COMMUNITY DETECTION AND SPREAD OF INFORMATION IN THE US CONGRESS

Felice Francario, Lorenzo Caltran, Sharare Zolghadr

The objective

of this analysis is to understand the **interaction patterns**, **key influencers**, **community structures**, and **information flow** within the Twitter interaction network of the 117th United States Congress House of Representatives

Dataset Description

Twitter Interaction Network for the US Congress

Step	Details
Data Type	Social network connectivity
Data Source	https://github.com/gsprint23/CongressionalTwitterNetwork Members' official Twitter handles were obtained from House Press Gallery.
Data Collection	Twitter API was used to retrieve all tweets by members of Congress.
Time Frame	Tweets were collected between February 9, 2022, and June 9, 2022.
Inclusion Criteria	Only members who issued 100 or more tweets during this period were included in the network.
Total Members Included	Out of 535 representatives, 475 met the inclusion criteria and were included in the Twitter interaction network.

Data Preprocessing

Dataset statistics

Directed	Yes.
Node features	No.
Edge features	Yes.
Nodes	475
Edges	13,289
Isolated nodes	No.

Using Twitter's API V2, **interactions** such as **retweets, quote tweets, replies, and mentions among members** were tracked.

Influence probabilities were calculated by normalizing these interactions against the total tweets issued by each member.

edge weights are empirically obtained "probabilities of influence" between all pairs of Congresspeople.

Methodology

Basic Network Statistics

degree distribution, diameter, and average path length

Centrality Measures

Identify key influencers

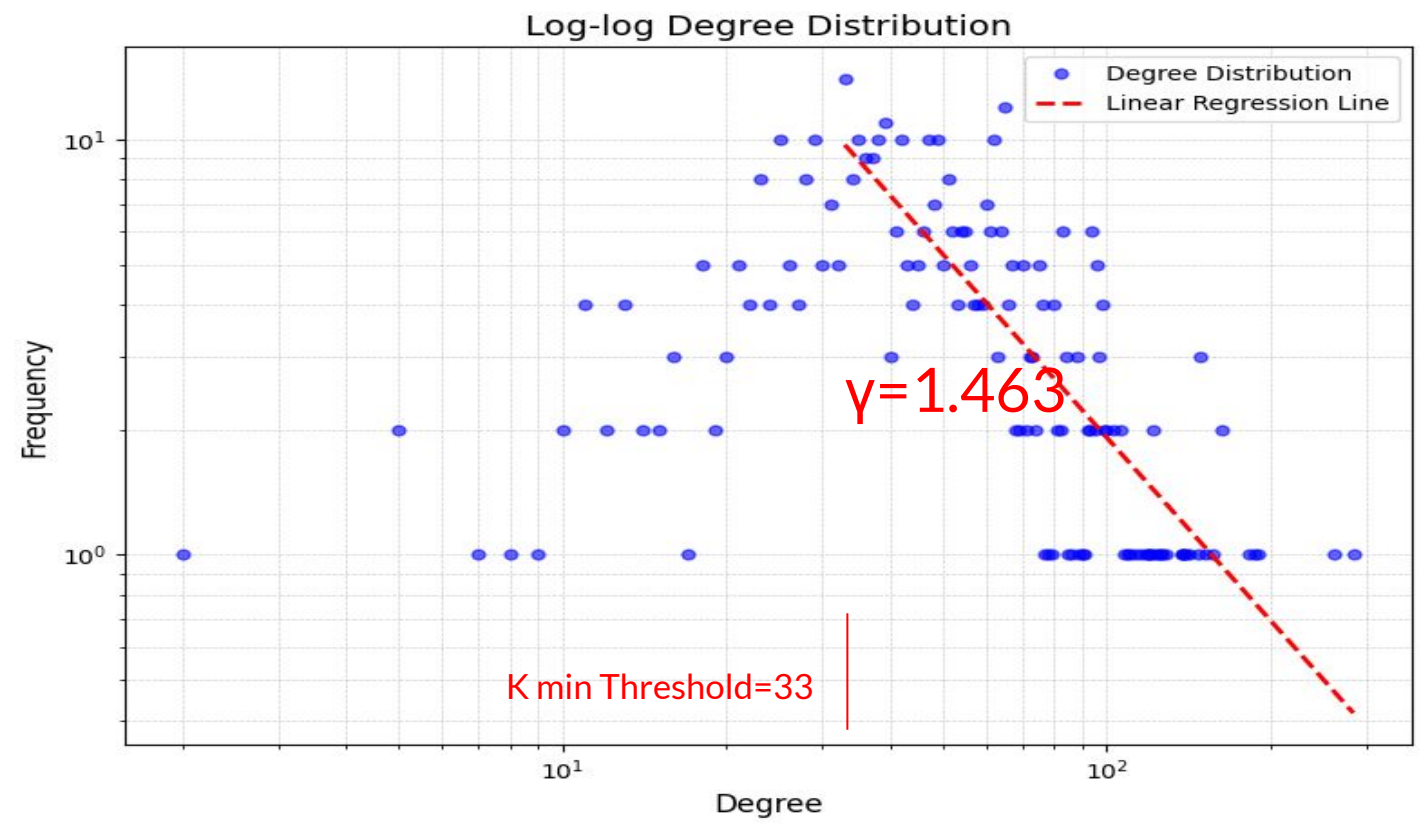
Community Detection

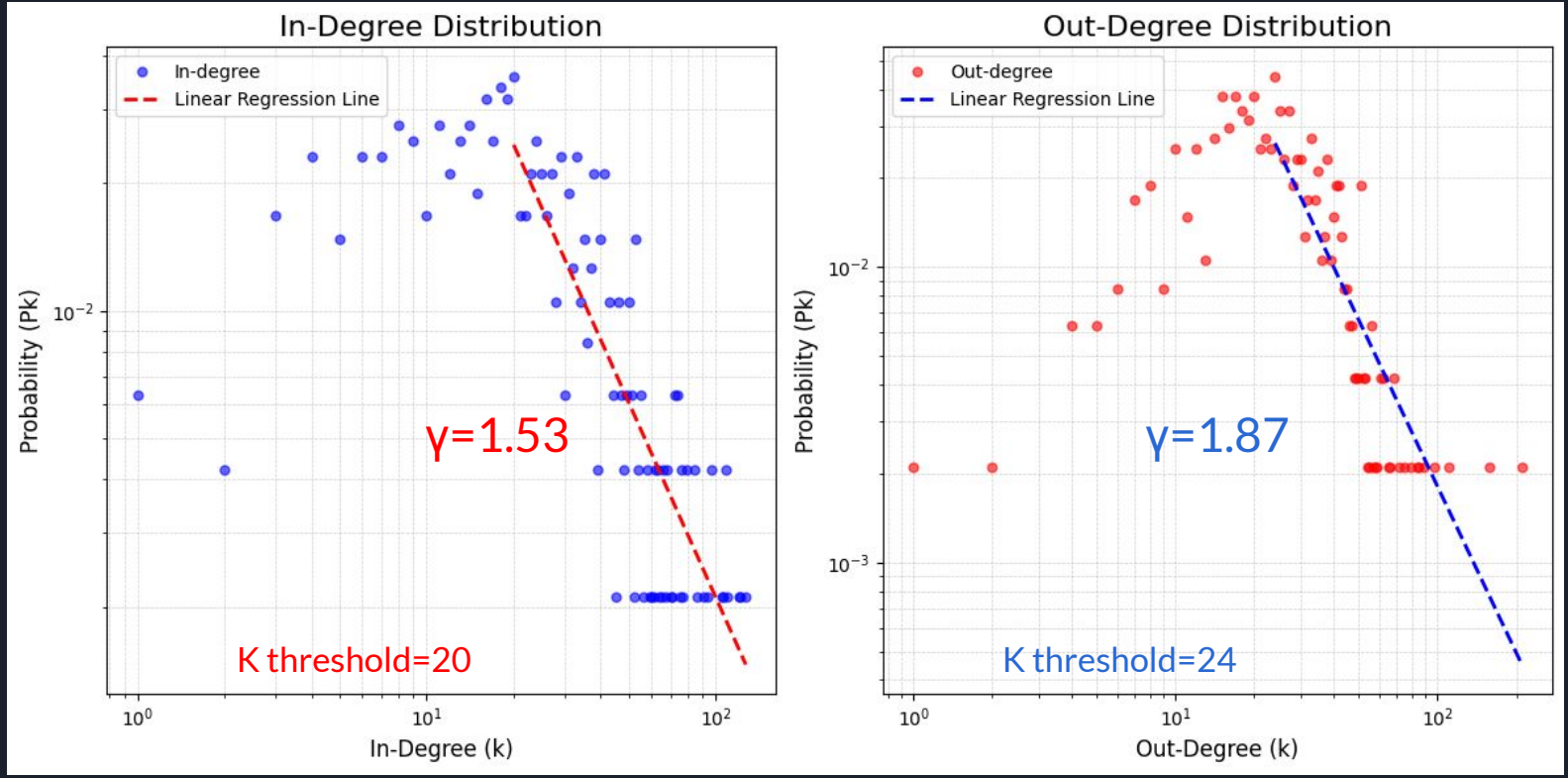
Identify groups of closely interacting members

