A Unified Weakly Supervised Framework for Community Detection and Semantic Matching

Wenjun Wang[1,2], Xiao Liu[1,2], Pengfei Jiao[1,2], Xue Chen[1,2], and Di Jin[1,2(✉)]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
{wjwang,xiaoxiao,pjiao,jindi}@tju.edu.cn, chenxuemail@163.com
[2] Tianjin Key Laboratory of Advanced Networking (TANK), Tianjin 300350, China

Abstract. Due to the sparsity of network, some community detection methods only based on topology often lead to relatively low accuracy. Although some methods have been proposed to improve the detection accuracy by using few known semi-supervised information or node content, the research of community detection not only pursues the enhancement of community accuracy, but also pays more attention to the semantic description for communities. In this paper, we proposed a unified nonnegative matrix factorization framework simultaneously for community detection and semantic matching by integrating both semi-supervised information and node content. The framework reveals two-fold community structures as well as their coupling relationship matrix, which helps to identify accurate community structure and at the same time assign specific semantic information to each community. Experiments on some real networks show that the framework is efficient to match each community with specific semantic information, and the performance are superior over the compared methods.

Keywords: Community detection · Nonnegative matrix factorization
Semi-supervised learning · Semantic matching

1 Introduction

The complex network is constituted by a group of entities and their interactive relationships. These direct or indirect interactions can partition the network into several functional communities, making which interact densely in each community and sparsely between them. For example, the protein network is partitioned into different functional units via the interaction among protein molecule. Therefore, the identification of these communities is helpful to understand how the network works and how the functional unit interacts. However, for many networks in real world, due to its community structure is very vague, it is very difficult to identify solely using the observed interactions. How to integrate the structural and semantic information to identify more accurate community structure and

D. Phung et al. (Eds.): PAKDD 2018, LNAI 10939, pp. 218–230, 2018.
https://doi.org/10.1007/978-3-319-93040-4_18

simultaneously assign an appropriate semantic description to each community is a worthy studying heat.

The early community detection methods only use network topology, including hierarchical clustering [1], spectral clustering [2], modularity optimization [3,4] and methods based on generative model [5]. However, for networks with sparse connections and vague community structure, these methods almost fail to accurately identify its community structure.

In order to uncover the vague community hidden in networks, it is necessary to exploit additional available prior information, and some semi-supervised community detection methods [6–10] have been proposed. Specifically, combined with both node labels and pairwise constraints, Eaton and Mansbach proposed a semi-supervised spin-model for community detection, which penalizes the term that violates the guidance and rewards the term that agrees with the guidance [6]. Based on latent space graph regularization, Yang *et al.* utilized must-link constraints to derive a unified semi-supervised community detection framework [8]. Zhang *et al.* directly used the pairwise constraints to modify the adjacency matrix of networks, and proposed a semi-supervised community detection framework [9,11]. Considering that the heterogeneity of node degree and community size may lower the utilization of prior constraints, Liu *et al.* developed a semi-supervised NMF community detection method with node popularity [10]. Indeed, the integration of semi-supervised prior information and network topology plays a vital role in assisting to reveal the vague community structure, but for very sparse networks, the semi-supervised prior cannot be effectively used, and usually has lower utilization. Moreover, it ignores the specific semantic of each community.

In addition, the node contents are often available. For example, a user of a social network often has a person profile with content information such as age, male, education background and profession; a paper in citation network often provides some contents information including author, title, abstract and key words. It is generally assumed that nodes of more similar contents information are more likely to belong to the same community. Therefore, node contents have been widely used to guide the community detection and depict the community semantic [12–14]. The early content-based methods handle the network topologies and content separately, and most of the methods just use node contents to improve the community detection accuracy and compensate the insufficiency of sparse topology. For example, by combining the user similarity, message similarity and user interaction, Pei *et al.* proposed a nonnegative matrix tri-factorization clustering framework to identify the community structure in a social network [15]. Recently, some researchers often use node contents to describe the semantic explanation for community, so as to further understand why some certain nodes belong to the same community, and what characteristics the community owns. From the perspective of content propagation, Liu *et al.* combined the topological structure as well as the content information to detect the community structure, and adopted the stable status of random walk to describe the semantic information of communities [16]. By integrating network topology and

semantic information of nodes, Wang *et al.* proposed a novel nonnegative matrix factorization (NMF) model [17], and by defining two sets of parameters, the community membership matrix and community attribute matrix respectively, to infer the community structure and its corresponding semantic interpretation.

