Social Networks 54 (2018) 41–49

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

A new scalable leader-community detection approach for community detection in social networks

Sara Ahajjam*, Mohamed El Haddad, Hassan Badir

Laboratory of Information and Communication Technologies, National School of Applied Sciences, ENSA, Tangier, Morocco

ARTICLE INFO

Article history:

Keywords:
Leader
Community detection
Big graph
Centrality
Social network
Similarity
Big data
Graph theory

ABSTRACT

Studying social influence in networks is crucial to understand how behavior spreads. An interesting number of theories were elaborated to analyze how innovations and trends get adopted. 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. The leaders or influential are responsible for the dissemination of information and the propagation of influence. In this paper, we propose a new scalable and a deterministic approach for the detection of communities using leaders nodes named Leader-Community Detection Approach LCDA. The proposed approach has two main steps. The first step is the leaders' retrieval. The second step is the community detection using similarity between nodes. Our algorithms provide good results compared to ground truth membership community.

1. Introduction

Complex networks represent complex systems in different areas. They can be modeled by a graph, where nodes represent the actors of the system, connected by edges to describe different types of relationships. Discovering communities is a fundamental problem in network science, which has attracted vast attention in recent years (Xie et al., 2013; de Arruda et al., 2014). Several research studies have addressed the problem of community detection. Community detection aims to find clusters as sub-graphs within a given network (Social Network Analysis, 2017), with the purpose of finding the communities using the information embedded in the network topology. A community is defined as a set of nodes highly interconnected, and loosely connected to other nodes in the network.

Within certain communities, some nodes play more important roles in diffusion of information, ideas, and innovation within those communities. They are the catalysts of influence. Therefore, it has motivated many researchers to look for an efficient method to find the most influential members in social networks. For example, it is in trading companies and banks interest to find active and influential parties in their existing network and potentially extend their network to include other parties (Wang et al., 2011a).

Identifying influential or leaders nodes in networks can be regarded as ranking important nodes, and it has become one of the main problems in network-based information retrieval and mining (Domingos and Richardson, 2001). In epidemic spreading, it is vital to find the important nodes to understand the dynamic processes, which could shed some light on immunizing modular networks. In biological systems, key nodes are identified in the communities for the purpose of treatment, for example in the case of lung cancer, the treatment option entails destroying cancer cells while protecting normal cells (Domingos and Richardson, 2001).

Identification of influential nodes relies on quantitative characterization of the node in terms of their importance to community structure, centrality in particular. Centrality aims to identify the most important actors in a social network, which determines the social influence and power of each actor within such network. Centrality can be local or global (Gao et al., 2014; Networks, 2010; Renoust, 2014). The local methods, which include degree and betweenness centrality, use the local features of the node to determine its importance. Degree centrality measures node involvement in a network and is determined by the number of nodes that a focal node is connected to. The betweenness centrality assesses centrality in a network of nodes based on shortest paths.

However, most network node researchers fail to consider global topological structures. Closeness centrality could be used to over-

* Corresponding author.
E-mail addresses: ahajjamsara@gmail.com (S. Ahajjam), elhaddad.mohamed@gmail.com (M. El Haddad), hbadir@gmail.com (H. Badir).

https://doi.org/10.1016/j.socnet.2017.11.004
0378-8733/

come such limitation, and it is assessed based on inverse of the sum of the shortest distances between each node and every other node in the network. Also, K-shell decomposition analysis shows that network nodes in core layers are capable of spreading to broader areas compared to those in peripheral layers. Since these central nodes have the ability to diffuse their influence to the whole network faster than the rest of the nodes, they would be regarded as the most influential spreaders.

This paper deals with two important research subjects in computer science: the community detection and the leader detection in complex networks. The majority of the proposed algorithms detect the community first followed by identification of the leader of a given community. The main limitation associated with such methodology is that prior knowledge of the number and the size of the final partitions are required. In this paper, a scalable and deterministic approach is presented to detect communities in social networks using leaders' nodes. In this approach, leader nodes of the networks that are responsible for the dissemination of influence are detected, then communities will be formed around the leaders using similarity between nodes, i.e. using different edge based similarity measures. In Section 2, relevant work and studies will be reviewed to provide border perspective and insight to the proposed subject. In Section 3, fundamentals of proposed algorithms will be presented in detail. In Section 4, experimental results will be presented, and the final section offers concluding remarks and further application.

2. Preliminaries & related works

Leader detection and community detection in complex networks, particularly in social networks have been the subject for wide research studies in recent years. The Centrality and prestige measures are used to determine the importance of a node in undirected and directed networks respectively, where prestigious nodes, the influential or the leader nodes are identified. Prestige and centrality provide reliable analysis of the nature of relationship and connection between nodes, which is important to understand a wide range of phenomena.

In computer-science literature, several methods have been applied to analyze user influence in social network and community leader detection in online social networks. The leader detection approaches are divided into two main groups: global and local methods. The global methods emphasize on all the network topology (betweenness centrality), while the local methods focus on local positions, i.e. individual nodes (degree centrality). In the context of social science, the influence can be defined as bargaining power, control over information, or level of persuasiveness (Khorasgani et al., 2010). Khorasgani et al. suggested a new approach to detect leader nodes that consider outliers, i.e. nodes that are not associated with any leader. This algorithm is inspired by K-means algorithm. In K-means algorithm, k nodes will be selected randomly, and other nodes will be assembled at their closest leaders to form communities. For each community, new leader will be identified gathering other new followers until no node moves. For each community, the centrality of each member is calculated and the node with the highest degree will be appointed as the new leader (Kernighan and Lin, 1970). Another approach of Bae et al. used the coreness centrality to estimate the spreading influence of a node. The coreness centrality is estimated by the k-shell indexes of the adjacent neighbors of the node. Therefore, the coreness centrality of a node is presented as the sum of the k-shell values of its adjacent nodes (Bae and Kim, 2014).

