International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

https://dx.doi.org/10.12785/ijcds/110153

Analysis of Social Media Complex System using Community Detection Algorithms

Hafiz Abid Mahmood Malik

Faculty of Computer Studies, Arab Open University Bahrain

Received 06 Jun. 2021, Revised 31 Dec. 2021, Accepted 11 Jan. 2021, Published 20 Jan. 2022

Abstract: The increasing significance of social networks has led to information propagation and community formation being an interesting domain in data science. The data gathered from big social networks exhibit different community structures. These communities attract various users who grow complex networks. The main goal is to identify the impacting nodes responsible for community data flow. The Twitter network edges are considered in the study, which plays a vital role in representing the activities and relationships developed among the community members. Different communities evolve when the network is analyzed using different community detection algorithms. The network statistics are used for analysis by calculating the weighted degree distribution of nodes in this study. The network is analyzed according to persistent clusters using community detection algorithms like the Spinglass, Walktrap, Fastgreedy, Leading Eigenvector, Multilevel, Edge Betweenness, and Label Propagation. It is found that these measures are very useful in community detection and observing the spread of information in social networks.

Keywords: Community Detection Algorithms, Social Networks, Complex System, Twitter network

1. Introduction

As social media networks have gained popularity, more people join the Social Networking Services (SNS). It has inspired a great deal of research into online social networks to study human behavior and interactions. The resulting sentiments that spread through these dynamically growing networks have a global impact on the thought process of those who are a part of such social networks [1], [2], [3]. These networks are very large and have a high growth rate with the passing time. Hence, the need to study the structure and growth patterns of such networks that have become mainstream for global audiences. Being large networks, it comes up with computational challenges. On the other side, analyzing these big datasets give a better understanding of human social behavior. These networks have also become a major source of marketing, advertising, and promotions. They draw the attention of large companies, corporations, and organizations to invest resources in research and analysis of such networks. There are a lot of examples of complex networks, such as networks of airports, scientific collaboration networks, the network of diseases, social media networks, crowd networks, and communication networks [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] The different social media networks have millions of users, and a high growth rate tends to provide improved services to their users. These include discussion forums, business and interest-related groups, games, music, videos. The exchange of information results in a large time duration spent on social networks by the users. The situation in SNS (forums, social networks, blogs, etc.) can be analyzed considering two factors; the number of users and the strength of interaction among these users. The more people on the social networks increase the informational traces about different behavioral aspects. The impact of online discussion forums, crowdsourcing, intelligent networks, virtual cooperation, and smart technologies points toward observing and analyzing a mainstream platform in today's world [14]. The primary objective is to identify the nodes that are main source for the community's data flow. The study considers the Twitter network's edges, which are critical in depicting the activities and relationships created among community members. In this study, the Twitter network is analyzed using different community detection algorithms. The network statistics are used for analysis by calculating the weighted degree distribution of nodes. The network is analyzed according to persistent clusters using community detection algorithms like the Spinglass, Walktrap, Fastgreedy, Leading Eigenvector, Multilevel, Edge Betweenness, and Label Propagation.

2. Literature Review

Social media networks are also used for large-scale analysis in applications that use image tagging, auto suggests, data indexing, etc. Datasets from online social networks exhibit well-defined community structures with many user participation. Therefore, resulting in complex and huge

E-mail address: hafiz.malik@aou.org.bh

graphs that are computationally challenging. The nodes that act as major information flow sources are essential in the network. They need to be identified due to their significance. They play a key role as connectors in the network and sometimes between different networks [15], [16]. The network science theory provides many methods and models to understand and analyze large-scale networks such as Facebook, Instagram, Twitter, LinkedIn, etc. The different models rank nodes based on various measures identifying the most influential nodes in a network using different centrality measures. This identification also makes it easier for newly developed nodes to connect to influential and authentic resources in the network. Large organizations also use these results in marketing and advertising to invest in fewer resources and spread their word. Identifying influential nodes has also become one of the important problems in network science, information retrieval, and data mining [17]. Another important measure in any network is any node's diversity. It shows how diversely any given node connects with the other nodes in the network domain. It is important to study the nature of the network, but there is not much research done considering it as an important part of understanding and studying the network semantics [18], [19], [20]. The huge data extracted in raw form from the social networks also depicts other characteristics like big data, semi-structured, high quality. A reflection of human society attracts many researchers towards social network analysis. But other than this, mining and analyzing this data is a non-trivial task with two limitations; the incompletion and dynamism of data. The information obtained from social networks is incomplete since only partial information can be collected from such platforms. Secondly, this information has a high rate of dynamicity, as it keeps changing rapidly with time [21], [22], [23], [24], [25].

