

Communities

Nina-Marie Cielo^{3*} Cameron Eller^{2†} Jake Pellett[‡] Thomas Schupp^{1§}

2023-09-27

Abstract

This study explores the potential variations in congressional AFL-CIO scorecards among communities created by donor networks. The Null Hypothesis (H₀) posits that these communities exhibit the same mean union score, while the Alternative Hypothesis (H_a) suggests differences in mean union scores. To investigate these hypotheses, a comprehensive analysis of various donor network communities is conducted, employing statistical methods to assess and compare their union scores. We find that communities created by donor networks meaningfully represent different mean AFL-CIO scores.

1 Introduction

In the ever-evolving landscape of political campaigns, understanding the intricate web of connections between individual donors, political action committees (PACs), and candidates is paramount. This paper presents a comprehensive data analysis utilizing R Studio to unravel the dynamics of political contributions and community structures. The initial step involved coding to classify data into distinct categories, distinguishing between individuals and PACs based on a designated column.

The dataset was divided into two subsets: ‘individuals_df’ and ‘pac_df.’ For ‘individuals_df,’ a network analysis was conducted, establishing edges between candidates based on individual donations. Connections were weighted, reflecting the cumulative contributions from shared individual donors. A similar approach was employed for ‘pac_df,’ where connections between candidates were determined by PAC donations, and weights represented the financial support provided.

Following the network analyses, the research delved into exploring candidate communities within the network. The investigation extended to examining these communities’ scorecard data, aiming to discern patterns and similarities among groups of candidates. This multifaceted approach offered a nuanced understanding of the relationships within and between different political entities.

The analytical phase included hypothesis testing, focusing on comparing the mean union scores of candidate communities. To evaluate whether communities exhibited similar or different means, ANOVA tests were employed, complemented by the use of standard error and other relevant statistical measures. Specifically, the paper delves into comparing means within communities with PACs, providing insights into potential disparities in political contributions.

*American University

†American University

‡American University

§American University

Furthermore, the analysis goes beyond statistical comparisons, delving into the nuanced characteristics of different candidate groups. Descriptive methods are utilized to elucidate the distinctive traits and behaviors within these political communities, shedding light on the intricate interplay of individual and PAC contributions.

This scientific data analysis paper, rooted in R Studio, provides a robust exploration of political contributions and community structures. By combining network analyses, hypothesis testing, and descriptive approaches, the study offers a comprehensive perspective on the complex relationships that define contemporary political landscapes.

2 Political Donations and Network Analysis

The last year has been marked by a growing number of coordinated labor movements and union action across several industries in the United States. Our research explores the nature of connections between labor union legislative scorecards from members of congress and the networks created by prolific wide-reaching large campaign contributors.

There are a number of political implications of the expectations placed upon a candidate by a donor, including but not limited to ideological transparency and the degree of a representative agenda. Conversely, the actions that a congressperson takes to attract donations to further their own interests is also a deep topic of study. La Raja, Raymond J., and Brian Schaffner (2015) theorize that candidates who receive very large funding from individuals are very likely to be ideologically centrist, while candidates who seek funding from large amounts of small donations are more likely to be ideologically extreme. Additionally, the theory extends to corporations and business groups to claim that they will tend to favor moderate and conservative officeholders who will bargain over specifics rather than overarching ideological change. In our research, we expect to find that large donations from corporations will skew toward conservative networks in congress, while large donations from individuals will flow to centrist networks.

The academic literary body surrounding lobbying and political donations is dense, but there is a distinct difference in the post Citizens United v Federal Election Commission (2010) world. Citizens United further emphasized an already disproportionate influence in massive donations and dark money in elections. Hansen and Rocca (2019) draw a clear line in analysis of individual campaign contributors in the pre and post Citizens landscape to argue that a flood of extremely large donations to super PACs have contributed to a reduction in academic research surrounding political donations. Our investigation seeks to construct a web of large donations from available sources to understand the financial connections between current members of congress and highly active large donors. The primary difference in a post Citizens analysis is that many of the largest donations, largely from CEOs, is that the data has become obfuscated by inflated dark money values.

