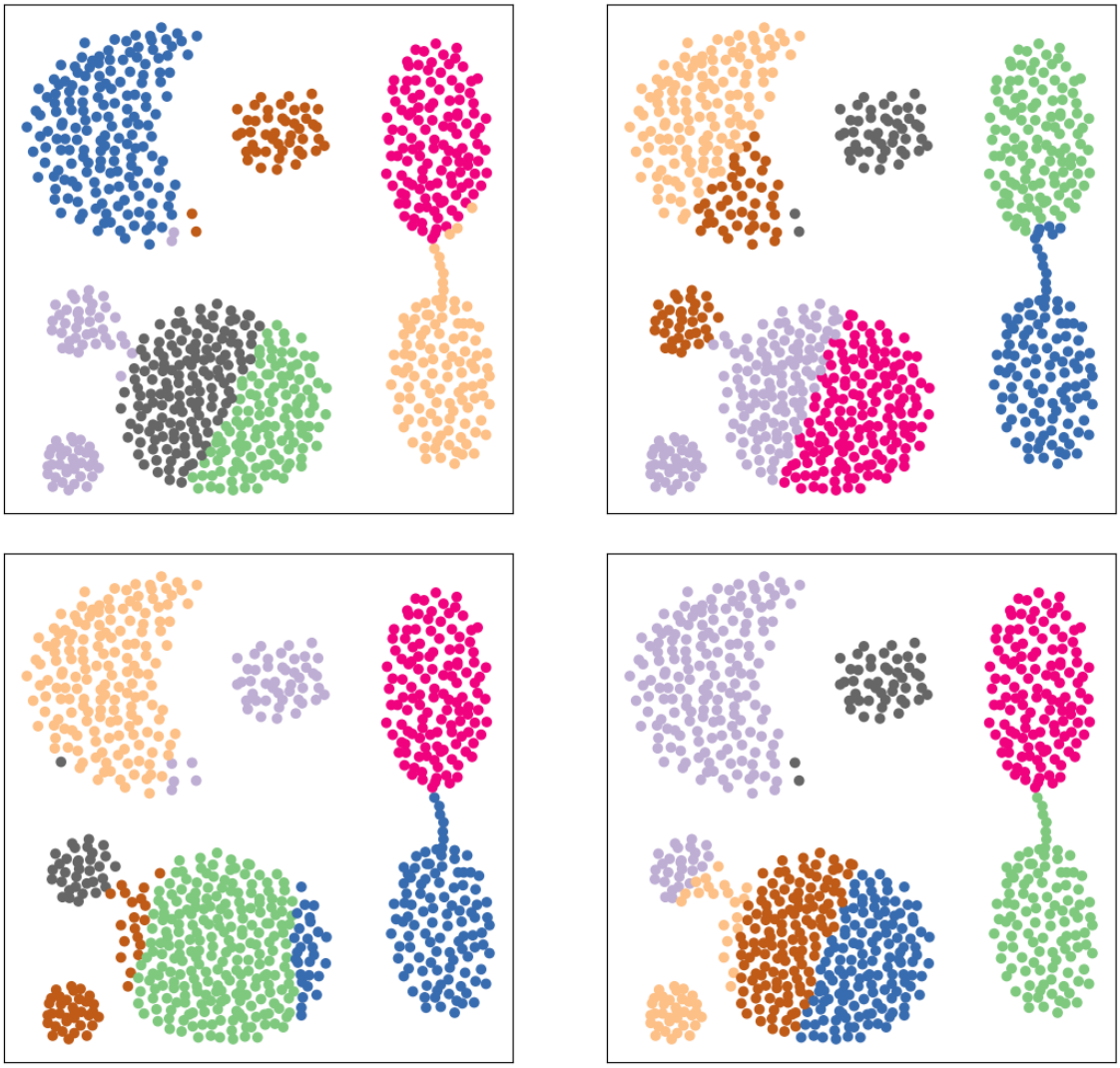


Figure 7.5: Example Runs of OSAE on 2-Dimensional Data

predicting voting behavior on the sampled dataset.

When generating the results given in the table, the sample of votes was the same across the three different classifiers for each individual trial within a cycle. In the table, "Eigs" refers to directly clustering on the row-normalized top-$k$ eigenvectors of the spectral decomposition. The number of clusters and eigenvectors is $k = 8$ for each of the following trials. "HFSC" is the result of predicting using the hierarchical

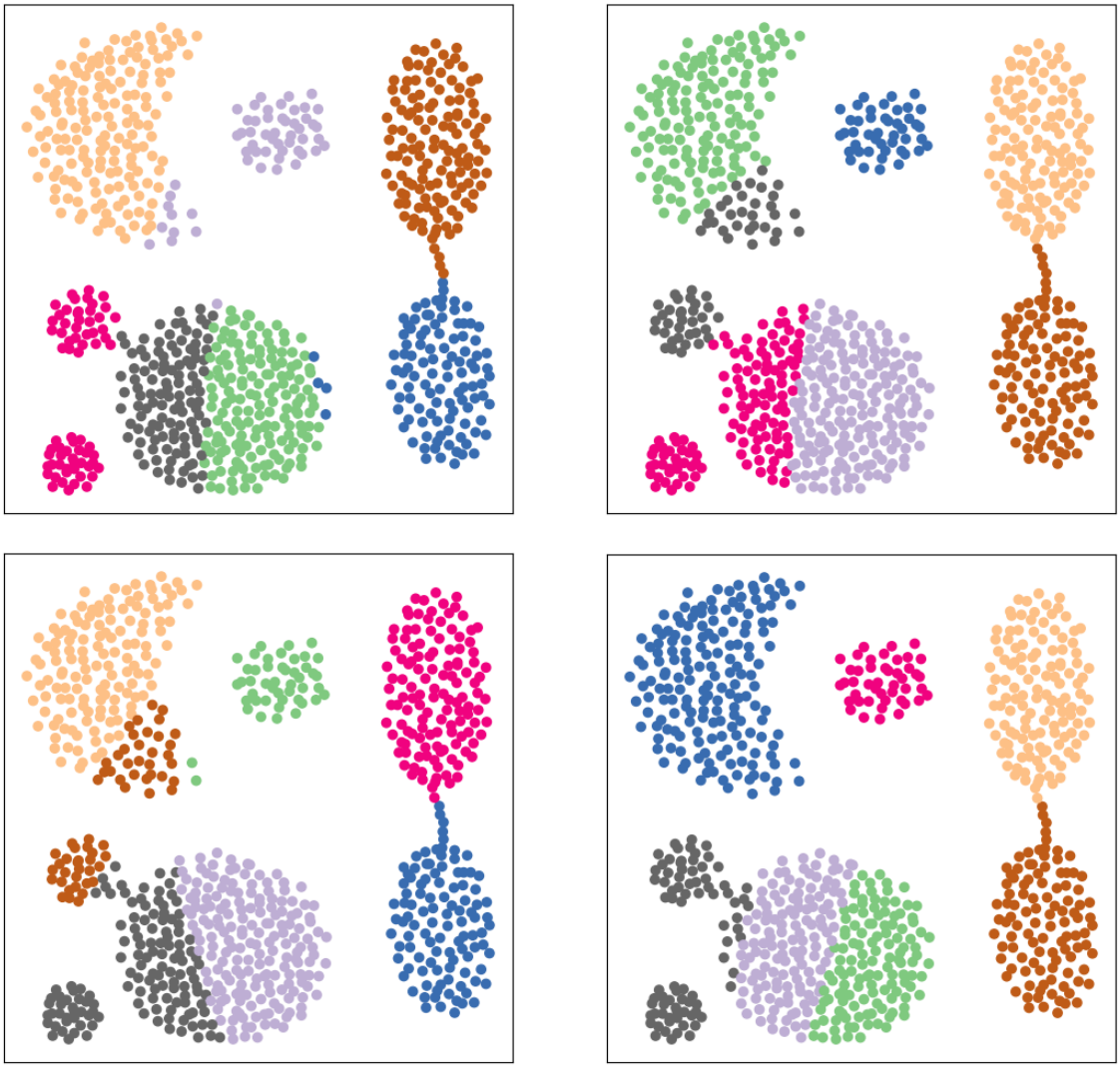


Figure 7.6: Additional Runs of OSAE on 2-Dimensional Data

community assignments after performing FCM on the eigenvectors. "CFScore" is using just the ideological metric as provided by Bonica [78].

The results from the table are as expected based on the larger experiments performed in Chapter 6. Clustering directly on the top-$k$ eigenvectors is effective in predicting voting behavior. Using the community hierarchy from HFSC did not perform quite as well as the raw top-$k$ eigenvectors, but the performance is still

Table 7.1: Prediction baselines on each cycle

| Cycle | Eigs | HFSC | CFScore |
|---|---|---|---|
| 1980 | $0.815 \pm 5.27E-04$ | $0.806 \pm 1.06E-03$ | $0.799 \pm 4.21E-04$ |
| 1982 | $0.823 \pm 3.96E-04$ | $0.816 \pm 4.68E-04$ | $0.806 \pm 2.63E-04$ |
| 1984 | $0.834 \pm 3.80E-04$ | $0.827 \pm 4.37E-04$ | $0.814 \pm 3.88E-04$ |
| 1986 | $0.824 \pm 2.64E-04$ | $0.818 \pm 5.83E-04$ | $0.808 \pm 3.54E-04$ |
| 1988 | $0.861 \pm 4.02E-04$ | $0.857 \pm 3.60E-04$ | $0.845 \pm 2.90E-04$ |
| 1990 | $0.820 \pm 3.19E-04$ | $0.803 \pm 8.33E-04$ | $0.826 \pm 4.68E-04$ |
| 1992 | $0.827 \pm 5.03E-04$ | $0.815 \pm 4.06E-04$ | $0.825 \pm 1.99E-04$ |
| 1994 | $0.835 \pm 5.19E-04$ | $0.800 \pm 1.46E-03$ | $0.842 \pm 4.07E-04$ |
| 1996 | $0.819 \pm 4.87E-04$ | $0.778 \pm 1.13E-02$ | $0.848 \pm 2.23E-04$ |
| 1998 | $0.838 \pm 6.33E-04$ | $0.826 \pm 6.34E-04$ | $0.854 \pm 3.11E-04$ |
| 2000 | $0.860 \pm 2.83E-04$ | $0.852 \pm 4.31E-04$ | $0.872 \pm 2.00E-04$ |
| 2002 | $0.892 \pm 4.52E-04$ | $0.881 \pm 6.50E-04$ | $0.901 \pm 3.06E-04$ |
| 2004 | $0.922 \pm 3.81E-04$ | $0.920 \pm 3.94E-04$ | $0.914 \pm 2.42E-04$ |
| 2006 | $0.896 \pm 5.39E-04$ | $0.888 \pm 4.95E-04$ | $0.894 \pm 2.32E-04$ |
| 2008 | $0.914 \pm 2.33E-04$ | $0.911 \pm 3.13E-04$ | $0.908 \pm 2.26E-04$ |
| 2010 | $0.922 \pm 2.23E-04$ | $0.912 \pm 7.93E-03$ | $0.928 \pm 1.95E-04$ |
| 2012 | $0.897 \pm 3.56E-04$ | $0.893 \pm 3.01E-04$ | $0.883 \pm 2.96E-04$ |

comparable. As shown, eigenvalue decomposition and hierarchical fuzzy spectral clustering outperforms the CFScore ideological metric during earlier cycles. This is likely due to CFScore being trained on the entire dataset of campaign contributions. The much higher proportion of data in later cycles would skew the CFScore. These results will serve as a point of comparison to the graph embedding generated from

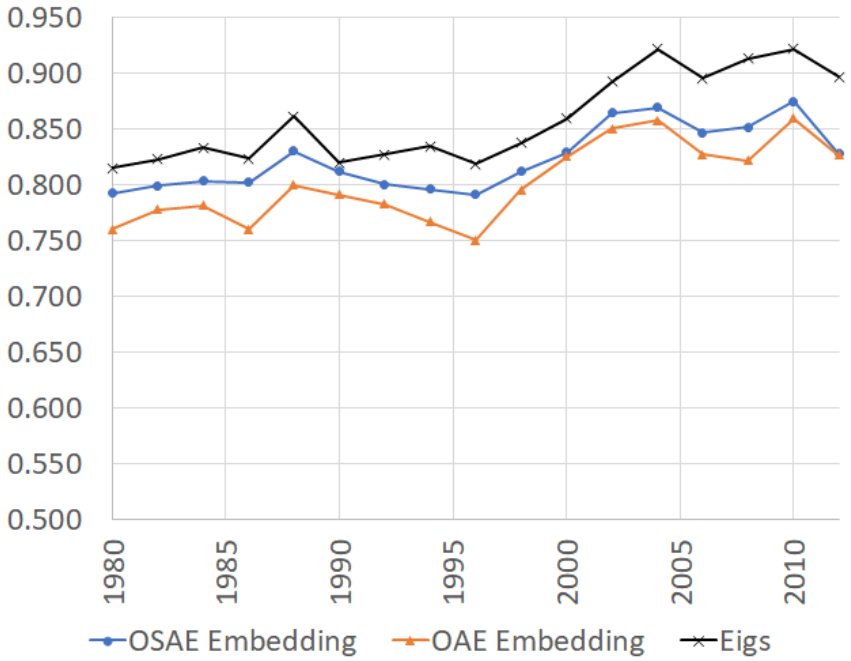



Figure 7.7: Vote Prediction using Spectral Embedding

the orthogonal spectral autoencoder.

The results of vote prediction on the spectral embedding of $\mathbf{Z}$ is shown in Figure 7.7. As can be seen, in general OSAE outperforms OAE in predicting behaviors among the legislators. The average performance per cycle of using the eigenvectors directly in prediction are also shown as a baseline.

Performing hierarchical fuzzy spectral clustering on the graph embeddings gives the results shown in Figure 7.8. The out-of-bag accuracy for this classifier shows it is not quite as effective as using the spectral embeddings themselves, although the results are still very close. This mirrors the results from Chapter 6 where HFSC was similar in effectiveness to using eigenvectors directly.

Figure 7.9 shows the results of the vote classification when using fuzzy c-means to cluster the spectral embedding at $k = 8$. As before, the embeddings of OSAE outperform those of OAE, although not always to a significant degree. In 2010 where

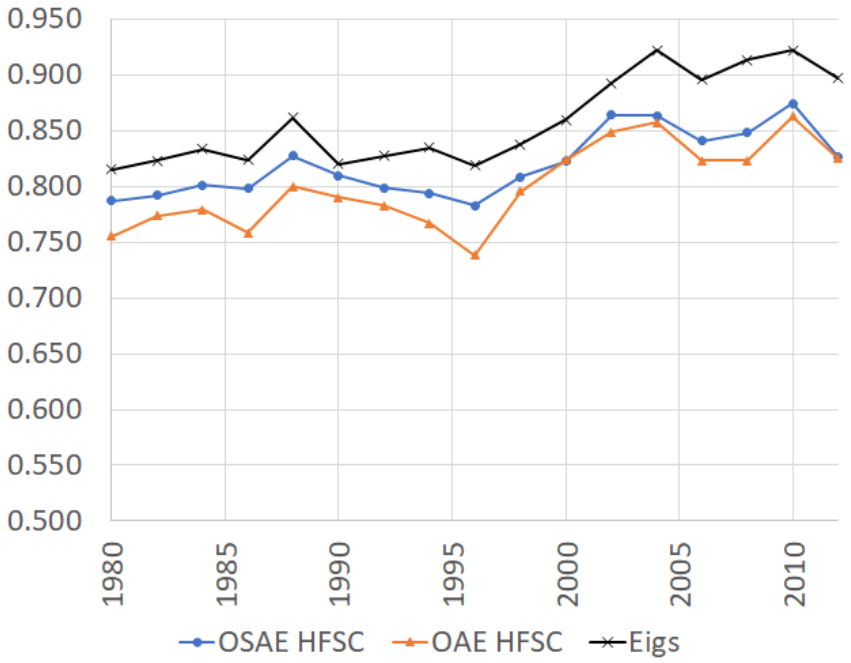



Figure 7.8: Vote Prediction using HFSC of Spectral Embedding

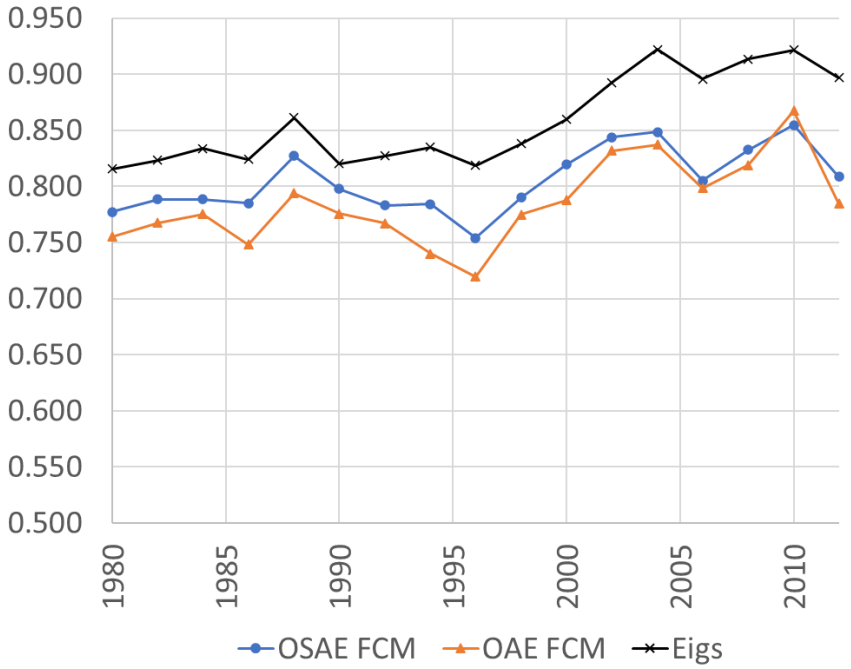



Figure 7.9: Vote Prediction using Fuzzy C-Means of Spectral Embedding

Table 7.2: Accuracy of vote prediction by cycle using spectral embedding

| Cycle | $OSAE_{Emb}$ | $OAE_{Emb}$ | $OSAE_{HFSC}$ | $OAE_{HFSC}$ | $OSAE_{FCM}$ | $OAE_{FCM}$ |
|---|---|---|---|---|---|---|
| 1980 | 0.792 | 0.760 | 0.787 | 0.755 | 0.777 | 0.755 |
| 1982 | 0.799 | 0.778 | 0.792 | 0.774 | 0.789 | 0.768 |
| 1984 | 0.804 | 0.782 | 0.801 | 0.779 | 0.789 | 0.775 |
| 1986 | 0.802 | 0.760 | 0.798 | 0.758 | 0.785 | 0.748 |
| 1988 | 0.830 | 0.800 | 0.827 | 0.800 | 0.827 | 0.794 |
| 1990 | 0.812 | 0.791 | 0.810 | 0.790 | 0.798 | 0.776 |
| 1992 | 0.801 | 0.783 | 0.799 | 0.783 | 0.783 | 0.767 |
| 1994 | 0.796 | 0.767 | 0.794 | 0.767 | 0.784 | 0.740 |
| 1996 | 0.791 | 0.750 | 0.783 | 0.738 | 0.754 | 0.720 |
| 1998 | 0.812 | 0.796 | 0.808 | 0.795 | 0.790 | 0.775 |
| 2000 | 0.830 | 0.825 | 0.823 | 0.824 | 0.819 | 0.788 |
| 2002 | 0.865 | 0.850 | 0.864 | 0.849 | 0.844 | 0.832 |
| 2004 | 0.869 | 0.858 | 0.863 | 0.858 | 0.848 | 0.837 |
| 2006 | 0.847 | 0.827 | 0.841 | 0.823 | 0.805 | 0.799 |
| 2008 | 0.852 | 0.822 | 0.848 | 0.823 | 0.833 | 0.819 |
| 2010 | 0.875 | 0.860 | 0.874 | 0.863 | 0.855 | 0.868 |
| 2012 | 0.828 | 0.827 | 0.826 | 0.825 | 0.809 | 0.785 |

OAE had better out-of-bag accuracy in the random forests, it was not statistically significantly better than OSAE. Still, this shows that the orthogonalization in OAE is useful in finding a network embedding for predicting behavior of the communities.

The full side-by-side results for each of the uses of the embeddings are shown in Table 7.2. These results help highlight once again that predictions using HFSC are quite close to the performance of the spectral embeddings themselves and can

significantly outperform just using FCM clustering.

## 7.5 Conclusion

In this chapter we defined the Orthogonal Spectral Autoencoder. This neural network allows embedding a graph into a lower dimensional space approximating the spectral decomposition. Results show the orthogonal spectral loss function is effective at obtaining a graph embedding that is useful in predicting some behavior of the individuals in the graph. The results were comparable to that of the full spectral decomposition. Based on the preliminary results, this avenue appears promising for refinement. There are many areas where improvements may be possible, either in modifying the structure of the autoencoder, or additional hyperparameter searching. Some of these will be explored in future work.

# CHAPTER EIGHT

# CONCLUSION

In the previous chapters we provided details on novel architectures for finding hierarchical and overlapping clusters in social networks. These architectures were validated against multiple campaign finance social networks that exhibited important features of the community assignment embeddings. The following sections summarizes these results and provides avenues for future work.

## 8.1 Summary

We introduced an approach for hierarchical fuzzy spectral clustering. This approach iteratively added eigenvectors and used fuzzy c-means to create a hierarchy of fuzzy community assignments. The hierarchical portion of the algorithm created the hierarchy by attaching child nodes to their parents based on a fuzzy Jaccard similarity. The effectiveness of this algorithm was shown on two small benchmark datasets in addition to a real-world campaign finance network. These results showed HFSC found communities that corresponded to known ground truths. Additionally, we showed that the discovered communities were crucial in identifying varying behaviors of the individuals within social networks. This was true for the hierarchical communities, where children shared some behaviors with their parent, but had different behavior between the siblings. It was also true for vertices in the overlap between communities as it was shown they behaved differently than those vertices who were entirely within a single community. We used the hierarchical communities on a campaign finance network to isolate $k = 2$ communities that corresponded to party alignment. The overlap of these two communities showed that the fuzzy communities

provide additional insight into the behaviors of the individuals since the donors in the overlap historically donated to both parties in equal amounts. The children of the $k = 2$ show how the communities at $k = 4$ differ from their parents in that the child communities split into sibling communities that heavily favored party for one and favored winning candidates of the party for the other sibling. These results show the effectiveness of HFSC in creating community assignments that highlight behaviors of the underlying networks.

We subsequently used fuzzy similarity metrics to connect communities found in snapshots of data through time. After the initial community discovery using HFSC on individual time steps, the links between communities in adjacent time steps were added based on that similarity metric. We analyzed the ability of this algorithm to track communities through time on multiple state campaign finance networks, detailing how individuals and communities change behavior over time. The results showed that the campaign finance communities persisted through time in the campaign finance networks despite the relatively high churn of vertices in the graph at each time step. These communities were validated against ideological scores as well as their behavior detailed through time.

During those experiments, the issue of interpretability of the communities was apparent. Enumerating the ways in which behavior was shared among communities, as well as in how they differed was time consuming. We utilized association rule mining on transaction data partitioned by the hierarchical fuzzy spectral clustering to improve the interpretability of the community assignments. The results on rules found within a state campaign finance network showed the automatic rule finding was beneficial in interpreting the community structure.

In the following chapter we showed the generalizability of the community detection by using the community assignments in a prediction task. This task required

predicting behavior of individual within a social network. This behavior was one that was not directly described in the underlying data that generated the network. This predictive task took the form of predicting voting of legislators in the United States legislature. We combined the community assignments with a data set bill ideological estimates and the history of Yea and Nay votes over a period of 12 different snapshots. The results showed the hierarchical fuzzy community assignments were generalizable and effective at predicting votes.

In the final chapter we introduced a novel graph embedding structure, Orthogonal Spectral Autoencoder, in order to resolve two issues inherent to the spectral decomposition step necessary for HFSC. This allowed for embedding the graph without performing spectral decomposition over the entire adjacency matrix. This neural network architecture naturally allowed for projecting new data points into the spectral domain without having to recalculate the Laplacian matrix or spectral decomposition. This allows for out-of-sample clustering of the new data. While the approximate graph embedding did not perform quite as well as the full spectral decomposition, the results were comparable.

## 8.2 Future Work

There are many ways in which improvements to hierarchical fuzzy spectral clustering could be made. The algorithms connecting communities through time could be improved by utilizing information in adjacent time steps to smooth the changes in the community assignments. In addition to the smoothing effect, such improvements could more easily match communities through time, as well as better identify areas where communities merge, split, or otherwise change.

Additional interpretability improvements could be made by modifying the item set detection. Instead of complete partitions, a weighted frequent itemset algorithm

may be used to find the important rules in the data. Developing better pruning to reduce the extraneous rules would also aid in rapid interpretation of the communities in social networks.

The introduction of the Orthogonal Spectral Autoencoder in particular opens many avenues for new research. Additional research in hyperparameter and structure tuning may yield a better approximation without further modifications to the underlying algorithm. Preliminary work has already begun on replacing the spectral loss function with an optimization function that approximates the eigenvectors without performing the direct decomposition. Additions to the neural network architecture, such as graph convolution networks could tag the communities via embedding the node and edge heterogeneous information. These graph convolution networks could also be used in a recurrent network to capture the dynamic properties of the social networks.

# REFERENCES

[1] B. Zhang and S. Horvath, "A general framework for weighted gene Co-Expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. Article 17+, 2005.

[2] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, and S. E. Petersen, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[3] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, Jul 2001.

[4] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities," *IEEE Computer*, vol. 35, pp. 66–71, 2002.

[5] D. Lusseau and M. E. Newman, "Identifying the role that animals play in their social networks," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 271, pp. S477–S481, 2004.

[6] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review*, vol. E 69, no. 026113, p. 026113, 2004.

[7] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients," *SIGCOMM Comput. Commun. Rev.*, vol. 30, p. 97–110, Aug. 2000.

[8] P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao, "A graph based approach to extract a neighborhood customer community for collaborative filtering," in *Proceedings of the Second International Workshop on Databases in Networked Information Systems*, (Berlin, Heidelberg), p. 188–200, Springer-Verlag, 2002.

[9] R. Guimerà and L. A. Nunes Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, Feb 2005.

[10] P. Csermely, "Strong links are important, but weak links stabilize them," *Trends in Biochemical Sciences*, vol. 29, no. 7, pp. 331–334, 2004.

[11] S. Fortunato, "Community detection in graphs," *CoRR*, vol. abs/0906.0612, 2009.

[12] V. Blondel, J. Guillaume, R. Lambiotte, and E. Mech, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, p. P10008, 2008.

[13] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[14] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences - ISCIS*, vol. 10, pp. 284–293, 2004.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856, MIT Press, 2001.

[16] A. Pothen, H. Simon, and K. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM Journal on Matrix Analysis and Applications*, vol. 11, no. 3, pp. 430–452, 1990.

