Towards a new approach for community detection algorithm in social networks

Sara AHAJJAM, Hassan BADIR , Mohamed EL HADDAD
Laboratory of Technologies and Communication
National School of Applied Sciences
Tangier, Morocco
Ahajjamsara@gmail.com
hbadir@gmail.com
elhaddad.mohamed@gmail.com

***Abstract*— In the recent years, social networks emerged rapidly and it's has become more complex. Social networks play an important role in the dissemination of information and the spread of influence. Several research studies are interested to the detection of the structure of complex networks, otherwise, to the community detection and leader detection. The major drawback of most of the proposed algorithms is that they require knowledge of number of communities to detect. Our approach proposes an algorithm for the detection of communities in social networks, especially the detection of leader nodes (influencer's nodes) without a priori knowledge of the number of communities or leaders to detect.**

Keywords—Community detection; leader node; centrality; complex networks; graph theory.

I. INTRODUCTION

Complex networks are a powerful tool for understanding the mechanisms of various systems. They are modeled by graphs with vertices denote the actors of phenomenon and links denote the interactions between vertices. These graphs represents different systems such as collaboration networks, citation network, protein interaction networks, WWW and social networks,.., etc. These networks are complex graphs with high local density and low overall density, they play a fundamental role in the diffusion of information, ideas and innovation, this advantage has been the subject of various parts that have moved towards these networks to achieve advertising goals (ads on Facebook), educational (LinkedIn), or political (Election of USA on Twitter). The key property of a real network is its community structure. The communities are groups of nodes, with more links connecting to nodes of the same group and comparatively fewer links connecting to nodes of different groups. Recent studies have verified that the way in which such nodes are organized plays a fundamental role in spreading processes [1]. Study the influence of role models can help us to better understand why some trends or innovations are adopted more quickly than others and how we can help advertisers and marketers to design more effective campaigns. This fact caused many researchers to look for an efficient method for finding top-k most influential people through social networks.

We are interested to study the problematic of detection of communities and leaders' nodes in complex network. Those nodes have high connectivity with the others nodes, and represent an optimization of the network while maintaining the same characteristics of the network. The major drawback of most of the proposed approaches is that they require knowledge of k leader and communities to detect. In this paper, we introduce a new approach to detect leaders' nodes and communities in the network without a prior knowledge of k nodes to detect. This problem has many applications such as: opinion propagation, studying acceptance of political movements or acceptance of technology in economics.

Actually, identifying influential nodes in networks, also regarded as ranking important nodes has become one of the three main problems in network-based information retrieval and mining [2]. In biological systems, we might like to identify the nodes that are keys to communities and protect them or disrupt them, such as in the case of lung cancer [2]. In epidemic spreading, we would like to find the important nodes to understand the dynamic processes, which could yield an efficient method to immunize modular networks [2]. Such strategies would greatly benefit from a quantitative characterization of the node importance to community structure. For example, suppose that we need to advertise a product in a country or we need to propagate news. For this purpose, we need to choose some people as a starting point and maximize the news or the products influence in the target society. The problem was introduced in [3] for the first time. After that in [4] the authors formalized the problem as follows: given a weighted graph in which nodes are people and edge weights represent influence of the people on each other, it is desired to find *k* starting nodes that their activation leads to maximum propagation In particular, we will focus our attention in one topological feature: centrality. Since those central nodes can diffuse their influence to the whole network

faster than the rest of nodes and they are the most influential spreaders.

II. Centrality Measures

The variety of measures of centrality comes from the fact that the importance of a node depends on other parameters such as connectivity and orientation in the graph and the nature of measurement of the entire network [5]. The work of Linton Freeman is probably one of the most important contributions to the analysis of social networks and networking in general. There are three varieties of measures of centrality [6], we'll cite the main ones:

- **Degree Centrality:**

It is defined as the number of links incident upon a vertex which means the number of edges a vertex has. For a graph *G :=(V, E)* with *n* vertices, the degree centrality C_d *(i,g)* for vertex is:

$$C_d(i,g) = \frac{d_i(g)}{n-1} = \frac{|N_i(g)|}{n-1} \quad (1)$$

Where: $d_i(g)$ is the degree of the node i.