However, most of these newly proposed methods have three potential problems. Firstly, users tend to form a community due to their interactions. For sparse network, the relatively vague community structure is difficult to accurately identify, and node contents cannot assign appropriate semantic topic for each community when the identified community structure is wrong. Secondly, they generally believe that network topology and node content share the same community membership, but there may be more than one semantic topic for each community. Therefore, although the above methods can identify accurate community structure, they cannot assign correct semantic interpretation to a community. Finally, most of the existing methods utilize network topology and node contents separately, ignoring the relation between topology and content.

In this paper, for sparse networks we propose a unified weakly supervised framework for community detection and semantic matching (WSCDSM). Firstly, we incorporate network topology with must-link prior to derive an accurate topology-driven community (TC) membership, and then utilize node content information to obtain a semantic-driven community (SC) membership. Finally, by introducing a coupling matrix to portray the matching relation between TC and SC community structure, we integrate the above two process into WSCDSM framework to simultaneously detect community structure and match semantic. In our framework, two types of auxiliary information are seamlessly integrated to reveal the vague community structure and help to understand the practical semantic of communities. Consequently, the prior information and node contents are not only more effectively utilized, but also can complement some missing information of each other. We adopt an iterative method to train the TC (SC) community membership and its coupling relationship. Experimental results on several real networks validate that the proposed framework not only improves, as expected, the detection accuracy of vague communities, but also assign an appropriate semantic interpretation to each community.

The contributions of this work are as follows:

(1) Integrating with topological and content information as well as semi-supervised prior, we proposed a unified framework simultaneously for community detection and semantic matching. In this framework, we introduce coupling matrix to depict the relationship between community and semantic topic. Besides, it can also adjust the semantic information of each community.
(2) On the basis of using semi-supervised prior to improve the community accuracy, our proposed framework can integrate content information to compensate the insufficiency of topological information, and further assign more appropriate semantic information to each community.
(3) Our proposed framework is superior over the compared methods in most cases, and the improvement is more obvious on vary sparse network.

2 Proposed WSCDSM Framework

Considering an undirected attributed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{S})$ of n nodes $\mathbf{V}$ and e edges $\mathbf{E}$, which often can be represented by a binary-valued adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and an attribute matrix $\mathbf{S} \in \mathbb{R}^{n \times m}$ where m indicates the dimension of attributes each node has. $a_{ij} = 1$ if there is an edge between nodes v_i and v_j and $s_{ij} = 1$ if node v_i has the j-th attribute, and 0 otherwise. Our main task of this paper is to partition the network $\mathcal{G}$ into k communities with well matched semantic interpretation, and the goal is twofold:

(1) Partition the nodes into TC communities based on network topology and must-link prior, and separate the nodes into SC clusters based on nodes content;
(2) Finding the best matching relationship between the two type communities so as to best describe and understand the practical meaning of each community.

