Wikipedia Voting Network Analysis

720084369
Department of Computer Science
University of Exeter
Exeter, UK
sp915@exeter.ac.uk

***Abstract*—This study analyzes the Wikipedia Administrator voting network to understand the social dynamics and decision-making factors involved. Using network analysis techniques, we examine individual node centrality, clustering effects, and community exploration to identify key influencers. This research sheds light on how influence and authority are distributed among individuals and how they impact rules and norms. This has implications for sociology, political science, and computer science, and contributes to our understanding of decision-making processes in social networks or online forums.**

***Index Terms*—Wikipedia Administrator, Request for Adminship (RfA), centrality measures, community detection, clustering effects**

I. Introduction: Contextualization and Motivation

Online communities and social networks are an essential part of people's lives nowadays and in recent years, these platforms have seen higher consumer engagement than ever. In particular, Wikipedia is one of the world's biggest and most important online encyclopedias, with information generated and maintained by a global voluntary network. The Wikipedia Administrator voting network is a subset of this wider community and is based on the Request for Adminship (RfA) process that helps Wikipedia select new administrators from volunteering users. It is crucial to the site's governance as it helps elect the administrators who will mandate the content that Wikipedia pushes out.

Network analysis has developed as a useful tool for studying the structure and dynamics of social networking sites, and it has been used to study a broad range of phenomena, from online communities to economic and political structures. This research aims to contribute to the expanding body of knowledge on internet forums and social networks by using network analysis techniques to the Wikipedia Administrator voting network. More so, the analysis will help in identifying the voting patterns in the election, and will also provide an outlook on the overall structure of the voting network. The analysis aims to identify the key players of the election, and further help Wikipedia in improving their governance and transparency towards selecting new administrators. This study sheds light on the impact of digital communities on society, politics, and economy. By examining the Wikipedia Administrator voting network, it increases awareness about the power dynamics and social networks that affect decision-making processes [1].

A. Related Work

The Wikipedia Voting network has been subject to various research and studies. Cabunducan [2] focused on implementation of logistic regression model on the network-level characteristics of a candidate's supporters while Burke and Kraut [3] analysed the importance of a candidate's certain attribute to be promoted in Wikipedia's Request for Adminship (RfA). Bonato and Bellomi [4] on the other hand use HITS and PageRank to gain insights on the macro-structure of the the organization and also identify any cultural biases related to specific topics. This paper [5] uses the Request for Adminship (RfA) procedure in the Polish-language Wikipedia to perform analysis on the Wikipedia Voting data using multidimensional behavioral social networks to classify the votes in RfA. This paper [6] closely resembles this study as it also uses community extractions and centrality measures to view the structure of network, but this implementation focuses on article data rather than voting data for RfA. The dataset used in these papers vary from this study as they have used article editing data, or Wikipedia data for different languages compared to English. The study in hand relates to all these studies as it's focused on network analysis on social network which is Wikipedia in this case. This paper prioritizes using network science to view the structure of the voting network and clustering effects on the voting patters which has not been focused by other studies. This paper will help provide Wikipedia the information on how communities are formed inside their organization and will help identify the clustering effects on votes that can be used to assess any bias or relationships inside the administration of the organization.

II. Description of Methods

The data is the Wikipedia Voting Network[1] acquired from the Stanford Large Network Dataset Collection[2]. The data was collected by Jure Leskovec and colleagues at Stanford University. The network holds all of Wikipedia's voting data from its foundation till January 2008. A directed edge from node $i \rightarrow j$ signifies that user i voted for user j.The methodology followed for the analysis along with the details of tasks to be followed is presented below:

a Network Visualization and basic topological attributes: Once the data is ingested, processed and ready for anal-

[1]https://snap.stanford.edu/data/wiki-Vote.html
[2]https://snap.stanford.edu/data/

ysis, visualizations of the network along with the giant component. The visualization will showcase the general structure of the network and how it is laid out. Besides visualizing the network, the general topological attributes of the network such as number of nodes, edges, shortest path for all pair of nodes, density of graph and other features are also discussed.

b Centrality Measures: Various centrality measures are implemented to identify the important and influential nodes in the network. For the analysis, degree centrality, betweenness centrality, closeness centrality and eigenvector centrality are used. These measures identify the most influential nodes, nodes that can spread information efficiently and also nodes that act as bridges that connect different communities.

c Community detection and clustering effects: Community detection algorithms identify groups of nodes that are closely linked to each other, aiding in identifying clusters of voters with shared opinions or values. Clustering effects examine how clusters impact voting patterns, with nodes inside a cluster tending to vote similarly. The Newman-Girvan modularity technique and Louvain method are used for community detection and clustering effects. These methods identify similarities in voter behavior based on voting similarity to others in the network.

