Project Report: Analysis of Proposal Voting Behavior Using Complex Networks

Abstract

This report presents a structured analysis of voting behaviors among decentralized addresses concerning proposal outcomes. We employ statistical filtering, clustering, regression analysis, and counterfactual analysis to uncover patterns and evaluate the decentralization of influence across the network.

Contents

1 Introduction

This report primarily describes the theoretical analysis methods used in this project and the specific steps involved.

2 Data Filtering

2.1 Objective

To enhance the reliability of the analysis by filtering out inactive addresses that do not significantly influence proposal results.

2.2 Methodology

1. **Define Activity Level**: Each address's activity level is quantified by the total number of votes cast. An address is deemed active if it has participated in at least 5 votes. This threshold is set based on preliminary analyses indicating that addresses with fewer than 5 votes exhibit a negligible impact on proposals.

$$\text{Active} \quad \text{if} \quad V_a \geq 5$$

Where V_a represents the total number of votes cast by address a.

2. **Filter Inactive Addresses**: Addresses failing to meet the threshold are removed from the dataset. This filtering step is crucial to reduce variance in the analysis results and improve the robustness of subsequent findings. The filtering can be expressed as:

$$D_{filtered} = D \setminus \{a \in D \mid V_a < 5\}$$

Where D is the set of all addresses and $D_{filtered}$ is the subset of active addresses.

3 Cluster Analysis

3.1 Objective

To identify communities of addresses with similar voting behaviors and analyze the structure of the voting network.

3.2 Methodology

1. **Modeling the Network**: Each address is represented as a node in a graph. An edge between two nodes indicates a similarity in their voting behavior. We define the similarity S_{ij} between addresses i and j based on the number of votes they cast identically across proposals.

$$S_{ij} = \sum_{k=1}^{N} \delta(v_{ik}, v_{jk})$$

Where δ is the Kronecker delta function, which equals 1 if the addresses voted the same way on proposal k and 0 otherwise.

2. **Edge Weight Filtering**: To enhance the robustness of our network analysis, we apply a threshold to the edge weights, setting weights below a specified value to zero. This process helps eliminate noise and focus on significant connections:

$$w_{ij} = \begin{cases} S_{ij} & \text{if } S_{ij} \geq T \\ 0 & \text{if } S_{ij} < T \end{cases}$$

Here, T is the threshold for edge weight significance, determined by the distribution of S_{ij}.

3. **Community Detection using Louvain Algorithm**: The Louvain algorithm is employed for community detection, optimizing modularity Q, which measures the strength of division of a network into modules (or communities):

$$Q = \frac{1}{2m} \sum_{(i,j) \in E} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Where:

- A is the adjacency matrix representing edge weights,
- m is the total edge weight of the graph,
- k_i and k_j are the degrees of nodes i and j,
- c_i and c_j denote the communities of nodes i and j.

This algorithm iteratively optimizes the partitioning of the network to maximize modularity, allowing us to identify clusters of addresses that exhibit similar voting patterns.

4. **Output**: After executing the Louvain algorithm, we determine the community each address belongs to, providing insights into clusters of similar voting behavior. The analysis of these communities can reveal underlying patterns of collaboration or alignment among addresses.

4 Logistic Regression

4.1 Objective

To estimate the impact of each address's voting behavior on the final outcome of the proposal, considering the binary nature of the outcomes.

4.2 Methodology

1. **Model Specification**: Given that voting outcomes are binary (1 for approval, 0 for disapproval), we utilize logistic regression to model the probability of a proposal being approved based on the voting results of each address:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Here, p is the probability of a proposal being approved, x represents the voting result of the address (1 for "For" and 0 for "Against"), and β_1 captures the relationship between the predictor and the log-odds of approval.

2. **Log-Odds and Coefficient Interpretation**: In logistic regression, the log-odds, or logit, is defined as:

$$\text{log-odds} = \log\left(\frac{p}{1-p}\right)$$

- **Interpretation of Coefficients**: - β_0: The intercept represents the log-odds of approval when all predictor variables (x) are zero. - β_1: The coefficient for the predictor variable x. A positive β_1 indicates that as x increases (i.e., the address votes "For"), the log-odds of proposal approval increase, suggesting a higher likelihood of approval. Conversely, a negative β_1 indicates that as x increases, the log-odds of approval decrease.