Graph theory is a powerful tool for analyzing complex systems of interacting agents in the political context. The precedent set by the research of Porter et al. (2005) to project bipartite networks onto one-mode networks, where the strength of connections can be quantified using the normalized interlock, in measuring social networks in congressional committees allows for an effective method of congressional network analysis. Our chosen method of clustering is the walktrap algorithm, discussed in greater detail in the following sections.

Here we go deeper into the intellectual debate, the political and social context of our investigation. To give the reader a clear sense of why we are writing this paper, we describe the relevant scholarly, technical, or popular literature. We give this section a meaningful *substantive* title; it is not entitled “Literature Review”, for example. We cite at least three published, peer-reviewed scholarly works.

For example, we could cite @mooree20 or @moorav12¹, which we discussed in class.² We only cite others' work in our paper when it enhances the reader's understanding of what we, the authors of this paper, are doing. We connect everything we cite to *our* investigation; this is our original research, not a book report or an annotated bibliography.

We do not cite paper titles or journal names, unless our paper is about someone else's paper or about the set of articles in a journal. We do not cite authors' first, given names. We can refer to either what an author does, or what a paper does, but we should be consistent. For example, "@moorav12 argue that we should ..." refers to what the authors do; "@moorav12 argues that we should ..." refers to what the paper – @moorav12 – does.

In order to integrate citations into the References section below, we add entries into our file `main.bib`. This is a plain-text file that we edit in RStudio. We store `main.bib` in the same folder as our paper's `.qmd` and `.pdf` files. Its entries are formatted so that they can be knit to `.pdf`; see <https://j.mp/2UzTXEZ> for example entries for articles, books, and miscellaneous. We can get these entries automatically from Google Scholar by turning on BibTeX in the Google Scholar Settings - Bibliography Manager. Perhaps we use a tool like free, open-source BibDesk to help us manage the `.bib` file.

3 Data and Methods

For this study, two primary datasets were utilized—Open Secrets donation data and Union Scores for the year 2022. The Open Secrets data underwent cleaning to isolate donors contributing over \$1000 to House and Senate members in 2022. The names were standardized, and the 'candidate_party' variable was simplified to 'D' for Democrats, 'R' for Republicans, and 'I' for Independents. A new variable, 'pac_or_individual,' was introduced to categorize PACs ('p') and individuals ('i'). Notably, we integrated two initially separate datasets for individuals and PACs.

We had many difficulties navigating this dataset. First of all, the files were extraordinarily large. Before cleaning, the original individual donations file was 15GB. We also had to use joins and merges to connect disparate datasets containing different information so that we could get the information we wanted. For example, the individual donors dataset did not have candidate names, only their ID numbers. So, we had to cross-reference with another source that corresponded ID numbers to candidate names.

Simultaneously, the Union Scores data for 2022 was processed, encompassing both yearly and lifetime scores. Ultimately, we settled on just selecting lifetime scores. Titles preceding names were removed, and percentage-based score values were converted to a standard format. The House and Senate data frames were merged, and encoding discrepancies between the Union Scores (latin1) and Open Secrets (utf-8) datasets were addressed.

To connect these datasets, we had to conduct a tremendous amount of data cleaning. In particular, we had to individually clean members of the AFL-CIO scorecard, since some members had suffixes, middle names, or nicknames associated to them. This undertaking was difficult and took a long time, but ultimately we wrangled the data the way we wanted it.

We leveraged the `igraph` package in R to construct network graphs, visualizing connections between Congress members based on donor contributions. Originally, we had a bipartite graph: directed connections from donors to Congresspeople. The method we used to project this is that two Congresspeople are connected if they have a mutual donor. The weight of the connection is the

¹There should always be a space before the "(" in a citation date.

