Analyzing a Congressional Twitter Interaction Network Using MCL, Spectral Clustering, and PageRank

Ian J. Wallace
University of Central Florida
Orlando, Florida
ian.wallace@knights.ucf.edu

***Abstract*—This paper analyzes a directed, weighted graph of the influence of members of Congress on Twitter. The weights indicate the probability of influence. We used Markov clustering and spectral clustering to see if we could cluster members based on party affiliation. The clustering analysis reveals that both MCL and spectral clustering effectively categorize Congress members into two clusters aligning with the two major political parties, based on the clustering evaluation methods we used. Additionally, we used the PageRank algorithm to determine if there is any correlation between the formal leadership positions in Congress and the influence they have on Twitter. Our results indicated that there is no correlation between PageRank ranking and official leadership roles, except for an intriguing outlier, Kevin McCarthy, the House Minority Leader. This study contributes to the evolving discourse on politics and social media, emphasizing the need for continued exploration of the intricate dynamics between digital influence and formal leadership positions within the U.S. Congress.**

I. Introduction and Problem Statement

Twitter has become an important part of politics in the United States. Nearly all politicians, including members of the United States Congress, use it to communicate their positions and ideas to both voters and other politicians. The Twitter accounts of members from the 117th United States Congress, which includes both the House of Representatives and the Senate, represent a rich source of information. In this paper we will use graph clustering techniques (Markov clustering (MCL) and spectral clustering) to see if it is possible to discern political party affiliation from this this communication. In addition, we will use the PageRank algorithm to rank the member's of Congress based on their relative importance based on this network and compare it to formal Congressional leadership positions.

The primary focus of this research is to leverage MCL and spectral clustering algorithms to delineate the congressional Twitter network into two distinct classes. We hope to see if these clusters correlate to known political factions. For this paper, we will only be using these clustering algorithms to partition the network into two clusters which will then be compared against the two major political parties in the United States, Republicans and Democrats.

Beyond network clustering, this paper will also look at the comparison of PageRank-derived rankings with formal congressional leadership positions. By correlating the digital influence of congress members with their official roles, such as Speaker of the House, Senate/-House Majority Leader, Senate/House Minority Leader, and party whips, we seek to examine the correlation between digital prominence and formal leadership roles within the legislative branches.

II. Dataset Overview

The dataset utilized in this paper is called "Twitter Interaction Network for the US Congress" and was found on the Stanford Network Analysis Platform (SNAP).[1] The original source of the dataset is from Fink et. al. [1] The dataset takes the form of a directed, weighted network in which the edge weights are empirically obtained "probabilities of influence" between all pairs of Congresspeople. For this dataset, "influence" was considered to be every time a member of Congress re-tweeted, quote tweeted, replied to, or mentioned other Congressional members. The probability of influence was found by normalizing the summed influence by the number of tweets issued by each Congressperson.

[1]Twitter Interaction Network for the US Congress: https://snap.stanford.edu/data/congress-twitter.html

TABLE I
MEMBERS OF CONGRESS IN LEADERSHIP POSITIONS, ORDERED IN TERMS OF IMPORTANCE

Position	Name	Twitter Username	Political Party
Speaker of the House	Nancy Pelosi	SpeakerPelosi	D
Senate Majority Leader	Chuck Schumer	SenSchumer	D
Senate Minority Leader	Mitch McConnell[2]	N/A	R
House Majority Leader	Steny Hoyer	LeaderHoyer	D
House Minority Leader	Kevin McCarthy	GOPLeader	R
President pro tempore	Patrick Leahy	SenatorLeahy	D
Senate Majority Whip	Dick Durbin	SenatorDurbin	D
House Majority Whip	Jim Clyburn	WhipClyburn	D
Senate Minority Whip	John Thune	SenJohnThune	R
House Minority Whip	Steve Scaliuse	SteveScalise	R

The original dataset contains a digraph with weighted edges, where the edge weights are the aforementioned "probabilities of influence". The nodes in the dataset are labeled with the Twitter usernames of each Congressperson. This data is included in the `congress_network_data.json` file. Additionally, we modified the `congress_network_data.json` file to include an additional label called "partyList", which uses an "R" or "D" to label that Congressperson with their respective political party. There is also a `congress.edgelist` file that was included. This file describes the graph as a set of edges and weights.

