Atrributed Graph Embedding Based on Multi-objective Evolutionary Algorithm for Overlapping Community Detection

Xiangyi Teng
School of Artificial Intelligence
Xidian University
Xi'an, China
tengxiangyi@stu.xidian.edu.cn

Jing Liu
School of Artificial Intelligence
Xidian University
Xi'an, China
neouma@mail.xidian.edu.cn

Abstract*—Graph embedding methods aim to represent nodes in the network into a low-dimensional and continuous vector space while preserving the topological structure and varieties of relational information maximally. Nowadays the structural connections of networks and the attribute information about each node are more easily available than before. As a result, many community detection algorithms for attributed networks have been proposed. However, the majority of these methods cannot deal with the overlapping community detection problem, which is one of the most significant issues in the real-world complex network study. In addition, it is quite challenging to make full use of both structural and attribute information instead of only focusing on one part. To this end, in this paper we innovatively combine the graph embedding with multi-objective evolutionary algorithms (MOEAs) for overlapping community detection problems in attributed networks. As far as I am concerned, MOEA is first used to integrate with graph embedding methods for overlapping community detection. We term our method as MOEA-GE$_{OV}$, which can automatically determine the number of communities without any prior knowledge and consider topological structure and vertex properties synchronously. In MOEA-GE$_{OV}$, two objective functions concerning community structure and attribute similarity are carefully designed. Moreover, a heuristic initialization method is proposed to get a relatively good initial population. Then a novel encoding and decoding strategy is designed to efficiently represent the overlapping communities and corresponding embedded representation. In the experiments, the performance of MOEA-GE$_{OV}$ is validated on both single and multiple attribute real-world networks. The experimental results of community detection tasks demonstrate our method can effectively obtain overlapping community structures with practical significance.

Keywords—graph embedding, community detection, multiobjective evolutionary algorithm, overlapping community

I. Introduction

Networks naturally exist in many aspects of real-world scenarios and provide a significant tool to characterize and analyze various kinds of modern systems, such as collaboration networks [1], social networks [2] and protein-protein interaction networks [3]. Traditionally, network analysis heavily relies on the representation in the form of adjacent matrix. However, this direct representation often cannot well reflect the underlying organizations and functions of complex systems. Recently, learning embedded representations of graphs becomes a very active area [4-6], since it leads to improved results in data mining and many downstream machine learning tasks such as node classification, link prediction and visualization. For example, DeepWalk [7] employs the truncated random walk sequence to search for the neighbors of a node and feeds to the Skip-Gram model. LINE [6] formally establishes the first- and second-order proximities, and proposes that the second-order proximity is effective to compensate for the data sparsity issue. GraRep [8] considers a special relational matrix and uses the singular value decomposition to reduce the dimensionality of the relational matrix to obtain the k-step network representation. In virtually all cases, most embedding methods can be regarded as seeking to identify the single role of each node in the graphs, which is contradicted to the fact that the very existence of overlap is one of the most important characteristics of complex networks. This observation naturally allows us to realize the close relationship between the research areas of graph embedding and overlapping community detection.

Moreover, with the rapid growth of information available to us, each node in the graph usually has one or more attributes describing their properties. Such graphs can be modelled as attributed networks [9], in which node attributes always play the same important role as the topological structure information. However, most community detection algorithms for attributed networks use either the topological structure or vertex attribute proximity since they overlook the fact that structure and attribute information have mutual influence in the process of clustering.

To deal with the above challenges, in this paper, we propose a novel network embedding method based on multi-objective evolutionary algorithm (MOEA) leveraging both topological structure and attribute information for overlapping community detection in attributed networks (MOEA-GE$_{OV}$). As we all know, structural connections and attributes of nodes are two totally different types of information, and to some

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

extent, they are even contradictory. Multi-objective evolutionary algorithms (MOEAs) [10-12], inspired by the biological mechanism of evolution and heredity, are effective tools to address optimization problems with conflicting objectives. In MOEA-GE$_{OV}$, a modified extended modularity EQ_{OV}, dealing with overlapping community detection problems, is proposed as the first objective to measure the quality of the community structure obtained by our algorithm. To better preserve the attribute similarity between each node, we innovatively define the attribute similarity matrix. Then we carefully design the second objective function maximally preserving the attribute similarity. Moreover, a heuristic initialization method taking advantage of the node attribute information is proposed to get a relatively good initial population. Then a novel encoding and decoding strategy is designed to efficiently represent the overlapping communities and corresponding embedded representation. MOEA-GE$_{OV}$ runs under the framework of NSGA-II and automatically determine the number of communities without any prior knowledge. In the experiments, the performance of MOEA-GE$_{OV}$ is validated on both synthetic and real-world attributed networks. We employ three popular measurements including modularity density D, entropy E, and generalized NMI to evaluate the performance of MOEA-GE$_{OV}$. The experimental results of community detection tasks demonstrate our method can effectively obtain overlapping community structures with practical significance. Our contributions in this paper can be summarized as follows:

- MOEA is first used to integrate with graph embedding methods for overlapping community detection in attributed networks. Two objective functions concerning community structure and attribute similarity synchronously are carefully designed.
- A novel heuristic initialization method taking advantage of node attribute information is proposed to get a relatively good initial population and therefore speed up the whole evolution process. In addition, we propose a two-part encoding and decoding strategy to represent overlapping communities of single individual and corresponding embedded space without the need to assign the number of communities in advance.
- We conduct extensive experiments on both synthetic and real-world attributed networks with single and multiple attribute. The experimental results in terms of three popular evaluation metrics demonstrate that the proposed methods efficiently discover overlapping community structures with practical significance.

