IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON GREEN SIGNAL PROCESSING FOR WIRELESS COMMUNICATIONS AND NETWORKING

Received December 6, 2017, accepted January 9, 2018, date of publication January 16, 2018, date of current version March 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2794324

Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network

MULUNEH MEKONNEN TULU[1,2], RONGHUI HOU[1], AND TALHA YOUNAS[1]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]Department of Electrical and Computer Engineering, Addis Ababa Science and Technology University, Addis Ababa 16417, Ethiopia

Corresponding author: Ronghui Hou (rhhou@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571351 and Grant 61401326, in part by the important National Science & Technology Specific Projects under Grant 2015ZX03002006003, and in part by the Natural Science Basic Research Plan in Shaanxi Province of China Program under Grant 2016JM6028 and Grant 2016JQ6054.

ABSTRACT Applying effective methods to identify important nodes in a complex network is highly invaluable. Recently, in a complex network, finding a powerful leader of the community to spread information quickly throughout the network is the concern of many researchers. In this paper, to identify influential nodes in a large and complex network, community-based mediator (CbM), which considers the entropy of a random walk from a node to each community is proposed as a metrics. CbM describes how the node is essential to connect two or more than two communities of the network. Correlations between CbM and other classical methods used to identify influential nodes are discussed. The performance of CbM is evaluated by susceptible-infected-recovered (SIR) model. In SIR model, the node is the most powerful node in the network, if the percentage of infected node is more while the node is used as the source of infection. Simulation results show that the proposed method performs better than the existing methods to spread information quickly and it can also introduce new influential nodes that other methods failed to identify.

INDEX TERMS Complex network, community-based mediator, influential nodes, susceptible-infected-recovered model.

I. INTRODUCTION

With significant theoretical and practical importance, the studies on epidemic, social and technological networks become one of the most attractive domains in many branches of sciences [1]. From the existence of network science to its current dramatic progress, finding the influential nodes to spread information in the complex networks is a crucial issue for researchers [2]. Nodes that are more likely to be infected and to infect a larger number of nodes in a network are influential nodes [3].

Understanding and analyzing the network topology has become an essential part to select important nodes in a complex network [4]. The significance of a node can have different meanings depending on its application. Betweenness centrality, degree centrality, and closeness centrality are the three common measures of node centrality formalized by Freeman [5]. Betweenness centrality connects and controls the interaction between the two nonadjacent nodes, but it has failed to apply it in large and complex networks. Degree centrality is defined as a number of nodes that a focal node is connected to, but it has failed to consider the global structure of the network. Closeness centrality is a node which can disseminate information to others very effectively, but it did not consider the neighbourhood nodes. It has shortest paths to all other nodes.

In last decade years, significant attention is given to selecting influential nodes to accelerate the diffusion of the information in complex networks. Eigenvector is one of the methods discovered to identify influential nodes in a complex network [6], [7]. It takes into account the impact of a single node in a network as the impact of all other nodes. PageRank [8] is introduced to select influential nodes based on random-walk method. Moreover, in [9] the authors introduced ClusterRank which considers nodes clustering coefficient. Meanwhile, N. Salamanos *et al.* [10] come up with Rank Degree which is based on graph sampling. In the past three years, several centrality measures are also proposed to identify influential nodes, such as by combining the existing centrality measures [11], Weighted LeaderRank [12], Neighborhood coreness centrality [13], Evidence theory and local structure [14], VoteRank [15], Pareto Shell decomposition [16], Efficiency centrality [17], Local Entropy (LE) [18], and Entropy of Betweenness Centrality (EBC) [19]. Of all these methods, none of them can be used in all kinds of networks perfectly. Also, some of them are not so good to select influential nodes as a group to spread information in a network. Therefore, identifying the influential nodes is still an open issue.

A node having a larger number of less influential neighbours may be less influential than a node having a few highly influential neighbours in the center of the network [18], [19]. Considering this fact, Qi Zhang *et al.* [18] proposed a local structure of complex network to quantify a node's influence based on the degree of neighbour nodes. Their main idea is to use the influence of the local network to replace the node's influence on the whole network. However, this method may fail to rank the influential nodes within each community/or cluster accurately. Chen D.B *et al.* [9] proposed a method to identify influential nodes to spread information in the networks by the role of clustering and who also proposed to increase the credentials of influential nodes by the path diversity [20]. In the network, the connections between the communities are scattered, while nodes in each community connect with each other firmly, since the actual network usually has a community structure [21]. Therefore, for ranking nodes based on their importance to disseminate information in the complex network, we can consider the community property of a node. For example, Hu *et al.*(2013) proposed an improved influential node selection method based on the centrality of K-shell which validated in the SIR Model [22]. The importance measure of a node is the number of communities that can be linked to a node [23], [24].

For large and complex networks, the number of communities depends on the community detection algorithm. Therefore, only considering the number of communities that are directly linked to the node is not quite enough to measure the importance of a node. To overcome these, Zhiying Zhao *et al.* [24] proposed Community-based Centrality (CbC), which is used to identify the influential spreaders based on the network community structure. It is true that influential node is a node that maximizes the spread of information in the network, but it may not the only way to define it.

Here in this paper, we proposed influential node selection method, which we call Community-based Mediator (CbM) Method. It reflects the influence of a node by considering the entropy of the random walk from a node to each community. Nodes selected by CbM are key nodes to spread information in the network quickly. On the other hand, absence of these nodes from the network highly affects the spreading speed of information within the network. Therefore, this kind of nodes plays a great role in the communities by receiving and passing information.

II. PRELIMINARIES

Degree centrality, closeness centrality, and betweenness centrality are the most common traditional way to select the influential nodes in the networks. Therefore, to measure the impact of a node in the network, centrality plays a great role [25].

A. DEGREE CENTRALITY

The number of neighbours or edges the node has in a network is simply expressed as a degree of a node. The nodes with many friends in networks have a high probability to disseminate information when compared to those nodes with a few friends. Degree centrality can be expressed as:

$$C_D(i) = d_i = \sum_j a_{ij} \tag{1}$$

where a_{ij} is equal to 1 if and only if node i is connected to node j, otherwise it is zero [26].

B. BETWEENNESS CENTRALITY

One of the popular methods used to identify the powerful nodes in network is betweenness centrality. It counts the shortest path through the node. Betweenness centrality can be defined as:

$$C_B(i) = \frac{v(i)}{\sum \partial st}, \quad (s \neq i \neq t \in v) \tag{2}$$

where $v(i)$ is the number of the shortest paths through the vertex i and ∂st is the number of the shortest paths from vertex s to vertex t.

C. COMMUNITY-BASED CENTRALITY

It is proposed to calculate the importance of node by considering the link connecting nodes within the community and out the community. It defined as:

$$CbC_i = \sum_{h=1}^{c} d_{ih} \frac{S_h}{N} \tag{3}$$

where d_{ih} is the number of links between node i and other nodes in community h, c is the number of communities in the network, S_h is the number of nodes in community h (the size of community h), and N is the total number of nodes in the network [24].

D. OBJECT DIFFUSION MECHANISM IN NETWORKS

We use SIR model to show the effectiveness of our proposed method to select influential nodes to disseminate information in complex network. The process is as follows.