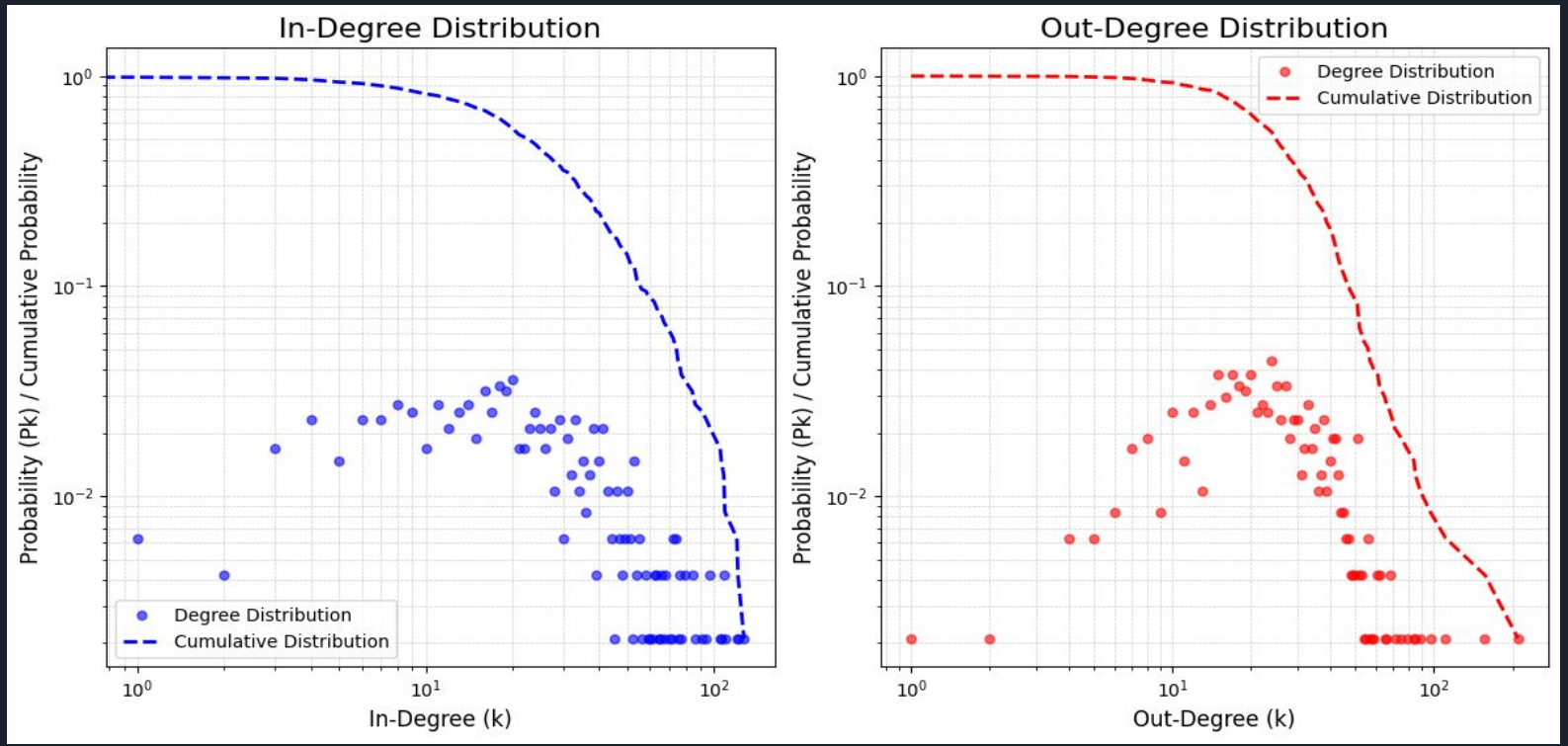
Influence and Information Flow

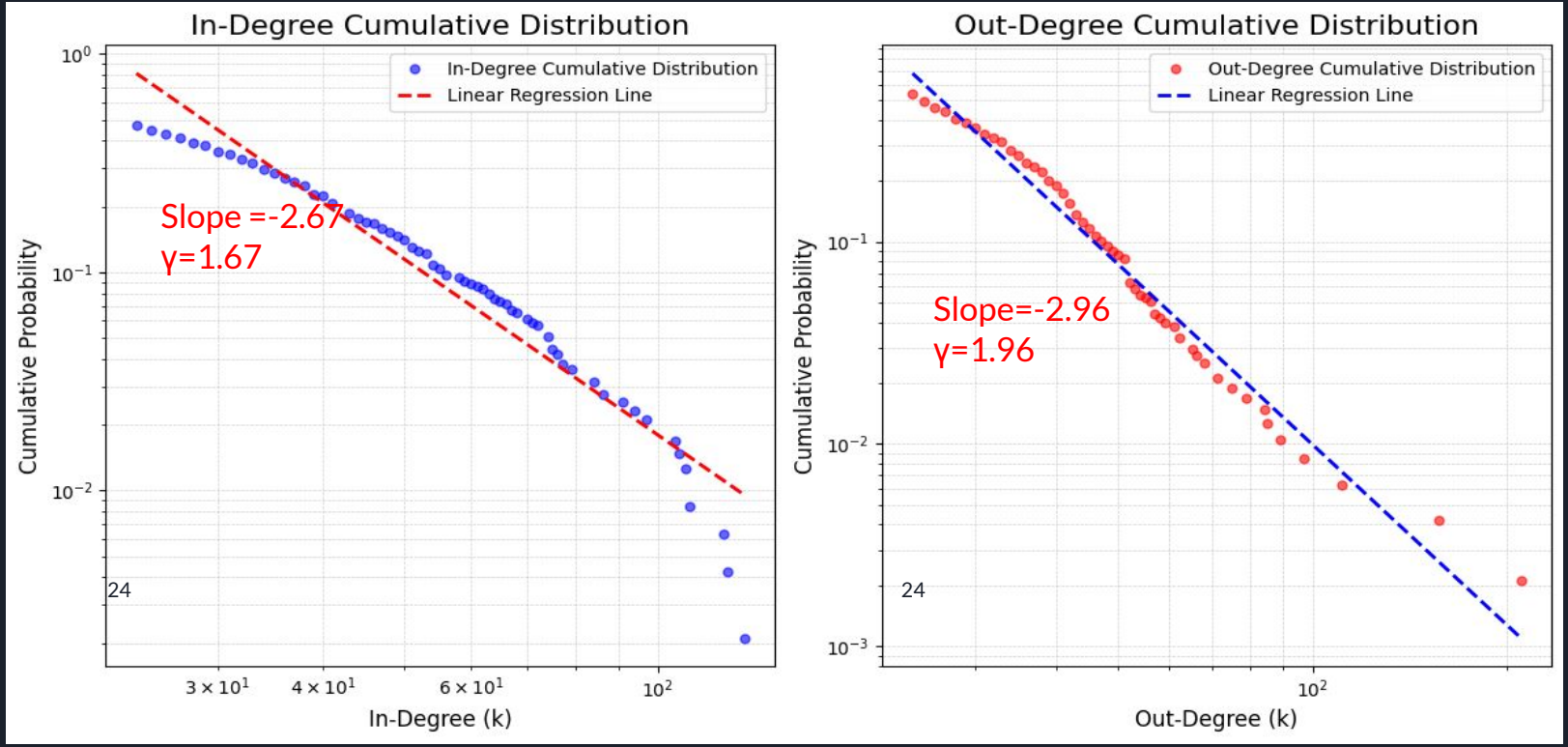
Identify influential members who have a significant impact on the network.

Degree Distribution




In-Degree Distribution
In-degree
Linear Regression Line
Probability (Pk)
10^{-2}
γ=1.53
K threshold=20
10^0
10^1
10^2
In-Degree (k)
Out-Degree Distribution
Out-degree
Linear Regression Line
Probability (Pk)
10^{-2}
10^{-3}
γ=1.87
K threshold=24
10^0
10^1
10^2
Out-Degree (k)


In-Degree Distribution
Probability (Pk) / Cumulative Probability
In-Degree (k)
Degree Distribution
Cumulative Distribution
Out-Degree Distribution
Probability (Pk) / Cumulative Probability
Out-Degree (k)
Degree Distribution
Cumulative Distribution


In-Degree Cumulative Distribution
In-Degree Cumulative Distribution
Linear Regression Line
Slope =-2.67
γ=1.67
Cumulative Probability
In-Degree (k)
24
Out-Degree Cumulative Distribution
Out-Degree Cumulative Distribution
Linear Regression Line
Slope=-2.96
γ=1.96
Cumulative Probability
Out-Degree (k)
24

ML estimate of γ

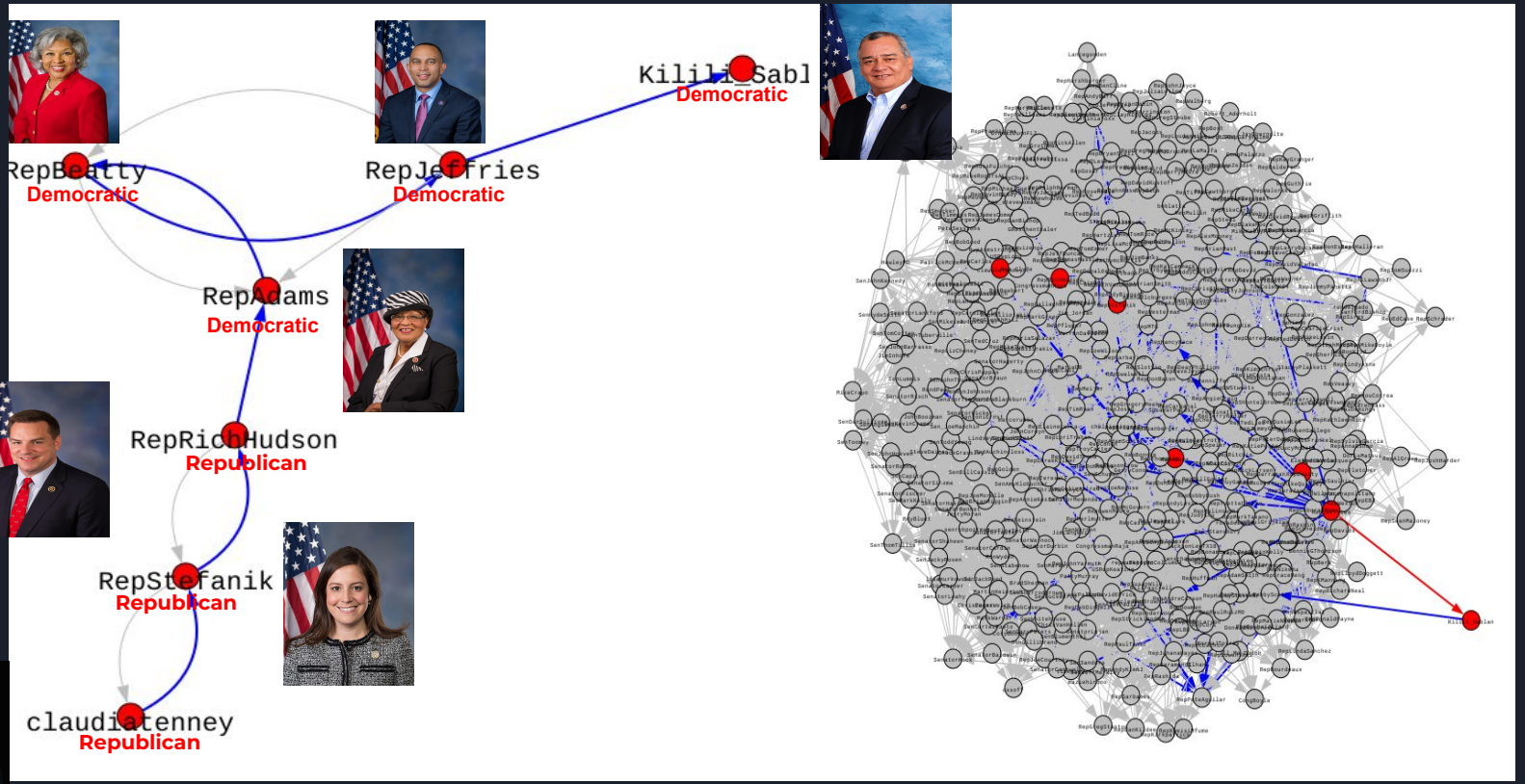
To estimate the exponent γ of the degree distribution using the Maximum Likelihood Estimation (MLE) method, we use the following formula:

$$\gamma = 1 + \frac{n}{\sum_i \ln\left(\frac{k_i}{k_{\min}}\right)}$$

In-degree γ=2.63
Out-degree γ=3.01
These γ values indicate that both distributions are heavy-tailed, but the out-degree distribution decays faster than the in-degree distribution.

Property	Value
Average clustering coefficient	0.2242
Network diameter	6
Average path length	2.3572
Is the graph connected	No
Graph order	475
Graph size	13289
Number of possible edges (N*(N-1))	225150
Graph density	0.0590

Path between most distant nodes



Homophily

The tendency of individuals to associate and bond with similar others.

Assortivity coefficient: $$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

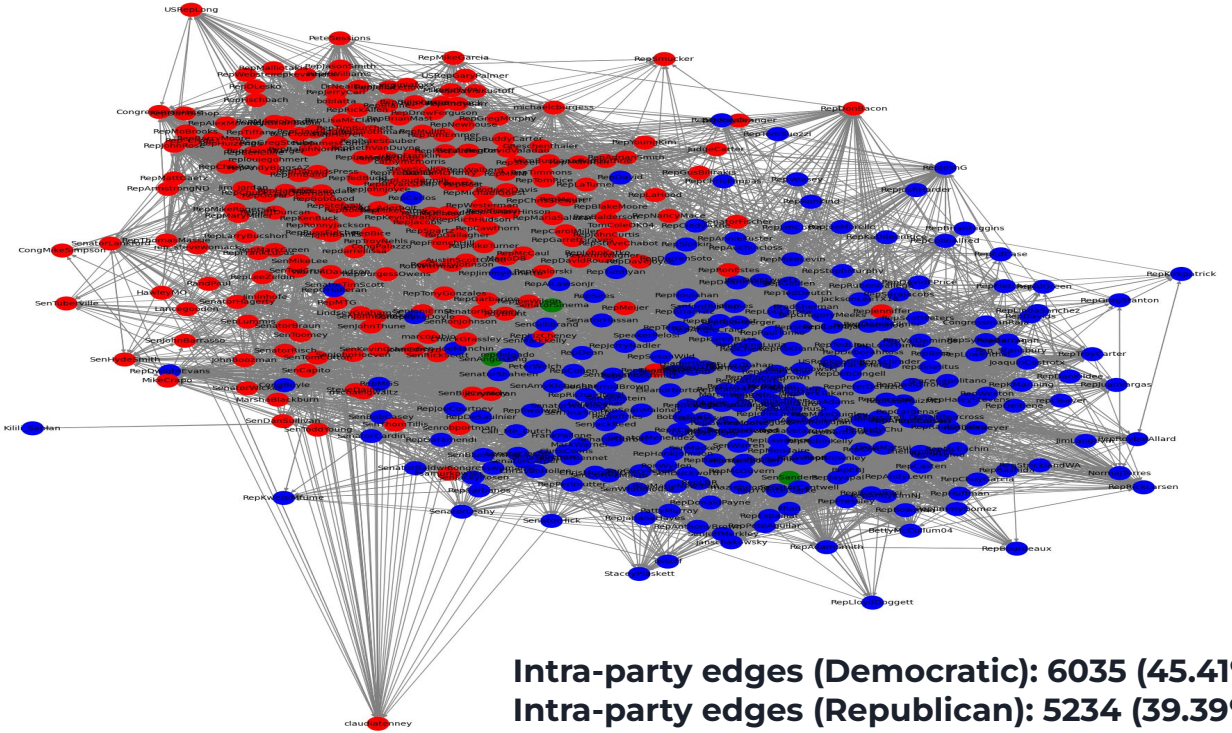
Attribute assortativity based on political party= 0.697

