GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

DETECTION OF LEADER'S NODES IN COMPLEX NETWORKS

Sara Ahajjam [*1] Hassan Badir[2] and Mohamed El Haddad[3]

[*1, 2, 3] Laboratory "Technologies of Information and Communication", ENSA, Tangier

ABSTRACT

Complex networks are a powerful tool for understanding the mechanisms of various systems. These networks are complex graphs with high local density and low overall density. We are interested to study the problematic of community detection and more specifically the detection of 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 communities to detect. In this paper, we introduce a new approach to detect communities and leaders nodes in the network without a prior knowledge of k nodes to detect, and by taking into consideration a fundamental property of this type of networks, which is the overlap between the communities.

Keywords- *Complex networks, community detection, centrality, leader node, overlapping.*

I. INTRODUCTION

Complex networks have been used in very different application domains, such as physics, biology, social science or computer networks, it can be described in terms of graphs, whose vertices denote players phenomenon and links denote the interactions between these entities. These networks have a high local density and low overall density.

The problem of modeling and studying the structure of networks attracted the attention of a huge amount of work in several areas, and led to the discovery of unexpected properties: the distances are often low and the distribution of degrees is very heterogeneous. In addition, although generally very sparse, they behave locally as cliques: two nodes connected to a node have a very high probability of being connected by two knots against random will almost never do. Another way to see this property is to consider that these networks consist of communities: A community is formally defined as a sub-structure present into the network that represents connections among users, in which the density of relationships within the members of the community is much greater than the density of connections among communities. From a structural perspective, this is reflected by a graph which is very sparse almost everywhere but dense in local areas, corresponding to clusters (i.e. communities).

The discovery of these communities would considerably help for understanding the structure of the network; it is a task that can be closer to unsupervised classification. Community detection is also important for identifying modules and their boundaries, allows a classification of vertices, according to their structural position in the modules. Vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities.

These interaction networks, i.e. social networks 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 A new approach to detecting nodes leading USA on Twitter). What makes these networks a powerful tool to influence the choices and directions of their users. We are looking to examine how e-commerce websites can take advantage of services and detailed data of social networks to identify "influential" individuals those who are likely to effectively encourage product adoption amongst their peers and susceptible peers that are likely to respond positively to influence. The notion that influential exist and that such individuals are catalysts for promoting diffusion of opinions, innovations and products is a popular one [1]. The complementary notion, however, that susceptible may be even more critical to product diffusion and the formation of widespread contagions, has been systematically understudied in the literature [2]. As Watts and Dodds note: "In the models that we have studied, in fact, it is generally the case that most social change is driven not by influential but by easily influenced individuals influencing other easily influenced individuals".

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. Hence the detection of the influence or detecting leaders of these networks i.e. specifically detect leaders trying to select a set of users that could influence many others to choose a product, where the influence works mainly by 'word-of-mouth'.

II. 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 communities' detection algorithms 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 [3]. Reihaneh Rabbany Khorasgani et al. suggest a new approach to detect leader's nodes that takes into account the nodes that are not associated to 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 [4]. Another algorithm proposed based on partitioning of graphs, 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 [5]. The results of these two algorithms vary according to the size and number of partitions which are introduced. This method of leader nodes detection requires a prior knowledge of this information. 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 [6]. 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 measures 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. There are two types of cuts used in the graphs of spectral clustering: RatioCut and NCUT. These functions are minimized when nodes are grouped into larger communities with few inter-edges. 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 [7]. The same approach in the eagle algorithm, we detect the cliques, and we apply an agglomeratif algorithm to merge cliques [8]. Another research 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 $k \geq 15$ and it is of complexity O (n3) [9]. More recently, in [9], 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. Experimental results demonstrate that UBLF, compared with CELF, reduces Monte Carlo simulations and reduces the execution time when the size of seed set is small. 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, [14] 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 to community using the spectrum of the graph 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 [16].

Community and leader nodes detection approaches are diverse. Therefore, these methods lead to incomplete coverage of the network. Each proposed algorithm brings a new idea or improvement of existing algorithms. But, it is impossible to be compliant with all types of networks. We will propose a new approach to detect communities and leader nodes in complex networks, especially social networks.