First, we choose one node or some nodes as a source of infection and set the infection time. Let N is the total

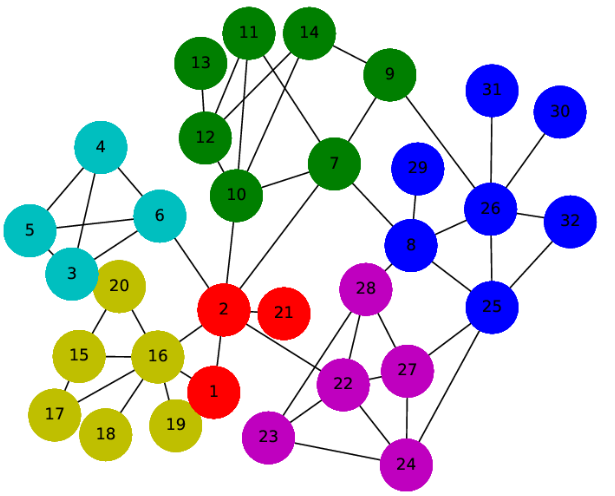


FIGURE 1. **Community structure of a toy network.**

number of nodes, $S(t)$ is a susceptible number of nodes, $I(t)$ is infected number of nodes, and $R(t)$ is recovered number of nodes at time t. This relationship can be expressed as:

$$N = S(t) + I(t) + R(t), \quad \forall t \tag{4}$$

From this, we set differential equations.

$$\frac{dS}{dt} = -\beta S(t)I(t) \tag{5}$$

$$\frac{dI}{dt} = (\beta S(t) - k)I(t) \tag{6}$$

$$\frac{dR}{dt} = kI(t) \tag{7}$$

where β is infection rate (on contacts) and k is the recovery rate.

The ratio of infected nodes to a total number of nodes indicates how the node is powerful to spread information in the complex network.

E. SHANNON ENTROPY

In 1948 the idea of information entropy was introduced by Claude Shannon [27]. The definition used in statistical thermodynamics is directly analogous to the definition of entropy used in information theory. It is the average amount of information produced by a probabilistic stochastic source of data [28]. Generally, Shannon Entropy, which described by probability theory is used to measure uncertainty in the system [27]. It defined as:

$$H_{shannon} = -\sum_{i=1}^{n} p_i \log(p_i) \tag{8}$$

III. COMMUNITY-BASED MEDIATOR NODES SELECTION METHOD

Most networks naturally divided into modules or communities [29]. Here in this part, strongly connected n-mobile node networks with weight matrix are considered [5], [30]. $A = [a_{ij}]$, i.e., $a_{ij} > 0$ indicates the weight of the edge from node i to j, when the network is binary (unweighted) it set to 1 while 0 if the edge does not exist. For all $i = 1, 2, \ldots, n$, we assume $a_{ii} = 0$. For a directed network the internal and external strength of node i is denoted by $s_i^{in} = \sum_j a_{ji}$ and $s_i^{out} = \sum_j a_{ij}$, respectively, and the total strength is defined as $\delta_i = \sum_j (a_{ij} + a_{ji})/2$. For undirected network, it defined as $\delta_i = \sum_j a_{ji} = \sum_j a_{ij}$.

A. COMMUNITY-BASED MEDIATOR NODES

We take into consideration that each node has internal and external density. In the same community, the ratio of the sum of edges of the node i within the community to the total edges of a node i in the network is the internal density of the node. In the same manner, the external density of the node i defined as the ratio of the sum of edges of the node i connected to other communities to total edges of node i in the network. Internal density is expected to be larger than external density of the node. We can get the impact of a node to share or disseminate information within the community from internal density and with other communities from the external density of the node. Therefore, the importance of nodes can be calculated by both characteristics of densities and the size of networks. We can assume in social networks if a person has many friends in different communities, he can play significant roles to receive and diffuse information around his circle to a large extent or more quickly than others [24]. As a result, our proposed scheme (CbM) considers the external and internal density of the node, and a number of friends the node has in the network to calculate the impact of the node to receive and diffuse information within and across the communities. We proposed the following three steps to calculate the CbM of the node i.

Step 1: Calculate the internal and external density of the node via the following formula:

$$\rho_i^{in} = \frac{\sum_j a_{ij}}{d_i}, \quad i \in h,\ j \in h \tag{9}$$

From Eq.(9), it is clear that a_{ij} indicates the friends or degrees of node i within community h, d_i implies the total friends of node i in the network and ρ_i^{in} indicates the internal density of node i. Node i which belongs to community h and its edge from community h to other community is external density of node i. It is calculated as follow:

$$\rho_{i_{h1}}^{ex} = \frac{\sum_j a_{ij}}{d_i}, \quad i \in h,\ j \in h_1 \tag{10}$$

where $\rho_{i_{h1}}^{ex}$ is external density from node i which belongs to community h to other node j which belongs to community h_1. $a_{ij}, i \in h, j \in h_1$ is the sum of outgoing edges from node i in community h to other nodes in community h_1.

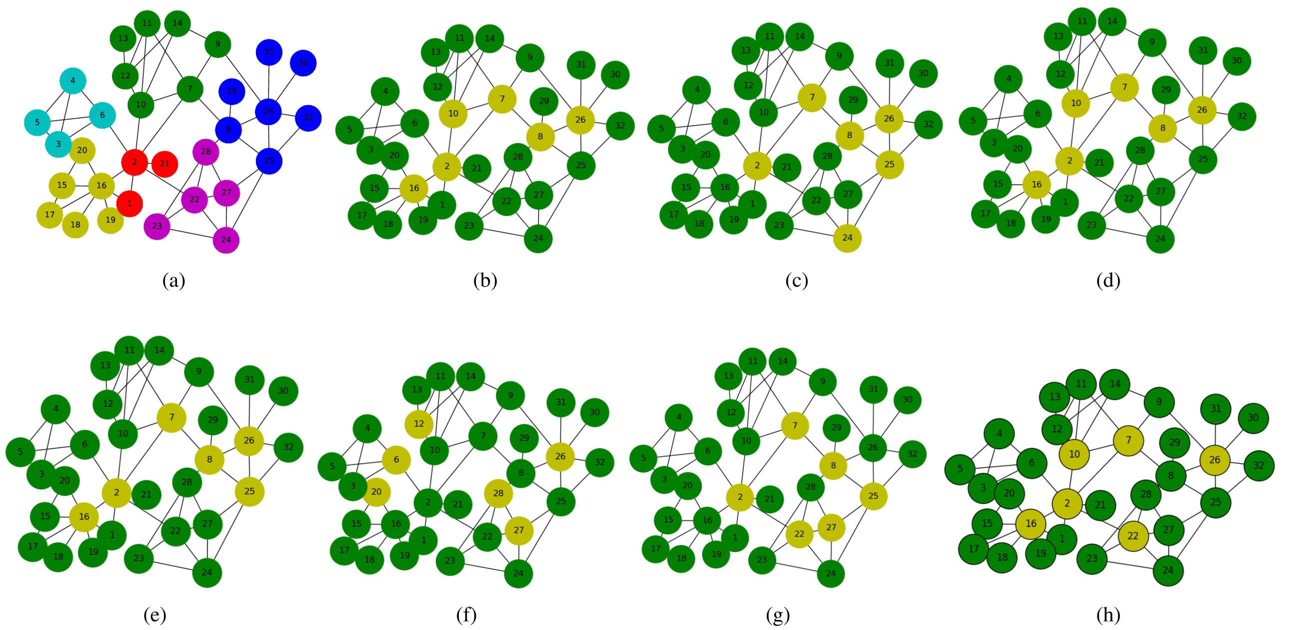


FIGURE 2. Applying different methods to identify influential nodes in a toy network: (*a*) The different colour of node shows the community of the node; from (*b*) to (*h*), nodes with green colour are ordinary or normal nodes while nodes with yellow colour are influential nodes.

