Team 014: Analyzing and Visualizing Campaign Finances

Taylor Graham, Niv Balamurali, Sharon Arulpragasam, Beatrice Lee, Federico Reinoso, Kaitlyn Baek

INTRODUCTION

For our project, we decided to pull campaign donation data from OpenSecrets, a non-profit organization that tracks and publishes data on campaign finance and lobbying. Our team aimed to create a set of interactive visualizations that would create an intuitive interface for understanding campaign finance data. We noticed that even though donations have a significant impact on the results of our elections and motivations of our politicians, publicly available records are often shared and analyzed in formats that are inaccessible to the average voter. Inspired by the clustering done by Wahl and Sheppard [16], we decided to create a network out of the donations and explore it for hidden hierarchical structure that could be highlighted in an interactive visualization. Our goal was that this analysis would reveal patterns and trends, such as identifying major clusters of political donation behaviors or highlighting the relationships between donor demographics and the types of initiatives they support.

PROBLEM DEFINITION

Campaign contributions play a significant role in American politics. Understanding the motivations behind individual donations and the impact these contributions have on political outcomes remains a challenge to this day. Our objective is to categorize donors and recipients into distinct groups based on their characteristics to identify underlying patterns and trends in political donations. This classification will allow us to gain valuable insights such as if donors with similar backgrounds favor certain types of political causes or candidates. Beyond mere classification, we aim to uncover and analyze trends within these groups, such as geographic patterns of donations or changes in donor behavior over time.

LITERATURE REVIEW

Demographics is a formula that could be vital for politicians, campaign managers, and political strategists. Rhodes, Schaffner, and La Raja [13] stress the importance of understanding individual's donor strategies in midterm elections and how they play a central role in funding American political campaigns and organizations, yet their motivations are not well understood. Donors also hope to gain access and influence during legislative deliberation through their contributions [11]. Schnakenburg and Turner [15] emphasize the importance that campaign finances have to influence policy decisions through affecting elections or the choices of politicians when in office for citizens. This was a point emphasized by Congleton [4], donors naturally have a greater interest in promoting the success of a particular candidate's position which is significantly closer to the donor's ideal point than another candidate. They see donor behavior to be influenced by electoral motivation and signaling to policymakers.

There is research utilizing creating graphs of campaign data and identifying hierarchical clustering of campaign donations to predict legislative votes but these remain inaccessible to the typical voters [17]. There is also existing research on the different types of small donors, indicating gender, ethnicity, and more [2], but this research falls short of having a broad classification capability. Other analysts have used clustering methods to classify political finance regulatory systems of other countries, which provides inspiration for our own deep dive into specific data in Georgia [8]. Voters typically consider researching candidates' platforms to be taxing, signaling that without easy access to honest finance data voter's are unlikely to take action to look into a campaign's funding [12]. In a paper discussing the dangers of hidden donors it is discussed that in a world where campaign finance data is strictly censored by campaign's themselves the concept of a 'hidden donor' supporting campaigns will always be a threat to honest disclosure of information to voters [1]. The limitations of the current practice leave voters without a reliable and convenient source to view election data from. Transparency of campaign finances and donors greatly impacts elections and can bring to attention potential corruption [6]. Research shows that when legislators are audited for their campaign finances, voters respond negatively if these legislators have not been honest [19]. In a paper regarding if voters truly care about the funding a campaign receives it was shown that if a campaign publicly speaks about their

funding voter's are likely to consider this information when voting [14].

If we chose to utilize classification, decision trees are found to be appropriate due to their intuitive nature and applicability to both categorical and numerical features [3]. The entropy-based subtree splitting used in C4.5 is seen to be more efficient than ID-3 due to its generally increased accuracy and decreased execution time [9]. However, our most promising and interesting approach is inspired by Wahl, Sheppard, and Shanahan [16] who created a graph out of a different donation dataset, and applied fuzzy hierarchical spectral clustering which successfully identified communities of donors and recipients. Their end goal was to predict donor voting habits, but their approach to creating hierarchical sets of donors aligns closely with our goals. We hope to utilize a similar fuzzy hierarchical spectral clustering in order to cluster a larger dataset. In Wahl and Sheppard's [16] approach they successfully clustered a dataset of campaign finance data from Alaska, we hope that in a larger dataset we will reveal more generalizable patterns which we can display in a visualization.