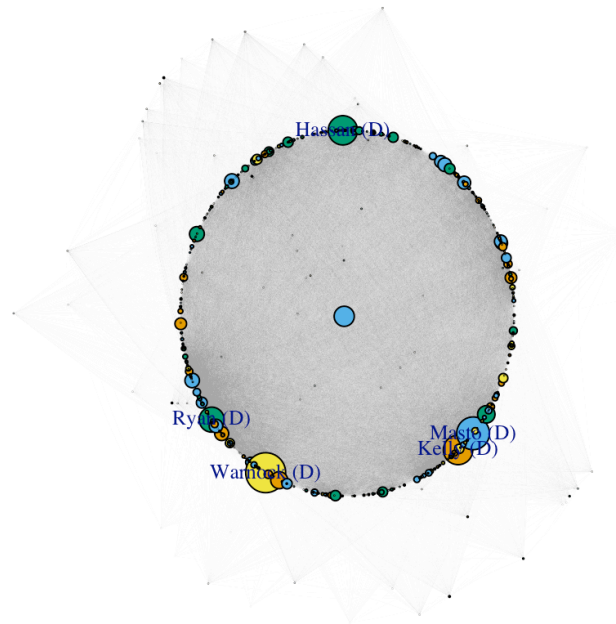
²To cite a paper within parentheses, use, e.g., [moore12].

sum of the contributions of the mutual donor to both candidates. Nodes, representing Congress members, were sized proportionally to the total amount donated.

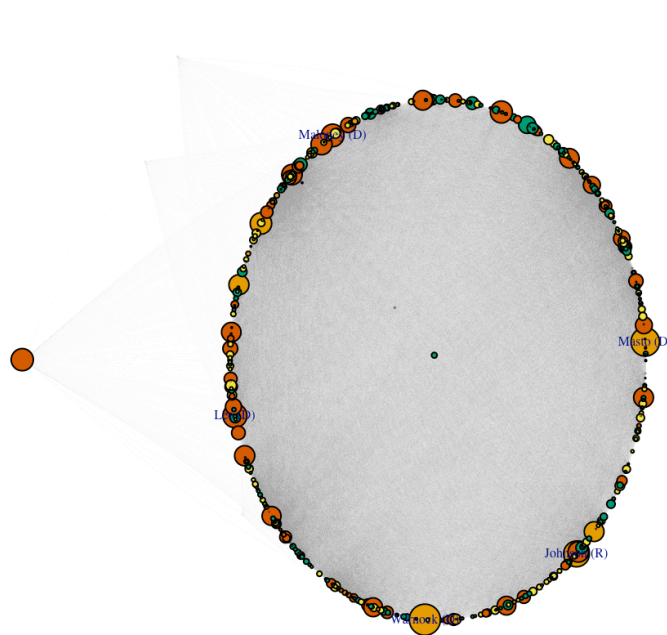
This took an extraordinarily long time to learn, but the results were fascinating. The graphs that we created were extremely complex and with more time, we would learn more methods of representing the information. This is our first time using graph theory, as we were deeply inspired by Dr. Noel, and we used the notes provided to us by him. Moreover, we were not aware of the LGL (large graph layout), which sped up our processing time by orders of magnitude. Originally, when we were running the code to draw these graphs, it took us 15 hours to draw a graph. With LGL, it sped it up into minutes. We also achieved performance speed ups by vectorizing the operations we used, which is demonstrated in the code we have in our GitHub.

A walktrap function was employed to identify groups within the network. We chose a number of steps that provided us with a number of groups that aligned with our expectations (more than 2, but not too many). Many configurations of the steps parameter led to either far too many groups or simply two (Democrats and Republicans). The communities we landed on were then transformed into a data frame, enabling the association of union scores to specific Congress members.

This graphical representation facilitated a comprehensive exploration of relationships within and between these donor groups. The graphs could be a lot cleaner and more thoughtful.



