Robust Detection of Communities with Multi-Semantics in Large Attributed Networks

Di Jin[1], Ziyang Liu[1], Dongxiao He[1], Bogdan Gabrys[2], Katarzyna Musial[2]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[2] Advanced Analytics Institute, School of Software, Faculty of Engineering and IT, University of Technology Sydney, PO Box 123. Broadway NSW 2007, Australia
{jindi, liuziyang, hedongxiao}@tju.edu.cn,
{Bogdan.Gabrys, Katarzyna.Musial-Gabrys}@uts.edu.au

Abstract. In this paper, we are interested in how to explore and utilize the relationship between network communities and semantic topics in order to find the strong explanatory communities robustly. First, the relationship between communities and topics displays different situations. For example, from the viewpoint of semantic mapping, their relationship can be one-to-one, one-to-many or many-to-one. But from the standpoint of underlying community structures, the relationship can be consistent, partially consistent or completely inconsistent. Second, it will be helpful to not only find communities more precise but also reveal the communities' semantics that shows the relationship between communities and topics. To better describe this relationship, we introduce the transition probability which is an important concept in Markov chain into a well-designed nonnegative matrix factorization framework. This new transition probability matrix with a suitable prior which plays the role of depicting the relationship between communities and topics can perform well in this task. To illustrate the effectiveness of the proposed new approach, we conduct some experiments on both synthetic and real networks. The results show that our new method is superior to baselines in accuracy. We finally conduct a case study analysis to validate the new method's strong interpretability to detected communities.

Keywords: Community Detection, Social Networks, Semantics, Transition Probability, Nonnegative Matrix Factorization.

1 Introduction

Network science is a modern and significant discipline in many fields, such as social and computer science. Networks, consisting of nodes and edges which connect a pair of nodes, always occur in a variety of contexts [1]. The real-world networks usually share the same characteristic: they exhibit strong community structure. The property of community structure is: in which network nodes are joined together in tightly knit groups, between which there are only looser connections [2]. For example, in Facebook, users who have consistent interests often gather together and form a community but there are only few connections between such communities. Community structure reveals the fundamental functional modules of a network and enables us to better understand the interactive behavior of the network.

Community detection has developed rapidly in recent years and various community

detection methods, which mainly focus on network topology, have been proposed, e.g., the agglomerative or divisive algorithms [3], modularity optimization based methods [4], and spectral algorithms [5]. Further, it is well known that a node may belong to multiple communities (i.e. overlapping community). As a result, lots of methods were developed to detect overlapping community, such as *k*-clique community detection algorithms [6], local expansion and optimization algorithms [7] and probabilistic model-based algorithms [8]. Except for network topology, node attributes or link attributes are also taken into account when discovering communities [9-11]. In addition to improving community detection, researchers have realized that community detection should not only find community structure but also describe communities semantically by the use of abundant verbal information in the textual content. These descriptions can reveal why some nodes form a community and enable people to better understand the functions or meanings of communities, and in a way, this has much more practical value in real-world applications. Some methods have been proposed which combine topology and content information and give reasonable and interpretable communities [12, 13].

However, some problems still occur and need to be solved when network topology and node contents are integrated. One of the most important issues is the mismatch problem of topology and content. Traditional methods [12-14] typically assume that the network topology and node contents share the same community membership, but in many real social networks, this assumption does not always hold. For example, in a Twitter network, social links usually directly reflect which users gather into a community, while users may generate diverse and disordered content information. Thus, the community membership derived by network topology probably differs from the cluster membership derived by node contents.

For the above problem, it is necessary to extract useful content information to assist topology information in detecting more actual and accurate communities. In this paper, we propose a new generative model different from the traditional generative model and design a new community detection method, referred to as Robust and Strong Explanatory Community Detection (RSECD). To be specific, based on nonnegative matrix factorization (NMF), we are able to obtain the community membership matrix for network topology and cluster membership matrix for node contents. More importantly, there exists some implicit relation between network communities and content clusters, thus we introduce a transition probability matrix to depict it. As a result, even though the content information does not match with topology information, our method can still obtain accurate detection results by using the transition matrix with a suitable prior. At last, we put network topology, node content and transition matrix into a unified NMF framework, and optimize them altogether by designing effective updating rules in order to achieve an integral balance of them.

In the experiments, we use artificial networks to analyze the parameter in the objective function and to demonstrate the effectiveness and robustness of our approach. Next, we conduct experiments on seven real-world network datasets and compare RSECD with eight baseline methods in terms of both disjoint community and overlapping community evaluation metrics. Experimental results show that RSECD can significantly improve the performance in all comparisons, which further illustrates our approach's robustness. And finally, in order to verify that RSECD is strong-explanatory to communities, we use a case study on a musical social network to semantically explain the hidden meanings of some topics and tell the 'true stories' behind communities.

