IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2021, accepted August 26, 2021, date of publication August 30, 2021, date of current version September 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3109124

Leader-Based Community Detection Algorithm in Attributed Networks

DAN-DAN LU
School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China
e-mail: dandanlu@email.cufe.edu.cn

ABSTRACT The community structure plays an indispensable role in developing the deep structure of complex networks. In recent years, some researchers have realized the importance of leader nodes in the community detection process. However, most of the existing leader-based algorithms only use the topological information of networks or attribute information to supplement the topological information, resulting in a significant loss of information integrity. In this paper, we propose a leader-based method that combines topological and attribute information (TALB), uses attribute information among nodes in the network to construct an attribute similarity matrix, and then combines it with network topological information to establish dependency relationships among nodes. As a result, a dependency tree is formed, and the final result of the community division is obtained. Experiments on synthetic networks and real networks show that our proposed method is more effective and practical than the existing leader-based algorithms.

INDEX TERMS Community detection, leader-based, attribute similarity matrix.

I. INTRODUCTION

With the development of complex network theory, attempts have been made to apply this new theory to study various large complex systems in the real world. Community structure [1] has become a trending research topic because it helps better analyze the structure and function of complex networks [2], [3].

Existing algorithms [4]–[8] for community detection based on different objective functions tend to treat every node in the network equally and ignore the existence of leader nodes, which reduces the algorithm accuracy to a certain extent. In recent years, some researchers have recognized the importance of leader nodes in the community detection process [9]–[11]. Leader-based models, which consider a community to be a set of nodes around a leader, have been used to identify community structure. In human societies, for example, each class in a school is regarded as a community, with the monitor of each class being the leader. In the natural world, each wolf pack is regarded as a community, and the alpha wolf is the leader. In the stock market, all the stocks in each plate are regarded as a community, and the most influential stock is the leader that influences the rise of the other stocks [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Yang Tang.

Leader-based algorithms are diverse, and each brings new ideas or improvements to existing algorithms. However, as the network continues to evolve, nodes in the network are associated with data, such as information about individuals (age, occupation, gender, interests, etc.). When identifying communities of nodes, using attribute data helps yield more accurate results. Most algorithms ignore the attributes of the nodes. Several existing algorithms that combine the node attribute information fail because when topological information is missing, only node attribute information is used as a filler. Then, the network topological information cannot be combined at the beginning of the algorithm, making the attribute information useless, which leads to unsatisfactory final community detection results.

To address these problems, this paper proposes a new leader-based algorithm that combines topology and attribute information (TALB) to detect the community structure and identify the community leaders. Our method introduces node attribute information and combines it with network topology information to assess the importance of nodes from a local perspective. The main contributions are as follows:

1) TALB is the first leader-based algorithm to combine network topology information with attribute information in the data preprocessing phase for community detection.
2) A new node-relationship matrix C is proposed. Finding nodes in the same community can be determined

directly based on the positivity or negativity of the elements in this matrix, which greatly simplifies the calculation.

3) Systematic experiments have been carried out in both artificial and real-world networks, and the results show the superiority of the proposed method.

The rest of this paper is as follows: Section 2 briefly introduces some related leader-based algorithms. Section 3 details the notations and symbols used in this paper; Section 4 gives the specific steps of TALB algorithm. Section 5 gives the conclusion.

II. RELATED WORK

In recent years, some researchers have proposed another topology for complex networks, in which each group is made up of a leader and its surrounding followers. Based on this, several leader-based algorithms have emerged. Such algorithms are roughly divided into two steps: the first step is searching for leader nodes in the network; the second step is assigning them to the remaining nodes.

Top leaders [13] was one of the earliest methods of using the leaders in a network to determine the structure of a community. The method is inspired by the k-means algorithm, which randomly selects the leaders in the network and then assigns the remaining nodes. However, the random selection of leaders can cause large fluctuations in the accuracy of the results. Many algorithms have been designed to overcome these disadvantages. For example, LICOD [14] determines leaders in the network based on the importance of the nodes (e.g., centrality metrics), and then assigns the communities to which they belong based on the membership degree of the remaining nodes. HDA [15] helps find pairs of leader nodes in the network, and auto-leader [16] constructs dependency trees based on the leaders.

Although the above methods optimize the leader-based algorithm from different perspectives, their community detection is based only on the topological information of networks, leading to lopsided results. In the era of big data, descriptions are becoming increasingly diverse. In addition to the topological information of the network, the node attribute information is increasingly important for determining to which community the node belongs. In contrast to topology information, node attribute information is focused on the node itself, providing useful information for community exploration.

To improve the effectiveness of community detection, many algorithms [17]–[19] have been derived by combining the node attribute information. For instance, aLBCD [20] uses attribute information to assign a leader to nodes when topological information is missing. However, this method also has limitations: the attribute information is only regarded as the complement to the topological information, and the two are not entirely combined at the beginning of the algorithm. To compensate for these shortcomings, we propose a new leader-based algorithm, TALB, which combines network topological information with attribute information. Furthermore, the validity of this proposed algorithm is verified by our data on both synthetic and real networks.

III. LEADER-FOLLOWER COMMUNITY DETECTION ALGORITHM: TALB

This section will describe the related content of TALB in detail, as shown in Fig (1).