[17] S. Bandyopadhyay, "Automatic determination of the number of fuzzy clusters using simulated annealing with variable representation," in *Foundations of Intelligent Systems* (M.-S. Hacid, N. Murray, Z. Ras, and S. Tsumoto, eds.), vol. 3488 of *Lecture Notes in Computer Science*, pp. 594–602, Springer Berlin Heidelberg, 2005.

[18] A. Devillez, P. Billaudel, and G. V. Lecolier, "A fuzzy hybrid hierarchical clustering method with a new criterion able to find the optimal partition," *Fuzzy Sets and Systems*, vol. 128, no. 3, pp. 323 – 338, 2002.

[19] J. Liu, "Fuzzy modularity and fuzzy community structure in networks," *The European Physical Journal B*, vol. 77, no. 4, pp. 547–557, 2010.

[20] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[21] V. Torra, "Fuzzy c-means for fuzzy hierarchical clustering," in *Proceedings of the 14th IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 646–651, May 2005.

[22] J. Xie, B. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *Proceedings of the 11th International Conference on Data Mining Workshops (ICDMW)*, pp. 344–349, Dec 2011.

[23] S. Wahl and J. Sheppard, "Hierarchical fuzzy spectral clustering in social networks using spectral characterization," in *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 305–310, 2015.

[24] S. Wahl and J. Sheppard, "Fuzzy spectral hierarchical communities in evolving political contribution networks," in *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS*, pp. 371–376, AAAI Press, 2017.

[25] S. Wahl and J. Sheppard, "Association rule mining in fuzzy political donor communities," in *Proceedings of the 14th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 231–245, 2018.

[26] S. Wahl, J. Sheppard, and E. Shanahan, "Legislative vote prediction using campaign donations and fuzzy hierarchical communities," in *18th IEEE International Conference On Machine Learning And Applications, ICMLA*, pp. 718–725, IEEE, 2019.

[27] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[28] G. K. Orman, V. Labatut, and H. Cherifi, "Qualitative comparison of community detection algorithms," in *Digital Information and Communication Technology and Its Applications*, pp. 265–279, 2011.

[29] M. Porter, J. Onnela, and P. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[30] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[31] A. Pothen, H. Simon, and K. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM Journal on Matrix Analysis and Applications*, vol. 11, no. 3, pp. 430–452, 1990.

[32] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 483–490, 2007.

[33] R. M. Karp, *Reducibility among Combinatorial Problems*, pp. 85–103. Boston, MA: Springer US, 1972.

[34] D. Zuckerman, "Linear degree extractors and the inapproximability of max clique and chromatic number," in *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '06, (New York, NY, USA), p. 681–690, Association for Computing Machinery, 2006.

[35] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5249–5253, April 2004.

[36] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "Monic: Modeling and monitoring cluster transitions," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 706–711, Association for Computing Machinery, 2006.

[37] T. Aynaud and J.-L. Guillaume, "Static community detection algorithms for evolving networks," in *WiOpt'10: Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, (Avignon, France), pp. 508–514, 2010.

[38] S. Jung and A. Segev, "Analyzing future communities in growing citation networks," *Knowledge Based Systems*, vol. 69, pp. 34–44, 2014.

[39] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, "Maximizing modularity is hard," 2006. cite arxiv:physics/0608255 Comment: 10 pages, 1 figure.

[40] S. Sarkar, J. A. Henderson, and P. A. Robinson, "Spectral characterization of hierarchical network modularity and limits of modularity detection," *PLoS ONE*, vol. 8, p. e54383, 01 2013.

[41] H.-W. Shen, X.-Q. Cheng, and J.-F. Guo, "Quantifying and identifying the overlapping community structure in networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 07, p. P07042, 2009.

[42] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, p. 7821–7826, Jun 2002.

[43] A. Clauset, M. E. J. Newman, , and C. Moore, "Finding community structure in very large networks," *Physical Review E*, pp. 1– 6, 2004.

[44] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[45] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Applied Intelligence*, vol. 48, no. 12, pp. 4743–4759, 2018.

[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[47] H. Sahbi, "A particular Gaussian mixture model for clustering and its application to image retrieval," *Soft Computing*, vol. 12, no. 7, pp. 667–676, 2007.

[48] R. Lehoucq and D. Sorensen, "Deflation techniques for an implicitly restarted Arnoldi iteration," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 4, pp. 789–821, 1996.

[49] J. L. Kalla and D. E. Broockman, "Campaign contributions facilitate access to congressional officials: A randomized field experiment," *American Journal of Political Science*, vol. 60, no. 3, pp. 545–558, 2016.

[50] I. Pastine and T. Pastine, "Politician preferences, law-abiding lobbyists and caps on political contributions," *Public Choice*, vol. 145, pp. 81–101, Oct 2010.

[51] J. Fox and L. Rothenberg, "Influence without bribes: A noncontracting model of campaign giving and policymaking," *Political Analysis*, vol. 19, no. 3, pp. 325–341, 2011.

[52] P. Akey, "Valuing changes in political networks: Evidence from campaign contributions to close congressional elections," *Review of Financial Studies*, vol. 28, no. 11, pp. 3188–3223, 2015.

[53] J. G. Gimpel, F. E. Lee, and J. Kaminski, "The political geography of campaign contributions in American politics," *Journal of Politics*, vol. 68, no. 3, pp. 626–639, 2006.

[54] J. H. Kirkland, "Ideological heterogeneity and legislative polarization in the united states," *Political Research Quarterly*, vol. 67, no. 3, pp. 533–546, 2014.

[55] P. Flavin, "Campaign finance laws, policy outcomes, and political equality in the American states," *Political Research Quarterly*, vol. 68, no. 1, pp. 77–88, 2015.

[56] J. Benz, J. H. Kirkland, V. Gray, D. Lowery, J. Sykes, and M. Deason, "Mediated density: The indirect relationship between us state public policy and pacs," *State Politics & Policy Quarterly*, vol. 11, no. 4, pp. 440–459, 2011.

[57] T. Rubenzer, "Campaign contributions and US foreign policy outcomes: An analysis of Cuban American interests," *American Journal of Political Science*, vol. 55, no. 1, pp. 105–116, 2011.

[58] E. N. Powell and J. Grimmer, "Money in exile: Campaign contributions and committee access," *The Journal of Politics*, vol. 78, no. 4, pp. 974–988, 2016.

[59] J. G. Gimpel, F. E. Lee, and M. Parrott, "Business interests and the party coalitions," *American Politics Research*, vol. 42, no. 6, pp. 1034–1076, 2014.

[60] J. Gilbert and R. Oladi, "Net campaign contributions, agricultural interests, and votes on liberalizing trade with china," *Public Choice*, vol. 150, no. 3/4, pp. 745–769, 2012.

[61] J. Huber and M. Kirchler, "Corporate campaign contributions and abnormal stock returns after presidential elections," *Public Choice*, vol. 156, pp. 285–307, Jul 2013.

[62] A. R. Brown, "Does money buy votes? The case of self-financed gubernatorial candidates, 1998–2008," *Political Behavior*, vol. 35, pp. 21–41, Mar 2013.

[63] K. Benoit and M. Marsh, "Incumbent and challenger campaign spending effects in proportional electoral systems," *Political Research Quarterly*, vol. 63, no. 1, pp. 159–173, 2010.

[64] T. M. Holbrook and A. C. Weinschenk, "Campaigns, mobilization, and turnout in mayoral elections," *Political Research Quarterly*, vol. 67, no. 1, pp. 42–55, 2014.

[65] S. J. Basinger, D. M. Cann, and M. J. Ensley, "Voter response to congressional campaigns: new techniques for analyzing aggregate electoral behavior," *Public Choice*, vol. 150, pp. 771–792, Mar 2012.

[66] M. J. Streb and B. Frederick, "When money cannot encourage participation: Campaign spending and rolloff in low visibility judicial elections," *Political Behavior*, vol. 33, pp. 665–684, Dec 2011.

[67] D. P. Christenson, C. D. Smidt, and C. Panagopoulos, "Deus ex machina: Candidate web presence and the presidential nomination campaign," *Political Research Quarterly*, vol. 67, no. 1, pp. 108–122, 2014.

[68] M. H. Crespin and J. L. Deitz, "If you can't join 'em, beat 'em: The gender gap in individual donations to congressional candidates," *Political Research Quarterly*, vol. 63, no. 3, pp. 581–593, 2010.

[69] S. Ansolabehere, J. M. Snyder, and C. Stewart, "Candidate positioning in U.S. house elections," *American Journal of Political Science*, vol. 45, no. 1, pp. 136–159, 2001.

[70] J. H. Aldrich, R. K. Gibson, M. Cantijoch, and T. Konitzer, "Getting out the vote in the social media era: Are digital tools changing the extent, nature and impact of party contacting in elections?," *Party Politics*, vol. 22, no. 2, pp. 165–178, 2016.

[71] R. La Due Lake and R. Huckfeldt, "Social capital, social networks, and political participation," *Political Psychology*, vol. 19, no. 3, pp. 567–584, 1998.

[72] E. Quintelier, D. Stolle, and A. Harell, "Politics in peer groups: Exploring the causal relationship between network diversity and political participation," *Political Research Quarterly*, vol. 65, no. 4, pp. 868–881, 2012.

[73] M. S. Mizruchi, "Similarity of political behavior among large American corporations," *American Journal of Sociology*, vol. 95, no. 2, pp. 401–424, 1989.

[74] G. Vonnahme, "A preferential attachment model of campaign contributions in state legislative elections," *Public Choice*, vol. 159, pp. 235–249, Apr 2014.

[75] R. Carroll, J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal, "Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap," *Political Analysis*, vol. 17, no. 3, p. 261–275, 2009.

[76] K. Poole, *Spatial Models of Parliamentary Voting*. Cambridge University Press, 01 2005.

[77] D. L. McFadden, *Quantal Choice Analysis: A Survey*, vol. 5, pp. 363–390. National Bureau of Economic Research, 1976.

[78] A. Bonica, "Mapping the ideological marketplace," *American Journal of Political Science*, vol. 58, no. 2, pp. 367–386, 2014.

[79] J. C. Bezdek, "Cluster validity with fuzzy sets," *Journal of Cybernetics*, vol. 3, no. 3, pp. 58–73, 1973.

[80] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[81] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[82] Y. Lin and S. Chen, "A centroid auto-fused hierarchical fuzzy c-means clustering," *IEEE Transactions on Fuzzy Systems*, vol. PP, pp. 1–1, 04 2020.

[83] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó, "Fuzzy communities and the concept of bridgeness in complex networks," *Phys. Rev. E*, vol. 77, p. 016107, Jan 2008.

[84] T. Havens, J. Bezdek, C. Leckie, K. Ramamohanarao, and M. Palaniswami, "A soft modularity function for detecting fuzzy communities in social networks," *IEEE Transactions on Fuzzy Systems*, vol. 21, pp. 1170–1175, Dec 2013.

[85] M. Bolla, "Spectra and structure of weighted graphs," *Electronic Notes in Discrete Mathematics*, vol. 38, pp. 149–154, 2011. The Sixth European Conference on Combinatorics, Graph Theory and Applications, EuroComb 2011.

[86] R. R. Nadakuditi and M. E. J. Newman, "Graph spectra and the detectability of community structure in networks," *Physical Review Letters*, vol. 108, May 2012.

[87] S. Chauhan, M. Girvan, and E. Ott, "Spectral properties of networks with community structure," *Phys. Rev. E*, vol. 80, p. 056114, Nov 2009.

[88] S. Sarkar and A. Dong, "Community detection in graphs using singular value decomposition," *Phys. Rev. E*, vol. 83, p. 046114, Apr 2011.

[89] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek, "Spectra of 'real-world' graphs: Beyond the semicircle law," *Phys. Rev. E*, vol. 64, p. 026704, Jul 2001.

[90] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[91] W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[92] G. Rossetti and R. Cazabet, "Community discovery in dynamic networks: a survey," *CoRR*, vol. abs/1707.03186, 2017.

[93] W. Wang, P. Jiao, D. He, D. Jin, L. Pan, and B. Gabrys, "Autonomous overlapping community detection in temporal networks: A dynamic Bayesian nonnegative matrix factorization approach," *Knowledge-Based Systems*, vol. 110, pp. 121–134, 2016.

[94] A. Bonica, "Database on ideology, money in politics, and elections: Public version 1.0 [accessed 2013-6-27]," 2013.

[95] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[96] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pp. 80–86, AAAI Press, 1998.

[97] J. B. Lewis, K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet, *Voteview: Congressional Roll-Call Votes Database [accessed 2018-04]*. Voteview, 2019. `https://voteview.com/`.

[98] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[99] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, Mar 1986.

[100] L. Raileanu and K. Stoffel, "Theoretical comparison between the Gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, pp. 77–93, 05 2004.

[101] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[102] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1 of *ICDAR '95*, (Washington, DC, USA), pp. 278–282, IEEE Computer Society, 1995.

[103] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[104] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, and U. Montreal, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, p. 153–160, 01 2007.

[105] M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *Aiche Journal*, vol. 37, pp. 233–243, 1991.

[106] F. Ratle, J. Weston, and M. L. Miller, "Large-scale clustering through functional embedding," in *Machine Learning and Knowledge Discovery in Databases* (W. Daelemans, B. Goethals, and K. Morik, eds.), (Berlin, Heidelberg), pp. 266–281, Springer Berlin Heidelberg, 2008.

[107] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[108] A. Choromanska, T. Jebara, H. Kim, M. Mohan, and C. Monteleoni, "Fast spectral clustering via the Nyström method," in *Algorithmic Learning Theory* (S. Jain, R. Munos, F. Stephan, and T. Zeugmann, eds.), (Berlin, Heidelberg), pp. 367–381, Springer Berlin Heidelberg, 2013.

[109] F. Pourkamali-Anaraki, "Scalable spectral clustering with Nyström approximation: Practical and theoretical aspects," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 242–256, 2020.

[110] Y. Han and M. Filippone, "Mini-batch spectral clustering," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3888–3895, 2017.

[111] S. Bonnabel, "Stochastic gradient descent on Riemannian manifolds," *IEEE Transactions on Automatic Control*, vol. 58, p. 2217–2229, Sep 2013.

[112] A. Jansen, G. Sell, and V. Lyzinski, "Scalable out-of-sample extension of graph embeddings using deep neural networks," *Pattern Recognition Letters*, vol. 94, pp. 1–6, 2017.

[113] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, vol. 48 of *ICML'16*, p. 478–487, JMLR, 2016.

[114] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu, "Deep spectral clustering using dual autoencoder network," in *IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4061–4070, 2019.

[115] M. T. Law, R. Urtasun, and R. S. Zemel, "Deep spectral clustering learning," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1985–1994, PMLR, 2017.

[116] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.

[117] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, (Cambridge, MA, USA), p. 177–184, MIT Press, 2003.

[118] K. Levin, F. Roosta, M. Mahoney, and C. Priebe, "Out-of-sample extension of graph adjacency spectral embedding," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 2975–2984, PMLR, 10–15 Jul 2018.

[119] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted Kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, 2010.

[120] K. Bhissy, F. Faleet, and W. Ashour, "Spectral clustering using optimized gaussian kernel function," *International Journal of Artificial Intelligence and Application for Smart Devices*, vol. 2, pp. 41–56, 05 2014.

[121] X. Zhang, J. Li, and H. Yu, "Local density adaptive similarity measurement for spectral clustering," *Pattern Recognition Letters*, vol. 32, no. 2, pp. 352–358, 2011.

[122] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, p. 1601–1608, MIT Press, 2004.

# APPENDICES

# APPENDIX A

## DONOR PATTERNS FOR NON-INDIVIDUALS IN STATES

The tables below show the results of hierarchical fuzzy spectral clustering on data provided by the National Institute on Money in Politics for each individual state in 2012. Each network is comprised of non-individuals who donated to more than one candidate. The largest connected component was retained, and hierarchical fuzzy spectral clustering applied to the generated networks. While Connecticut is included in these results, the resulting graph is small and sparse due to the restrictions imposed when creating the graph. Each table contains summary information regarding the donation history for donors within the found clusters. This summary includes all historical data those donors across all states in order to illustrate possible differing donation strategies among donors. For these examples, a threshold $\lambda = 0.3$ on the membership values is used to assign a community.

Table A.1: Donor History by Community in AK

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $6,231,206 | 0.576 | I | $13,356,519 | 3.522 | Lost | $2,671,707 | 0.202 |
| $c_{2,0}$ | R | $10,821,130 | 1.737 | N | $3,792,117 | 0.284 | Won | $13,207,194 | 4.943 |
| $c_{2,1}$ | D | $27,187,814 | 9.163 | I | $19,441,452 | 1.753 | Lost | $9,311,200 | 0.515 |
| $c_{2,1}$ | R | $2,967,028 | 0.109 | N | $11,088,941 | 0.570 | Won | $18,083,491 | 1.942 |
| $c_{2,0} \cap c_{2,1}$ | D | $316,847 | 1.420 | I | $395,911 | 2.615 | Lost | $114,579 | 0.269 |
| $c_{2,0} \cap c_{2,1}$ | R | $223,054 | 0.704 | N | $151,415 | 0.382 | Won | $425,896 | 3.717 |
| $c_{4,0}$ | D | $25,613,250 | 8.561 | I | $19,065,511 | 1.929 | Lost | $8,197,997 | 0.466 |
| $c_{4,0}$ | R | $2,991,708 | 0.117 | N | $9,884,314 | 0.518 | Won | $17,574,767 | 2.144 |
| $c_{4,1}$ | D | $2,971,304 | 0.450 | I | $7,298,018 | 3.154 | Lost | $1,706,765 | 0.243 |
| $c_{4,1}$ | R | $6,596,901 | 2.220 | N | $2,313,532 | 0.317 | Won | $7,012,928 | 4.109 |
| $c_{4,2}$ | D | $1,378,731 | 23.496 | I | $297,804 | 0.256 | Lost | $1,102,884 | 3.081 |
| $c_{4,2}$ | R | $58,679 | 0.043 | N | $1,164,765 | 3.911 | Won | $357,994 | 0.325 |
| $c_{4,3}$ | D | $3,459,270 | 0.814 | I | $6,283,352 | 4.247 | Lost | $949,159 | 0.149 |
| $c_{4,3}$ | R | $4,250,760 | 1.229 | N | $1,479,631 | 0.235 | Won | $6,376,948 | 6.719 |

Table A.2: Donor History by Community in AL

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $34,485,704 | 0.541 | I | $80,323,659 | 4.398 | Lost | $17,586,843 | 0.255 |
| $c_{2,0}$ | R | $63,706,690 | 1.847 | N | $18,264,635 | 0.227 | Won | $69,069,634 | 3.927 |
| $c_{2,1}$ | D | $33,447,824 | 4.189 | I | $26,682,042 | 1.761 | Lost | $14,520,483 | 0.657 |
| $c_{2,1}$ | R | $7,985,424 | 0.239 | N | $15,148,532 | 0.568 | Won | $22,108,843 | 1.523 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,347,249 | 0.425 | I | $3,530,416 | 3.525 | Lost | $1,114,123 | 0.392 |
| $c_{2,0} \cap c_{2,1}$ | R | $3,167,732 | 2.351 | N | $1,001,565 | 0.284 | Won | $2,838,704 | 2.548 |
| $c_{4,0}$ | D | $31,249,475 | 13.929 | I | $20,471,555 | 1.535 | Lost | $12,266,109 | 0.703 |
| $c_{4,0}$ | R | $2,243,447 | 0.072 | N | $13,340,692 | 0.652 | Won | $17,450,943 | 1.423 |
| $c_{4,1}$ | D | $33,543,754 | 0.564 | I | $76,283,859 | 4.445 | Lost | $15,820,918 | 0.238 |
| $c_{4,1}$ | R | $59,504,215 | 1.774 | N | $17,160,010 | 0.225 | Won | $66,372,234 | 4.195 |
| $c_{4,2}$ | D | $942,350 | 0.339 | I | $2,889,820 | 3.226 | Lost | $1,269,750 | 0.620 |
| $c_{4,2}$ | R | $2,782,245 | 2.952 | N | $895,775 | 0.310 | Won | $2,047,195 | 1.612 |
| $c_{4,3}$ | D | $2,080,379 | 0.269 | I | $7,303,229 | 2.907 | Lost | $2,889,425 | 0.497 |
| $c_{4,3}$ | R | $7,730,057 | 3.716 | N | $2,512,207 | 0.344 | Won | $5,809,661 | 2.011 |

Table A.3: Donor History by Community in AR

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $78,622,556 | 2.084 | I | $89,025,514 | 3.110 | Lost | $23,636,478 | 0.292 |
| $c_{2,0}$ | R | $37,725,817 | 0.480 | N | $28,621,244 | 0.321 | Won | $80,890,589 | 3.422 |
| $c_{2,1}$ | D | $88,114,927 | 0.653 | I | $194,101,446 | 6.326 | Lost | $25,838,692 | 0.149 |
| $c_{2,1}$ | R | $134,923,801 | 1.531 | N | $30,684,267 | 0.158 | Won | $173,135,986 | 6.701 |
| $c_{2,0} \cap c_{2,1}$ | D | $24,680,610 | 0.742 | I | $52,009,877 | 7.993 | Lost | $5,794,934 | 0.128 |
| $c_{2,0} \cap c_{2,1}$ | R | $33,276,511 | 1.348 | N | $6,507,051 | 0.125 | Won | $45,212,127 | 7.802 |
| $c_{4,0}$ | D | $79,409,745 | 0.718 | I | $168,528,211 | 7.350 | Lost | $20,100,069 | 0.135 |
| $c_{4,0}$ | R | $110,653,982 | 1.393 | N | $22,927,875 | 0.136 | Won | $149,395,090 | 7.433 |
| $c_{4,1}$ | D | $57,979,598 | 2.896 | I | $57,597,684 | 2.712 | Lost | $17,148,743 | 0.323 |
| $c_{4,1}$ | R | $20,023,915 | 0.345 | N | $21,238,922 | 0.369 | Won | $53,012,859 | 3.091 |
| $c_{4,2}$ | D | $25,336,914 | 19.730 | I | $15,378,144 | 1.340 | Lost | $9,327,911 | 0.598 |
| $c_{4,2}$ | R | $1,284,190 | 0.051 | N | $11,473,785 | 0.746 | Won | $15,596,465 | 1.672 |
| $c_{4,3}$ | D | $15,923,977 | 0.460 | I | $41,722,561 | 4.525 | Lost | $7,546,470 | 0.200 |
| $c_{4,3}$ | R | $34,586,570 | 2.172 | N | $9,220,123 | 0.221 | Won | $37,815,840 | 5.011 |

Table A.4: Donor History by Community in AZ

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $171,550 | 1.796 | I | $234,180 | 7.116 | Lost | $23,780 | 0.119 |
| $c_{2,0}$ | R | $95,540 | 0.557 | N | $32,910 | 0.141 | Won | $199,910 | 8.407 |
| $c_{2,1}$ | D | $82,005,227 | 1.086 | I | $127,180,257 | 3.997 | Lost | $26,559,550 | 0.228 |
| $c_{2,1}$ | R | $75,544,659 | 0.921 | N | $31,821,648 | 0.250 | Won | $116,627,976 | 4.391 |
| $c_{2,0} \cap c_{2,1}$ | D | $171,550 | 1.812 | I | $233,480 | 7.129 | Lost | $23,780 | 0.119 |
| $c_{2,0} \cap c_{2,1}$ | R | $94,680 | 0.552 | N | $32,750 | 0.140 | Won | $199,050 | 8.370 |
| $c_{4,0}$ | D | $33,545,933 | 12.278 | I | $22,182,368 | 1.534 | Lost | $12,011,405 | 0.561 |
| $c_{4,0}$ | R | $2,732,169 | 0.081 | N | $14,461,703 | 0.652 | Won | $21,415,110 | 1.783 |
| $c_{4,1}$ | D | $16,288,066 | 0.683 | I | $34,326,596 | 5.374 | Lost | $4,671,557 | 0.146 |
| $c_{4,1}$ | R | $23,855,814 | 1.465 | N | $6,387,439 | 0.186 | Won | $32,072,341 | 6.865 |
| $c_{4,2}$ | D | $23,958,651 | 0.592 | I | $56,062,909 | 6.266 | Lost | $7,735,460 | 0.153 |
| $c_{4,2}$ | R | $40,499,007 | 1.690 | N | $8,947,790 | 0.160 | Won | $50,638,475 | 6.546 |
| $c_{4,3}$ | D | $32,712,701 | 1.817 | I | $40,159,567 | 3.663 | Lost | $9,226,313 | 0.258 |
| $c_{4,3}$ | R | $18,003,450 | 0.550 | N | $10,962,699 | 0.273 | Won | $35,783,542 | 3.878 |

Table A.5: Donor History by Community in CA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $462,193,658 | 1.342 | I | $622,936,188 | 3.277 | Lost | $134,959,301 | 0.217 |
| $c_{2,0}$ | R | $344,442,457 | 0.745 | N | $190,093,268 | 0.305 | Won | $621,631,277 | 4.606 |
| $c_{2,1}$ | D | $843,949 | 1.678 | I | $948,145 | 1.470 | Lost | $515,599 | 0.479 |
| $c_{2,1}$ | R | $502,919 | 0.596 | N | $644,964 | 0.680 | Won | $1,077,510 | 2.090 |
| $c_{2,0} \cap c_{2,1}$ | D | $185,200 | 2.750 | I | $181,700 | 2.374 | Lost | $39,750 | 0.182 |
| $c_{2,0} \cap c_{2,1}$ | R | $67,350 | 0.364 | N | $76,550 | 0.421 | Won | $218,500 | 5.497 |
| $c_{4,0}$ | D | $17,600 | 0.140 | I | $109,638 | 0.542 | Lost | $198,975 | 1.760 |
| $c_{4,0}$ | R | $125,745 | 7.145 | N | $202,375 | 1.846 | Won | $113,038 | 0.568 |
| $c_{4,1}$ | D | $164,844,588 | 16.600 | I | $107,950,768 | 1.578 | Lost | $40,148,123 | 0.308 |
| $c_{4,1}$ | R | $9,930,533 | 0.060 | N | $68,428,242 | 0.634 | Won | $130,421,379 | 3.249 |
| $c_{4,2}$ | D | $657,249 | 2.023 | I | $684,475 | 1.760 | Lost | $287,374 | 0.366 |
| $c_{4,2}$ | R | $324,824 | 0.494 | N | $388,839 | 0.568 | Won | $785,940 | 2.735 |
| $c_{4,3}$ | D | $377,947,946 | 1.107 | I | $570,523,492 | 3.691 | Lost | $112,038,400 | 0.200 |
| $c_{4,3}$ | R | $341,469,715 | 0.903 | N | $154,565,697 | 0.271 | Won | $559,562,532 | 4.994 |