When analyzing the data from OpenSecrets, we could potentially encounter issues such as incomplete and inaccurate data due to Federal Election Commision (FEC) reporting thresholds according to Williams, Gulati, and Zeglen [18]. An article by Christopher Hitchcock and Elliott Sober [7], discusses the distinction between prediction and accommodation within scientific theories, particularly focusing on the concept of overfitting. As datasets grow in size and complexity, traditional data processing tools and methods can struggle to perform analyses in a timely and cost-effective manner [5].

METHOD FOR ALGORITHM

Cleaning

Originally, we planned to analyze and create visualizations for the campaign finance data of all 50 states. However, upon inspecting the bulk data, we found that for just a single state, the file of the campaign finance data can contain millions of rows (with each row representing one donation), making it unrealistic for us to perform analyses on all 50 states. Therefore, we decided to focus on the data from a single state, specifically Georgia, to fit our time constraints for this project. The original bulk data file of Georgia campaign finances contains over 2 million rows detailing individual donations.To consolidate the data, we eliminated candidates that received only one donation as well as donors that only gave once during the election cycle. To additionally filter the data to fit our time constraints, we will focus on donation pairs (contributor and recipient) that had a total donation amount greater than $2000 to filter out trivial donations and limit the size of our dataset.

After consolidating and filtering our data we then began formatting the data appropriately for the clustering algorithm. Utilizing the methods Wahl, Sheppard, and Shanahan [16] used, we began by turning our data into a graph. In order to run the algorithms next steps we require a nonsingular adjacency matrix, to satisfy this constraint we made the graph representing our data undirected. We represented the data as a graph where nodes represent either a donor or a candidate and edges represent a donation made from a donor to a candidate.

Spectral Characterization

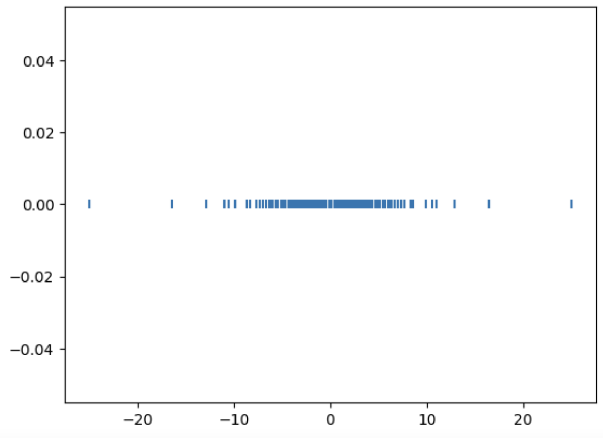

Figure 1: Eigen-spectrum of the adjacency matrix

In order to understand the hierarchical structure of our graph as well as how many clusters exist at each level, we are using the same spectral characterization technique detailed by Wahl, Sheppard, and Shanahan [16]. It involves analyzing the eigen-spectrum of the adjacency matrix in order to identify outliers which represent the individual clusters in each level. We can see in Fig. 1 that there are large spaces after the 2nd, 3rd, 6th, and 8th largest eigenvalues (by magnitude), indicating that we should have as many clusters at each level. Once we have an understanding of the amount

of clusters, we will perform spectral clustering as described by Ng, Jordan, and Weiss [10], but replace their recommended K-means step to the algorithm with fuzzy c-means clustering.

Spectral clustering

From the results obtained from the spectral characterization we aimed to use spectral clustering to create meaningful clusters from our data. Using the results from spectral characterization a similarity graph of our data was constructed from the eigenvalues found of the original adjacency matrix. These eigenvalues are then constructed into their own adjacency matrix that is used to conduct spectral clustering. The Laplacian matrix of the adjacency matrix constructed in the previous step was found and then underwent a series of operations to find and normalize its eigenvalues and eigenvectors. The final step to spectral clustering is putting the resulting matrix from those operations through the clustering algorithm of our choosing.

Fuzzy c-mean Clustering

We chose to utilize fuzzy c-means clustering because of the likelihood that clusters of election donor data would exhibit overlap. Fuzzy c-means clustering is advantageous because nodes are not restricted to belonging only to one cluster, each data point will receive a membership score detailing the percentage of membership that node has to that cluster. This allows us to reflect a more accurate picture of the data rather than an oversimplification that an algorithm like K-means clustering might create.