Some members of Congress identify as politically independent and therefore they are not officially a member of either the Republican or Democrat political parties. These members include Senator Bernie Sanders, Senator Kyrsten Sinema, and Senator Angus King. However, based on mainstream political understanding and a report from the Congressional Research Service, all three of those Senators caucus with Democratic party, therefore we labeled them as Democrats for the purposes of this research. [2]

In order for probabilities of influence to be meaningful, the authors of this dataset required that a given Congressional member issue at least 100 tweets over a time-frame of February 9, 2022 to June 9, 2022 in order to be included in this dataset. This resulted in 475 Congressional members out of the 535 total members being included in the network. This requirement leads to the dataset missing an important member of Congress, Senate Minority Leader Mitch McConnell, who did not meet the Tweet threshold in order to be included in the dataset. He is included in Table I for reference, but he is not included in the network.

III. METHODOLOGY

In this paper, we are examining two separate problems on the same dataset. The first problem we are looking at is dividing the dataset into two separate clusters and evaluating how they align with the values the data has been labeled with, which is the two major political factions, the Republican and the Democratic parties. The second problem we will be looking at is using PageRank to rank the nodes of the dataset and compare it to formal Congressional leadership positions.

A. Clusters Compared To Political Parties

To perform the clustering of our network, we will be using two clustering algorithms Markov clustering (MCL) and spectral clustering. For this project, we wrote our own implementation of MCL which is defined by the `markov_clustering()` function in the source code. This function takes in a NumPy adjacency matrix, a value for the inflation parameter, and a number of iterations to run the expansion, inflation, and normalization steps of MCL. We will be using the implementation of the spectral clustering algorithm that is included with the SkLearn Python package. Both MCL and spectral clustering algorithms expect un-directed graphs and MCL expects unweighted, un-directed graphs. Therefore, we used the function called `create_undirected_graph()` create an un-directed, unweighted graph for use by the clustering algorithms. The code for this section can be found in the `clustering.py` file.

To evaluate the clusters generated by each algorithm against the known labels, we used the Normalised Mutual Information (NMI), the Adjusted Mutual Information (AMI), and the Adjusted Rand Information metrics.

[3] All of these metrics provides scores from 0.0 to 1.0. The NMI and AMI metrics both the concept of mutual information (MI), which is a function that measures the agreement of the predicted labels and the ground truth labels. The ARI is a function that measures the similarity of the predicted labels and the ground truth labels. We used the implementation of these metrics that can be found in the SkLearn Python package. [4]

We will also test some hyperparameters for each algorithm to see how it influences the performance. For MCL, we will be testing it with different values for the inflation parameter. We will only include the results from this algorithm if they generate an output with two clusters, since we are evaluating the clusters against a ground truth with only two labels. For spectral clustering, we will try two different values for the assign_labels parameter: discretize and cluster_qr. The cluster_qr method directly extracts clusters from eigenvectors in spectral clustering and is stated in the documentation to potentially outperform discretization.

B. Network Influence Ranking Compared to Congressional Leadership Positions

To rank Congressional members in terms of influence in the network, we used the PageRank algorithm. For PageRank, we selected a standard value of 0.85 for the damping factor and an epsilon of 1×10^{-6}. Our PageRank implementation then returns an array of ranks, with rank 1 being located at the first index, and an array of scores that corresponds to those rankings. The code for this section can be found in the `pagerank.py` file.

