International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-9, Issue-5, October 2022
https://doi.org/10.55524/ijirem.2022.9.5.16
Article ID IJIRD-1202, Pages 113-118
www.ijirem.org

Influence Maximization in Online Social Networks Using Community Structures

Naima Bashir[1], and Dr. Ashish Oberoi[2]

[1]M. Tech Scholar, Department of Computer Science & Engineering, RIMT University, Mandi Gobindgarh, Punjab, India
[2]Professor, Department of Computer Science & Engineering, RIMT University Mandi Gobindgarh, Punjab, India

ABSTRACT- Online social networks (OSNs) have dominated modern life on a global scale. The immense popularity of online social networks increases day by day as they help us in modelling various types of processes like viral marketing, rumor controlling, collaborative filtering, market prediction and controlling diseases spread. In the realm of complex networks research, influence maximisation in social networks has long been a challenging task. Influence maximisation is the method of identifying k-seed nodes or influential nodes in order to increase overall influence in a network. Ranking the nodes using network node-centrality metrics is one of the conventional techniques for finding prominent nodes in a social network. However, estimating global centrality metrics like betweenness centrality is computationally exhaustive and typically not scalable for very large size networks such as a country's whole population. In this paper, we provide a novel approach for extracting communities from the underlying social network to identify prominent nodes aka "influential nodes”. Experimental results indicate that the seed nodes identified by the proposed approach have high betweenness centrality in the social network thus rendering the proposed approach significant.

KEYWORDS- Information Diffusion, Influence maximisation, Centrality Measures, Betweenness Centrality, Community Structures.

I. INTRODUCTION

With the rise of social networks, tens of millions of individuals now communicate and produce enormous volumes of data. There has been a lot of interest in information exchange due to the growing usage of social networks, as a piece of data may swiftly spread through "word-of-mouth" distribution among friends and acquaintances. This phenomenon of information dissemination has been helpful in various applications[1], including viral marketing, controlling rumours and technological advancements. As a result, researchers from a variety of fields have focused on information diffusion through online social networks, including computer science, mathematics, medicine, sociology, and many more. Due of potential financial gain of influence maximisation, it has lately garnered a lot of attention as a key algorithmic issue in information diffusion work. The goal of influence maximisation (IM) is to select k-nodes (seed nodes or actors) in a social network that may have the greatest overall impact on the network. Kempe et al. [2] first comprehensively investigate influence maximisation as a discrete optimization problem. Influence maximisation has substantial research obstacles despite having a wide range of possible applications. The first difficulty is simulating the information dissemination mechanism in a network, which has a big impact on how quickly each influence seed set spreads in the context of influence maximisation. Second, there are several theoretical issues with the impact maximisation problem in general. It has been established that, for the majority of diffusion models, finding the best solution for influence maximisation is NP-hard. These theoretical findings suggest that it is very challenging to retrieve an optimal seed set when scaling to very large social network. Third, a variety of node centrality measures, such as betweenness centrality, are employed to identify the influential nodes in the network. Each measure identifies certain nodes as influential based on the nature of process. However, it is computationally expensive and involves storing the entire network in memory to compute global centrality metrics like betweenness centrality. As a result, the majority of influential-node detection algorithms based on centrality measurements are ineffective in detecting very vast social networks. Real-world social networks have communities built into their fundamental structure as one of its defining features. In a social network, communities are made up of nodes that are functionally, cognitively, and physically closer to one another than they are to other nodes outside of the community. The problem of identifying densely linked groups inside social networks is vital to the field of community detection since these groups typically act as the functional units of a networked system. Community detection from social networks has gained a lot of attention nowadays, and the area is still expanding quickly. As shown in Fig. 1 C1, C2, C3 and C4 are communities of a network. In these communities, the intra-density is more than inter- density which means the ratio of number of edges by number of nodes is more within a community than across communities.