Table A.6: Donor History by Community in CO

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $78,185,123 | 2.174 | I | $83,803,125 | 2.715 | Lost | $24,852,684 | 0.318 |
| $c_{2,0}$ | R | $35,968,415 | 0.460 | N | $30,865,344 | 0.368 | Won | $78,229,814 | 3.148 |
| $c_{2,1}$ | D | $42,466,738 | 0.557 | I | $98,718,086 | 4.718 | Lost | $20,361,434 | 0.233 |
| $c_{2,1}$ | R | $76,268,501 | 1.796 | N | $20,925,503 | 0.212 | Won | $87,215,502 | 4.283 |
| $c_{2,0} \cap c_{2,1}$ | D | $21,295,477 | 0.811 | I | $41,546,229 | 6.696 | Lost | $5,521,860 | 0.152 |
| $c_{2,0} \cap c_{2,1}$ | R | $26,257,692 | 1.233 | N | $6,204,760 | 0.149 | Won | $36,427,443 | 6.597 |
| $c_{4,0}$ | D | $54,280,283 | 9.202 | I | $36,734,215 | 1.547 | Lost | $18,695,045 | 0.502 |
| $c_{4,0}$ | R | $5,898,518 | 0.109 | N | $23,747,822 | 0.646 | Won | $37,227,987 | 1.991 |
| $c_{4,1}$ | D | $2,408,341 | 0.110 | I | $17,302,589 | 2.373 | Lost | $9,089,230 | 0.677 |
| $c_{4,1}$ | R | $21,841,569 | 9.069 | N | $7,291,473 | 0.421 | Won | $13,418,359 | 1.476 |
| $c_{4,2}$ | D | $4,873,156 | 0.909 | I | $7,703,093 | 2.971 | Lost | $1,767,610 | 0.258 |
| $c_{4,2}$ | R | $5,358,292 | 1.100 | N | $2,592,461 | 0.337 | Won | $6,861,954 | 3.882 |
| $c_{4,3}$ | D | $39,137,658 | 0.722 | I | $80,867,136 | 6.189 | Lost | $10,868,592 | 0.148 |
| $c_{4,3}$ | R | $54,239,001 | 1.386 | N | $13,066,422 | 0.162 | Won | $73,365,682 | 6.750 |

Table A.7: Donor History by Community in CT

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $5,390 | 0.596 | I | $8,870 | 1.537 | Lost | $5,150 | 0.543 |
| $c_{2,0}$ | R | $9,050 | 1.679 | N | $5,770 | 0.651 | Won | $9,490 | 1.843 |
| $c_{2,1}$ | D | $12,987,096 | 0.854 | I | $24,655,426 | 6.741 | Lost | $3,163,494 | 0.143 |
| $c_{2,1}$ | R | $15,209,527 | 1.171 | N | $3,657,647 | 0.148 | Won | $22,054,806 | 6.972 |
| $c_{2,0} \cap c_{2,1}$ | D | $0 | NA | I | $0 | NA | Lost | $0 | NA |
| $c_{2,0} \cap c_{2,1}$ | R | $0 | NA | N | $0 | NA | Won | $0 | NA |
| $c_{4,0}$ | D | $25 | 0.004 | I | $3,675 | 1.329 | Lost | $2,545 | 0.653 |
| $c_{4,0}$ | R | $6,415 | 256.600 | N | $2,765 | 0.752 | Won | $3,895 | 1.530 |
| $c_{4,1}$ | D | $4,290 | 10.094 | I | $3,260 | 1.970 | Lost | $1,400 | 0.398 |
| $c_{4,1}$ | R | $425 | 0.099 | N | $1,655 | 0.508 | Won | $3,515 | 2.511 |
| $c_{4,2}$ | D | $3,060 | NA | I | $1,475 | 0.931 | Lost | $1,080 | 0.545 |
| $c_{4,2}$ | R | $0 | NA | N | $1,585 | 1.075 | Won | $1,980 | 1.833 |
| $c_{4,3}$ | D | $12,985,111 | 0.854 | I | $24,655,886 | 6.741 | Lost | $3,163,774 | 0.143 |
| $c_{4,3}$ | R | $15,211,892 | 1.171 | N | $3,657,567 | 0.148 | Won | $22,054,906 | 6.971 |

Table A.8: Donor History by Community in DE

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $49,302,417 | 0.761 | I | $100,548,273 | 6.991 | Lost | $14,002,157 | 0.157 |
| $c_{2,0}$ | R | $64,780,247 | 1.314 | N | $14,383,165 | 0.143 | Won | $88,960,053 | 6.353 |
| $c_{2,1}$ | D | $83,628,947 | 1.331 | I | $122,664,162 | 4.960 | Lost | $21,381,708 | 0.192 |
| $c_{2,1}$ | R | $62,828,631 | 0.751 | N | $24,732,871 | 0.202 | Won | $111,164,213 | 5.199 |
| $c_{2,0} \cap c_{2,1}$ | D | $35,749,106 | 0.903 | I | $67,306,486 | 7.946 | Lost | $8,378,331 | 0.140 |
| $c_{2,0} \cap c_{2,1}$ | R | $39,582,378 | 1.107 | N | $8,470,122 | 0.126 | Won | $59,706,953 | 7.126 |
| $c_{4,0}$ | D | $56,800,432 | 0.925 | I | $103,544,029 | 6.689 | Lost | $14,417,455 | 0.156 |
| $c_{4,0}$ | R | $61,379,789 | 1.081 | N | $15,480,330 | 0.150 | Won | $92,164,854 | 6.393 |
| $c_{4,1}$ | D | $22,841,284 | 14.945 | I | $16,062,871 | 1.866 | Lost | $6,761,364 | 0.437 |
| $c_{4,1}$ | R | $1,528,349 | 0.067 | N | $8,606,722 | 0.536 | Won | $15,472,932 | 2.288 |
| $c_{4,2}$ | D | $41,407,115 | 0.828 | I | $80,405,994 | 6.950 | Lost | $10,845,785 | 0.151 |
| $c_{4,2}$ | R | $50,005,976 | 1.208 | N | $11,569,824 | 0.144 | Won | $71,770,034 | 6.617 |
| $c_{4,3}$ | D | $896,352 | 0.195 | I | $3,388,763 | 1.555 | Lost | $1,855,280 | 0.546 |
| $c_{4,3}$ | R | $4,604,434 | 5.137 | N | $2,179,473 | 0.643 | Won | $3,396,076 | 1.830 |

Table A.9: Donor History by Community in FL

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $648,096 | 0.600 | I | $1,584,383 | 2.737 | Lost | $372,425 | 0.226 |
| $c_{2,0}$ | R | $1,080,700 | 1.667 | N | $578,820 | 0.365 | Won | $1,647,328 | 4.423 |
| $c_{2,1}$ | D | $194,741,889 | 0.753 | I | $372,580,657 | 4.403 | Lost | $64,770,104 | 0.191 |
| $c_{2,1}$ | R | $258,665,258 | 1.328 | N | $84,616,868 | 0.227 | Won | $339,882,967 | 5.248 |
| $c_{2,0} \cap c_{2,1}$ | D | $339,345 | 0.387 | I | $898,037 | 2.204 | Lost | $189,600 | 0.189 |
| $c_{2,0} \cap c_{2,1}$ | R | $876,975 | 2.584 | N | $407,520 | 0.454 | Won | $1,003,207 | 5.291 |
| $c_{4,0}$ | D | $175,552,537 | 0.688 | I | $356,614,284 | 4.621 | Lost | $58,605,624 | 0.180 |
| $c_{4,0}$ | R | $255,314,179 | 1.454 | N | $77,172,389 | 0.216 | Won | $325,301,080 | 5.551 |
| $c_{4,1}$ | D | $632,371 | 1.115 | I | $999,551 | 2.207 | Lost | $326,873 | 0.338 |
| $c_{4,1}$ | R | $566,945 | 0.897 | N | $452,869 | 0.453 | Won | $967,873 | 2.961 |
| $c_{4,2}$ | D | $18,886,050 | 5.971 | I | $15,463,158 | 2.129 | Lost | $5,987,107 | 0.423 |
| $c_{4,2}$ | R | $3,162,904 | 0.167 | N | $7,262,860 | 0.470 | Won | $14,157,626 | 2.365 |
| $c_{4,3}$ | D | $18,451,966 | 7.141 | I | $14,949,135 | 2.187 | Lost | $5,698,052 | 0.419 |
| $c_{4,3}$ | R | $2,583,784 | 0.140 | N | $6,834,466 | 0.457 | Won | $13,605,827 | 2.388 |

Table A.10: Donor History by Community in GA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $185,161,097 | 0.414 | I | $366,200,801 | 1.361 | Lost | $262,926,440 | 0.823 |
| $c_{2,0}$ | R | $447,190,453 | 2.415 | N | $269,035,421 | 0.735 | Won | $319,395,546 | 1.215 |
| $c_{2,1}$ | D | $4,277,540 | 0.967 | I | $7,263,474 | 3.885 | Lost | $1,832,268 | 0.289 |
| $c_{2,1}$ | R | $4,423,022 | 1.034 | N | $1,869,696 | 0.257 | Won | $6,349,895 | 3.466 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,490,691 | 0.680 | I | $3,171,124 | 5.129 | Lost | $580,695 | 0.206 |
| $c_{2,0} \cap c_{2,1}$ | R | $2,192,841 | 1.471 | N | $618,317 | 0.195 | Won | $2,812,494 | 4.843 |
| $c_{4,0}$ | D | $1,801,099 | 1.242 | I | $2,655,115 | 2.997 | Lost | $874,197 | 0.377 |
| $c_{4,0}$ | R | $1,450,596 | 0.805 | N | $885,879 | 0.334 | Won | $2,317,897 | 2.651 |
| $c_{4,1}$ | D | $61,616,592 | 1.808 | I | $72,968,794 | 3.056 | Lost | $20,657,723 | 0.318 |
| $c_{4,1}$ | R | $34,088,798 | 0.553 | N | $23,875,574 | 0.327 | Won | $64,907,617 | 3.142 |
| $c_{4,2}$ | D | $11,915,765 | 2.238 | I | $13,878,204 | 3.835 | Lost | $3,222,953 | 0.277 |
| $c_{4,2}$ | R | $5,324,555 | 0.447 | N | $3,618,374 | 0.261 | Won | $11,629,756 | 3.608 |
| $c_{4,3}$ | D | $134,388,233 | 0.309 | I | $321,869,190 | 1.288 | Lost | $246,380,244 | 0.879 |
| $c_{4,3}$ | R | $435,293,182 | 3.239 | N | $249,861,239 | 0.776 | Won | $280,241,706 | 1.137 |

Table A.11: Donor History by Community in HI

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $2,638,248 | 0.150 | I | $14,522,002 | 2.496 | Lost | $7,840,631 | 0.684 |
| $c_{2,0}$ | R | $17,558,132 | 6.655 | N | $5,817,827 | 0.401 | Won | $11,462,052 | 1.462 |
| $c_{2,1}$ | D | $47,274,085 | 1.002 | I | $79,022,708 | 4.862 | Lost | $14,330,164 | 0.201 |
| $c_{2,1}$ | R | $47,193,318 | 0.998 | N | $16,253,268 | 0.206 | Won | $71,287,481 | 4.975 |
| $c_{2,0} \cap c_{2,1}$ | D | $623,090 | 0.246 | I | $2,587,591 | 4.329 | Lost | $630,584 | 0.272 |
| $c_{2,0} \cap c_{2,1}$ | R | $2,532,141 | 4.064 | N | $597,739 | 0.231 | Won | $2,318,557 | 3.677 |
| $c_{4,0}$ | D | $17,489,799 | 1.361 | I | $22,947,689 | 2.987 | Lost | $6,760,165 | 0.323 |
| $c_{4,0}$ | R | $12,851,908 | 0.735 | N | $7,683,370 | 0.335 | Won | $20,931,724 | 3.096 |
| $c_{4,1}$ | D | $60,289 | 26.795 | I | $33,550 | 1.076 | Lost | $20,750 | 0.550 |
| $c_{4,1}$ | R | $2,250 | 0.037 | N | $31,189 | 0.930 | Won | $37,739 | 1.819 |
| $c_{4,2}$ | D | $2,563,298 | 0.146 | I | $14,480,002 | 2.505 | Lost | $7,810,881 | 0.684 |
| $c_{4,2}$ | R | $17,554,482 | 6.848 | N | $5,779,927 | 0.399 | Won | $11,421,152 | 1.462 |
| $c_{4,3}$ | D | $38,388,470 | 0.866 | I | $71,699,662 | 6.088 | Lost | $10,272,231 | 0.159 |
| $c_{4,3}$ | R | $44,340,973 | 1.155 | N | $11,777,054 | 0.164 | Won | $64,446,193 | 6.274 |

Table A.12: Donor History by Community in IA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $76,855,763 | 0.654 | I | $164,720,790 | 5.252 | Lost | $28,005,873 | 0.189 |
| $c_{2,0}$ | R | $117,569,381 | 1.530 | N | $31,361,832 | 0.190 | Won | $148,193,906 | 5.292 |
| $c_{2,1}$ | D | $64,096,047 | 7.524 | I | $48,602,790 | 1.976 | Lost | $19,864,704 | 0.441 |
| $c_{2,1}$ | R | $8,519,086 | 0.133 | N | $24,599,921 | 0.506 | Won | $45,016,162 | 2.266 |
| $c_{2,0} \cap c_{2,1}$ | D | $6,975,872 | 2.036 | I | $7,725,878 | 2.861 | Lost | $2,306,996 | 0.315 |
| $c_{2,0} \cap c_{2,1}$ | R | $3,426,573 | 0.491 | N | $2,700,251 | 0.350 | Won | $7,326,153 | 3.176 |
| $c_{4,0}$ | D | $56,228,356 | 14.480 | I | $39,130,558 | 1.817 | Lost | $17,194,771 | 0.475 |
| $c_{4,0}$ | R | $3,883,084 | 0.069 | N | $21,541,577 | 0.551 | Won | $36,229,379 | 2.107 |
| $c_{4,1}$ | D | $14,243,306 | 0.488 | I | $32,897,511 | 3.049 | Lost | $9,076,089 | 0.294 |
| $c_{4,1}$ | R | $29,206,553 | 2.051 | N | $10,787,956 | 0.328 | Won | $30,912,857 | 3.406 |
| $c_{4,2}$ | D | $17,827,599 | 1.451 | I | $24,046,123 | 3.834 | Lost | $5,847,307 | 0.269 |
| $c_{4,2}$ | R | $12,287,966 | 0.689 | N | $6,271,723 | 0.261 | Won | $21,758,603 | 3.721 |
| $c_{4,3}$ | D | $66,890,469 | 0.662 | I | $147,765,801 | 6.835 | Lost | $19,716,670 | 0.150 |
| $c_{4,3}$ | R | $101,000,423 | 1.510 | N | $21,618,882 | 0.146 | Won | $131,481,793 | 6.669 |

Table A.13: Donor History by Community in ID

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $72,054,345 | 0.688 | I | $150,311,673 | 5.327 | Lost | $23,560,714 | 0.173 |
| $c_{2,0}$ | R | $104,734,825 | 1.454 | N | $28,215,833 | 0.188 | Won | $136,335,510 | 5.787 |
| $c_{2,1}$ | D | $39,741,691 | 15.785 | I | $28,215,980 | 1.932 | Lost | $11,955,699 | 0.465 |
| $c_{2,1}$ | R | $2,517,743 | 0.063 | N | $14,602,667 | 0.518 | Won | $25,706,475 | 2.150 |
| $c_{2,0} \cap c_{2,1}$ | D | $7,349,467 | 8.826 | I | $5,595,833 | 2.086 | Lost | $2,110,343 | 0.426 |
| $c_{2,0} \cap c_{2,1}$ | R | $832,685 | 0.113 | N | $2,682,549 | 0.479 | Won | $4,953,249 | 2.347 |
| $c_{4,0}$ | D | $9,512,687 | 23.213 | I | $6,969,118 | 2.317 | Lost | $2,400,199 | 0.381 |
| $c_{4,0}$ | R | $409,803 | 0.043 | N | $3,008,122 | 0.432 | Won | $6,297,241 | 2.624 |
| $c_{4,1}$ | D | $3,307,030 | 1.492 | I | $4,196,084 | 2.953 | Lost | $1,134,967 | 0.287 |
| $c_{4,1}$ | R | $2,216,582 | 0.670 | N | $1,420,733 | 0.339 | Won | $3,950,730 | 3.481 |
| $c_{4,2}$ | D | $39,474,990 | 15.143 | I | $28,058,048 | 1.920 | Lost | $11,899,485 | 0.462 |
| $c_{4,2}$ | R | $2,606,883 | 0.066 | N | $14,612,899 | 0.521 | Won | $25,779,639 | 2.166 |
| $c_{4,3}$ | D | $61,543,900 | 0.603 | I | $141,049,375 | 5.835 | Lost | $20,372,614 | 0.159 |
| $c_{4,3}$ | R | $102,126,018 | 1.659 | N | $24,172,293 | 0.171 | Won | $127,884,832 | 6.277 |

Table A.14: Donor History by Community in IL

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $15,827,551 | 3.041 | I | $16,056,603 | 2.897 | Lost | $4,609,219 | 0.320 |
| $c_{2,0}$ | R | $5,205,134 | 0.329 | N | $5,543,027 | 0.345 | Won | $14,424,322 | 3.129 |
| $c_{2,1}$ | D | $375,017,586 | 1.294 | I | $554,878,715 | 4.725 | Lost | $109,552,706 | 0.228 |
| $c_{2,1}$ | R | $289,734,099 | 0.773 | N | $117,435,278 | 0.212 | Won | $480,518,701 | 4.386 |
| $c_{2,0} \cap c_{2,1}$ | D | $10,876,921 | 2.813 | I | $11,362,723 | 3.130 | Lost | $2,991,256 | 0.293 |
| $c_{2,0} \cap c_{2,1}$ | R | $3,867,326 | 0.356 | N | $3,629,749 | 0.319 | Won | $10,204,742 | 3.412 |
| $c_{4,0}$ | D | $163,041,830 | 4.762 | I | $154,015,609 | 3.260 | Lost | $42,110,231 | 0.321 |
| $c_{4,0}$ | R | $34,241,052 | 0.210 | N | $47,241,854 | 0.307 | Won | $131,164,128 | 3.115 |
| $c_{4,1}$ | D | $247,636,179 | 1.000 | I | $436,649,227 | 7.025 | Lost | $62,233,639 | 0.166 |
| $c_{4,1}$ | R | $247,554,294 | 1.000 | N | $62,159,423 | 0.142 | Won | $373,900,447 | 6.008 |
| $c_{4,2}$ | D | $1,454,310 | 0.059 | I | $10,270,230 | 0.630 | Lost | $13,341,576 | 1.148 |
| $c_{4,2}$ | R | $24,710,568 | 16.991 | N | $16,307,248 | 1.588 | Won | $11,625,320 | 0.871 |
| $c_{4,3}$ | D | $17,912,863 | 4.937 | I | $17,115,783 | 3.373 | Lost | $4,175,202 | 0.278 |
| $c_{4,3}$ | R | $3,627,995 | 0.203 | N | $5,074,937 | 0.297 | Won | $15,014,409 | 3.596 |

Table A.15: Donor History by Community in IN

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $117,997,990 | 0.594 | I | $274,424,272 | 6.195 | Lost | $37,616,716 | 0.153 |
| $c_{2,0}$ | R | $198,635,595 | 1.683 | N | $44,298,769 | 0.161 | Won | $246,107,013 | 6.542 |
| $c_{2,1}$ | D | $107,337,869 | 11.333 | I | $76,119,451 | 1.831 | Lost | $36,261,182 | 0.550 |
| $c_{2,1}$ | R | $9,471,564 | 0.088 | N | $41,566,026 | 0.546 | Won | $65,980,129 | 1.820 |
| $c_{2,0} \cap c_{2,1}$ | D | $956,756 | 0.735 | I | $1,892,831 | 4.558 | Lost | $356,301 | 0.212 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,301,430 | 1.360 | N | $415,316 | 0.219 | Won | $1,680,405 | 4.716 |
| $c_{4,0}$ | D | $105,099,561 | 14.532 | I | $72,520,865 | 1.785 | Lost | $35,393,653 | 0.564 |
| $c_{4,0}$ | R | $7,232,478 | 0.069 | N | $40,617,055 | 0.560 | Won | $62,792,004 | 1.774 |
| $c_{4,1}$ | D | $54,622,800 | 0.512 | I | $135,958,529 | 5.082 | Lost | $20,785,521 | 0.165 |
| $c_{4,1}$ | R | $106,764,709 | 1.955 | N | $26,751,391 | 0.197 | Won | $125,647,694 | 6.045 |
| $c_{4,2}$ | D | $11,679,310 | 3.570 | I | $12,294,084 | 4.415 | Lost | $2,699,051 | 0.309 |
| $c_{4,2}$ | R | $3,271,429 | 0.280 | N | $2,784,717 | 0.227 | Won | $8,727,724 | 3.234 |
| $c_{4,3}$ | D | $83,103,288 | 0.681 | I | $182,351,826 | 7.694 | Lost | $21,856,956 | 0.136 |
| $c_{4,3}$ | R | $121,956,785 | 1.468 | N | $23,700,131 | 0.130 | Won | $160,494,121 | 7.343 |

Table A.16: Donor History by Community in KS

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $31,382,439 | 14.873 | I | $20,966,162 | 1.632 | Lost | $11,285,916 | 0.574 |
| $c_{2,0}$ | R | $2,109,961 | 0.067 | N | $12,850,552 | 0.613 | Won | $19,674,667 | 1.743 |
| $c_{2,1}$ | D | $139,480,540 | 0.803 | I | $264,985,896 | 5.270 | Lost | $43,154,592 | 0.180 |
| $c_{2,1}$ | R | $173,655,433 | 1.245 | N | $50,283,768 | 0.190 | Won | $239,389,010 | 5.547 |
| $c_{2,0} \cap c_{2,1}$ | D | $25,589,740 | 12.350 | I | $18,296,974 | 1.892 | Lost | $8,273,774 | 0.486 |
| $c_{2,0} \cap c_{2,1}$ | R | $2,072,111 | 0.081 | N | $9,668,641 | 0.528 | Won | $17,021,444 | 2.057 |
| $c_{4,0}$ | D | $13,981,351 | 24.433 | I | $7,734,849 | 1.124 | Lost | $5,998,595 | 0.761 |
| $c_{4,0}$ | R | $572,238 | 0.041 | N | $6,882,418 | 0.890 | Won | $7,880,713 | 1.314 |
| $c_{4,1}$ | D | $35,913,755 | 10.371 | I | $26,574,820 | 2.017 | Lost | $10,843,154 | 0.438 |
| $c_{4,1}$ | R | $3,462,797 | 0.096 | N | $13,174,855 | 0.496 | Won | $24,769,352 | 2.284 |
| $c_{4,2}$ | D | $79,661,362 | 0.567 | I | $192,622,074 | 6.607 | Lost | $25,368,786 | 0.147 |
| $c_{4,2}$ | R | $140,593,932 | 1.765 | N | $29,152,549 | 0.151 | Won | $172,825,497 | 6.813 |
| $c_{4,3}$ | D | $41,226,812 | 0.708 | I | $86,952,284 | 6.550 | Lost | $11,303,252 | 0.143 |
| $c_{4,3}$ | R | $58,238,924 | 1.413 | N | $13,275,766 | 0.153 | Won | $79,313,365 | 7.017 |

Table A.17: Donor History by Community in KY

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $124,374,074 | 1.249 | I | $180,336,438 | 3.973 | Lost | $38,845,423 | 0.241 |
| $c_{2,0}$ | R | $99,602,966 | 0.801 | N | $45,387,628 | 0.252 | Won | $161,203,396 | 4.150 |
| $c_{2,1}$ | D | $13,500 | 0.004 | I | $2,318,308 | 2.132 | Lost | $2,290,457 | 2.069 |
| $c_{2,1}$ | R | $3,357,281 | 248.687 | N | $1,087,563 | 0.469 | Won | $1,107,013 | 0.483 |
| $c_{2,0} \cap c_{2,1}$ | D | $13,500 | 0.279 | I | $38,675 | 1.655 | Lost | $14,716 | 0.316 |
| $c_{2,0} \cap c_{2,1}$ | R | $48,441 | 3.588 | N | $23,366 | 0.604 | Won | $46,575 | 3.165 |
| $c_{4,0}$ | D | $28,390,364 | 0.614 | I | $66,625,435 | 7.872 | Lost | $8,204,242 | 0.141 |
| $c_{4,0}$ | R | $46,225,266 | 1.628 | N | $8,463,623 | 0.127 | Won | $58,083,469 | 7.080 |
| $c_{4,1}$ | D | $603,027 | 0.095 | I | $4,742,445 | 2.080 | Lost | $3,263,382 | 1.021 |
| $c_{4,1}$ | R | $6,325,221 | 10.489 | N | $2,279,543 | 0.481 | Won | $3,195,010 | 0.979 |
| $c_{4,2}$ | D | $63,121,551 | 18.935 | I | $40,543,024 | 1.526 | Lost | $21,479,688 | 0.551 |
| $c_{4,2}$ | R | $3,333,541 | 0.053 | N | $26,560,655 | 0.655 | Won | $38,989,017 | 1.815 |
| $c_{4,3}$ | D | $43,212,696 | 0.688 | I | $94,985,795 | 8.113 | Lost | $10,787,238 | 0.130 |
| $c_{4,3}$ | R | $62,832,887 | 1.454 | N | $11,708,054 | 0.123 | Won | $82,773,894 | 7.673 |

Table A.18: Donor History by Community in LA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $4,669,561 | 0.421 | I | $11,073,659 | 2.279 | Lost | $3,405,658 | 0.298 |
| $c_{2,0}$ | R | $11,101,047 | 2.377 | N | $4,858,974 | 0.439 | Won | $11,427,573 | 3.355 |
| $c_{2,1}$ | D | $29,033,073 | 0.612 | I | $64,663,196 | 5.250 | Lost | $9,707,142 | 0.164 |
| $c_{2,1}$ | R | $47,466,530 | 1.635 | N | $12,316,571 | 0.190 | Won | $59,343,873 | 6.113 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,737,468 | 0.253 | I | $5,939,015 | 2.204 | Lost | $1,756,483 | 0.290 |
| $c_{2,0} \cap c_{2,1}$ | R | $6,857,021 | 3.947 | N | $2,694,450 | 0.454 | Won | $6,055,779 | 3.448 |
| $c_{4,0}$ | D | $1,669,768 | 0.274 | I | $5,591,973 | 2.542 | Lost | $1,816,626 | 0.350 |
| $c_{4,0}$ | R | $6,085,606 | 3.645 | N | $2,199,876 | 0.393 | Won | $5,186,221 | 2.855 |
| $c_{4,1}$ | D | $6,718,832 | 0.608 | I | $13,655,382 | 3.177 | Lost | $2,834,698 | 0.206 |
| $c_{4,1}$ | R | $11,046,823 | 1.644 | N | $4,298,192 | 0.315 | Won | $13,729,711 | 4.843 |
| $c_{4,2}$ | D | $25,350,055 | 0.680 | I | $54,424,811 | 6.369 | Lost | $7,154,127 | 0.146 |
| $c_{4,2}$ | R | $37,261,481 | 1.470 | N | $8,545,864 | 0.157 | Won | $49,106,573 | 6.864 |
| $c_{4,3}$ | D | $1,677,050 | 0.572 | I | $3,626,056 | 3.314 | Lost | $801,050 | 0.219 |
| $c_{4,3}$ | R | $2,933,181 | 1.749 | N | $1,094,175 | 0.302 | Won | $3,659,031 | 4.568 |