- Strong Positive assortivity
- High Homophily

Implications:

- Political Polarization
- Information flow

INTERACTIONS BETWEEN THE PARTIES



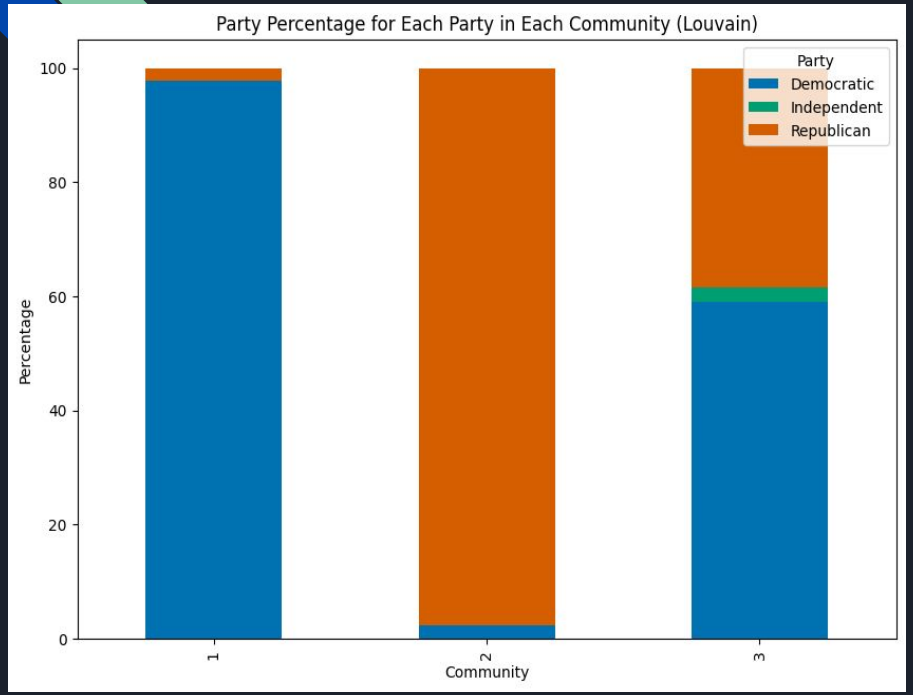
Intra-party edges (Democratic): 6035 (45.41%)
Intra-party edges (Republican): 5234 (39.39%)
Inter-party edges: 2020 (15.20%)

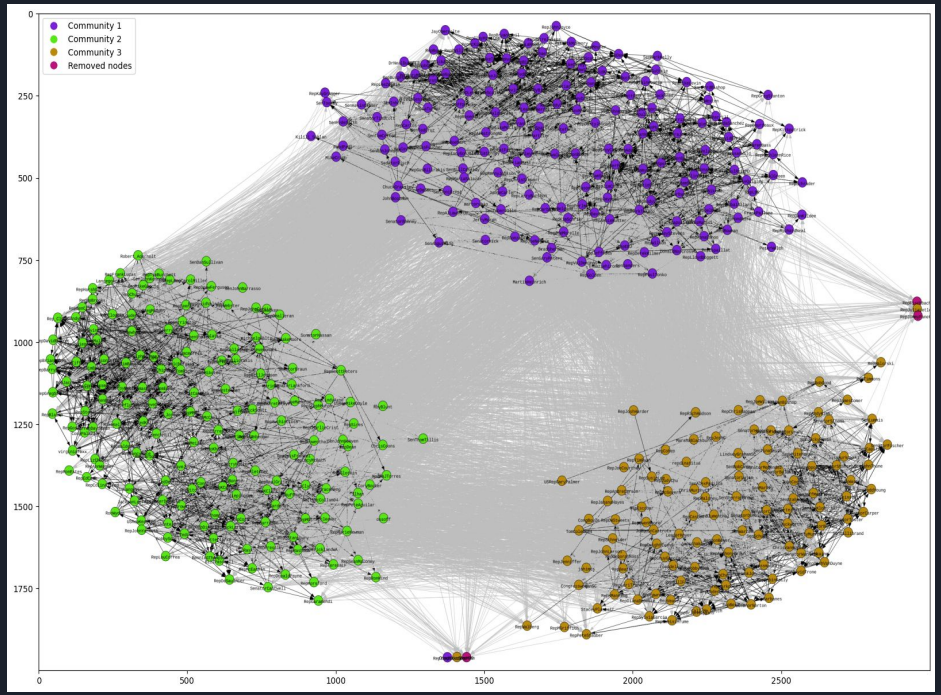
COMMUNITY DETECTION

In this section we implemented the following community detection algorithms:

1. Louvain
2. Infomap
3. Girvan-Newman (to detect hierarchical community structure)
4. Spectral clustering with nearest neighbor similarity
5. Spectral clustering with Gaussian similarity
6. Spectral clustering with cosine similarity

LOUVAIN ALGORITHM

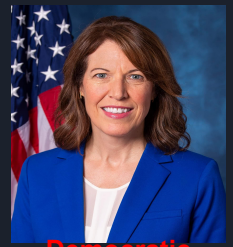




The algorithm identifies 4 communities, after removing one smaller community (3 nodes) we can observe 3 main communities and there appears to be a strong correlation between the community structure and the political party.

Isolated Communities (three nodes)

- Rep Cindy Acne
- Sanford Bishop
- Rip Chris Pappas

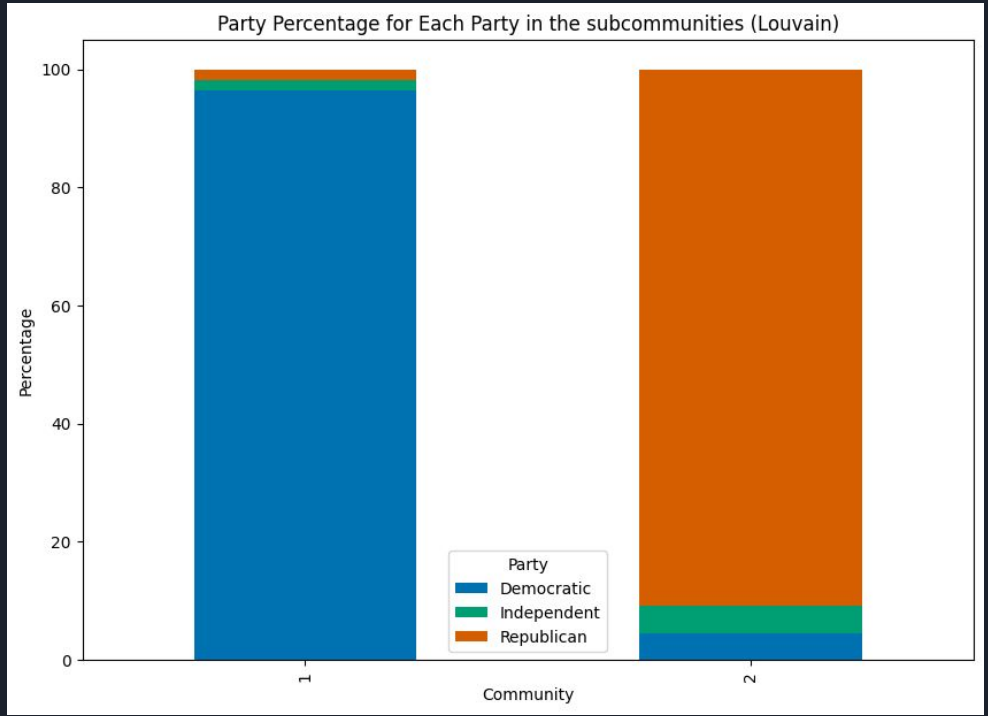

Democratic


Democratic

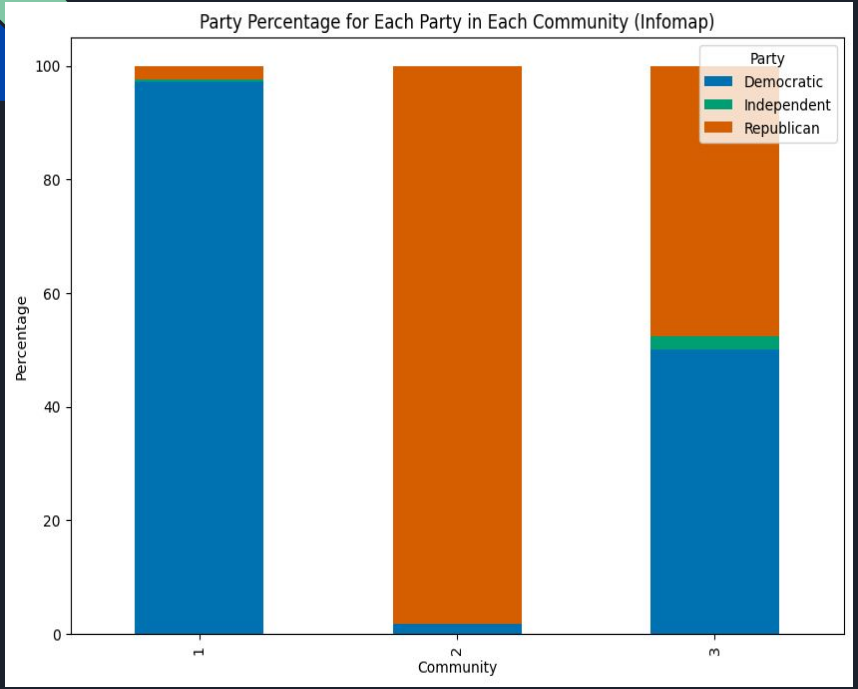

Democratic

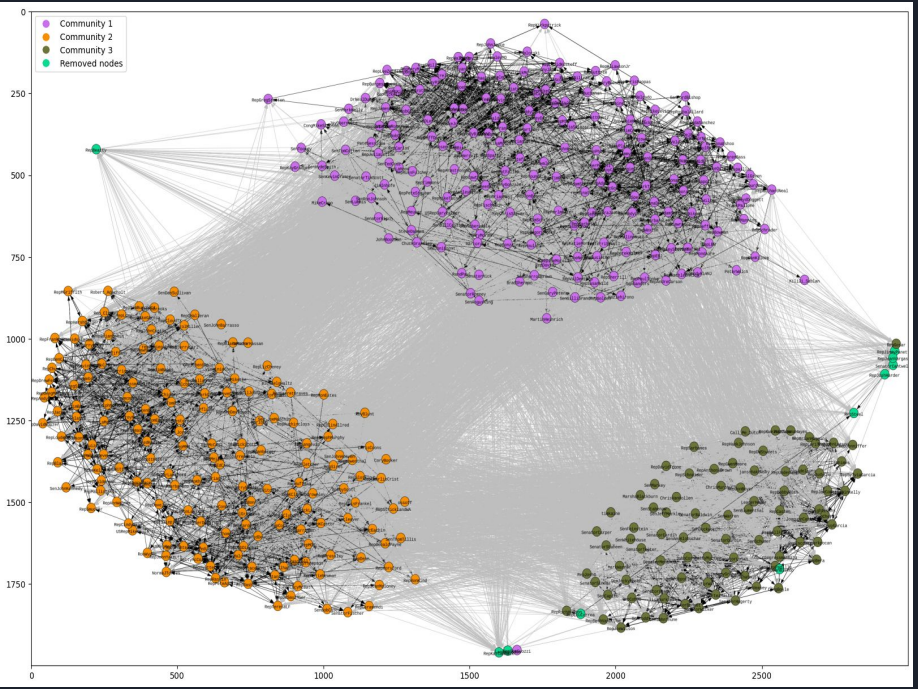
SUB-COMMUNITIES ANALYSIS

We applied Louvain algorithm to the last community, to understand the types of interactions between the two political parties inside the mixed community.