Step 2: Calculate the entropy of internal and external density of the node i via the following formula:

$$H_i = \left[-\sum \rho_i{}^{in} log(\rho_i{}^{in})\right] + \left[-\sum \rho_{ih_1}{}^{ex} log(\rho_{ih_1}{}^{ex})\right] \tag{11}$$

where H_i is the entropy of node i.

Step 3: Calculate the CbM of the node i via the following formula:

$$CbM_i = H_i \times \frac{d_i}{\sum_{i=1}^{N} d_i} \tag{12}$$

Where CbM_i is the community-based mediator value of node i, d_i is the total number of degrees of node i, and $\sum_{i=1}^{N} d_i$ is the sum of total degree of the networks. For generalization, if the external and internal density of the node is equal, then CbM of the node will be its normalized degree, i.e., $CbM_i = \frac{d_i}{\sum_{i=1}^{N} d_i}$.

Currently, a diversity of community detection algorithms have been proposed in [21], [30]–[32]. Increasing modularity value is a confirmation of the good community detection partition since high modularity values resulted in the occurrence of cluster nodes with comparatively large intra-community edges. In this paper α-partition proposed by C. Piccardi [30] is adopted. To make it clear, we consider a toy network with 32 nodes and 52 edges as given in Fig. 1. Some centrality indices and CbM of nodes in a toy network are listed in Table 1. The network is divided via α-partition into six communities ($c = 6$) by adjusting α value to 0.4.

TABLE 1. The influence of each node measured by different methods in a toy network.

ID	Degree	Betweenness	CbC	CbM	Page Rank	Eigen Vector	LE
1	2	0	0.2813	0.0192	0.017	0.122	1.4186
2	7	21	1.0938	0.1505	0.024	0.327	2.8329
3	3	0	0.3750	0	0.017	0.057	1.9878
4	3	0	0.3750	0	0.021	0.057	1.9878
5	3	0	0.3750	0	0.03	0.057	1.9878
6	4	0	0.4688	0.0312	0.06	0.12	2.2261
7	5	17.5	0.9688	0.0659	0.02	0.283	2.522
8	5	13.167	1.0313	0.0659	0.02	0.267	2.4566
9	3	8.333	0.6563	0.0265	0.02	0.155	1.9328
10	5	8	0.9688	0.0347	0.035	0.255	2.5186
11	3	0	0.6563	0	0.039	0.165	1.9713
12	4	7	0.8750	0	0.042	0.142	2.18
13	1	0	0.2188	0	0.035	0.034	0.7219
14	3	4	0.6563	0	0.026	0.134	1.9656
15	3	1	0.5625	0	0.039	0.072	1.7783
16	7	10	1.1250	0.0581	0.027	0.176	2.6415
17	2	0	0.3750	0	0.021	0.06	1.3844
18	1	0	0.1875	0	0.021	0.043	0.5436
19	1	0	0.1875	0	0.021	0.043	0.5436
20	2	0	0.3750	0	0.055	0.06	1.3844
21	1	0	0.0938	0	0.02	0.079	0.5436
22	5	9.5	0.7188	0.0347	0.02	0.316	2.5321
23	3	0	0.4688	0	0.021	0.2	1.9772
24	4	12.5	0.6875	0.0312	0.03	0.257	2.2983
25	5	15.833	0.9688	0.0467	0.034	0.275	2.5186
26	6	21.167	1.3125	0.0375	0.04	0.225	2.546
27	4	2	0.6875	0.0312	0.043	0.267	2.3131
28	4	0	0.6875	0.0312	0.071	0.255	2.2983
29	1	0	0.2188	0	0.021	0.065	0.65
30	1	0	0.2188	0	0.028	0.054	0.5917
31	1	0	0.2188	0	0.028	0.054	0.5917
32	2	0	0.4375	0	0.037	0.121	1.4605

As we can observe from Table 2, the most six nodes of a toy network identified by CbM is also identified by degree centrality and betweenness centrality. The top five nodes

TABLE 2. The most six influential nodes in a toy network.

Method	1	2	3	4	5	6
Degree	2	16	26	7	8	10
Betweenness	26	2	7	25	8	24
CbC	26	16	2	8	7	10
CbM	2	7	8	16	25	26
Pagerank	28	6	20	27	12	26
Eigenvector	2	22	7	25	8	27
LE	2	16	26	22	7	10

identified by CbM is a combination of top five nodes identified by degree and betweenness centrality. From this, we can say CbM holds properties of degree and betweenness centrality for a toy network. The details of the top six influential nodes selected by different methods are shown in Fig. 2.

IV. EXPERIMENT RESULTS AND ANALYSIS

A. DATASET DESCRIPTION

1) ZACHARY'S KARATE CLUB NETWORK

Karate club social network was studied by Wayne W. Zachary from 1970 to 1972 [33]. The conflict between Mr. Hi (node 1) and Mr. John A. (node 33) divided the 34 members of sports club into two groups.

2) AMERICAN FOOTBALL NETWORK

American football network [21] is a network which has $N = 115$ teams or nodes, and the game between two teams represents the relation or edges between them.

3) SYNTHETIC NETWORKS

The Barabsi-Albert (BA) network was built with the standard "preferential attachment" algorithm [34] to demonstrate the impact of CbM on unclear community structure network. Also, Synthetic network I and Synthetic network II is generated to validate the impact of internal strength or density of communities on CbM.

4) DOLPHIN NETWORK

Dolphin network is an undirected social network of regular communications among 62 dolphins in a community living off Doubtful Sound, New Zealand [35].

5) AIRPORT NETWORK

Airport network is directed and weighted network, which is the largest connected flight in the worldwide level between 2868 destinations [36]. The number of routes between the two airports is the weight of directed edges.

6) INTERNET NETWORK

At the level of autonomous systems, internet network is a symmetrized snapshot of the structure of the Internet (for July 22, 2006). It is undirected and binary social network which consists of $N = 11745$ [37].

To evaluate the effectiveness of proposed method to select influential nodes to speed up the dissemination of information in the complex network, we considered real and synthetic

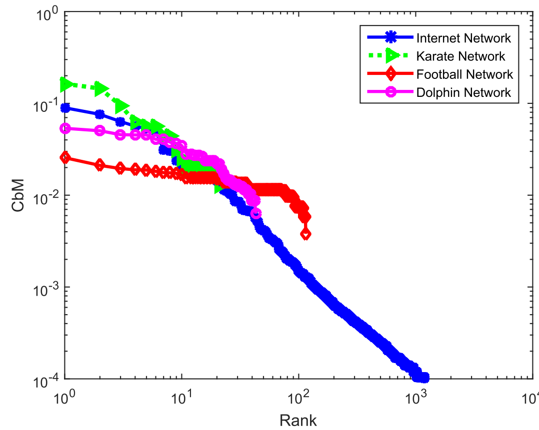


FIGURE 3. CbM distribution for four different real data networks.

TABLE 3. The most six influential nodes in Zachary's Karate Club network.

Method	1	2	3	4	5	6
Degree	34	1	33	3	2	4
Betweenness	1	34	33	3	32	9
CbC	34	1	33	3	2	4
CbM	1	34	3	33	2	32
PageRank	34	33	32	17	15	7
Eigenvector	34	1	3	33	2	9
LE	34	1	33	3	2	4

networks. Some related studies exist in [38], [39] that deal with the complex networks follows a power law at least asymptotically. That is, many nodes make a small impact on the network and small nodes make a dominant impact if we consider CbM as the strength of a node to the community. Fig. 3 shows the distribution of CbM for four real data networks. From this, we can say that for most networks, CbM selection distribution is downward sloping. Therefore, CbM selection follows power law distribution asymptotically. Table 4 shows the basic topological properties of the all synthetic and real networks used for the experiment.