- **Betweenness Centrality:**

Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not. For a graph G :=(V, E) with n vertices, the betweenness C_b (i) for vertex is computed as follows:

1. For each pair of vertices (s, t), compute all shortest paths between them.
2. For each pair of vertices (s, t), determine the fraction of shortest paths that pass through the vertex in question (here, vertex v).
3. Sum this fraction over all pairs of vertices (s, t).

The betweenness centrality is:

$$C_b(i,g) = \frac{2}{(n-1)(n-2)} \sum_{k \neq j, i \notin \{j,k\}} \frac{P_i(kj)}{P(kj)} \quad (2)$$

With : $\frac{P_i(kj)}{P(kj)}$ is the probability that i falls on a randomly selected geodesic connecting k and j.

- **Closeness Centrality:**

In graph theory closeness is a centrality measure of a vertex within a graph. Vertices that are 'shallow' to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. Closeness is preferred in network analysis to mean shortest-path length, as it gives higher values to more central vertices, and so is usually positively associated with other measures such as degree [7].

The closeness centrality is:

$$C_c(i,g) = \frac{n-1}{\sum_{i \neq j} d(i,j;g)} \quad (3)$$

Where: d(i,j ;g) is the geodesic distance between i and j.

- **Eigenvector Centrality:**

It simulates a mechanism in which each vertex affects all of its neighbors simultaneously [16]. Eigenvector centrality is a sort of extended degree centrality which is proportional to the sum of the centralities of the vertex's neighbors. A vertex has large value of eigenvector centrality score either if it is connected to many other vertices or if it is connected to others that themselves have high eigenvector centrality [17]. The eigenvector centrality score of the *i*th vertex in the network is defined as the *i*th component of the eigenvector corresponding to the greatest eigenvalue of the following characteristic equation:

$$A\chi = \lambda\chi \quad (4)$$

Where: A is the adjacency matrix of the network, λ is the largest eigenvalue of A, and x is the corresponding eigenvector.

III. Related works

The community detection algorithms have been the subject of several research papers. Most studies classify articles and research methods depending on the type of the algorithm. The community detection algorithms are belonging to two main types of approaches namely graph partitioning and classification. The major drawback of methods based on the partitioning of graphs is that they require a prior knowledge of the number and size of groups to determine [9]. Also, the leader detection approaches are divided to two mains types: global and local methods. The global method deals with all the network topology (betweenness centrality), while the local ones treat with local position, i.e. with the node (degree centrality). Reihaneh Rabbany Khorasgani et al. suggest a new approach to detect leaders nodes that takes into account the nodes that are not associated with no leaders. This algorithm is inspired from k-means, the k nodes to be detected will be randomly selected. Other nodes will be assembled at their closest leaders to form communities, and then find new leaders for each community around which gather followers until no node moves. For each community, the centrality of each member is calculated and the node with the highest degree is chosen as the new leader [10]. Another algorithm of leaders' nodes detection in complex networks proposed by Kernighan and Lin based on partitioning of graphs. This algorithm tries to find a section of the graph minimizing the number of edges between partitions by trading vertices between these partitions. The results of this algorithm are generated by introducing the size of each partition [11]. The results of these two algorithms vary according to the size and number of partitions which are introduced. Other proposed studies use classification. The classification was introduced to