III. PROBLEM FORMULATION

Social network is represented by a social graph which is an undirected graph G = (V, E) where the nodes are users. There is an undirected edge between user's u and v representing a social tie between the users. The tie may be explicit in the form of declared friendship, or it may be derived on the basis of shared interests between users.
There are a number of conflicting ideas and theories about how trends and innovations get adopted and spread. The traditional view assumes that a minority of members in a society possess qualities that make them exceptionally persuasive in spreading ideas to others. These exceptional individuals drive trends on behalf of the majority of ordinary people. They are loosely described as being informed, respected, and well-connected; they are called the leaders, innovators in the diffusion of innovations theory [10], and hubs, connectors, or mavens in other work [11]. The theory of influential is intuitive and compelling. By identifying and convincing a small number of influential individuals, a viral campaign can reach a wide audience at a small cost. The theory spread well beyond academia and has been adopted in many marketing businesses [12] [13].
We seek to find the partition P = {C_1, C_2,…. C_n} of all nodes forming communities that can overlap with each other, i.e. the node can belong to a community or more. For example, if a computer science article was referenced in the computer section and the community of mathematical article, so the two communities overlap ($\square C_i \cap C_j \neq \emptyset$, $i \neq j$) community. And for each community, we detect the central node or leader, as in figure 1.

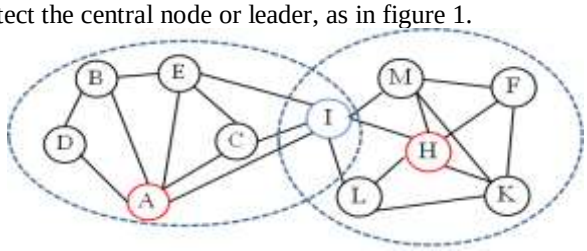

FIG. 1 – Example of two communities overlap in the blue node, A and H are central nodes

IV. OUR CONTRIBUTION

There are two key contributions of this work. First, using insights from social networks, we propose a certain natural internal structure for a community. Second, we provide an algorithm that uses this internal structure to find these communities and leader of each community.
Using social networks, we define two structural properties one would expect a community to possess. These properties are essentially internal to the community, and are invariant to the density of inter-community edges:

- First, the most of community nodes are connected to each other. While in the first level, everyone in a community should know each other, or more formally, the community should be a clique.
- Second, each community should possess distinguishing members. More formally, each community must have at least one node with no neighbors in any other community. Those nodes are what provide to the community a distinguishable identity.

Our contribution has roughly three stages.

1. **Detecting maximal cliques-** inspired from, Kerbosch Bron (1973) algorithm, we break the graph into maximal cliques:

- A clique is a subset of the graph whose vertices are adjacent. The maximal clique of a graph is a clique that supports the largest number of vertex and is not a subset of another clique.
- Neglecting subordinate maximal cliques: they are the maximal cliques that have nodes of another large maximum clique.

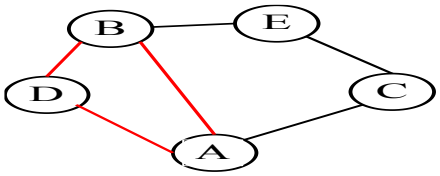

FIG.2 – A maximal clique made up of nodes A, B,D.

2. **Detecting communities**

- We will apply agglomerative approach. The agglomerative methods attempt to recursively merge small communities into larger by making choices based on a measure of proximity between communities, with the starting point of the atomic communities containing only one node. In our case, we will consider each clique found as a node.
- Select randomly two nodes; and we will calculate the similarity function of each one.
- Join repeatedly communities in pairs, choosing at each step the join that results in the largest increase (or decrease) the smallest Q. We will stop once the modularity is maximal and does not change more.

$$EQ = \frac{1}{2m}\sum_{i}\sum_{x\in C_i, y\in C_i}\frac{1}{OxOy}\left[A_{xy} - \frac{K_x K_y}{2m}\right]$$

With: is the number of communities to which x belongs. is the element is the adjacency matrix of the network. is the degree of x. **m** is the total number of arcs in the network.

- Detecting overlapping communities. For duplication, we will use the modular overlap to assign the node to the community that suits him.