The remainder of this paper is organized as follows. Section II introduce the preliminaries to network embedding and overlapping community detection. Then, MOEA-GE$_{OV}$ is described in detail in Section III. In Section IV, we provide the experimental results of both synthetic and real-world attribute networks. Section V concludes the work in this paper.

II. Prelimineries

A. Network Embedding

For an attributed network G, mathematically we can regard it as a 3-tuple (V, E, A), where V= {v_1, v_2, …, v_n} denotes the set of n nodes, and E denotes the set of edges with $E=\{(v_i,v_j)\,|\,v_i\in V, v_i\in V, \text{ and } i\neq j\}$. A = {a_1, a_2, …, a_t} is the set of t attributes of each node. We first review the standard setting of network representation learning. The purpose of network embedding is to learn a mapping $\Phi\colon v\in V\rightarrow R^{n\times d}$, which encodes the latent structural role of a node in the graph. This, in practice, can be achieved by representing the mapping Φ as an $n \times d$ matrix of free parameters that are learned by solving an optimization problem. Perozzi *et al.* [7] first introduced a representation model by optimizing a function of its co-occurrences with other nodes in short truncated random walks.

B. Overlapping Commnity Deteciton

Community detection is a significant domain in network research, whose aim is to partition all nodes in the network into several communities with links being dense within each community but sparse between them. Suppose $\boldsymbol{C}$ = {C_1, C_2, …, C_k} be a set of k communities obtained from G, which meets the following requirements:

$$C_i\subset V,\ i=1,\ 2,\ \dots,\ k \tag{1}$$

$$C_i\neq\varnothing,\ i=1,\ 2,\ \dots,\ k \tag{2}$$

$$\forall i\neq j \text{ and } i,j\in\{1,2,...,k\},\quad C_i\neq C_j \tag{3}$$

$$\bigcup_{\mathrm{i}=1}^{k}C_i=V \tag{4}$$

In order to make communities obtained more practical and meaningful, here each community in $\boldsymbol{C}$ is a proper subset of V. Based on the number of communities a vertex belongs to, we can categorize community detection problems as two kinds. If all communities in $\boldsymbol{C}$ satisfy:

$$\forall i\neq j \text{ and } i,j\in\{\,1,2,\ ...,\ k\},\quad C_i\bigcap C_j=\varnothing \tag{5}$$

that is, each node only belongs to one community. Thus, we call it separated community detection problems. Otherwise, if there exists the following situation:

$$\exists i\neq j \text{ and } i,j\in\{1,2,...,k\},\quad C_i\bigcap C_j\neq\varnothing \tag{6}$$

then there must be at least one node belonging to more than one community and we call it overlapping community detection problems. In this paper, our concentration focuses basically on how to discover overlapping community structures in attributed networks.

III. THE PROPOSED ALGORITHM MOEA-GE$_{OV}$

A. Two Objectives of MOEA-GE$_{OV}$

In the field of community detection research, one of the most well-known functions to evaluate the quality of clusters is modularity Q [13]. Formally, modularity Q is expressed as:

$$Q = \frac{1}{2m}\sum_{q=1}^{k}\sum_{i\in C_q, j\in C_q}\left(\boldsymbol{A}_{ij} - \frac{d_i d_j}{2m}\right) \tag{7}$$

where k denotes the number of communities, $\boldsymbol{A}$ denotes the adjacency matrix of the original network, m represents the total number of edges, and d_i and d_j denote the degrees of nodes v_i and v_j, respectively. However, it cannot evaluate the overlapping communities effectively and precisely. Later, a new version of modularity, EQ_{ov}, was proposed in [14] to deal with this problem:

$$EQ_{OV} = \frac{1}{2m}\sum_{q=1}^{k}\sum_{i\in C_q, j\in C_q}\frac{1}{O_i O_j}\left(\boldsymbol{A}_{ij} - \frac{d_i d_j}{2m}\right) \tag{8}$$

where O_i and O_j denote the number of communities that nodes v_i and v_j belong to. We apply EQ_{ov} as our first objective to be maximized. The higher the value EQ_{ov} is, the more distinct the overlapping community structures are.