INFOMAP ALGORITHM



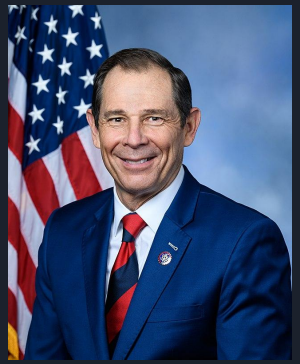


The Infomap algorithm yields 8 communities, 5 smaller modules were removed (one composed by 6 nodes and the remaining ones composed by 1), resulting in 3 main communities

Isolated Communities

- RepLy Cheney (Republican)
- Rep John Curtis (Republican)
- RepLy Granger (Republican)
- RepElaineLuria (Democratic)
- RepO Halleran (Democratic)
- Rip Chris Pappas (Democratic)
- RepTomSuozzi (Democratic)
- RenAn Wagner (Republican)
- Rep Joe Wilson (Republican)
- Rob Wittmann (Republican)

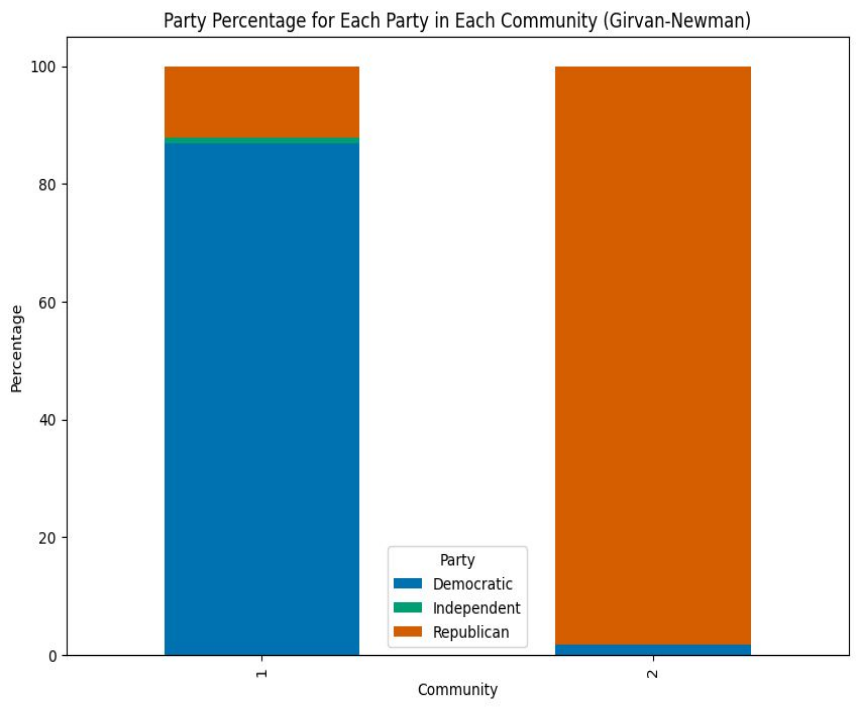
LIZZ CHENEY
Republican

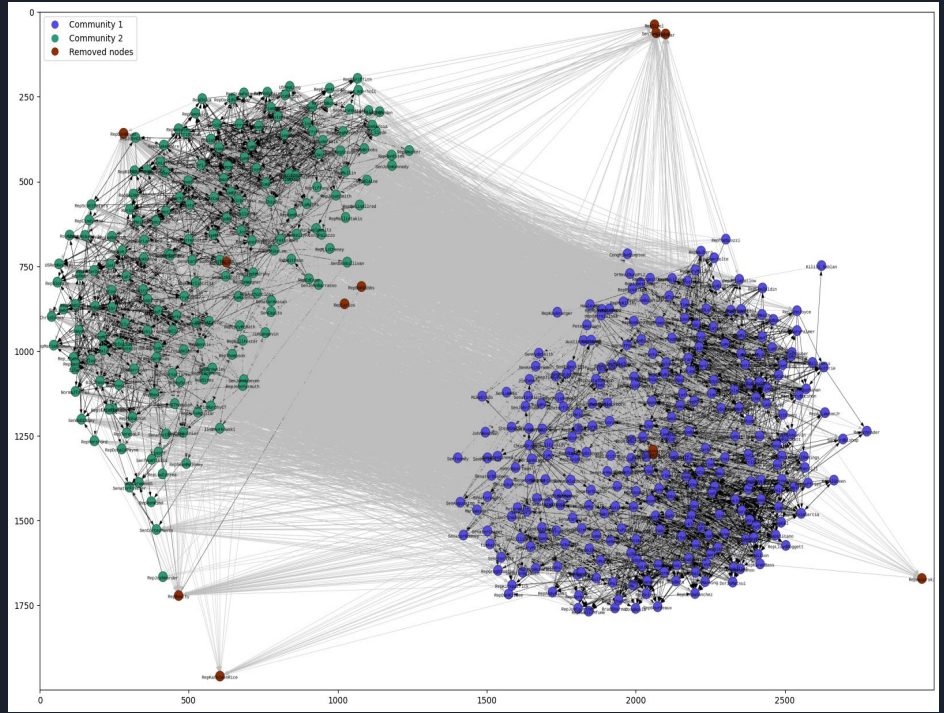


JOHN CURTIS
Republican

KAY GRANGER
Republican

GIRVAN-NEWMAN ALGORITHM





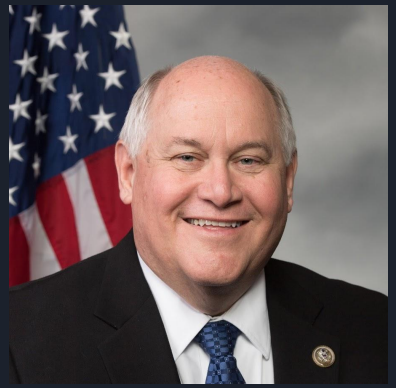
The Girvan-Newman algorithm results in 14 communities, but most of them are single-node communities, in fact the main communities are only two.

Isolated Communities

- SenDen Sullivan (Republican)
- RepEdCase (Democratic)
- RepRonEstes (Republican)
- RepLay Granger (Republican)
- Rep Josh Harder (Democratic)
- Rep Kirkpatrick (Democratic)
- RepAlLawsonJr (Democratic)
- RepO Halleran (Democratic)
- Kilili Sablan (Democratic)
- Rep Schrader (Democratic)
- Con Mike Simpson (Republican)
- RepTomSuozzi (Democratic)

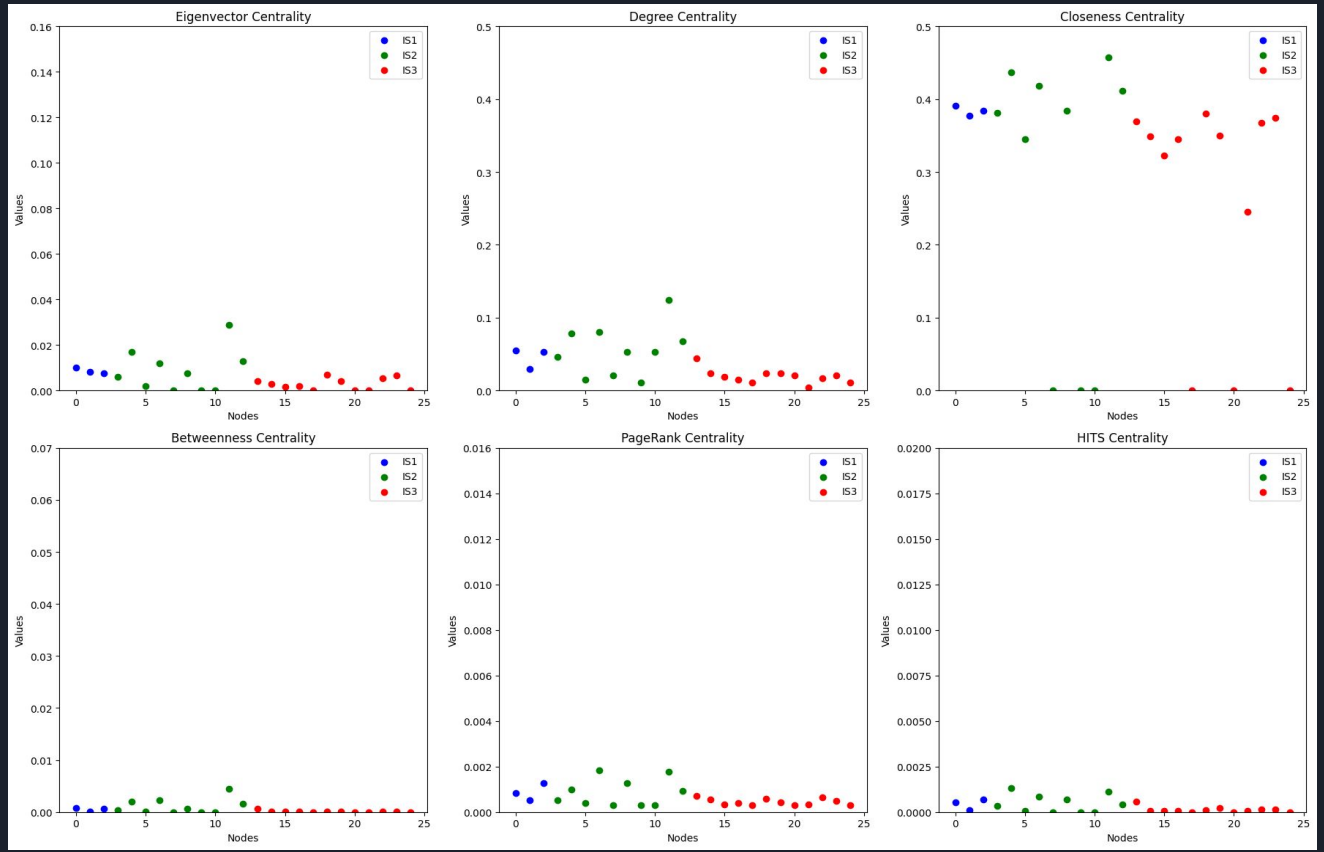
DANIEL SULLIVAN
Republican

ED CASE
Democratic



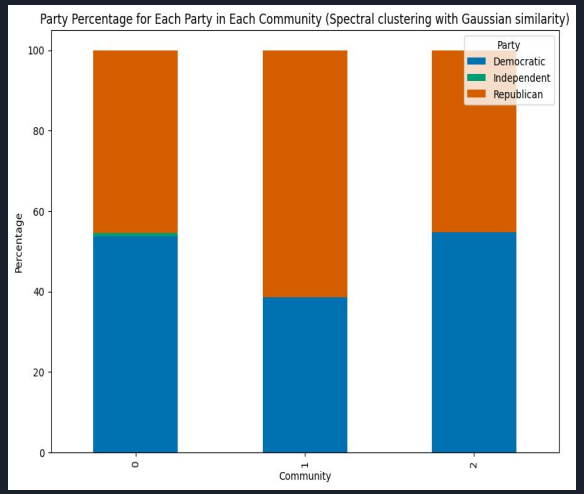
RON ESTES
Republican

Centrality Measurements for Isolated Nodes after applying different community detection methods (Louvain, Infomap, Girvan-Newman)