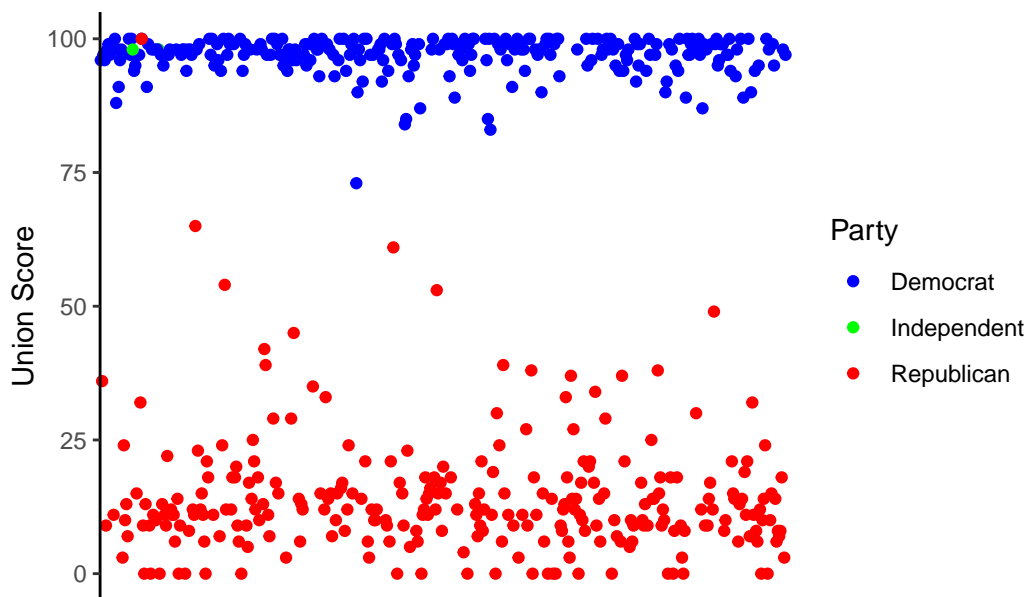
Above is the graph of the individual donation network. The candidates with the most prolific donor connections are highlighted by showing their names. Communities are distinguished by colors.



Above is the graph of the PAC donation network. The candidates with the most prolific donor connections are highlighted by showing their names. Communities are distinguished by colors.

For statistical insights, we implemented code to perform ANOVA on the identified communities, assessing differences in mean Union Scores. We also conducted additional analyses, such as graphical representations of score and donation distributions.

Congressional Union Scores



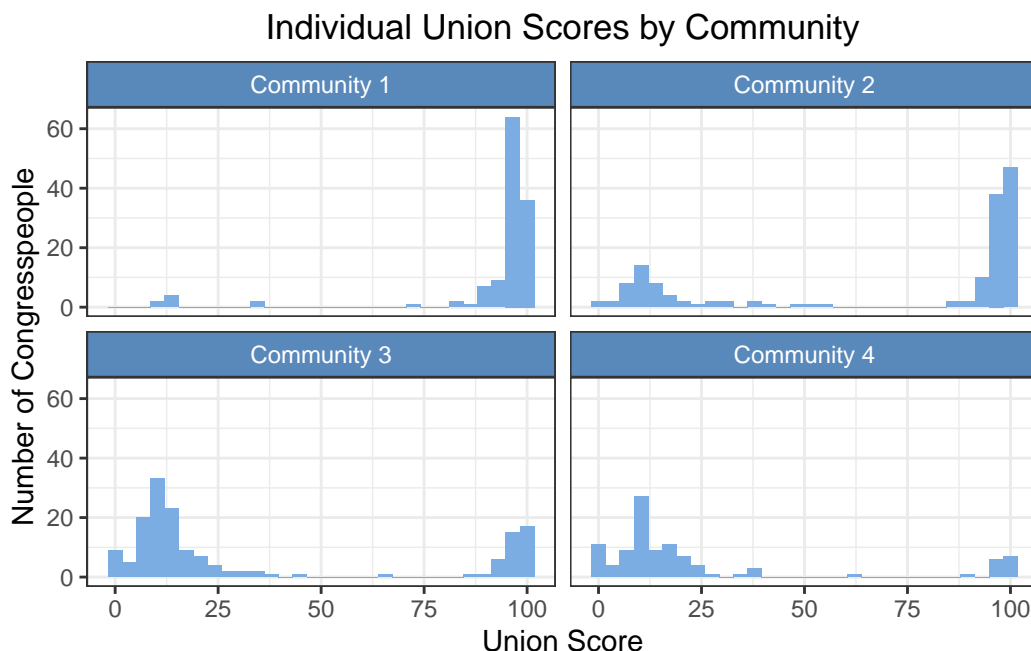
4 [Our Results Section Title Here]