The existing methodologies used for community detection can be divided into two main categories, i.e. graph partitioning and classification. The major drawback of partitioning of graphs is that prior knowledge of the number of groups to be detected (Newman, 2013; Bader et al., 2013; Fortunato, 2011) is required. For example, in the study carried out by Kernighan and Lin, the algorithm of leader nodes detection in complex networks is based on the 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, and results are generated by introducing the size of each partition (Fortunato, 2011), and result can vary significantly given the assumption of size and number of partitions of the graphs.

Classification methodology is also used by researchers to analyze data and partition based on a measure of similarity between partitions. Approaches based on a partitioned classification require prior knowledge of the number of communities to be detected (Fortunato, 2011; Yang et al., 2014; Wu et al., 2013). The major problem to overcome is to select an appropriate distance for better data classification (Yang et al., 2014), which explains why classification methods are generally more appropriate for networks with hierarchical structure. Agglomerative methods and divisive methods are two main groups of methods used in hierarchical clustering. The agglomerative methods attempt to recursively merge small communities into larger communities based on a measure of proximity between communities. It is a bottom up approach, each node will be defined as a cluster, and clusters will be combined at each step, until pairs of clusters are merged as one moves up the hierarchy. The divisive methods attempt to identify the inter-links and remove them to gradually isolate communities, i.e. they initially put all nodes in a complete graph and they remove the links between nodes with the lowest similarity (de Arruda et al., 2014). Some other studies utilize the spectral classification; in the approach proposed by Yang et al., the community is defined based on three properties: community structure to define strong and weak communities, community membership to detect members of the communities, and overlapping property, which considers the number of connections between the node and the corresponding community. The result obtained by these methods will heavily depend on the choice of the similarity measure used initially (Yang et al., 2014).

Some other research studies propose methods that combine the two predefined subjects: leader detection and community detection. In the Leader-Follower algorithm, internal structure of a community should be defined, i.e. the community should be a clique and is formed with a leader and at least one "loyal follower", with loyal follower defined as a node that only has neighbors within a single community. The leader is a node whose distance is less than at least one of its neighbors. Nodes will be allocated to the community in which a majority of their neighbors belongs by destroying the links arbitrarily. However, removing parasite communities will lead to loss of information (Wu et al., 2013).

Qin et al. proposed a novel centrality guided clustering for detecting leader nodes and communities by choosing the vertex with the highest centrality as a starting point. The approach is carried out with three steps: grouping will cluster vertices in G into different groups, merging groups with a large percentage of overlap, and contract those vertices in the same groups to form a new vertex (Zhou et al., 2014). Zhang et al. proposed a greedy algorithm based on user preferences (GAUP) to operate the top-k influential users, based on the model Extended Independent Cascade. 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 would only provide a good result if k exceeds a certain threshold $k \geq 15$ (Xia et al., 2014).

Recent study also found that the location of the node in the network topology is another important factor when estimating the spreading ability (Fu et al., 2014). According the study, the location of a node is identified through the k-shell decomposition method, by which the network is divided into several layers. Each node is corresponding to 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, Degree centrality and k-shell decomposition are suitable for measuring the spreading ability of nodes quickly but not very accurately. Previous studies (Khorasgani et al., 2010; Wang et al., 2011b) use both global and local methods of centrality measures to effectively identify the influential spreaders in large-scale social networks. The main idea is that it reduces the scale of the network by eliminating the node located in the peripheral layer (namely relatively small k_s value) that will not have much spreading potency comparing with the core node, and vice versa. This algorithm uses the k-decomposition centrality to deal only with the nodes in the core of the network. Hence, it reduces the scale of the network by ignoring nodes with small k_s value and their associated links, and retains the nodes in the core layers.

The global methods (i.e. betweenness centrality and closeness centrality) are used to rank the most influential spreaders. A novel approach to detect communities and important nodes of the detected communities using the spectrum of the graph defines the important nodes to the community as the relative changes in the c largest eigenvalues of the network adjacency matrix upon their removal. There are two types of nodes, the core nodes namely central nodes that are the most important for the community, and the bridges node that connect the communities to each other's. The main drawback of this approach, it needs to know the number of partitions in the network to obtain a better result, and it cannot identify the important nodes in small communities (Shah and Zaman, 2010).

Community and leader detection approaches are diverse. Each proposed algorithm brings a new idea or improvement to the existing algorithms. In this paper, a new approach will be proposed to detect communities and leader nodes in complex networks, without prior knowledge of the number of communities to detect.

3. Proposed algorithms

3.1. General approach

Community and influential detection in networks is crucial to understand how behavior spreads. An interesting number of theories are used to analyze how innovations and trends are conveyed and adopted to targeted parties. The traditional view assumes that a small number of members in a society possess superior qualities, which make them exceptionally persuasive in terms of conveying ideas to others. These exceptional individuals impose more influence on the society than majority of the so-called ordinary people. They are called the influential and innovators in the diffusion of innovations theory, and hubs, connectors, or mavens in other existing studies (Spizzirri, 2011). The theory of leaders is intuitive and compelling. In order to detect the catalyst of this influence, we need to detect the central nodes that are responsible for the dissemination of influence, and form communities around those nodes that facilitate the spread of influence. In fact, centrality represents an important factor (Xia et al., 2014) within social network analysis, which measures the relative importance of a vertex within the graph.

Assume a social network that is modeled as an undirected and unweighted graph $G = (V, E)$, with V being the vertex set. Each vertex in G 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. $n = |V|$ represents the number of network nodes and $m = |E|$, the number of edges. The network structure is represented as an adjacency matrix $A = \{a_{ij}\}$ and $a_{ij} \in R$, where $a_{ij} = 1$, if a link exists between nodes i and j, otherwise $a_{ij} = 0$.

In the context of social network, there is an explosion of massive quantity of varied and unstructured information, i.e. big data, which careful analysis of such data is required to obtain useful information. This network will be modeled as a big graph. Firstly, it will be stored in NoSQL databases, specifically a scalable graph database named Neo4j. An ETL will be used to extract the adjacency matrix of the graph from the Neo4j Database, and will be imported to the RHadoop platform for the purpose of large-scale data processing as in Fig. 1.