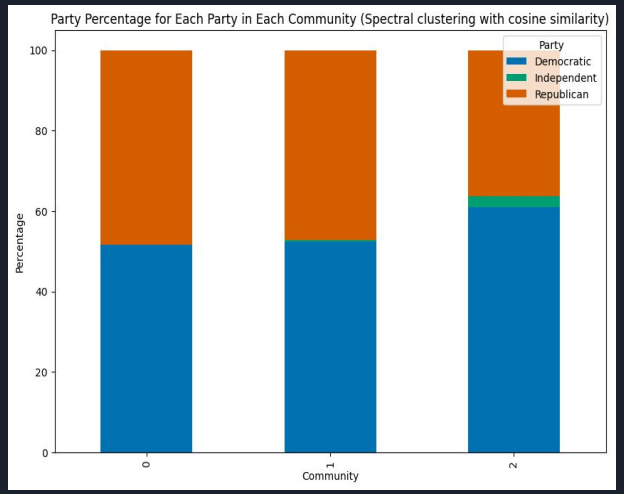


SPECTRAL CLUSTERING

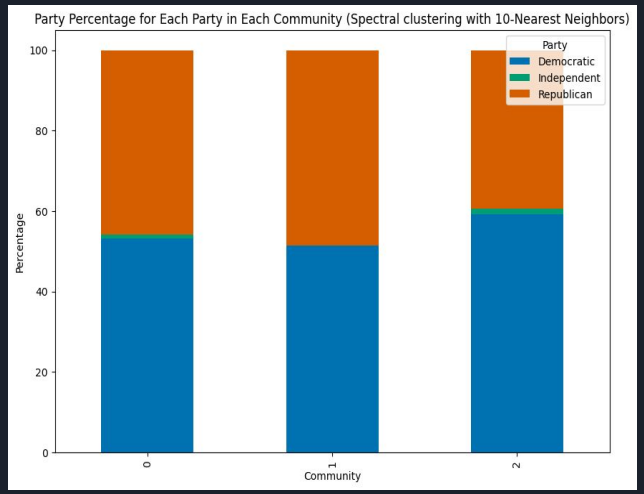
Unlike the previous algorithms, the communities yielded by spectral clustering don't show correlation with the political parties



Gaussian similarity

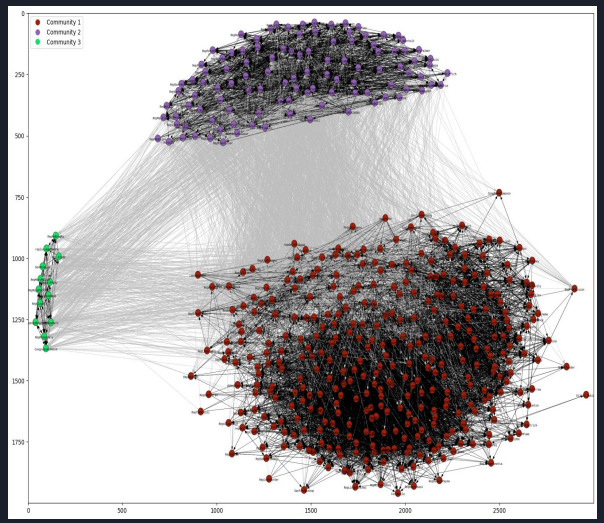


Cosine similarity

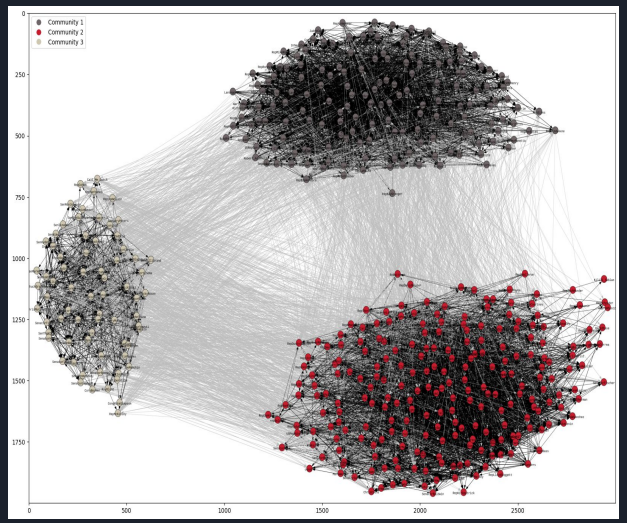


Nearest neighbors (k=10)

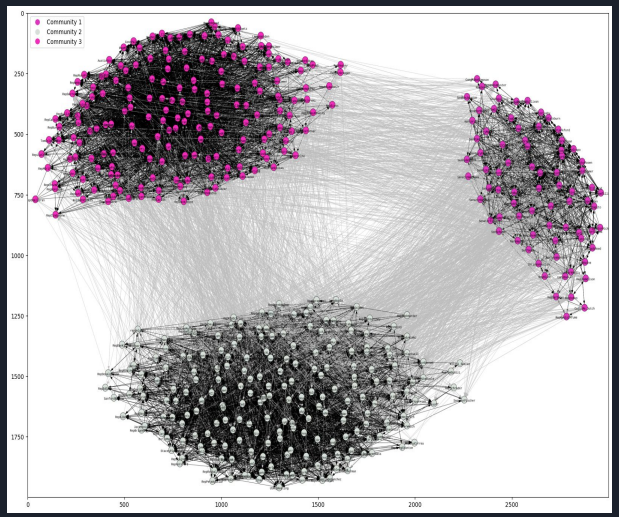
SPECTRAL CLUSTERING VISUALIZATION



Gaussian similarity



Cosine similarity



Nearest neighbors (k=10)

COMMUNITY DETECTION RESULTS

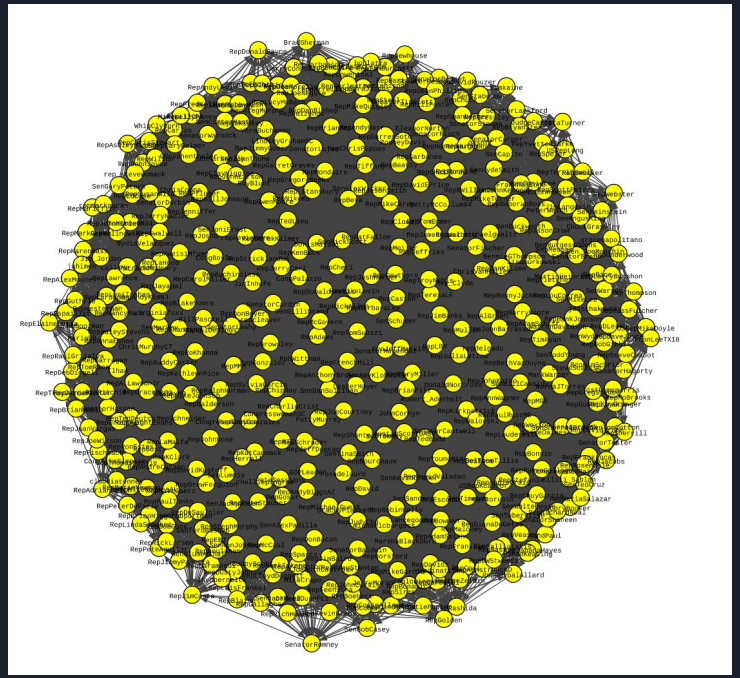
Algorithm	N of communities	Modularity	Map equation
Louvain	3	0.432480	9.273039
Infomap	3	0.055118	7.962810
Girvan-Newman	2	0.380121	6.378135
SC (Gaussian similarity)	3	0.008578	7.978734
SC (Cosine similarity)	3	0.031200	8.512298
SC (10-Nearest Neighbors)	3	0.027824	8.482418

SMALL WORLD

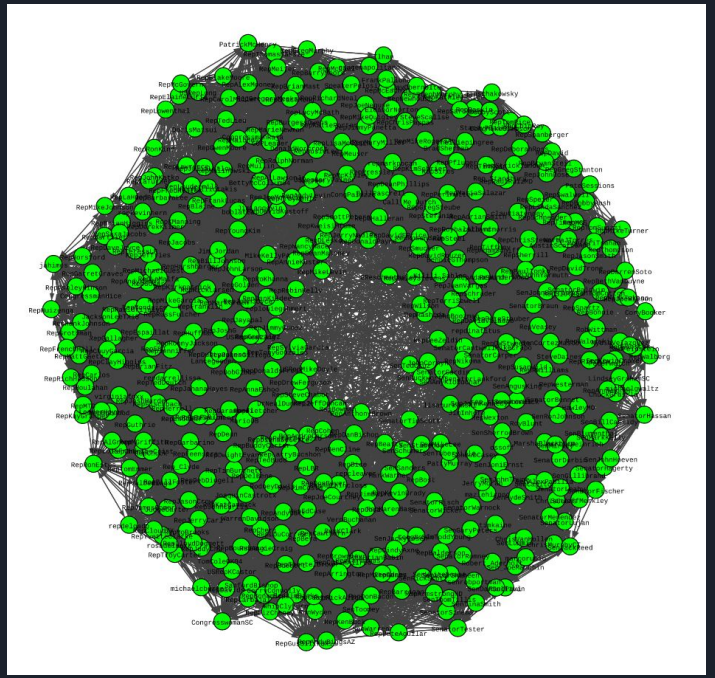
Our graph exhibits higher clustering than a random graph and lower than the Watts-Strogatz model and higher average shortest path length than the Erdős–Rényi graph and than the Watts-Strogatz model

Graph	Clustering coefficient	Average Shortest Path Length
US Congress Twitter	0.22422040490798673	2.3548922056384742
Erdős–Rényi graph	0.11621478188367941	2.121394625805019
Watts-Strogatz graph	0.5535555918084096	2.094639129469243

SMALL WORLD VISUALIZATION



Erdős–Rényi graph



Watts-Strogatz graph

SMALL WORLD METRICS

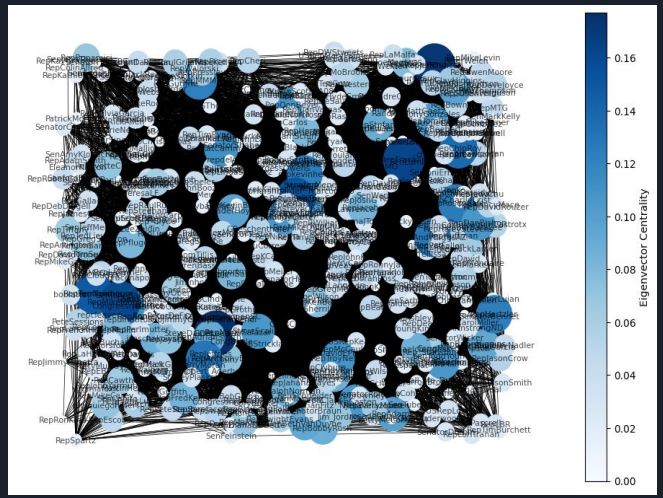
The network is pretty clustered and exhibits small-world properties to a degree, but it is not as efficient in terms of average path lengths compared to an ideal small-world network. The alternative metric indicates that the network does have significant small-world characteristics, combining high clustering with efficient path lengths, though it is not a classic small-world network.