Traditionally, network analysis heavily relies on the representation in the form of adjacency matrix $\boldsymbol{A}$. However, it cannot well reflect the underlying properties of nodes in the graph. In order to better preserve the attribute similarity information in the embedding space, we first make a definition of the attribute similarity matrix S^A as follows:

Definition 1. (*Attribute Similarity Matrix*) The attribute similarity matrix S^A is the initial source to record the attribute similarity between two nodes. For single attribute networks, the element (i, j), which is the ith row and jth column in S^A is set 1 if node v_i and node v_j have the same attribute and 0 otherwise. For multi-attribute networks, the element (i, j) in S^A is defined as the cosine similarity of the attribute vectors of node v_i and node v_j.

For a single attribute b, S^A_{ij} is easy to be defined as

$$S^A_{ij} = \begin{cases} 1 & if\ b_i = b_j \\ 0 & otherwise \end{cases} \tag{9}$$

For multi-attribute networks, S^A_{ij} is defined as

$$S^A_{ij} = \frac{\vec{i}\cdot\vec{j}}{|\vec{i}|\times|\vec{j}|} \tag{10}$$

where $\vec{i}$ and $\vec{j}$ are sets of attributes of two nodes. Attribute similarity matrix S^A can precisely represent the attribute similarity of two nodes in the network. In the next part, a novel heuristic initialization method taking advantage of S^A is proposed to get a relatively good initial population.

Next, to calculate the attribute similarity between nodes in the embedding space, we define the attribute similarity matrix in the embedding space, S^{NE}, as follows:

$$S^{NE}_{ij} = \frac{\vec{I_i}\cdot\vec{I_j}}{|\vec{I_i}|\times|\vec{I_j}|}\cdot\boldsymbol{\delta}_{ij} \tag{11}$$

$$\boldsymbol{\delta}_{ij} = \begin{cases} 1 & v_i \text{ and } v_j \text{ are in the same community} \vee S^A_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where I_i and I_j are the learned representation for nodes v_i and v_j.

In the embedding space, inspired by the intuitive idea from [15], we propose our second objective in (13) to measure the attribute similarity of the obtained communities which is also to be maximized:

$$similarity(S^A, S^{NE}) = \frac{S^A_{t=1}\cdot S^{NE}_{t=1}}{|S^A_{t=1}|\times|S^{NE}_{t=1}|} \tag{13}$$

where $S^A_{t=1}$ ($S^{NE}_{t=1}$) means to connect each line of S^A (S^{NE}) end to end and expand into a one-dimensional vector.

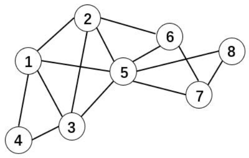
B. Representation and Intialization

The representation of EAs makes a great influence on the evolutionary process. In our algorithm, each individual contains two part, the embedding representations of nodes $\boldsymbol{I} = \{I_1, I_2,\ldots, I_n\} \in R^{n\times d}$ and community labels of all nodes $\boldsymbol{X} = \{X_1, X_2,\ldots, X_n\}$, where d is the embedding dimension and X_i denotes the community labels of node v_i. There are usually two efficient direct representations, the label-based [16] and locus-based representations [17]. For the label-based representation, suppose there is totally k communities, then every gene x_i randomly takes a number from integer set $\{1, 2, \ldots, k\}$ as the label representing the cluster it participates in. While in the locus-based representation, each gene X_i can have an allele value j in the set $\{1, 2, \ldots, n\}$. A value j assigned to X_i indicates there is a connection between nodes v_i and v_j and they are exactly in the same cluster.

In MOEA-GE$_{OV}$, we use hybrid representation to make full use of their advantages. First, locus-based representation is employed in the initialization process. In order to get a relatively good initial population, a heuristic method taking advantage of the node attribute information is proposed efficiently. Specifically, we employ the roulette selection based on attribute similarity to choose a relatively suitable allele value for each node. Fig. 1 gives an illustrative example of the heuristic method for initialization. Let us take node 2 for example. From the attribute similarity matrix obtained, we can see there are totally five nodes, namely nodes 1, 3, 4, 5 and 8, that have the same attribute value for at least one dimension. Therefore, taking the roulette selection, the possibilities of being chosen for these five nodes are 0.12, 0.37, 0.07, 0.23 and 0.21, respectively. Then we transform them into label-based representation since it is much easier for evolutionary operators to employ. Finally, we perform a random initialization on $\boldsymbol{I}$ to let nodes with the same community label be closer in the embedding space as follows. The value of each element in $\boldsymbol{I}$ is randomly initialized between 0 and 1, and then is updated according to (14).

$$I_i = I_i + \beta \cdot (I_{\max} - I_i) \tag{14}$$

where I_{max} is the representation of the node with maximum degree in the community that v_i belongs to and β is positive parameter controlling the extent of move in the embedding space. Here we set $\beta = 0.6$ in the study.



G=(V,E,A)

node								
1	0	0.31	0	0	0.91	0.66	0	0.31
2	0.31	0	0.94	0.16	0.59	0	0	0.53
3	0	0.94	0	0.37	0	0	0.16	0
4	0	0.16	0.37	0	0	0	0.22	0
5	0.91	0.59	0	0	0	0.66	0	0.78
6	0.66	0	0	0	0.66	0	0	0.22
7	0	0	0.16	0.22	0	0	0	0
8	0.31	0.53	0	0	0.78	0.22	0	0

Attribute Similarity Matrix of G (t=32)

Fig. 1. An illustrative example of the heuristic method for initialization.

