Examining political diversity in Twitter followings using embedding

Tatsuhito Yoshihara
Keio University
Tatsu.yoshihara@keio.jp

ABSTRACT

In an age where social media provides widespread access to political information, it becomes vital to assess the ideological balance in the political figures followed by the general populace. This study employs the methodology developed by Waller and Anderson (2019) to analyze whether Twitter users predominantly follow politicians with a specific ideological leaning or maintain a diverse range. By adapting the Embedding methodology, this research achieved vector representations of politicians in Japan and the United States, revealing varied GS-scores in user following behaviors. Consistent with Waller and Anderson's observations, this study found that the GS-scores tend to be lower for users following more politicians in both countries. Additionally, it was observed that the GS-score distribution for followers of the Republican Party is broader compared to that of the Democratic Party. This approach not only leverages GS-scores to evaluate the extent of political diversity in users' following patterns but also allows for the customization of the definition of community similarity.

1. INTRODUCTION

To diversify the information obtained on social media, it is advisable to follow politicians with varying political viewpoints rather than exclusively following politicians who share the same political beliefs. This can be elucidated by employing the GS-score from previous research. While previous studies used Reddit data, this study focuses on Twitter, which is also a highly significant online platform in comprehending political trends of society.

Hypothesis.

Our hypothesis posits that individuals who follow a greater number of politicians tend to be more passionate about politics, leading them to follow politicians who align with their own political ideologies and consequently yielding higher GS-scores. Conversely, individuals with less interest in politics are inclined to follow only well-known politicians, irrespective of their party affiliations, resulting in lower GS-scores.

2. METHODOLOGY

This work aims to investigate whether sets of politicians that Twitter users follow exhibit political diversity, employing the GS-score methodology. Since complete data akin to Reddit dataset is not available on Twitter, the methodology of partisan dimension by Weller and Anderson (2021) to measure political diversity was precluded. To overcome this problem, I created a network where politicians with close ideology are connected, and applied Node2vec to get vector representation of politicians.

2.1 Data

I collected Twitter ID of followers of 401 politicians[1] who are the members of house of Representatives in Japan and 439 politicians[2] serving in the United States House of Representatives through Twitter API from February 6th to February 19th. Twitter accounts of U.S. politicians were collected from official website. As for the Japanese politicians, I have collected their accounts by visiting each party's official websites. Some twitter accounts were missing or had problems, therefore I filtered them out by hand. The followers of Republicans, on average, follow 2.9 politicians while followers of Democrats follow 2.4 politicians on average (see Appendix A for more details).

Table 1 | Number of followers of politicians

Number of followers of politicians in the U.S.	50712258
Number without duplication	18316737
Number of followers of Japanese politicians	10997465
Number without duplication	4807457

2.2 Politician Embedding

Using the degree of overlap in the follower sets that each politician has, I managed to construct two networks respectively for the U.S. and Japan. Thick edges are stretched between politicians with close political identities and thin edges are stretched between those with opposing political identities. By applying Node2vec to the network, the proximity between politicians can be measured and a GS-score can be obtained. Thus, the utilization of index vectors in the previous research, a feature only accessible with complete datasets, becomes unnecessary in this approach. Advantage of this method is that the proximity/closeness of each politician can be defined to suit one's own preferences by constructing a network with desired properties.

Calculating the similarity between two politicians requires determining the extent of overlap among their followers, and there are multiple metrics available for this purpose. The most famous of which are Jaccard Similarity, Sorensen Coefficient, Overlap Coefficient and Normal Count (which is simply the number of followers that two politicians share). The following section is a comparison of Jaccard Similarity, Overlap Coefficient and Normal Count.

Jaccard Similarity, Overlap Coefficient and Normal Count

$$Jaccard\ Similarity(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$Overlap\ Coefficient(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}$$

$$Normal\ Count(X,Y) = |X \cap Y|$$

Normal Count focuses on absolute number of followers in common while Overlap Coefficient restrict attention to proportion of overlap as shown above. Jaccard Similarity holds an intermediate meaning among these metrics. Normal Count, by definition, gives a thick edge to a pair of two famed politicians while it gives a thin edge to a pair of lesser-

1, Members of the House of Representatives in Japan.