3. Data Specification

For this study, 3500 tweets are collected about a specific matter and from the population of these tweets. Fifty users and their tweets are selected as samples by which the network is formalized. A simple random sample technique is utilized to attain the samples.

A. Network Statistics Used for Visual Analysis

Fifty nodes that have been filtered are used in this network. At the same time, there are 669 edges between these nodes. The 'edges' are the connections among users. The network is based on the people who tweet, their followers, and those who retweeted, showing the network's diameter in which the tweet information propagated. There are different colors and widths of edges to represent the weight of edges. The thickest and darkest edges have the maximum weight, and thinner and lighter shaded edges have lesser weights.

Data items of tweets consist of the following fields:

- Name: The name of the user who tweeted.
- Username: The username s/he uses on Twitter as their identity.
- Description: The description of the tweet (i.e., the main topic or message conveyed).
- Followers: The number of followers that a user has on Twitter describes the vastness of their circle.
- Number Statuses: Number of tweets s/he did on the topic.
- Time: The date and time of the tweet.
- Tweets: The actual tweet string as tweeted by the user. It also includes a keyword RT if the tweet was not done by the user but is a retweet of some other user.

B. Degree Distribution of Nodes

The degree of a node indicates the number of different usernames connected to a given username. Table I depicts the degree distribution for the tweets network.

TABLE I. Degree Distribution of Nodes

Min	Max	1st Quadrant	Median	Mean	3rd Quadrant
1.00	21.25	29.00	24.78	29.00	47.00

4. Methodology

In the user tweet network, we have designed our research approach to revolve around the following research patterns collectively:

- Correlation Analysis
- Observational Research
- Meta-Analytic Conclusion

A. Correlational Analysis

Correlational method is commonly used to ascertain the relationship between the events being analyzed. It typically denotes the degree of linear dependency between variables and the effect of a change in one of the relationships on the other variables. Correlation models and analysis techniques are used in environments where we need a predictive analysis based on the current understanding and observations of the system. In this system, the correlation among different users is calculated by observing the impact of one user on the entire system in terms of influential nodes [10].

B. Observational Research

The observational technique is used to simulate the network using various strategies for community building. Following that, the findings are analyzed to determine the optimal strategy for detecting community development in order to maximize coverage and assure information dissemination throughout the network. The observational study is concerned with concluding for the entire data population through the analysis of a sample [10].

C. Meta-Analytic Conclusion

After the correlation and observational analysis of the network, it is concluded using the meta-analysis strategy, a statistical analysis that combines the results from many different scientific studies. The basic concept behind such analysis is that some common features and characteristics. It exists behind similar networks and graphs such that they exhibit common traits to a certain extent. They start varying in different characteristics that are a tradeoff between the exhibitable features of the network [10].

5. Results and Discussions

A. *Analysis of Weighted User Tweet Network According to Persistent Clusters with Filtration, Spinglass Community Detection*

In the below analysis, the Spinglass approach is followed, which is related to statistical physics. This approach is based on the Potts model, where each node will be considered in one of the crystalline spin states. The connections or edges show the tendency of the nodes to maintain their spin state or to change the state. We further follow this concept to analyze the network for various spin steps. The final spin states that the network displays, in the end, are considered to be the final structure of the community formed in the network. Figure 1 represents, user tweet network (50% filtered) with clusters represented according to the spin-glass community. The network is shown by using a heat map to show the cluster formation within the network, with the node size depicting the influence of nodes (users) and the edge weights denoting the strength of connections.

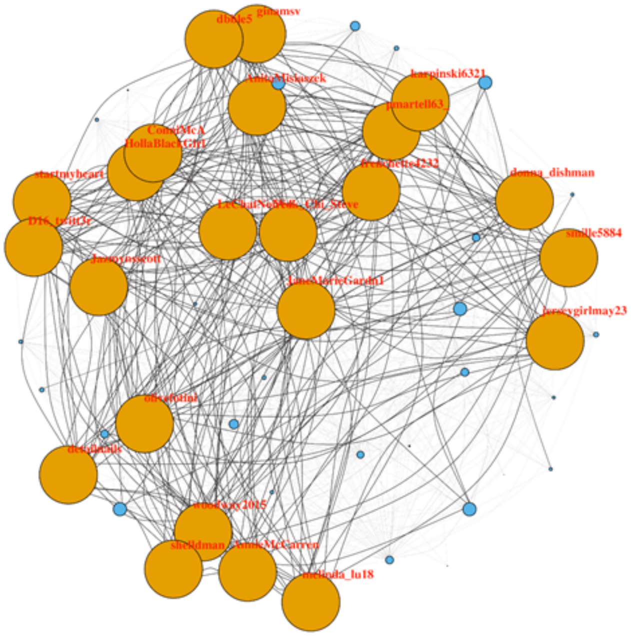


Figure 1. User tweet network (50% filtered) with clusters represented according to the Spinglass community