2 Related work

Various community detection methods, which only take the network topology into account, have been proposed. For example, hierarchical clustering methods [3] which include agglomerative and divisive hierarchical algorithms. Optimal modularity approaches (such as spectrum optimization method [5]) can find communities by the use of modularity optimization. Another approach [4] applies modularity into graphs of different networks by correcting modularity, such as symbolized networks. By mapping a network into a Laplacian matrix and calculating its eigenvector values, spectral methods can find each node's corresponding community accurately.

With in-depth analysis and research of complex network, the content information of complex networks shows its value and some community detection methods, which integrate the content information with network topology, have been developed. For instance, a subgraph overlapping clustering algorithm combining network structure and content information is proposed [9]. This method applies expectation-maximization (EM) algorithm to maximize likelihood function to generate stationary candidate subgraphs, and then uses k-means algorithm to cluster edges in order to obtain the overlapping community structure. A new generative probabilistic model is proposed which is learned by using a nested expectation-maximization algorithm and can describe the generalized communities [10]. In [11], a co-learning strategy is developed to jointly train the two parts (communities and semantics) in the model by combining a nested EM algorithm and belief propagation.

Recently, researchers have also realized that community detection should not only find communities, but also use rich verbal information in the text to give semantic description of communities. The description information reveals why some nodes gather into a community and helps people better understand the functions or implications of communities. For example, the approach in [12] using nonnegative matrix factorization integrates two tasks of community detection and user profiling into a unified model, and then achieves community profiling by a linear operator integrating the profiles of users. A joint community profiling and detection (CPD) model [13] is proposed which describes communities by published content and friendship links of users. In addition, the method SCI [14], which can detect and describe communities, has also been proposed. This method uses nonnegative matrix factorization to integrate topology and content information into a unified model, and achieves relatively high detection accuracy in comparison with other methods. More importantly, SCI can not only detect communities, but also analyzes the semantics of detected communities. In general, this type of method has more practical value than others without semantics.

However, the methods mentioned above mainly focus on how to effectively fuse topology structure and content information to improve the performance of community detection while do not further consider how to detect communities more robustly, especially when the node contents do not match well with network communities. Moreover, most of these methods can only interpret each community using a single topic, which is far from satisfactory in many real applications.

3 RSECD: The Network Model

Our proposed RSECD approach extends the previous SCI approach by introducing a transition probability matrix with a suitable prior to represent the hidden relationship between network communities and content clusters. In this section, firstly, we illustrate the difference between traditional generative model and our proposed new generative

model; then we give some notations. Finally, we elaborate how to model RSECD.

3.1 Traditional Generative Model vs. New Generative Model

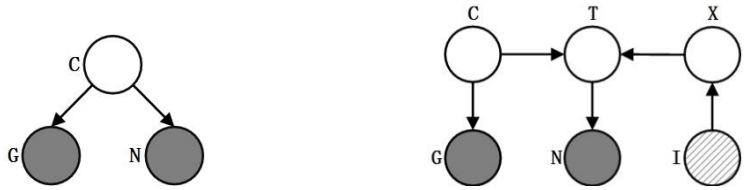


Fig. 1. A comparison of traditional generative model and our proposed new model. **(a)** is the traditional generative model where community structure C directly generates network topology G and node contents N. **(b)** is RSECD's generative model where node contents N implicates topic cluster T (not community structure C) and topic cluster T is generated by community structure C and transition probability matrix X together. In addition, identity matrix I, as the transition matrix's prior, plays a key guiding role in fusing these two types of information.

Most of community detection methods [9-14] follow traditional generative model which generally assumes that network topology and node contents share the same community structure (as shown in Fig. 1(a)). While in many real-world networks, network topology and node contents may implicate different community structures, so that we modify the traditional generative model and design a more reasonable generative model, as shown in Fig. 1(b). In this new model, node contents N implicates topic cluster T (not community structure C) and topic cluster T is generated by community structure C and transition probability matrix X together.

3.2 Notations

For an undirected network G with n nodes and e edges, we represent it by a binary-valued adjacency matrix $A \in R^{n \times n}$. Each node i has its attributes S_i, which may be the semantic information of the node. S_i is in the form of an m-dimensional binary-valued vector. All of S_i form an attribute matrix $S \in R^{n \times m}$. The community detection task is: when A and S are observed, on topology, we need to find k different communities; on content cluster with semantics, we need to find k' different topics and infer the semantics for each community. Because all of the baseline algorithms assume that the number of communities is equal to that of topics, we still assume $k = k'$ in this paper. However, our RSECD algorithm can also apply equally to $k = k'$.