2, Members of the United States House of Representatives.

known politicians. On the other hand, Overlap Coefficient only values the proportion of overlap, and this allows politicians from the same party to have thick edges. The following figures are the networks of US politicians using Jaccard Similarity, Overlap Coefficient and Normal Count, allowing us to pick a network that successfully maps relevant politicians.

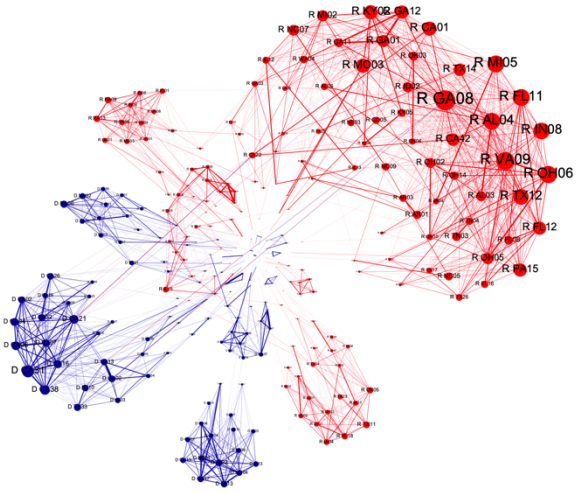

Fig. 1 | Network of politicians in the U.S. using Jaccard Similarity. 313 politicians with 1544 edges. An edge is only formed if its weight exceeds a threshold of 0.125. Both the size of the nodes and their name labels are scaled according to the degree of each node.

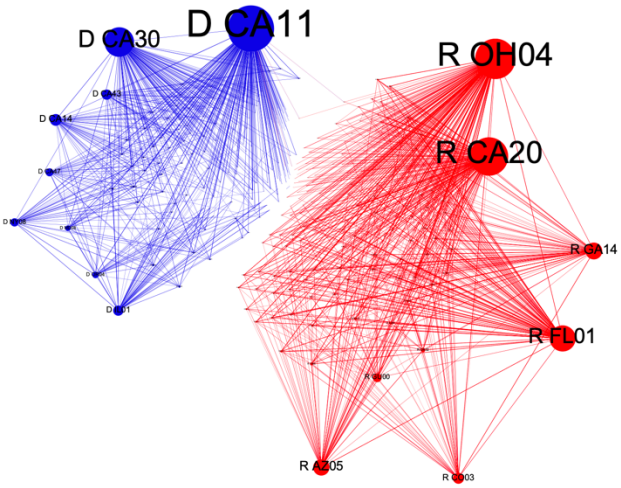

Fig. 2 | Network of politicians in the U.S. using Overlap Coefficient. 365 politicians with 1186 edges. An edge is only formed if its weight exceeds a threshold of 0.55. Both the size of the nodes and their name labels are scaled according to the degree of each node (see Appendix B for the network with threshold: 0).

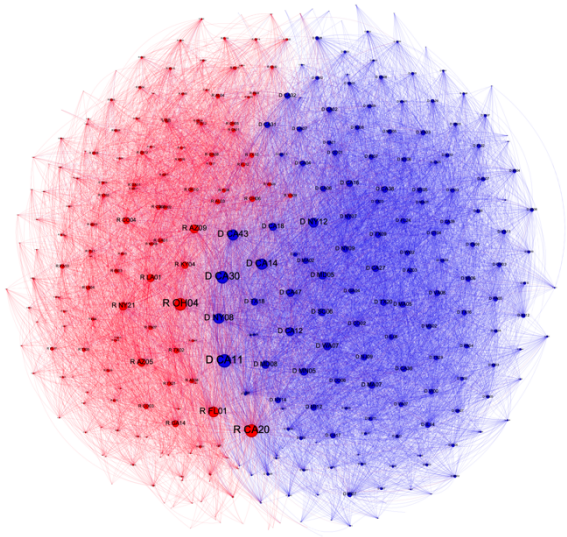

Fig. 3 | Network of politicians in the U.S. using Normal Count. 353 politicians with 13472 edges. An edge is only formed if its weight exceeds a threshold of 5000. Both the size of the nodes and their name labels are scaled according to the degree of each node.

Why network with Overlap Coefficient?