analyze the data and partition based on a measure of similarity between partitions. The problem of communities detection can be seen as a problem of data classification for which we need to select an appropriate distance [12]. Indeed, the classification methods are generally appropriate for some networks that have a hierarchical structure. The result obtained by these methods depends on choice of similarity measure that used initially. Blondel et al. have proposed the Louvain method that put each node in a vertex. Other approaches are based on partitioned classification which is like the partitioning of the graph requires prior knowledge of size and number of communities to detect. Another study focuses on the spectral classification. In the Leader-Follower algorithm, we define some internal structure of a community. A community should be a clique and is formed of a leader and at least one "loyal follower" which is a node in the community without neighbors in any other community. The leader is a node whose distance is less than at least one of its neighbors. The nodes will be allocated to the community in which a majority of its neighbors belong by destroying the links arbitrarily. However, parasites communities i.e. leaders without loyal follower assigned will be removed from the network. This can cause a loss of information [13]. Yunlong Zhang et al propose a greedy algorithm based on user preferences (GAUP) to operate the top-k influential users, based on the model Extended Independent Cascade (EIC said that an active node v is active in t-1, has only one chance to activate all inactive neighbors). During each cycle i, the algorithm adds a record in the selected set such that the vertex S with the current set S maximizes propagation of the influence. This means that the vertex selected in round i is the one that maximizes the incremental propagation influence in this cycle. This algorithm calculates the user's preferences for different subjects, and combines traditional greedy algorithms and preferences calculated by LSI user and calculates an approximate solution of the problem of maximizing the influence of a specific topic. This algorithm provides a good result if k exceeds a certain threshold k≥ 15 and it is of complexity $O(n^3)$ [14]. More recently, in [14], the authors derive an upper bound for the spread function under the LT model. They propose an efficient UBLF algorithm by incorporating the bound into CELF. Recent research found that the location of the node in the network topology is another important factor when estimating the spreading ability. According to that, [15] propose a new approach to identify the location of node through the k-shell decomposition method, by which the network is divided into several layers. Each node corresponding one layer and the entire network formed the core-periphery structure. K-shell decomposition method indicates that the inner the layer is, the more important the node. However, in practical applications there are often too many nodes having the same index value by employing these two methods to distinguish which node is more powerful. Generally speaking, *DC* and k-shell decomposition are suitable to measure the spreading ability of nodes quickly but not very accurate. Another proposed algorithm use both global and local methods of centrality measures to effectively identifying the influential spreaders in large-scale social networks. The main idea, that it reduces the scale of network by eliminating the node located in the peripheral layer (namely relatively small *ks* value) that will not have much spreading potency comparing with the core node in general, and vice versa. This algorithm uses the k-decomposition centrality to deal only with the nodes in the core of the network. Hence, it reduce the scale of the network by ignoring the nodes whose *ks* value is small and the links connected them and retain the nodes in the core layers. At last, the global methods (i.e. betweenness centrality and closeness centrality) are used to rank the most influential spreaders [15]. A novel approach to detect communities and important nodes of the detected communities using the spectrum of the graph. It defines the importance nodes to community as the relative changes in the c largest eigenvalues of the network adjacency matrix upon their removal. It has two types of nodes, the core nodes who are the central nodes and the most important for the community, and the bridges node who connect the communities to each other's. The main drawback of this approach, it is that to have a better result, they need to know the number of partitions in the network and it cannot identify the important nodes in the small communities when the communities are in very different size has the same size. It cannot identify the important nodes in the small communities when the communities are in very different size [17].

Community and leader nodes detection approaches are diverse. Each proposed algorithm brings a new idea or improvement of existing algorithms. We will propose a new approach to detect communities and leader nodes in complex networks without a priori knowledge of number of communities to detect.

IV. PROPOSED ALGORITHM

Identifying social influence in networks is critical to understanding how behaviors spread. In order to detect the catalyst of this influence, we need to detect the central nodes that are responsible for the dissemination of influence. Analysis on social network datasets reveals that in each community, there is usually some member (or leader) who plays a key role in that community. In fact, centrality is an important concept [13] within social network analysis, which measures the relative importance of a vertex within the graph. Different from others methods, our approach detect leaders, and build communities around these leaders without a priori knowledge of k leader to detect.