A. PRELIMINARIES

Given an undirected and weightless graph $G = (V, E)$, $V = |n|$ is the set of nodes and $E = |e|$ is the set formed by the edge between any two nodes in the graph. In this paper, we apply the adjacency matrix $A \in \mathbb{R}^{n\times n}$ to represent the topological information of the network. $A_{ij} = 1$ if and only if there exists edge between nodes i and j, otherwise $A_{ij} = 0$. Similarly, the attribute information of nodes in the network also constitutes an attribute information matrix $W \in \mathbb{R}^{n\times m}$, where n is the number of nodes and m is the number of features. In this matrix, W_{ij} is the j-th feature of node i. In simple terms, $W_{ij} = 1$ if node i has attribute j, otherwise $W_{ij} = 0$. To better describe the relationship between points, we give the following definition.

Definition 1: Topological neighbor

For a node $u \in V$ in an undirected and weightless graph $G = (V, E)$, the topological neighbours of it are a set $\Pi(u)$ of nodes that have a link with node u:

$$\Pi(u) = \{u\} \cup \{v \in V | (u, v) \in E\}. \tag{1}$$

Definition 2: Jaccard similarity coefficient [21]

For any two nodes $u, v \in V$ in an undirected and weightless graph $G = (V, E)$, the jaccard similarity coefficient of nodes u and v is:

$$J(u, v) = \frac{\Pi(u) \cap \Pi(v)}{\Pi(u) \cup \Pi(v)}. \tag{2}$$

It can be seen that the higher the value of the Jaccard similarity coefficient, the higher the similarity between the nodes. Further, we can use (2) to obtain the node-relationship matrix $R \in \mathbb{R}^{n\times n}$ from the adjacency matrix A of the network, where R_{ij} takes the value of $J(i, j)$.

Definition 3: Pearson correlation coefficient [22]

The metric is measures the correlation between two variables (the $1 \times m$-dimensional attribute vector corresponding to each node), calculated as follows:

$$P(u, v) = \frac{cov(u, v)}{\delta u \delta v}, \tag{3}$$

where $cov(u, v)$ is the covariance of the two variables and $\delta u(\delta v)$ is the standard deviation of $u(v)$.

Obviously the larger the absolute value of the Pearson correlation coefficient, the stronger the correlation between the nodes. Similarly, we can derive the attribute relationship matrix $S \in \mathbb{R}^{n\times n}$ between nodes from (3), where S_{ij} takes the value of $P(i, j)$.

At this point we have done a preliminary pre-processing of the network: we have constructed the node-relationship

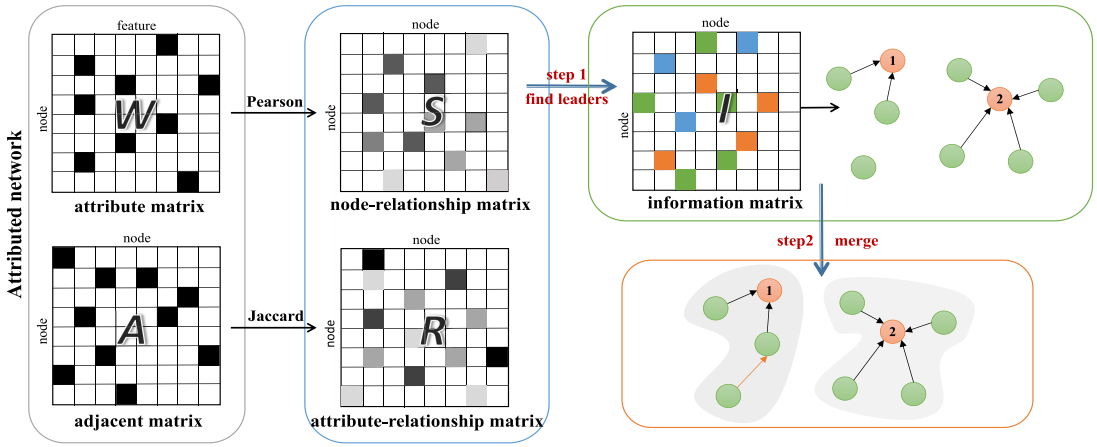


FIGURE 1. The flow chart of the algorithm TALB. The algorithm mainly includes two steps: 1) find the leader to form a local community (In the green box, each node finds its own local leader (nodes 1 and 2)); 2) merge the branches to get the final clustering result (In the orange box, isolated points are dealt with and the final community structure is formed).

matrix R and the attribute relationship matrix S of the network. Obviously both matrices express the relationship between nodes in the network in different ways, and this relationship can also be seen as the edge weights between nodes.

We next give the definition of the information matrix I, which is a combination of the two matrices.

Definition 4: Information matrix

Given an undirected and weightless graph $G = (V, E)$, the topological information as well as the attribute information of the network is combined to obtain a new information matrix $I \in \mathbb{R}^{n\times n}$:

$$I = R + \alpha S, \tag{4}$$

where α is a parameter that controls the weight of attribute information.

Definition 5: Node's leadership

Given an undirected and weightless graph $G = (V, E)$, the leadership of each node $v \in V$ is the sum of the correlations of that node with all its neighbours:

$$L(u) = \sum_{v\in\prod(u)} I(u, v). \tag{5}$$

B. MODEL DESCRIPTION

Based on the above definition, we combine topological information and attribute information of networks to give the new leader-based community detection method TALB. The method is divided into two main steps.