TABLE 4. The basic topological features of all synthetic and real networks used for the experiment. n and m denote the total numbers of nodes and edges, respectively. $< k >$, $k_{\max}$ and $k_{\min}$ represent average, maximum and minimum degree, respectively.

Network	n	m	$< k >$	$k_{\max}$	$k_{\min}$
Zachary's Karate Club	34	156	4.5882	17	1
American football	115	1234	10.7304	14	6
Barabási-Albert (BA) network	1000	1998	1.9980	86	1
Synthetic network I	15	32	2.133	4	1
Synthetic network II	15	50	3.333	5	2
Dolphin	62	317	5.1129	12	1
Airport	2868	44361	15.4676	352	1
Internet	1145	57126	4.8639	1651	1

B. CbM WITH CLEAR COMMUNITY STRUCTURE NETWORK

1) ZACHARY'S KARATE CLUB NETWORK

To show the property of CbM with clear community structure network, we used Zachary's Karate Club network [33].

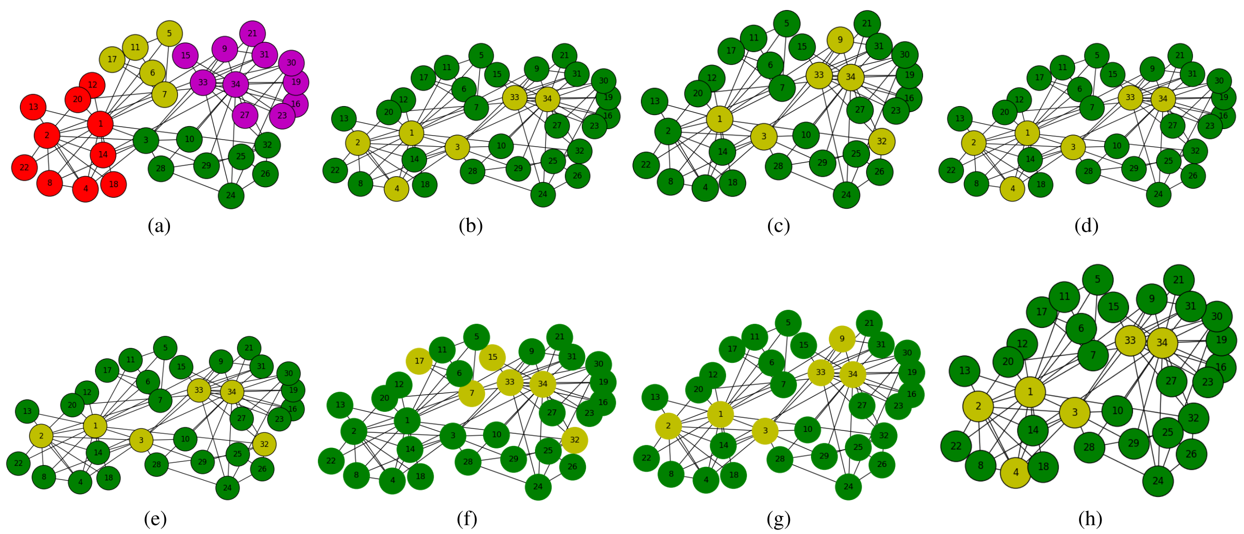


FIGURE 4. Applying different methods to identify influential nodes in Zachary Karate Club network: (*a*) The different colour of node shows the community of the node; from (*b*) to (*h*), nodes with green colour are ordinary or normal nodes while nodes with yellow colour are influential nodes.

The network is divided via α-partition into four communities ($c = 4$) by setting α value to 0.5. The modularity value of the network is 0.41. The details visualization of Zachary's Karate Club network structure and influential nodes selected by different methods are shown in Fig. 4. The top six influential nodes identified by degree centrality, betweenness centrality, CbC, CbM, PageRank, Eigenvector, and LE are shown in Table 3.

The results of our test on the Zachary's Karate Club network show that the CbM can identify the nodes that play a great role in the communities by receiving and passing information in the network. From Table 3, we observe that for Zachary's Karate Club network, all nodes identified by CbC and LE are also identified by degree centrality. As a result, we can say CbC and LE holds only the property of degree centrality for the network mentioned above. However, our proposed method and eigenvector hold the properties of degree and betweenness centrality but not PageRank method.

2) AMERICAN FOOTBALL NETWORK

To show the property of CbM with clear structure network, we also used an American football network [21]. We have applied α-partition community detection and make 12 different communities, which is matched with ground truth community detection. Table 5 shows the top 20% of influential nodes of football network, which again indicates the effectiveness of CbM to selecte influential nodes in the networks with clear community structure.

C. CbM WITH WEAK AND STRONG INTERNAL DENSITIES OF COMMUNITIES

To demonstrate the property of CbM with weak and strong internal densities of communities, we generated the network shown in Fig. 5(a) and 5(b). Fig. 5 shows the structure of synthetic network I and II. The network is partitioned to three communities by α-partition. Table 7 demonstrates the member and internal strength of each community, and also Table 8 demonstrates the quality measure (modularity) of each community structure and influential nodes selected by CbM. The internal strength of the community for the synthetic network II is greater than the synthetic network I. Also, the Modularity measure of synthetic network II is greater than the synthetic network I. However, Top influential nodes selected by CbM for both networks (synthetic network I and II) are the same. From this point of view, our proposed method is not affected by the internal strength of community rather than affected by community structure (number of community), internal and external strength of the node.

TABLE 5. 20% of influential nodes in football network which are selected by CbM.

Rank	ID	CbM	Rank	ID	CbM
1	1	0.0256	12	4	0.0154
2	81	0.0211	13	7	0.0154
3	83	0.0195	14	8	0.0154
4	87	0.0189	15	16	0.0154
5	59	0.0188	16	33	0.0154
6	70	0.0182	17	50	0.0154
7	64	0.0176	18	89	0.0154
8	25	0.0175	19	115	0.0154
9	105	0.0173	20	13	0.0149
10	113	0.0166	21	17	0.0149
11	2	0.0154	22	20	0.0149

D. CbM WITH UNCLEAR COMMUNITY STRUCTURE NETWORK

To show the property of CbM with unclear community structure network, we have generated 1000 nodes using Barabas-Albert model [34]. It starting from a connected node pair, the remaining 998 nodes were iteratively added one