B. *Walktrap Community Detection*

The Walktrap community detection works on random walks. If we take random walks on the graph, we probably stay within the community because very few edges lead outside the community. After all, the community structure is highly connected inside and has rare connections outside. The results of this concept are used to merge different graphs (communities) using a bottom-up approach. Figure 2 shows the user tweet network (65% filtered) with clusters represented according to the Walktrap community. It has generated the network by using weighted degrees with colors corresponding to different characteristics of clusters in the Walktrap community.

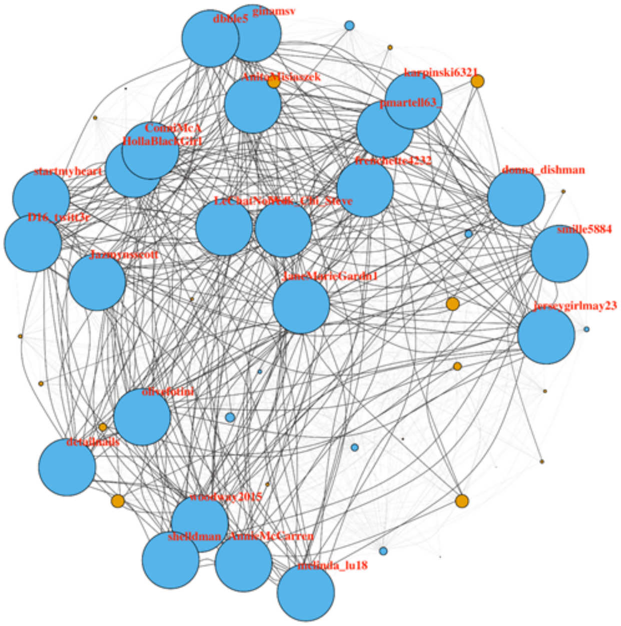


Figure 2. User tweet network (65% filtered) with clusters represented according to the Walktrap community

C. *Fastgreedy Community Detection*

The Fastgreedy approach optimizes the network modularity with a greedy behavior using the bottom-up approach. All the nodes are assigned to different communities, and the community formation process goes on iteratively by merging communities. Each iteration is feasible and optimal (each merged community increases the modularity with the largest possible value). This process goes on iteratively up to the point where there is no further increase possible. The modularity of the network results in the output of the community formation. The network is represented in the heat map and dendrogram for better visual analysis. The Fastgreedy algorithm is preferred because it gives fast results and does not have any parameters for approximation purposes. In Figure 3, the network is generated after 50% filtered, using weighted degrees with colors according to different cluster formations in the Fastgreedy community. The colors of nodes represent the communities with the edge weights denoting the weakness or strength of connections among the nodes in communities.

D. *Leading Eigenvector Community Detection*

This community detection algorithm follows the top-down approach and aims at optimizing the community

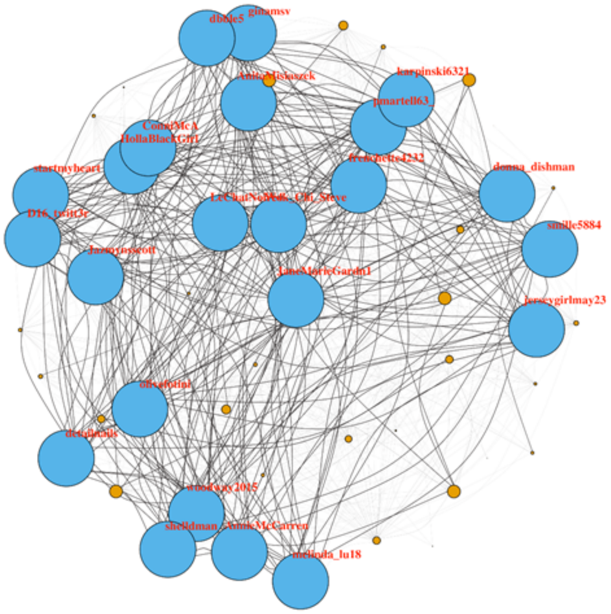


Figure 3. User tweet network (50% filtered) with clusters represented according to the Fastgreedy community