1) DETERMINE THE LEADER OF EACH NODE

Start by traversing each node in the network, calculating the leadership $L(u)$ of each node and comparing with $L(v)$, where $v \in \Pi(u)$. All the nodes that meet the condition $L(v) > L(u)$ are considered in this paper as the set of candidate leaders for node u.

For a node u, there may be several neighbour nodes with leadership $L(\cdot) > L(u)$ at the same time. In order to select the appropriate leader from the leader candidate set, an additional discriminant is selected in terms of whether two nodes belong to the same community. Considering that the Pearson correlation coefficient is negative when the relationship between two variables is negative, this paper uses the information matrix I to further refine the leader candidate set. If $I_{ij} < 0$, the two nodes are considered not to belong to the same community, which actively demonstrates that node i, j cannot act as leader for each other.

In addition to this, we also apply attractive force [16] to further optimise leaders.

Definition 6: Attractive force

Given an undirected and weightless graph $G = (V, E)$, the attractive force of node $v \in \Pi(u)$ to node u is formulated as follows:

$$f(u, v) = \frac{deg(u)}{deg(v)} \cdot \frac{L(u)}{d(u, v)^2}, \tag{6}$$

where $deg(u)$ is the degree of node u, and $d(u, v) = \dfrac{1}{I(u, v)}$.

After calculating the above conditions, we can determine for each node its unique local leader.

Definition 7: Local leader

Given an undirected and weightless graph $G = (V, E)$, the local leader of node $v \in V$ is:

$$L_{loc}(v) = \{u|\arg\max_{u\in N(v)} f(u, v)\}, \tag{7}$$

where $N(v) = \{u|u \in \Pi(v) \cap L(u) > L(v) \cap I(u, v) \geq 0\}$.

Note that after performing the above calculations, it is possible that $L_{loc}(v)$ is an empty set. The particular reason for the circumstance is that the leadership $L(u)$ is a local extreme value.

At the end of the above calculation, we can obtain an initial set of local leaders, and we do not need to know the real number of communities of the network in advance.

2) MERGE SMALL BRANCHES TO GET THE FINAL RESULT

After completing the above steps, we obtain a preliminary clustering result formed by the local information of each node. However, as we have only utilised the local information, there may be some fine branches that have been misclassified due to the ambiguity of the original data, such as isolated points. Therefore, for more accurate results, we have to merge the small communities formed initially around the local leaders (we only consider the relationship between each leader in the merging process).

First, if an outlier exists, we treat it as a follower of the node with the greatest leadership among its neighbouring nodes. If there are multiple neighbours with the greatest leadership, the neighbour with the greatest degree is chosen as its local leader.

Then we use the elements in the information matrix I to determine whether the two leader need to be merged, which in this paper is set to merge if $I(u, v) > 1$ (the threshold value depends on the actual situation).

Up to this point, we can get the final classification result of the communities. The number of communities is the number of the updated leader nodes. Notably we experimented with a number of merging methods in our experiments and found that the effect on the results was not very significant, as we already obtained better results when clustering the nodes using local information.

Algorithm (1) gives the pseudo-code of our algorithm in detail.

C. ALGORITHM COMPLEXITY

TALB has two stages: leader detection, and merging branches. We first look at the running time of each phase and then combine these running times to get the overall running time of TALB.

In the first step, the leadership of each node in the network is calculated, a process that traverses each node in the network with $O(n)$ time complexity. Then, the leadership nodes are determined, which is done by searching all the local top leadership nodes. This can be done by a simple breadth-first search with a complexity of $O(n \log n)$. In summary, the complexity of this step is $O(n \log n + n)$, approximately equal to $O(n \log n)$.

In the second step, the isolated points are first determined and its leaders are reassigned. The complexity of this procedure is $O(N)$, where $N = n - |L_{loc}|$. Finally, merge each leader-based generated small branch with time complexity $O(K \log(K))$, where K is the number of leaders in the network.

Summarizing the above two steps, the total time complexity of our algorithm is $O(n \log n + K \log(K))$.

IV. EXPERIMENTS

In this section, we have conducted systematic experiments in both synthetic and real-world networks respectively to verify the effectiveness and stability of our proposed method, and the experimental results show that TALB is superior to both currently existing leader-based methods.

Algorithm 1 TALB Algorithm

Input: $A \in \mathbb{R}^{n\times n}$, $W \in \mathbb{R}^{n\times m}$, α
Output: Clustering results

```
compute Jaccard similarity coefficient J, and get the node-relationship matrix R;
compute Jaccard correlation coefficient P, and get the attribute relationship matrix S;
for each node v ∈ V do
    compute the Leadership L(v);
end for
for each node v ∈ V do
    for each node u ∈ Π(v) do
        compute the attractive force f(u, v)
    end for
    compute the local leader L_loc(v);
end for
merging leaders
for each node i ∈ V do
    if node i has no followers then
        reassign the community of this node;
    end if
end for
for each node i ∈ L_loc do
    compute the similarity between nodes;
    if similarity > threshold then
        Merge the communities led by the two nodes;
    end if
end for
```

TABLE 1. Parameter settings in the LFR-benchmark network.

