Communities Identification Using Nodes Features

Sara Ahajjam(✉), Hassan Badir, Rachida Fissoune, and Mohamed El Haddad

Laboratory of Technologies of Information and Communication, National School of Applied Sciences of Tangier, Tangier, Morocco
ahajjam.sara-etu@uae.ac.ma,
{Hassan.badir,Fissoune.rachida,elhaddad.mohamed}@uae.ma

Abstract. The network sciences have provided significant strides for understanding complex systems. Those systems are represented by graphs. One of the most relevant features of graphs representing real systems is clustering, or community structure. The communities are clusters (groups) of nodes, with more edges connecting to nodes of the same cluster and comparatively fewer edges connecting to nodes of different clusters. It can be considered as independent compartments of a graph. There are two possible sources of information we can use for the community detection: the network structure, and the attributes and features of nodes. In this paper, we use the features of nodes to detect communities. There are nodes in network that are more able and susceptible to diffuse information and propagate influence. The main purpose of our approach is to find leader nodes of networks and to form community around those nodes. Unlike to most existing researches studies, the proposed algorithm doesn't require a priori knowledge of k number of communities to be detected.

Keywords: Community detection · Influential node · Complex networks · Centrality · Classification

1 Introduction

Graphs become extremely useful as the representation of a wide variety of systems in different areas (biological, information, and social networks). Graph analysis is becoming crucial to understand the features of these complex systems.

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]. Studying 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.

F. Esposito et al. (Eds.): ISMIS 2015, LNAI 9384, pp. 303–312, 2015.
DOI: 10.1007/978-3-319-25252-0_33

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 [5, 6]. 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.

2 Overview

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) [7], while the local ones treat with local position, i.e. with the node (degree centrality) [8]. 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(n3) [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. 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, [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 reduce 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 k_s 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 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.

3 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 users 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, and hubs, connectors, or mavens in other work [16]. The theory of leaders is intuitive and compelling. By identifying and convincing a small number of influential/leader 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, e.g., RoperASW and

Tremor [11]. We need to detect those influential/leader nodes that are responsible for the dissemination of information and form communities around those nodes whose facilitate the spread of influence once we need to.

4 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 is based 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].

$$\mathrm{Ax} = \lambda \mathrm{x} \tag{1}$$

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 in a list L, 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 add it to the set of communities detected. After, we deal with the second node with the highest centrality until all the vertices (nodes) will be treated.

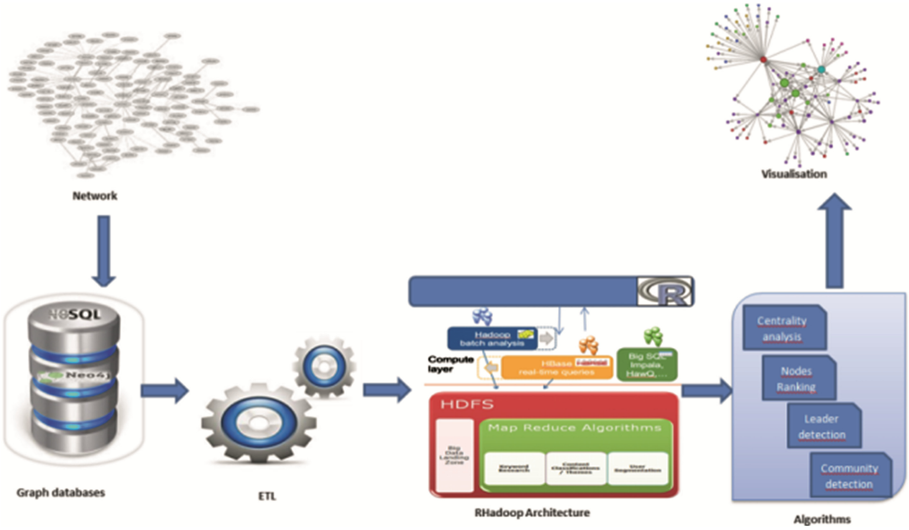

Fig. 1. Architecture of the proposed solution

5 Results and 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 [22] (Table 1).

Table 1. Datasets properties

Datasets	*Nodes*	*Edges*	*Real Communities*
Zachary Karaté Club	34	78	2
Word adjacencies	112	425	2

Adjective and noun adjacencies: This is also a famous network widely used as a benchmark to validate community detection algorithms. It's a network of common adjective and noun adjacencies for the novel "David Copperfield" by Charles Dickens, as described by M. Newman. Nodes represent the most commonly occurring adjectives and nouns in the book. Edges connect any pair of words that occur in adjacent position in the text of the book [23].

Figures 3 and 4 show 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) [24] and Leading Eigenvalue Algorithm (LEA) [23] using different metrics. For each network we calculate the quality of partition using the modularity Q.