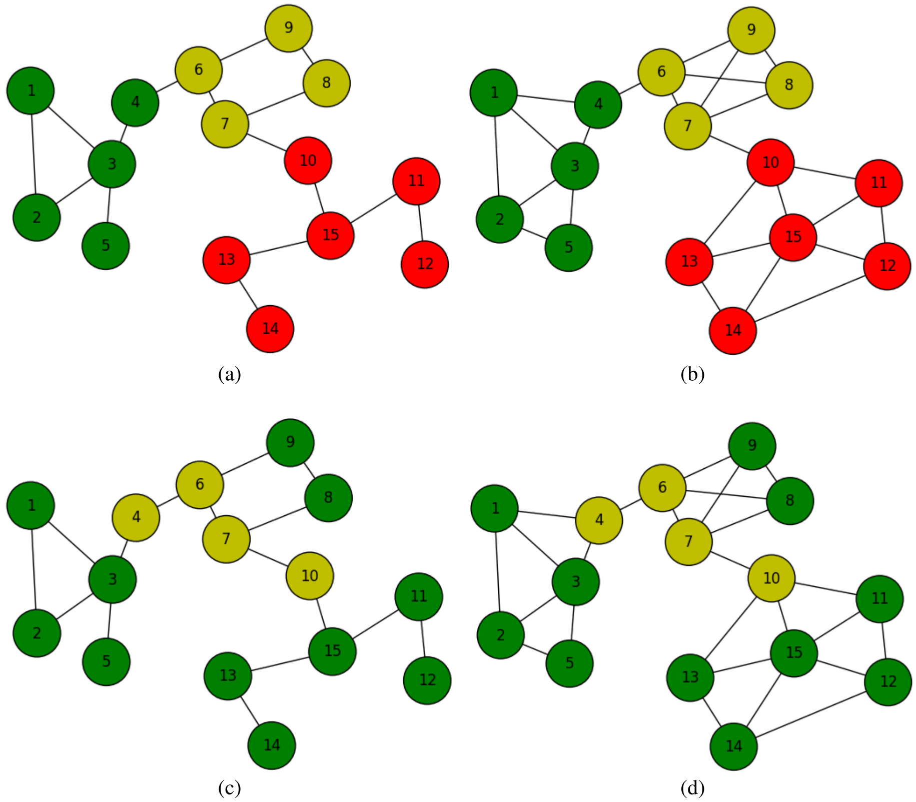


FIGURE 5. Applying CbM method to identify influential nodes in synthetic network I and II which are given in (a) and (b), respectively: In (a) and (b), the three different colors show the three communities of the network; In (c) and (d), nodes with green color are ordinary or normal nodes while nodes with yellow color are influential nodes.

TABLE 6. 3% of influential nodes in generated network which are selected by CbM.

Rank	Node ID at c=1	Node ID at c=170	Rank	Node ID at c=1	Node ID at c=170
1	4	4	16	15	44
2	30	3	17	19	47
3	3	30	18	24	7
4	1	1	19	33	19
5	20	10	20	34	21
6	6	6	21	44	12
7	10	20	22	47	40
8	8	8	23	11	118
9	2	2	24	12	49
10	16	99	25	29	179
11	179	16	26	61	22
12	22	79	27	196	34
13	79	81	28	40	11
14	81	15	29	67	29
15	99	33	30	233	196

at a time, by attaching each of their 2 edges to a node of the current network, randomly selected with a probability proportional to its current degree ($k = 2 \times 997/n \approx 4$). The degree distribution is shown in Fig.6.

We have applied CbM with unclear community network by assuming it as a one community. Also, we have partitioned the generated network into 170 communities using α-partition and applied CbM to identify influential nodes. Table 6 shows

TABLE 7. Community members and internal densities for synthetic networks.

Network ‘	Community	C_1	C_2	C_3
Synthetic Network I	Nodes	1,2,3,4,5	6,7,8,9	10,11,12, 13,14,15
	internal density	0.923	0.82	0.909
Synthetic Network II	Nodes	1,2,3,4,5	6,7,8,9	10,11,12, 13,14,15
	internal density	0.9412	0.821	0.946

the top 3% of influential nodes of 1000 generated network, which indicates the effectiveness of CbM to select influential nodes in the networks with unclear community structure. Most of top 3% influential nodes, which are identified by considering the network as 1 community and 170 communities, are the same. Therefore, it is possible to apply proposed method on networks with unclear community structure to select influential nodes.

E. EVALUATION WITH SUSCEPTIBLE-INFECTED-REMOVED MODEL

To demonstrate the effectiveness of the proposed method, we utilize SIR model. We applied it to infect the most influential nodes distinguish by various methods in the Dolphin network. Each infected node has one opportunity to infect

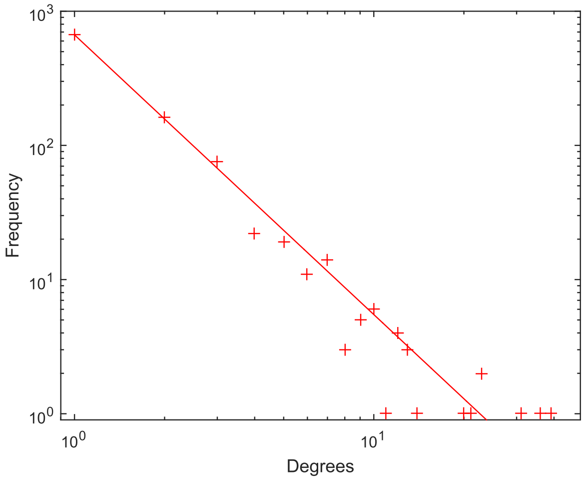


FIGURE 6. Degree distribution of 1000 generated networks.

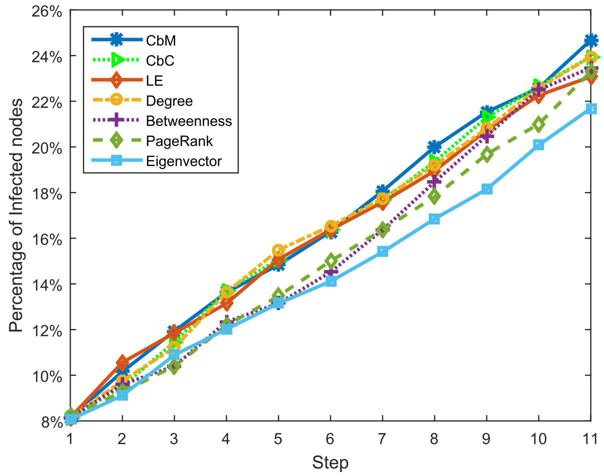


FIGURE 7. Impact of top 5 nodes selected by four different methods without removing any nodes in Dolphin network.

TABLE 8. CbM selection for synthetic networks.

Networks	α	Modularity	Influential Nodes selected by CbM
Synthetic Networks I	α=0.8	$Q = 0.54498$	4, 6, 7, 10
Synthetic Networks II	α=0.81	$Q = 0.5768$	4, 6, 7, 10

TABLE 9. The most ten influential nodes in Dolphin network.

Method	1	2	3	4	5	6	7	8	9	10
Degree	15	38	46	34	52	18	21	30	58	2
Betweenness	37	2	41	38	8	18	21	55	52	58
CbC	15	34	38	46	52	30	39	44	41	18
CbM	21	41	15	38	46	37	51	2	16	30
PageRank	52	56	58	48	51	61	53	28	46	50
Eigenvector	15	38	46	34	51	30	52	17	41	22
LE	15	46	38	34	21	52	30	58	2	14

other neighbour nodes in a step with the probability β, and it is assumed that at a fixed rate γ each infected individual is changed to the "removed" status. The source node is more influential as the more number of nodes are infected in the network. The top ten influential nodes of Dolphin network identified by degree centrality, betweenness centrality, CbC, CbM, PageRank, eigenvector, and LE are shown in Table 9.

To observe the importance of the presence and absence of top nodes selected by different methods, we considered Table 10. It shows less number of nodes infected when the top five nodes selected by CbM are removed. Nodes selected by all methods infect more number of nodes in the network if there is a presence of nodes selected by CbM. Therefore, we can say that nodes selected by CbM are the most mediator nodes, which connect communities to each other, and also the presence of this node maximizes, and the absence of it minimizes the spread of information in the communities.