3.3 Modeling Network Topology

Our network topology model is based on the following intuitive properties: 1) if two nodes belong to the same community, they are more likely to be connected; 2) if two nodes have similar community memberships, they have a high probability to be linked. We define the propensity of node i belonging to community c as u_{ic}. Then we have a community membership of all nodes denoted as $U = (u_{ic})_{n \times k}$. Based on the first propensity, we can use $u_{ic}u_{jc}$ to represent the expected number of edges between nodes i and j in community c. Based on the second propensity, we can achieve that the expected number of edges between nodes i and j in the whole network is $\sum_{c=1}^{k} u_{ic}u_{jc}$. Considering all nodes, we have the following loss function:

$$\min_{U \geq 0} \| A - UU^T \|_F^2 \tag{1}$$

3.4 Modeling Node Attributes

We define the propensity of topic t having attribute q as c_{qt} and the propensity of

node i belonging to topic t as v_{it}. Then we have an attribute membership of all topics denoted as $C = (c_{qt})_{m\times k}$ and a topic cluster membership of all nodes denoted as $V = (v_{it})_{n\times k}$. In addition, we define the propensity of a node i having attribute q as s_{iq}, which is an element of attribute matrix S. We suppose that if node i belongs to topic t, node i and topic t will have similar attributes information. It can be represented as $s_{iq} = \sum_{t=1}^{k} v_{it} c_{tq}$. Then we have the following loss function:

$$\min_{C\geq 0, V\geq 0} \| S - VC^{T} \|_{F}^{2} \tag{2}$$

3.5 Modeling Transition Probabilities

Transition probability is an important concept of Markov chain and is defined as the probability of transferring from one state to another. We introduce transition probabilities to represent the relationship between network communities and topic clusters. Here the probability transferring from community c to topic t is defined as x_{ct}, the probability vector transferring from community c to any topic is defined as x_c (x_c satisfies a probability distribution) and the probability matrix transferring from any community to any topic is defined as X. Moreover, to effectively guide the fusion of topology and content, we employ identity matrix I as the prior of X. Then we have the following loss function:

$$\min_{X\geq 0} \| UX - V \|_{F}^{2} + \| X1_{k}^{T} - 1_{k}^{T} \|_{F}^{2} + \| I - X \|_{F}^{2} \tag{3}$$

where $1_{k}^{T} \in R^{k\times 1}$ and all of its elements are 1.

3.6 The Unified Model

By combining the objective functions of the above formulas (including (1) to (3)), we obtain RSECD's overall loss function:

$$\min_{\substack{U\geq 0, V\geq 0,\\ C\geq 0, X\geq 0}} L = \| A - UU^{T} \|_{F}^{2} + \alpha \| S - VC^{T} \|_{F}^{2} + \| UX - V \|_{F}^{2} + \| X1_{k}^{T} - 1_{k}^{T} \|_{F}^{2} + \| I - X \|_{F}^{2} \tag{4}$$

where α is a balance parameter between network topology and node contents.

Our RSECD model can deal with the topology and content's mismatch problem in networks well. To be specific, 1) when topology matches with content very well, the first two parts of unified model (network topology model and node attributes model) work so that topology and content can reinforce each other in order to find more exact community structure. 2) When only some parts of content match with network topology, RSECD can also extract useful material from content information to assist topology information in detecting more actual and accurate communities by the mapping and tractive function of transition matrix X. 3) When content does not match with topology at all, matrices U and V are almost orthogonal, thus matrix X is close to a random matrix and the final result is equal to that of using only topology information. In addition, the optimized X essentially represents the mapping relationship between communities and topics, so that we can also use X to explain the detected communities. So our RSECD is robust and strong-explanatory to community detection. We will further use extensive experiments (including a case study) to demonstrate these cases.

4 Optimization

Since the objective function in (4) is not convex, it is hard to obtain the global optimal solution. Fortunately, the local minima of (4) can be obtained using the Majorization - Minimization framework [16]. Here we describe an algorithm that iteratively updates U with V, C, X fixed, updates V with U, C, X fixed, updates C with U, V, X fixed, and

updates X with U, V, C fixed, which guarantees that our objective does not increase and the parameters keep nonnegative (with any nonnegative initial seeds) after each iteration. The specific formulas are shown in the following subproblems.

4.1 U-Iteration

When updating U, we need to solve the following problem:

$$\min_{U\geq 0} L(U) = \| A - UU^T \|_F^2 + \| UX - V \|_F^2 \tag{5}$$

An arbitrary matrix M satisfies $\| M \|_F^2 = \mathrm{tr}(MM^T)$, so we transform this problem as:

$$L(U) = \mathrm{tr}(A^TA - A^TUU^T - UU^TA + UU^TUU^T) + \mathrm{tr}(X^TU^TUX - X^TU^TV - V^TUX + V^TV) \tag{6}$$

We then take a derivative with respect to U and get the following formula:

$$\frac{\partial L(U)}{\partial U} = -2(A^T + A)U + 2(UX - V)X^T + 4UU^TU \tag{7}$$

In order to reduce computational cost, we use a multiplicative update algorithm based on the Oja's iterative learning rule [15] to update U. We decompose (7) into two sets:

$$\nabla_U L(U) = \nabla_+ - \nabla_- \tag{8}$$

where ∇_+ (∇_-) is the sum of all positive (negative) components, then we have:

$$U_{new} = U_{old} \frac{\nabla_-}{\nabla_+} \tag{9}$$

In (7), the negative terms are $2A^TU$, $2AU$, $2VX^T$ and the positive terms are $2UXX^T$, $4UU^TU$. So we have the updating rule of U as:

$$u_{ij} \leftarrow u_{ij} \left(\frac{A^TU + AU + VX^T}{UXX^T + 2UU^TU}\right)_{ij} \tag{10}$$