Table A.19: Donor History by Community in MA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $1,405,346 | 0.309 | I | $3,870,167 | 1.789 | Lost | $1,621,156 | 0.415 |
| $c_{2,0}$ | R | $4,541,654 | 3.232 | N | $2,163,254 | 0.559 | Won | $3,901,805 | 2.407 |
| $c_{2,1}$ | D | $60,402,169 | 2.722 | I | $67,194,067 | 4.222 | Lost | $13,856,940 | 0.228 |
| $c_{2,1}$ | R | $22,191,544 | 0.367 | N | $15,916,568 | 0.237 | Won | $60,863,768 | 4.392 |
| $c_{2,0} \cap c_{2,1}$ | D | $249,000 | 1.526 | I | $370,550 | 8.897 | Lost | $50,850 | 0.157 |
| $c_{2,0} \cap c_{2,1}$ | R | $163,200 | 0.655 | N | $41,650 | 0.112 | Won | $324,700 | 6.385 |
| $c_{4,0}$ | D | $400 | 0.031 | I | $2,060 | 0.186 | Lost | $8,070 | 1.586 |
| $c_{4,0}$ | R | $12,759 | 31.896 | N | $11,099 | 5.388 | Won | $5,089 | 0.631 |
| $c_{4,1}$ | D | $60,402,169 | 2.722 | I | $67,194,067 | 4.222 | Lost | $13,856,940 | 0.228 |
| $c_{4,1}$ | R | $22,191,544 | 0.367 | N | $15,916,568 | 0.237 | Won | $60,863,768 | 4.392 |
| $c_{4,2}$ | D | $1,156,046 | 0.265 | I | $3,496,347 | 1.662 | Lost | $1,554,002 | 0.435 |
| $c_{4,2}$ | R | $4,358,125 | 3.770 | N | $2,104,220 | 0.602 | Won | $3,572,755 | 2.299 |
| $c_{4,3}$ | D | $175 | 0.008 | I | $2,385 | 0.113 | Lost | $20,359 | 6.536 |
| $c_{4,3}$ | R | $23,174 | 132.423 | N | $21,089 | 8.842 | Won | $3,115 | 0.153 |

Table A.20: Donor History by Community in MD

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $121,206,819 | 1.098 | I | $198,347,456 | 5.720 | Lost | $31,463,573 | 0.181 |
| $c_{2,0}$ | R | $110,385,020 | 0.911 | N | $34,674,436 | 0.175 | Won | $173,444,742 | 5.513 |
| $c_{2,1}$ | D | $28,742,640 | 1.539 | I | $39,227,632 | 4.684 | Lost | $6,979,755 | 0.203 |
| $c_{2,1}$ | R | $18,670,363 | 0.650 | N | $8,374,843 | 0.213 | Won | $34,314,762 | 4.916 |
| $c_{2,0} \cap c_{2,1}$ | D | $16,061,726 | 1.050 | I | $26,854,967 | 5.805 | Lost | $3,907,244 | 0.164 |
| $c_{2,0} \cap c_{2,1}$ | R | $15,294,331 | 0.952 | N | $4,626,040 | 0.172 | Won | $23,853,073 | 6.105 |
| $c_{4,0}$ | D | $52,537,099 | 2.770 | I | $56,562,810 | 3.673 | Lost | $13,544,081 | 0.279 |
| $c_{4,0}$ | R | $18,968,770 | 0.361 | N | $15,397,890 | 0.272 | Won | $48,534,502 | 3.583 |
| $c_{4,1}$ | D | $25,690,956 | 1.387 | I | $37,116,769 | 5.103 | Lost | $6,157,975 | 0.190 |
| $c_{4,1}$ | R | $18,519,834 | 0.721 | N | $7,273,543 | 0.196 | Won | $32,367,745 | 5.256 |
| $c_{4,2}$ | D | $46,846,172 | 0.760 | I | $95,833,652 | 7.241 | Lost | $12,148,267 | 0.144 |
| $c_{4,2}$ | R | $61,648,065 | 1.316 | N | $13,235,257 | 0.138 | Won | $84,399,103 | 6.947 |
| $c_{4,3}$ | D | $19,231,343 | 0.755 | I | $40,062,979 | 8.050 | Lost | $5,193,784 | 0.150 |
| $c_{4,3}$ | R | $25,472,312 | 1.325 | N | $4,976,858 | 0.124 | Won | $34,649,716 | 6.671 |

Table A.21: Donor History by Community in ME

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $9,139,779 | 1.552 | I | $11,373,459 | 2.893 | Lost | $3,521,290 | 0.333 |
| $c_{2,0}$ | R | $5,888,252 | 0.644 | N | $3,931,572 | 0.346 | Won | $10,585,946 | 3.006 |
| $c_{2,1}$ | D | $13,361,671 | 0.497 | I | $34,934,839 | 6.142 | Lost | $5,224,790 | 0.169 |
| $c_{2,1}$ | R | $26,897,823 | 2.013 | N | $5,687,743 | 0.163 | Won | $30,981,152 | 5.930 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,987,176 | 0.614 | I | $4,695,376 | 8.466 | Lost | $557,925 | 0.133 |
| $c_{2,0} \cap c_{2,1}$ | R | $3,235,775 | 1.628 | N | $554,625 | 0.118 | Won | $4,179,401 | 7.491 |
| $c_{4,0}$ | D | $13,941,704 | 0.631 | I | $32,482,223 | 8.013 | Lost | $3,657,126 | 0.128 |
| $c_{4,0}$ | R | $22,110,079 | 1.586 | N | $4,053,449 | 0.125 | Won | $28,536,342 | 7.803 |
| $c_{4,1}$ | D | $2,479,367 | 0.245 | I | $10,054,582 | 3.700 | Lost | $2,335,697 | 0.255 |
| $c_{4,1}$ | R | $10,131,091 | 4.086 | N | $2,717,174 | 0.270 | Won | $9,175,374 | 3.928 |
| $c_{4,2}$ | D | $6,495,629 | 2.982 | I | $5,704,958 | 1.773 | Lost | $2,874,640 | 0.534 |
| $c_{4,2}$ | R | $2,178,027 | 0.335 | N | $3,217,647 | 0.564 | Won | $5,381,345 | 1.872 |
| $c_{4,3}$ | D | $4,160 | 0.010 | I | $184,922 | 0.779 | Lost | $195,822 | 1.248 |
| $c_{4,3}$ | R | $417,344 | 100.323 | N | $237,282 | 1.283 | Won | $156,917 | 0.801 |

Table A.22: Donor History by Community in MI

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $156,357,241 | 1.059 | I | $242,211,275 | 3.617 | Lost | $52,540,906 | 0.240 |
| $c_{2,0}$ | R | $147,613,350 | 0.944 | N | $66,963,985 | 0.276 | Won | $219,328,462 | 4.174 |
| $c_{2,1}$ | D | $652,426 | 1.029 | I | $1,097,657 | 1.938 | Lost | $557,250 | 0.510 |
| $c_{2,1}$ | R | $634,178 | 0.972 | N | $566,462 | 0.516 | Won | $1,093,308 | 1.962 |
| $c_{2,0} \cap c_{2,1}$ | D | $125,015 | 1.171 | I | $215,281 | 1.883 | Lost | $136,945 | 0.757 |
| $c_{2,0} \cap c_{2,1}$ | R | $106,760 | 0.854 | N | $114,314 | 0.531 | Won | $180,875 | 1.321 |
| $c_{4,0}$ | D | $4,527,941 | 2.001 | I | $4,735,679 | 1.985 | Lost | $1,692,107 | 0.389 |
| $c_{4,0}$ | R | $2,262,946 | 0.500 | N | $2,385,977 | 0.504 | Won | $4,351,566 | 2.572 |
| $c_{4,1}$ | D | $86,189,788 | 0.600 | I | $200,031,149 | 5.961 | Lost | $28,064,179 | 0.160 |
| $c_{4,1}$ | R | $143,698,292 | 1.667 | N | $33,558,596 | 0.168 | Won | $175,891,914 | 6.267 |
| $c_{4,2}$ | D | $306,240 | 0.842 | I | $579,983 | 2.411 | Lost | $228,735 | 0.392 |
| $c_{4,2}$ | R | $363,520 | 1.187 | N | $240,560 | 0.415 | Won | $583,983 | 2.553 |
| $c_{4,3}$ | D | $67,800,981 | 18.672 | I | $40,427,982 | 1.241 | Lost | $23,802,757 | 0.567 |
| $c_{4,3}$ | R | $3,631,064 | 0.054 | N | $32,584,211 | 0.806 | Won | $41,954,867 | 1.763 |

Table A.23: Donor History by Community in MN

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $65,518,584 | 3.259 | I | $57,847,903 | 2.020 | Lost | $23,423,914 | 0.434 |
| $c_{2,0}$ | R | $20,104,427 | 0.307 | N | $28,636,013 | 0.495 | Won | $53,985,075 | 2.305 |
| $c_{2,1}$ | D | $1,550 | 0.016 | I | $33,750 | 0.504 | Lost | $60,378 | 1.567 |
| $c_{2,1}$ | R | $98,003 | 63.228 | N | $66,903 | 1.982 | Won | $38,525 | 0.638 |
| $c_{2,0} \cap c_{2,1}$ | D | $150 | 0.023 | I | $600 | 0.098 | Lost | $4,333 | 6.666 |
| $c_{2,0} \cap c_{2,1}$ | R | $6,583 | 43.887 | N | $6,133 | 10.222 | Won | $650 | 0.150 |
| $c_{4,0}$ | D | $4,779,603 | 4.395 | I | $4,422,478 | 3.021 | Lost | $1,465,053 | 0.393 |
| $c_{4,0}$ | R | $1,087,535 | 0.228 | N | $1,464,059 | 0.331 | Won | $3,726,495 | 2.544 |
| $c_{4,1}$ | D | $5,782,767 | 0.583 | I | $13,204,575 | 5.106 | Lost | $2,332,804 | 0.204 |
| $c_{4,1}$ | R | $9,917,754 | 1.715 | N | $2,586,037 | 0.196 | Won | $11,418,607 | 4.895 |
| $c_{4,2}$ | D | $10,471,997 | 0.627 | I | $22,642,982 | 4.666 | Lost | $4,088,930 | 0.202 |
| $c_{4,2}$ | R | $16,707,299 | 1.595 | N | $4,852,326 | 0.214 | Won | $20,231,243 | 4.948 |
| $c_{4,3}$ | D | $53,084,769 | 20.682 | I | $33,246,177 | 1.449 | Lost | $18,468,523 | 0.576 |
| $c_{4,3}$ | R | $2,566,747 | 0.048 | N | $22,939,183 | 0.690 | Won | $32,053,480 | 1.736 |

Table A.24: Donor History by Community in MO

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $87,415,306 | 0.536 | I | $210,326,112 | 5.063 | Lost | $39,535,667 | 0.213 |
| $c_{2,0}$ | R | $163,032,534 | 1.865 | N | $41,543,403 | 0.198 | Won | $185,354,635 | 4.688 |
| $c_{2,1}$ | D | $168,090,569 | 1.157 | I | $258,016,511 | 4.495 | Lost | $48,862,198 | 0.210 |
| $c_{2,1}$ | R | $145,278,193 | 0.864 | N | $57,396,243 | 0.222 | Won | $232,317,120 | 4.755 |
| $c_{2,0} \cap c_{2,1}$ | D | $61,306,856 | 0.652 | I | $136,070,185 | 6.774 | Lost | $18,072,369 | 0.150 |
| $c_{2,0} \cap c_{2,1}$ | R | $93,973,701 | 1.533 | N | $20,086,942 | 0.148 | Won | $120,405,524 | 6.662 |
| $c_{4,0}$ | D | $114,043,222 | 0.662 | I | $251,073,642 | 6.793 | Lost | $32,371,720 | 0.145 |
| $c_{4,0}$ | R | $172,237,365 | 1.510 | N | $36,963,116 | 0.147 | Won | $223,214,224 | 6.895 |
| $c_{4,1}$ | D | $63,607,849 | 12.558 | I | $44,441,701 | 1.789 | Lost | $20,140,999 | 0.474 |
| $c_{4,1}$ | R | $5,065,188 | 0.080 | N | $24,836,239 | 0.559 | Won | $42,513,729 | 2.111 |
| $c_{4,2}$ | D | $4,176,831 | 0.113 | I | $27,851,029 | 2.070 | Lost | $15,256,262 | 0.664 |
| $c_{4,2}$ | R | $36,949,622 | 8.846 | N | $13,455,223 | 0.483 | Won | $22,970,436 | 1.506 |
| $c_{4,3}$ | D | $27,507,787 | 1.649 | I | $35,066,128 | 3.742 | Lost | $7,682,895 | 0.235 |
| $c_{4,3}$ | R | $16,684,778 | 0.607 | N | $9,370,861 | 0.267 | Won | $32,718,695 | 4.259 |

Table A.25: Donor History by Community in MS

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $12,553,477 | 0.821 | I | $24,432,422 | 6.927 | Lost | $3,057,484 | 0.140 |
| $c_{2,0}$ | R | $15,286,469 | 1.218 | N | $3,527,250 | 0.144 | Won | $21,806,915 | 7.132 |
| $c_{2,1}$ | D | $7,495,422 | 0.382 | I | $22,650,205 | 4.017 | Lost | $4,440,037 | 0.206 |
| $c_{2,1}$ | R | $19,636,505 | 2.620 | N | $5,638,605 | 0.249 | Won | $21,516,616 | 4.846 |
| $c_{2,0} \cap c_{2,1}$ | D | $0 | NA | I | $0 | NA | Lost | $0 | NA |
| $c_{2,0} \cap c_{2,1}$ | R | $0 | NA | N | $0 | NA | Won | $0 | NA |
| $c_{4,0}$ | D | $12,553,477 | 0.821 | I | $24,432,422 | 6.927 | Lost | $3,057,484 | 0.140 |
| $c_{4,0}$ | R | $15,286,469 | 1.218 | N | $3,527,250 | 0.144 | Won | $21,806,915 | 7.132 |
| $c_{4,1}$ | D | $5,424,663 | 0.488 | I | $13,679,598 | 3.854 | Lost | $2,692,055 | 0.202 |
| $c_{4,1}$ | R | $11,112,325 | 2.048 | N | $3,549,340 | 0.259 | Won | $13,320,053 | 4.948 |
| $c_{4,2}$ | D | $300,835 | 0.322 | I | $989,467 | 2.881 | Lost | $290,361 | 0.323 |
| $c_{4,2}$ | R | $934,880 | 3.108 | N | $343,447 | 0.347 | Won | $900,153 | 3.100 |
| $c_{4,3}$ | D | $2,316,380 | 0.262 | I | $9,237,786 | 3.681 | Lost | $1,930,905 | 0.219 |
| $c_{4,3}$ | R | $8,828,986 | 3.812 | N | $2,509,313 | 0.272 | Won | $8,816,767 | 4.566 |

Table A.26: Donor History by Community in MT

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $85,270,697 | 2.917 | I | $82,268,986 | 2.485 | Lost | $27,627,618 | 0.366 |
| $c_{2,0}$ | R | $29,232,553 | 0.343 | N | $33,106,412 | 0.402 | Won | $75,568,663 | 2.735 |
| $c_{2,1}$ | D | $32,914,998 | 0.511 | I | $81,877,003 | 4.899 | Lost | $13,743,141 | 0.184 |
| $c_{2,1}$ | R | $64,383,485 | 1.956 | N | $16,714,123 | 0.204 | Won | $74,738,882 | 5.438 |
| $c_{2,0} \cap c_{2,1}$ | D | $8,059,669 | 0.710 | I | $17,423,185 | 7.961 | Lost | $2,088,981 | 0.136 |
| $c_{2,0} \cap c_{2,1}$ | R | $11,358,090 | 1.409 | N | $2,188,526 | 0.126 | Won | $15,411,980 | 7.378 |
| $c_{4,0}$ | D | $16,563,784 | 0.508 | I | $40,914,900 | 4.659 | Lost | $6,876,592 | 0.180 |
| $c_{4,0}$ | R | $32,611,980 | 1.969 | N | $8,782,122 | 0.215 | Won | $38,235,845 | 5.560 |
| $c_{4,1}$ | D | $71,767,032 | 2.330 | I | $76,742,439 | 2.887 | Lost | $22,042,678 | 0.315 |
| $c_{4,1}$ | R | $30,795,432 | 0.429 | N | $26,586,025 | 0.346 | Won | $70,081,251 | 3.179 |
| $c_{4,2}$ | D | $10,385,248 | 0.427 | I | $28,686,701 | 4.281 | Lost | $5,548,617 | 0.217 |
| $c_{4,2}$ | R | $24,332,406 | 2.343 | N | $6,701,333 | 0.234 | Won | $25,584,244 | 4.611 |
| $c_{4,3}$ | D | $15,252,791 | 20.000 | I | $9,215,290 | 1.333 | Lost | $5,987,202 | 0.671 |
| $c_{4,3}$ | R | $762,626 | 0.050 | N | $6,913,226 | 0.750 | Won | $8,916,906 | 1.489 |

Table A.27: Donor History by Community in NC

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $51,769,921 | 12.905 | I | $36,272,515 | 1.815 | Lost | $18,095,752 | 0.552 |
| $c_{2,0}$ | R | $4,011,545 | 0.077 | N | $19,981,527 | 0.551 | Won | $32,782,543 | 1.812 |
| $c_{2,1}$ | D | $107,467,660 | 0.622 | I | $246,695,555 | 7.009 | Lost | $32,123,285 | 0.146 |
| $c_{2,1}$ | R | $172,813,218 | 1.608 | N | $35,198,388 | 0.143 | Won | $219,611,778 | 6.837 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,972,874 | 1.750 | I | $2,593,979 | 4.687 | Lost | $554,200 | 0.248 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,127,615 | 0.572 | N | $553,430 | 0.213 | Won | $2,238,359 | 4.039 |
| $c_{4,0}$ | D | $1,603,314 | 0.131 | I | $8,265,790 | 1.451 | Lost | $3,073,534 | 0.302 |
| $c_{4,0}$ | R | $12,249,778 | 7.640 | N | $5,697,272 | 0.689 | Won | $10,185,464 | 3.314 |
| $c_{4,1}$ | D | $6,356,762 | 0.942 | I | $11,528,916 | 6.997 | Lost | $1,630,034 | 0.163 |
| $c_{4,1}$ | R | $6,747,319 | 1.061 | N | $1,647,805 | 0.143 | Won | $10,003,824 | 6.137 |
| $c_{4,2}$ | D | $49,009,025 | 20.029 | I | $32,850,437 | 1.728 | Lost | $17,351,930 | 0.582 |
| $c_{4,2}$ | R | $2,446,862 | 0.050 | N | $19,005,631 | 0.579 | Won | $29,800,516 | 1.717 |
| $c_{4,3}$ | D | $102,253,371 | 0.644 | I | $233,704,860 | 8.098 | Lost | $28,464,259 | 0.139 |
| $c_{4,3}$ | R | $158,853,551 | 1.554 | N | $28,860,146 | 0.123 | Won | $205,274,613 | 7.212 |

Table A.28: Donor History by Community in ND

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $67,741,718 | 65.329 | I | $32,838,603 | 0.908 | Lost | $28,316,314 | 0.816 |
| $c_{2,0}$ | R | $1,036,936 | 0.015 | N | $36,157,633 | 1.101 | Won | $34,722,282 | 1.226 |
| $c_{2,1}$ | D | $46,993,949 | 0.328 | I | $151,133,436 | 3.762 | Lost | $37,143,077 | 0.276 |
| $c_{2,1}$ | R | $143,484,392 | 3.053 | N | $40,174,929 | 0.266 | Won | $134,820,570 | 3.630 |
| $c_{2,0} \cap c_{2,1}$ | D | $2,325 | 3.100 | I | $1,800 | 1.412 | Lost | $1,025 | 0.526 |
| $c_{2,0} \cap c_{2,1}$ | R | $750 | 0.323 | N | $1,275 | 0.708 | Won | $1,950 | 1.902 |
| $c_{4,0}$ | D | $40,558,246 | 0.311 | I | $135,816,736 | 3.795 | Lost | $33,895,241 | 0.281 |
| $c_{4,0}$ | R | $130,410,642 | 3.215 | N | $35,786,005 | 0.263 | Won | $120,545,939 | 3.556 |
| $c_{4,1}$ | D | $9,218,937 | 0.524 | I | $21,747,148 | 4.061 | Lost | $4,042,770 | 0.202 |
| $c_{4,1}$ | R | $17,610,172 | 1.910 | N | $5,355,352 | 0.246 | Won | $20,001,948 | 4.948 |
| $c_{4,2}$ | D | $3,602,621 | 58.037 | I | $1,438,226 | 0.640 | Lost | $2,345,587 | 1.852 |
| $c_{4,2}$ | R | $62,075 | 0.017 | N | $2,245,780 | 1.561 | Won | $1,266,619 | 0.540 |
| $c_{4,3}$ | D | $66,897,686 | 68.381 | I | $32,354,782 | 0.906 | Lost | $27,744,666 | 0.807 |
| $c_{4,3}$ | R | $978,311 | 0.015 | N | $35,719,697 | 1.104 | Won | $34,382,823 | 1.239 |

Table A.29: Donor History by Community in NE

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $26,330,817 | 14.188 | I | $18,794,817 | 1.627 | Lost | $9,411,263 | 0.526 |
| $c_{2,0}$ | R | $1,855,913 | 0.070 | N | $11,550,208 | 0.615 | Won | $17,891,986 | 1.901 |
| $c_{2,1}$ | D | $63,096,062 | 0.680 | I | $139,592,832 | 6.529 | Lost | $18,004,296 | 0.143 |
| $c_{2,1}$ | R | $92,768,841 | 1.470 | N | $21,379,969 | 0.153 | Won | $126,031,089 | 7.000 |
| $c_{2,0} \cap c_{2,1}$ | D | $344,850 | 0.730 | I | $790,565 | 4.411 | Lost | $99,040 | 0.126 |
| $c_{2,0} \cap c_{2,1}$ | R | $472,215 | 1.369 | N | $179,225 | 0.227 | Won | $788,575 | 7.962 |
| $c_{4,0}$ | D | $2,874,096 | 0.629 | I | $7,330,511 | 4.406 | Lost | $1,165,712 | 0.165 |
| $c_{4,0}$ | R | $4,566,621 | 1.589 | N | $1,663,759 | 0.227 | Won | $7,050,042 | 6.048 |
| $c_{4,1}$ | D | $25,985,667 | 20.045 | I | $17,634,429 | 1.619 | Lost | $9,048,131 | 0.547 |
| $c_{4,1}$ | R | $1,296,368 | 0.050 | N | $10,890,683 | 0.618 | Won | $16,531,804 | 1.827 |
| $c_{4,2}$ | D | $31,500,738 | 0.671 | I | $70,446,954 | 6.232 | Lost | $9,141,692 | 0.145 |
| $c_{4,2}$ | R | $46,952,248 | 1.491 | N | $11,304,913 | 0.160 | Won | $63,001,131 | 6.892 |
| $c_{4,3}$ | D | $33,261,054 | 0.661 | I | $74,273,861 | 6.725 | Lost | $9,802,418 | 0.146 |
| $c_{4,3}$ | R | $50,351,730 | 1.514 | N | $11,045,013 | 0.149 | Won | $66,997,585 | 6.835 |

Table A.30: Donor History by Community in NH

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $29,912,728 | 9.445 | I | $21,109,643 | 1.717 | Lost | $10,257,385 | 0.513 |
| $c_{2,0}$ | R | $3,166,914 | 0.106 | N | $12,292,017 | 0.582 | Won | $19,992,009 | 1.949 |
| $c_{2,1}$ | D | $61,923,295 | 0.727 | I | $130,284,428 | 7.389 | Lost | $16,556,695 | 0.143 |
| $c_{2,1}$ | R | $85,199,104 | 1.376 | N | $17,632,756 | 0.135 | Won | $115,925,828 | 7.002 |
| $c_{2,0} \cap c_{2,1}$ | D | $1,486,695 | 0.810 | I | $3,048,387 | 10.724 | Lost | $255,404 | 0.099 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,836,409 | 1.235 | N | $284,267 | 0.093 | Won | $2,569,111 | 10.059 |
| $c_{4,0}$ | D | $28,987,020 | 14.651 | I | $19,162,151 | 1.581 | Lost | $10,098,497 | 0.551 |
| $c_{4,0}$ | R | $1,978,491 | 0.068 | N | $12,123,979 | 0.633 | Won | $18,342,061 | 1.816 |
| $c_{4,1}$ | D | $49,496,324 | 0.779 | I | $100,926,093 | 7.947 | Lost | $12,112,361 | 0.136 |
| $c_{4,1}$ | R | $63,508,805 | 1.283 | N | $12,699,366 | 0.126 | Won | $89,198,366 | 7.364 |
| $c_{4,2}$ | D | $25,442,031 | 0.665 | I | $55,941,987 | 6.913 | Lost | $7,757,110 | 0.155 |
| $c_{4,2}$ | R | $38,234,621 | 1.503 | N | $8,091,872 | 0.145 | Won | $50,205,666 | 6.472 |
| $c_{4,3}$ | D | $116,190 | 0.327 | I | $293,220 | 1.643 | Lost | $170,281 | 0.705 |
| $c_{4,3}$ | R | $355,092 | 3.056 | N | $178,462 | 0.609 | Won | $241,601 | 1.419 |