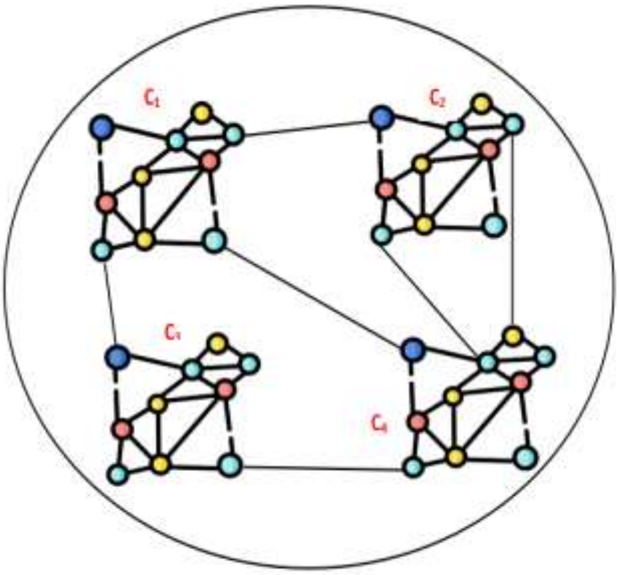

Figure 1: Network consists of four communities C1, C2, C3 and C4

By combining strategies and tools from several academic fields, including applied physics, bioinformatics, mathematics, social sciences and computer science numerous methodologies have been developed. However, which algorithms are reliable and ought to be applied, though, is yet unclear. The issue of reliability is complicated because it requires accepted definitions of community and partitioning, which are currently lacking. This effectively implies that, despite the vast literature on the subject, there is still no agreement among experts on what a network containing communities looks like. All of these disadvantages can be minimised if we can effectively extract information from community structures.

II. LITERATURE SURVEY

Social network research has recently focussed on influence maximisation problem. Finding the first users who will influence the greatest number of additional users in a network is known as influence maximisation. Bryan Wilder et al[3]formulates an algorithm which queries individual nodes to learn their links. The algorithm starts by locating a seed set which will be influential as the global optimum by using certain queries. The proposed ARISEN algorithm uses querying leveraging community structure by finding seed set of influential nodes. It was applied for undirected graph. Seed set of k nodes are selected with aim of increasing expected size of resulting cascade influence. Initially the seed nodes are only active nodes in network then each node tries to activate connected neighbor with a probability which is assumed to be same for all edges.

Yuxin Zhao [4] proposed a community structure-based approach for locating influential nodes in a network. The approach employs an influence maximisation algorithm based on label propagation, which consists of two steps: identifying the seed set and then modifying the seed set further in the second step i.e., in label phase.

To more precisely determine a node's influence, Jiali Dong [5] proposed an effective approach based on semi-local centrality, a ranking mechanism. Using a random walk as a starting point, this approach captures significant surrounding nodes. It is founded on the premise that if a node is surrounded by many influential nodes, the node has a high possibility of being influential too. Starting from the source node v, it does a random walk several times to all the paths in its neighborhood, collecting all the nodes in the path. The acquired set cardinality is utilised to determine the centrality of the source node.

Qian Wang et al. [6] proposed a clustering degree algorithm (CDA)-based strategy to identifying the most significant nodes in a weighted complex network. Utilizing Kendall's tau coefficient, CDA may identify influential spreaders with great accuracy. In accordance with the degree and node strength, a weighted node's degree has been established. The topological network structure is used to compute the clustering degree in order to determine the neighbours' contributions and capacity to propagate information.

Amrita et al. [6] suggested a weighted k shell degree neighbourhood technique that does not need the completeness of the network structure in order to discover influential spreaders that are effective in the spreading process.