4.2 V-Iteration and C-Iteration

When updating V, we need to solve the following problem:

$$\min_{V\geq 0} L(V) = \alpha \| S - VC^T \|_F^2 + \| UX - V \|_F^2 \tag{11}$$

In order to iterate V, we transform this problem into the following equation:

$$L(V) = \alpha \cdot \mathrm{tr}(S^TS - S^TVC^T - CV^TS + CV^TVC^T) + \mathrm{tr}(X^TU^TUX - X^TU^TV - V^TUX + V^TV) \tag{12}$$

We then take a derivative with respect to V and get the next formula:

$$\frac{\partial L(V)}{\partial V} = -2\alpha SC - 2UX + 2\alpha VC^TC + 2V \tag{13}$$

Similar to (10), we then obtain the updating rule of V as:

$$v_{ij} \leftarrow v_{ij} \left(\frac{\alpha SC + UX}{\alpha VC^TC + V}\right)_{ij} \tag{14}$$

When updating C, similar to the steps from (11) to (14), we obtain the updating rule of C as:

$$c_{ij} \leftarrow c_{ij} \left(\frac{S^TV}{CV^TV}\right)_{ij} \tag{15}$$

4.3 X-Iteration

When updating X, we need to solve the following problem:

$$\min_{X\geq 0} L(X) = \| UX - V \|_F^2 + \| I - X \|_F^2 + \| X1_k^T - 1_k^T \|_F^2 \tag{16}$$

To iterate X, we can transform this problem into the following equation:

$$L(\mathrm{X}) = \mathrm{tr}(\mathrm{X}^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathrm{U}\mathrm{X} - \mathrm{X}^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathrm{V} - \mathrm{V}^{\mathrm{T}}\mathrm{U}\mathrm{X} + \mathrm{V}^{\mathrm{T}}\mathrm{V}) + \mathrm{tr}(1_k \mathrm{X}^{\mathrm{T}}\mathrm{X}1_k^{\mathrm{T}} - 1_k \mathrm{X}^{\mathrm{T}}1_k^{\mathrm{T}} - 1_k \mathrm{X}1_k^{\mathrm{T}} + 1_k 1_k^{\mathrm{T}}) + \mathrm{tr}(\mathrm{I} - \mathrm{X} - \mathrm{X}^{\mathrm{T}} + \mathrm{X}^{\mathrm{T}}\mathrm{X}) \quad (17)$$

We then take a derivative with respect to X and get the next formula:

$$\frac{\partial L(\mathrm{X})}{\partial \mathrm{X}} = -2\mathrm{U}^{\mathrm{T}}\mathrm{V} - 2\mathrm{I} - 2\mathrm{M} + 2\mathrm{U}^{\mathrm{T}}\mathrm{U}\mathrm{X} + 2\mathrm{X}\mathrm{M} + 2\mathrm{X} \quad (18)$$

where $\mathrm{M} \in \mathrm{R}^{k \times k}$ and its elements are all 1. In (21), the negative terms are $2\mathrm{U}^{\mathrm{T}}\mathrm{V}$, 2I, 2M and positive terms are $2\mathrm{U}^{\mathrm{T}}\mathrm{U}\mathrm{X}$, 2XTM, 2X. So we obtain the updating rule of X:

$$x_{ij} \leftarrow x_{ij} \left(\frac{\mathrm{U}^{\mathrm{T}}\mathrm{V} + \mathrm{I} + \mathrm{M}}{\mathrm{U}^{\mathrm{T}}\mathrm{U}\mathrm{X} + \mathrm{X}\mathrm{M} + \mathrm{X}}\right)_{ij} \quad (19)$$

5 Experiments

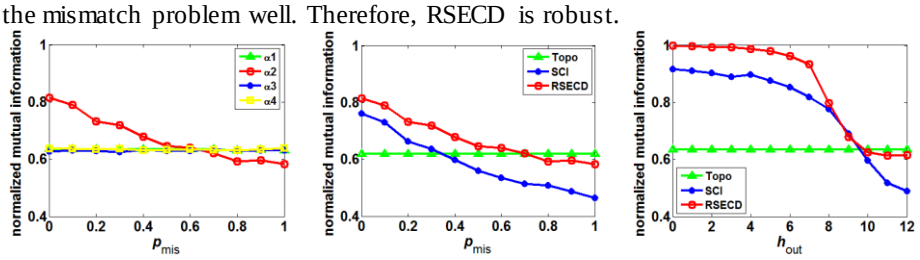
Here we first use artificial networks to analyze the influence of parameter α in the objective function and demonstrate that our approach can solve the mismatch problem well. We then compare our method with eight state-of-the-art algorithms on seven real datasets in terms of four well-known metrics. And finally, we discuss a case study analysis to show that our method has a strong explanatory capability to communities.