Notation	value
-N: number of nodes	1000
-k: average degree	20
-maxk: maximum degree	50
-mu: mixing parameter (the increment is 0.1)	0.1-0.8
-minc: minimum for the community sizes	20
-maxc: maximum for the community sizes	100

A. DESCRIPTION OF EXPERIMENTAL DATA

1) SYNTHETIC NETWORKS

The standard Lancichinetti-Fortunato-Radicchi (LFR) [23] benchmark network model is commonly used to evaluate community detection algorithms as it provides a good approximation to the real network. The parameters we used to generate the network are detailed in Tabel(1). Specifically, the complexity of the network structure will increase as the mixing parameter μ increases.

On the basis of the LFR-benchmark network, we construct for each node its own m-dimensional 0-1 vector of attribute information, where m is the number of attributes.

In this paper, we will give each community 50 corresponding attributes. Apparently, nodes belonging to the same community should have the same attributes with high probability. Further, assuming that there are h communities in a network, there will be $50 \times h$ attributes corresponding to each node. In more details, we use a binomial distribution with probability ρ_{in} to generate the 0-1 attributes of the node which belong to the community, and a binomial distribution with probability ρ_{out} to generate the remaining attributes. In subsequent experiments, we chose three different values for both ρ_{in} and ρ_{out} respectively, where $\rho_{in} = \{0.8, 0.7, 0.6\}$ and $\rho_{out} = \{0.2, 0.3, 0.4\}$. Obviously a larger ρ_{in} (ρ_{out})makes the network structure clearer(more ambiguous).

2) REAL-WORLD NETWORKS

Four common real-world networks have been selected for this paper and are detailed below.

1) Zachary karate network [24]: this is a social network constructed by Wayne Zachary, which in essence consists of the university karate club in Wayne Zachary's country. In the three years between 1970 and 1972, an interesting event took place in the club that led to the network becoming researchable. A difference in philosophy led to a dispute between the club's director and the karate instructor, and the members formed two new internal organisations around the leaders they supported. Let us abstract and analyse this community and we will see that the nodes in the network represent each member, and that when two members become friends (e.g. watching a film together, going to a party) we can establish a link between the corresponding nodes.
2) Football Network [25]: The Football Network is actually a network based on the NCAA's American college football league in 2000. The league divided the 115 teams participating in the tournament into 12 conferences (12 communities) to compete. The nodes in the network represent rugby teams (115 nodes in total), and we construct a link (i.e. an edge) between any two teams whenever they have played a match. There are 616 edges in the network, which means that a total of 616 matches were played during the entire league.
3) The Dolphin Network [26]: a social network that is strangely not made up of humans, but of 62 nodes representing 62 wide-nosed dolphins. As one of the species with a complex social network other than humans, this network of dolphins inhabiting New Zealand's bays is no less complex than a human network. Led by two leaders, the dolphins were divided into two families (the number of communities was 2). When one of the two dolphins communicates with each other beyond a defined closeness value, we establish contact between the corresponding nodes.
4) Polbooks network[1]: the nodes of this network are made up of all the political-related books on Amazon US, and according to the bottom of the page "People who bought this book also bought..." V. Krebs classifies these books into three groups ("liberal", "conservative" and "centrist") according to their classification on Amazon. "centrist").
5) WebKB network[2]: The dataset is a collection of citation networks of four universities (Cornell, Texas, Washington, and Wisconsin). Every university's network is divided into 5 classes, including Course, Student, Faculty, Project and Staff. Each publication has 1703 attributes which are represented by binary vectors, meaning the exist or absence of words.

TABLE 2. Specific parameters of 4 real networks.

Dataset \ Parameters	Vertices	Edges	Communities
Karate	34	78	2
Football	115	613	12
Dolphins	62	159	2
Polbooks	105	441	3
Cornell	195	304	5
Washington	230	446	5
Wisconsin	265	530	5
Texas	187	328	5

The specific parameters for each network are shown in Table (2).

B. BASELINE METHODS AND EVALUATION METRICS

The experiments in this paper use the more commonly used methods known as Topleaders [13], HDA [15] and autoLeader [16] for comparison experiments.

We use both NMI and Kappa metrics to evaluate the performance of these method.

1) Normalized mutual information (NMI) [27]: NMI is often used to measure the similarity of two clustering results. For two different community label A and B, NMI is defined as follows:

$$NMI = \frac{-2\sum_{i=1}^{k}\sum_{j=1}^{k} C_{ij}\log\frac{C_{ij}N}{C_{i.}C_{.j}}}{\sum_{i=1}^{k} C_{i.}\log(\frac{C_{i.}}{N}) + \sum_{i=1}^{k} C_{.j}\log(\frac{C_{j}}{N})} \quad (8)$$

where N is the number of nodes, k is the number of communities, C_{ij} is the number of nodes in community i assigned to community j and $C_{i.}$ is the sum of the i-th row of matrix C, log is the natural logarithm. Note that the value of NMI is between [0, 1].
2) Kappa [28]: Kappa considers the influence of random effects and the calculation formula of kappa is as follows:

$$kappa = \frac{accuracy - expected\ accuracy}{1 - expected\ accuracy} \quad (9)$$