2.1 Modeling TC Communities

In this subsection, we utilize must-link constraint to derive an accurate TC community structure. Must-link constraint is a kind of commonly used prior information, which depicts whether two nodes belong to the same community and is helpful to improve the accuracy of community structure. We random select a few of must-link constraints and denote them as $\mathcal{C}_{ml}$. The corresponding must-link constraint matrix $\mathbf{M} \in \mathbf{R}^{n \times n}$ is defined as:

$$(\mathbf{M})_{ij} = \begin{cases} 1, if\ i = j, \\ 2, if\ (v_i, v_j) \in \mathcal{C}_{ml}, \\ 0, others. \end{cases}$$

Assume the TC community membership of all nodes in the network to be $\mathbf{H} \in \mathbf{R}^{n \times k}$, and h_{iz} represents the propensity that node v_i belongs to the z-th TC community. If two nodes belong to the same community, it is often believed that they have similar community membership and close with each other in their geometrical distance. In order to keep this property, we use the following graph regularization to incorporate the must-link constraint for helping reveal the TC community structure:

$$\begin{aligned} &\min \sum_{ij} \|\mathbf{h}_i - \mathbf{h}_j\|^2 M_{ij} \\ &s.t.\ \ \mathbf{H} \geq \mathbf{0}. \end{aligned} \tag{1}$$

2.2 Modeling SC Communities

Define the semantic driven community membership to be $\mathbf{W} \in \mathbf{R}^{n \times k}$ where w_{ir} denotes the propensity that node v_i belongs to the r-th SC community. For each SC community, it carries some common semantic information which are summarized from the nodes' contents. On one hand, nodes in the same community usually have common contents. For another, if the contents of a

node are highly similar to the semantic information of one SC community, the node may belong to this SC community with a high propensity. Therefor, nodes of the similar content may have high propensity constitute one SC community. Assume the common semantic matrix to be $\mathbf{C} \in \mathbf{R}^{m \times k}$, and $\mathbf{c}_r$ is the contents distribution of community r. Then for a node v_i, its propensity belonging to the r-th SC community can be written as:

$$W_{ir} = \mathbf{s}_i \cdot \mathbf{c}_r$$

where $\mathbf{s}_i$ represents the contents of node v_i.

In addition, we realize that each node has multiple contents, but only a small number of contents are relevant to each community and most of contents are background information. For this case, we adopt an l_1 norm to keep the sparse semantic interpretation of each community. Further more, in order to keep the balance of these sparse contents, it needs to impose a constraint $\sum_{r=1}^{k} \|\mathbf{c}(:,r)\|_1^2$ on $\mathbf{C}$. We can derive the SC community detection model as follows:

$$\begin{aligned} &\min \|\mathbf{W} - \mathbf{SC}\|_F^2 + \xi \sum_{r=1}^{k} \|\mathbf{c}(:,r)\|_1^2 \\ &s.t.\ \ \mathbf{C} \geq \mathbf{0}. \end{aligned} \tag{2}$$

2.3 The Unified Model: Matching TC with SC Communities

According to the above defined TC community membership $\mathbf{H}$ and SC community membership $\mathbf{W}$, we introduce a coupling matrix $\mathbf{\Lambda} \in \mathbf{R}^{k \times k}$ to measure how to match semantic information with topological communities, and simultaneously use the relationship of this three matrices to generate the observed network.

In our proposed WSCDSM framework, for any node v_i, it generates a link with node v_j based on the following rule:

(1) According to the SC community structure, node v_i has one kind of common content l with propensity w_{il};
(2) Then the l-th SC community assign its semantic information to the k-th TC community with coupling probability λ_{lk};
(3) As a result, the probability of existing a link between node v_i with common content l and node v_j of the k-th TC community is $w_{il}\lambda_{lk}h_{jk}$.

Summing over all the l and k, we derive the expect number of edge between nodes v_i and v_j is:

$$\widehat{a}_{ij} = \sum_{lk} w_{il}\lambda_{lk}h_{jk}.$$

Using the square error to measure the difference between expected and observed network, it can be further written in matrix formulation:

$$\begin{aligned} &\min \|\mathbf{A} - \mathbf{W\Lambda H}^T\|_F^2 \\ &s.t.\ \ \mathbf{W}, \mathbf{\Lambda}, \mathbf{H} \geq \mathbf{0}. \end{aligned} \tag{3}$$