5.1 Artificial Networks

We use the Newman's model [2] to generate artificial benchmark networks. Each network has 128 nodes which have been divided into 4 communities. Each node has z_{in} edges connecting to the nodes of the same community and z_{out} edges connecting to the nodes of different communities ($z_{in} + z_{out} = 16$). In addition, all nodes are partitioned into 4 clusters corresponding to 4 communities. To be specific, for each node in the sth cluster, we use a binomial distribution with mean $p_{in} = h_{in}/h$ to generate a h-dimensional binary vector as its $((s - 1) \times h + 1)$-th to $(s \times h)$-th attributes and use a binomial distribution with mean $p_{out} = h_{out}/(3h)$ to generate its rest attributes. In our experiment, we set $h = 50$, $z_{out} = h_{out} = 8$ and use normalized mutual information (NMI) [19] as the metric. To simulate real-world networks' mismatch problem, we use p_{mis} (ranging from 0 to 1) to reveal the mismatch rate between network topology and node contents. For example, if $p_{mis} = 0.8$, then in this network, there are 20 percent of nodes whose contents match with topology and 80 percent of nodes whose contents do not match with topology. In the first experiment, based on experience, we consider four choices for parameter α ($\alpha = 1$, $\|\mathrm{A}\|_{\mathrm{F}}^2$, $1/\|\mathrm{S}\|_{\mathrm{F}}^2$, or $\|\mathrm{A}\|_{\mathrm{F}}^2 / \|\mathrm{S}\|_{\mathrm{F}}^2$) and respectively compute the average NMI values under them. The results are shown in Fig. 2(a), when p_{mis} is less than 0.6 (this corresponds to most cases in real-world networks), the result under $\alpha = \|\mathrm{A}\|_{\mathrm{F}}^2$ is greater than the others, so we conclude that choosing $\alpha = \|\mathrm{A}\|_{\mathrm{F}}^2$ as the default value may be better than the other three choices.

Next, to illustrate RSECD's robustness, we compare three methods—Topo, SCI and RSECD. Topo is a variant of RSECD using topology information alone. SCI is a NMF-based method using topology and content information together but did not consider the mismatch problem [14]. As shown in Fig. 2(b), Topo keeps a stable detection accuracy no matter how p_{mis} changes because the topology information existing in the network is fixed. When p_{mis} is less than 0.3, as SCI combines topology and content information together, it has higher accuracy than Topo. However, because SCI fails to solve the mismatch problem, when p_{mis} is greater than 0.4, the performance of SCI gradually weakens and is worse than Topo. RSECD, as the extended work of SCI, has better

performance than Topo and SCI when p_{mis} is less than 0.7. Moreover, when p_{mis} is larger than 0.7 (i.e., a high mismatch rate in the network), RSECD is just slightly worse than Topo but much better than SCI. In summary, the result demonstrates that: 1) when content match with topology well, RSCED can better combine topology and content to find communities; 2) when content does not match with topology, RSECD can also solve the mismatch problem well. Therefore, RSECD is robust.



(a) (b) (c)

Fig. 2. Results on artificial networks. **(a)** is the NMI results under 4 different choices of parameter α. **(b)** is 3 different methods' NMI results when the mismatch rate p_{mis} varies from 0 to 1. **(c)** is 3 different methods' NMI results when h_{out} varies from 0 to 12 under p_{mis}=0.

Finally, because the cluster structure implicated by content information may be indistinct in the real-world networks, we design a third experiment. In this part, we set $p_{mis} = 0$ and relieve the constraint $h_{out} = 8$, making h_{out} vary from 0 to 12. The larger h_{out} is, the higher distinct degree is. The final result is shown in Figure 2(c). As we can see, RSECD's accuracy is almost always higher than that of SCI. Even though when the cluster structure is very indistinct, RSECD's accuracy does not decline too much and is very close to that of Topo.

5.2 Real-World Networks

Datasets. We use 7 real networks [17, 18] with node attributes and ground-truth community labels. These datasets are often used in the field of community detection by researchers and their detailed information is shown in Table 1. In this table, the number of attributes represents the total number of attributes in the network.

Table 1. Datasets used.

Dataset	Communities	Nodes	Edges	Attributes	Ground Truth
Facebook	14	226	3,417	131	√
Cornell	5	877	1,608	1,703	√
Texas	5	877	1,608	1,703	√
Washington	5	877	1,608	1,703	√
Wisconsin	5	877	1,608	1,703	√
Citeseer	6	3,312	4,732	3,703	√
Uai2010	19	3,363	45,006	4,972	√

Metrics. To test RSECD's performance, we conduct a quantitative analysis of the final detection results using two types of metrics (disjoint community metrics and overlapping community metrics). For disjoint community metrics, we choose accuracy (AC) [19] and normalized mutual information (NMI) [19]. AC is used to measure the percentage of correct labels obtained. In clustering applications, NMI is used to measure how similar two sets of clusters are. For overlapping community metrics, we choose F-score [20] and Jaccard similarity [20]. Both of them are common metrics which are used to quantify the performance in terms of the agreement between the ground-truth communities and the detected communities.