III. OVERLAPPING COMMUNITY DETECTION METHODS

Given the characteristics and development of community structures, community detection has attracted a lot of interest. Various techniques connected to community detection have since been put out in literature. Communities in social organisations frequently lead to important utilitarian gatherings, making the challenge of identifying them quite alluring. The complexity of network detection is also influenced by several variables, such as the global or local characteristics of network, communities it covers, the dynamic nature of the system etc. The various community detection algorithms used in our work are:

- The Clique Percolation Method is used by CFinder [7] to detect k-cliques in a network and identify communities that overlap. This method finds clusters of closely spaced overlapping nodes in a network. The most popular technique for finding covering networks is the CPM (also known as CFinder), which depends on permeating k-factions from a basic system.
- COPRA (Community Overlap Propagation Algorithm) [8] , an adaptation of Raghavan, Albert, and Kumara's label propagation algorithm [9]. The main improvement is the addition of details about many communities throughout the labelling and spreading process.
- SLPA [10]is a community detection algorithm in which nodes can be either information consumers or producers and have multiple labels. Without removing the previously saved label, a node continues to collect data on the observed labels. There is a direct correlation between the frequency at which a node notices a label and the frequency at which other nodes also detect it. The minimum possibility that a label will appear before it is erased from the node's cache is required as a baseline parameter.
- Demon [11] is a realistic solution to community detection that is based on the modular architecture of

networks. Using the label propagation method, each node begins by determining and choosing the communities that are present in its near surrounds. By totalling up all the votes, these small communities are combined to form a bigger group, which results in the construction of overlapping modules. However, a minimal threshold parameter is required for this method.

- AFOCS [12] is a two-stage method for detecting and examining the expansion of overlapping network communities in large dynamic systems. Using this method, it is possible to track the evolution of networks over time in a dynamic system where each network is defined by a number of significant developmental milestones.

All of the methods outlined above detect nodes that are shared by several communities. These nodes are referred to as overlapping nodes, as depicted in Figure 2. These nodes are of prime importance for various applications like in viral marketing, rumour detection etc. Fig.2 shows an overlapped node shared between three communities. This node has an overlapping membership of 3. Overlapping membership of a node tells us about the number of communities a node is shared.

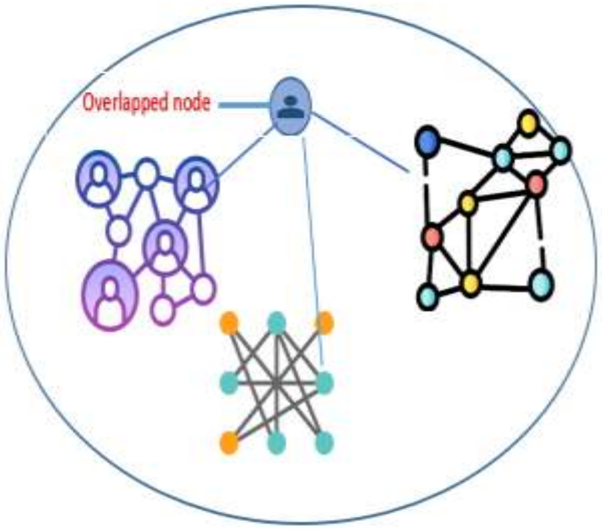

Figure 2: Overlapping node and communities

IV. PROPOSED METHODOLOGY

Social networks are expanding at an alarming rate, and their value is rising as they aid us in modelling all sorts of processes by probing social structure utilising various ideas from graph theory and network theory. The data from the graph might be difficult to grasp due to the overcomplexity of networks. The suggested methodology is predicated on the idea that every underlying network includes communities. Communities can be hierarchical, dynamic, or static. A number of graph centrality metrics, including betweenness centrality, can be used to identify influential nodes. Depending on the type of the process, each metric captures the node as influential. One of the primitive and global centrality measure is betweenness centrality, which has several major issues such as knowing whole network information in advance, the entire network must be in memory, it cannot be used for dynamic networks, and so on. But it is commonly acknowledged that users of social networks may be recognised by the different communities they belong to, giving rise to the concept of community overlap. They may belong to more than one functional unit, and they may also serve as a bridge for information flow. The suggested approach has two levels of operation:

- Level-1: - In the first level, nodes which are having high betweenness centrality are treated as influential. In order of decreasing betweenness centrality, these nodes are selected. To identify overlapping communities in a network, we employ community identification algorithms as COPRA, CFinder, SLPA, DEMON, and AFOCS.
- Level-2: - In this level, we extract overlapping nodes from overlapping community structures and calculate their overlapping membership. We correlate the overlapping membership count and betweenness centrality and check whether they are correlating positively or not. From this we can infer that high betweenness nodes have high overlapping membership. Thus, instead of using betweenness as a metric to determine which nodes are influential, we may use overlapping membership to identify such nodes in the network.