Matsuo et al. (2005) indicates that, among various metrics used for evaluating the strength of relationships through co-occurrence, the Simpson coefficient has been the most effective. In our analysis, well-known politicians tend to have a certain number of followers in common on Twitter regardless of their political ideologies. In this case, using Normal Count, any of two famous politicians would share thick edges and eventually similar vectors are given to them. On the other hand, Overlap Coefficient successfully divides politicians based on their ideology and creates clusters for each political parties as shown in Fig.2.

Not only it groups politicians into two, but it only builds connection among similar politicians based on the overlap in two follower sets. As for Jaccard Similarity, the network in the Figure 1 has split into several clusters and Node2vec would not provide embedding where politicians of same ideology have similar vector.

PCA on Politician Embedding

Using the Overlap Coefficient, I created a Network with threshold of 0 (see Appendix B) and applied Node2vec to obtain vector representations. However, since complete dataset is not available on Twitter, default parameters were employed. Unlike in Reddit data analysis, where the entire dataset can be leveraged and analogies can be drawn, our dataset lacks this advantage.

To check validity of the vector representation of politicians, I utilized the TensorFlow Embedding Projector to transform the 64-dimensional distributed representations to three dimensions through Principal Component Analysis. As shown in the Figure 4, in this embedding, politicians affiliated with the Democratic Party are depicted in blue, while those affiliated with the Republican Party are depicted in orange. It seems politicians from the same party are positioned closely to each other and thus, it can be inferred that the embeddings capture political identity to a certain extent.

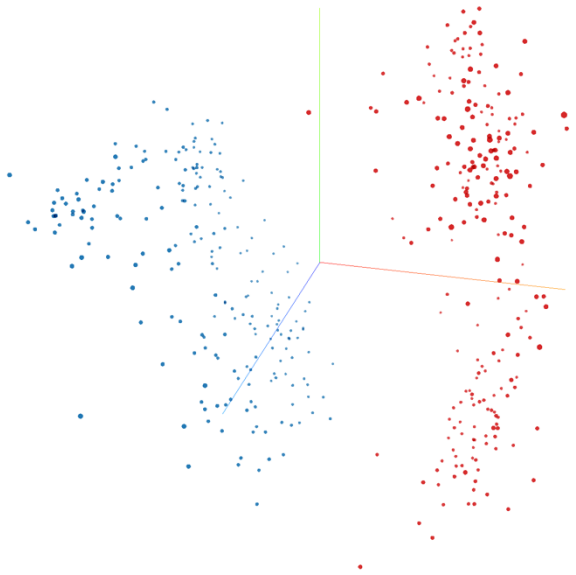

Fig. 4 | Visualization of Embedding for US politicians by Principal Component analysis. Blue-colored points represent Democrats and orange-colored points represent Republicans. Total variance described: 29.0%.

2.3 Measuring GS-score

In this analysis, user contribution is limited to a binary value, representing whether a user follows a respective politician or not, while on the other hand, the level of user contribution in the prior research was defined as the number of comments a user makes. Consequently, the formula for calculating the GS-score is simplified as follows:

$$GS(u_i) = \frac{1}{J}\sum_j \frac{\boldsymbol{c_j}\cdot\boldsymbol{\mu_i}}{|\mu_i|},\ \boldsymbol{\mu_i} = \sum_j \boldsymbol{c_j}$$

J represents the total number of politicians that user u_i follows and $\boldsymbol{c_j}$ corresponds to the vector representation of the j-th politician (note all vectors are standardized to a unit length via normalization). $\boldsymbol{\mu_i}$ denotes u_i's *center of mass*.

3. RESULT

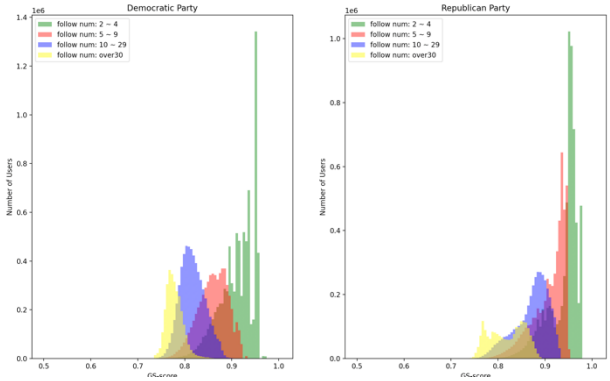

Fig. 5 | Distribution of the GS-score for Democrats (left) and Republicans (right). The distributions of GS-scores over users, broken down by number of politicians they follow.