To evaluate how the PageRank rankings of Congressional members in leadership positions compares to the rankings of normal members, we used the chi-square test of independence. We did this in order to determine if having a high PageRank is associated with being a Congressional leader. Our null hypothesis for the test is that their is no association between being a leader and having a higher PageRank ranking. To perform the test, we discretized the 475 ranks into 10 groups. We then calculated a contingency table, using the Pandas `crosstab()` function, between the discretized ranking groups and the `isLeader` categorical variable. Using our calculated contingency table, we passed the table to the SciPy `chi2_contingency` function, which output both the chi-square value and the P-value. We then compared the P-value to see if it was less than the critical value ($\alpha < 0.05$). Additionally, we calculated some standard descriptive statistics (mean, standard deviation, Z-score, etc.). To calculate these statistics, we took the NumPy array outputs of the `pagerank()` function and created a Pandas dataframe to calculate the descriptive statistics and the Z-scores for the PageRank score values. [5]

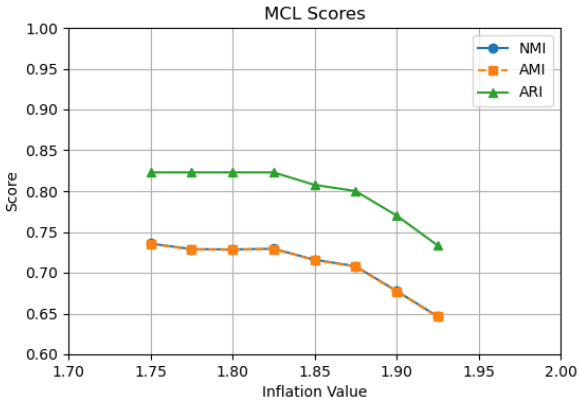

Fig. 1. MCL scores

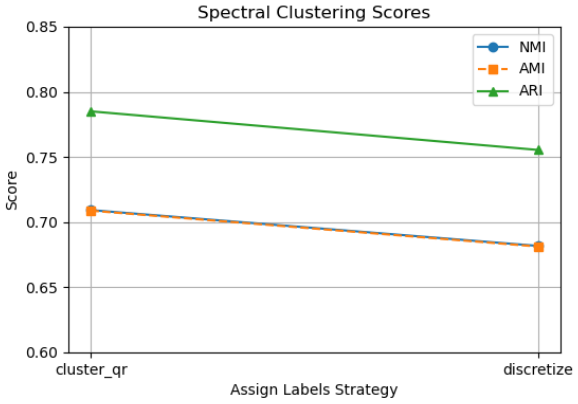

Fig. 2. Spectral clustering scores

IV. Results

In this section, we present the results obtained from our experiments, including both the analysis the of clustering algorithms to discern political parties and comparing the most important Congressional members according to PageRank compared to official Congressional leadership positions.

A. Clusters Compared To Political Parties

As discussed earlier, we ran the two clustering algorithms multiple times with different values for some of the hyperparameters. For the MCL algorithm, we found that the clustering results achieved the highest results across all metrics for a value of 1.75. In Figure 1, we show an inflation value of between 1.75 - 1.825 had

TABLE II
THE BEST SCORES FROM EACH ALGORITHM

Clustering Algorithm	NMI	AMI	ARI
MCL (inflation=1.75)	0.73579	0.73539	0.82295
Spectral Clustering (cluster_qr)	0.7093	0.70885	0.78509

a similar score, but then the scores quickly fell of for inflation values greater than 1.85. As seen in Table II, the NMI and AMI essentially identical. The ARI achieved the highest value in the study with a value of 0.82295.

In Figure 2, we compare the results for spectral clustering using the two different assign labels strategies, cluster_qr and discretize. We found cluster_qr performed the best. As seen in Table II, both NMI and AMI were nearly the same and ARI had the highest value. The ARI metric had the highest value with 0.78508.

B. Network Influence Ranking Compared to Congressional Leadership Positions

After running our PageRank implementation on our dataset, we obtained the rankings for each of the Congress members who designated as being in a leadership position (Table I). These rankings can be found in Table III. Using these rankings, we ran the chi-square test as described earlier and calculated a P-value of 0.46. Since that value is greater than the critical value, we could not reject our null hypothesis. Therefore, from this test we find that there is no significant association between being a leader and having a high PageRank ranking.

Figure 3 shows how the leader PageRank scores compare to non-leader Congressional members. We found that the mean rank of a leader Congress member to be 168.67, which is higher then the overall mean rank of 238. However, as you can see none of the Congressional leaders really stand out among the regular members. This fits with the conclusion we drew from the chi-square test.