[1]http://www-personal.umich.edu/~mejn/

[2]http://linqs.cs.umd.edu/projects/projects/lbc/

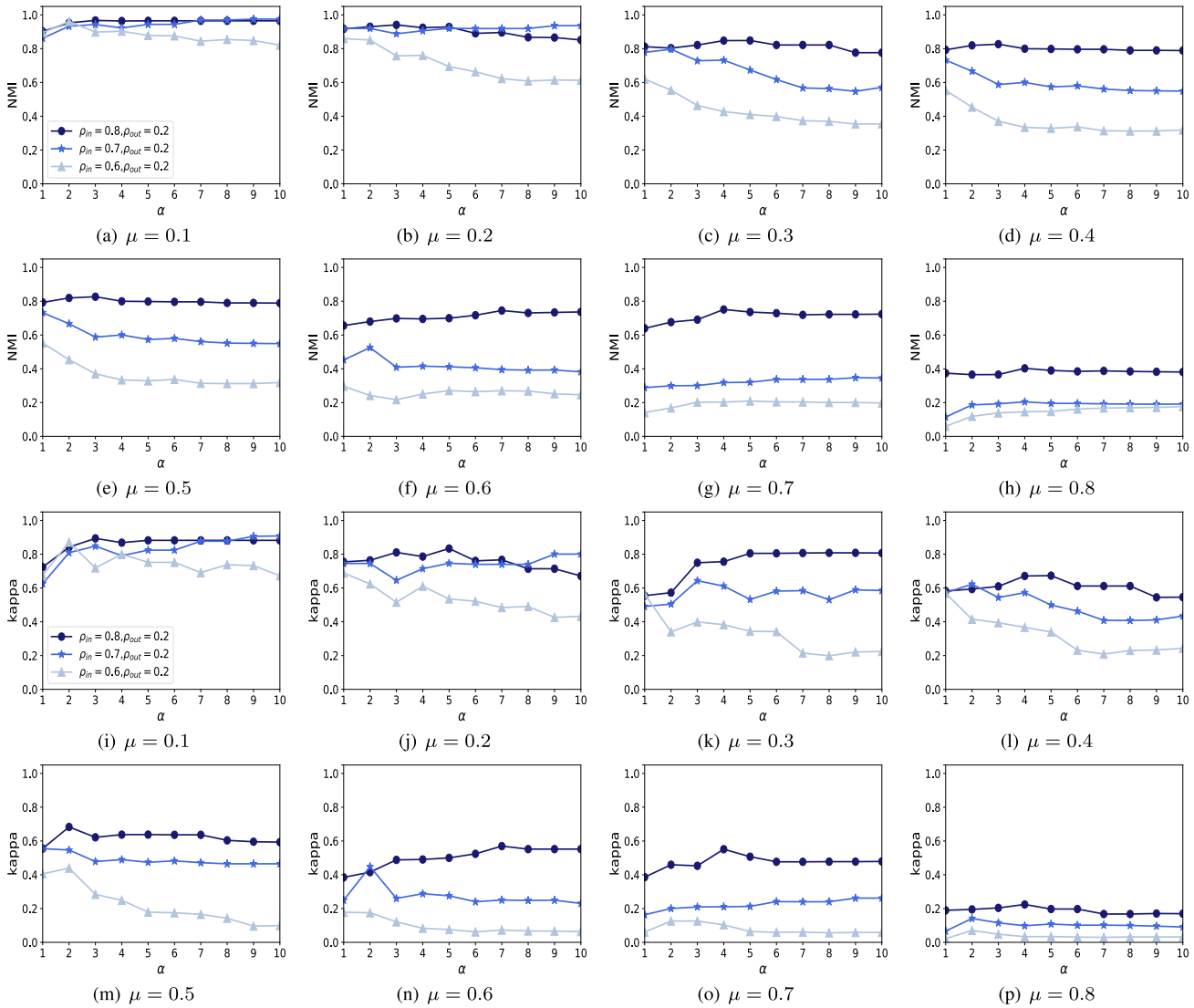


FIGURE 2. The effects by different values of α on TALB in LFR-bnechmark networks. Subplots (a)-(h) show the values of NMI taken for different μ, and (i)-(p) show the values of kappa.

TABLE 3. Confusion matrix of binary classification.

		Actual class	
		Positive	Negative
Predicted class	Positive	TP(True Positive)	FP(False Positive)
	Negative	FN(False Negative)	TN(True Negative)

where $accuracy = \dfrac{TP+TN}{TP+TN+FP+FN}$, which is defined by the confusion matrix showed in Table(3).

3) Purity [29]: purity is the number of correctly assigned nodes divided by the total number of nodes in V. The purity ranges from 0 (completely inconsistent) to 1 (completely consistent). It is calculated using the following formula:

$$purity(M, Q) = \frac{1}{n} \times \Sigma_j \max_i |M_j \bigcap Q_j| \qquad (10)$$

where M and Q are two different partitions of the same data, and V is the set of all nodes.

Based on these three metrics, we conduct the following experiments. The detailed results are as follows.

C. COMMUNITY DETECTION RESULTS

1) SYNTHETIC NETWORKS

First, we conducted experiments on LFR-benchmark networks that were generated by different parameters μ to investigate the effects of different parameters α on the results of TALB, as shown in Fig (2). From these figures, it can be observed that: 1) As the network becomes more complex (i.e., the ρ_{in} decreases and the ρ_{out} increases), the community detection becomes increasingly difficult, thus reducing the TALB's accuracy. 2) Even in highly noisy networks($\rho_{in} = 0.6$, $\rho_{out} = 0.4$), the community detection results of TALB

TABLE 4. The performance of the various methods in different LFR-benchmark networks (bold numbers represent the best results).