Table A.31: Donor History by Community in NJ

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $7,038,910 | 0.682 | I | $15,349,793 | 7.151 | Lost | $1,945,478 | 0.140 |
| $c_{2,0}$ | R | $10,326,637 | 1.467 | N | $2,146,554 | 0.140 | Won | $13,896,308 | 7.143 |
| $c_{2,1}$ | D | $160,298,141 | 1.559 | I | $222,545,179 | 5.311 | Lost | $36,232,730 | 0.177 |
| $c_{2,1}$ | R | $102,828,154 | 0.641 | N | $41,899,549 | 0.188 | Won | $204,438,130 | 5.642 |
| $c_{2,0} \cap c_{2,1}$ | D | $637,110 | 1.312 | I | $1,011,465 | 9.036 | Lost | $73,675 | 0.071 |
| $c_{2,0} \cap c_{2,1}$ | R | $485,540 | 0.762 | N | $111,935 | 0.111 | Won | $1,038,675 | 14.098 |
| $c_{4,0}$ | D | $110,242,012 | 1.496 | I | $155,268,232 | 5.208 | Lost | $25,896,936 | 0.181 |
| $c_{4,0}$ | R | $73,708,954 | 0.669 | N | $29,811,516 | 0.192 | Won | $143,280,666 | 5.533 |
| $c_{4,1}$ | D | $63,062,832 | 4.059 | I | $61,931,586 | 3.613 | Lost | $14,211,834 | 0.250 |
| $c_{4,1}$ | R | $15,536,286 | 0.246 | N | $17,139,934 | 0.277 | Won | $56,900,139 | 4.004 |
| $c_{4,2}$ | D | $6,624,520 | 0.667 | I | $14,643,463 | 7.147 | Lost | $1,886,103 | 0.143 |
| $c_{4,2}$ | R | $9,936,912 | 1.500 | N | $2,048,769 | 0.140 | Won | $13,158,118 | 6.976 |
| $c_{4,3}$ | D | $22,503,897 | 1.124 | I | $37,708,190 | 7.628 | Lost | $4,302,627 | 0.123 |
| $c_{4,3}$ | R | $20,022,229 | 0.890 | N | $4,943,391 | 0.131 | Won | $34,947,257 | 8.122 |

Table A.32: Donor History by Community in NM

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $102,633,713 | 0.676 | I | $223,196,003 | 6.711 | Lost | $29,535,288 | 0.148 |
| $c_{2,0}$ | R | $151,919,890 | 1.480 | N | $33,257,297 | 0.149 | Won | $199,529,157 | 6.756 |
| $c_{2,1}$ | D | $60,320,143 | 10.020 | I | $41,326,640 | 1.622 | Lost | $20,986,852 | 0.528 |
| $c_{2,1}$ | R | $6,020,229 | 0.100 | N | $25,473,528 | 0.616 | Won | $39,736,201 | 1.893 |
| $c_{2,0} \cap c_{2,1}$ | D | $3,636,419 | 1.317 | I | $5,470,412 | 5.774 | Lost | $848,648 | 0.175 |
| $c_{2,0} \cap c_{2,1}$ | R | $2,760,718 | 0.759 | N | $947,380 | 0.173 | Won | $4,857,176 | 5.723 |
| $c_{4,0}$ | D | $3,742,370 | 0.249 | I | $13,453,945 | 2.422 | Lost | $4,038,552 | 0.298 |
| $c_{4,0}$ | R | $15,032,503 | 4.017 | N | $5,555,803 | 0.413 | Won | $13,572,454 | 3.361 |
| $c_{4,1}$ | D | $17,744,567 | 0.949 | I | $31,979,631 | 6.899 | Lost | $4,498,730 | 0.164 |
| $c_{4,1}$ | R | $18,706,789 | 1.054 | N | $4,635,449 | 0.145 | Won | $27,495,091 | 6.112 |
| $c_{4,2}$ | D | $56,557,048 | 17.700 | I | $35,711,940 | 1.459 | Lost | $20,108,303 | 0.579 |
| $c_{4,2}$ | R | $3,195,317 | 0.056 | N | $24,479,566 | 0.685 | Won | $34,729,404 | 1.727 |
| $c_{4,3}$ | D | $93,608,267 | 0.705 | I | $201,664,035 | 7.625 | Lost | $24,300,448 | 0.136 |
| $c_{4,3}$ | R | $132,864,439 | 1.419 | N | $26,449,446 | 0.131 | Won | $178,761,471 | 7.356 |

Table A.33: Donor History by Community in NV

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $158,965,056 | 1.300 | I | $225,025,227 | 3.775 | Lost | $47,543,067 | 0.228 |
| $c_{2,0}$ | R | $122,306,458 | 0.769 | N | $59,605,989 | 0.265 | Won | $208,251,253 | 4.380 |
| $c_{2,1}$ | D | $8,731,174 | 0.266 | I | $31,046,607 | 2.883 | Lost | $11,613,868 | 0.442 |
| $c_{2,1}$ | R | $32,825,328 | 3.760 | N | $10,770,372 | 0.347 | Won | $26,273,418 | 2.262 |
| $c_{2,0} \cap c_{2,1}$ | D | $4,437,348 | 0.679 | I | $9,190,367 | 4.925 | Lost | $1,357,219 | 0.159 |
| $c_{2,0} \cap c_{2,1}$ | R | $6,533,880 | 1.472 | N | $1,865,939 | 0.203 | Won | $8,525,645 | 6.282 |
| $c_{4,0}$ | D | $35,636,945 | 0.722 | I | $72,771,322 | 5.394 | Lost | $10,735,308 | 0.163 |
| $c_{4,0}$ | R | $49,352,688 | 1.385 | N | $13,491,350 | 0.185 | Won | $65,906,928 | 6.139 |
| $c_{4,1}$ | D | $78,654,444 | 0.855 | I | $146,693,913 | 5.742 | Lost | $21,356,570 | 0.162 |
| $c_{4,1}$ | R | $91,942,866 | 1.169 | N | $25,545,624 | 0.174 | Won | $131,895,928 | 6.176 |
| $c_{4,2}$ | D | $70,498,664 | 9.667 | I | $48,663,872 | 1.615 | Lost | $23,344,450 | 0.472 |
| $c_{4,2}$ | R | $7,292,713 | 0.103 | N | $30,126,580 | 0.619 | Won | $49,420,421 | 2.117 |
| $c_{4,3}$ | D | $6,339,268 | 0.215 | I | $26,196,703 | 2.667 | Lost | $10,829,362 | 0.495 |
| $c_{4,3}$ | R | $29,448,192 | 4.645 | N | $9,821,335 | 0.375 | Won | $21,891,598 | 2.022 |

Table A.34: Donor History by Community in NY

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $483,225 | 0.293 | I | $1,194,102 | 1.216 | Lost | $1,051,995 | 0.948 |
| $c_{2,0}$ | R | $1,650,714 | 3.416 | N | $982,367 | 0.823 | Won | $1,109,524 | 1.055 |
| $c_{2,1}$ | D | $295,219,012 | 1.195 | I | $445,189,439 | 4.399 | Lost | $91,539,282 | 0.222 |
| $c_{2,1}$ | R | $247,092,123 | 0.837 | N | $101,193,213 | 0.227 | Won | $412,876,484 | 4.510 |
| $c_{2,0} \cap c_{2,1}$ | D | $181,400 | 0.308 | I | $540,375 | 2.283 | Lost | $387,800 | 0.997 |
| $c_{2,0} \cap c_{2,1}$ | R | $588,068 | 3.242 | N | $236,743 | 0.438 | Won | $388,918 | 1.003 |
| $c_{4,0}$ | D | $197,075,437 | 1.812 | I | $245,638,942 | 3.920 | Lost | $51,670,230 | 0.223 |
| $c_{4,0}$ | R | $108,785,405 | 0.552 | N | $62,667,078 | 0.255 | Won | $231,800,994 | 4.486 |
| $c_{4,1}$ | D | $6,879,527 | 0.217 | I | $23,879,340 | 1.591 | Lost | $16,921,310 | 0.796 |
| $c_{4,1}$ | R | $31,718,127 | 4.611 | N | $15,009,776 | 0.629 | Won | $21,271,009 | 1.257 |
| $c_{4,2}$ | D | $161,430,736 | 0.875 | I | $307,367,202 | 7.514 | Lost | $39,081,879 | 0.140 |
| $c_{4,2}$ | R | $184,540,350 | 1.143 | N | $40,907,173 | 0.133 | Won | $278,926,754 | 7.137 |
| $c_{4,3}$ | D | $398,839 | 0.119 | I | $1,554,493 | 0.695 | Lost | $2,339,541 | 1.631 |
| $c_{4,3}$ | R | $3,338,540 | 8.371 | N | $2,236,296 | 1.439 | Won | $1,434,748 | 0.613 |

Table A.35: Donor History by Community in OH

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $117,544,949 | 10.620 | I | $85,241,256 | 1.856 | Lost | $41,259,359 | 0.549 |
| $c_{2,0}$ | R | $11,068,296 | 0.094 | N | $45,934,000 | 0.539 | Won | $75,100,745 | 1.820 |
| $c_{2,1}$ | D | $135,438,609 | 0.601 | I | $317,869,465 | 6.508 | Lost | $46,264,730 | 0.169 |
| $c_{2,1}$ | R | $225,418,505 | 1.664 | N | $48,845,417 | 0.154 | Won | $274,120,954 | 5.925 |
| $c_{2,0} \cap c_{2,1}$ | D | $5,140,967 | 1.158 | I | $7,872,409 | 3.995 | Lost | $1,739,284 | 0.239 |
| $c_{2,0} \cap c_{2,1}$ | R | $4,439,420 | 0.864 | N | $1,970,528 | 0.250 | Won | $7,286,157 | 4.189 |
| $c_{4,0}$ | D | $62,083,311 | 0.689 | I | $134,711,997 | 6.731 | Lost | $18,746,931 | 0.161 |
| $c_{4,0}$ | R | $90,083,971 | 1.451 | N | $20,013,368 | 0.149 | Won | $116,790,889 | 6.230 |
| $c_{4,1}$ | D | $81,148,254 | 18.718 | I | $56,033,545 | 1.793 | Lost | $28,838,212 | 0.603 |
| $c_{4,1}$ | R | $4,335,263 | 0.053 | N | $31,253,367 | 0.558 | Won | $47,808,194 | 1.658 |
| $c_{4,2}$ | D | $98,660,519 | 0.533 | I | $251,488,703 | 6.879 | Lost | $35,089,901 | 0.162 |
| $c_{4,2}$ | R | $185,253,218 | 1.878 | N | $36,559,534 | 0.145 | Won | $216,065,284 | 6.157 |
| $c_{4,3}$ | D | $74,880,366 | 8.445 | I | $56,713,304 | 1.973 | Lost | $24,281,000 | 0.463 |
| $c_{4,3}$ | R | $8,866,475 | 0.118 | N | $28,744,175 | 0.507 | Won | $52,474,576 | 2.161 |

Table A.36: Donor History by Community in OK

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $800 | 0.001 | I | $267,691 | 0.459 | Lost | $250,000 | 0.437 |
| $c_{2,0}$ | R | $850,191 | 1.062E3 | N | $583,300 | 2.179 | Won | $571,991 | 2.288 |
| $c_{2,1}$ | D | $104,307,442 | 0.712 | I | $211,684,987 | 5.190 | Lost | $33,717,577 | 0.176 |
| $c_{2,1}$ | R | $146,470,889 | 1.404 | N | $40,785,128 | 0.193 | Won | $191,611,024 | 5.683 |
| $c_{2,0} \cap c_{2,1}$ | D | $0 | NA | I | $98,841 | 0.371 | Lost | $98,250 | 0.369 |
| $c_{2,0} \cap c_{2,1}$ | R | $365,591 | NA | N | $266,750 | 2.699 | Won | $266,341 | 2.711 |
| $c_{4,0}$ | D | $35,995,280 | 0.492 | I | $94,035,012 | 6.020 | Lost | $12,478,812 | 0.147 |
| $c_{4,0}$ | R | $73,169,015 | 2.033 | N | $15,621,197 | 0.166 | Won | $84,807,010 | 6.796 |
| $c_{4,1}$ | D | $20,956,162 | 13.744 | I | $14,832,050 | 1.869 | Lost | $6,681,686 | 0.489 |
| $c_{4,1}$ | R | $1,524,754 | 0.073 | N | $7,935,051 | 0.535 | Won | $13,661,094 | 2.045 |
| $c_{4,2}$ | D | $54,812,673 | 0.582 | I | $129,432,539 | 6.276 | Lost | $16,916,874 | 0.144 |
| $c_{4,2}$ | R | $94,162,794 | 1.718 | N | $20,622,121 | 0.159 | Won | $117,665,346 | 6.956 |
| $c_{4,3}$ | D | $8,211,271 | 1.131 | I | $12,522,610 | 4.190 | Lost | $2,795,397 | 0.245 |
| $c_{4,3}$ | R | $7,262,906 | 0.885 | N | $2,988,847 | 0.239 | Won | $11,403,837 | 4.080 |

Table A.37: Donor History by Community in OR

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $31,650 | 63.300 | I | $22,000 | 0.330 | Lost | $17,150 | 0.242 |
| $c_{2,0}$ | R | $500 | 0.016 | N | $66,580 | 3.026 | Won | $70,930 | 4.136 |
| $c_{2,1}$ | D | $171,425,635 | 0.875 | I | $297,890,883 | 4.053 | Lost | $62,079,180 | 0.229 |
| $c_{2,1}$ | R | $195,912,111 | 1.143 | N | $73,498,025 | 0.247 | Won | $271,464,553 | 4.373 |
| $c_{2,0} \cap c_{2,1}$ | D | $7,500 | NA | I | $3,950 | 0.244 | Lost | $3,500 | 0.210 |
| $c_{2,0} \cap c_{2,1}$ | R | $0 | NA | N | $16,200 | 4.101 | Won | $16,650 | 4.757 |
| $c_{4,0}$ | D | $1,626,070 | 0.061 | I | $16,962,663 | 1.442 | Lost | $13,726,569 | 1.020 |
| $c_{4,0}$ | R | $26,640,043 | 16.383 | N | $11,762,591 | 0.693 | Won | $13,463,633 | 0.981 |
| $c_{4,1}$ | D | $81,662,944 | 0.638 | I | $180,814,833 | 5.894 | Lost | $26,545,625 | 0.162 |
| $c_{4,1}$ | R | $127,957,378 | 1.567 | N | $30,675,229 | 0.170 | Won | $163,608,029 | 6.163 |
| $c_{4,2}$ | D | $44,175,965 | 9.538 | I | $31,865,471 | 1.747 | Lost | $12,912,275 | 0.399 |
| $c_{4,2}$ | R | $4,631,724 | 0.105 | N | $18,241,983 | 0.572 | Won | $32,346,356 | 2.505 |
| $c_{4,3}$ | D | $70,617,449 | 1.024 | I | $118,765,705 | 5.486 | Lost | $16,302,193 | 0.152 |
| $c_{4,3}$ | R | $68,950,251 | 0.976 | N | $21,650,840 | 0.182 | Won | $107,565,246 | 6.598 |

Table A.38: Donor History by Community in PA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $149,638,238 | 0.610 | I | $340,136,508 | 5.966 | Lost | $56,734,601 | 0.191 |
| $c_{2,0}$ | R | $245,213,271 | 1.639 | N | $57,010,454 | 0.168 | Won | $297,486,327 | 5.243 |
| $c_{2,1}$ | D | $151,624,395 | 4.616 | I | $118,112,859 | 1.754 | Lost | $53,605,945 | 0.463 |
| $c_{2,1}$ | R | $32,849,300 | 0.217 | N | $67,330,689 | 0.570 | Won | $115,780,389 | 2.160 |
| $c_{2,0} \cap c_{2,1}$ | D | $16,885,302 | 0.919 | I | $27,376,235 | 3.408 | Lost | $7,477,733 | 0.302 |
| $c_{2,0} \cap c_{2,1}$ | R | $18,374,434 | 1.088 | N | $8,033,151 | 0.293 | Won | $24,780,565 | 3.314 |
| $c_{4,0}$ | D | $60,800,441 | 12.011 | I | $37,685,149 | 1.318 | Lost | $23,095,177 | 0.608 |
| $c_{4,0}$ | R | $5,062,144 | 0.083 | N | $28,598,074 | 0.759 | Won | $37,992,954 | 1.645 |
| $c_{4,1}$ | D | $117,777,822 | 6.056 | I | $89,002,626 | 1.816 | Lost | $38,969,172 | 0.452 |
| $c_{4,1}$ | R | $19,447,071 | 0.165 | N | $49,018,365 | 0.551 | Won | $86,240,063 | 2.213 |
| $c_{4,2}$ | D | $137,081,746 | 0.671 | I | $303,275,194 | 7.575 | Lost | $42,692,901 | 0.163 |
| $c_{4,2}$ | R | $204,334,834 | 1.491 | N | $40,037,013 | 0.132 | Won | $262,332,319 | 6.145 |
| $c_{4,3}$ | D | $34,699,562 | 0.396 | I | $98,945,070 | 4.062 | Lost | $23,663,560 | 0.273 |
| $c_{4,3}$ | R | $87,730,404 | 2.528 | N | $24,356,882 | 0.246 | Won | $86,731,143 | 3.665 |

Table A.39: Donor History by Community in RI

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $1,595,237 | 0.505 | I | $3,642,215 | 3.117 | Lost | $987,252 | 0.301 |
| $c_{2,0}$ | R | $3,160,874 | 1.981 | N | $1,168,620 | 0.321 | Won | $3,280,219 | 3.323 |
| $c_{2,1}$ | D | $79,690,867 | 3.353 | I | $76,480,417 | 2.763 | Lost | $22,789,085 | 0.321 |
| $c_{2,1}$ | R | $23,764,090 | 0.298 | N | $27,675,543 | 0.362 | Won | $71,002,040 | 3.116 |
| $c_{2,0} \cap c_{2,1}$ | D | $92,800 | 4.852 | I | $101,600 | 8.311 | Lost | $17,250 | 0.179 |
| $c_{2,0} \cap c_{2,1}$ | R | $19,125 | 0.206 | N | $12,225 | 0.120 | Won | $96,325 | 5.584 |
| $c_{4,0}$ | D | $18,364,498 | 38.962 | I | $10,853,806 | 1.340 | Lost | $6,070,350 | 0.509 |
| $c_{4,0}$ | R | $471,340 | 0.026 | N | $8,099,932 | 0.746 | Won | $11,914,870 | 1.963 |
| $c_{4,1}$ | D | $1,502,387 | 0.481 | I | $3,535,115 | 3.091 | Lost | $959,947 | 0.302 |
| $c_{4,1}$ | R | $3,123,930 | 2.079 | N | $1,143,526 | 0.323 | Won | $3,175,580 | 3.308 |
| $c_{4,2}$ | D | $37,856,156 | 17.133 | I | $27,970,066 | 2.234 | Lost | $10,623,583 | 0.420 |
| $c_{4,2}$ | R | $2,209,545 | 0.058 | N | $12,521,364 | 0.448 | Won | $25,305,435 | 2.382 |
| $c_{4,3}$ | D | $24,867,550 | 1.052 | I | $40,757,435 | 5.132 | Lost | $6,875,706 | 0.187 |
| $c_{4,3}$ | R | $23,638,710 | 0.951 | N | $7,941,547 | 0.195 | Won | $36,732,834 | 5.342 |

Table A.40: Donor History by Community in SC

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $101,535,588 | 0.642 | I | $230,025,796 | 7.376 | Lost | $29,067,526 | 0.144 |
| $c_{2,0}$ | R | $158,048,921 | 1.557 | N | $31,184,381 | 0.136 | Won | $201,761,931 | 6.941 |
| $c_{2,1}$ | D | $2,990,430 | 1.584 | I | $3,398,973 | 2.222 | Lost | $1,407,296 | 0.453 |
| $c_{2,1}$ | R | $1,888,135 | 0.631 | N | $1,529,707 | 0.450 | Won | $3,108,583 | 2.209 |
| $c_{2,0} \cap c_{2,1}$ | D | $946,584 | 0.869 | I | $1,631,769 | 3.863 | Lost | $365,414 | 0.239 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,089,802 | 1.151 | N | $422,364 | 0.259 | Won | $1,532,051 | 4.193 |
| $c_{4,0}$ | D | $1,790,822 | 0.403 | I | $4,971,590 | 3.639 | Lost | $1,203,365 | 0.278 |
| $c_{4,0}$ | R | $4,439,218 | 2.479 | N | $1,366,283 | 0.275 | Won | $4,325,637 | 3.595 |
| $c_{4,1}$ | D | $51,399,851 | 0.700 | I | $110,850,903 | 7.618 | Lost | $13,669,505 | 0.142 |
| $c_{4,1}$ | R | $73,378,076 | 1.428 | N | $14,552,110 | 0.131 | Won | $96,366,491 | 7.050 |
| $c_{4,2}$ | D | $90,970,148 | 0.665 | I | $203,374,886 | 7.885 | Lost | $24,034,101 | 0.135 |
| $c_{4,2}$ | R | $136,832,501 | 1.504 | N | $25,793,718 | 0.127 | Won | $178,593,651 | 7.431 |
| $c_{4,3}$ | D | $2,104,080 | 2.547 | I | $1,857,297 | 1.677 | Lost | $1,053,363 | 0.635 |
| $c_{4,3}$ | R | $825,940 | 0.393 | N | $1,107,638 | 0.596 | Won | $1,659,228 | 1.575 |

Table A.41: Donor History by Community in SD

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $34,099,617 | 39.554 | I | $19,704,685 | 1.277 | Lost | $11,952,173 | 0.589 |
| $c_{2,0}$ | R | $862,102 | 0.025 | N | $15,425,442 | 0.783 | Won | $20,293,985 | 1.698 |
| $c_{2,1}$ | D | $51,360,286 | 0.696 | I | $106,971,152 | 5.561 | Lost | $16,573,094 | 0.172 |
| $c_{2,1}$ | R | $73,846,165 | 1.438 | N | $19,235,228 | 0.180 | Won | $96,633,950 | 5.831 |
| $c_{2,0} \cap c_{2,1}$ | D | $7,786,855 | 19.521 | I | $5,758,133 | 2.326 | Lost | $1,890,340 | 0.352 |
| $c_{2,0} \cap c_{2,1}$ | R | $398,893 | 0.051 | N | $2,476,065 | 0.430 | Won | $5,373,558 | 2.843 |
| $c_{4,0}$ | D | $25,530,792 | 65.582 | I | $13,485,220 | 1.074 | Lost | $9,825,788 | 0.686 |
| $c_{4,0}$ | R | $389,299 | 0.015 | N | $12,550,329 | 0.931 | Won | $14,325,442 | 1.458 |
| $c_{4,1}$ | D | $20,223,936 | 0.696 | I | $43,465,188 | 7.093 | Lost | $5,516,411 | 0.144 |
| $c_{4,1}$ | R | $29,054,805 | 1.437 | N | $6,127,873 | 0.141 | Won | $38,201,627 | 6.925 |
| $c_{4,2}$ | D | $8,571,735 | 17.798 | I | $6,224,671 | 2.160 | Lost | $2,126,905 | 0.356 |
| $c_{4,2}$ | R | $481,623 | 0.056 | N | $2,881,688 | 0.463 | Won | $5,979,803 | 2.812 |
| $c_{4,3}$ | D | $30,395,055 | 0.535 | I | $75,096,740 | 5.812 | Lost | $11,373,534 | 0.167 |
| $c_{4,3}$ | R | $56,798,693 | 1.869 | N | $12,921,888 | 0.172 | Won | $68,033,381 | 5.982 |

Table A.42: Donor History by Community in TN

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $106,902,225 | 0.688 | I | $227,586,463 | 6.272 | Lost | $32,159,426 | 0.161 |
| $c_{2,0}$ | R | $155,287,273 | 1.453 | N | $36,284,462 | 0.159 | Won | $199,947,505 | 6.217 |
| $c_{2,1}$ | D | $35,724,535 | 14.939 | I | $25,236,325 | 1.891 | Lost | $10,932,374 | 0.478 |
| $c_{2,1}$ | R | $2,391,415 | 0.067 | N | $13,345,490 | 0.529 | Won | $22,887,583 | 2.094 |
| $c_{2,0} \cap c_{2,1}$ | D | $8,724,038 | 24.360 | I | $6,131,233 | 2.033 | Lost | $2,413,383 | 0.422 |
| $c_{2,0} \cap c_{2,1}$ | R | $358,128 | 0.041 | N | $3,016,383 | 0.492 | Won | $5,722,798 | 2.371 |
| $c_{4,0}$ | D | $37,356,171 | 13.590 | I | $26,628,068 | 1.897 | Lost | $11,779,887 | 0.493 |
| $c_{4,0}$ | R | $2,748,765 | 0.074 | N | $14,033,971 | 0.527 | Won | $23,870,404 | 2.026 |
| $c_{4,1}$ | D | $5,322,249 | 0.797 | I | $10,166,360 | 5.209 | Lost | $2,068,093 | 0.242 |
| $c_{4,1}$ | R | $6,674,749 | 1.254 | N | $1,951,838 | 0.192 | Won | $8,541,619 | 4.130 |
| $c_{4,2}$ | D | $70,924,395 | 0.687 | I | $154,269,350 | 7.426 | Lost | $19,373,408 | 0.143 |
| $c_{4,2}$ | R | $103,177,801 | 1.455 | N | $20,774,440 | 0.135 | Won | $135,368,117 | 6.987 |
| $c_{4,3}$ | D | $36,919,826 | 0.517 | I | $93,694,061 | 6.118 | Lost | $12,611,413 | 0.154 |
| $c_{4,3}$ | R | $71,420,834 | 1.934 | N | $15,313,970 | 0.163 | Won | $81,885,075 | 6.493 |

Table A.43: Donor History by Community in TX

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $280,373,449 | 0.743 | I | $545,315,209 | 4.704 | Lost | $105,731,641 | 0.216 |
| $c_{2,0}$ | R | $377,111,259 | 1.345 | N | $115,914,858 | 0.213 | Won | $488,680,578 | 4.622 |
| $c_{2,1}$ | D | $12,637,745 | 0.580 | I | $26,169,595 | 3.328 | Lost | $7,073,731 | 0.279 |
| $c_{2,1}$ | R | $21,783,521 | 1.724 | N | $7,863,896 | 0.300 | Won | $25,385,967 | 3.589 |
| $c_{2,0} \cap c_{2,1}$ | D | $5,130,534 | 0.840 | I | $7,942,375 | 2.382 | Lost | $2,851,784 | 0.368 |
| $c_{2,0} \cap c_{2,1}$ | R | $6,110,724 | 1.191 | N | $3,334,283 | 0.420 | Won | $7,742,291 | 2.715 |
| $c_{4,0}$ | D | $4,548,594 | 0.453 | I | $11,871,811 | 4.718 | Lost | $2,339,794 | 0.208 |
| $c_{4,0}$ | R | $10,033,408 | 2.206 | N | $2,516,030 | 0.212 | Won | $11,266,527 | 4.815 |
| $c_{4,1}$ | D | $18,570,940 | 0.795 | I | $27,393,183 | 1.888 | Lost | $11,541,097 | 0.410 |
| $c_{4,1}$ | R | $23,362,868 | 1.258 | N | $14,511,728 | 0.530 | Won | $28,148,295 | 2.439 |
| $c_{4,2}$ | D | $269,555,486 | 0.731 | I | $532,883,432 | 4.876 | Lost | $99,903,849 | 0.210 |
| $c_{4,2}$ | R | $368,886,621 | 1.368 | N | $109,277,253 | 0.205 | Won | $476,365,686 | 4.768 |
| $c_{4,3}$ | D | $2,390,848 | 1.116 | I | $3,575,972 | 4.544 | Lost | $960,038 | 0.280 |
| $c_{4,3}$ | R | $2,142,754 | 0.896 | N | $787,014 | 0.220 | Won | $3,430,889 | 3.574 |