Given an input dataset, the dataset is modeled as an undirected and unweighted graph $G = (V, E)$. V is the vertex set. Each vertex in V represents an element in the dataset. $|(G)|$ represents the number of vertices in G (or elements in the dataset). E is the edge set. Each edge represents a relationship between a pair of elements. Our approach has three steps as in "Fig. 1":

- **Nodes centrality:** For each node v in the network G, calculate the eigenvector centrality. Eigenvector centrality or Gould's index of accessibility [17] is a measure that describes how well connected an individual isbased on direct and indirect relationships

(i.e., it takes into account the connections of the individuals the focal individual is connected to [18]. Because eigenvector centrality is proportional to an individual's neighbors' centralities[19], more influential individuals will be more connected with other influential individuals. Lastly, embeddedness quantifies how isolatable an individual is or how involved in the network structure an individual is [20]. If all of an individual's connections with other individuals are severed, the individual would be isolated. Thus, higher embeddedness values mean that it is more difficult to isolate an individual [21].

$$Ax = \lambda x \qquad (1)$$

With: A is the adjacency matrix of the network and λ is the eigenvalue.

- **Nodes ranking:** we rank the nodes by the high centrality score, and choose the leader V_1 which is the node with the highest centrality.
- **Form community:** we calculate neighborhood function to find the neighbors of the leader node which is the node with the highest centrality score. We assign neighbors to the detected leader node to form a community.
- We remove the community i.e. the leader node and its neighbors from the network and we deal with the second node with the highest centrality until all the vertices (nodes) will be treated.

V. Results & Evaluations

To test our community detection using leader node algorithm, we ran the proposed algorithm on two networks described above:

Zachary's karate club network. This is a well-known benchmark network for testing community detection algorithms. The network is made up of 34 nodes and 78 edges, where every node represents a member of a karate club at an American university. If two members are observed to have social interactions within or away from the karate club, they are connected by an edge. Later, because of a dispute arising between the club's administrator and instructor, the club is eventually split into two factions centered on the administrator and the instructor, respectively.

TABLE I: Datasets properties

Datasets	Nodes	Edges	Real Communities
Zachary Karaté Club	34	78	2
Dolphins Social networks	62	159	2

- **Lusseau's bottlenose dolphin social network**: This is also a famous network widely used as a benchmark to validate community detection algorithms. It contains 62 nodes that represent bottlenose dolphins living in Doubtful Sound, New Zealand, and 159 edges that represent associations between dolphin pairs observed to co-occur more often than expected occasionally.

Fig. 2 and Fig. 3 shows the communities structure in the network for Zachary karate club and Dolphins social network respectively. We compared our community detection algorithm using leader nodes with other community detection algorithm: Label Propagation Algorithm (LPA) [22] and Leading Eigenvalue Algorithm (LEA) [23] using different metrics. For each network we calculate the quality of partition using the modularity Q.

$$Q = \sum_{i=1}^{k}(e_{ii} - a_i^2) \qquad (2)$$

where the first term, $\sum_{i=1}^{k} e_{ii}$ is the proportion of edges inside the communities, and the second term $\sum_{i=1}^{k} a_i^2$ represents the expected value of the same quantity in a random network constructed by keeping the same node set and node degree distribution, but connecting the edges between nodes randomly.

Also to evaluate our algorithm, we use the Adjusted Rand Index, the measure penalizes false negatives and false positives. Let a,b,c and d denote the number of pairs of nodes that are respectively in the same community in both G and R, in the same community in G but in different communities in R, in different communities in G but in the same community in R, and in different communities in both G and R. Then the ARI is computed by the following formula:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \qquad (3)$$

And we use the Normalized Mutual Information (NMI):

$$NMI(X,Y) := \frac{2I(X,Y)}{H(X) + H(Y)} \qquad (4)$$

where I(X,Y) The mutual information corresponds to the quantity of information shared by the variables. Its lower bound is, representing the independence of the variables (they share no information). The upper bound corresponds to a complete redundancy; however this value is not fixed.

The table below presents the result of our algorithm and the Label Propagation Algorithm and Leading Eigenvector Algorithm using the cited metrics.

The results in table 2 show that for Zachary Karaté Club dataset our algorithm provides the best result for ARI and NMI comparing to LPA and LEA algorithms, while for the modularity that present the quality of founded clusters is quite good compared to LEA which provide the highest one.