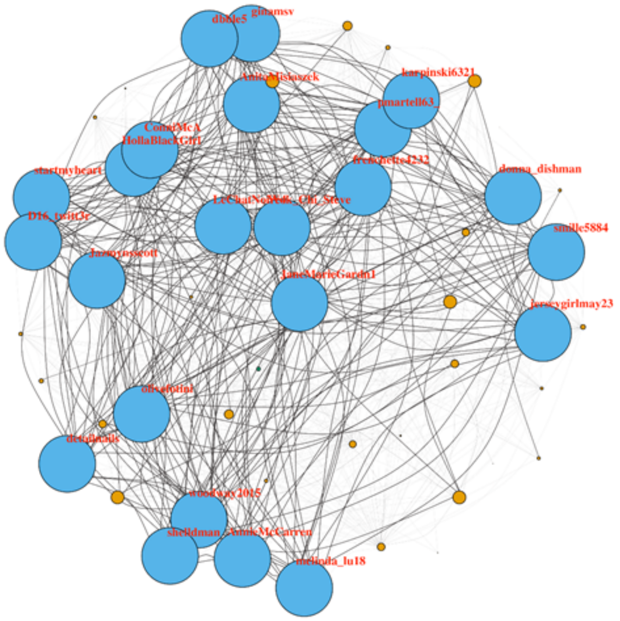


Figure 4. User tweet network (50% filtered) with clusters represented according to leading Eigenvector community

structure in terms of modularity. With each iteration, we split the network graph so that there is a significant increase in the network's modularity. The leading eigenvector is utilized for the network modularity matrix to determine the network. The algorithm avoids further splitting highly and strongly connected communities. In Figure 4, the user tweet network after 50% filtered is shown. The clusters are represented according to the leading eigenvector community. The network is generated by using weighted degrees with colors corresponding to different community clusters in the leading eigenvector community. The heatmap represents the more influential nodes in the network with larger node sizes and those with lesser influence with smaller node sizes.

E. Multilevel Community Detection

Multilevel community analysis has been utilized to study the network characteristics of community formation. It is started with the assumption that this weighted network has N number of nodes. Each node is considered to be in a different community, resulting in the number of communities equal to the number of nodes (i.e. N). Then, we perform an intensive computation for each ith node by considering all its neighbors' j' and finding the resulting modularity if 'i' was removed from its community and placed in the community of 'j'. Then, we move 'i' to the 'j' community that gives the maximum modularity gain. If no such condition arises, with a positive gain in modularity, we leave 'i' in its own community. We repeat the process iteratively until no further positive gain is possible in the modularity. In this iterative procedure, any given node may be (is usually) considered more than once up to several counts. The result of the algorithm is also based on the sequence in which we consider the nodes. Although it has a little effect there is a variation in the output. In Figure 5, the network is plotted using the weighted degree with the colors representing clusters in the multilevel community.

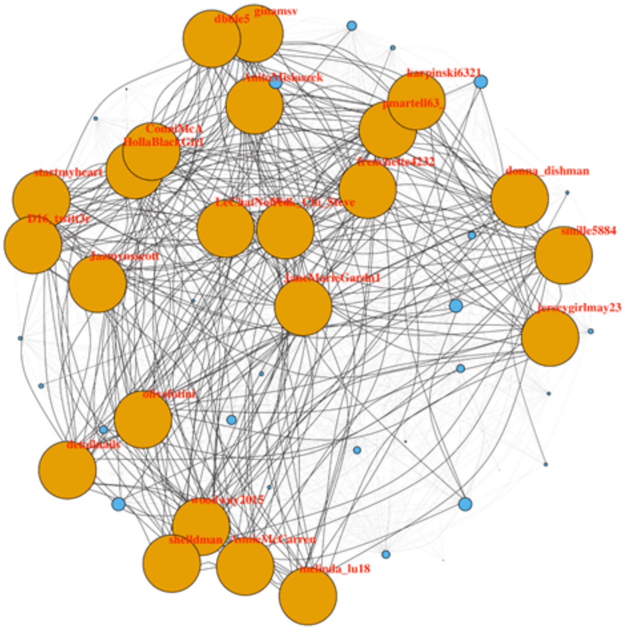


Figure 5. User tweet network (50% filtered) with clusters represented according to multilevel community