Table A.44: Donor History by Community in UT

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $36,596,582 | 14.744 | I | $24,894,727 | 1.706 | Lost | $11,827,031 | 0.490 |
| $c_{2,0}$ | R | $2,482,069 | 0.068 | N | $14,594,387 | 0.586 | Won | $24,153,996 | 2.042 |
| $c_{2,1}$ | D | $101,526,047 | 0.757 | I | $200,620,736 | 5.459 | Lost | $31,072,428 | 0.171 |
| $c_{2,1}$ | R | $134,049,613 | 1.320 | N | $36,750,431 | 0.183 | Won | $181,435,202 | 5.839 |
| $c_{2,0} \cap c_{2,1}$ | D | $2,770,577 | 6.286 | I | $2,297,463 | 2.352 | Lost | $535,248 | 0.199 |
| $c_{2,0} \cap c_{2,1}$ | R | $440,750 | 0.159 | N | $977,014 | 0.425 | Won | $2,694,904 | 5.035 |
| $c_{4,0}$ | D | $8,860,553 | 10.682 | I | $6,227,325 | 1.771 | Lost | $2,763,967 | 0.440 |
| $c_{4,0}$ | R | $829,501 | 0.094 | N | $3,516,429 | 0.565 | Won | $6,285,389 | 2.274 |
| $c_{4,1}$ | D | $85,119,826 | 0.640 | I | $190,737,203 | 6.563 | Lost | $25,291,515 | 0.148 |
| $c_{4,1}$ | R | $133,032,197 | 1.563 | N | $29,064,128 | 0.152 | Won | $170,522,548 | 6.742 |
| $c_{4,2}$ | D | $32,825,723 | 18.551 | I | $22,084,263 | 1.719 | Lost | $10,664,777 | 0.511 |
| $c_{4,2}$ | R | $1,769,519 | 0.054 | N | $12,850,296 | 0.582 | Won | $20,878,920 | 1.958 |
| $c_{4,3}$ | D | $19,978,906 | 3.092 | I | $17,809,334 | 2.020 | Lost | $6,673,817 | 0.366 |
| $c_{4,3}$ | R | $6,460,650 | 0.323 | N | $8,816,251 | 0.495 | Won | $18,215,496 | 2.729 |

Table A.45: Donor History by Community in VA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $153,582,752 | 0.773 | I | $305,667,205 | 6.248 | Lost | $47,583,033 | 0.175 |
| $c_{2,0}$ | R | $198,651,556 | 1.293 | N | $48,925,890 | 0.160 | Won | $271,187,351 | 5.699 |
| $c_{2,1}$ | D | $7,499,091 | 1.384 | I | $9,494,653 | 2.684 | Lost | $2,375,821 | 0.232 |
| $c_{2,1}$ | R | $5,418,633 | 0.723 | N | $3,537,874 | 0.373 | Won | $10,260,231 | 4.319 |
| $c_{2,0} \cap c_{2,1}$ | D | $781,770 | 0.611 | I | $1,774,846 | 5.853 | Lost | $383,827 | 0.227 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,279,344 | 1.636 | N | $303,222 | 0.171 | Won | $1,690,741 | 4.405 |
| $c_{4,0}$ | D | $7,635,933 | 1.910 | I | $7,874,003 | 2.052 | Lost | $2,740,774 | 0.328 |
| $c_{4,0}$ | R | $3,998,608 | 0.524 | N | $3,836,437 | 0.487 | Won | $8,345,026 | 3.045 |
| $c_{4,1}$ | D | $451,314 | 0.198 | I | $2,126,087 | 3.384 | Lost | $771,523 | 0.400 |
| $c_{4,1}$ | R | $2,277,022 | 5.045 | N | $628,248 | 0.295 | Won | $1,928,712 | 2.500 |
| $c_{4,2}$ | D | $70,882,098 | 1.162 | I | $110,598,192 | 4.997 | Lost | $20,753,310 | 0.210 |
| $c_{4,2}$ | R | $61,002,142 | 0.861 | N | $22,134,898 | 0.200 | Won | $98,992,761 | 4.770 |
| $c_{4,3}$ | D | $100,283,458 | 0.621 | I | $232,900,107 | 7.615 | Lost | $30,721,930 | 0.150 |
| $c_{4,3}$ | R | $161,455,602 | 1.610 | N | $30,585,470 | 0.131 | Won | $205,159,405 | 6.678 |

Table A.46: Donor History by Community in VT

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $40,000,763 | 0.769 | I | $81,682,206 | 7.471 | Lost | $10,155,625 | 0.140 |
| $c_{2,0}$ | R | $52,003,291 | 1.300 | N | $10,933,935 | 0.134 | Won | $72,299,481 | 7.119 |
| $c_{2,1}$ | D | $81,251,612 | 2.161 | I | $92,960,719 | 3.460 | Lost | $22,074,319 | 0.264 |
| $c_{2,1}$ | R | $37,604,793 | 0.463 | N | $26,869,642 | 0.289 | Won | $83,510,295 | 3.783 |
| $c_{2,0} \cap c_{2,1}$ | D | $28,741,480 | 0.828 | I | $56,726,731 | 7.915 | Lost | $6,513,956 | 0.131 |
| $c_{2,0} \cap c_{2,1}$ | R | $34,717,843 | 1.208 | N | $7,166,812 | 0.126 | Won | $49,663,757 | 7.624 |
| $c_{4,0}$ | D | $32,886,802 | 0.806 | I | $65,715,396 | 7.778 | Lost | $7,755,765 | 0.133 |
| $c_{4,0}$ | R | $40,779,115 | 1.240 | N | $8,449,381 | 0.129 | Won | $58,421,969 | 7.533 |
| $c_{4,1}$ | D | $1,512,947 | 0.340 | I | $4,872,164 | 4.248 | Lost | $1,197,654 | 0.271 |
| $c_{4,1}$ | R | $4,446,421 | 2.939 | N | $1,147,003 | 0.235 | Won | $4,414,549 | 3.686 |
| $c_{4,2}$ | D | $35,262,465 | 20.273 | I | $23,901,052 | 1.788 | Lost | $10,306,157 | 0.454 |
| $c_{4,2}$ | R | $1,739,353 | 0.049 | N | $13,368,954 | 0.559 | Won | $22,703,604 | 2.203 |
| $c_{4,3}$ | D | $72,900,147 | 1.876 | I | $88,943,879 | 3.747 | Lost | $19,793,015 | 0.249 |
| $c_{4,3}$ | R | $38,867,047 | 0.533 | N | $23,734,834 | 0.267 | Won | $79,367,425 | 4.010 |

Table A.47: Donor History by Community in WA

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $144,263,691 | 0.722 | I | $297,147,724 | 5.963 | Lost | $46,476,765 | 0.175 |
| $c_{2,0}$ | R | $199,715,633 | 1.384 | N | $49,833,854 | 0.168 | Won | $265,424,990 | 5.711 |
| $c_{2,1}$ | D | $81,435,401 | 7.887 | I | $60,627,626 | 1.865 | Lost | $25,147,831 | 0.420 |
| $c_{2,1}$ | R | $10,324,868 | 0.127 | N | $32,502,538 | 0.536 | Won | $59,910,532 | 2.382 |
| $c_{2,0} \cap c_{2,1}$ | D | $11,505,965 | 2.137 | I | $13,803,726 | 4.125 | Lost | $3,577,234 | 0.298 |
| $c_{2,0} \cap c_{2,1}$ | R | $5,383,496 | 0.468 | N | $3,346,183 | 0.242 | Won | $11,999,964 | 3.355 |
| $c_{4,0}$ | D | $76,551,146 | 13.357 | I | $52,515,674 | 1.692 | Lost | $23,817,235 | 0.453 |
| $c_{4,0}$ | R | $5,731,021 | 0.075 | N | $31,035,493 | 0.591 | Won | $52,551,252 | 2.206 |
| $c_{4,1}$ | D | $38,934,515 | 0.896 | I | $71,965,131 | 6.472 | Lost | $9,451,076 | 0.145 |
| $c_{4,1}$ | R | $43,457,220 | 1.116 | N | $11,118,888 | 0.155 | Won | $65,308,001 | 6.910 |
| $c_{4,2}$ | D | $118,805,276 | 0.712 | I | $251,205,663 | 6.817 | Lost | $32,582,054 | 0.145 |
| $c_{4,2}$ | R | $166,887,680 | 1.405 | N | $36,847,403 | 0.147 | Won | $224,881,185 | 6.902 |
| $c_{4,3}$ | D | $13,724,910 | 0.372 | I | $40,370,174 | 3.755 | Lost | $11,792,766 | 0.331 |
| $c_{4,3}$ | R | $36,878,219 | 2.687 | N | $10,751,179 | 0.266 | Won | $35,666,660 | 3.024 |

Table A.48: Donor History by Community in WI

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $59,894,805 | 15.724 | I | $40,531,574 | 1.684 | Lost | $19,714,059 | 0.548 |
| $c_{2,0}$ | R | $3,809,096 | 0.064 | N | $24,075,667 | 0.594 | Won | $35,984,131 | 1.825 |
| $c_{2,1}$ | D | $57,326,408 | 0.583 | I | $135,824,919 | 6.405 | Lost | $18,743,600 | 0.156 |
| $c_{2,1}$ | R | $98,384,044 | 1.716 | N | $21,207,587 | 0.156 | Won | $120,056,802 | 6.405 |
| $c_{2,0} \cap c_{2,1}$ | D | $2,353,534 | 2.140 | I | $2,985,702 | 5.848 | Lost | $743,088 | 0.402 |
| $c_{2,0} \cap c_{2,1}$ | R | $1,100,000 | 0.467 | N | $510,567 | 0.171 | Won | $1,847,881 | 2.487 |
| $c_{4,0}$ | D | $54,979,584 | 0.565 | I | $132,842,093 | 6.417 | Lost | $18,004,242 | 0.152 |
| $c_{4,0}$ | R | $97,284,044 | 1.769 | N | $20,700,853 | 0.156 | Won | $118,211,901 | 6.566 |
| $c_{4,1}$ | D | $23,916,915 | 17.589 | I | $16,444,287 | 1.782 | Lost | $7,329,345 | 0.484 |
| $c_{4,1}$ | R | $1,359,746 | 0.057 | N | $9,226,758 | 0.561 | Won | $15,157,416 | 2.068 |
| $c_{4,2}$ | D | $27,747,143 | 11.419 | I | $20,296,272 | 1.980 | Lost | $8,612,065 | 0.519 |
| $c_{4,2}$ | R | $2,429,805 | 0.088 | N | $10,252,745 | 0.505 | Won | $16,583,280 | 1.926 |
| $c_{4,3}$ | D | $19,208,736 | 10.757 | I | $12,774,014 | 1.484 | Lost | $7,248,531 | 0.627 |
| $c_{4,3}$ | R | $1,785,735 | 0.093 | N | $8,606,952 | 0.674 | Won | $11,563,515 | 1.595 |

Table A.49: Donor History by Community in WV

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $58,907,989 | 0.646 | I | $133,547,525 | 7.518 | Lost | $17,088,670 | 0.146 |
| $c_{2,0}$ | R | $91,234,606 | 1.549 | N | $17,762,752 | 0.133 | Won | $116,737,601 | 6.831 |
| $c_{2,1}$ | D | $84,388,861 | 5.441 | I | $72,091,258 | 2.509 | Lost | $22,811,084 | 0.346 |
| $c_{2,1}$ | R | $15,509,133 | 0.184 | N | $28,731,644 | 0.399 | Won | $65,919,047 | 2.890 |
| $c_{2,0} \cap c_{2,1}$ | D | $9,610,121 | 0.897 | I | $18,599,240 | 10.549 | Lost | $1,766,869 | 0.110 |
| $c_{2,0} \cap c_{2,1}$ | R | $10,718,698 | 1.115 | N | $1,763,169 | 0.095 | Won | $16,002,559 | 9.057 |
| $c_{4,0}$ | D | $42,180,436 | 1.673 | I | $52,898,862 | 3.496 | Lost | $12,766,549 | 0.273 |
| $c_{4,0}$ | R | $25,210,254 | 0.598 | N | $15,132,022 | 0.286 | Won | $46,814,112 | 3.667 |
| $c_{4,1}$ | D | $62,050,437 | 14.569 | I | $44,353,754 | 1.946 | Lost | $17,871,427 | 0.432 |
| $c_{4,1}$ | R | $4,259,115 | 0.069 | N | $22,786,715 | 0.514 | Won | $41,344,418 | 2.313 |
| $c_{4,2}$ | D | $5,911,107 | 0.729 | I | $12,425,855 | 7.615 | Lost | $1,651,293 | 0.155 |
| $c_{4,2}$ | R | $8,107,265 | 1.372 | N | $1,631,697 | 0.131 | Won | $10,647,208 | 6.448 |
| $c_{4,3}$ | D | $51,228,240 | 0.671 | I | $113,970,282 | 7.819 | Lost | $14,063,970 | 0.141 |
| $c_{4,3}$ | R | $76,311,442 | 1.490 | N | $14,575,496 | 0.128 | Won | $99,479,862 | 7.073 |

Table A.50: Donor History by Community in WY

| **C** | Party | Total | Ratio | Inc | Total | Ratio | Status | Total | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| $c_{2,0}$ | D | $35,830,320 | 0.576 | I | $84,939,099 | 6.101 | Lost | $11,921,889 | 0.156 |
| $c_{2,0}$ | R | $62,179,740 | 1.735 | N | $13,922,138 | 0.164 | Won | $76,230,556 | 6.394 |
| $c_{2,1}$ | D | $9,737,211 | 1.971 | I | $10,644,068 | 2.470 | Lost | $3,656,496 | 0.367 |
| $c_{2,1}$ | R | $4,939,454 | 0.507 | N | $4,309,170 | 0.405 | Won | $9,968,500 | 2.726 |
| $c_{2,0} \cap c_{2,1}$ | D | $2,328,613 | 0.529 | I | $5,999,011 | 6.292 | Lost | $718,722 | 0.130 |
| $c_{2,0} \cap c_{2,1}$ | R | $4,401,518 | 1.890 | N | $953,420 | 0.159 | Won | $5,547,609 | 7.719 |
| $c_{4,0}$ | D | $63,150 | 0.192 | I | $226,925 | 1.369 | Lost | $76,075 | 0.240 |
| $c_{4,0}$ | R | $329,550 | 5.219 | N | $165,775 | 0.731 | Won | $316,625 | 4.162 |
| $c_{4,1}$ | D | $13,229,296 | 0.596 | I | $30,432,763 | 5.531 | Lost | $4,270,834 | 0.152 |
| $c_{4,1}$ | R | $22,187,686 | 1.677 | N | $5,502,496 | 0.181 | Won | $28,163,749 | 6.594 |
| $c_{4,2}$ | D | $24,107,370 | 0.556 | I | $58,370,882 | 6.198 | Lost | $8,378,451 | 0.162 |
| $c_{4,2}$ | R | $43,323,769 | 1.797 | N | $9,417,607 | 0.161 | Won | $51,842,457 | 6.188 |
| $c_{4,3}$ | D | $6,393,438 | 1.295 | I | $9,062,999 | 3.568 | Lost | $2,200,332 | 0.267 |
| $c_{4,3}$ | R | $4,938,548 | 0.772 | N | $2,540,177 | 0.280 | Won | $8,248,228 | 3.749 |

# APPENDIX B

## VOTE PREDICTION RESULTS

The following figures show the results of classification using both the random forests and decision trees on Yea and Nay votes in the United States legislature, as described in Chapter 6.

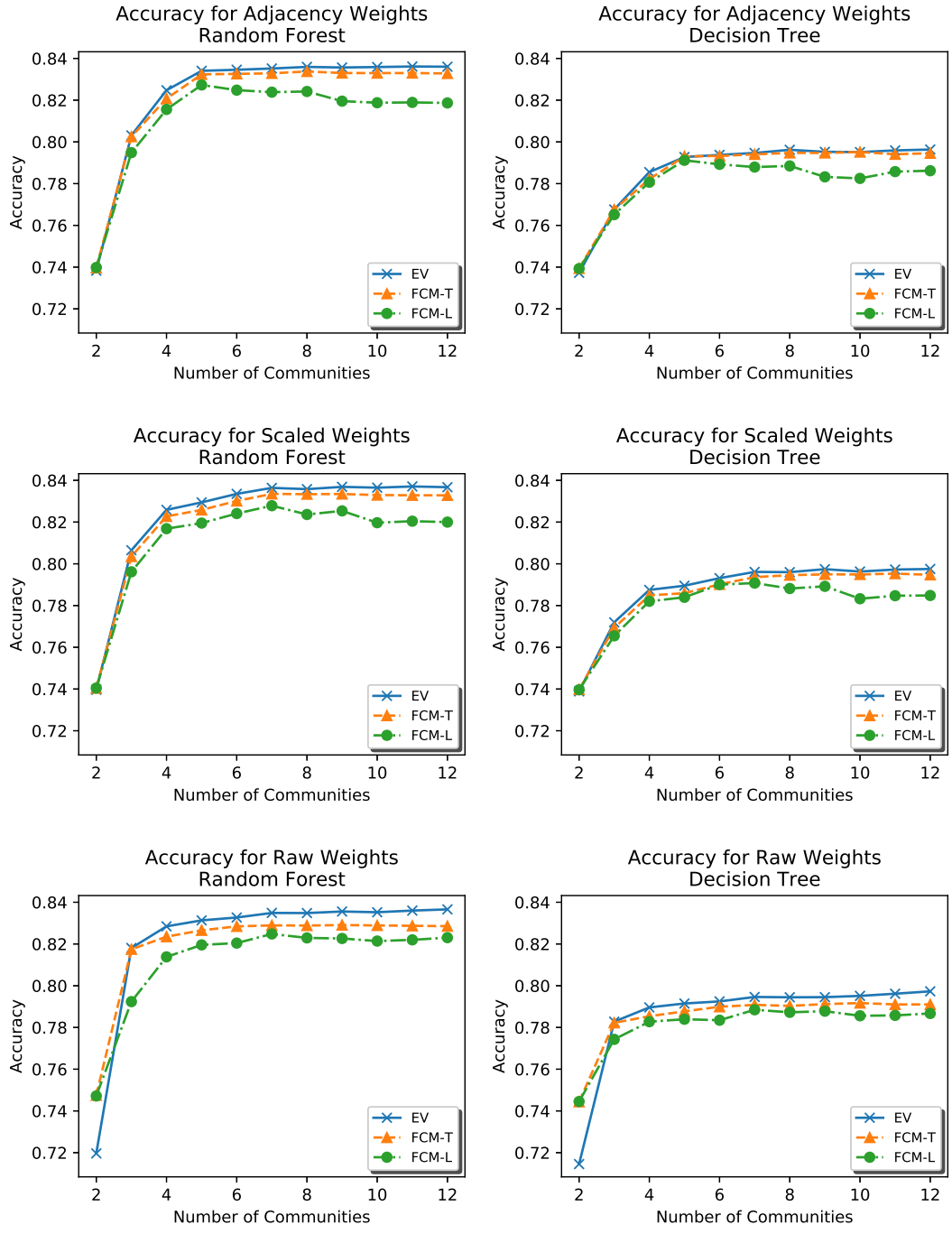



Figure B.1: Decision Tree and Random Forest Vote Prediction for 1980 using Hierarchical Fuzzy Spectral Clustering

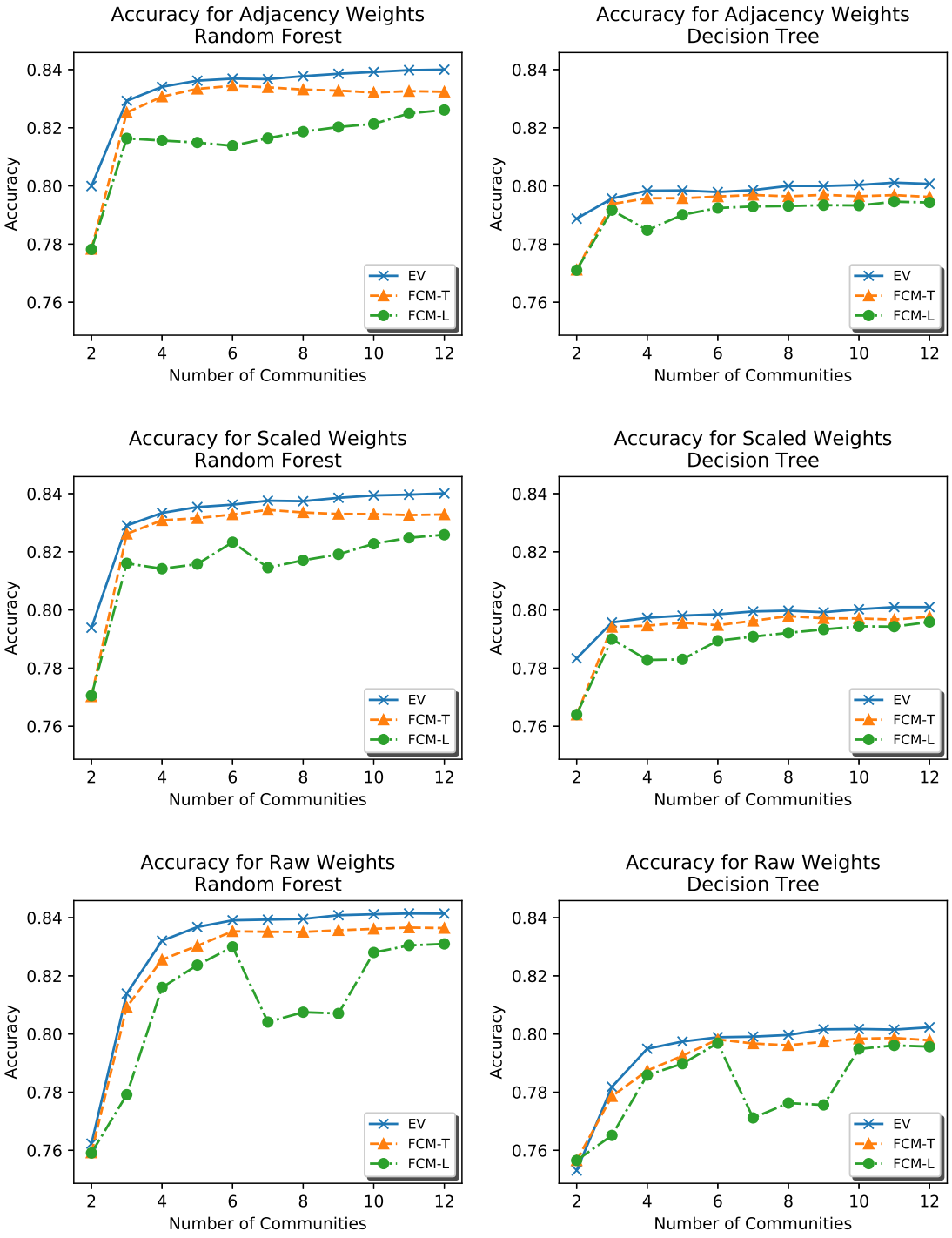



Figure B.2: Decision Tree and Random Forest Vote Prediction for 1982 using Hierarchical Fuzzy Spectral Clustering

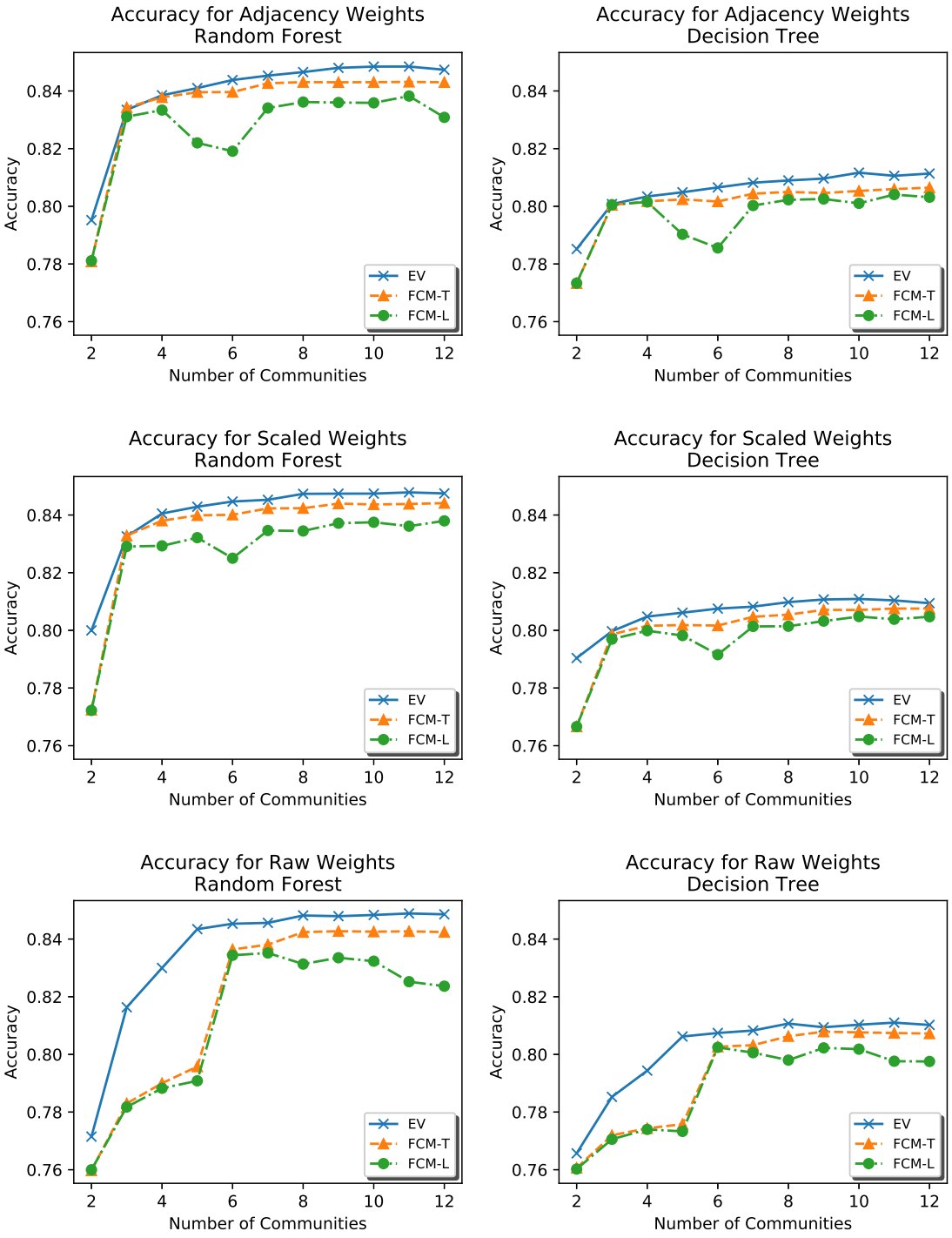



Figure B.3: Decision Tree and Random Forest Vote Prediction for 1984 using Hierarchical Fuzzy Spectral Clustering