There are two main steps in the proposed approach:

3.1.1. Leader detection

a) Nodes centrality:

Drawing inspiration from sociometric studies, in a complex social network system, a powerful leader is chosen by individuals, which themselves are powerful and are chosen by some other powerful individuals within the network (high sociometric status), whereas the popular leader will be chosen by individuals, which are more likely to be ordinary and not frequently chosen by others (Moreno, 1955). In this study, eigenvector centrality will be used to detect leaders nodes. For each node v in the network G, eigenvector centrality "Eq. (1)" will be calculated (Ahajjam et al., 2015a). Eigenvector centrality or Gould's index of accessibility (Ruhnau, 2000) is defined as a measure to describe how well connected an individual is based on direct and indirect relationships (i.e., it takes into account the connections of individuals the focal individual is connected to) (Newman, 2004). Nodes that are connected to other high degree nodes are considered as highly central, and can be quantified by a relative score recursively defined as a function of the number and strength of connections to its neighbors and as well as those neighbors' centralities, it uses the eigenvector corresponding to the largest eigenvalue of the graph adjacency matrix. Because eigenvector centrality is proportional to an individual's neighbors' centralities (Moody and White, 2003), more influential individuals will be more connected with other influential individuals. Lastly, centrality quantifies how isolated or how involved an individual in the network structure (Fuong et al., 2015; Zachary, 1977).

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

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

b) Nodes ranking:

Nodes will be ranked according to their centrality scores, and leader V_l is defined as the node with the highest centrality, i.e. the leader V_1 is the node with the largest eigenvalue in the adjacency matrix A. Fig. 2 shows an example of a graph and Table 1 is the ranked list of its vertices by centrality.

3.1.2. Community detection

For each leader, similarity function will be calculated to find the neighbors of the leader node with the highest centrality score. Similar nodes will be assigned to the detected leader node to form a community based on a certain threshold k, which for this study, $k \geq 0.5$.

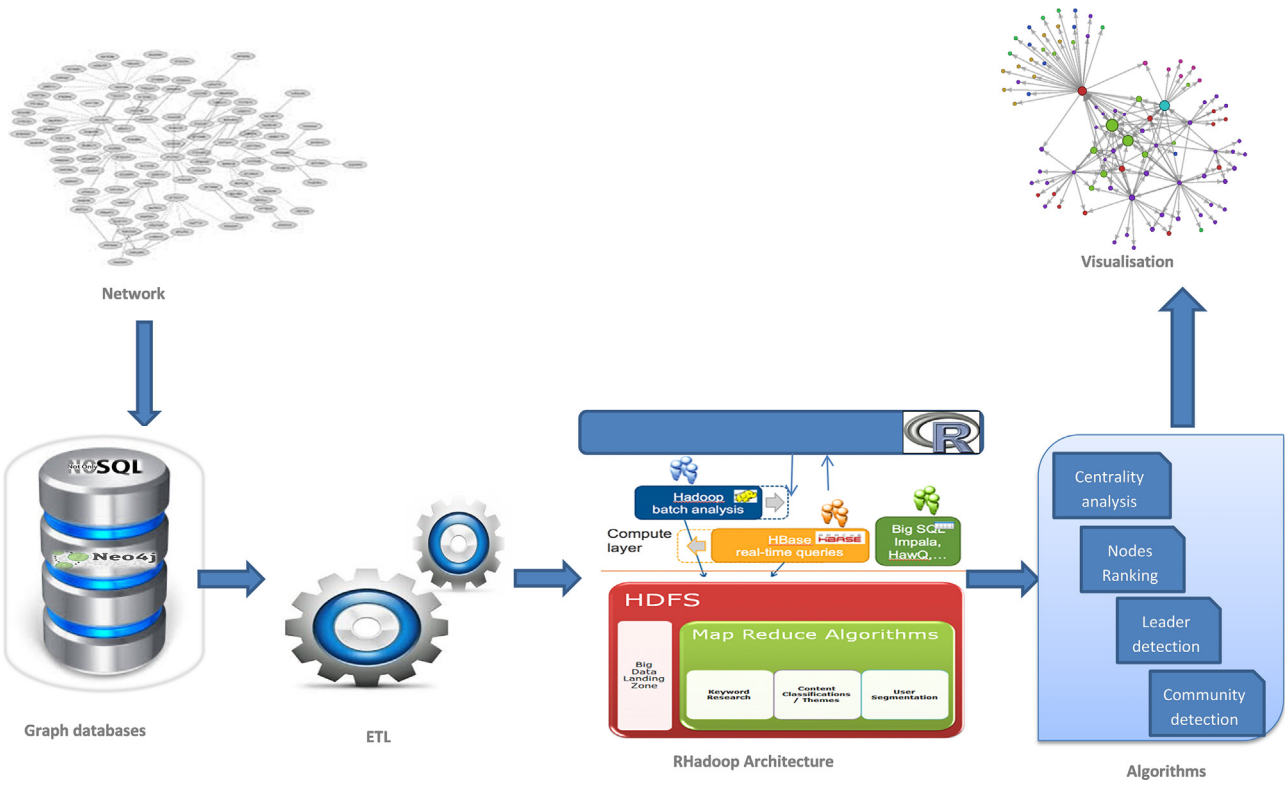


Fig. 1. Architecture of the Leader-Community Detection Approach.

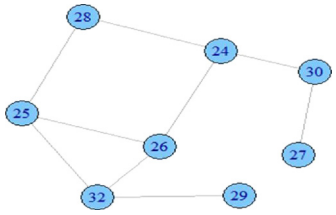


Fig. 2. The initial graph.

Table 1
The initial ranked List.