C. Indirect Decoding Strategy

Direct representations alone cannot achieve the goal of encoding the division of a network. Usually, we require a decoder to help with the whole representation process. In [18], a heuristic search, which is a process of optimizing the community fitness function proposed in [19] to represent overlapping communities has shown effective performance. Suppose C_i be a community, and the community fitness function of C_i can be defined as

$$F(C_i) = \frac{k_{in}^{C_i}}{(k_{in}^{C_i} + k_{out}^{C_i})^{\alpha}} \tag{15}$$

where $k_{in}^{C_i}$ and $k_{out}^{C_i}$ denote the internal and external degrees of all nodes in C_i, respectively, and α here is a resolution parameter. For the sake of simplicity, we set α=1 in this study. Fig. 2 gives an illustrative example of the decoding process and the corresponding embedded representation. All the qualified nodes will participate into multiple communities if the value of the corresponding fitness functions increases. As a result, the position of each qualified node in the embedding space needs to be shifted slightly to the center of new community it belongs to. Mathematically, the embedding representation $\boldsymbol{I}$ is updated as follows:

$$I_i = I_i + \eta \cdot (I_{\text{center}} - I_i) \tag{16}$$

where I_{center} is the representation of the node nearest to the center of new community and η is a positive parameter controlling the shift range. Here we set η as 0.4 in this study.

D. Evolutionary Operators

For crossover operator, we apply the classic simulated binary crossover (SBX) operator [20] in our algorithm. SBX is especially useful in problems with multiple optimal solutions and the lower and upper bounds of the search space are unknown in advance. The detailed parameter setting about SBX can be referred to [15].

For the mutation operator, traditional operators such as random mutation, Gaussian mutation and polynomial mutation operators are all suitable for this problem. Here in this paper, for the sake of simplicity and low computational cost, we choose the random mutation operator.

Algorithm 1: MOEA-GE$_{OV}$

```
Input:
    Popsize: number of individuals in a single population;
    $G_{max}$: maximum number of generations;
    $p_c$: crossover probability;
    $p_m$: mutation probability;
    $d$: dimension of the embedding space
    $G = (V,E,A)$: attribute network;
Output:
    A set containing all the non-dominated solutions;

Series of Symbols:
  $P_t$: parent population containing popsize individuals from the tth generation;
  $Q_t$: offspring population containing popsize individuals from the tth generation;
  $U_t$: combination of parents and offspring populations containing 2× Popsize individuals from the tth generation;
  $F$: a non-dominated set that consists of a variety of rank levels;

$t \leftarrow 1$;
Initialize $P_t$ using the locus-based representation with a heuristic method, then transform $P_t$ into the label-based representation and perform a random initialization on the embedding space;
For each individual in $P_t$, use the decoding strategy to obtain a set of overlapping communities and the corresponding embedded representations, then calculate the two objective functions;
while ($t < G_{max}$) do
    $Q_t \leftarrow \varnothing$;
    while ($|Q_t| <$ Popsize) do
        $[X_1, X_2] \leftarrow$ binary-tournament-selection ($P_t$);
        $[X_{1'}, X_{2'}] \leftarrow$SBX crossover ($X_1, X_2, p_c$);
        $[X_{1''}, X_{2''}] \leftarrow$random mutation ($X_{1'}, X_{2'}, p_m$);
        $Q_t \leftarrow Q_t \cup [X_{1''}, X_{2''}]$;
    end while
    For each individual in $Q_t$, use the decoding strategy to obtain a set of overlapping communities and the corresponding embedded representations, then calculate the two objective functions;
    $U_t \leftarrow P_t \cup Q_t$;
    $F \leftarrow$ fast-nondominated-sort ($U_t$);
    $P_{t+1} \leftarrow \varnothing$;
    $i \leftarrow 1$;
    while ($|P_{t+1}| + |F_i| <$ Popsize) do
        $P_{t+1} \leftarrow P_{t+1} \cup F_i$;
        $i \leftarrow i + 1$;
    end while
    Calculate the crowding distance of each individual in $F_i$;
    Sort $F_i$ in descending order according to the crowding distance;
    $P_{t+1} \leftarrow P_{t+1} \cup F_i$ [1:( Popsize - $|P_{t+1}|$)];
    $t \leftarrow t + 1$;
end while
```