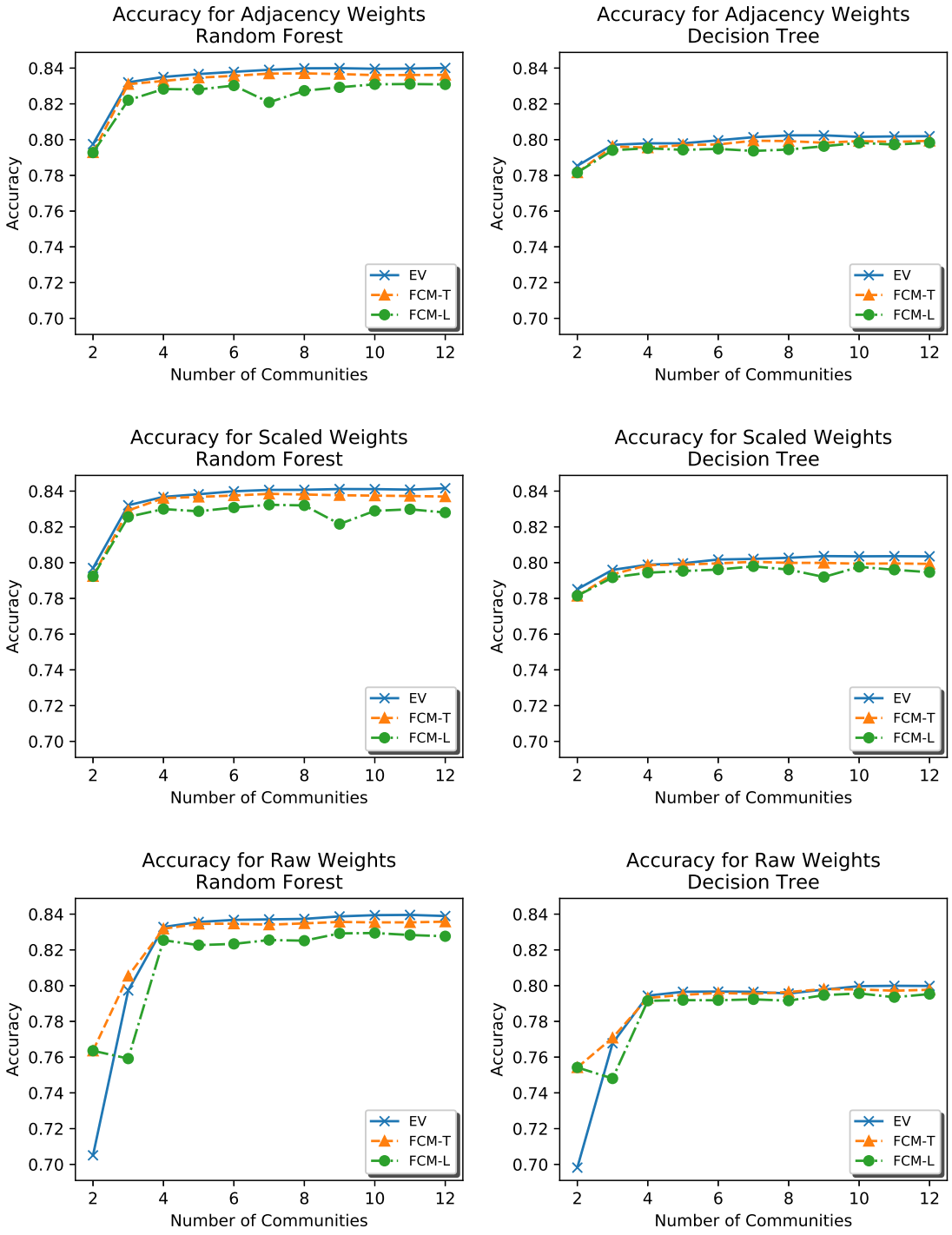



Figure B.4: Decision Tree and Random Forest Vote Prediction for 1986 using Hierarchical Fuzzy Spectral Clustering

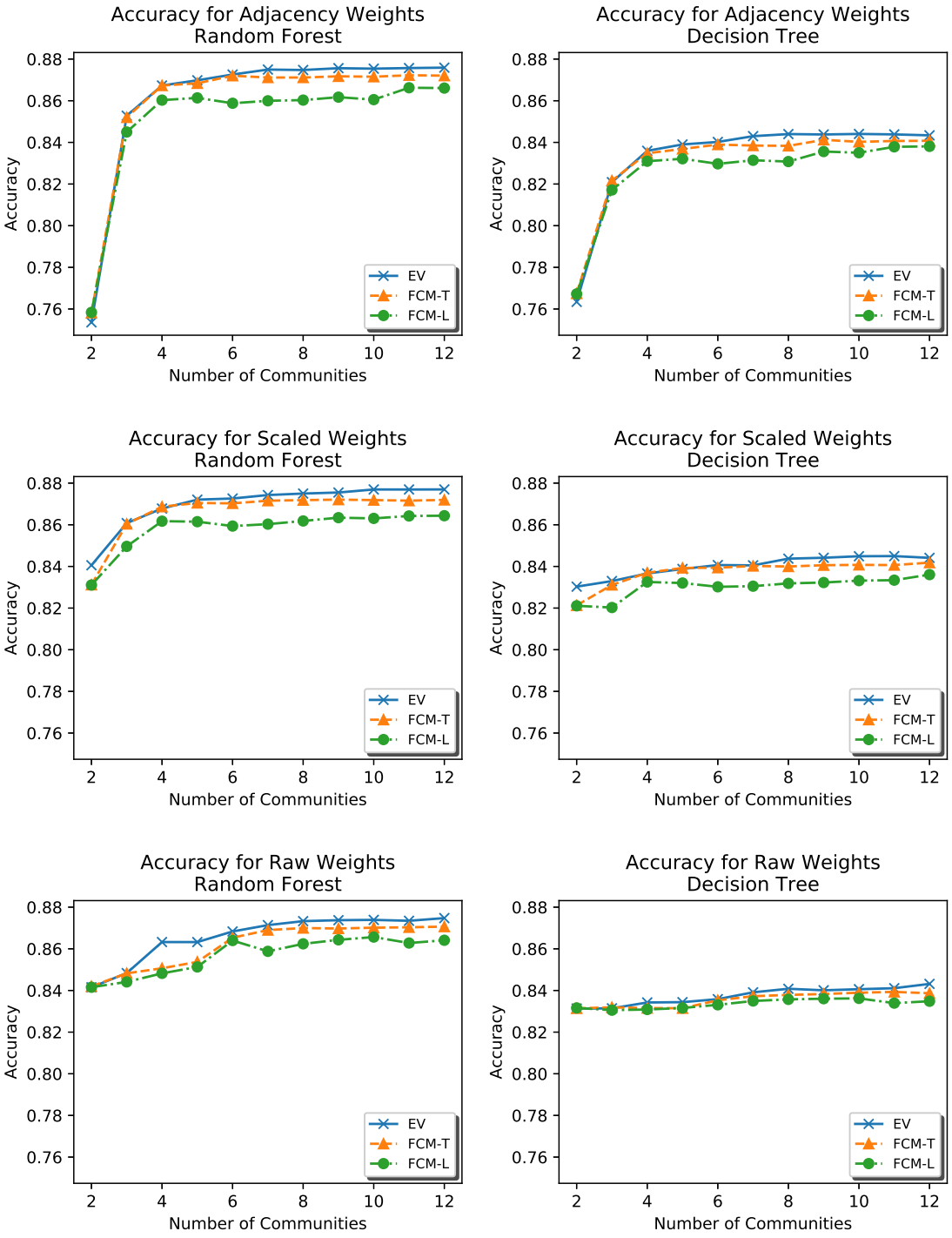



Figure B.5: Decision Tree and Random Forest Vote Prediction for 1988 using Hierarchical Fuzzy Spectral Clustering

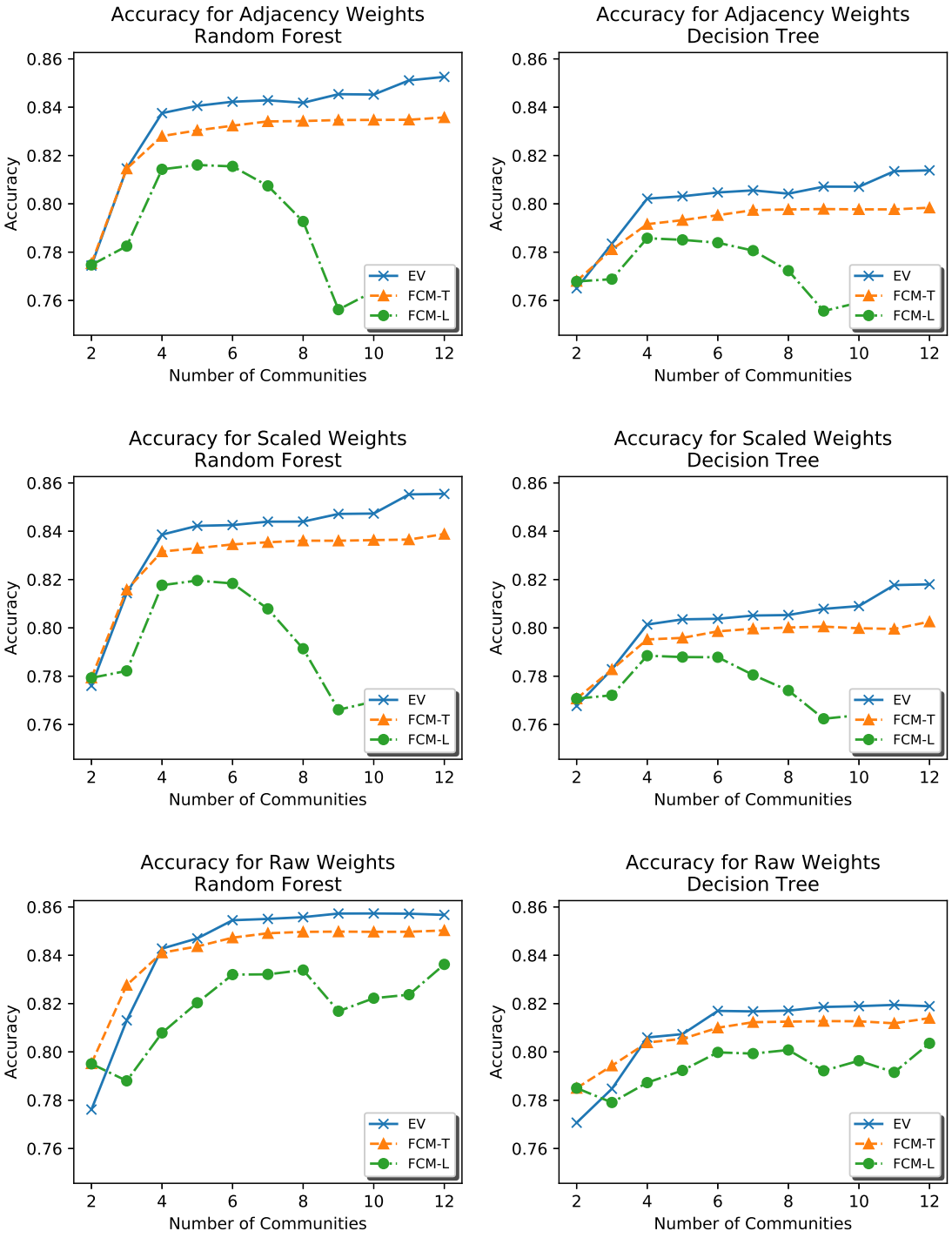



Figure B.6: Decision Tree and Random Forest Vote Prediction for 1990 using Hierarchical Fuzzy Spectral Clustering

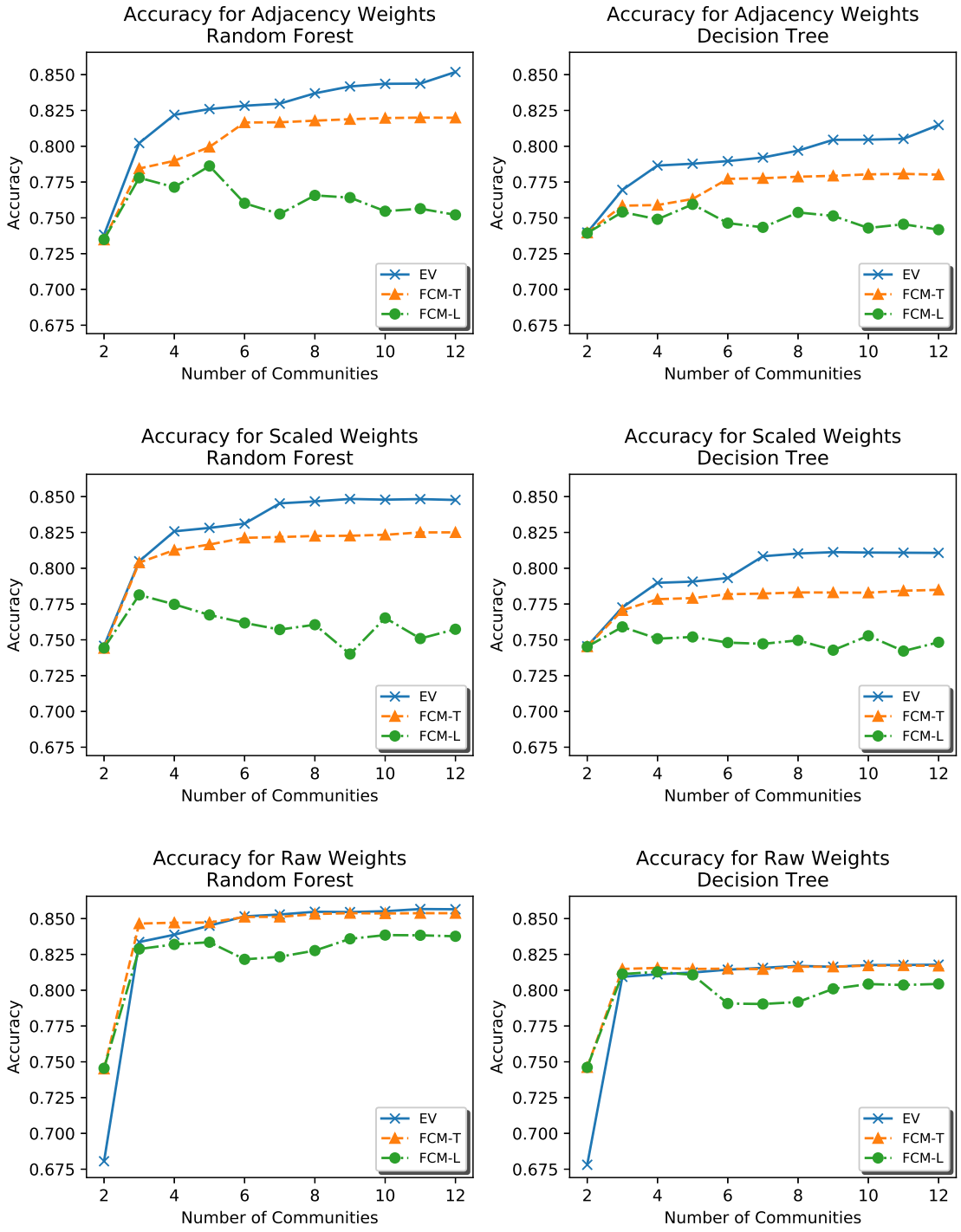



Figure B.7: Decision Tree and Random Forest Vote Prediction for 1992 using Hierarchical Fuzzy Spectral Clustering

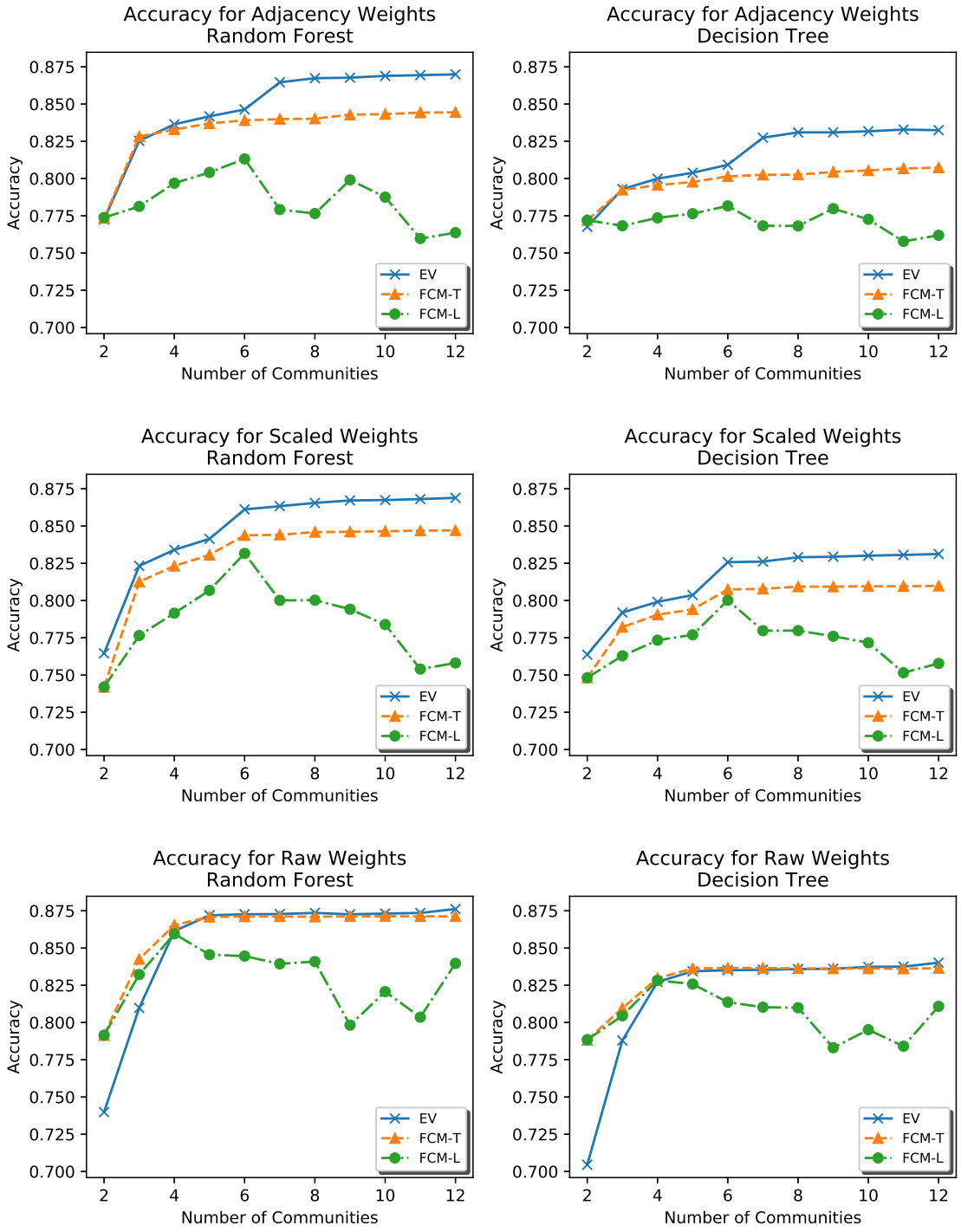



Figure B.8: Decision Tree and Random Forest Vote Prediction for 1994 using Hierarchical Fuzzy Spectral Clustering

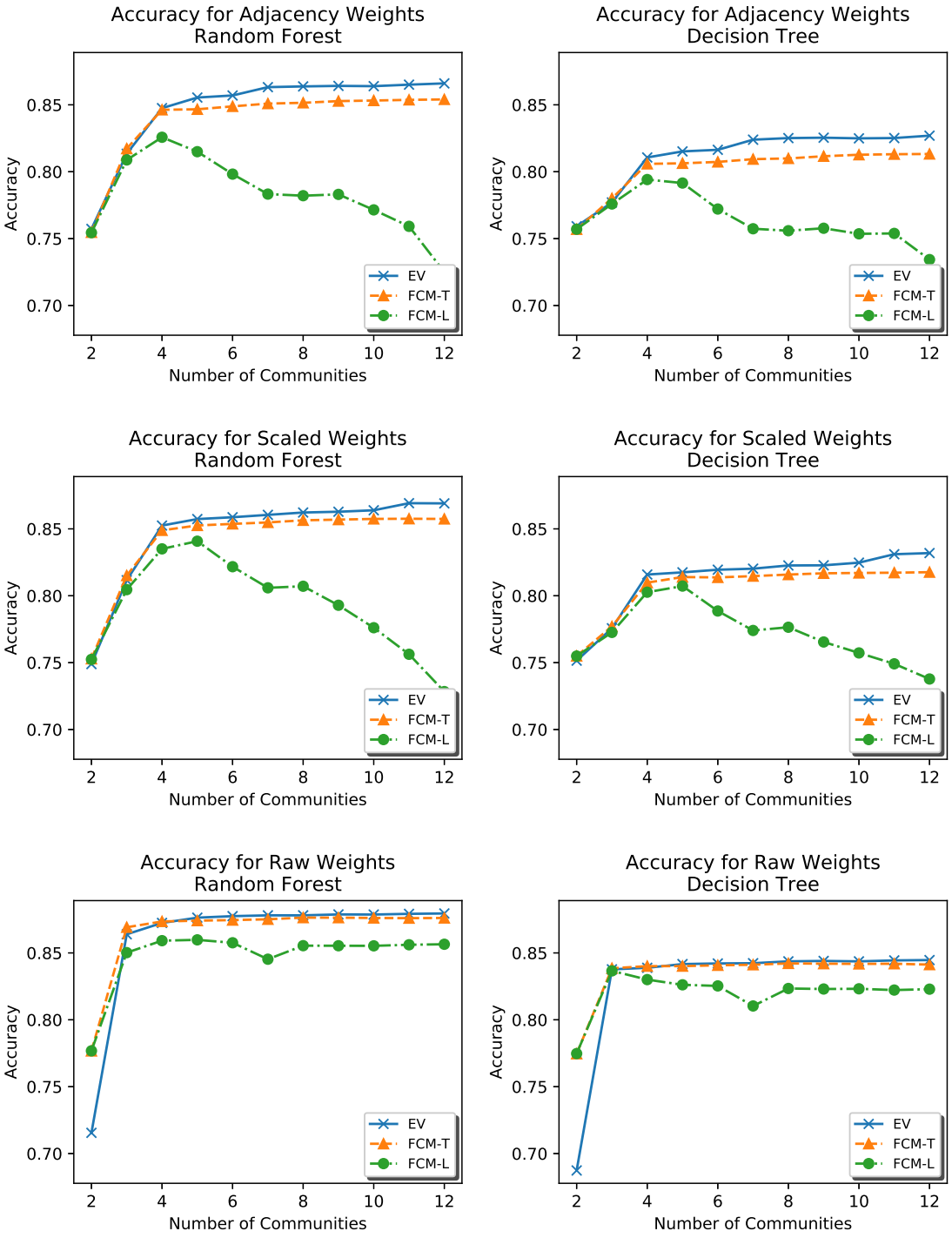



Figure B.9: Decision Tree and Random Forest Vote Prediction for 1996 using Hierarchical Fuzzy Spectral Clustering

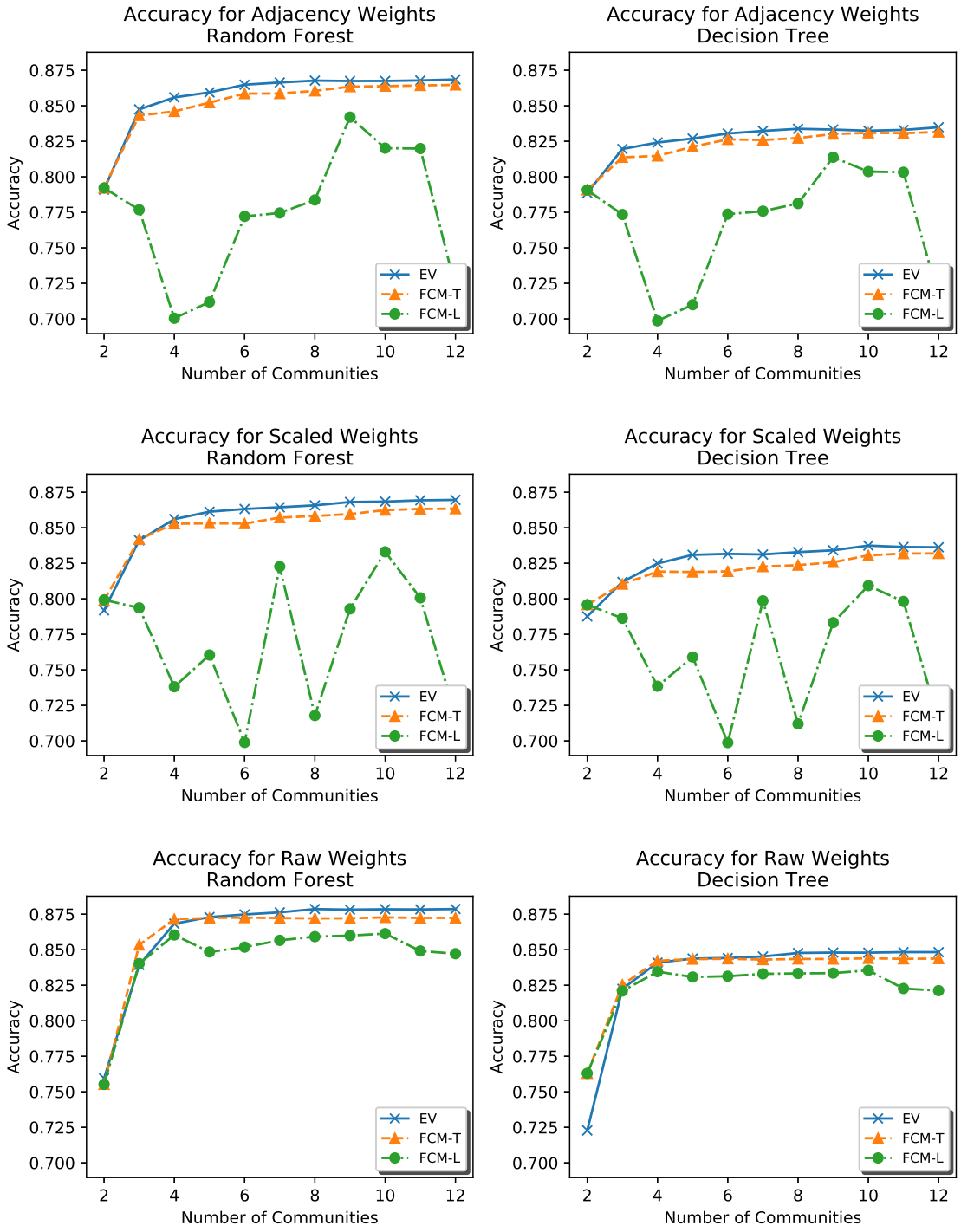



Figure B.10: Decision Tree and Random Forest Vote Prediction for 1998 using Hierarchical Fuzzy Spectral Clustering