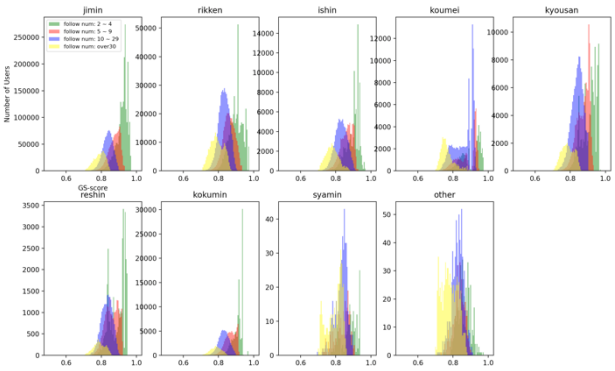

Fig. 6 | Distribution of the GS-score for followers of Japanese politician. The name of the political party is labeled at the top of each column.

4. Discussion & Limitation

In line with the previous research, GS-score lowers as they follow more politicians, leading to the conclusion that increased political diversity is associated with following a larger number of politicians. However, it is worth noting that there are instances where users following 30 or more politicians and have higher GS-scores than those following only 2 to 4 politicians. Consequently, it is evident that our hypothesis, which posited that individuals fervently interested in politics tend to follow politicians with preferred ideologies results in higher GS-scores, was found to be incorrect. The Republican Party exhibits a broader range of GS-scores across various user demographics and its histogram skews right as opposed to that of the Democratic Party. The Figure 5 shows a smoother, more gradual distribution for the Republican Party, while on the other hand, the histogram for Democratic Party exhibits sharp peaks for red, blue, and yellow shades (see Appendix C for distribution of GS-score of users who follows over 30 politicians). Due to the unavailability of suitable analogies from prior research on semantic relationships in Reddit communities by Waller and Anderson (2019), it was not feasible to conduct a hyperparameter search for embeddings. Consequently, ensuring the validity of the analysis based on these embeddings has proven challenging. However, as mentioned in the section of politician Embedding, one can benefit from the flexibility of defining political similarity among politicians to suit the specific goals of the analysis.

Appendix A

The Twitter accounts of members of Japan's House of Representatives are primarily adopted from the official website (https://pressgallery.house.gov/member-data/members-official-twitter-handles) of their political parties. However, in cases where the information is not available on official websites, I searched for them on Google using "the person's name + Twitter" or "the person's name _ office _ Twitter". Among the multiple accounts displayed, I determined the authenticity by checking whether other official Twitter accounts of House of Representatives members, Senate members, or municipal council members follow them. For members of the U.S. House of Representatives, I used the Twitter accounts listed https://pressgallery.house.gov/member-data/members-official-twitter-handles as of February 5, 2022. There were instances where non-existent accounts were listed, or a different politician's Twitter account was assigned to a wrong politician, so I verified these through additional searches. AS of February 12, 2024, Anthony D'Esposito, a member of the Republican Party, holds two Twitter accounts, and the unverified account has more followers than the verified one. In cases where a politician has two or more accounts, the verified account was adopted.

Table2 | Followers of Republicans and Democrats

Total number of followers of Republicans	20168405
Number of users who follow Republicans	6954760
Total number of followers of Democrats	30543853
Number of users who follow Democrats	12862118
Proportion of Republican Party followers who only follow the Democratic Party	78.43%
Proportion of Republican Party followers who only follow the Democratic Party	88.34%

Table3 | Number of followers of each Party in Japan

Liberal Democratic Party of Japan (“jimin”)	7679810
Constitutional Democratic Party of Japan (“rikken”)	1777185
Japan Innovation Party (“ishin”)	366195
Komeito (“komei”)	268090
Japanese Communit Party (“kyosan”)	440008
Reiwa Shinsengumi (“reshin”)	101263
Democratic Party For the people (“kokumin”)	360913
Social Democratic Party (“syamin”)	1381
Others including independent	2587

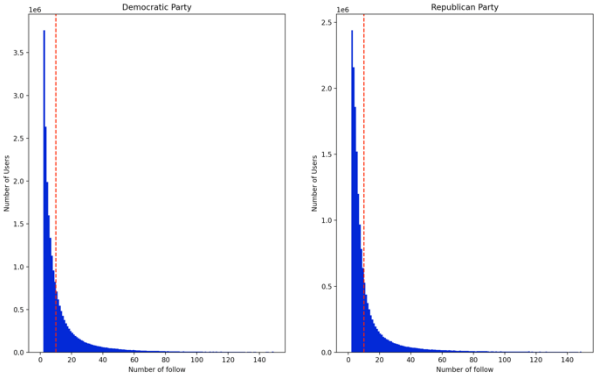

Fig. 7 | Distribution of number of followers of politician in the US. Vertical axis is the number of users, horizontal axis is the number of politicians followed by the users.