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

```
Input: undirected, unweighted graph G=(V,E)
Output: Set C=($C_1$,$C_2$,…,$C_n$)
    $i = 0$
    While L≠0
    Calculate the centrality score of each vertex V ∈ $G$,
    Loop
    Nodes ranking: Order V via their centrality scores, such that L = ($V_1$, $V_2$, . . . , $V_n$) with Cent ($V_1$)
  ≥ Cent ($V_2$) ≥ · ·· ≥Cent ($V_n$).
    $i = i + 1$
    Select where $V_{i1}$ is the first vertex in the vertex list $Q$.
    Create a new group $Ci$= {$V_{i1}$},
    New L = L− {$V_{i1}$}
    L = New L
    Community detection: Calculate the neighborhood function of $V_{i1}$ to find the candidate neighbors set "neighbors N($V_j$)".
    insert into N($V_j$).
    New L=L-N($V_j$)
    End loop
```

Fig. 2. Pseudo-code of the proposed algorithm

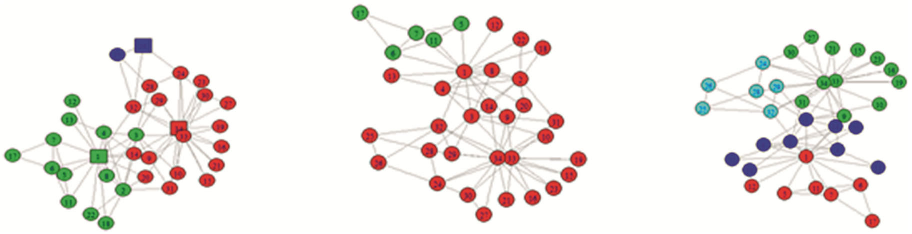

Fig. 3. Community structure in Zachary Karaté Club provided by our algorithm where the leaders are represented by square, by LPA algorithm and LEA algorithm respectively.

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:

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

And we use the Normalized Mutual Information (NMI):

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

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. And for the second dataset, our algorithm provides the best result for the three metrics NMI, ARI and modularity.

Table 2. 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**
Word adjacencies	LPA	4	0	0	−1.101
	LEA	5	0.243	0.008	−0.013
	Proposed algorithm	22	**0.584**	**0.109**	**−0.0002**

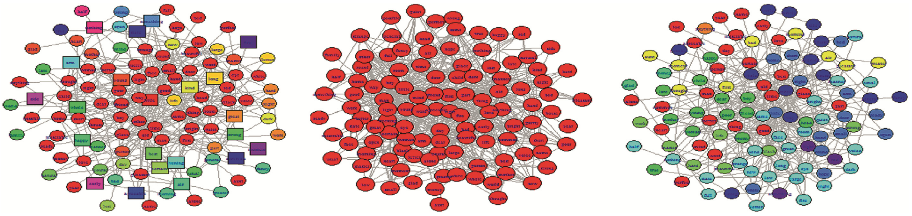

Fig. 4. Community structure in Word adjacencies network provided by our algorithm where the leaders are represented by square, by LPA algorithm and LEA algorithm respectively.

6 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. de Arruda, G.F., Barbieri, A.L., Rodríguez, P.M., Rodrigues, F.A., Moreno, Y., da Fontoura Costa, L.: Role of centrality for the identification of influential spreaders in complex networks. Phys. Rev. E **90**(3), 032812 (2014)
2. Wang, Y., Di, Z., Fan, Y.: Identifying and characterizing nodes important to community structure using the spectrum of the graph. PLoS ONE **6**(11), e27418 (2011)
3. Domingos, P., Richardson, M.: 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, pp. 57–66 (2001)
4. Kempe, D., Kleinberg, J., Tardos, É.: 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, pp. 137–146 (2003)
5. Shen, H., Cheng, X., Cai, K., Hu, M.-B.: Detect overlapping and hierarchical community structure in networks. Phy. A **388**(8), 1706–1712 (2009)
6. Renoust, B.: Analysis and Visualisation of Edge Entanglement in Multiplex Networks. University of Massachusetts Lowell (2014)
7. Gor, H.R., Dhamecha, M.V.: A survey on community detection in weighted social network. Int. J. **2**(1) (2014)
8. Wu, Q., Qi, X., Fuller, E., Zhang, C.-Q.: Follow the leader: a centrality guided clustering and its application to social network analysis. Sci. World J. **2013**, e368568 (2013)
9. Pons, P.: Detection communities in real networks, Paris 7 (2010). (P. Pons, Détection de communautés dans les grands graphes de terrain, Paris 7, 2010)

10. Khorasgani, R.R., Chen, J., Zaïane, O.R.: Top leaders community detection approach in information networks. In: Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010, p. 228 (2013). ISSN: 2319-7323
11. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. **49**(2), 291–307 (1970)
12. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2011)
13. Shah, D., Zaman, T.: Community Detection in Networks: The Leader-Follower Algorithm. arXiv:1011.0774, November 2010
14. Zhou, J., Zhang, Y., Cheng, J.: Preference-based mining of top- influential nodes in social networks. Future Gener. Comput. Syst. **31**, 40–47 (2014)
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 (2014)
16. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: the million follower fallacy. ICWSM **10**, 10–17 (2010)
17. Wang, Y., Di, Z., Fan, Y.: Detecting important nodes to community structure using the spectrum of the graph. arXiv:1101.1703, January 2011
18. Ruhnau, B.: Eigenvector-centrality—a node centrality? Soc. Netw. **22**, 357–365 (2000)
19. Newman, M.E.J.: Analysis of weighted networks. Phy. Rev. E **70**(5), 056131 (2004)
20. Moody, J., White, D.R.: Structural cohesion and embeddedness: a hierarchicalconcept of social groups. Am. Sociol. Rev. **68**, 103–127 (2003)
21. Fuong, H., Maldonado-Chaparro, A., Blumstein, D.T.: Are social attributes associated with alarm calling propensity? Behav. Ecol. **26**, 587–592 (2015)
22. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
23. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phy. Rev. E **74**(3), 036104 (2006)
24. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phy. Rev. E **76**(3), 036106 (2007)