V. RESULTS AND DISCUSSIONS

Due to the necessity to compute global network measures like betweenness centrality, conventional overlapping community detection algorithms are not easily scalable to large-scale social networks. Due to this problem, our attention went towards finding a new metric for influence maximisation problem. We use 5 overlapping community detection methods as shown in Fig.3. The number of communities identified by overlapping methods varies greatly. The number of communities identified by any algorithm doesn't talk about the optimality of an algorithm. It just tells us about how many modules are present in network.

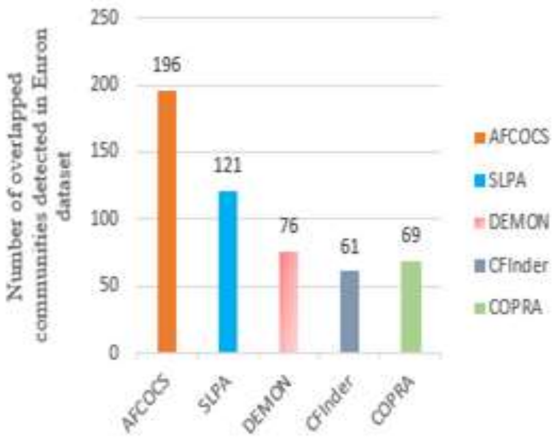

Figure 3: Number of communities formed using Enron dataset

Table 1: Statistics of Enron dataset and number of communities formed by various methods

Number of Nodes	13750
Number of edges	175253
Average Degree	12.746
Number of Connected Components	10
Average Clustering Coefficient	0.207
Number of Communities formed by AFOCS, SLPA, DEMON, CFinder and COPRA algorithm respectively	196, 121, 76, 61, 69

The Enron dataset comprises of about 13k nodes with 175k edges. The average degree of a node is approximately 13 which gives us a fair idea of how nodes are well connected in a network. In Table 1 information regarding various parameters about Enron dataset are shown and the number of communities formed by various community detection algorithms is also shown. The data in Table 1 tells us about the overall composition of a network whether network is sparsely or densely connected.

Figure 4 shows lot of fluctuations in community formation. It is just because of the fact that Facebook dataset is densely connected and finding communities in such a network is computationally expensive. One of the primitive algorithms is CFinder which did not find much communities in a network. So, CFinder cannot be used for dense networks. This is the major drawback of CFinder. It can’t be used for dynamic networks as well. Community formation in dense networks by CFinder leads to loss of so many nodes which in the real sense is loss of information. So, we can infer that CFinder isn’t an optimal algorithm in case of dense networks. SLPA also shows poor performance for identifying communities in Facebook network. The number of parameters used in SLPA also affect the community formation. DEMON and AFOCS both show good result in community formation. The advantage of DEMON over others is that it can be used for dynamic networks as well. It can be seen from the Fig. 4 that number of communities formed by DEMON algorithm is more in case of Facebook dataset which is much denser than Enron dataset.

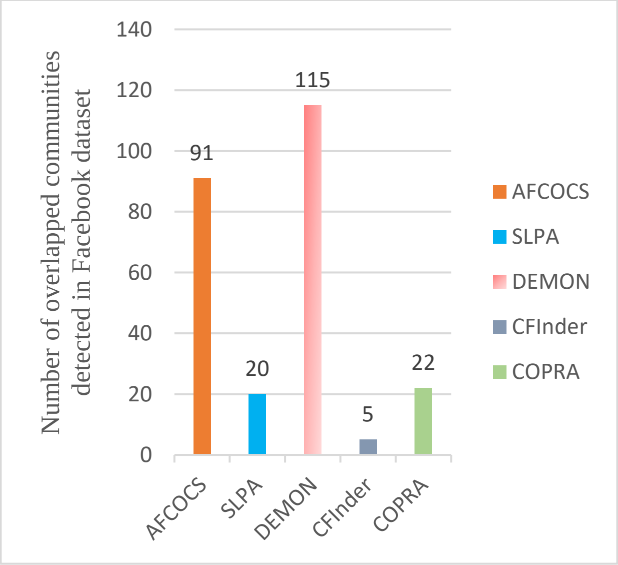

Figure 4: Number of communities formed using Facebook dataset

Table 2: Statistics of Facebook Dataset and number of communities formed by various methods