The one exception to this is the member with the top PageRank score. This member is Kevin McCarthy (GOPLeader). He had a PageRank score with a Z-score of 7.5, making him a very important node in this network. It would be challenging to not draw an association between his outlier score and his position as House Minority Leader. We believe that future research is needed why he is ranked higher than other leaders.

[2]Mitch McConnell is not included in the dataset, because he did not meet the 100 Tweet threshold during the specified time frame.

TABLE III
THE BEST SCORES FROM EACH ALGORITHM

Rank	Score	Twitter Username
1	0.016722	GOPLeader
18	0.006805	LeaderHoyer
41	0.004939	SteveScalise
137	0.002333	WhipClyburn
139	0.002324	SpeakerPelosi
190	0.001867	SenatorDurbin
277	0.001294	SenSchumer
355	0.000931	SenatorLeahy
360	0.000903	SenJohnThune

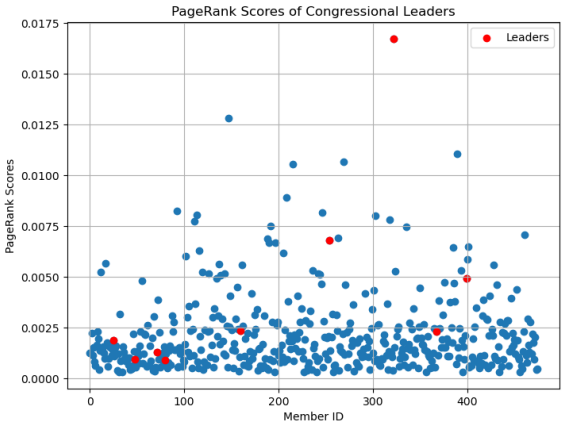

Fig. 3. PageRank scores with Congressional leaders marked

Perhaps gathering more data from more than one session of Congress could provide more insight.

V. CONCLUSION

In this paper, we presented an analysis of graph that represents the social interaction of members of the 117th Congress over Twitter. Using clustering techniques, such as Markov clustering (MCL) and spectral clustering, we sought to see if we could identify political affiliations based on Twitter interactions. Additionally, we used the PageRank algorithm to rank Congress members by digital influence and compared these rankings with official leadership positions.

The results of our clustering analysis indicate that both MCL and spectral clustering can effectively categorize Congress members into two clusters that closely align with the Republican and Democratic parties. This suggests that the social interaction on Twitter somewhat reflects political affiliations. It would seem to indicate

the Congress members on Twitter tend to interact only with others within their own party.

In contrast, the relation between PageRank rankings and formal leadership positions shows no strong correlation. The chi-square test we performed suggests that holding a leadership position doesn't strongly relate to having a high PageRank. However, the one exception to this is Keivn McCarthy, the House Minority Leader, who has a very high PageRank relative to other members. We believe that his status of being a number of standard deviations above the mean, even among other Congressional leaders, deserves more analysis. This could be due to a number of factors, including factors related to Kevin McCarthy as an individual, instead of being solely related to his leadership position. As stated before, a study that uses collected Twitter data from previous and future Congresses might be able to provide more insight.

In conclusion, our work provides insights into the relationship between politics and social media. Future studies might choose to try clustering the Congressional network in different ways to see if there is any comparison between those generated clusters and other formal groups in Congress, such as committees. Another expansion to the PageRank part of the study might be finding another grouping of Congress members to analyze, such as Congress members who are considered important in the context of American politics, but who do not hold any formal position. This might explain the above-average PageRank of certain members.

REFERENCES

[1] C.G. Fink, N. Omodt, S. Zinnecker, and G. Sprint, "A Congressional Twitter network dataset quantifying pairwise probability of influence." Data in Brief, 2023

[2] Congressional Research Service, "Membership of the 117th Congress: A Profile." 2022. https://crsreports.congress.gov/product/pdf/R/R46705

[3] L. Danon, A. Díaz-Guilera, J. Duch and A. Arenas, "Comparing community structure identification", Journal of Statistical Mechanics: Theory and Experiment, vol. 2005, no. 9, Sep. 2005.

[4] Scikit Learn. "Clustering performance evaluation" https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

[5] Pandas API Reference. "pandas.DataFrame.describe" https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html