By combining the model (3) with the models to derive TC community (1) and SC community (2), we obtain our proposed WSCDSM framework as follows:

$$\begin{aligned} \min\ & \|\mathbf{W}-\mathbf{SC}\|_F^2+\xi\sum_{r=1}^{k}\|\mathbf{c}(:,r)\|_1^2+\alpha\|\mathbf{A}-\mathbf{W\Lambda H}^T\|_F^2 \\ & +\frac{\mu}{2}\sum_{ij}\|\mathbf{h}_i-\mathbf{h}_j\|^2M_{ij}+\gamma\|\mathbf{\Lambda}\|_1 \\ s.t.\ & \ \mathbf{W},\mathbf{H},\mathbf{\Lambda},\mathbf{C}\geq\mathbf{0} \end{aligned} \tag{4}$$

where the parameters α and μ are, respectively, used to adjust the contribution of network topology and must-link prior. The parameter ξ and γ respectively control the sparsity of community common contents and coupling relationship.

3 Optimization

Due to the objective function in (4) is not convex with respect to $\mathbf{W}$, $\mathbf{H}$, $\mathbf{\Lambda}$ and $\mathbf{C}$, it is unreasonable to find its global minimum. Here we use an iteration algorithm to derive the update rule for each matrix by fixing other matrices.

Firstly, the update of $\mathbf{W}$ can be realized by optimizing the following W-subproblem with $\mathbf{H}$, $\mathbf{\Lambda}$ and $\mathbf{C}$ fixed:

$$\begin{aligned} &\min\ \|\mathbf{W}-\mathbf{SC}\|_F^2+\alpha\|\mathbf{A}-\mathbf{W\Lambda H}^T\|_F^2 \\ &s.t.\ \ \mathbf{W}\geq\mathbf{0}. \end{aligned} \tag{5}$$

For the problem (5), we introduce a Lagrange multiplier matrix $\mathbf{\Psi}$ for the constraint $\mathbf{W}\geq\mathbf{0}$, and set the derivative of $\mathcal{L}$ with respect to $\mathbf{W}$ to $\mathbf{0}$, we obtain:

$$2\mathbf{W}-2\mathbf{SC}-2\alpha\mathbf{AH\Lambda}^T+2\alpha\mathbf{W\Lambda H}^T\mathbf{H\Lambda}^T+\mathbf{\Psi}=\mathbf{0}.$$

Using the KKT condition $\Psi_{ik}w_{ik}=0$, we obtain the following update rule for $\mathbf{W}$:

$$W_{ik}\leftarrow W_{ik}\cdot\frac{(\alpha\mathbf{AH\Lambda}^T+\mathbf{SC})_{ik}}{(\alpha\mathbf{W\Lambda H}^T\mathbf{H\Lambda}^T+\mathbf{W})_{ik}}, \tag{6}$$

Similarly, the update rules for $\mathbf{H}$ and $\mathbf{\Lambda}$ are as follows:

$$H_{ik}\leftarrow H_{ik}\cdot\frac{(\alpha\mathbf{A}^T\mathbf{W\Lambda}+\mu\mathbf{MH})_{ik}}{(\alpha\mathbf{H\Lambda}^T\mathbf{W}^T\mathbf{W\Lambda}+\mu\mathbf{QH})_{ik}}, \tag{7}$$

$$\mathbf{\Lambda}\leftarrow\mathbf{\Lambda}\cdot\frac{\alpha\mathbf{W}^T\mathbf{AH}}{\alpha\mathbf{W}^T\mathbf{W\Lambda H}^T\mathbf{H}+\gamma\mathbf{E}}, \tag{8}$$

where $\mathbf{E}$ is a $k\times k$ matrix with all element to be 1, and $\mathbf{Q}$ is a $n\times n$ diagonal matrix ($q_{ii}=\sum_j M_{ij}$ and $q_{ij}=0$ if $i\neq j$).