After organizing communities by donor networks for both large individual donors and large PAC donors, we tested the mean union scores of the communities using an Analysis of Variance (ANOVA) Test. Using an ANOVA depends on certain assumptions being fulfilled.

The ANOVA Assumptions are that 1. Each community has a normal population distribution of union score, 2. The union scores of congresspeople are independent, 3. The union scores between communities have an equal variance.

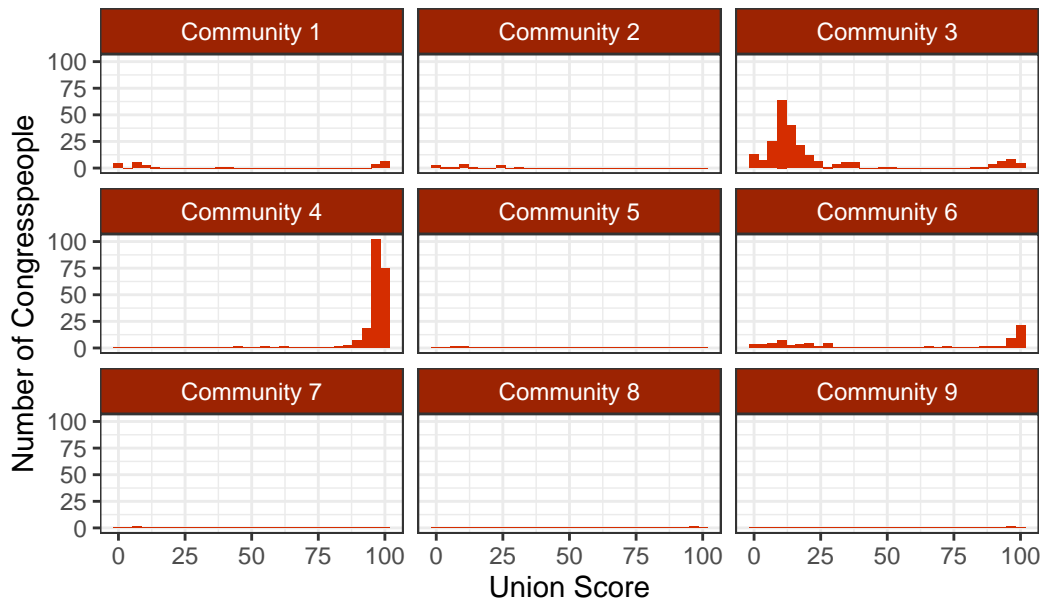
We begin by addressing these assumptions.

First, we examine the union score histograms of the large individual donor communities. Community 1 and Community 4 both seem to be unimodal and normally distributed, however, both communities have some skew. Community 2 and Community 3 are both bimodal with peaks on both ends of the union score spectrum. They do not seem to fit the normality assumption very well.