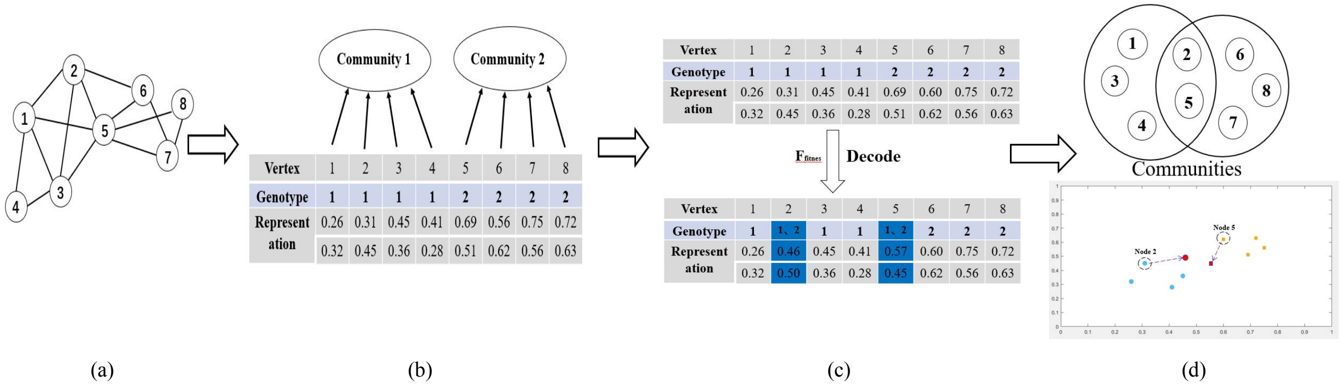


Fig. 2. Illustrative examples of the decoding process (schema). (a) Original graph. (b) One possible genotype and the corresponding embedded representation after the initialization (each vertex belongs to only one community now). (c) One possible genotype and the corresponding embedded representation after the decoding process (here community fitness function is the criterion for each vertex to join other communities). (d) Obtained communities and the shift of the overlapping nodes in the embedding space according to the genotype after the decoding process.

E. Implementation of MOEA-GE$_{OV}$

NSGA-II [10] has repeatedly been proved to be an efficient EA for multiobjective optimization problems. This algorithm generates nominated sets with different levels and provides evenly distributed solutions. Therefore, MOEA-GE$_{OV}$ is implemented under the framework of NSGA-II. Algorithm 1 summarizes the details of MOEA-GE$_{OV}$. We employ a trial-and-error procedure and then select the parameter values giving good results. Thus, the size of population is set as 100, the maximum number of generations 50, the crossover rate 0.9 and the mutation rate 0.1. For single attributed networks, the dimension of embedding space *d* is 2 and for the others is 16. All experiments are conducted under the same parameter setting.

IV. EXPERIMENTS

A. Experimental Setup

We use both synthetic benchmark networks and real-world attributed networks in our experiments. The synthetic network is generated by the LFR benchmark [21], and real-world networks include one single attribute network and three multi-attribute networks. The detailed descriptions about these networks are given as follows.

1) Political Books Network [22]: This network contains political American books published around 2004 national presidential election. All books are purchased from the website Amazon.com. If there is a link between two books, it means customers frequently buy these books all together. There is a single attribute for each book showing their political tendencies for three choices: conservative, liberal, or neutral.

2) LFR benchmark network [21]: The parameters of the LFR benchmark network used in the experiment are set as follows: number of nodes n=3000, average degree $\overline{d}$ =5, maximum degree *maxd*=20, mixing parameter μ=0.1 , which denotes the fraction of edges of a node linking to other communities, power-law degree distribution with exponent t_1=2, power-law community size distribution with exponent t_2=1, minimum for the community size *minc*=20, maximum for the community size *maxc*=60, number of overlapping nodes O_n=30% of the total number of nodes, and number of memberships of the overlapping nodes O_{mem}=2.

3) Ego Facebook Networks with Multi-attribute: These datasets are obtained from ten ego-networks, consisting of 193 circles and 4039 users [23]. The Ego-network consists of a set of nodes which are directly connected to the ego. All the 10 ego-networks have multi-attribute and we take three of them in our experiment. The detailed statistic information about three networks are shown in Table I.

TABLE I

STATISTICAL INFORMATION ABOUT EGO FACEBOOK NETWORKS WITH MULTIPLE ATTRIBUTES

Network	Nodes	Edges	Attribute dimension	<*k*>
Ego_0	347	2519	224	14.5
Ego_686	170	1656	63	19.5
Ego_3980	58	143	42	4.9

B. Evaluation Measurements

Two most commonly used evaluation metrics, namely, density *D* and entropy *E*, are employed to evaluate the performance of MOEA-GE$_{OV}$.

$$D=\sum_{q=1}^{k}\frac{m_q}{m} \tag{17}$$

where m_q denotes how many links in community C_q, m represents the total number of edges, and k stands for the number of clusters. The density *D*, to some extent, is the reflection of the proportion of community intra-links over the total number of edges in the network. The higher the density *D* is, the more evident the obtained community structures is.