Number of Nodes	4039
Number of Edges	60264
Average Degree	29.84
Number of connected Components	1
Average Clustering Coefficient	0.625
No. of Communities formed by AFOCS, SLPA, DEMON, CFinder and COPRA algorithm respectively	91, 20, 115, 5, 22

Facebook dataset comprised of 4k nodes and 60k edges. The average degree of node is 30 which means every node has approximately 30 connections with other nodes which tells us about that Facebook network is too dense. The parameters of Facebook dataset and number of communities formed by various methods are shown in Table 2.

Fig 5 and Fig.6 shows that nodes which lie in top 0.5 % and top 1% in accordance with betweenness centrality are influential nodes. We infer that top 0.5 % and 1% nodes with high betweenness possess high overlapping membership. We infer that top nodes possess high overlapping membership in accordance with betweenness centrality. We measure an overlapped node's likelihood of being listed among a social network's top influential nodes. In this regard, we determine the chances that overlapping nodes will be in the top 0.5% and 1% of the network's influential nodes. The findings in Fig. 5,6,7,8 support the hypothesis that highly overlapping nodes in a social network directly correspond to nodes that are very influential since these nodes are more likely to rank in the top 0.5% to top 1% of significant nodes in the network.

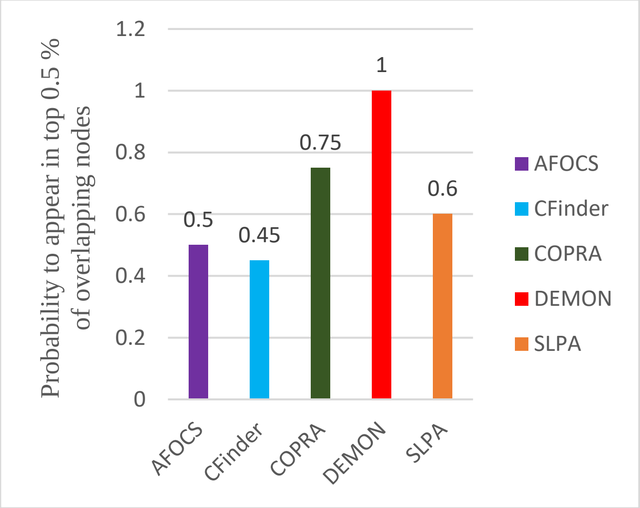

Figure 5: Occurrence of top 0.5 % overlapping nodes in accordance with betweeness centrality in Enron dataset

Total Number of Nodes in Enron dataset: - 13750
Number of Nodes in 0.5% of total Nodes: - 69