Node	Centrality
26	1.0000000
25	0.9716986
32	0.8813582
24	0.7658899
28	0.6634684
30	0.3423563
29	0.3365316
27	0.1307229

Finding similar objects can be used to predict links in data networks (Lu and Zhou, 2011). Detecting similarity between objects has its application in clustering, collaborative filtering, and search engines (Ganesan et al., 2003). In the study carried out by Ganesan et al., several similarity measures are used, and its utility depends on domain applications (Rawashdeh and Ralescu, 2015). Semantic similarity is the similarity between concepts using knowledge such as Wordnet or between words in the semantic web (Ilakiya et al., 2012). The global structural similarities aim to evaluate the similarity between two nodes in the context of the whole network.

In this study, due to the nature of the graph, which is undirected and unweighted, edge similarities will be used for the purpose of analysis. Similarity between nodes will be computed using Jaccard, Salton, HDI, and HPI similarities.

- Jaccard Similarity:

$$Sim_{Jaccard}(X, Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{|\Gamma(X) \cup \Gamma(Y)|} \quad (2)$$

- Salton Similarity:

$$Sim_{Salton}(X, Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\sqrt{K_x \times K_y}} \quad (3)$$

- Hub Depressed Index Similarity "HDI":

$$Sim_{HDI}(X, Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\max\{K_x, K_y\}} \quad (4)$$

- Hub Pressed Index Similarity "HPI":

$$Sim_{HPI}(X, Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{\min\{K_x, K_y\}} \quad (5)$$

Where $\Gamma(X)$ denotes the set of neighbors of X, and K_x is the degree of node X.

The proposed approach does not allow overlapping, therefore the deletion of communities is based on the assumption that one node can only belong to one community; we can use a political voting network as an example, where the person can vote for only one party. The second assumption is that members of a community share lots of neighbours with a leader; using the same example of political networks, the members of a party share a lot of neighbours (others members) with the president of the party. And this is can be the case for others networks.

Using the main steps of this approach, we will present two algorithms.

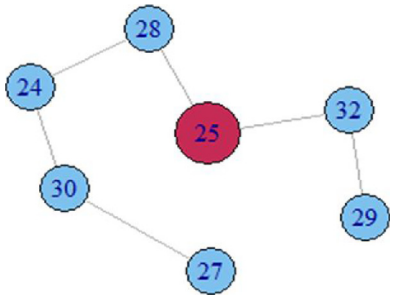


Fig. 3. The graph resulted from LCDA1 algorithm showing the second leader.

Input: undirected, unweighted graph G=(V,E)
Output: Set C=(C_1,C_2,…,C_n)
1: i = 0
2: Calculate the centrality score of each vertex V ∈ G,
3: While Q≠0
4: Loop
5: **Nodes ranking:** Order V via their centrality scores, such that Q = (V_1, V_2, . . . , V_n) with Cent (V_1) ≥ Cent (V_2) ≥ · ·· ≥Cent (V_n).
6: i = i + 1
7: **Select** where V_{i1} is the first vertex in the vertex list Q.
8: Create a new group C_i= {V_{i1}},
9: New Q = Q − {V_{i1}}
10: Q = New Q
11: **Community detection:** Calculate the similarity function of V_{i1} to find the **candidate nodes set "Similars S(V_j)"**.
12: **insert into** S(V_j).
13: C_i = C_i + S(V_j)
14: New Q=Q-S(V_j)
15: End loop

Fig. 4. Pseudo-code of the first algorithm.

3.2. First algorithm: leader-community detection algorithm (LCDA1)

The proposed algorithm *LCDA1* is a new version of the LeadersRank algorithm (Ahajjam et al., 2015b). In the LeadersRank algorithm, the neighborhood of order one of the leader *L1* = 26 will be selected to detect community (Table 1). Whereas for the proposed algorithm, similarity functions cited previously will be used to find the similar nodes of the leader to form communities. Once the first community is detected, nodes will be deleted from the network and the same process will be applied to the others nodes (Fig. 4). In fact, the second leader node will be chosen *L2* = 25 from the existed ranking list and second community will be formed as it is shown in Fig. 3, until all nodes of the network will be dealt with.

Eigenvector centrality considers nodes that are connected to other high degree nodes as highly central. Since the leader with the highest eigenvalue will be detected before forming the community, nodes will be attributed to the community based on the number of shared neighbors with the leader, so the centrality of those nodes will be always less than the leader. In fact, the identified leader will always be a community leader.

3.3. Second algorithm: leader-community detection algorithm (LCDA2)

The proposed algorithm *LCDA2* is a new version of the algorithm (Ahajjam et al., 2017) in which the similarity between nodes is used to find the similar nodes of the leader nodes for detecting the communities (Fig. 6). Once first community formed by the Leader *L1* and its neighbors (similar) is detected, it will be deleted from the network and the whole process will be repeated again. Therefore, the remaining network will be dealt with the same way as the first one. Centrality will be computed for all the remaining nodes, and nodes with highest eigenvector centrality will be selected as the second leader and detects the community. In this case, the second leader was *L2* = 28 (Fig. 5). As it is shown in Table 2, the centrality of the nodes has changed.

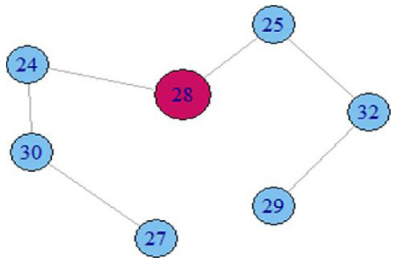


Fig. 5. The graph resulted from LCDA2 algorithm showing the second leader.

Input: undirected, unweighted graph G=(V,E)
Output: Set C=(C_1,C_2,…,C_n)
1: i = 0
2: Loop
3: While Q≠0
4: Calculate the centrality score of each vertex V ∈ G,
5: **Nodes ranking:** Get V with max(Cent(V)) from Q , such that Q = (V_1, V_2, … , V_n) with (Cent (V_1),Cent (V_2), …, Cent (V_n)) .
6: i = i + 1
7: **Select** where V_{i1} is the first vertex in the vertex list Q.
8: Create a new group C_i= {V_{i1}},
9: New Q = Q − { V_{i1}}
10: Q = New Q
11: **Community detection:** Calculate the similarity function of V_{i1} to find the **candidate nodes set "Similars S(V_j)"**.
12: **insert into** S(V_j).
13: C_i = C_i+ S(V_j)
14: New Q=Q-S(V_j)
15: End loop

Fig. 6. Pseudo-code of the second algorithm.