Let the set of values of attribute *b* be {1, 2, …, l}, and entropy *E* is expressed as

$$E=\sum_{q=1}^{k}\frac{r_q}{n}\cdot entropy(q) \tag{18}$$

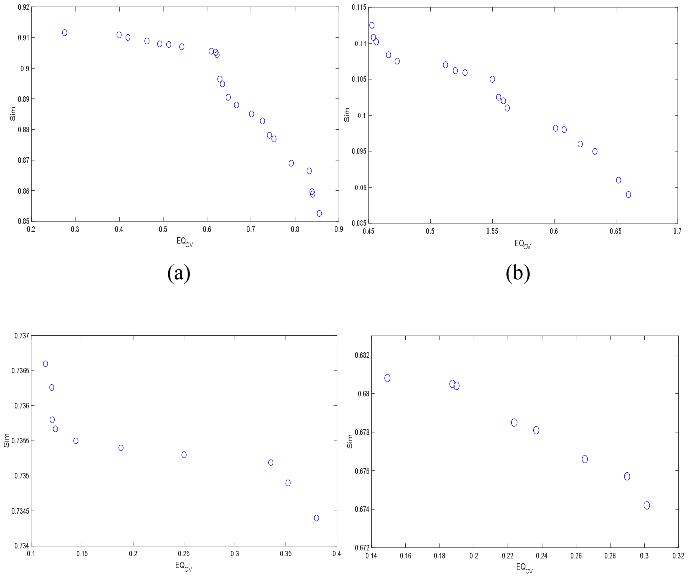
$$entropy(q) = -\sum_{i} p_{iq} \cdot \log(p_{iq}) \tag{19}$$

where r_q denotes how many nodes in community C_q, n denotes the number of nodes in graph, and p_{iq} is the percentage of nodes in community C_q whose attribute values are i. The obtained clusters with smaller E shows that the nodes inside clusters are more homogenous.

To further verify the performance of MOEA-GE$_{\text{OV}}$, generalized normalized mutual information (gNMI) [24] is used to estimate the similarity between the network with true community structure and a detected one. gNMI (A, B) is expressed as:

$$\text{gNMI}(A,B) = \frac{-2\sum_{i=1}^{k_A}\sum_{j=1}^{k_B} R_{ij}\log(R_{ij}\cdot n / R_{i.}R_{.j})}{\sum_{i=1}^{k_A} R_{i.}\log(R_{i.}/n) + \sum_{j=1}^{k_B} R_{.j}\log(R_{.j}/n)} \tag{20}$$

where n represents the total number of nodes in the network, R denotes the confusion matrix of which element R_{ij} records the number of nodes in community C_i from division A that are also in community C_j from division B. k_A (k_B) denotes the number of communities in division A(B), $R_{i.}$ ($R_{.j}$) stands for the sum of the elements from row i (column j).



(a) (b)

(c) (d)

Fig. 3. PFs obtained by MOEA-GE$_{\text{OV}}$ for both synthetic and real-world attributed networks. (a) Political Books. (b) LFR network. (c) Ego_686. (d) Ego_3980.

C. Pareto Fronts obtained by MOEA-GE$_{OV}$

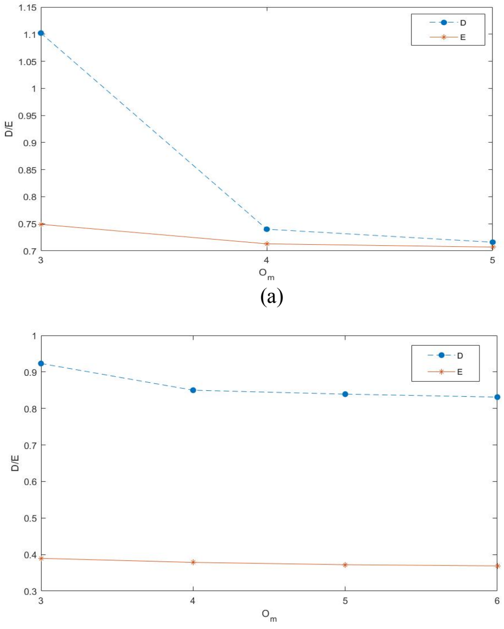
Since MOEA-GE$_{\text{OV}}$ is a multiobjective evolutionary algorithm, the results obtained are a set of nondominated solutions. To verify the good performance of MOEA-GE$_{\text{OV}}$, we plot the final PFs for four networks obtained by our method, which are shown in Fig. 3. As can be seen, MOEA-GE$_{\text{OV}}$ clearly obtains lots of nondominated solutions and evenly distributed PFs.

From Fig. 3 (a) and (b), we can obviously observe that nearly all the nondominated solutions obtained in single attributed networks have EQ_{OV} greater than 0.4. In two multi-attribute networks, all values of $Similarity(S^A, S^{NE})$ are greater than 0.67, which means the communities obtained are quite homogenous.

D. Experimental Results in term of D and E

In this subsection, we first study the relationship between D and E as the number of communities varying. Then we compare our method with two popular overlapping community detection algorithms Cfinder [25] and SLPA [26] in terms of D and E. CFinder is an implementation of clique percolation methods, and SLPA is a label propagation algorithm.

In the Political Books network and Ego_3980 multi-attribute network, the relationship between D/E (average value of the final generation) and the number of overlapping communities (O$_m$) are illustrated in Fig. 4. From Fig. 4, it is interesting to observe that Density D and Entropy E have the same variation trends when the number of detected communities increases. Since the larger value of D reflects more obvious community structure and the smaller value of E means vertices inside clusters tend to be more homogenous, the same variation trend means the relationship between D and E is conflicting. To some great extent, it also proves that topological structure and node attribute are complementary information for community detection.