The fuzzy c-means algorithm iterates through steps including: calculating the centers based on the current membership and threshold values and updating the membership matrix until the threshold value has been met. Iterations are complete when the distance between the cluster centers to each of their data points has been minimized.

Time complexity for this algorithm is high and contributed to the decision to look only at total donations over $2000 in order to manage the time taken to run clustering. We implemented fuzzy c-means clustering by utilizing a fuzzy c-means python package. Using our results from the spectral characterization we were able to sort our data into a clustered hierarchy containing levels of 2, 3, 6, and 8 clusters with each node having a fuzzy membership score that will allow us to visualize the overlapping of clusters and hierarchies present in our complex unstructured dataset.

Jaccard Similarility Score for Hierarchy Construction

After clustering our data into 2, 3, 6, and 8 clusters that represent the levels of our hierarchy we used Jaccard Similarity scores to attach parent clusters to child clusters. Jaccard similarity uses the intersection and union of each level of clusters to deliver a score for each cluster representing their similarity. After calculating a Jaccard similarity between 2 and 3, 3 and 6, and 6 and 8, we used the results to construct a hierarchy.

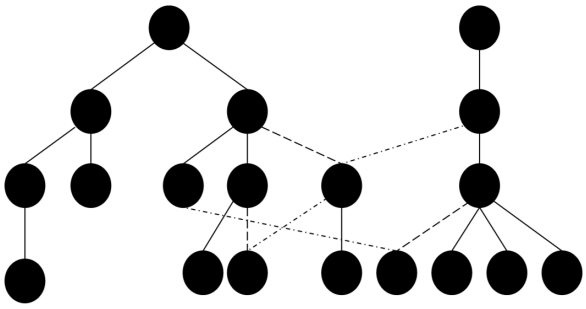

Figure 2: Hierarchical structure revealed by Jaccard similarity scores of clusters

The above structure was identified by our algorithm. Each circle represents a cluster center of the data. Solid lines represent links while dashed lines represent weak links between clusters with the darker the weight, the stronger the connection.

Decision Tree

In order to understand the characteristics of our clusters with respect to the information we already had access to for donations (gender, amount, political party, etc.), we decided to create decision tree classifiers that would attempt to predict the cluster of our contributors. We hoped that upon creating a decently accurate decision tree classifier, we could examine its structure (where and why it split) in order to learn more about the types of contributors that made up each of our clusters. Unfortunately, the decision tree, whose parameters were tuned for every set of labels, only ever achieved a maximum of 65% accuracy with our binary labels

and even less with all others. This led us to conclude that there was no significant relationship between the clusters produced by our network algorithm and the characteristics of donors we had access to.

METHOD FOR VISUALIZATION

Proposed Methods

When we first started the project we didn't have a clear idea of what type of results our algorithm would give us, and we were excited with the idea of creating a choropleth map. We soon determined this would not allow us to see structures we were looking for in our data. We next created a zoomable circles visualization that displayed hierarchical data, but we ended up giving up on this idea because it did not allow us to visualize fuzzy clustering. To visualize fuzzy clustering, we started creating a visualization of nested venn diagrams, but ultimately found this would not give viewers a big picture perspective of our results.

Sunburst Visualization and Hover Functionality

We pivoted from the previous ideas to a zoomable sunburst visualization because it could show the big picture of our clustering results while still allowing us to show our fuzzy clusters distinctly. In our representation of the data we use red and blue to define individual clusters, while shades of purple show places where clustering is fuzzy. The zoom feature allows users to gain more understanding of how the clusters work at each level of the hierarchy. Ultimately, we found that the sunburst gave the most valuable information while still being simple enough for an everyday user to grasp. We decided to use a mouseover action over the sunburst to allow for users to see more information about the clusters to the right of the overview visualization. Here, users can understand more of what makes up the components of each of our clusters. These graphs should provide our visualization users with information like gender ratios, employment status, and contribution amount of a cluster.

Overall, we attempted to create a visualization that does not overwhelm the user, but still displays a large magnitude of information. Through the interactive components of the visualization, the user can easily navigate through information about over 20 clusters.