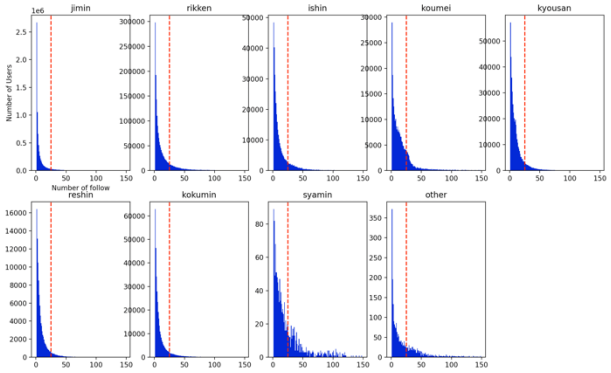

Fig. 8 | Distribution of number of followers of politician in Japan. Vertical axis is the number of users, horizontal axis is the number of politicians followed by the users.

Appendix B

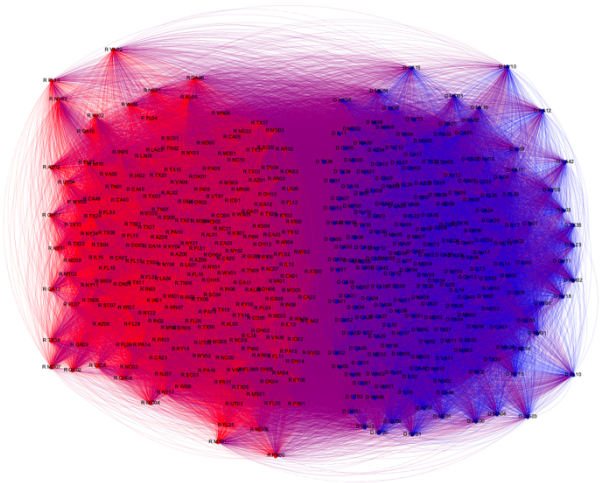

Fig. 9 | Network of politicians in the U.S. using Overlap Coefficient. 439 politicians with 96141 edges. An edge is only formed if its weight exceeds a threshold of 0. Both the size of the nodes and their name labels are scaled according to the degree of each node.

Appendix C

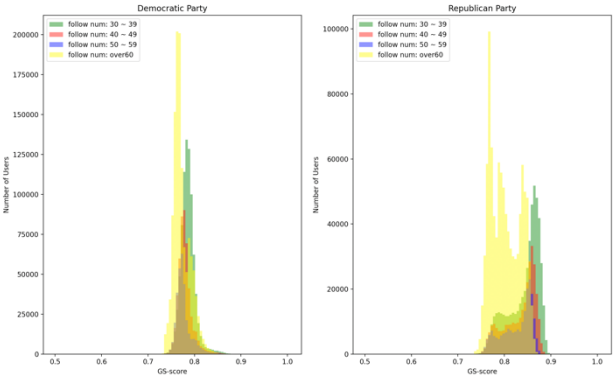

Fig. 10: Distribution of the GS-score for Democrats (left) and Republicans (right). The distributions of GS-scores over users who follow over 30 politicians, broken down by number of politicians they follow.

REFERENCES

Matsuo, Y., Tomobe, H., Hasida, Koiti., Nakashima, H., Ishizuka, M. (2005). Social Network Extraction from the Web information. Journal of the Japanese Society for Artificial Intelligence, 20(1), 46-56. https://www.jstage.jst.go.jp/article/tjsai/20/1/20_1_46/_pdf/-char/ja

Waller, I., & Anderson, A. (2019). Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. The World Wide Web Conference. https://doi.org/10.1145/3308558.3313729

Waller, I., & Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. Nature, 600, 264-268. https://www.nature.com/articles/s41586-021-04167-x