Analyzing the Wikipedia Voting network is challenging due to its size and complexity, influenced by factors such as user engagement and social influence. Network analysis techniques are used to overcome these challenges and gain insights into the network's structure and dynamics, using centrality measurements, community identification, clustering effects, and network visualization to reveal new patterns and trends. The data lacks features on the voters or nominees, barring the study to focus on additional attributes that affect the centrality of the nodes, or clustering effects for community detection. Additional data can be collected to delve deeper on clustering of communities, to identify biases or existing relationships that impacted the voting results.

III. Discussion about Experiments

Here, the results of the analysis are presented with algorithms, proofs and discussion.

A. Network visualization and topological attributes of Network

The network has a total of 7115 nodes/participants and 100762 edges/votes. The average degree of a node in the network is found to be about 28 indicating an average of 28 neighbours for each node in the network. The shortest path for all pairs of nodes is computed, and it was found that in order to reach from one node to another, approximately 3.2 edges will be traversed on average. The density of the network is 0.00398 specifying that the network is a sparse one as $density < 1$, which is anticipated from a network on votes as voters are expected to vote for significantly less nominees than the listed nominees. Since the network was found to have many disconnected components, the diameter of network was computed for the Giant Connected Component (GCC) discarding the smaller disconnected components. The diameter was found to be 7 implying that nodes need to travel 7 or lesser edges to connect to any node in the GCC.The Giant Connected Component (GCC) of the Wikipedia Voting network is presented in Figure 1.

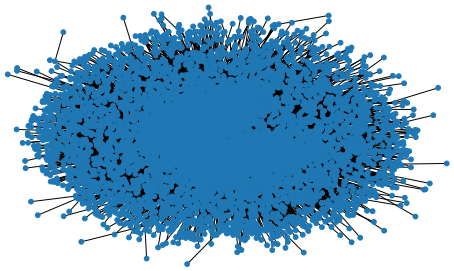

Fig. 1. Giant Connected Component (GCC) of Wikipedia Voting Network

B. Centrality Measures

Implementation of Centrality Measures in the network identified the 5 nodes that have high centrality value for each centrality measure in the network, and can be shortlisted as the 5 individuals who had the most impact in the voting patterns and outcomes. These 5 spotlight nodes are **2565, 766, 457, 1549** and **11**. Individuals with high degree centrality have more connections to other individuals in the network and signify a heavy influence on others and impact on the voting outcomes. About 15% of the total voters are connected to 2565, with other spotlight nodes each sharing votes with 10% of the total participants.Node 2565 is involved in votes from about 1065 voters. Nodes 766, 11, 1549 and 457 are linked to upwards of 700 votes each. Betweenness centrality identifies the individuals that act as bridges connecting different groups of voters and can be highlighted as individuals that can facilitate communication and collaboration among different groups in the administration. Majorly the whole network has betweenness centralities below 0.01, which indicates the network is sparse and mostly nodes do not act as bridges in shortest paths. Nodes 2565, 766, 457, 1549, 1166, and 11 have high closeness and eigenvector centrality, making them more central, connected, and influential in the network. These nodes have better access to information and resources, can spread information and build consensus, and impact voting outcomes. The majority of nodes have closeness centrality between 0.2 and 0.45, while most nodes have eigenvector centrality between 0 and 0.04. These influential nodes play a significant role in decision-making and information spread, and their behavior can be both collaborative and divisive.

C. Clustering effects

The tendency of nodes in a network to form groups or clusters based on similar traits or behaviours is referred to as clustering effects. Firstly, the assortativity coefficient of the network is -0.083 which signifies a negative assortativity meaning that individuals tend to vote less for individuals that are similar to themselves in terms of a certain attribute. The average clustering coefficient is 0.1408 with majority of the nodes have a clustering coefficient between 0 and 0.4. About 200 nodes have a clustering coefficient of 1. This implies the sparsity of the Wikipedia voting network and suggests a more decentralized network structure with fewer cohesive subgroups or clusters. There are 608389 unique triangles in the network with each node being part of a triangle in average of about 256 times. Interestingly, the median for a node being part of a triangle is just 1; such a huge difference in the mean and median triangles per node suggests that the majority of nodes in the network belong to extremely few triangles , while some nodes skew the average as they might be part of huge number of nodes while majority others have very less connections. The histogram of clustering coefficient for the Wikipedia voting network is showcased in Figure 2.