Metric	Value
Normalized Clustering Coefficient (γ)	0.30063371314162046
Normalized Path Length (λ)	2.0628197997775306
Small World Index (σ)	0.1457392028009635
Alternative Metric (ω)	0.7558976848812724

CENTRALITY MEASURES

In this section we analyzed the following centrality measures:

1. Eigenvector centrality
2. Closeness centrality
3. Betweenness centrality
4. Degree centrality
5. PageRank
6. HITS
7. Viral centrality

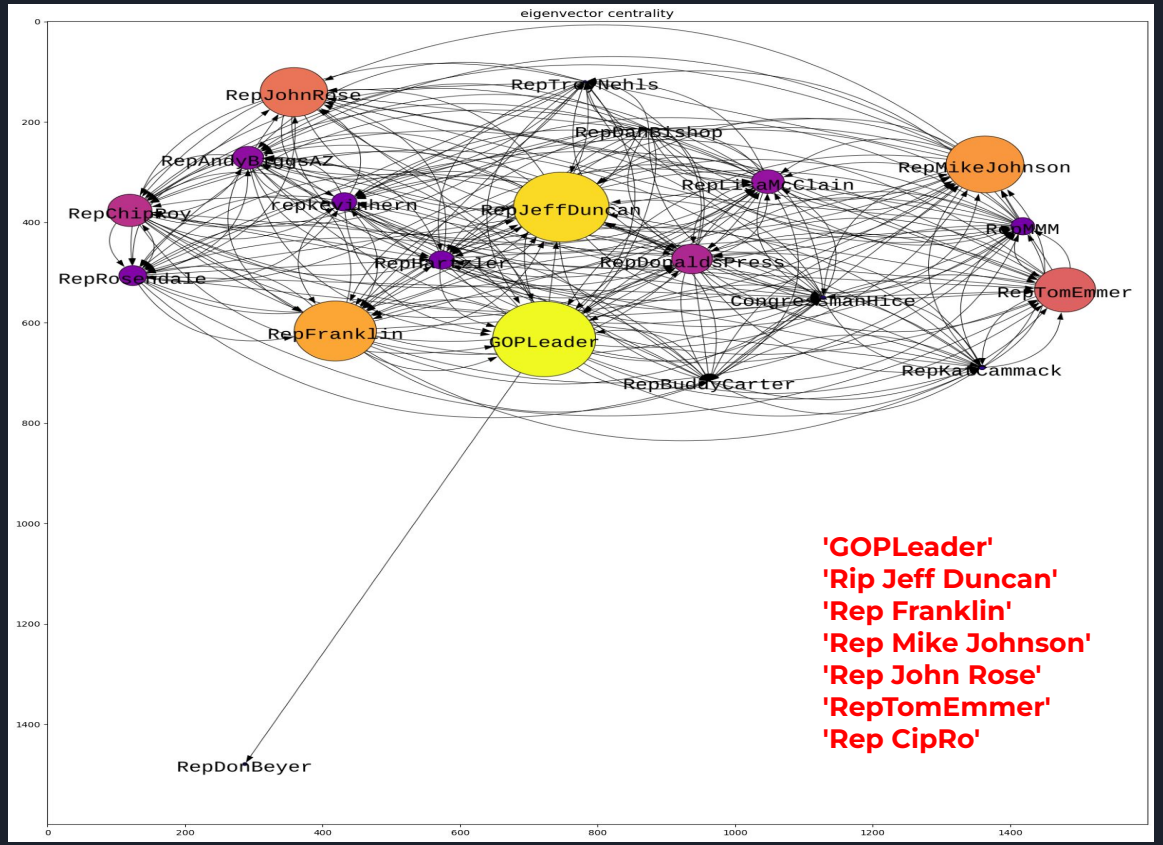


EIGENVECTOR CENTRALITY

MITCH MCCONNELL (GOP LEADER)
Republican

JEFF DUNCAN
Republican

SCOTT FRANKLIN
Republican

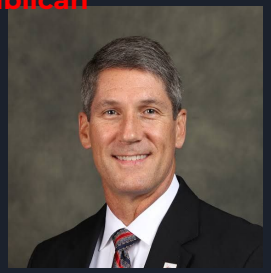


'GOPLeader'
'Rip Jeff Duncan'
'Rep Franklin'
'Rep Mike Johnson'
'Rep John Rose'
'RepTomEmmer'
'Rep CipRo'