The odds ratio, which provides a multiplicative factor for the odds, is expressed as e^{β_1}: - If $e^{\beta_1} > 1$: The odds of approval increase with a positive vote. - If $e^{\beta_1} < 1$: The odds of approval decrease with a positive vote. - If $e^{\beta_1} = 1$: There is no effect of the address's vote on the odds.

3. **Effect Estimation**: After estimating the logistic regression model, we compute individual effects for each address based on their voting behavior. The estimated coefficients provide insight into how each address's vote influences the overall outcome. The predicted probability of proposal approval can be expressed as:

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

4. **Community Average Effect**: For addresses belonging to the same community, we assign the average effect across all members of that community:

$$\text{Effect}_{community} = \frac{1}{N_c} \sum_{i \in C} \text{Effect}_i$$

Where N_c is the number of addresses in community C, allowing for the aggregation of influence within communities.

5 Counterfactual Analysis

5.1 Objective

To assess the potential impact of reversing individual addresses' votes on the overall proposal outcomes.

5.2 Methodology

1. **Counterfactual Processing**: We perform a counterfactual analysis by reversing each address's vote (e.g., changing "For" to "Against" and vice versa) to simulate how this change could affect the final proposal outcome. This hypothetical scenario allows us to analyze the potential shifts in voting influence:

$$\text{Counterfactual Outcome} = f(\text{Reversed Votes})$$

The significance of this analysis lies in identifying key nodes whose influence might drastically alter outcomes when their votes are changed.

2. **Influence Assessment**: If an address's vote is found to be critical (e.g., determining the outcome of a proposal), reversing it could lead to significant changes in the proposal's acceptance. This highlights the importance of certain nodes in maintaining or disrupting decentralization.

6 Centralization Metrics

6.1 Objective

To evaluate the distribution of effects and assess overall centralization.

6.2 Metrics Defined

1. **Variance (σ^2)**: Measures the dispersion of the effects across addresses:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(effect_i - \mu)^2$$

A higher variance indicates a lower degree of decentralization.

2. **Skewness (γ_1)**: Assesses the asymmetry of the distribution of effects:

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$$

A skewness close to zero indicates a symmetrical distribution, while positive or negative values indicate right or left skew, respectively.

3. **Kurtosis (γ_2)**: Evaluates the tails of the distribution:

$$\gamma_2 = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

Higher kurtosis indicates more extreme values in the distribution, suggesting potential influence concentration.

4. **Gini Coefficient (G)**: A measure of inequality within the distribution of effects:

$$G = \frac{A}{A + B}$$

Where A is the area between the Lorenz curve and the line of equality, and B is the area under the Lorenz curve. A Gini coefficient of 0 indicates perfect equality, while a coefficient of 1 indicates maximum inequality.

5. **Herfindahl-Hirschman Index (HHI)**: A measure of concentration in the distribution of effects:

$$HHI = \sum_{i=1}^{N} (effect_i^2)$$

Higher values suggest greater centralization of influence.

6. **Entropy (H)**: Indicates the diversity of effects:

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

Where p_i is the proportion of each address's effect. Higher entropy indicates a more even distribution of influence.

7. **Coefficient of Variation (CV)**: A relative measure of variability:

$$CV = \frac{\sigma}{\mu}$$

Expresses the standard deviation as a percentage of the mean, providing insight into the relative dispersion of effects.

6.3 Interpretation

By analyzing these metrics over time, we can determine shifts in centralization. For instance, if the variance increases from one period to the next, it may indicate that the influence is becoming more concentrated among fewer addresses, thus diminishing the decentralization of power. In an ideally decentralized scenario, we would expect each address to have an equal effect, resulting in a variance of zero.

7 Improvement

In this report, the model primarily addresses single-choice scenarios, where voting options are limited to "For" and "Against." Through certain data transformations, it can also handle multiple-choice situations, where one or more options can be selected. In this case, we can treat each option as a separate

single-choice scenario, with "For" and "Against" corresponding to each option, thus reverting to the single-choice problem.

Additionally, the network clustering in this model assumes that all addresses are within the same house. However, it is also capable of processing addresses that appear across multiple houses. To achieve this, we can use the count of consistent voting across different houses as edge weights when constructing the similarity matrix.

Regarding flexibility, the model could be improved with more information to adjust the similarity matrix, enhancing clustering outcomes. Moreover, incorporating better metrics to evaluate the divergence or concentration of a given vector (in this case, the effect of each address) could serve as a new centralization indicator for future iterations of the model.

Future work should focus on refining these aspects to increase the model's applicability and robustness.