The performance of the top 5 nodes selected by different methods without removing nodes from the network is shown in Fig. 7. From it, we demonstrated that nodes selected by CbM perform best to infect more nodes in the network mainly after step 7. The nodes selected by degree centrality and CbC perform better and so as LE, betweenness centrality and PageRank too, but nodes selected by eigenvector always perform the worst.

Table 11 demonstrates the influence of individual top five nodes selected by seven different methods. Nodes 15, 21, 34, 37, and 46 perform the best to infect neighbour nodes at each step. Node 21 which performs the best in the network is introduced to the most top 5 influential nodes by only CbM and LE methods. Our proposed method to select influential nodes in a large and complex network has a better chance to introduce new nodes which have a good impact on a network than some nodes selected by traditional methods. The other four nodes selected by CbM method are the combination of nodes selected by degree and betweenness centrality. So, we can say in Dolphin network, CbM holds the property of both degree and betweenness centrality.

Node 8, 15, 21, 34, 37, and 46 are top six nodes to spread information in Dolphin network (Table 11). Out of this top six nodes, only CbM and LE identify four of them under their top-six selected nodes, which performs the best to select influential nodes. However, CbM outperforms LE by ranking influential nodes in a good manner. CbC, degree centrality and eigenvector selects three under their top six selected nodes, which perform next to CbM. Betweenness centrality selects only two nodes, and PageRank performs the worst to select influential nodes in Dolphin network. PageRank selects none of the nodes from top six influential nodes of Dolphin network.

F. THE CORRELATIONS BETWEEN CbM AND OTHER METHODS

In this section, we investigated the Pearson correlation between our proposed method and some traditional methods that used to find influential nodes in large and complex networks. The Pearson correlation coefficient between CbMs

TABLE 10. Percentage of infected nodes at step 10.

Re-moved	Betweenness	Degree	CbC	CbM	PageRank	EigenVector	LE
Betweenness		0.2169	0.1887	0.19274	0.1621	0.1621	0.2185
Degree	0.1766		0.1887	0.17823	0.1411	0.1258	0.2371
CbC	0.1823	0.2129		0.19758	0.1282	0.1113	0.21048
CbM	0.1452	0.1936	0.1589		0.1073	0.1073	0.1565
PageRank	0.1919	0.2016	0.1855	0.1597		0.1919	0.1968
EigenVector	0.15	0.1984	0.1597	0.1323	0.129		0.21048
LE	0.1532	0.1944	0.1593	0.1395	0.1258	0.2177	

TABLE 11. The influence of top 5 nodes selected by seven different methods in Dolphin network to infect neighbour nodes at each step (step 1 to step 50).

Node ID	Step 1	Step 2	Step 3	Step 4	Step 5	Step 10	Step 20	Step 30	Step 40	Step 50
2	0.01613	0.01774	0.02339	0.02419	0.03145	0.05726	0.11048	0.2	0.26613	0.34194
8	0.01613	0.02016	0.02339	0.02661	0.02903	0.05565	0.125	0.22177	0.2871	0.36129
15	0.01613	0.02177	0.03226	0.04194	0.04516	0.08387	0.18871	0.28952	0.35081	0.36774
21	0.01613	0.02177	0.02903	0.03387	0.03548	0.06048	0.16613	0.27097	0.35081	0.37177
34	0.01613	0.02177	0.02742	0.03387	0.04354	0.0879	0.17419	0.27742	0.33871	0.36855
37	0.01613	0.02097	0.02661	0.02823	0.03468	0.06694	0.17419	0.27742	0.33871	0.36855
38	0.01613	0.02258	0.025	0.02903	0.03387	0.07016	0.175	0.27097	0.3355	0.34516
41	0.01613	0.02016	0.02177	0.02581	0.03387	0.07258	0.16048	0.26452	0.34355	0.34919
46	0.01613	0.02419	0.02903	0.03629	0.04194	0.075	0.16613	0.25484	0.35161	0.37177
48	0.01613	0.01294	0.0194	0.021	0.0259	0.0389	0.0794	0.167	0.2701	0.285
51	0.01613	0.01774	0.0226	0.0306	0.0387	0.0645	0.17903	0.2629	0.3193	0.3274
52	0.01613	0.01774	0.02016	0.02661	0.03468	0.05403	0.13871	0.22339	0.29677	0.32903
56	0.01613	0.01774	0.0194	0.021	0.021	0.0306	0.0774	0.1742	0.2452	0.3032
58	0.01613	0.0226	0.0274	0.0306	0.0403	0.0645	0.1226	0.1968	0.2274	0.27903

TABLE 12. The Pearson correlation coefficient(r) between CbM and other methods (c-indicates the number of communities).

Method	Zachary network	Dolphin network	Airport network
	$c = 4$	$c = 6$	$c = 112$
	r	r	r
Degree	0.96	0.82	0.86
Betweenness	0.95	0.92	0.83
CbC	0.94	0.72	0.81

(by α-partition) and other methods (betweenness centrality, degree centrality, and CbC are compared in Table 12. The relation between CbM and other methods to select influential nodes illustrated in Table 12 shows their positive correlation. The details of correlation matrix between CbM and other methods in airport network is shown in Fig. 8. It demonstrates the histogram and correlation between CbM and other methods.

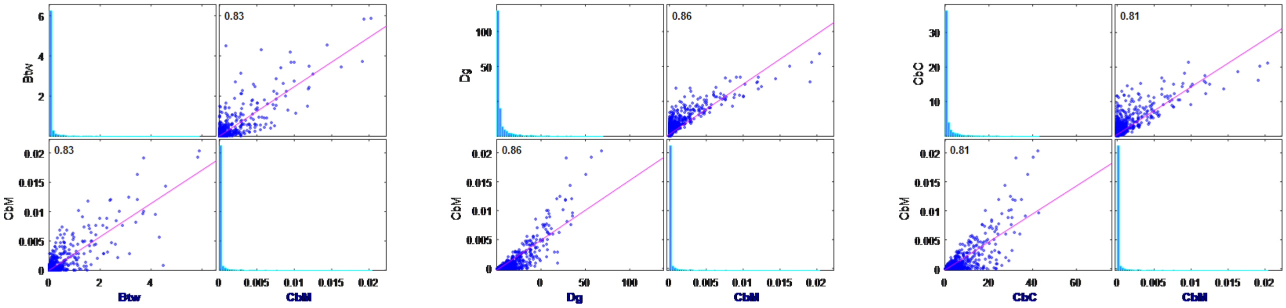
As shown in Fig. 8(a), betweenness centrality and CbM have a positive correlation. The node with large betweenness centrality also has high CbM value. In the airport network, as we see from Fig. 8, there are a number of nodes with high CbM, but its betweenness centrality is not large. Nodes with high CbM and large with betweenness centrality has good spreading capability. CbM has a higher capability to measures the importance of a node in a network than betweenness centrality since node with high CbM has higher spreading capability than a node with high betweenness centrality (see Table 11).

Fig. 8(b), illustrates the correlation between CbM and degree centrality. Degree centrality and CbM has a positive correlation. Therefore, a node with large degree centrality has higher CbM. However, there is also a node with high CbM but low with degree centrality. As shown in Table 9, a node with higher degree centrality and CbM has strong spreading capability. Nodes identified by CbM has a better spreading capability than nodes selected by degree centrality (Fig. 7). From Table 11, the first and top-ranked powerful node in dolphin network to spread information is node 21. It is selected and ranked first only in the top four influential nodes, which are identified by CbM. Other methods are failed to include this node under their top four influential nodes (Table 9).