Metrics	Methods	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$	$\mu = 0.5$	$\mu = 0.6$	$\mu = 0.7$	$\mu = 0.8$
NMI	Top leaders	0.703	0.668	0.5947	0.4288	0.3631	0.2158	0.1563	0.1096
	HDA	**0.9855**	0.9363	0.8808	0.8046	0.6364	0.3654	0.2211	0.0739
	autoLeader	0.8084	0.6938	0.5913	0.4293	0.3901	0.2757	0.2497	0.1129
	TALB82	0.9751	**0.9559**	**0.9323**	**0.8483**	**0.8267**	**0.745**	**0.7512**	**0.4026**
	TALB73	0.9672	0.9409	0.8019	0.7966	0.7322	0.5254	0.3375	0.2044
	TALB64	0.9612	0.8596	0.6814	0.6198	0.5527	0.2963	0.2092	0.1765
kappa	Top leaders	0.6448	0.6329	0.5766	0.3868	0.3611	0.1858	0.1351	0.0691
	HDA	**0.9337**	0.9091	0.8049	0.7173	0.5412	0.3271	0.1412	0.0444
	autoLeader	0.7846	0.6233	0.5679	0.4042	0.3762	0.2327	0.1995	0.0994
	TALB82	0.9082	**0.9108**	**0.8051**	**0.7456**	**0.6818**	**0.5697**	**0.5513**	**0.2241**
	TALB73	0.8938	0.9006	0.6431	0.6227	0.5547	0.4484	0.2403	0.1412
	TALB64	0.871	0.6887	0.5658	0.5691	0.4392	0.1751	0.1362	0.0704

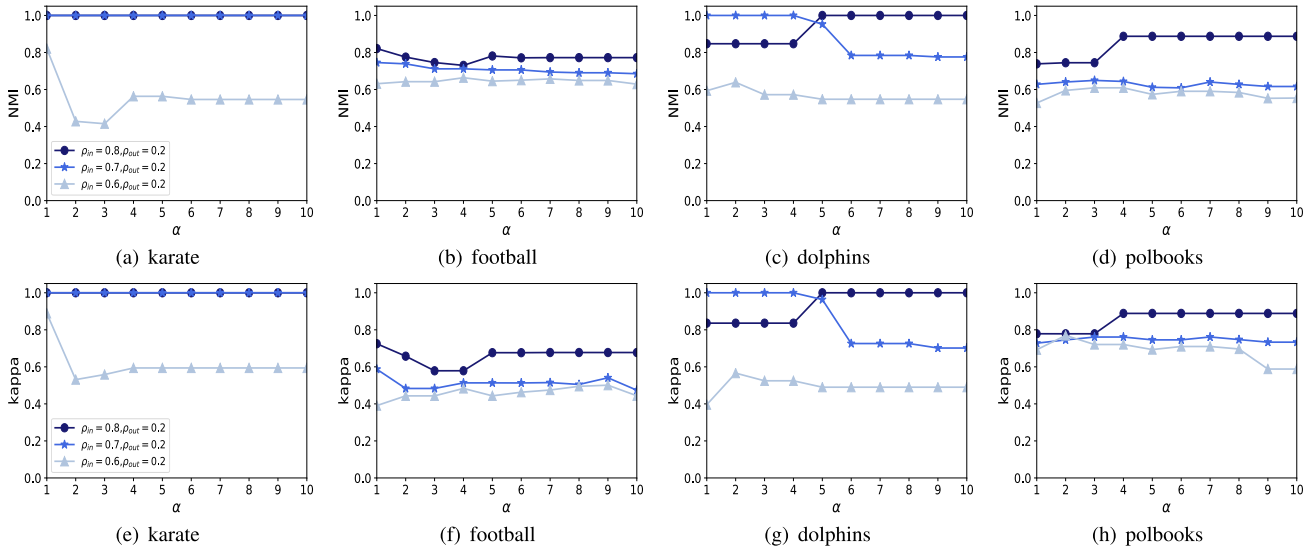


FIGURE 3. The effects by different values of α on TALB in real-world networks. Subplots (a)-(d) show the values of NMI taken for different μ, and (e)-(h) show the values of kappa.

proposed in this paper is acceptable. When the attribute information of the network is clear($\rho_{in} = 0.8$, $\rho_{out} = 0.2$), our algorithm can achieve an accuracy rate of up to 0.95. 3) Experimental results show that different values of the parameter α have little effect on the results in different data sets, which means that special manual adjustment of α will not be necessary for future practical calculations, and the value can be set to be 1.

Next, experiments are conducted to compare the accuracy of TALB with other leader-based models that are currently more widely used, which are shown in Table (4). Note that TALB82,73,64 in the table indicate different attribute information used. For example, TALB82 indicates that a vector of attributes is generated for each node with parameter $\rho_{in} = 0.8$,$\rho_{out} = 0.2$. As can be seen from this table, the proposed TALB algorithm performs almost optimally when the attribute information is clear, and when the attribute information is ambiguous, the results remain to be more accurate than the Top leaders that does not combine the attribute information for community detection. The stability and universality of TALB are also evident, which also shows that using the topological information of the network only for community detection is not sufficient.