Baselines. To illustrate RSECD's effectiveness, we choose three types of baseline algorithms including two topology-based methods (DCSBM [21] and BigCLAM [22]), one content-based method (AP [23]), and five methods using both topology and content (CESNA [24], DCM [25], PCL-DC [26], Block-LDA [27] and SCI [14]).

Setting. In the experiments, first for each network we uniformly set α to be $\| A \|_F^2$ based on previous parameter analysis. We then repeat RSECD algorithm 20 times with different random seeds. We obtain the result which corresponds to the smallest loss function value as the final result.

Table 2. Performance comparison of different methods using disjoint community metrics. Here "topo", "cont", "both" denote methods using topology, contents, and topology-and-contents.

Metrics	Methods		Datasets					
(%)	*Type*	*Name*	*Cornell*	*Texas*	*Washington*	*Wisconsin*	*Citeseer*	*Uai2010*
AC	topo	DCSBM	37.95	48.09	31.80	32.82	26.57	2.60
	both	PCL-DC	30.26	38.80	29.95	30.15	24.85	28.82
	both	Block-LDA	46.15	54.10	39.17	49.62	24.35	16.04
	both	SCI	36.92	49.73	46.09	46.42	29.53	29.51
	both	RSECD	**53.85**	**61.50**	**58.70**	**69.43**	**48.67**	**47.21**
NMI	topo	DCSBM	9.69	16.65	9.87	3.14	4.13	31.22
	cont	AP	25.27	31.02	31.79	32.48	13.28	41.60
	both	PCL-DC	7.23	10.37	5.66	5.01	2.99	26.92
	both	Block-LDA	6.81	4.21	3.69	10.09	2.42	5.70
	both	SCI	6.80	12.49	6.83	13.28	7.17	23.39
	both	RSECD	**30.24**	**32.67**	**35.10**	**45.32**	**22.34**	**45.73**

Table 3. Performance comparison of different methods using overlapping community metrics.

Metrics	Methods		Datasets						
(%)	*Type*	*Name*	*Cornell*	*Texas*	*Washington*	*Wisconsin*	*Facebook*	*Citeseer*	*Uai2010*
F-score	topo	DCSBM	34.08	36.14	32.83	29.47	44.92	26.83	30.12
	topo	BigCLAM	13.23	20.64	13.35	12.84	47.40	9.30	16.99
	cont	AP	21.10	23.59	24.11	20.53	23.60	12.92	13.23
	both	CESNA	23.48	23.54	21.91	23.17	52.51	3.38	32.32
	both	DCM	14.38	11.15	12.45	10.45	41.29	2.50	9.65
	both	PCL-DC	32.03	34.30	30.38	27.83	39.49	25.49	29.71
	both	Block-LDA	36.77	32.55	28.95	31.36	39.57	22.49	18.58
	both	SCI	26.94	30.99	28.06	27.06	24.94	26.18	29.66
	both	RSECD	**53.26**	**44.89**	**47.44**	**53.54**	**52.73**	**45.77**	**43.86**
Jaccard	topo	DCSBM	21.20	24.14	20.06	17.92	32.18	15.78	18.81
	topo	BigCLAM	7.18	12.18	7.25	7.01	34.25	5.01	9.87
	cont	AP	13.32	16.39	16.26	12.51	13.63	7.39	7.88
	both	CESNA	13.47	13.57	12.40	13.14	39.82	1.73	21.26
	both	DCM	7.95	6.03	6.72	5.54	33.60	1.27	5.77
	both	PCL-DC	19.02	21.56	18.99	16.27	26.99	14.75	19.17
	both	Block-LDA	24.29	22.51	18.20	20.31	26.61	12.80	11.08
	both	SCI	17.10	21.98	18.72	17.15	15.65	15.26	19.11
	both	RSECD	**37.12**	**33.32**	**34.04**	**41.47**	**41.67**	**31.49**	**32.39**

Results. We show the final results in Tables 2 and 3. It is worth noting that AP cannot compute accuracy (AC) value, and CESNA and DCM are only applicative to overlapping community metrics. In the tables, we use bold to mark the best results. Table 2 shows the comparison results in terms of AC and NMI. In AC, our method RSECD performs best among all the five methods. In NMI, RSECD still achieves the best results in comparison to the other methods. All the comparison results using different algorithms under overlapping community metrics are shown in Table 3. In these results, RSECD again has the best performance in comparison to the other tested approaches. In summary, the main reasons that our algorithm achieves such superior performance are as follows: 1) RSECD assumes that topology and content do not share the same

community structure, so that those harmful content information will not interfere with topology information's important role in community detection; 2) transition probability matrix, as a filter of content information, can retain beneficial content information which can assist topology information in detecting more actual, accurate communities and remove harmful content information which has wrong guidance in community detection. Therefore, RSCED can solve the mismatch problem well and the final performance results are relatively high and stable in any case.