Table 2
The ranked list of LCDA2 after deletion of the first leader.

Node	Centrality
28	1.0000000
24	0.9238795
25	0.9238795
30	0.7071068
32	0.3423563
29	0.3826834
27	0.3826834

4. Experimental results

In this section, the proposed algorithms will be tested in real life social networks in which a ground truth community' membership is known. Algorithms will be applied to the following datasets:

- 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. 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 (Zachary, 1977).

- American College football. This is a known network used for validating community detection algorithms. The football network represents the Network of American football games between Division IA colleges during regular season Fall 2000. It contains 115 nodes representing players and 616 edges that represent the interactions between players (Girvan and Newman, 2002).
- Political blogs: A network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance. The network is made up of 1490 nodes and 19090 edges (Adamic and Glance, 2005).

To compare the accuracy of the resulting community structures; modularity Q is used to measure the quality of community used extensively in community detection.

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2\right) \tag{6}$$

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

In addition, Adjusted Rand Index will be used, where such 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:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{7}$$

Then Normalized Mutual Information (NMI) will be used to measure the similarity between the true community structures and the detected ones:

$$NMI(X, Y) := \frac{2I(X, Y)}{H(X) + H(Y)} \tag{8}$$

Where I(X,Y) the mutual information corresponds to the quantity of information shared by the variables. The lower bound represents the independence of the variables (they share no information). The upper bound corresponds to a complete redundancy; the value is not fixed, and the H(X) is the entropy of clustering of X.

4.1. Comparison analysis

4.1.1. Similarity comparison

In order to choose well-performing algorithms, the results of the fourth cited similarities that have been tested in the two algorithms will be compared: *HPI*, *Jaccard*, *Salton* and *HDI*. Figs. 8 and 9 present the *ARI* and *NMI* results for the both algorithms using different similarities, which are tested in three datasets. The Hub Pressed Index similarity gives the best results compared to other similarities. Based on these results, the proposed algorithms *LCDA1* and *LCDA2* will use the *HPI* similarity to find the leader's similar nodes for community detection.

Table 3
Modularity results.

	Zachary Karaté Club	American College Football	Political blogs
Ground-truth Modularity	0.3582	0.5539	0.4111
LCDA1	0.3266	0.5162	0.3977
LeadersRank1	0.3180	0.0378	0.2359
LCDA2	0.3266	0.4842	0.4062
LeadersRank2	0.3014	0.3750	0.2388

4.1.2. Algorithms comparison

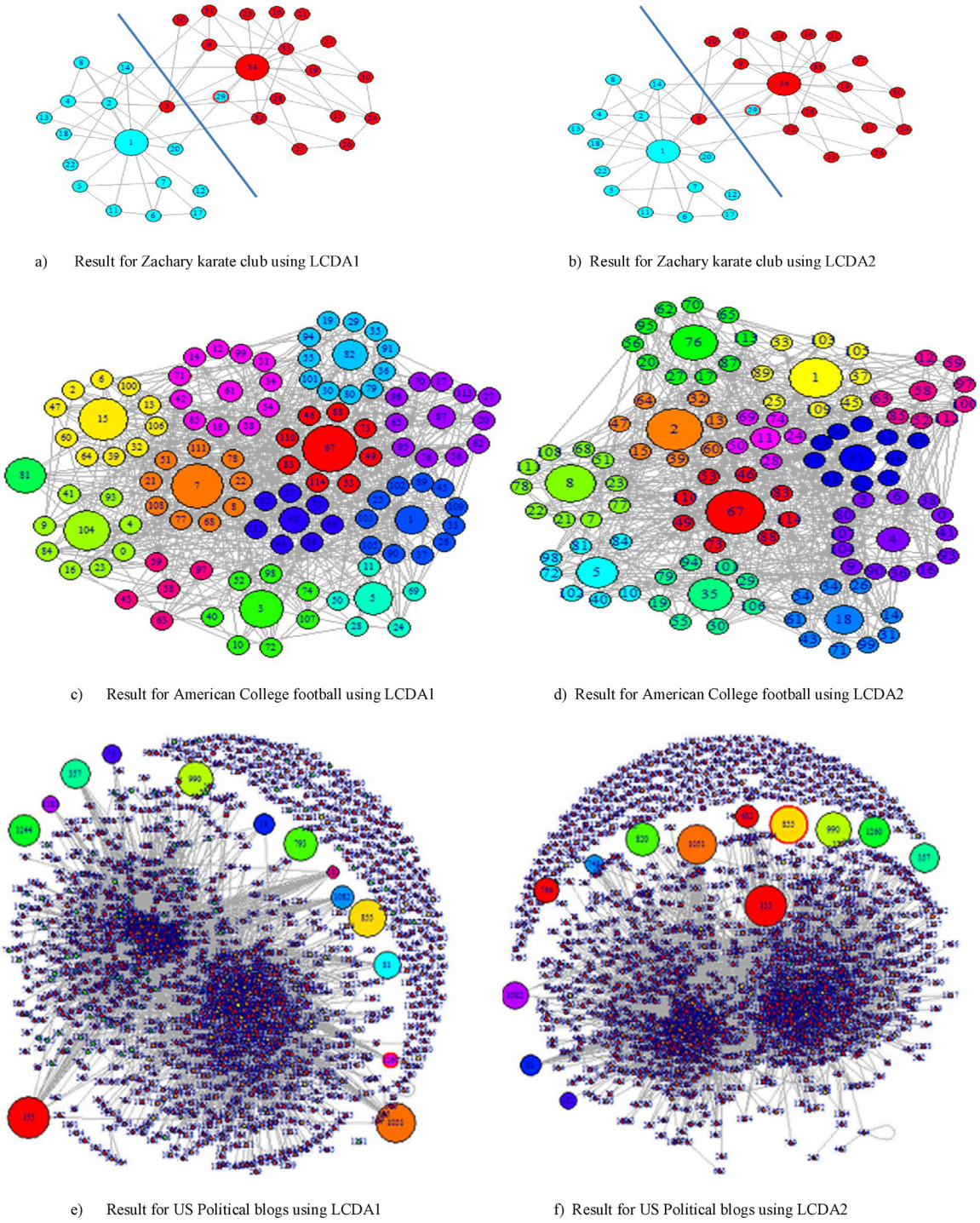
The detected community structure for each network is visualized in Fig. 7. Fig. 7a and b show a visualization of the detected communities on the Zachary Karate network. The detected partitions are two communities and it is similar to the original partition of the network. The colors of nodes represent the community to which they belong. The blue line divides the network to two communities as they would be in reality (ground truth). As it is shown in the graph, the only difference is the assignment of nodes 3 and 29. As indicated in Table 3, the proposed algorithms *LCDA1* and *LCDA2* provide a better modularity value i.e. 0.3266, compared to results produced by the *LeadersRank1* (Ahajjam et al., 2015b) and *LeadersRanks2* (Ahajjam et al., 2017).