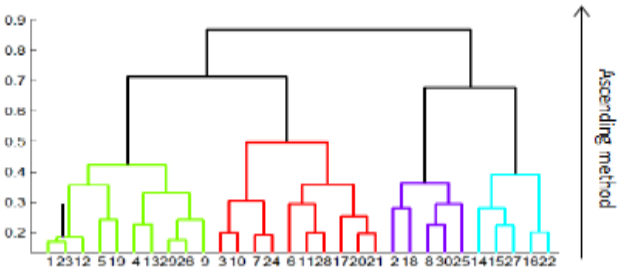

FIG.3 – Example of a dendrogram "agglomerative method".

Detecting leaders' nodes- For each community found, we calculate the centrality of each of its nodes, and we will choose the one with the high degree centrality as "Leader" of the community.

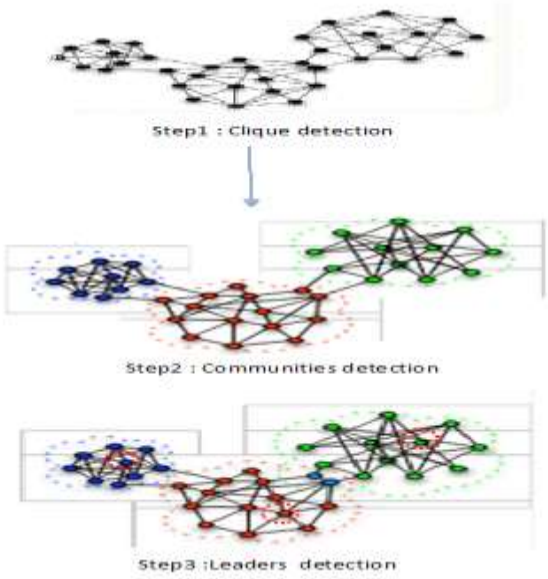

FIG.4 – Steps of our approach.

V. CONCLUSION

This paper presents a study of different detection algorithms communities and especially the leader nodes in social. 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 community leaders that unlike existing algorithms do not require prior knowledge of k nodes to detect leaders, and take into account the overlap of the communities which characterize the interaction network.

REFERENCES

1. *Aral, S., and Walker, D. Identifying influential and susceptible members of social networks. Science 337, 337–341, 2012.*
2. *Watts, D., and Dodds, P. Influentials, Networks, and Public Opinion Formation. Journal of Consumer Research, 2007.*
3. *Pons, P. Détection de communautés dans les grands graphes de terrain (Paris 7), 2010.*
4. *Khorasgani, R.R., Chen, J., and Zaïane, O.R. Top leader's community detection approach in information networks. Proceedings of the 4th Workshop on Social Network Mining and Analysis. ISSN: 2319-7323, p. 228, 2013.*
5. *Kernighan, B.W., and Lin, S. An Efficient Heuristic Procedure for Partitioning Graphs. Bell Syst. Tech. J. 49, 291–307, 1970.*
6. *Fortunato, S. Community detection in graphs. Phys. Rep. 486, 75–174, 2011.*
7. *Shah, D., and Zaman, T. Community Detection in Networks: The Leader-Follower Algorithm. ArXiv10110774 Phys. Stat, 2010.*
8. *Shen, H., Cheng, X., Cai, K., and Hu, M.-B. Detect overlapping and hierarchical community structure in networks. Phys. Stat. Mech. Appl. 388, 1706–1712, 2009.*
9. *Zhou, J., Zhang, Y., and Cheng, J. Preference-based mining of top-k influential nodes in social networks. Future Gener. Comput. Syst. 31, 40–47, 2014.*
10. *Rogers, E.M. Diffusion of innovations. (New York: Free Press of Glencoe), 1962.*
11. *Gladwell, M. The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books, 2002.*
12. *Berry, J., and Keller, E. The Influentials. Free Press. 2003.*
13. *Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P.K. Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM 10, 10–17. 2010.*
14. *M. KITSAK ET AL., Identification of Influential Spreaders in Complex Networks. Nature Physics, Vol. 6, pp. 888-893. 2011.*
15. *XIA, Y., REN, X., PENG, Z., ZHANG, J. & SHE, L. Effectively identifying the influential spreaders in large-scale social networks. Multimed Tools Appl 1–13. doi:10.1007/s11042-014-2256-z. 2014.*
16. *WANG, P., YU, X., LU, J. & CHEN, A. Identification of important nodes in artificial bio-molecular networks. in 2014 IEEE International Symposium on Circuits and Systems (ISCAS) 1267–1270. 2014.*