(a)

(b)

Fig. 4. Variation tendencies of D and E as O$_{\text{m}}$ increasing for two real-world networks. (a) Political Books, (b) Ego_3980.

Next, to further verify the performance of MOEA-GE$_{\text{OV}}$, we also compare MOEA-GE$_{\text{OV}}$ with CFinder and SLPA in terms of D and E. We run the experiments 10 times and records the result with the best setting for algorithms. Specifically, in CFinder, k varies from 3 to 6. In SLPA, parameter r varies from 0.01 to 0.5 in the step of 0.05. Since

MOEA-GE$_{\mathrm{OV}}$ attains many nondominated solutions in a single run, we select the solutions from PFs with the closest values of D or E compared with that of the two other algorithms. Table II shows the detailed experimental results.

As we can see from Table II, MOEA-GE$_{\mathrm{OV}}$ achieves the best results in terms of D and E. Especially, MOEA-GE$_{\mathrm{OV}}$ get the highest value of E in almost all networks except for the LFR network. For the political books network Ego_686 and Ego_3980, MOEA-GE$_{\mathrm{OV}}$ performs better in terms of both D and E.

Table II
COMPARISION RESULTS IN TERMS OF D AND E ON ATTRIBUTED NETWORKS

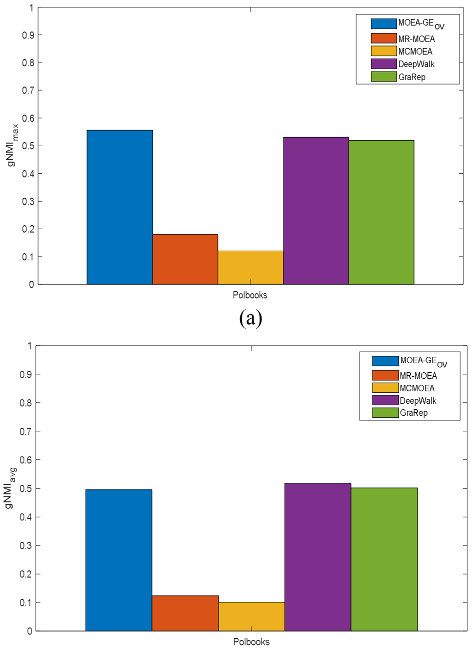
Network	Metric	MOEA-GE$_{\mathrm{OV}}$	CFinder	*SLPA*
Political books	D E	**1.1021** **0.7498**	0.9932 0.7733	0.9728 0.9320
LFR	D E	**1.2302** 2.2653	1.0596 2.9673	0.7958 **1.9583**
Ego_0	D E	0.6743 **0.2937**	**0.8033** 1.1920	0.6167 1.6812
Ego_686	D E	**0.9565** **0.3261**	0.5681 0.7645	0.8587 1.1534
Ego_3980	D E	**0.9231** **0.3896**	0.6489 0.7931	0.8358 1.6485

E. Experimental Results in term of gNMI

To further validate the performance of MOEA-GE$_{\mathrm{OV}}$, we use gNMI as an evaluation metric. gNMI originates from the information science and is widely used in network analysis. However, gNMI can only be applied to networks with ground truth community structure, here we take political books network as an example in this part.

For the comparison algorithms, two recently proposed MOEA-based overlapping community detection algorithms, namely, MR-MOEA [27] and MCMOEA [28], together with two famous graph embedding algorithms GraRep and DeepWalk are included. For a fair comparison, two MOEA baseline approaches use the same size and maximum number of generations as MOEA-GE$_{\mathrm{OV}}$. The tunable parameters of each algorithm are set as the suggestion of the corresponding paper. We repeat the process for 10 times and record the maximum gNMI (denoted as gNMI$_{\mathrm{max}}$) as well as the average gNMI (denoted as gNMI$_{\mathrm{avg}}$) of each algorithm. Fig. 5 gives the detailed comparison.

From Fig. 5, we can clearly see that MOEA-GE$_{\mathrm{OV}}$ obtains the highest values of maximum gNMI (NMI$_{\mathrm{max}}$) on the Political Books network. With regard to average gNMI (NMI$_{\mathrm{avg}}$), our method performs slightly worse than DeepWalk but better than MR-MOEA and MCMOEA. The reason why this phenomenon occurs is probably that MOEA-GE$_{\mathrm{OV}}$ tends to find more communities sometimes and the values of dominated solutions in the PFs are in a wide range.



(a)

(b)

Fig. 5. Detailed comparison in terms of (a) maximum gNMI and (b) average gNMI on Political Books.