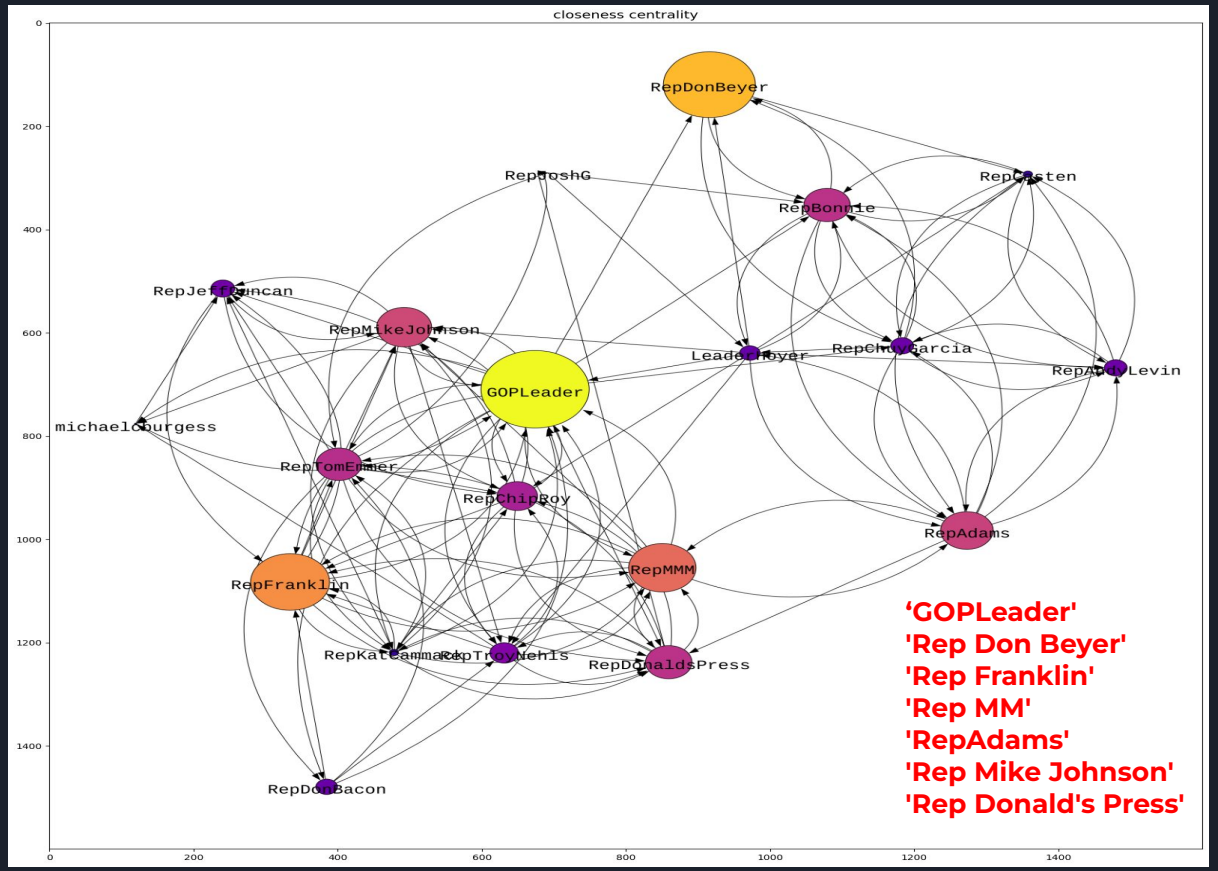
CLOSENESS CENTRALITY

MITCH MCCONNELL
Republican

DON BEYER
Democratic



SCOTT FRANKLIN
Republican

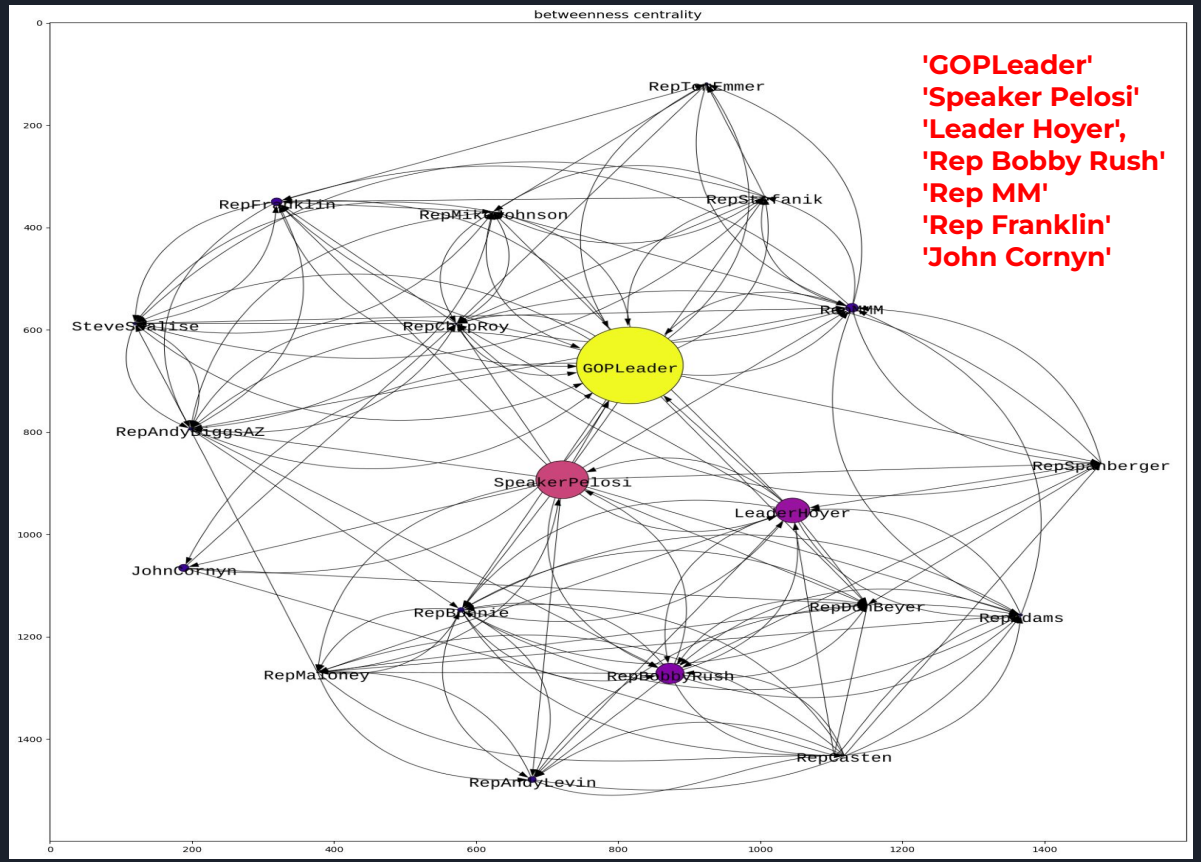


Betweenness Centrality

MITCH MCCONNELL
Republican

NANCY PELOSI
Democratic

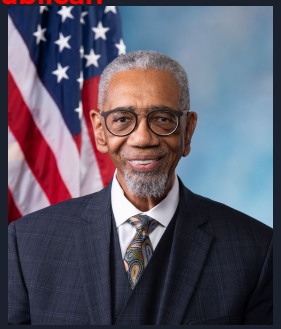
STENY HOYER
Democratic



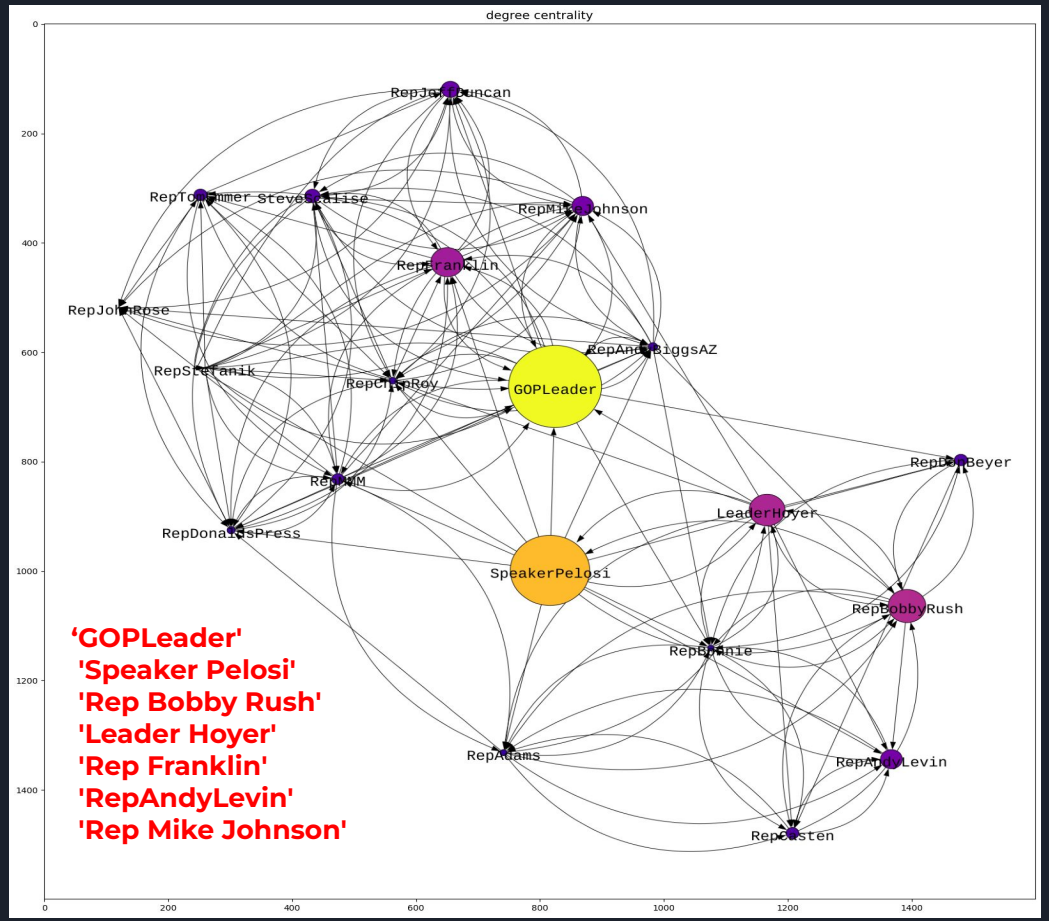
Degree Centrality

MITCH MCCONNELL
Republican

NANCY PELOSI
Democratic



BOBBY RUSH
Democratic

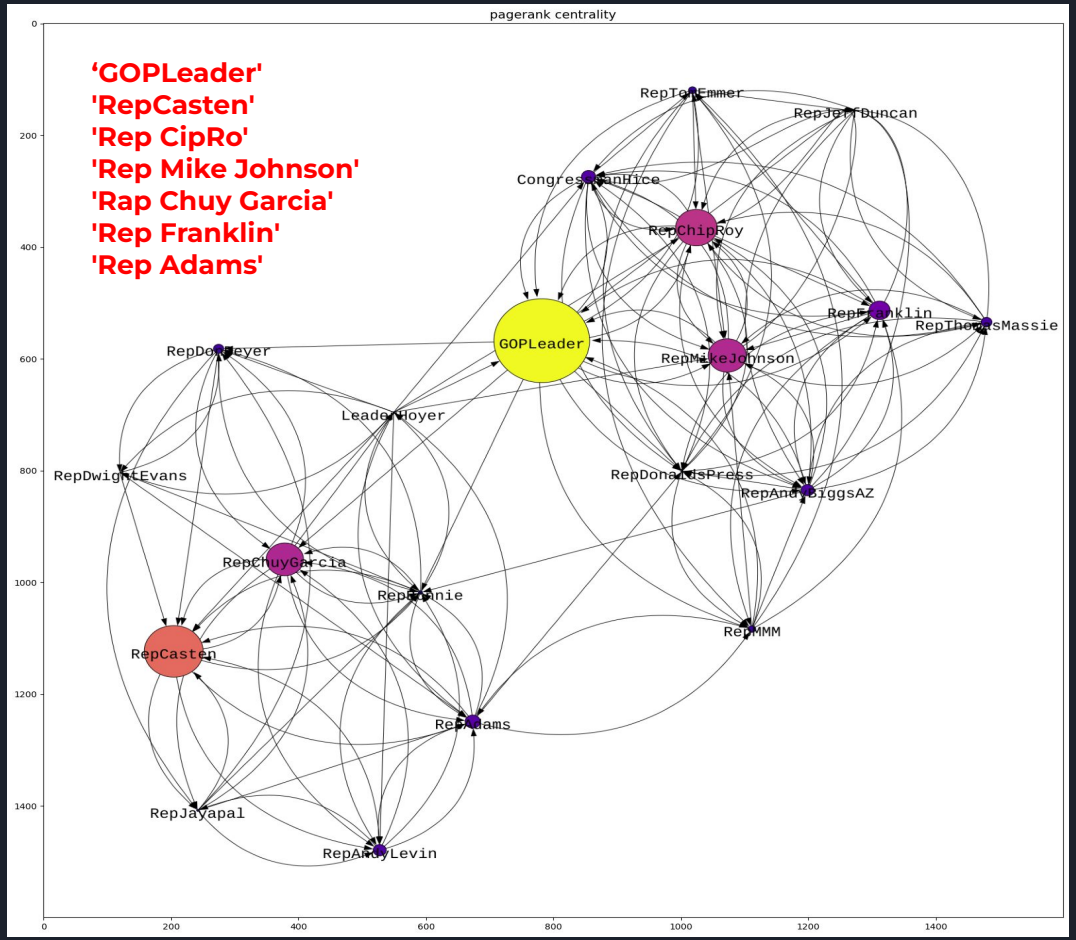


PageRank

MITCH MCCONNELL
Republican

SEAN CASTEN
Democratic

CHIP ROY
Republican

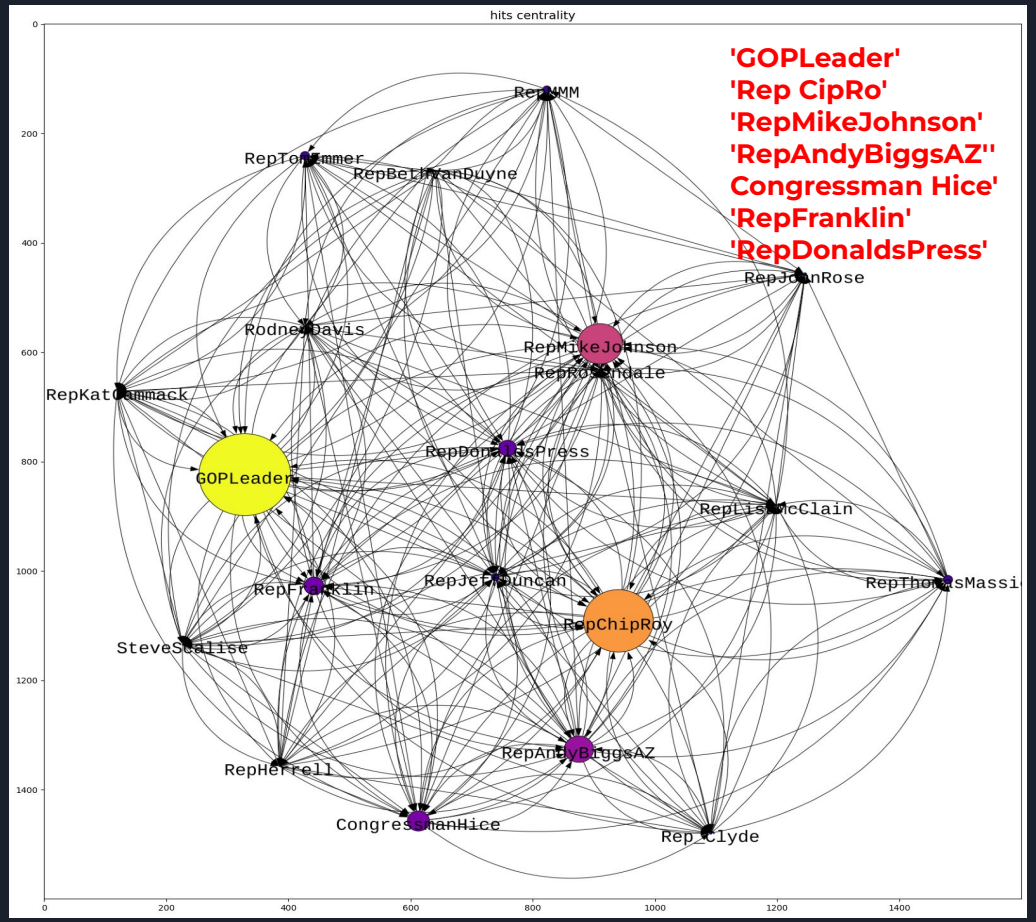


HITS (Authority)

MITCH MCCONNELL
Republican

CHIP ROY
Republican

MIKE JOHNSON
Republican



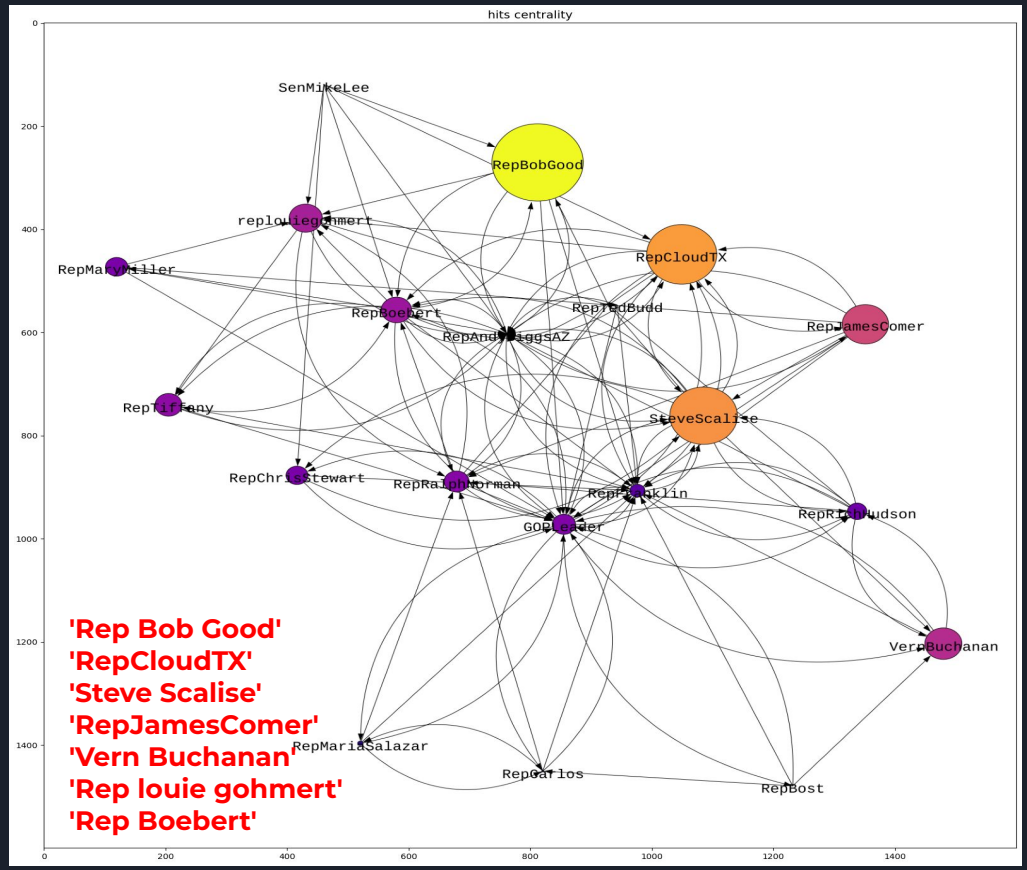
HITS(Hubs)

BOB GOOD
Republican

MICHAEL CLOUD
Republican

STEVE SCALISE
Republican

RepJamesComer
Republican



VIRAL CENTRALITY

Viral centrality identifies the most influential nodes in a network, particularly in the context of information spread. The idea is to find nodes that, when activated, can maximize the spread of information.

- ## Influence Probability:

  The probability that a node (or agent) can influence another node. This can be thought of as the likelihood that information or a disease will be transmitted from one node to another.

- ## Independent Cascade Model (ICM):

  A model often used to simulate the spread of information or disease in a network. In the ICM, once a node becomes active (infected), it has a single chance to activate (infect) each of its inactive neighbors based on a given probability.