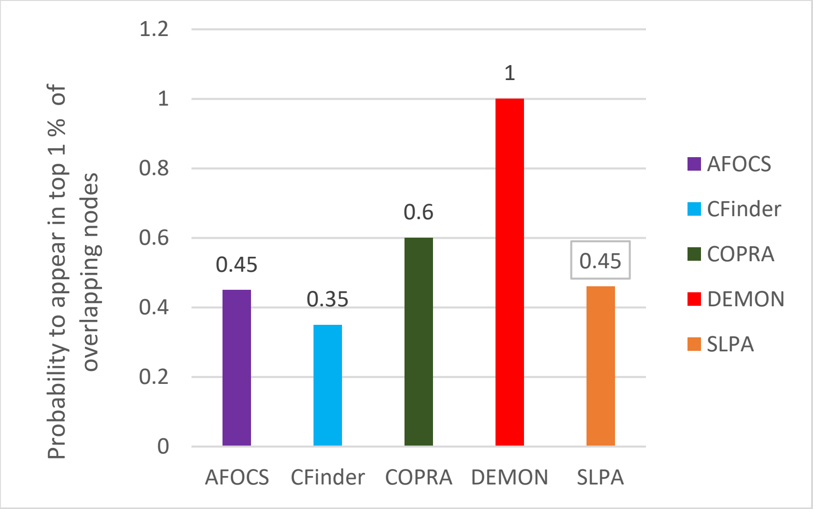

Figure 6: Occurrence of top 1% overlapping nodes in accordance with betweeness centrality in Enron dataset

Number of Nodes in 1% of total Nodes: - 138

Table 3: Results of various overlapping community detection algorithms as shown in Fig. 5 and Fig. 6 using Enron Dataset

	Community Detection Algorithms				
	AFOCS	CFinder	COPRA	DEMON	SLPA
Number of Overlapping Nodes in Top 0.5 % of total nodes.	345	31	52	69	42
Occurrence of top 0.5 % overlapping nodes	0.5	0.45	0.75	1.0	0.6
Number of Overlapping Nodes in Top 1% of total nodes.	62	48	83	138	62
Occurrence of top 1 % overlapping nodes in accordance	0.45	0.35	0.60	1.0	0.45

The count of overlapping nodes that lie in top 0.5% and top 1% of Enron network are shown in Table 3. The probabilities of nodes that lie in top 0.5% and top 1% of total nodes with high overlapping membership in accordance with the high betweenness centrality are also shown in Table 3.

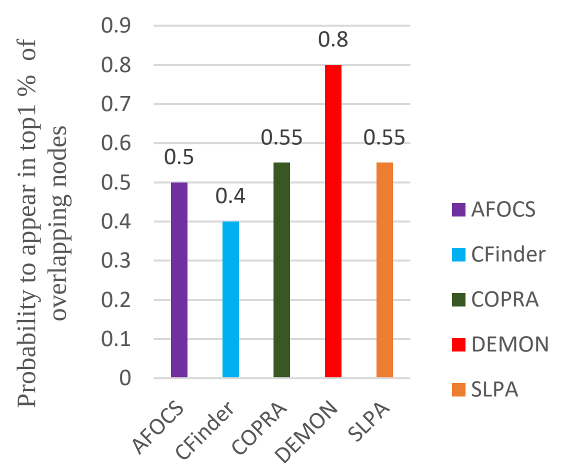

Figure 7: Occurrence of top 1% overlapping nodes in accordance with betweeness centrality in Facebook…

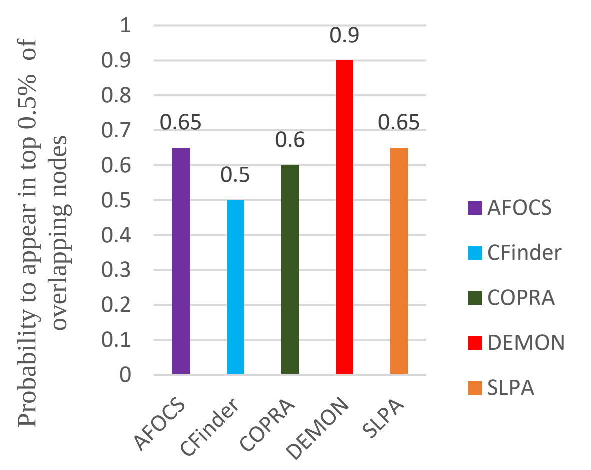

Figure 8: Occurrence of top 0.5% overlapping nodes in accordance with betweeness centrality in Facebook…

Total Number of Nodes in Enron dataset: - 4039
Number of Nodes in 0.5% of total Nodes: - 20

Number of Nodes in 1% of total Nodes: - 40

Table 4: Results of various overlapping community detection algorithms as shown in

	Community Detection Algorithms				
	AFOCS	CFinder	COPRA	DEMON	SLPA
Number of Overlapping Nodes in Top 0.5 % of total nodes.	13	10	12	18	13
Occurrence of top 0.5 % overlapping nodes	0.65	0.5	0.6	0.9	0.65
Number of Overlapping Nodes in Top 1% of total nodes.	20	16	22	32	22
Occurrence of top 1 % overlapping nodes in accordance	0.50	0.4	0.55	0.80	0.55

Fig. 7 and Fig. 8 using Facebook Dataset.