V. CONCLUSION

In this paper, we propose a novel network embedding method based on multi-objective evolutionary algorithm considering topological structure and attribute information simultaneously for overlapping community detection in attributed networks (MOEA-GE$_{\mathrm{OV}}$). To better preserve the attribute similarity between each node, we innovatively define the attribute similarity matrix and carefully design an objective function based on attribute similarity. Moreover, a heuristic initialization method taking advantage of the node attribute information is proposed to get a relatively good initial population. Then a novel encoding and decoding strategy is designed to efficiently represent the overlapping communities and corresponding embedded representation. MOEA-GE$_{\mathrm{OV}}$ can automatically determine the number of communities without any prior knowledge. The experimental results on both synthetic and real-world networks show the effectiveness of MOEA-GE$_{\mathrm{OV}}$ compared with several existing methods. In the future, our top priority will be put on implementing the algorithm in the parallel pattern and use the surrogate model to reduce the computational time.

ACKNOWLEDGEMENTS

This work was supported in part by the Key Project of Science and Technology Innovation 2030 supported by the Ministry of Science and Technology of China under Grant 2018AAA0101302 and in part by the General Program of

National Natural Science Foundation of China (NSFC) under Grant 61773300.

REFERENCES

[1] M. E. J. Newman, “The structure of scientific collaboration networks,” *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.

[2] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, “Community detection in social media,” *Data Min. Knowl. Disc.*, vol. 24, no. 3, pp. 515–554, 2012.

[3] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Pro. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821-7826, 2002.

[4] A. Grover and J. Leskovec, “Node2vec: Scalable feature learning for networks,” In *Proceedings of the 22nd ACM SIGKDD International Conference oKnowledge Discovery and Data Mining*, pp. 855-864, 2016.

[5] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, 290(5500): 2323-2326, 2000.

[6] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “Line: Large-scale information network embedding,” *In Proceedings of the 24th International Conference on World Wide Web*, pp. 1067-1077, 2015.

[7] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.701-710, 2014.

[8] S. Cao, W. Lu, and Q. Xu, “GraRep: Learning graph representations with global structural information,” In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 891-900, 2015.

[9] Z. Li, J. Liu and K. Wu, “A multiobjective evolutionary algorithm based on structural and attribute similarities for community detection in attributed networks,” *IEEE Trans. on Cybernetics*, vol. 48, no. 7, pp. 1963-1976, 2018.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[11] X. Teng, J. Liu and M. Li, “Overlapping community detection in directed and undirected attributed networks using a multiobjective evolutionary algorithm,” *IEEE Trans. on Cybernetics,* in press, 2019.

[12] N. Srinivas and K. Deb, “Multiobjective optimization using nondominated sorting in genetic algorithms,” *Evol. Comput.*, vol. 2, no. 3, pp. 221-248, 1994.

[13] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Art. no. 026113, 2004.

[14] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, “Detect overlapping and hierarchical community structure in networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, 2009.

[15] M. Li, J. Liu, P. Wu and X. Teng, “Evolutionary network embedding preserving both local proximity and community structure,” *IEEE Trans. Evol. Comput.*, in press, 2019.

[16] M. Tasgin, A. Herdagdelen, and H. Bingol, “Community detection in complex networks using genetic algorithms,” *CoRR*, vol. 2005, no. 3120, pp. 1067–1068, 2006.

[17] Y. J. Park and M. S. Song, “A genetic algorithm for clustering problems,” in *Proc. 3rd Annu. Conf. Genet. Program.*, pp. 568–575, 1998.

[18] J. Liu, W. Zhong, Hussein A. Abbass, and David G. Green, “Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms,” In *Proceedings of the IEEE Congress on Evolutionary Computation*, *Barcelona*, *Spain*, 2010.

[19] A. Lancichinetti, S. Fortunato, and J. Kertész, “Detecting the overlapping and hierarchical community structure of complex networks,” *New Journal of Physics*, vol. 11, no. 3, Art. no. 033015, 2009.

[20] R. B. Agrawal, K. Deb, and R. B. Agrawal, “Simulated binary crossover for continuous search space,” *Complex Systems*, 9(2): 115-148, 1995.

[21] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E*, vol. 78, no. 4, Art. no. 046110, 2008.

[22] V. Krebs. (2004). *Books About US Politics*. [Online]. Available: http://www.orgnet.com

[23] J. J. McAuley and J. Leskovec, “Learning to discover social circles in ego networks,” in *Proc. NIPS*, pp. 548–556, 2012.

[24] A. Lancichinetti, S. Fortunato, and J. Kertész, “Detecting the overlapping and hierarchical community structure of complex networks,” *New Journal of Physics*, vol. 11, no. 3, Art. no. 033015, 2009.

[25] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[26] J. Xie, B. K. Szymanski, and X. Liu, “Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process,” In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference*, pp. 344-349, 2011.

[27] L. Zhang, H. Pan, Y. Su, X. Zhang, and Y. Niu, “A mixed representation-based multiobjective evolutionary algorithm for overlapping community detection,” *IEEE Trans. on Cybernetics,* vol.47, no. 9, pp. 2703-2716, 2017.

[28] X. Wen, W. Chen, Y. Lin, T. Gu, and J. Zhang, “A maximal clique based multiobjective evolutionary algorithm for overlapping community detection,” *IEEE Trans. Evol. Comput.*, vol. 21, no. 3, pp. 363–377, Jun. 2017.