VIRAL CENTRALITY ALGORITHM

VC is designed to identify influential spreaders within networks for which probability of transmission between nodes can be quantified.

Initialization

- For each node in the network, set it as the seed node.
- Initialize vectors to keep track of which nodes are susceptible (not yet influenced) and which have been activated (influenced) in the previous and current time steps.

Spread Simulation

- Use the Independent Cascade Model to simulate the spread from the seed node.
- For each time step, calculate the probability that each node is influenced based on the influence probabilities of its neighbors.
- Update the activation status of each node.

Centrality Measure

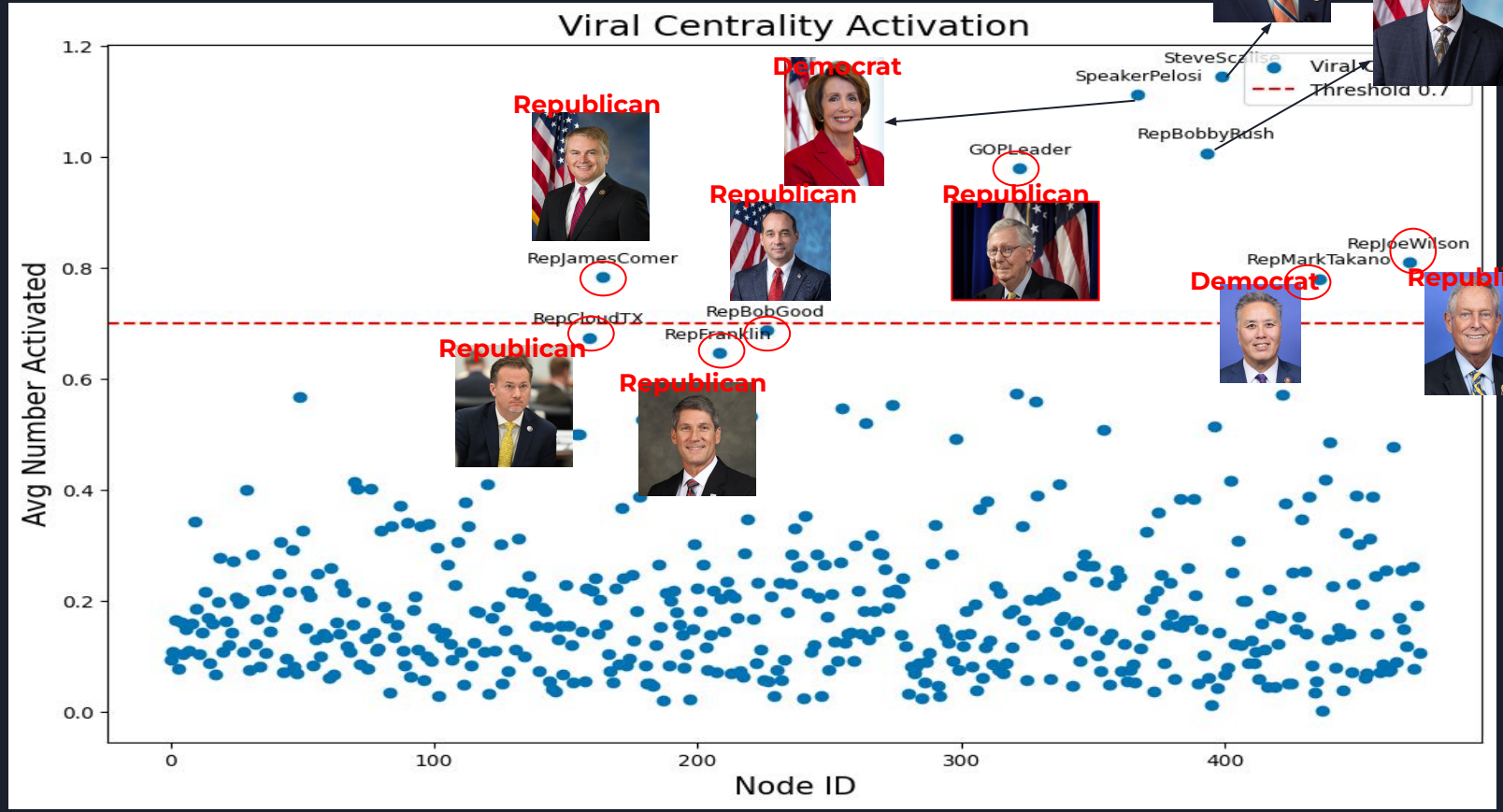
- The viral centrality of the seed node is determined by the total number of nodes it influences, minus one to exclude the seed node itself.

Algorithm 1 Viral Centrality

```
for each node as SeedNode do
    prob_susceptible = ones(TotalNodes) # initializing probabilities a given node is still susceptible
    prev_activated = zeros(TotalNodes) # initializing probabilities a given node was activated on previous time step
    cur_activated = zeros(TotalNodes) # initializing probabilities a given node is activated on current time step

    prev_activated[SeedNode] = 1 # set seed node to definitely being activated on previous time step...
    cur_activated[SeedNode] = 0 # ... and therefore definitely NOT activated on current time step

    t=1 #initialize time step
    while (termination condition) do

        perform breadth-first search to determine all nodes within t edges of SeedNode

        for each Node within reach of SeedNode do
            prob_uninfected = 1 # initialize probability that a given node is not infected on this time step
            for each Neighbor sending a connection to Node do
                # P[i, j] is probability of node j activating node i, given that node j was activated on previous time step
                prob_uninfected = prob_uninfected * (1-prev_activated[Neighbor]*P[Node, Neighbor])
            end for
            cur_activated[Node]=(1-prob_uninfected)*prob_susceptible[Node]
        end for

        for each Node in TotalNodes do # clean-up in preparation for next time step
            prev_activated[Node] = cur_activated[Node]
            prob_susceptible[Node] = prob_susceptible[Node] - cur_activated[Node]
        end for

        t = t + 1
    end while
    viral_centrality[SeedNode] = sum(1-prob_susceptible) - 1 #"-1" discounts initial activation of seed node
end for
```

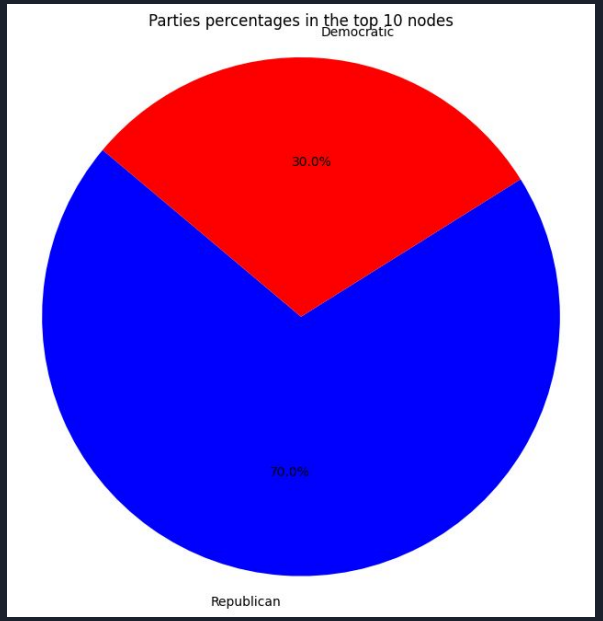
VIRAL CENTRALITY RESULTS

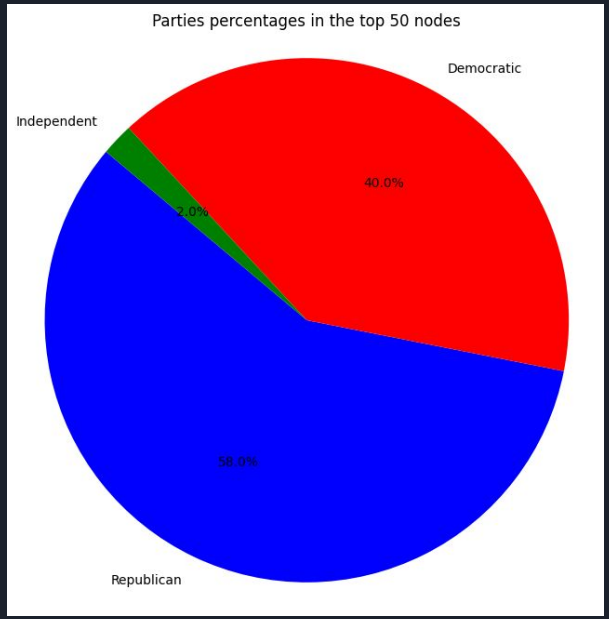
This plot shows the expected average number of activated nodes for each vertex in the gra



VIRAL CENTRALITY RESULTS

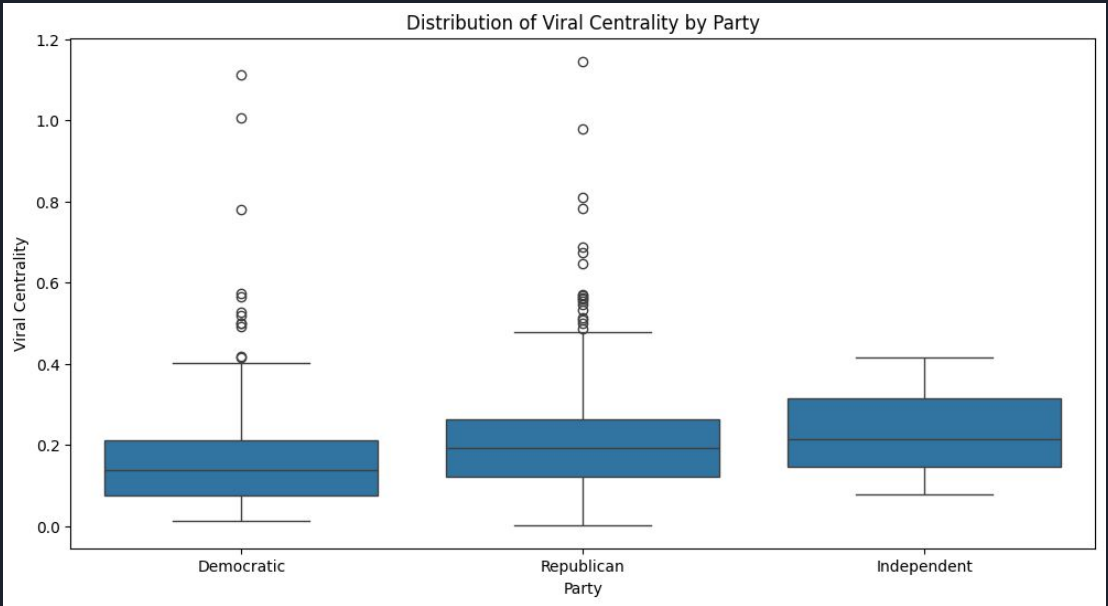
Despite having more democratic than republican congress members in our dataset (254 vs. 218), it appears that most of the nodes with highest viral centrality value belong to the republican party.





VIRAL CENTRALITY RESULTS

To have further confirmation of our results we performed further statistical tests, in particular we performed a t-test between the scores of the nodes in the republican party and the democratic one, which yielded t-statistic = -4.195871836468747 and p-value = 3.250697898112218e-05.



CONCLUSIONS

- Politicians tend to interact predominantly with others from the same party
- The small world metrics show division inside the network and an efficient spread of information, especially inside the political groups
- Based on most centrality measures, Mitch McConnell emerges as the most influential figure in the US Congress.
- Viral centrality efficiently identifies key spreaders in the network, with significant overlap among those with highest centrality.
- The analysis indicates that Republican politicians are the primary spreaders of information within the graph.
- The path from a Republican to a Democrat reflects efficient communication and connectivity among all Congress members.
- The isolated nodes have low overall centrality but show efficient communication within their smaller subgroups.
- As expected, the network comprises three main communities: one includes both Republicans and Democrats, while the others consist solely of Republicans or Democrats, highlighting the flow of information within and across party lines.