Fig. 7c and d visualize the result of the American college football network, each color represents a community and the nodes in the core of the communities are the leaders. The original division of the network contains 12 communities. For the *LCDA2*, the detected communities are 12, they are similar to the ground truth partition and the modularity value is 0.4497431. While for the *LCDA1*, the detected communities are 13, however the modularity value of 0.5162414 is better than the value produced by *LCDA2*; *LCDA2* detects similar communities as the original network. In addition, the modularity for both algorithms is impressive compared to LeadersRank1 and especially to LeadersRank2 as shown in Table 3. The leader of each community is the node with the bigger size. The size of the leader node represents its level of influence relative to others leaders.

The US political blogs network contains two communities where members belong to either the right or the left part of a certain political blog. For the *LeadersRank1* and *LeadersRank2*, the number of the detected communities is 68 and 58 respectively. However, 14 and 13 communities are detected using *LCDA1* (Fig. 7e) and *LCDA2* (Fig. 7f) algorithms respectively. As shown in Table 3, the ground-truth modularity and proposed algorithms modularity scores are very close. For the political blogs, the similarity between the modularity scores for *LCDA2* and ground truth modularity is 98%, while the similarity is 50% for the *LeadersRank2* modularity and ground truth modularity. The proposed algorithms provide a significant improvement on modularity compared to the *LeadersRank1* and *LeadersRank2* algorithms. In fact, the proposed algorithms have managed to achieve the optimization of modularity, relative to previous studies.

Results obtained using proposed algorithms will be compared with other algorithms in existing researches, such as *Infomap* (Rosvall et al., 2009), *Multi-level* or *Louvain* (Blondel et al., 2008), *Label propagation* (Raghavan et al., 2007), *Walktrap* (Pons and Latapy, 2005), *LeadersRank1* (Ahajjam et al., 2015b) and *LeadersRank2* (Ahajjam et al., 2017).

Fig. 10 summarizes the *NMI* and the *ARI* values for all the algorithms in previous three data sets in which a ground truth community's membership is known. In terms of *NMI*, the proposed



a) Result for Zachary karate club using LCDA1

b) Result for Zachary karate club using LCDA2

c) Result for American College football using LCDA1

d) Result for American College football using LCDA2

e) Result for US Political blogs using LCDA1

f) Result for US Political blogs using LCDA2

Fig. 7. Visualization of the LCDA1and LCDA2 algorithm on each social network dataset.
a) Result for Zachary karate club using LCDA1.
b) Result for Zachary karate club using LCDA2.
c) Result for American College football using LCDA1.
d) Result for American College football using LCDA2.
e) Result for US Political blogs using LCDA1.
f) Result for US Political blogs using LCDA2.

algorithms produce a better result compared to other methods, especially for the big dataset "US Political blogs"; the *LCDA2* algorithm produces the best result in this case. In terms of *ARI*, in Zachary karate Club case, the proposed algorithms perform better than the others methods, as well as for the other networks, the proposed algorithms remains competitive relative to others algorithms. Regardless of datasets size, it can be observed that the proposed algorithms can produce a community structure with a good *NMI* values. However, for the *ARI*, the proposed algorithms perform better with relative smaller size dataset. Regarding the US Political blogs dataset, the *LCDA2* produces a better *ARI* value com-

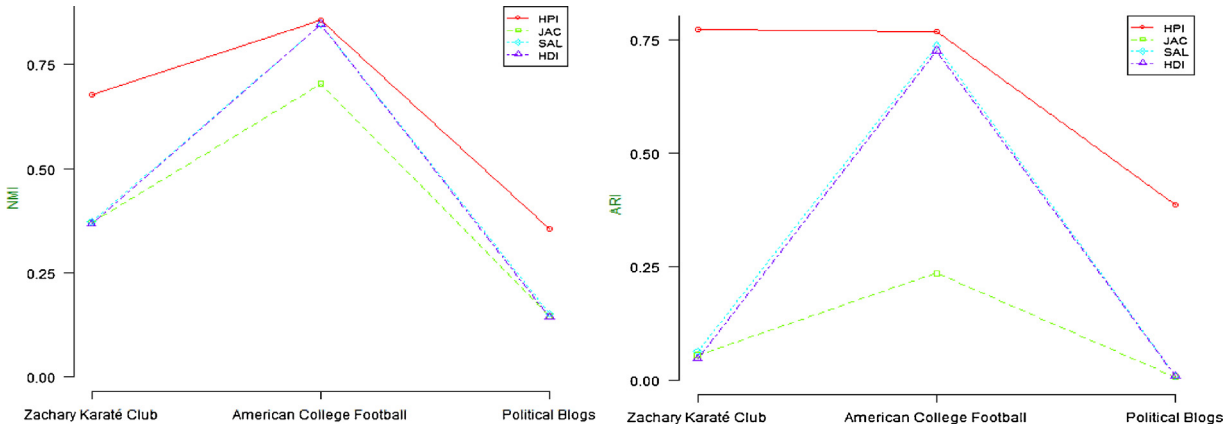


Fig. 8. NMI and ARI results for the first algorithm LCDA1 using HPI, Jaccard, Salton and HDI similarities.