F. Edge-Betweenness Community

Edge-Betweenness algorithm follows the concept of hierarchical decomposition. We consider the initial network and then start eliminating edges based on the decreasing behavior of Edge-Betweenness (absolute shortest paths passing through the edge). The edges that act as a connection between different communities have a higher probability of being present in multiple shortest paths depending on the fact that in many cases, they might be the only source to move from one community to another [26]. The Edge-Betweenness algorithm generates good results but is comparatively slow due to its computational intensity of i*j calculations of Edge-Betweenness values and further recalculations after any edge is removed. It can be observed from the generated graph that the Edge-Betweenness community (Figure 6) is not very useful in our case as it results in a community through edge weight filtration. The number of nodes with the shortest path is almost the same. The reason is the high connectivity and density of the network.

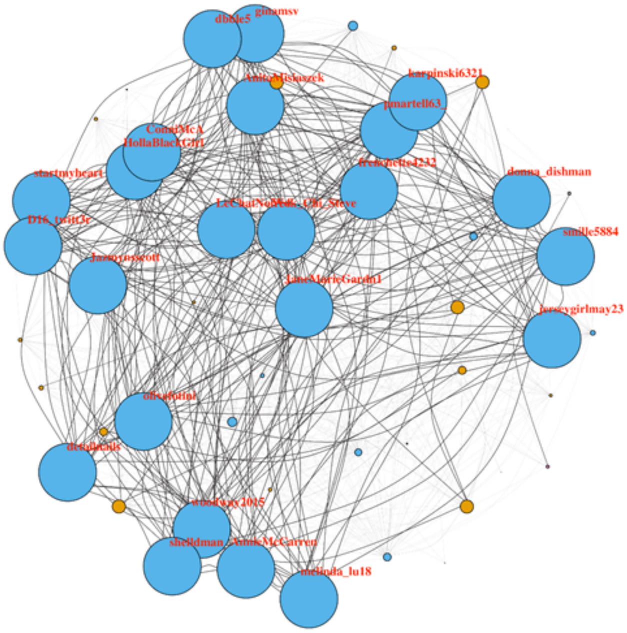


Figure 6. User tweet network (50% filtered) with clusters represented according to Edge-Betweenness community

G. Label Propagation Community

Label propagation algorithm, also known as connected component detection. It follows a simple concept of assigning each node one of the k labels. The algorithm then iteratively repeats the process of reassigning labels to nodes such that each node is synchronously reassigned its neighbor node's label having the highest frequency. The algorithm ends when each node has a label that is one of the most frequent from its neighborhood. Although the label propagation algorithm is fast, it does not produce results based on the initial network configuration. Thus, converging the network requires the iterations to be performed a large number of times (i.e. more than 1000 in some cases depending on the nature and behavior of the network). For the user tweet network, the algorithm does not produce the desired clusters of communities due to the high connectivity of the graph. Even after filtration, we have avoided this approach for its computation intensity.

H. Community Persistence Throughout Network Filtration

The network is analyzed using different community detection algorithms. Each produces a result with different variations, according to their concept and approach. Figure 7 compares four community detection methods (i.e., Multilevel community, Walktrap community, Fastgreedy community, and Leading Eigenvector community). While in Figure 8, a comparison among four community cluster distributions (i.e., Multilevel, Walktrap, Fastgreedy, and Leading Eigenvector) is shown. It increases the minimum threshold value for edge weights in our filtered network.

6. Conclusion

As the world grows to connect socially, it becomes more important to find the social influence of individuals in any network. By keeping in mind that influence can be used for different purposes. Such as advertising and marketing, campaigning, implementing a cause, running a successful business, promoting your thoughts. On the other side, the most influential nodes can also be used to destroy the network. The Walktrap, Spinglass, Fastgreedy, Leading Eigenvector, Multilevel and Edge-Betweenness algorithms are used for community detection and comparative analysis of cluster formation and information propagation within the network. The heat maps represented the strength of connections through edge weights and how to reach the other nodes through their size. The graphical illustrations visibly highlight the community structure formation when following each of the used algorithms. Another major finding is that the Leading Eigen Vector algorithm forms communities based on different numbers of clusters making the network. They are less dependable and vulnerable as breaking the link from a single cluster might lead to the complete isolation of any node. There are multiple clusters in the network. If a node breaks its link from one cluster, there always will be an alternate path to reaching that node through another cluster, increasing the network connectivity and reachability. With this comparative analysis, the community formation and targeting a certain set of users or audience will become convenient and quicker. One needs to target the most centralized and influential nodes in the network to spread information. In the future, other social networks (Facebook, Instagram, LinkedIn, etc.) can also be analyzed in this way.