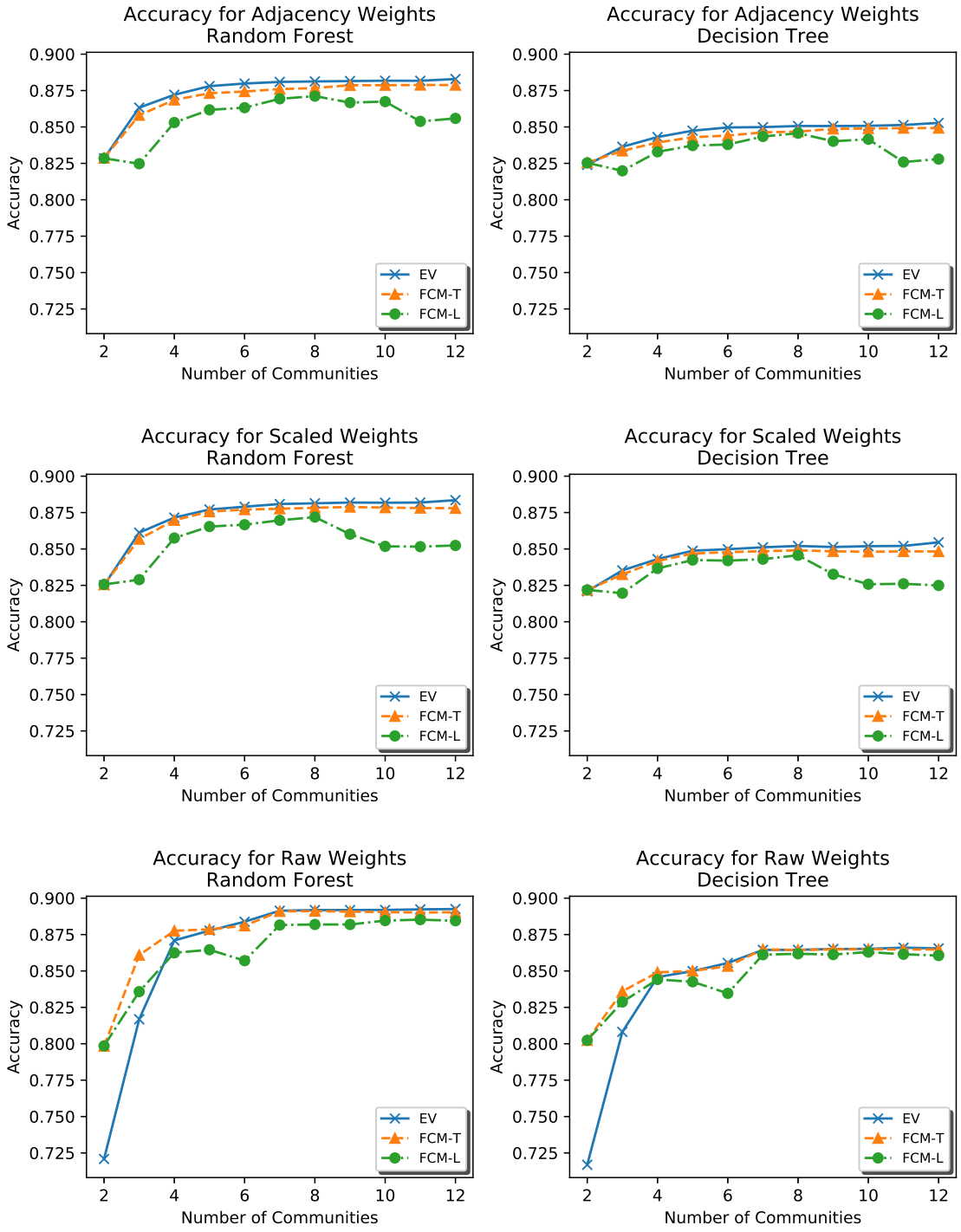



Figure B.11: Decision Tree and Random Forest Vote Prediction for 2000 using Hierarchical Fuzzy Spectral Clustering

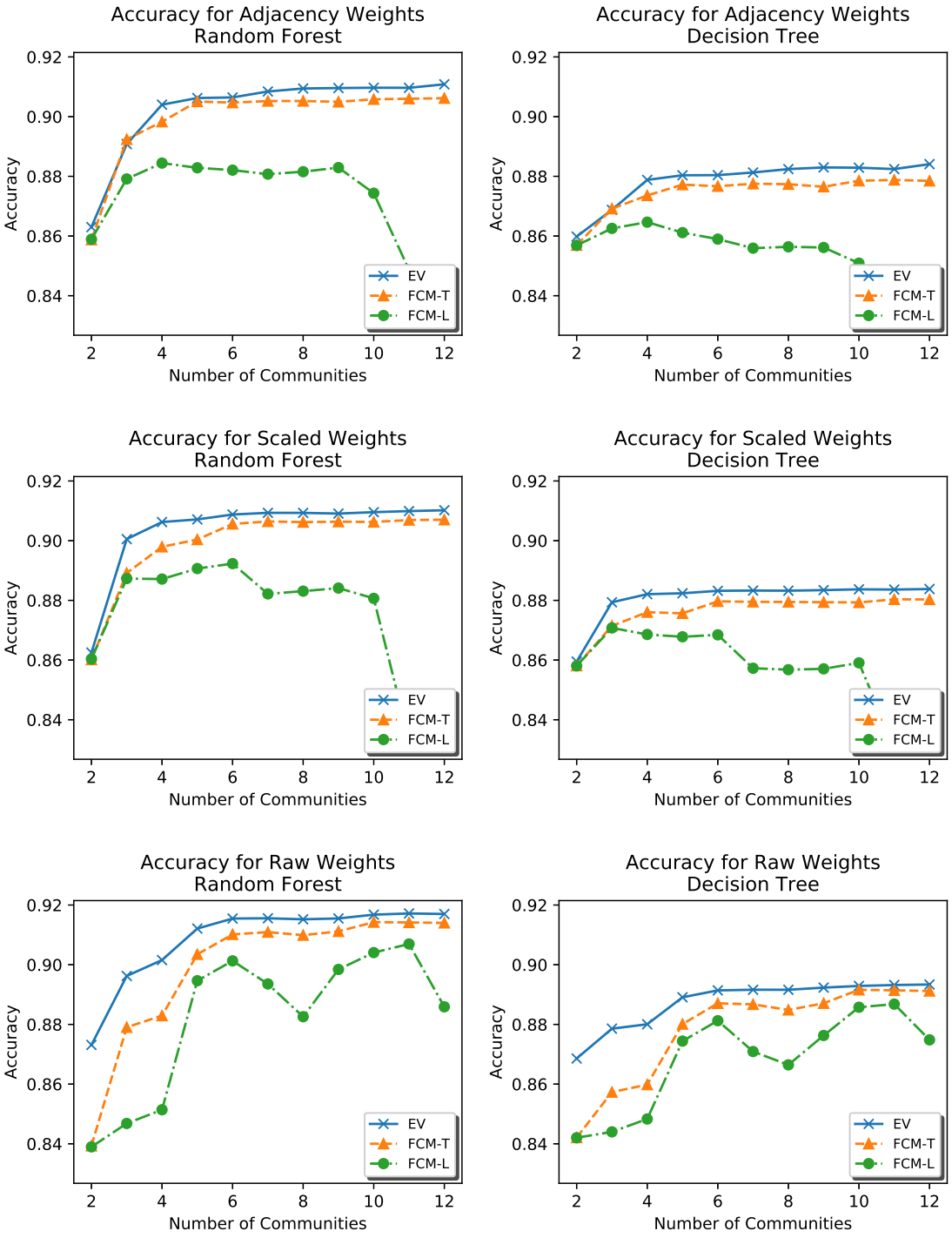



Figure B.12: Decision Tree and Random Forest Vote Prediction for 2002 using Hierarchical Fuzzy Spectral Clustering

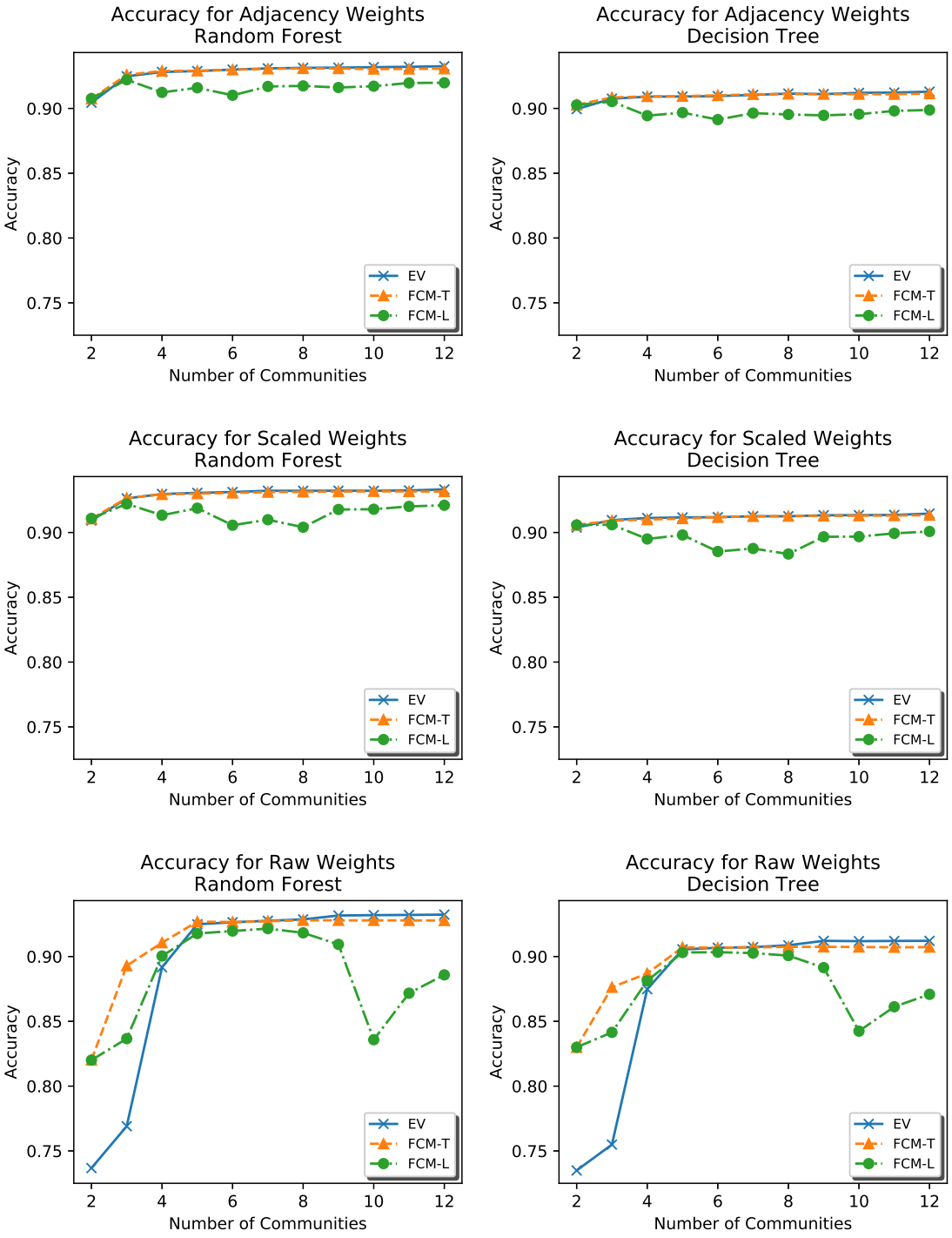



Figure B.13: Decision Tree and Random Forest Vote Prediction for 2004 using Hierarchical Fuzzy Spectral Clustering

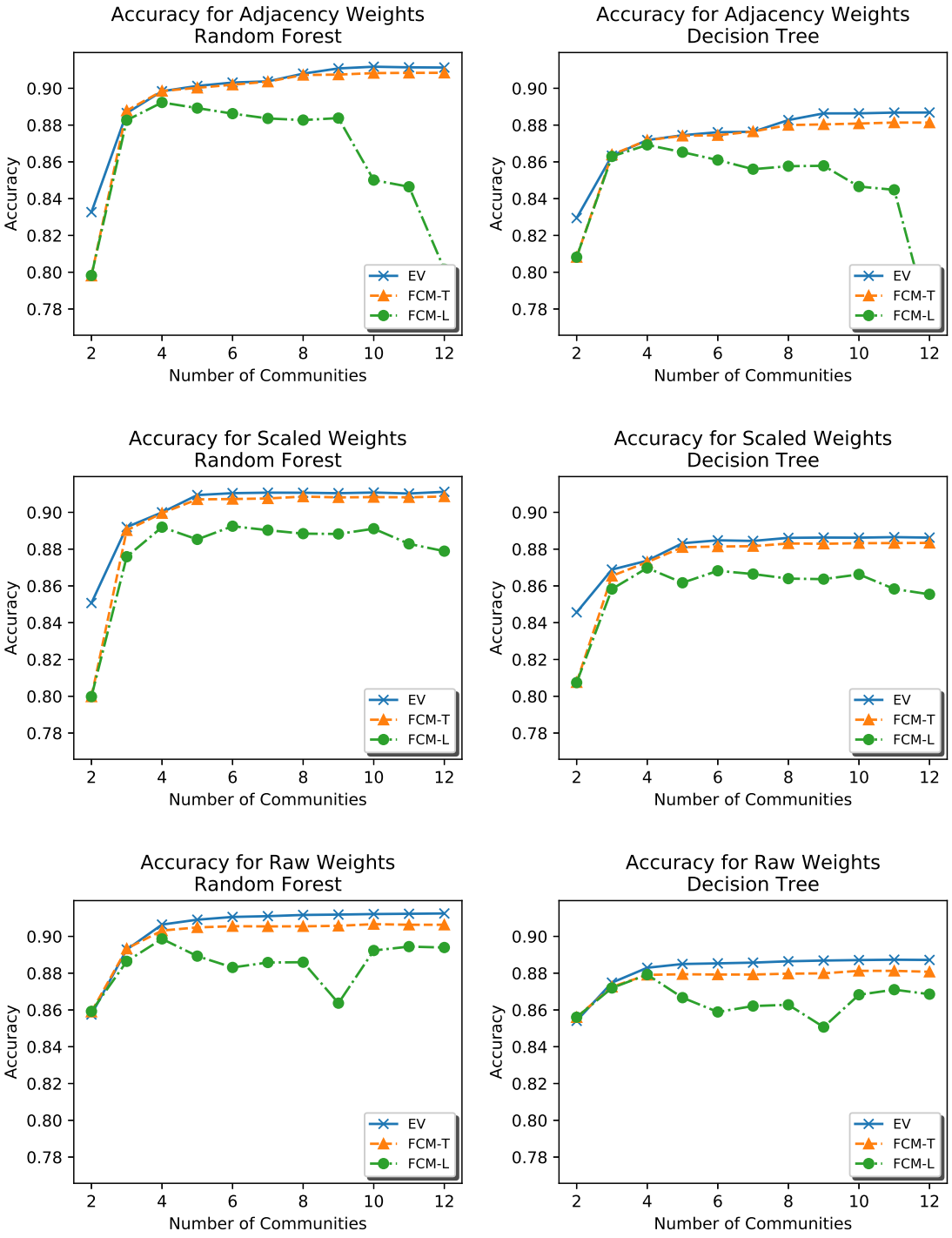



Figure B.14: Decision Tree and Random Forest Vote Prediction for 2006 using Hierarchical Fuzzy Spectral Clustering

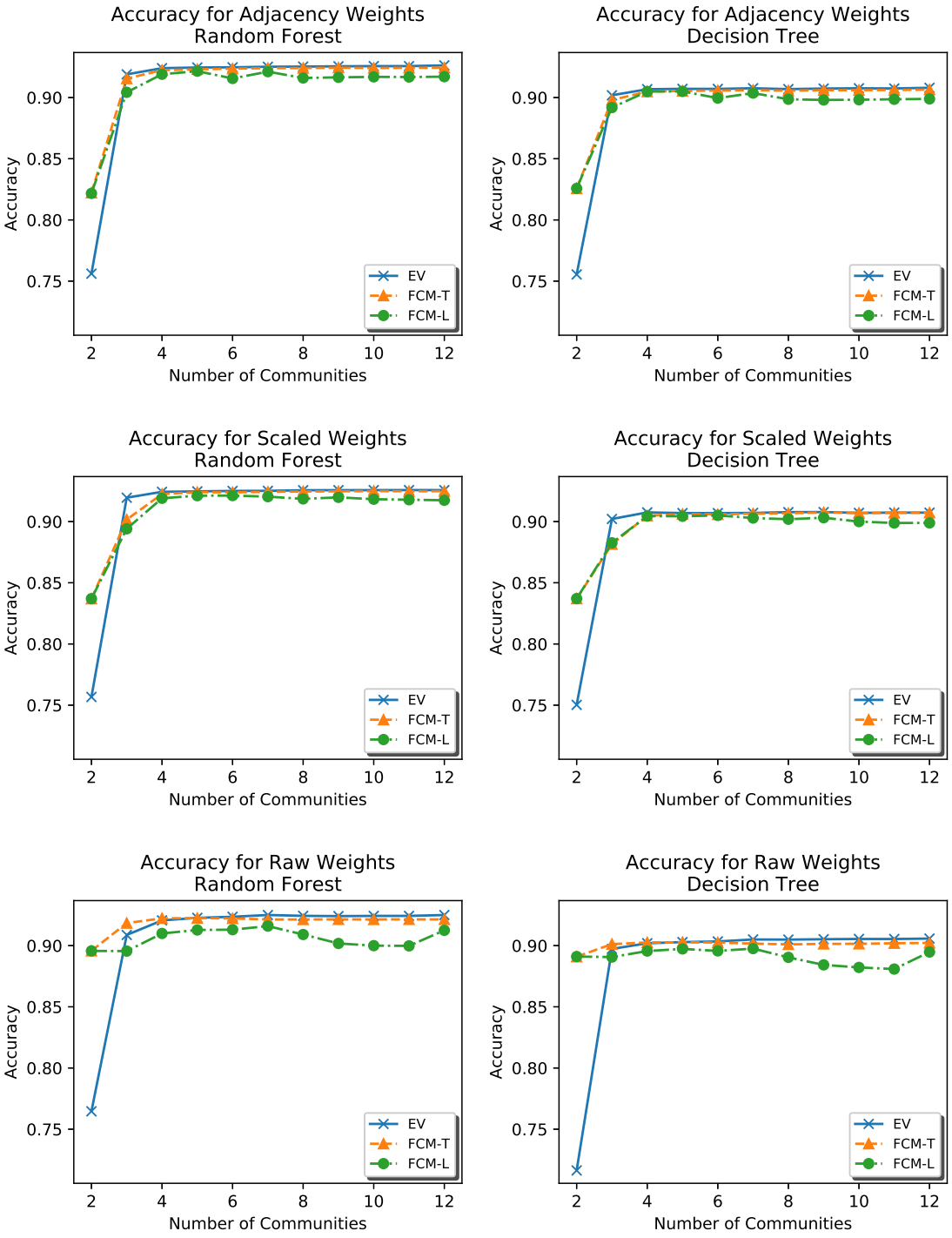



Figure B.15: Decision Tree and Random Forest Vote Prediction for 2008 using Hierarchical Fuzzy Spectral Clustering

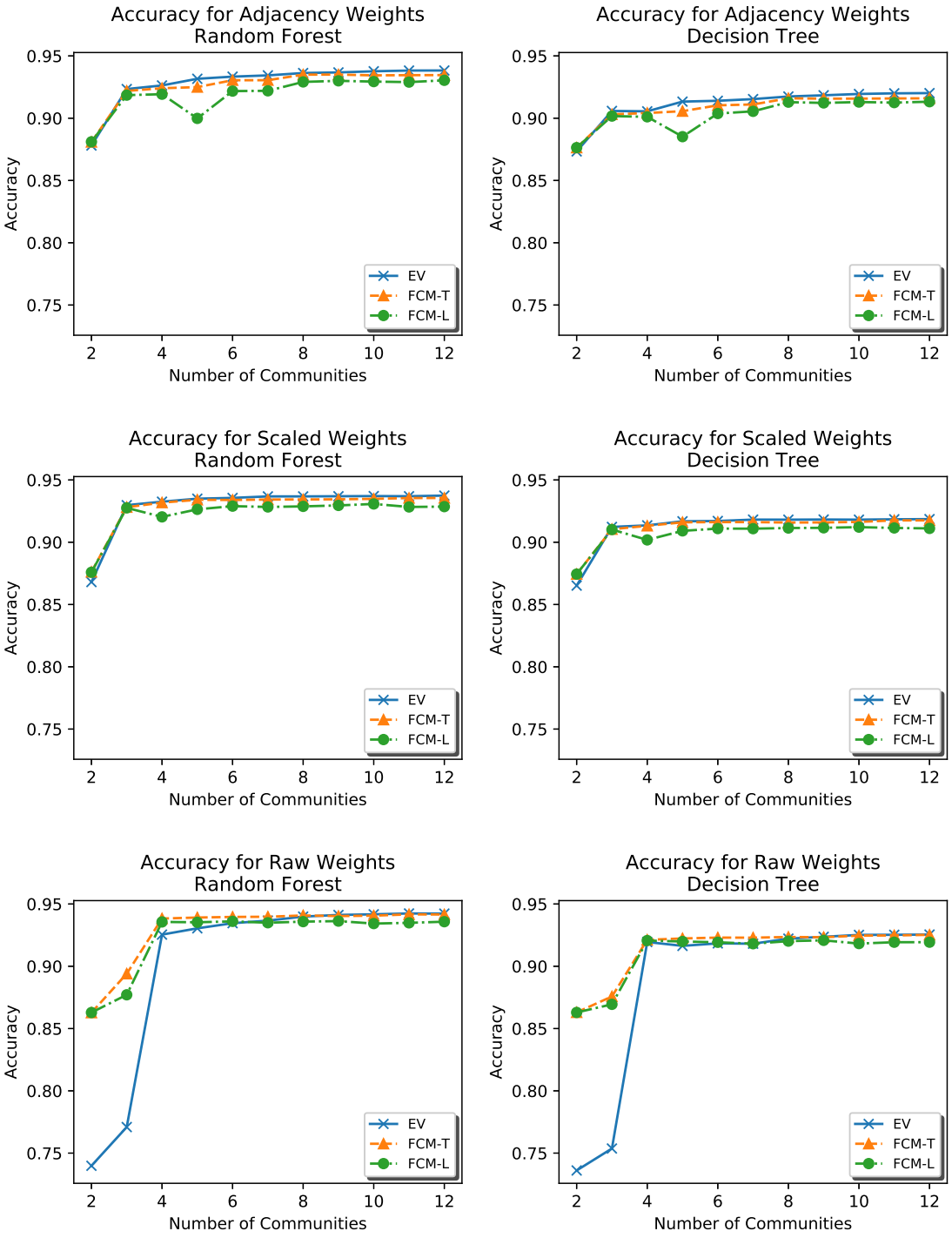



Figure B.16: Decision Tree and Random Forest Vote Prediction for 2010 using Hierarchical Fuzzy Spectral Clustering

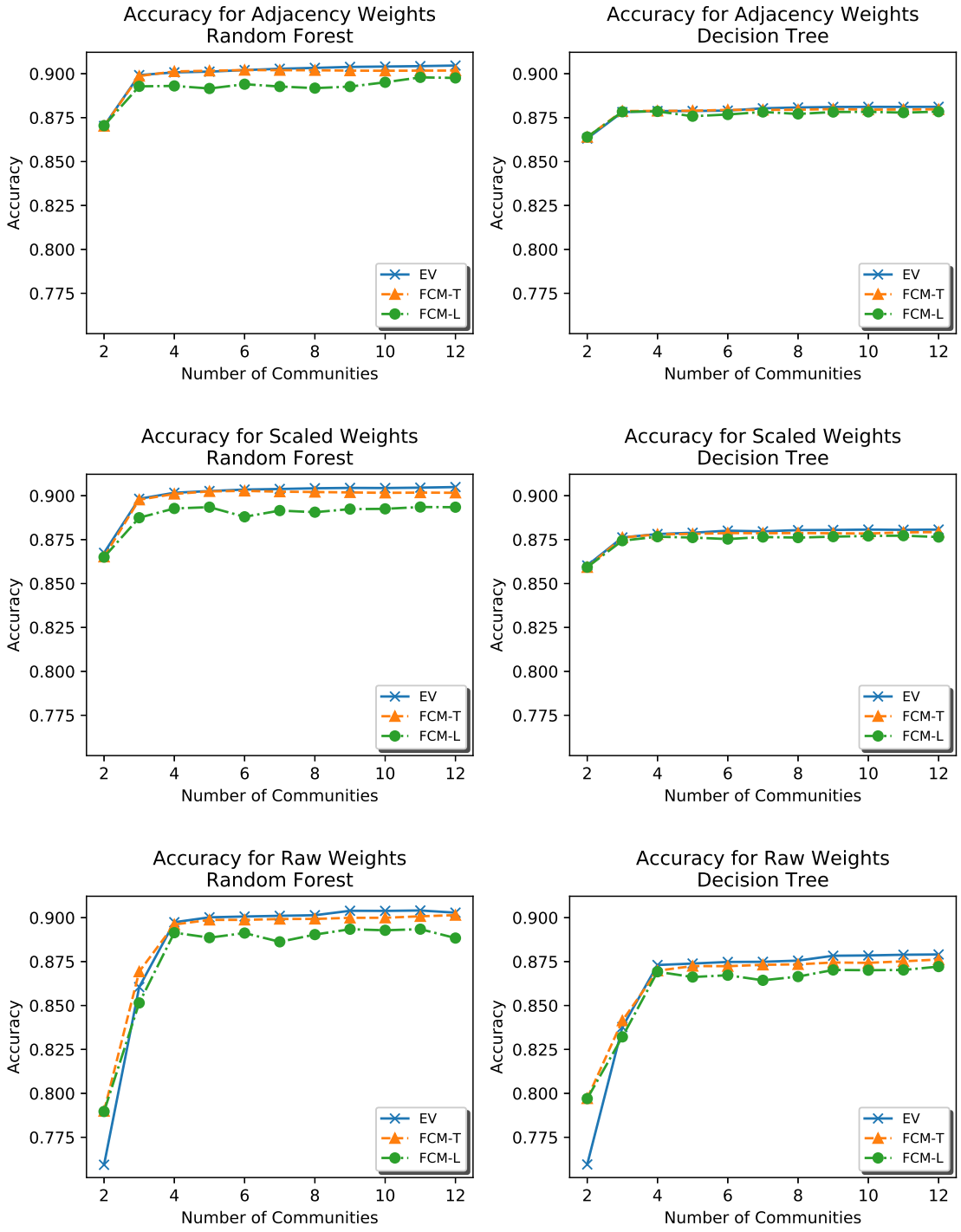



Figure B.17: Decision Tree and Random Forest Vote Prediction for 2012 using Hierarchical Fuzzy Spectral Clustering