Now, we examine the union score histograms of the large PAC donor communities. First, note that Community 2, Community 5, Community 7, Community 8, and Community 9 are all extremely small, between 1 and 10 members each. For that reason, we drop those communities from the analysis, as they are too small to gain meaningful insight. Community 3 and Community 4 are both unimodal and have some light skew. Meanwhile, Community 6 and Community 1 are more evenly distributed. Like the individual section, these do not seem to fit the normality assumptions very well.

PAC Union Scores by Community

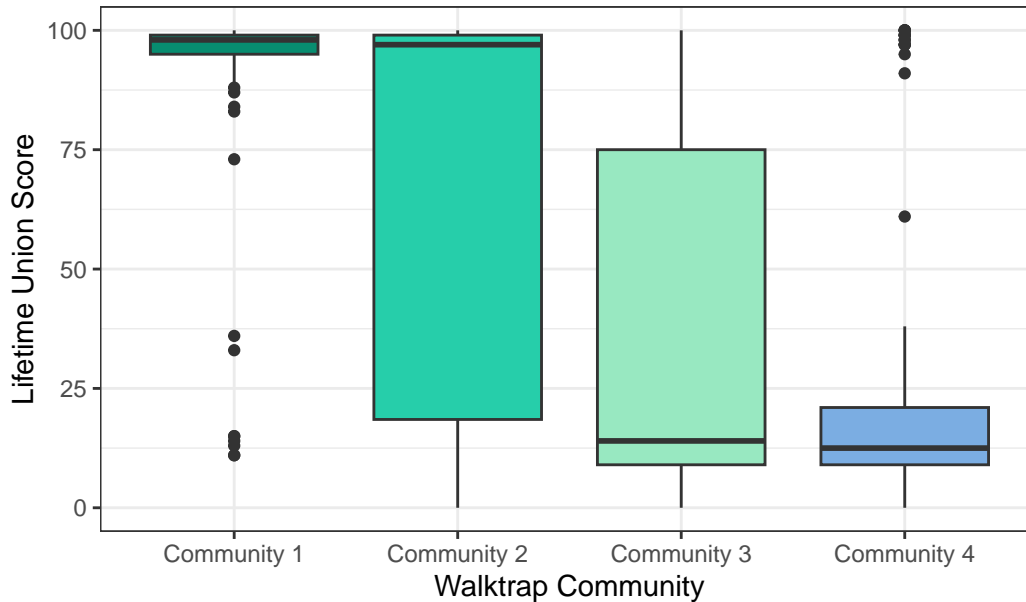


Even though the communities have skewed distributions, we treat them as normally distributed.

In regards to the independence of union scores, it is very unlikely that they are independent. Because of the nature of Congress, voting tends to occur in coalitions and by party line. However, to perform this analysis, we can assume that the observations are independent.

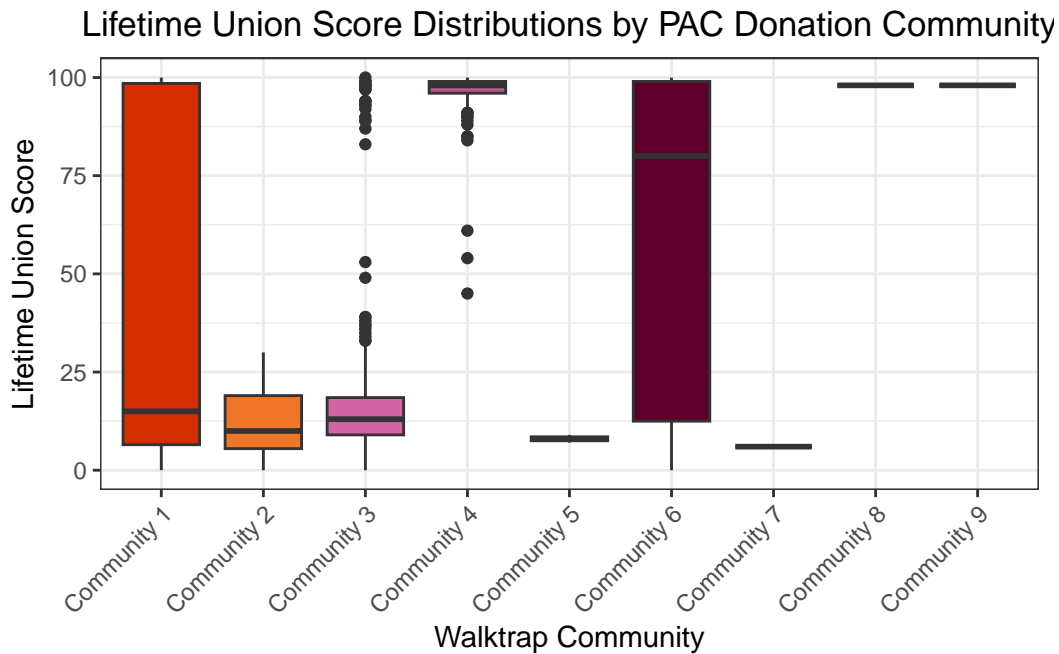
Now, we examine the variances of the communities using boxplots.