Table 4 shows the count of overlapping nodes that lie in top 0.5% and top 1% of Facebook network. The probabilities of nodes that lie in top 0.5% and top 1% of total nodes with high overlapping membership are also shown in Table 4. These top nodes have the power to influence the whole network. These nodes act as seed nodes in the information diffusion process, viral marketing and other such processes. Identification of these nodes in a very large-scale network reduces the complexity of information dissemination to a great extent.

VI. CONCLUSION

Information and ideas flow in social systems through communication between different social actors. As a result, identifying the key factors in information dissemination is critical. It's crucial to identify influential persons or organisations on social networks so that they may reach as many people as possible. Similar to this, community structure is crucial to comprehending the dynamics of a social system. Numerous research has been carried out in an effort to pinpoint important spreaders. We are attempting to concisely summarise numerous cutting-edge techniques and algorithms employed by different researches. One of our study objectives is to discover the best way for identifying prominent nodes in a network. We have tested various community detection algorithms on Facebook, Friendship and Enron datasets which contains set of nodes and respective edges. We find that overlapping membership count, rather than other global centrality measures such as betweenness centrality, may be used to identify influential nodes in a network. Furthermore, we don't require any prior knowledge of the whole network to calculate the overlapping membership count of a node, which decreases our computational overhead. We determine that the algorithm that best discovers overlapping nodes while being computationally inexpensive is optimal.

REFRENCES

[1] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, “Information diffusion in online social networks: A survey,” SIGMOD Rec, vol. 42, no. 2, 2013, doi: 10.1145/2503792.2503797.

[2] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” Theory of Computing, vol. 11, 2015, doi: 10.4086/toc.2015.v011a004.

[3] B. Wilder, N. Immorlica, E. Rice, and M. Tambe, “Maximizing influence in an unknown social network,” 2018. doi: 10.1609/aaai.v32i1.11585.

[4] Y. Zhao, S. Li, and F. Jin, “Identification of influential nodes in social networks with community structure based on label propagation,” Neurocomputing, vol. 210, 2016, doi: 10.1016/j.neucom.2015.11.125.

[5] J. Dong, F. Ye, W. Chen, and J. Wu, “Identifying Influential Nodes in Complex Networks via Semi-Local Centrality,” in Proceedings - IEEE International Symposium on Circuits and Systems, 2018, vol. 2018-May. doi: 10.1109/ISCAS.2018.8351889.

[6] A. Namtirtha, A. Dutta, and B. Dutta, “Weighted kshell degree neighborhood: A new method for identifying the influential spreaders from a variety of complex network connectivity structures,” Expert Syst Appl, vol. 139, 2020, doi: 10.1016/j.eswa.2019.112859.

[7] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, “CFinder: Locating cliques and overlapping modules in biological networks,” Bioinformatics, vol. 22, no. 8, 2006, doi: 10.1093/bioinformatics/btl039.

[8] J. Xie, S. Kelley, and B. K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study,” ACM Comput Surv, vol. 45, no. 4, 2013, doi: 10.1145/2501654.2501657.

[9] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Phys Rev E Stat Nonlin Soft Matter Phys, vol. 76, no. 3, 2007, doi: 10.1103/PhysRevE.76.036106.

[10] J. Xie, B. K. Szymanski, and X. Liu, “SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process,” 2011. doi: 10.1109/ICDMW.2011.154.

[11] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, “DEMON: A local-first discovery method for overlapping communities,” 2012. doi: 10.1145/2339530.2339630.

[12] N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai, “Overlapping communities in dynamic networks: Their detection and mobile applications,” 2011. doi: 10.1145/2030613.2030624.