Efficiency. As like standard nonnegative matrix factorization, the calculational complexity of RSECD is $O(T(n^2k + 2mnk + nk^2))$ where T is the number of iterations, n the number of nodes, k the number of communities ($k<<n$) and m the number of attributes. By taking into account the sparsity of the adjacency matrix A and attribute matrix S, RSECD needs $O(T(ek + 2e'k + nk^2)$ time where e is the number of edges ($e<<n$) and e' the number of nonzero elements in the attribute matrix S ($e'<<m$). Thus, the computational complexity of RSECD is near linear with the number of nodes. We also report RSECD's running time. It needs 2.893s (here "s" denotes seconds), 8.9s, 8.233s, 10.952s, 14.041s, 6248.029s and 5760.169s, respectively, on the datasets Facebook, Cornell, Texas, Washington, Wisconsin, Citeseer and Uai2010.

5.3 A Case Study on Lastfm

We select LASTFM dataset[1], which comes from a musical social network, as our dataset for the case study analysis. This dataset contains 1,892 users and the total number of attributes in the network is 11,946. These attributes reveal users' favorite songs or singers. LASTFM does not have the ground-truth of community labels. While, all the methods used in this work need the number of communities to be given. So, as did in [14], we use Louvain method [28] to set the number of communities in this network to 38. Two vivid examples to interpret the communities derived are shown in Figs. 3 and 4 in the form of word clouds. Word clouds can graphically show different attribute words' importance degree in one community in order to explain the current community's semantics. That is, in a word cloud, the size of a word is proportional to the probability that it belongs to this community.

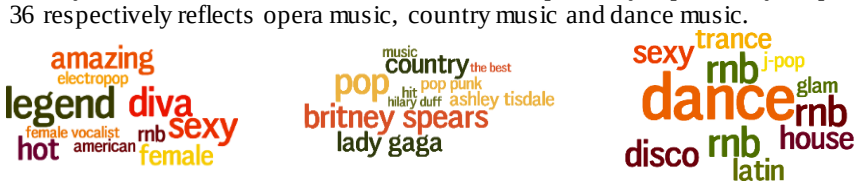
The first example is the 30th community which contains two dominant topics, i.e., topics 1 and 32. Topic 1, as shown in Fig. 3(a), is highly related to electronic pop music. The total of "electronic", "electropop" and "electronica" has a high proportion in all attribute words and illustrates that the theme of topic 1 is pop electronic music. In addition, "australian", "8-bit", "synth pop", "big beat" and "dark pop" are different styles of pop electronic music. On the other hand, topic 32, as shown in Fig. 3(b), mainly denotes synth pop music. Synth pop music origins from "new wave", "post-punk" and is popular in "80s". "new romantic" is a synth pop song of Taylor Swift. "depeche mode" is a British band in style of alternative dance and synth pop. "electroclash" is another name of "tech pop" which contains the style of synth pop. "synth" and "synth pop" also appear here. It is worth noting that, these two topics which corresponds to electronic pop music of multiple styles and synth pop music, respectively, both belong to electronic pop music although being the different branches. Therefore, the 30th community will be a group of fans adoring electronic pop music mainly including synth pop music.

[1] http://ir.ii.uam.es/hetrec2011/datasets.html

(a) (b)

Fig. 3. Word clouds for the 30th community. **(a)** denotes topic 1 and **(b)** denotes topic 32, both of which are dominant topics of the 30th community.

Our second example is the 16th community which contains three dominant topics, i.e, topic 13, 24 and 36. They are shown in Fig. 4(a), (b) and (c), respectively. Similar to the previous analysis, we found out that topic 13 is related to opera music (for example, "diva", "female vocalist" appear here); topic 24 is related to country music and pop music (for example, "country", "pop" appear here); and topic 36 is related to dance music (for example, "dance", "disco" appear here). Simultaneously, these three topics have the same theme, i.e., female singer. So, we can conclude that the 16th community's dominant topic is female singers and the three topics (topic 13, 24, 36) in this community all have their own accurate semantics, respectively. Specifically, topic 13, 24, 36 respectively reflects opera music, country music and dance music.



(a) (b) (c)

Fig. 4. Word clouds for the 16th community. This community contains three dominant topics, in which **(a)** denotes topic 13, **(b)** denotes topic 24 and **(c)** denotes topic 36.

6 Conclusion

In this paper, we proposed a new community detection method (RSECD) which is able to detect communities and in the same time analyze the semantics of founded communities. We introduced a nonnegative matrix factorization model to depict the relationships between nodes, topics and communities more accurately. A transition probability matrix with a suitable prior was also introduced to show their hidden relationships to improve the robustness of the new model, especially when node contents do not match well with network topology. Through artificial benchmark networks, we analyzed the influence of parameter α in the objective function and demonstrated RSECD's high level of robustness. On real-world networks, we showed that RSECD outperforms all of the baseline methods. Finally, the case study analysis on a musical social network showed how the semantic explanation of communities derived by RSECD works. This helps people to understand and interpret communities more precisely and in a human-readable form in many real applications.