Union Score Distributions by Individual Donation Community



In the boxplots of large individual donation communities, we can see that the variances of Community 1 and Community 4 seem to be approximately equal. The variances of Community 2 and Community 3 also seem to be approximately equal. However, the variances of Community 1 and

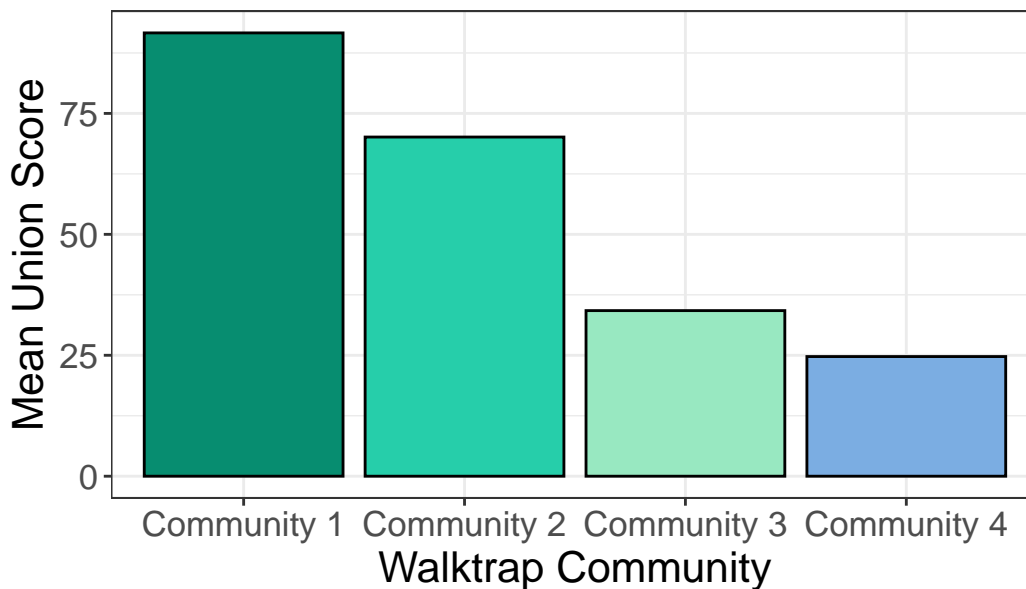
Community 4 are not equal to Community 2 and Community 3.



In the boxplots of large PAC donations communities, we find that Community 2, Community 3, and Community 4 have similar variances. We also find that Community 1 and Community 6 have similar variances. Also, Community 5, Community 7, Community 8, and Community 9 have similar variances. Again, the variances between these groupings of communities are quite different.

Though the ANOVA assumptions are violated, we still think that the test is appropriate to test the means of these communities. Future research could apply more complex non-parametric tests.