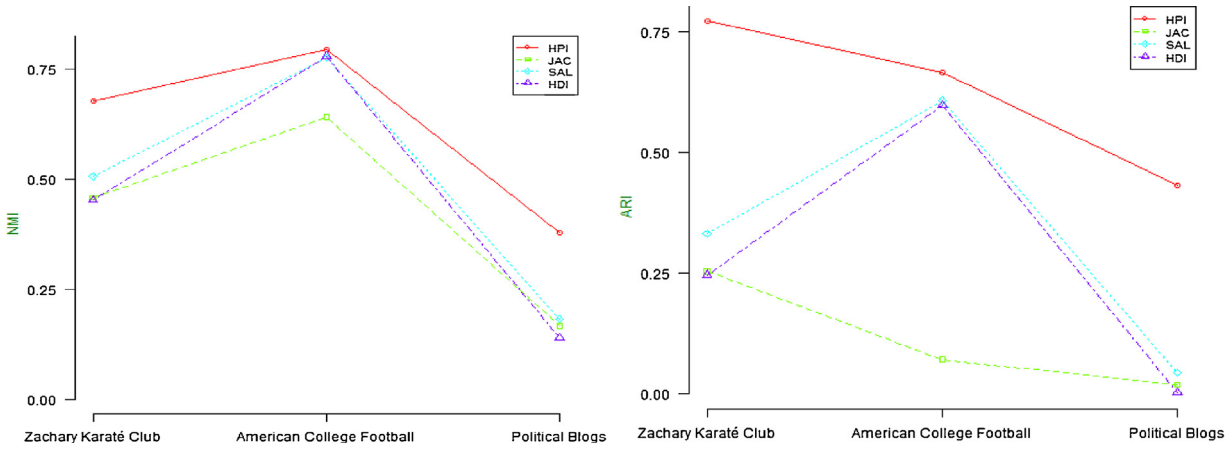


Fig. 9. NMI and ARI results for the second algorithm LCDA2 using HPI, Jaccard, Salton and HDI similarities.

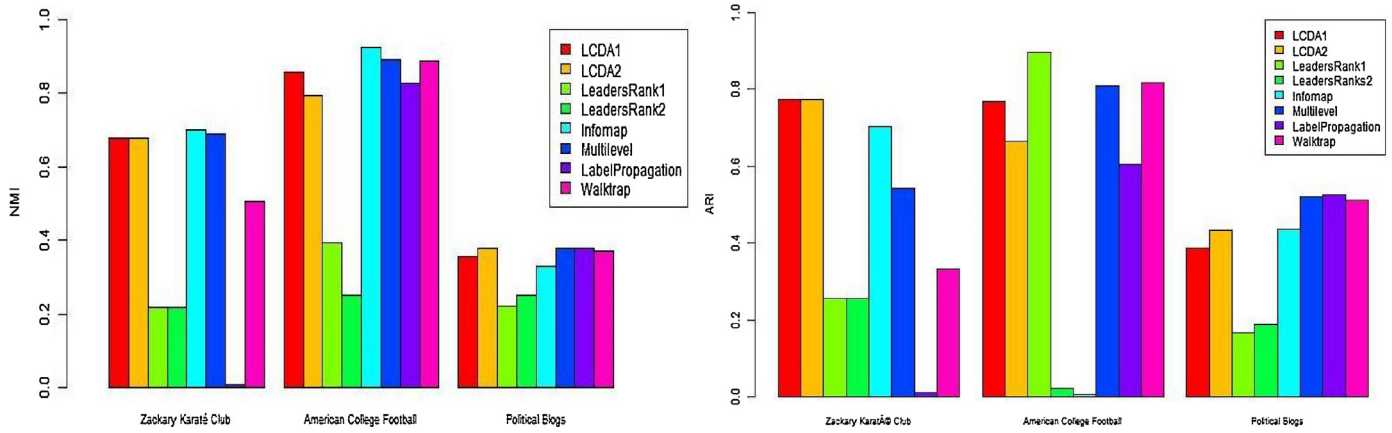


Fig. 10. NMI and ARI results for the second algorithm LCDA2 using HPI, Jaccard, Salton and HDI similarities.

pared to *LCDA1*, but the both algorithms failed to produce a result with a high *ARI* value compared to others algorithms.

As mentioned above, the proposed approach deals with undirected and unweighted networks. The proposed algorithms *LCDA1* and *LCDA2* represent improved versions of the *LeadersRank1* and *LeadersRank2* algorithms. The proposed algorithms are scalable and deterministic and its complexity is $O(n^2)$. By comparing these algorithms, it is noticeable that the improved versions (*LCDA1*, *LCDA2*) improve *ARI* and *NMI* values relative to the *LeadersRank1*, *LeadersRank2*, *LabelPropagation* algorithms. Since the number of the detected communities is the same as represented in ground truth, it can be argued that the difference in scores of *NMI* or *ARI* compared to other algorithms (*Infomap*, *Multilevel* and *Walktrap*) is due to the weights or the direction of the edges of the datasets, which are not considered by our approach.

5. Conclusion

The Leader-Community Detection Algorithms are introduced for finding community structure in social networks. The proposed

approach shows a major advantage, which does not require prior knowledge of the number of leaders and communities. The proposed approach begins with leader detection to find the most influential nodes of the network, followed by community detection. For each detected leader, its community is built by seeking similar nodes. Experimental results demonstrate that the performance of the proposed algorithms *LCDA1* and *LCDA2* is reliable in terms of accuracy and finding community structure based on modularity, *NMI* and *ARI*. While the *ARI* values are not as reliable for larger size networks, it still represents a significant improvement relative to previous versions of *LeadersRank* algorithms.