Acknowledgment

This work was supported by the National Key R&D Program of China (2017YFC0820106), the Natural Science Foundation of China (61502334, 61772361, 61673293) and the Elite Scholar Program of Tianjin University (2017XRG-0016).

References

1. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports 659, 1-44 (2016).
2. Girvan, M., Newman, M.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99(12), 7821-7826 (2002).
3. Jia, S., Gao, L., Gao, Y., *et al.*: Defining and identifying cograph communities in complex networks. New Journal of Physics 17(1), 013044 (2015).
4. Yang, L., Cao, X., He, D., *et al.*: Modularity based community detection with deep learning. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 2252-2258. New York, USA (2016).
5. Fanuel, M., Alaiz, C. M., Suykens, J. A., *et al.*: Magnetic eigenmaps for community detection in directed networks. Physical Review E 95(2), 022302 (2017).
6. Hao, F., Min, G., Pei, Z., *et al.*: *K*-clique community detection in social networks based on formal concept analysis. IEEE Systems Journal 11(1), 250-259 (2017).
7. Whang, J. J., Gleich, D. F., Dhillon, I. S., *et al.*: Overlapping community detection using neighborhood-inflated seed expansion. IEEE Transactions on Knowledge and Data Engineering 28(5), 1272-1284 (2016).
8. Jin, D., Wang, H., Dang, J., *et al.*: Detect overlapping communities via modeling and ranking node popularities. In: 30th AAAI Conference on Artificial Intelligence, pp. 172-178. Phoenix, Arizona, USA (2016).
9. Van Laarhoven, T., Marchiori, E.: Local network community detection with continuous optimization of conductance and weighted kernel k-means. Journal of Machine Learning Research 17(147), 1-28 (2016).
10. Jin, D., Wang, X., He, R., *et al.*: Robust detection of link communities in large social networks by exploiting link semantics. In: 32th AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA (2018).
11. He, D., Feng, Z., Jin, D., *et al.*: Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In: 31th AAAI Conference on Artificial Intelligence, San Francisco, California, USA (2017).
12. Akbari, M., Chua, T. S.: Leveraging behavioral factorization and prior knowledge for community discovery and profiling. In: Web Search and Data Mining (WSDM), pp. 71-79. UK (2017).
13. Cai, H., Zheng, V. W., Zhu, F. *et al.*: From community detection to community profiling. Proceedings of the Vldb Endowment 10(7), 817-828 (2017).
14. Wang, X., Jin, D., Cao, X., *et al.*: Semantic community identification in large attribute networks. In: 30th AAAI Conference on Artificial Intelligence, pp. 265-271. Phoenix, Arizona, USA (2016).
15. Oja, E.: Principal components, minor components, and linear neural networks. Neural Networks 5(6), 927-935 (1992).
16. Hunter, D. R., Lange, K. A.: A tutorial on mm algorithms. The American Statistician 58(1), 30-37 (2004).
17. Sen, P., Namata, G., Bilgic, M., *et al.*: Collective classification in network data. AI Magazine 29(3), 93-106 (2008).
18. Leskovec, J. 2016. Stanford Network Analysis Project. http://snap.stanford.edu.
19. Liu, H., Wu, Z., Li, X., *et al.*: Constrained nonnegative matrix factorization for image representation. IEEE Transactions on Software Engineering 34(7), 1299-1311 (2012).
20. Yang, J., Mcauley, J., Leskovec, J., *et al.*: Community detection in networks with node attributes. In: the IEEE International Conference on Data Mining series (ICDM), pp. 1151-1156. Dallas, Texas, USA (2013).
21. Karrer, B., Newman, M.: Stochastic blockmodels and community structure in networks. Phys. Rev.E 83(1), 016107 (2011).
22. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Web Search and Data Mining (WSDM), pp. 587–596. Rome, Italy (2013).
23. Frey, B. J., Dueck, D.: Clustering by Passing Messages Between Data Points. Science 315(5814), 972-976 (2007).
24. Yang, J., Mcauley, J., Leskovec, J., *et al.*: Community detectionin networks with node attributes. In: the IEEE International Conference on Data Mining series (ICDM), pp. 1151-1156. Dallas, Texas, USA (2013).
25. Pool, S., Bonchi, F., Van Leeuwen, M., *et al.*: Description-driven community detection. ACM Transactions on Intelligent Systems and Technology 5(2), 1-28 (2014).
26. Yang, T., Jin, R., Chi, Y., *et al.*: Combining link and content for community detection: a discriminative approach. In: 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 927-936. Paris, France (2009).
27. Balasubramanyan, R., Cohen, W. W.: Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In: SIAM International Conference on Data Mining (SDM), pp. 450-461. Mesa, Arizona, USA (2011).
28. Kido, G. S., Igawa, R. A., Barbon Jr, S.: Topic modeling based on louvain method in online social networks. In: Proc. of XII Brazilian Symposium on Information Systems, pp. 353-360. Florianópolis, SC (2016).