TABLE II: COMPARISON RESULTS OF ALGORITHMS.

Network	Algorithm	Communities	Modularity	NMI	ARI
Zachary Karaté club	LPA	2	0.132	0.002	-0.027
	LEA	4	**0.393**	0.006	-0.037
	Proposed algorithm	3	0.318	**0.216**	**0.255**
Dolphins social network	LPA	4	**0.519**	**0.555**	**0.445**
	LEA	5	0.491	0.539	0.344
	Proposed algorithm	16	0.345	0.047	-0.025

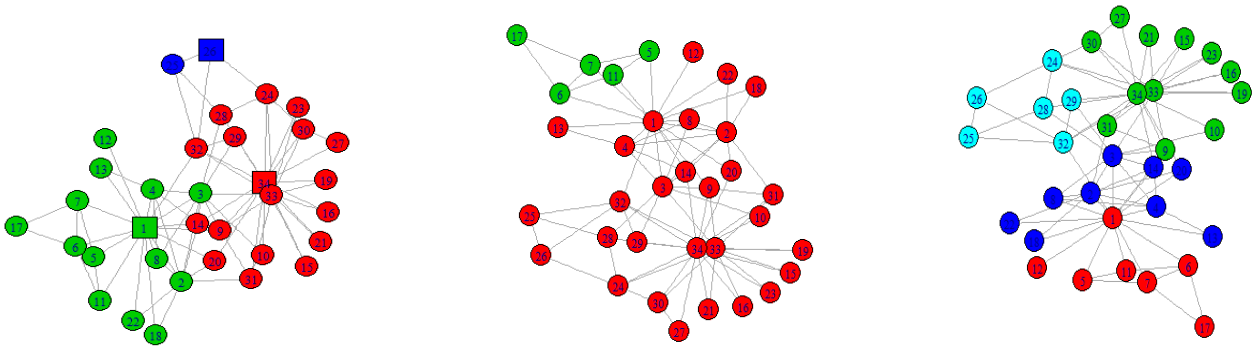

Figure 2.Community structure in Zachary Karaté Club provided from our algorithm where the leaders are represented by square, by LPA algorithm and LEA algorithm respectively.