Individual Network Mean Union Scores



Examining the means of the large individual donor communities, we see that the mean union score

of Community 1 and Community 2 is favorable towards unions. Conversely, Community 3 and Community 4 are less favorable towards unions.

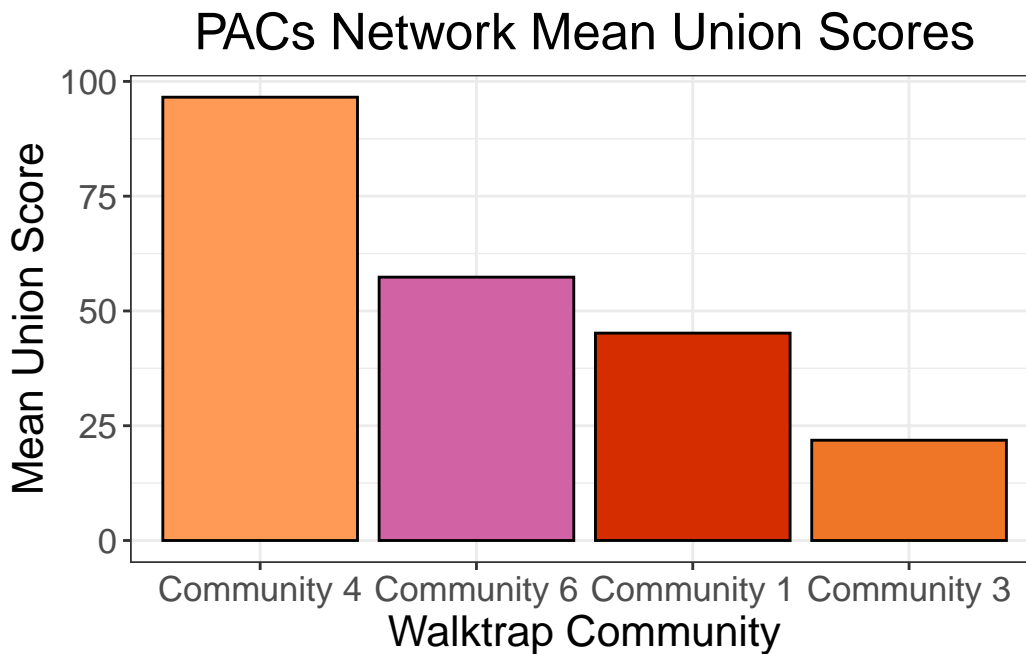
```
individual_anova = aov(lifetime_score ~ community, data = individuals)
summary(individual_anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
community	3	367608	122536	110.3	<2e-16 ***
Residuals	535	594267	1111		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We find that, with F -Value 110.3 and level of significance $p < 0.001$, we reject that the means of these communities are the same. This means that for the distinct communities of congresspeople generated by mutual large individual donors, voting patterns on bills related to unions are distinct. In particular, this shows that structure in relational campaign donation behavior encodes information about the voting patterns of congresspeople towards unions.

Plotting PAC means



```
pacs_anova = aov(lifetime_score ~ community, data = pacs)
summary(pacs_anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
community	3	609386	203129	329	<2e-16 ***
Residuals	520	321038	617		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We find that, with F -Value 329 and level of significance $p < 0.001$, we reject that means of the communities created by PAC donation networks are the same. This means that for the distinct communities of congresspeople generated by mutual large PAC donors, voting patterns on bills related to unions are distinct. In particular, this shows that structure in relational campaign donation behavior encodes information about the voting patterns of congresspeople towards unions.

Between the two groups of donors, these results show that union scores are legitimate ways of distinguishing communities created by large PAC donations and large individual donations networks. This is a meaningful result because it shows that donation patterns are important in finding out information about congressional voting behavior.

5 Discussion

We remind the reader what this paper was about, why it was important, and what we found. We reflect on limitations of the data or methods. If we have specific advice for someone picking up where we leave off, we provide that guidance. We avoid making trite statements like “more research should be done”.

6 References