2) REAL-WORLD NETWORKS

Similar to the experiments in the LFR-benchmark network, we first experimented with the effect of different values of the parameter α on the results of TALB, see Figure (3). As can be seen from the graph: 1) each data corresponds to a specific value of α for the highest value taken by itself, but the value of α does not have a significant impact on the results of the overall algorithm. Therefore, we set $\alpha = 1$ in the actual calculation when no other particular circumstances occur. 2) Also, the clearer attribute information enhances the community detection results, verifying the critical role

TABLE 5. The performance of the various methods in different real-world networks (bold numbers represent the best results).

Metrics	Methods	karate	football	polbooks	dolphins
NMI	Top leaders	0.8072	0.6914	0.5675	0.3003
	HDA	0.8389	0.6524	0.6577	0.8189
	LICOD	0.5196	0.7583	0.4801	0.5398
	autoLeader	0.8372	0.777	0.6069	0.6863
	TALB82	**1**	**0.8212**	**0.8874**	**1**
	TALB73	**1**	0.7456	0.6492	**1**
	TALB64	0.8209	0.6663	0.6091	0.6387
kappa	Top leaders	0.8412	0.6189	0.7023	0.3426
	HDA	0.8636	0.5262	0.7439	0.9136
	LICOD	0.7099	0.5994	0.6787	0.4816
	autoLeader	0.9412	0.5837	0.7289	0.5611
	TALB82	**1**	**0.7249**	**0.8886**	**1**
	TALB73	**1**	0.5871	0.7614	**1**
	TALB64	0.8889	0.5149	0.7207	0.5661

played by node attribute information. 3) When the attribute information is exceptionally noisy, the community structure contained in the attribute information is obscure, but TALB has an accuracy above 0.4, indicating that TALB can combine network topological information and node attribute information well.

Then, TALB is compared with several different algorithms in real networks, with the specific results shown in Tables (5) and (6). According to the results: 1) When the attribute information is clear, algorithm TALB presents high community detection results, which indicates that the information not only is a complement to network topological information, but also provides useful information orthogonally for the detection. 2) Our proposed method TALB also performs well in extremely ambiguous attribute information, reflecting the robustness of TALB from the side, i.e., the function of our algorithm does not rely exclusively on explicit and clear information from the node attributes. 3) The accuracy of TALB decreases with the blurring of attribute information, which further demonstrates that node attribute information of high quality plays an essential role in community detection.

In addition to this, we compare the computation time of several algorithms. It can be seen from the Fig (4) that although the computation time of Top leader is the shortest, the computation accuracy of this algorithm is not high. And our proposed algorithm TALB has a shorter computation time than the rest of the algorithms with guaranteed accuracy.

D. A CASE STUDY

In this section, we select a real Twitter dataset to analyze the practicality of our proposed method TALB. "Olympics"[3] is a dataset containing 464 users covering the athletes and

[3]http://mlg.ucd.ie/aggregation/index.html

TABLE 6. The performance of the various methods in different real-world networks (bold numbers represent the best results).

Metrics	Methods	cornell	texas	wisconsin	washington
NMI	Top leaders	0.0778	0.1087	0.0688	0.1219
	HDA	0.2234	0.2091	0.1868	0.2612
	LICOD	0.1847	0.1904	0.2031	0.2116
	autoLeader	0.2909	0.304	0.2525	0.1788
	TALB	**0.3293**	**0.3206**	**0.2994**	**0.303**
Purity	Top leaders	0.241	0.4828	0.349	0.5434
	HDA	0.4802	0.5508	0.4792	0.5743
	LICOD	0.3095	0.3372	0.5018	0.4716
	autoLeader	0.5948	0.6833	0.6528	0.6217
	TALB	**0.6256**	**0.7044**	**0.6991**	**0.6913**

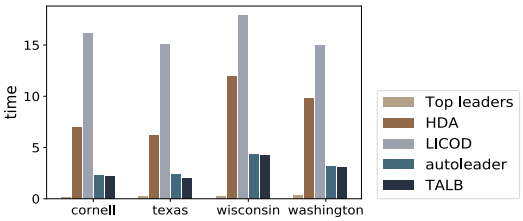
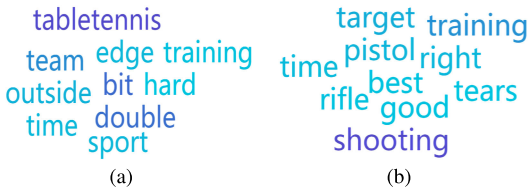


FIGURE 4. Comparison of computing times of several different algorithms.



(a) (b)

FIGURE 5. Word clouds for different communities. The darker the color of each word, the more frequently it appears in the community.

organizations involved in the 2012 Summer Olympics in London. The following relationships between each user form a topological network structure, while the list of tweets posted by each user becomes an attribute of each user with a total of 18,455 attributes.

We selected the first 10 attributes of each community based on the final obtained community structure. The semantics implied by each community can be visualized in Figure (5).