Acknowledgment

The author would like to thank Arab Open University Bahrain for the administrative support.

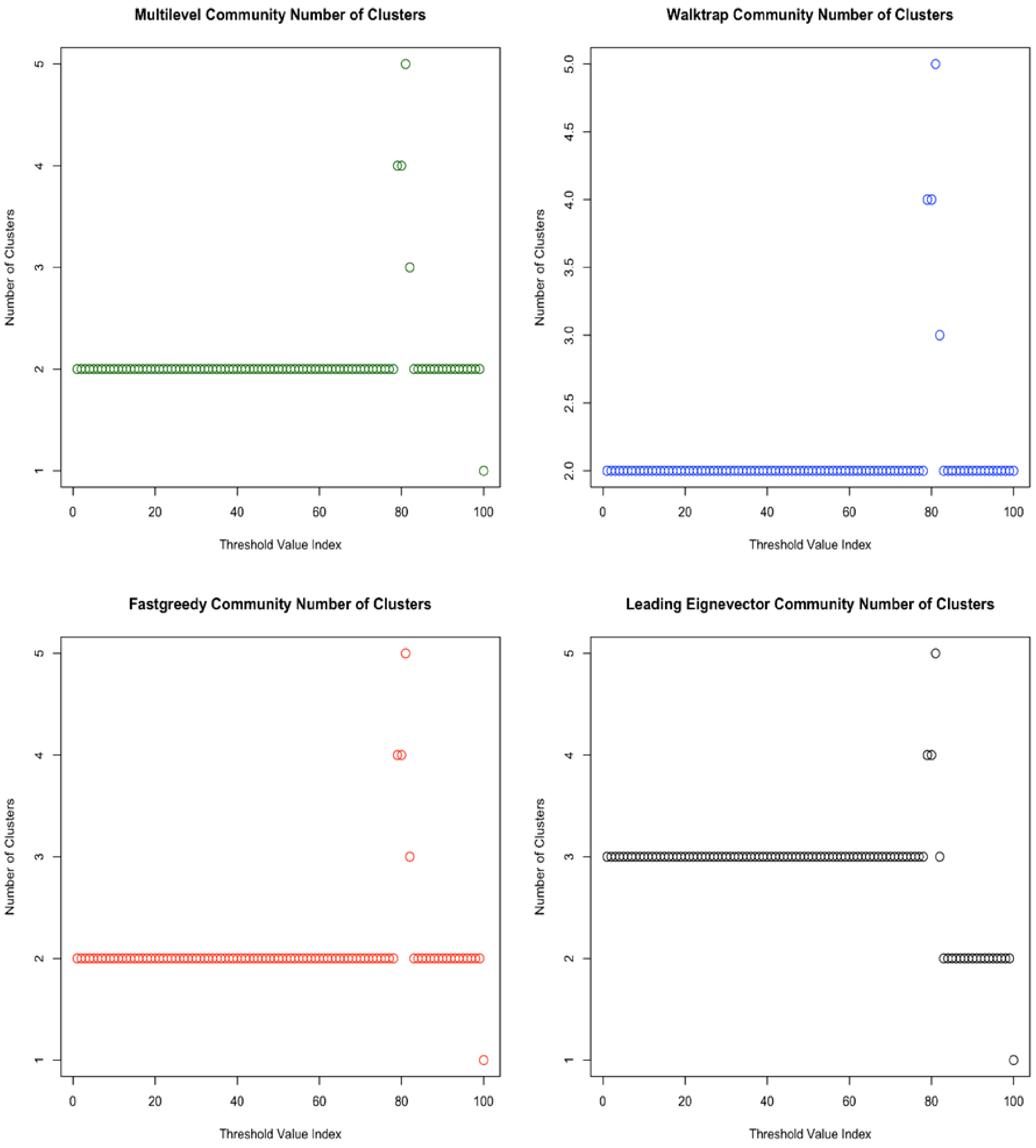


Figure 7. comparison among four community detection methods

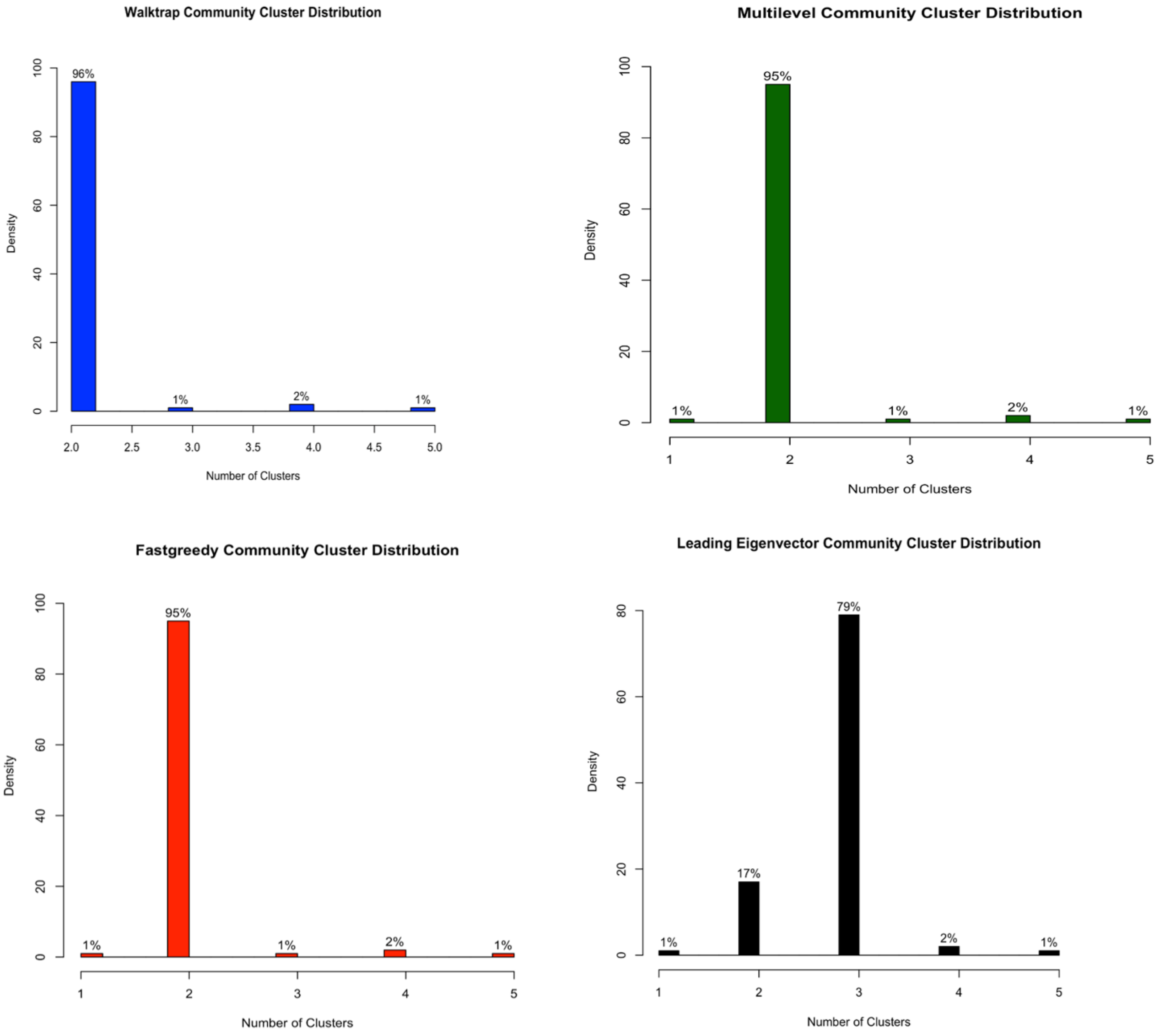


Figure 8. comparison among four community cluster distributions

References

[1] D. A. Gubanov and A. G. Chkhartishvili, “Conceptual approach to online social networks analysis,” *Upravlenie Bol’shimi Sistemami*, vol. 45, pp. 222–236, 2013.

[2] J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, “Information propagation in online social networks: a tie-strength perspective,” *Knowledge and Information Systems*, vol. 32, no. 3, pp. 589–608, 2012.

[3] L. Liu, F. Zhu, M. Jiang, J. Han, L. Sun, and S. Yang, “Mining diversity on social media networks,” *Multimedia Tools and Applications*, vol. 56, no. 1, pp. 179–205, 2012.

[4] H. A. M. Malik, N. Mahmood, M. H. Usman, K. Rziwan, and F. Abid, “Analysis of airport network in pakistan utilizing complex network approach,” *network*, vol. 10, no. 1, 2019.