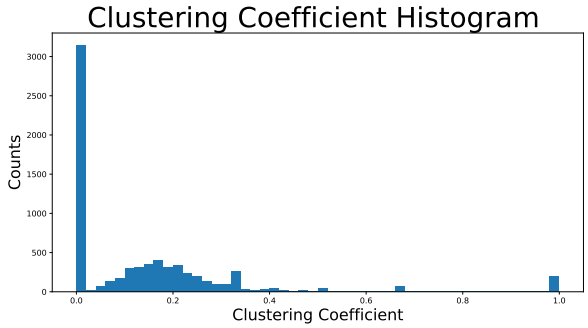

Fig. 2. Clustering Coefficient Histogram

D. Community Detection

Both the Girvan-Newman modularity technique and the Louvain method are implemented for community detection. The Girvan-Newman technique segregates communities based on the modularity score while the Louvain method iteratively optimizes the modularity score by merging or splitting communities. Despite their differences in algorithm, these methods segregate the different communities or sub-groups of voters in the Wikipedia Voting network that share similarities to each other. The similarities can be various factors, such as age, gender or matching interests in field of work. Community detection helped identify the size, quantity, and features of various communities, as well as their internal structure and links to other communities. It is useful to compare the composition and behaviour of the larger communities to those of smaller or more isolated communities, and to see if some groups are more influential or active in the voting network. Since the Giant Connected Component (GCC) only has about 100 nodes lesser than the entire network, the Louvain community detection method is implemented on the GCC. Figure 3 showcases the different communities identified.

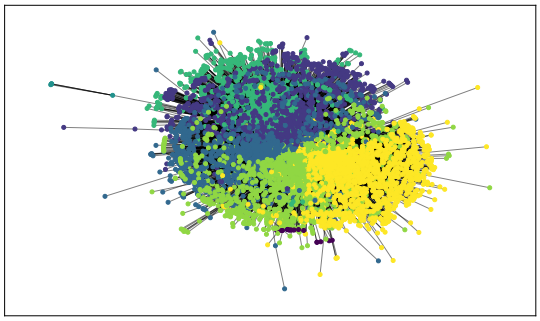

Fig. 3. Communities in Wikipedia Voting Network

The study provided an overview on how the voting network formed, and identified the existing communities and clustering effects on the organization. Despite the identification of communities in the network, the results lack backing on what cause the communities to form in such manner. Ingestion of additional information on the nodes, and relationship between nodes can help visualize the biases more and can provide more transparency on the Request for Adminship (RfA) process.

IV. Conclusion

The network science analysis of the Wikipedia Voting Network revealed its underlying structure, sub-communities, and factors that drive collaboration and decision-making. The study utilized centrality measures, community detection algorithms, clustering analysis, and network visualization tools. This research demonstrates the value of network science tools in understanding online networks and their dynamics. The findings can be applied to other online communities to develop new tools for facilitating collaboration and decision-making processes.

References

[1] P. Massa, "Social networks of wikipedia," in *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, HT '11, (New York, NY, USA), p. 221–230, Association for Computing Machinery, 2011.

[2] G. Cabunducan, R. Castillo, and J. B. Lee, "Voting behavior analysis in the election of wikipedia admins," in *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 545–547, 2011.

[3] M. Burke and R. Kraut, "Mopping up: Modeling wikipedia promotion decisions," in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, (New York, NY, USA), p. 27–36, Association for Computing Machinery, 2008.

[4] F. Bellomi and R. Bonato, "Network analysis for wikipedia," in *proceedings of Wikimania*, p. 81, 2005.

[5] M. Jankowski-Lorek, L. Ostrowski, P. Turek, and A. Wierzbicki, "Modeling wikipedia admin elections using multidimensional behavioral social networks," *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 787–801, 2013.

[6] T. Yamada, K. Saito, and K. Kazama, "Network analyses to understand the structure of wikipedia," in *Symposium on Network Analysis in Natural Sciences and Engineering*, p. 108, Citeseer, 2006.