It can be seen that Figure (5)(a) is a community formed by table tennis players. In addition to "tabletennis", which is an obvious word, "edge" and "outside" are also professional words used by table tennis players. Due to the specificity of the data, each club is composed of different athletes, so the word "traing" and "time", which reflects the hard work of the athletes, is present in both communities.

V. CONCLUSION

This paper takes a new perspective on observing complex networks, which suggests that exploring community structure can be done by starting from the leader in the network,

identifying the key nodes in the network, and then specifying the community affiliation of the remaining nodes (followers). However, most of the existing leader-driven methods perform community detection based only on the topological information of the network, which undoubtedly leads to one-sided results and fails to obtain the desired community; while very few leader-based community detection methods only use the attribute information of the nodes as a supplement when dealing with missing values of the network topology or isolated nodes at the end, without combining the two types of information well at the beginning of the algorithm.

Based on these problems, this chapter proposes a leader-driven algorithm TALB that combines network topological information and node attribute information. the algorithm constructs a information matrix I between nodes, and reassigns values to the correlation between nodes in the network. Experiments conducted in synthetic and real networks demonstrate the effectiveness of the algorithm.

This paper focuses on undirected and weightless networks, yet this ideal situation is rarely seen in real networks and the next step seeks to extend TALB to the detection of overlapping communities.

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[2] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 6, no. 3, pp. 115–135, May 2016.
[3] T. Squartini, A. Gabrielli, D. Garlaschelli, T. Gili, A. Bifone, and F. Caccioli, "Complexity in neural and financial systems: From time-series to networks," *Complexity*, vol. 2018, pp. 1–2, Apr. 2018.
[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.
[5] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, Dec. 2004, Art. no. 66111.
[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int. Conf. Knowledg Discovery Data Mining (KDD)*, 1996, pp. 226–231.
[7] D. Huang, J.-H. Lai, and C.-D. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, May 2016.
[8] Y. Li, C. Sha, X. Huang, and Y. Zhang, "Community detection in attributed graphs: An embedding approach," in *Proc. AAAI*, 2018, pp. 338–345.
[9] S. Ahajjam, M. El Haddad, and H. Badir, "A new scalable leader-community detection approach for community detection in social networks," *Social Netw.*, vol. 54, pp. 41–49, Jul. 2018.
[10] N. A. Helal, R. M. Ismail, N. L. Badr, and M. G. M. Mostafa, "Leader-based community detection algorithm for social networks," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery*, vol. 7, no. 6, 2017, Art. no. e1213.
[11] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Discovering leaders from community actions," in *Proc. 17th ACM Conf. Inf. Knowl. Mining (CIKM)*, 2008, pp. 499–508.
[12] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2001, pp. 57–66.
[13] R. R. Khorasgani, J. Chen, and O. R. Zaïane, "Top leaders community detection approach in information networks," in *Proc. 4th SNA-KDD Workshop Social Netw. Mining Anal.*, Jul. 2010, pp. 1–10.
[14] Z. Yakoubi and R. Kanawati, "LICOD: A leader-driven algorithm for community detection in complex networks," *Vietnam J. Comput. Sci.*, vol. 1, no. 4, pp. 241–256, Nov. 2014.
[15] K. Shen, L. Song, X. Yang, and W. Zhang, "A hierarchical diffusion algorithm for community detection in social networks," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Oct. 2010, pp. 276–283.
[16] H. Sun, H. Du, J. Huang, Y. Li, Z. Sun, L. He, X. Jia, and Z. Zhao, "Leader-aware community detection in complex networks," *Knowl. Inf. Syst.*, vol. 62, no. 2, pp. 639–668, Feb. 2020.
[17] K. Steinhaeuser and N. V. Chawla, "Identifying and evaluating community structure in complex networks," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 413–421, Apr. 2010.
[18] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "A model-based approach to attributed graph clustering," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2012, pp. 505–516.
[19] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 718–729, Aug. 2009.
[20] N. A. Helal, R. M. Ismail, N. L. Badr, and M. G. M. Mostafa, "An efficient algorithm for community detection in attributed social networks," in *Proc. 10th Int. Conf. Informat. Syst. (INFOS)*, 2016, pp. 180–184.
[21] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull Soc. Vaudoise Sci. Nat.*, vol. 37, pp. 547–579, 1901.
[22] R. A. Fisher, "Statistical methods, experimental design, and scientific inference," *Biometrics*, vol. 47, no. 3, p. 1206, Sep. 1991.
[23] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, Oct. 2008, Art. no. 46110.
[24] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropolog. Res.*, vol. 33, no. 4, pp. 452–473, Dec. 1977.
[25] 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, 2002.
[26] D. Lusseau and M. E. J. Newman, "Identifying the role that animals play in their social networks," *Proc. Roy. Soc. London. Ser. B, Biol. Sci.*, vol. 271, no. 6, pp. S477–S481, Dec. 2004.
[27] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, 2003.
[28] X. Liu, H.-M. Cheng, and Z.-Y. Zhang, "Evaluation of community detection methods," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1736–1746, Sep. 2020.
[29] R. R. Larson, "Introduction to information retrieval," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 4, pp. 852–853, 2010.

DAN-DAN LU received the master's degree from Shandong University of Finance and Economics, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Statistics and Mathematics, Central University of Finance and Economics, China. Her current research interest includes complex social network analysis.