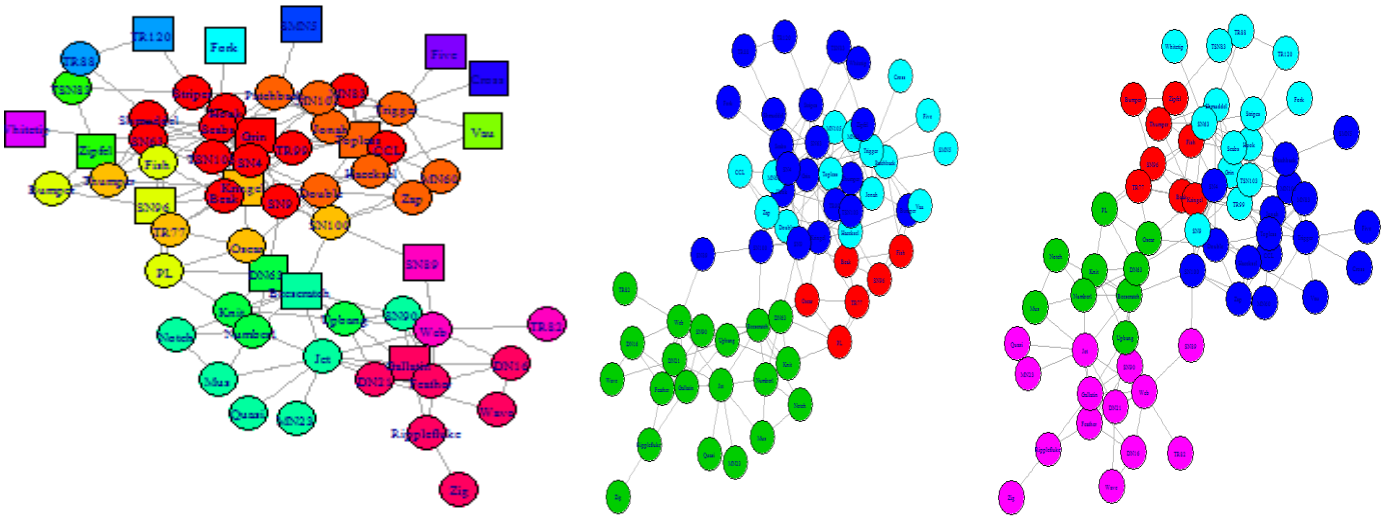

Figure 3.Community structure in Dolphins Social Networks provided from our algorithm where the leaders are represented by square, by LPA algorithm and LEA algorithm respectively.

VI. CONCLUSION

This paper presents a study of different detection algorithms communities and especially the leader nodes in complex networks have become increasingly important given the scientific and industrial challenges it represents. The idea is to group objects based on certain criteria. The interest shown by the research in this area is the fact that the dissemination of information i.e. the distribution of influence in complex networks is an element both strategic and particularly sensitive to their use. Thus, we have proposed a new approach for detecting communities using leaders' nodes who unlike the proposed algorithms do not require a priori knowledge of k nodes to detect leaders.

REFERENCES

[1] G. F. de Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. da F. Costa, "Role of centrality for the identification of influential spreaders in complex networks," Phys. Rev. E, vol. 90, no. 3, p. 032812, Sep. 2014.

[2] Y. Wang, Z. Di, and Y. Fan, "Identifying and Characterizing Nodes Important to Community Structure Using the Spectrum of the Graph," PLoS ONE, vol. 6, no. 11, p. e27418, Nov. 2011.

[3] P. Domingos and M. Richardson, "Mining the Network Value of Customers," in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2001, pp. 57–66.

[4] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the Spread of Influence Through a Social Network," in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 137–146.

[5] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," Physica A: Statistical Mechanics and its Applications, vol. 388, no. 8, pp. 1706–1712, Apr. 2009.

[6] B. Renoust, "Analysis and Visualisation of Edge Entanglement in Multiplex Networks," University of Massachusetts Lowell, 2014.

[7] H. R. Gor and M. V. Dhamecha, "A Survey on Community Detection in Weighted Social Network," International Journal, vol. 2, no. 1, 2014.

[8] Q. Wu, X. Qi, E. Fuller, and C.-Q. Zhang, "Follow the Leader: A Centrality Guided Clustering and Its Application to Social Network Analysis," The Scientific World Journal, vol. 2013, p. e368568, Oct. 2013.

[9] P. Pons, Detection communities in real networks. Paris 7, 2010. (P. Pons, Détection de communautés dans les grands graphes de terrain. Paris 7, 2010.)

[10] R. R. Khorasgani, J. Chen, and O. R. Zaïane, "Top leaders community detection approach in information networks," in in Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010. ISSN : 2319-7323, 2013, p. 228.

[11] B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," Bell System Technical Journal, vol. 49, no. 2, pp. 291–307, Feb. 1970.

[12] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3–5, pp. 75–174, Feb. 2011.

[13] D. Shah and T. Zaman, "Community Detection in Networks: The Leader-Follower Algorithm," arXiv:1011.0774 [physics, stat], Nov. 2010.

[14] J. Zhou, Y. Zhang, and J. Cheng, "Preference-based mining of top-influential nodes in social networks," Future Generation Computer Systems, vol. 31, pp. 40–47, Feb. 2014.

[15] Y. Xia, X. Ren, Z. Peng, J. Zhang, and L. She, "Effectively identifying the influential spreaders in large-scale social networks," Multimed Tools Appl, pp. 1–13, Sep. 2014.

[16] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy.," ICWSM, vol. 10, pp. 10–17, 2010.

[17] Y. Wang, Z. Di, and Y. Fan, "Detecting Important Nodes to Community Structure Using the Spectrum of the Graph," arXiv:1101.1703 [physics], Jan. 2011.