References

Adamic, L.A., Glance, N., 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, New York, NY USA, pp. 36–43.

Ahajjam, S., Badir, H., El Haddad, M., 2015a. Detection of leader's nodes in complex networks. Glob. J. Eng. Sci. Res., 40–44, vol. SESA 2014.

Ahajjam, S., Haddad, M.E., Badir, H., 2015b. LeadersRank: towards a new approach for community detection in social networks. 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), 1–8.

Ahajjam, S., Badir, H., Fissoune, R., Haddad, M.E., 2015. In: Esposito, F., Pivert, O., Hacid, M.-S., Rás, Z.W., Ferilli, S. (Eds.), Communities Identification Using Nodes Features, in Foundations of Intelligent Systems. Springer International Publishing, pp. 303–312.

Bader, D.A., Meyerhenke, H., Sanders, P., Wagner, D., 2013. Graph partitioning and graph. Cluster. Am. Math. Soc.

Bae, J., Kim, S., 2014. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Phys. Stat. Mech. Appl. 395, 549–559.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp., p. P10008.

de Arruda, G.F., Barbieri, A.L., Rodríguez, P.M., Rodrigues, F.A., Moreno, Y., Costa, L.F., 2014. Role of centrality for the identification of influential spreaders in complex networks. Phys. Rev. E 90 (September (3)), p. 32812.

Domingos, P., Richardson, M., 2001. 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.

Fortunato, S., 2011. Community detection in graphs. Phys. Rep. 486 (February (3–5)), 75–174.

Fu, Y.-H., Huang, C.-Y., Sun, C.-T., 2014. Identifying super-spreader nodes in complex networks. Math. Probl. Eng., p. e675713.

Fuong, H., Maldonado-Chaparro, A., Blumstein, D.T., 2015. Are social attributes associated with alarm calling propensity? Behav. Ecol., p. aru235.

Ganesan, P., Garcia-Molina, H., Widom, J., 2003. Exploiting hierarchical domain structure to compute similarity. ACM Trans. Inf. Syst. 21 (January (1)), 64–93.

Gao, S., Ma, J., Chen, Z., Wang, G., Xing, C., 2014. Ranking the spreading ability of nodes in complex networks based on local structure. Phys. Stat. Mech. Appl. 403, 130–147.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99 (12), 7821–7826.

Ilakiya, P., Sumathi, M., Karthik, S., 2012. A survey on semantic similarity between words in semantic web. 2012 International Conference on Radar Communication and Computing (ICRCC), 213–216.

Kernighan, B.W., Lin, S., 1970. An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. 49 (2), 291–307.

Khorasgani, R., Chen, J., Zaïane, O.R., 2010. Top leaders community detection approach in information networks. In: 4th SNA-KDD Workshop on Social Network Mining and Analysis, Washington DC.

Lu, L., Zhou, T., 2011. Link prediction in complex: a survey. Phys. Stat. Mech. Appl. 390 (6), 1150–1170.

Moody, J., White, D., 2003. Structural cohesion and embeddedness: a hierarchical concept of social groups. Am. Sociol. Rev. 68 (February (1)), 103–127.

Moreno, J.L., 1955. Moreno (J.L.)—~~Fondements de la sociométrie~~. Traduit d'après la seconde édition américaine (Who Shall Survive?) par H. Lesage et P.-H. Maucorps. Rev. Fr. Sci. Polit. 5 (3), 641–646.

Networks, 2010. An Introduction. Oxford University Press, Oxford, New York.

Newman, M.E.J., 2004. Analysis of weighted networks. Phys. Rev. E 70 (5).

Newman, M.E.J., 2013. Community detection and graph partitioning. EPL Europhys. Lett. 103 (July (2)), p. 28003.

Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks (long version), arXiv:physics/0512106.

Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76 (3).

Rawashdeh, A., Ralescu, A.L., 2015. Similarity measures in social networks—a brief survey. Greensboro NC In: presented at the Modern AI and Cognitive Science Conference, NC A&T State U, vol. 1353, pp. 153–159.

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

Rosvall, M., Axelsson, D., Bergstrom, C.T., 2009. The map equation. Eur. Phys. J. Spec. Top. 178 (1), 13–23.

Ruhnau, B., 2000. Eigenvector-centrality—a node-centrality? Soc. Netw. 22 (4), 357–365.

Shah, D., Zaman, T., 2010. Community detection in networks: the leader-follower algorithm. Phys. Stat. (November), ArXiv10110774.

Social Network Analysis—Community Detection and Evolution | Rokia Missaoui | Springer.

Spizzirri, L., 2011. Justification and application of eigenvector centrality. Algebra Geogr. Eig. Netw.

Wang, Y., Di, Z., Fan, Y., 2011a. Identifying and characterizing nodes important to community structure using the spectrum of the graph. PLoS One 6 (11), p. e27418.

Wang, Y., Di, Z., Fan, Y., 2011. Detecting Important Nodes to Community Structure Using the Spectrum of the Graph, ArXiv11011703 Phys.

Wu, Q., Qi, X., Fuller, E., Zhang, C.-Q., 2013. Follow the leader: a centrality guided clustering and its application to social network analysis. Sci. World J. 2013 (October), p. e368568.

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

Xie, J., Kelley, S., Szymanski, B.K., 2013. Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput. Surv. 45 (August (4)), 43:1–43:35.

Yang, J., McAuley, J., Leskovec, J., 2014. Community Detection in Networks with Node Attributes, ArXiv Prepr. ArXiv14017267.

Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33 (4), 452–473.

Zhou, J., Zhang, Y., Cheng, J., 2014. Preference-based mining of top- influential nodes in social networks. Future Gener. Comput. Syst. 31, 40–47.