Fig. 8(c) demonstrates the positive correlation between CbM and CbC. We can observe that there are many nodes with high CbC, but its CbM is not large. Node higher with both (CbM and CbC) has strong capability to spread information in the network (Table 11). Since nodes selected by CbM are the collection of the most influential nodes than selected by CbC, CbM is better than CbC to select important nodes in complex and large networks.

In Table 13, 1% of influential nodes identified by CbM, degree centrality, betweenness centrality, and CbC methods in airport network are demonstrated. CbM identifies more new influential nodes than others. Nodes 1602, 1629, 643, and 1395 are identified only by CbM. Moreover, CbC identifies only two new influential nodes (Node 1562 and 1701). From this point of view, CbM has more chance to introduce new intermediate nodes used to disseminate information in the network quickly than traditional methods.

In Table 14, 0.5% of influential nodes identified by CbM in internet network is described. We can apply our proposed



(a) (b) (c)

FIGURE 8. The histogram and correlation between CbM and other methods. a) The histogram and correlation between betweenness centrality and CbM. b) The histogram and correlation between degree centrality and CbM. c) The histogram and correlation between CbC and CbM.

TABLE 13. The influential nodes of Airport network identified by four different methods (1% influential nodes).

Rank	CbM ID	Degree ID	Betweenness ID	CbC ID
1	170	170	170	1535
2	585	585	585	170
3	232	254	946	585
4	254	1535	1707	254
5	946	230	1114	230
6	1712	723	1597	538
7	723	1673	232	1569
8	1754	946	113	232
9	538	1840	1535	659
10	659	232	1754	531
11	175	538	254	1754
12	113	659	1840	1776
13	1500	175	1630	1539
14	1535	531	1712	175
15	1597	1712	1673	1673
16	230	251	1500	1712
17	1776	1750	1503	1820
18	1388	1754	993	1500
19	993	260	79	1388
20	1840	1569	1532	218
21	1725	711	723	1701
22	1602	174	538	946
23	1629	218	1184	1562
24	1673	1776	1495	1597
25	643	113	1776	1750
26	1343	1343	1044	1725
27	1495	1597	1750	723
28	1395	1539	1725	260
29	711	268	1388	174

TABLE 14. The influential nodes of Internet network identified by proposed methods (0.5% influential nodes).

Rank	Node ID	CbM	Rank	Node ID	CbM
1	4	0.089522	31	42	0.008443
2	3	0.075909	32	14	0.007862
3	15	0.063414	33	13	0.006981
4	23	0.056432	34	18	0.006944
5	59	0.053419	35	46	0.006885
6	55	0.049818	36	28	0.006878
7	40	0.031514	37	45	0.006743
8	56	0.029966	38	61	0.00672
9	27	0.023949	39	296	0.006611
10	16	0.020336	40	35	0.006582
11	51	0.019078	41	37	0.006561
12	64	0.018487	42	26	0.005619
13	158	0.017795	43	19	0.005504
14	39	0.01763	44	112	0.005273
15	11	0.017605	45	1271	0.005031
16	12	0.017423	46	1761	0.004558
17	7	0.01689	47	63	0.004313
18	25	0.016057	48	189	0.004227
19	20	0.015824	49	219	0.004162
20	99	0.014053	50	34	0.004139
21	43	0.013918	51	161	0.004081
22	129	0.01258	52	72	0.004036
23	36	0.011454	53	1802	0.003825
24	24	0.011258	54	29	0.003764
25	53	0.010702	55	1454	0.00356
26	128	0.010225	56	1282	0.003437
27	1	0.009919	57	157	0.003432
28	38	0.00864	58	32	0.00333
29	21	0.008515	59	1272	0.003291
30	58	0.008466	60	1826	0.003286

method for large-scale network to identify influential nodes which are used to speed up the dissemination of information.

G. COMPUTATIONAL COMPLEXITY ANALYSIS OF CbM

The algorithm has four major parts. The first part of the algorithm will calculate the $N \times N$ node adjacent Markov matrix of each node and store it as matrix p, which is the probability matrix. Hence, this can be computed in $O(n)$ time. The second part is where it computes the strength of the node inside and outside of the community by multiplying p (probability matrix) and H (collecting matrix). Since this is nonsquare matrix multiplication, the complexity of this part is $O(mn < k >)$ (where k is number of clusters). However, k is usually far less than m and n. The third part of the algorithm will do the entropy of node. The running time of this part is $O(n)$. The final part calculates the CbM value of each node. Its running time is $O(1)$. Thus, the total complexity of the algorithm is the maximum of the four parts, and the asymptotic complexity will be $O(mn < k >)$.

In Fig. 9, the running times of CbM, LE, and CbC on real and synthetic nodes are presented, i.e, toy-32, karate-34, dolphin-62, football-115, airport-2868, and internet-11745 nodes. For this experiment, we have used a standard PC endowed with a 2.5 GHz Intel Core i7 CPU, 8GB of RAM

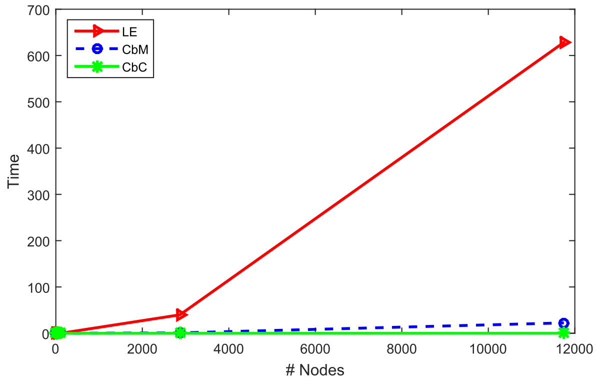


FIGURE 9. **A comparison of the running times of CbM, LE, and CbC.**

and window 10 operating system. On y-axis, we report time expressed in seconds, while in x-axis we report the synthetic and real network nodes. (toy-32, karate-34, dolphin-62, football-115, airport-2868, and internet-11745 nodes). Even though our algorithm is not designed to optimize computational time complexity, it performs much better than LE in running time and also outperforms both of them in identifying influential nodes. Therefore, it is reasonable to apply CbM in large-scale networks to select influential nodes.

V. CONCLUSION

In the large and complex network, identifying the powerful nodes to disseminate information throughout the network is challenging. In this paper, CbM is proposed to measure the importance of a node in large and complex networks. Also, the performance of CbM is compared with traditional measurements (Degree, Betweenness, CbC, PageRank, Eigenvector and LE). Depending on the capability of nodes to spread information in the network; our proposed method identify and rank them better than traditional methods. Nodes selected and ranked by CbM is critical nodes to connect communities. The removal of such kind of nodes from network highly decreases the spread of information in the communities than removing nodes selected by traditional methods. Therefore, nodes selected by CbM are the most intermediate nodes which receive and pass information in the network than other nodes. From the simulation results, we observed that node with high CbM has a greater impact to spread information in the network than a node with high degree, betweenness, CbC, pageRank, eigenvector and LE, i.e., the scale of the network infection is larger if the node with a higher CbM value is taken as the infection source. Our proposed method combines the influence of the degree and the betweenness of the nodes in the network. Finally, CbM outperforms the traditional methods to select influential nodes in the complex network and rank them well by their potential to spread information. In the future, community detection which is compatible with CbM will be conducting.

REFERENCES

[1] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nature Phys.*, vol. 8, pp. 32–39, Dec. 2012.

[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. (Dec. 2011). "Four degrees of separation." [Online]. Available: https://arxiv.org/abs/1111.4570

[3] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[4] M. Kas, C. L. Richard, and K. M. Carley, "Monitoring social centrality for peer-to-peer network protection," *IEEE Commun. Mag.*, vol. 51, no. 12, pp. 155–161, Dec. 2013.