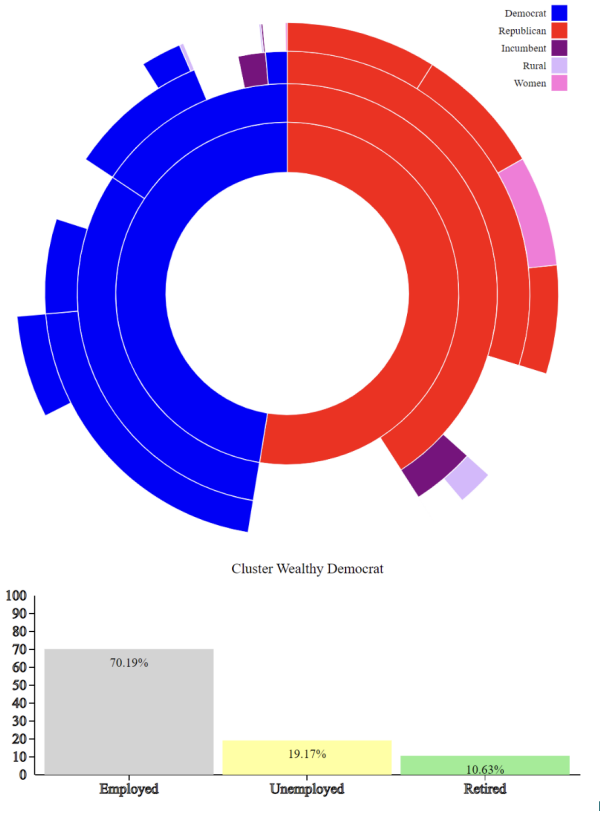

Figure 3: Sunburst Visualization with Bar Graph Examining Employment Status of the Wealthy Democrat Cluster

EXPERIMENTS/EVALUATION

Our goal for this experiment was to uncover underlying trends and patterns in campaign finance data and make these results easily accessible to voters. Our experiment was designed to answer questions such as "Are there any underlying structures in the campaign financial data in Georgia?", "Are there any interesting relationships between sub communities and hierarchies in campaign financial data?", "Is there any relation between this structure and the success of candidates?" and "What features within our data are the most relevant indicators of donation behavior?".

Utilizing spectral characterization and fuzzy c-means clustering, we were able to obtain results regarding how the data should be clustered. Furthermore, using Jaccard's similarity, we were able to uncover the hierarchical structure between the clusters of our data. However, we struggled to find the significance of our clusters. Despite knowing what nodes belonged to what

grouped, we were unable to determine what characteristic defined a specific cluster. We had expected to see results similar to Wahl, Sheppard, and Shanahan's [16] analysis of Alaska campaign finance data: specifically we expected that the top level of the hierarchy would be primarily split by party. However, upon a simple visual inspection of our results, we found that both clusters in the top level had significant percentages of both Democrats and Republicans, indicating that the top level of our hierarchy was not defined by party. Thus, we attempted to use a decision tree to help classify and label our clusters. However, aftering running accuracy tests on our decision tree algorithm, we found that for the top level of the hierarchy (binary), the accuracy was approximately 65%, which is not significantly higher than if the algorithm was to just assign random arbitrary labels. The lower levels of the hierarchy yielded even lower accuracy percentages, so we concluded that our decision tree was inaccurate and overfitted our data to fit some sort of label.

Through our experiment, we were able to determine that there indeed was a hierarchical structure within campaign finance data and what donations belonged together in the same clusters. However, we were unable to determine what the significance or meaning behind the clusters was, and ultimately we were unable to determine the underlying trends and relationships existing within campaign finance data. Although we were unable to return valuable results, we decided to create arbitrary labels and move forward with the visualization portion of the project. Therefore, our visualizations represent true percentages and breakdowns of each cluster and accurately represent the hierarchies we found in the data, but the labels of each cluster were made to fit what we had originally expected and do not hold any meaning, and exist only to show what we had hoped our visualization would be capable of displaying.

CONCLUSION

Utilizing hierarchical clustering we aimed to categorize donors and recipients based on their characteristics, hoping to reveal insights into why people donate to political campaigns and the effects of these contributions on political outcomes. Our literature review highlights the reasons we wanted to create this type of transparency in campaign finance data and the potential for our analysis to contribute valuable insights into donor behavior, campaign finance regulation, and voter engagement.