[5] H. A. M. Malik, N. Mahmood, M. H. Usman, and F. Abid, “Unweighted network study of pakistani airports,” in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2019, pp. 1–6.

[6] H. A. M. Malik, A. W. Mahesar, F. Abid, and M. R. Wahiddin, “Two-mode complex network modeling of dengue epidemic in selangor, malaysia,” in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*. IEEE, 2014, pp. 1–6.

[7] J. Li and J. Zhou, “Chinese character structure analysis based on complex networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 380, pp. 629–638, 2007.

[8] H. A. M. Malik, F. Abid, N. Mahmood, M. R. Wahiddin, and A. Malik, “Nature of complex network of dengue epidemic as a scale-free network,” *Healthcare Informatics Research*, vol. 25, no. 3,

pp. 182–192, 2019.

[9] H. A. M. Malik, F. Abid, A. R. Gilal, and A. S. Raja, "Use of cloud computing in hajj crowed management and complex systems," in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. IEEE, 2017, pp. 1–5.

[10] H. A. M. Malik, "Complex network formation and analysis of online social media systems," *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES*, 2021.

[11] H. A. M. Malik, F. Abid, M. R. Wahiddin, and Z. Bhatti, "Robustness of dengue complex network under targeted versus random attack," *Complexity*, vol. 2017, 2017.

[12] H. A. M. Malik, F. Abid, M. R. Wahiddin, and A. Waqas, "Modeling of internal and external factors affecting a complex dengue network," *Chaos, Solitons & Fractals*, vol. 144, p. 110694, 2021.

[13] H. A. M. Malik, A. W. Mahesar, F. Abid, A. Waqas, and M. R. Wahiddin, "Two-mode network modeling and analysis of dengue epidemic behavior in gombak, malaysia," *Applied Mathematical Modelling*, vol. 43, pp. 207–220, 2017.

[14] R. S. Monclar, J. Oliveira, F. F. de Faria, L. Ventura, J. M. de Souza, and M. L. M. Campos, "Using social networks analysis for collaboration and team formation identification," in *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2011, pp. 562–569.

[15] Y. Xia, X. Ren, Z. Peng, J. Zhang, and L. She, "Effectively identifying the influential spreaders in large-scale social networks," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8829–8841, 2016.

[16] A. V. Mantzaris, "Uncovering nodes that spread information between communities in social networks," *EPJ Data Science*, vol. 3, pp. 1–17, 2014.

[17] P. van den Berg, T. Arentze, and H. Timmermans, "A multilevel path analysis of contact frequency between social network members," *Journal of geographical systems*, vol. 14, no. 2, pp. 125–141, 2012.

[18] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2015.

[19] H. MALIK, A. Waqas, F. Abid, A. GILAL, A. MAHESSAR, and Y. Koondar, "Complex network of dengue epidemic and link prediction," *Sindh University Research Journal-SURJ (Science Series)*, vol. 48, no. 4, 2016.

[20] Y. Wang, W. Huang, L. Zong, T. Wang, and D. Yang, "Influence maximization with limit cost in social network," *Science China Information Sciences*, vol. 56, no. 7, pp. 1–14, 2013.

[21] B. Latour, *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford, 2007.

[22] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

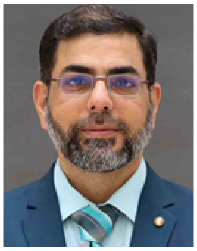
[23] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[24] A. C. Howe, D. B. Tindall, and M. C. Stoddart, "Drivers of tie formation in the canadian climate change policy network: Belief homophily and social structural processes," *Social Networks*, 2021.

[25] D. Tischer, "Collecting network data from documents to reach non-participatory populations," *Social Networks*, 2020.

[26] G. D. Gospodinov, "Isis tweet network analysis." [Online]. Available: https://www.kaggle.com/ggospodinov/tweet-analysis2/notebook

Author Biography



Hafiz Abid Mahmood Malik is a faculty member at Arab Open University Bahrain. He is 'Senior Fellow AdvanceHE (SFHEA)', UK. Being a member of IEEE, he served the IEEE Bahrain chapter as Vice-Chair for Students Activities. Dr. Malik has won gold and silver medal in some research poster competitions (national/international). He has received best paper awards in two international conferences. Dr. Malik is serving some international journals as Editor-in-chief and editorial board member. He has provided his volunteer services to several national/ international conferences as a Technical Committee member, Reviewer, Secretary, and acted as a chairperson of an international conference. Dr. Malik has also delivered his talk as a 'Keynote Speaker' at international conferences.