[5] L. C. Freeman, "Centrality in social networks conceptual clarification," *Soc. Netw.*, vol. 1, no. 3, pp. 215–239, 1979.

[6] S. P. Borgatti, "Centrality and network flow," *Soc. Netw.*, vol. 27, no. 1, pp. 55–71, 2005.

[7] N. Salamanos, E. Voudigari, and E. J. Yannakoudakis, "Erratum to: Identifying influential spreaders by graph sampling," in *Proc. 5th Int. Workshop Complex Netw. Their Appl.*, 2017, pp. 111–122.

[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[9] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: The role of clustering," *PLoS One*, vol. 8, no. 10, p. e77455, 2013.

[10] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Soc. Netw.*, vol. 11, no. 1, pp. 1–37, 1989.

[11] L. Fei, H. Mo, and Y. Deng, "A new method to identify influential nodes based on combining of existing centrality measures," *Modern Phys. Lett. B*, vol. 31, no. 26, p. 1750243, 2017.

[12] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Phys. A, Statist. Mech. Appl.*, vol. 404, pp. 47–55, Jun. 2014.

[13] J. Bae and S. Kim, "Identifying and ranking influential spreaders in complex networks by neighborhood coreness," *Phys. A, Statist. Mech. Appl.*, vol. 395, pp. 549–559, Feb. 2014.

[14] J. Ren, C. Wang, H. He, and J. Dong, "Identifying influential nodes in weighted network based on evidence theory and local structure," *Int. J. Innov. Comput. Inf. Control*, vol. 11, no. 5, pp. 1765–1777, 2015.

[15] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Sci. Rep.*, vol. 6, no. 1, p. 27823, 2016.

[16] S. Yeruva, T. Devi, and Y. S. Reddy, "Selection of influential spreaders in complex networks using Pareto Shell decomposition," *Phys. A, Statist. Mech. Appl.*, vol. 452, pp. 133–144, Jun. 2016.

[17] S. Wang, Y. Du, and Y. Deng, "A new measure of identifying influential nodes: Efficiency centrality," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 47, pp. 151–163, Jun. 2017.

[18] Q. Zhang, M. Li, Y. Du, and Y. Deng. (Dec. 2014). "Local structure entropy of complex networks." [Online]. Available: https://arxiv.org/abs/1412.3910

[19] M. Tulu, R. Hou, C. Li, and M. D. Amentie, "Cluster head selection method for content centric mobile social network in 5G," *IET Commun.*, pp. 1–8, Dec. 2017. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-com.2016.1433, doi:10.1049/iet-com.2016.1433.

[20] D.-B. Chen, R. Xiao, A. Zeng, and Y.-C. Zhang, "Path diversity improves the identification of influential spreaders," *Europhys. Lett.*, vol. 104, no. 6, p. 68006, 2014, doi: 10.1209/0295-5075/104/68006.

[21] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.

[22] Q. Hu, Y. Gao, P. Ma, Y. Yin, Y. Zhang, and C. Xing, "A new approach to identify influential spreaders in complex networks," in *Web-Age Information Management* (Lecture Notes in Computer Science), vol. 7923. Berlin, Germany: Springer, 2013, pp. 99–104.

[23] Z. Y. Zhao, H. Yu, Z. L. Zhu, and X. F. Wang, "Identifying influential spreaders based on network community structure," *Chin. J. Comput.*, vol. 37, pp. 753–766, Apr. 2014.

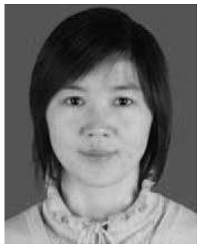
[24] Z. Zhao, X. Wang, W. Zhang, and Z. Zhu, "A community-based approach to identifying influential spreaders," *Entropy*, vol. 17, no. 4, pp. 2228–2252, 2015.

[25] Z. Gao, Y. Shi, and S. Chen, "Measures of node centrality in mobile social networks," *Int. J. Modern Phys. C*, vol. 26, no. 9, p. 1550107, 2015.

[26] D. Wei, X. Deng, X. Zhang, Y. Deng, and S. Mahadevan, "Identifying influential nodes in weighted networks based on evidence theory," *Phys. A, Statist. Mech. Appl.*, vol. 392, no. 10, pp. 2564–2575, 2013.

[27] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[28] *Entropy (Information Theory), Wikipedia, Free Encyclopedia*. Accessed: Oct. 2017. [Onlin]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory)
[29] M. Newman, A.-L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks*. New York, NY, USA: Princeton Univ. Press, 2003.
[30] C. Piccardi, "Finding and testing network communities by lumped Markov chains," *PLoS One*, vol. 6, no. 11, p. e27028, 2011.
[31] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 9, pp. 2658–2663, 2004.
[32] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[33] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.
[34] M. Pósfai *et al.*, "The Barabási-albert model," in *Network Science*, 2015. [Online]. Available: http://barabasi.com/f/622.pdf
[35] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecol. Sociobiol.*, vol. 54, no. 4, pp. 396–405 2003.
[36] Openflights.Org. *Airport, Airline, and Route Data*. Accessed: Jun. 15, 2017. [Online]. Available: http://openflights.org/data.html
[37] M.E.J.Newman. *Network Data*. Accessed: Oct. 12, 2017. [Online]. Available: http://www-personal.umich.edu/~mejn/netdata
[38] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001.
[39] A. Asztalos and Z. Toroczkai, "Network discovery by generalized random walks," *EPL (Europhys. Lett.)*, Aug. 2010. [Online]. Available: https://www.researchgate.net/publication/46581890_Network_Discovery_by_Generalized_Random_Walks, doi: 10.1209/0295-5075/92/50008.

MULUNEH MEKONNEN TULU received the B.Ed. degree in electrical/electronic technology from Adama University, Adama, Ethiopia, and the M.Eng. degree in signal and information processing from the Tianjin University of Technology and Education, Tianjin, China, in 2004 and 2013, respectively. He is currently pursuing the Ph.D. degree in communications and information systems with the State Key Laboratory of ISN, Xidian University, Xi'an, China. From 2004 to 2006, he served as Graduate Assistance I at Maichew Technical College, Tigray, Ethiopia. From 2007 to 2010, he was an Assistance Lecturer at the Asella Technical and Vocational College, Oromia, Ethiopia. Since 2013, he has been a Lecturer with the Department of Electrical and Computer Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia. His current research interests include 5G cellular networks and mobile social network analysis.



RONGHUI HOU received the B.Eng., M.Eng., and Ph.D. degrees in communication engineering from Northwestern Polytechnical University in 2002, 2005, and 2007, respectively. She was a Post-Doctoral Fellow with the Department of Electrical and Electronic Engineering, The University of Hong Kong, from 2007 to 2009. Since 2009, she has been with Xidian University, China, where she is currently a Professor with the Department of Telecommunication Engineering. Her research interests include network quality of service issues, routing algorithm design, and wireless networks.

TALHA YOUNAS received the Bachelor's degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2009 and the M.S. degree in electrical and electronics engineering from the University of Bradford, U.K., in 2011. He is currently pursuing the Ph.D. degree with the Department of Telecommunications Engineering, Xidian University, China with research focused on massive MIMO and cognitive radio networks. He is currently a Lecturer with the COMSATS Institute of Information Technology, Pakistan. His research interest includes wireless communications networks and improvement of bandwidth efficiency in multi-antenna systems.