Through our approach of using spectral characterization and clustering to reveal hierarchical clustering information in our data we found a supposed hierarchy that our data adhered to. Using this hierarchy we began work on a visualization that would accurately be able to reflect the nested nature of our clusters as well as the cluster overlap that our fuzzy c-means clustering algorithm identified. While we began to develop the visualization we searched our clusters for characteristics that defined our clusters. We had hoped to find evidence of clusters being clearly defined by political party or employment type but we struggled to find defining characteristics. We utilized a decision tree to see where potential branching characteristics should exist in our data. While the decision tree and further analysis revealed some small patterns in the data we could not define each level of our hierarchy by the characteristics that defined a cluster of data.

In order to construct a visualization that was usable for the scope of this project we opted to create labels for our clusters that aligned with our original expectations of the data rather than the ambiguous clusters our algorithm actually identified. While the visualization does contain the data, hierarchy, and clusters found using our algorithm the labels are not based on patterns we were able to discern with the results of our algorithm. When navigating the visualization any patterns that we found using our algorithm are noted by a disclaimer.

Overall we had hoped that by following the approach used by Wahl, Sheppard, and Shanahan [16] to successfully analyze Alaska's campaign finance data that we would be able to define similar clusters and discern clear meaning from our clusters. Unfortunately during the time scope of this project and the manageable complexity of our algorithm we were only able to cluster and identify structures that held no clear meaning. The algorithm holds promise but would require further investigation into modifications that could be made to the spectral analysis and clustering techniques used that could help discern a more meaningful set of patterns in the data.

Statement of Work

All members contributed an approximately even workload and number of hours to this project.

REFERENCES

[1] R Michael Alvarez, Jonathan N Katz, and Seo-young Silvia Kim. Hidden donors: The censoring problem in us federal campaign finance data. *Election Law Journal: Rules, Politics, and Policy*, 19(1):1–18, 2020.

[2] Laurent Bouton, Julia Cagé, Edgard Dewitte, and Vincent Pons. Small campaign donors. Technical report, National Bureau of Economic Research, 2022.

[3] Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

[4] Roger D Congleton. Campaign finances and political platforms: the economics of political controversy. *Public choice*, pages 101–118, 1989.

[5] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[6] Brent Ferguson and Chisun Lee. *Developing Empirical Evidence for Campaign Finance Cases*. New York: Brennan Center for Justice, 2016.

[7] Christopher Hitchcock and Elliott Sober. Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55(1), 2004.

[8] William CR Horncastle. Model based clustering of political finance regimes: Developing the regulation of political finance indicator. *Electoral Studies*, 79:102524, 2022.

[9] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2):13–19, 2014.

[10] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[11] Spencer Overton. The donor class: Campaign finance, democracy, and participation. *U. Pa. L. Rev.*, 153:73, 2004.

[12] David M Primo. Information at the margin: Campaign finance disclosure laws, ballot issues, and voter knowledge. *Election Law Journal*, 12(2):114–129, 2013.

[13] Jesse H Rhodes, Brian F Schaffner, and Raymond J La Raja. Detecting and understanding donor strategies in midterm elections. *Political Research Quarterly*, 71(3):503–516, 2018.

[14] Thomas S Robinson. When do voters respond to campaign finance disclosure? evidence from multiple election types. *Political Behavior*, 45(4):1309–1332, 2023.

[15] Keith E Schnakenberg and Ian R Turner. Helping friends or influencing foes: Electoral and policy effects of campaign finance contributions. *American Journal of Political Science*, 65(1):88–100, 2021.

[16] Scott Wahl and John Sheppard. Hierarchical fuzzy spectral clustering in social networks using spectral characterization. In *The twenty-eighth international flairs conference*. Citeseer, 2015.

[17] Scott Wahl, John Sheppard, and Elizabeth Shanahan. Legislative vote prediction using campaign donations and fuzzy hierarchical communities. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 718–725. IEEE, 2019.

[18] Christine B Williams, Jeff Gulati, and Mateusz Zeglen. Following the money: Uses and limitations of fec campaign finance data. *Interest Groups & Advocacy*, 9:317–329, 2020.

[19] Abby K Wood and Christian R Grose. Campaign finance transparency affects legislators' election outcomes and behavior. *American Journal of Political Science*, 66(2):516–534, 2022.