As for the common content matrix $\mathbf{C}$, it is equivalent to the problem of Wang *et al.* [17]. The corresponding update rule for $\mathbf{C}$ is:

$$\mathbf{C}\leftarrow\mathbf{C}\cdot\frac{\mathbf{S}_{new}^T\mathbf{W}_{new}}{\mathbf{S}_{new}^T\mathbf{S}_{new}\mathbf{C}}, \tag{9}$$

where $\mathbf{S}_{new} = \begin{pmatrix} \mathbf{S} \\ \sqrt{\xi}\mathbf{e}_{1\times m} \end{pmatrix}$, $\mathbf{W}_{new} = \begin{pmatrix} \mathbf{W} \\ \mathbf{0}_{1\times k} \end{pmatrix}$ and $\mathbf{e}_{1\times m}$ is a row vector with all elements equal to 1, $\mathbf{0}_{1\times k}$ is a zero vector.

4 Experimental Results

We evaluate our WSCDSM framework on several real networks with well known communities to validate its accuracy of community detection, and on an online music system *Last.fm* to visualize the semantic information of communities.

4.1 The Performance of Community Detection

The real networks used in the experiments are shown in Table 1.

Table 1. Some real-world networks used.

Dataset	Nodes (n)	Edges (e)	Attributes (m)	Communities (k)
Cora	2708	5429	1433	7
Citeseer	3312	4732	3706	6
Texas	187	328	1703	5
Cornell	195	304	1703	5
Washington	230	446	1703	5
Wisconsin	265	530	1703	5

The Cora and Citeseer networks are both paper citation networks with nodes representing publications and edges denoting that one publication is cited by the other publication. The other four networks are all webpage citation networks where nodes representing webpages gathered from four different universities and edges denoting that one webpage is cited by the other webpage. The node attributes of all six networks are binary-valued word attributes indicating whether each word in the vocabulary is present (indicated by 1) or absent (indicated by 0).

In order to validate the efficiency of prior information and content information for community detection, we compare with the following four types of methods: the first type is only topology-based SNMF method [18]; the second type is only attribute-based SMR method [19] and the third type is two methods based on both network topology and node content, including SCI [17] and NEMBP [20]. In addition, we also compare with one method extracted from our WSCDSM framework, but it ignores the coupling matrix and only combines with must-link constraint. This method is denoted as MLNMF.

In the specific experiments, the number of communities is set to be the same as the ground truth specified. During each experiment, we iterate 2000 times and run 20 times. As for the parameter setting, we set $\alpha = 10$, $\mu = 20$, $\xi = 100$, $\gamma = 5$

for Cora and Citeseer networks and $\gamma = 0.5$ for the other four small networks. For the comparative methods, their parameters are set to be their default values.

In this paper, we only focus on the detection of disjoint community structure, and adopt the normalized mutual information (NMI) and accuracy (AC) to measure the performance of all methods against the ground truth. The results of our WSCDSM framework as well as other 5 comparative methods on Cora and Citeseer networks are shown in Tables 2 and 3, and on the remaining networks are shown in Figs. 1 and 2. From the Tables 2 and 3 and Figs. 1 and 2, we find that due to the sparsity of network and vagueness of community structure, the method only based on topology (SNMF) or content (SMR) almost fail to accurately identify its community structure. However, the detection accuracy can be further improved by integrating both topology and content. In our WSCDSM framework, we believe that the content and topology don't share the same community structure, and on the basis of using few semi-supervised prior to improve the accuracy of community detection, content information can be more effectively utilized to make up for the insufficiency of topology. Therefore, WSCDSM framework outperforms the other five comparative methods on most of networks, especially for Cora and Cornell networks, the improvement is more obvious. Although the randomness of prior information causes that the results of WSCDSM are not always higher than NEMBP on Wisconsin network, it will achieve superior performance when proper prior information is integrated.

Table 2. Comparative results in terms of NMI, and the best results are in bold.

Information used	Method	Cora			Citeseer		
Only topology	SNMF	0.1994			0.0403		
Only content	SMR	0.0078			0.0032		
Topology+Content	SCI	0.1780			0.0922		
	NEMBP	0.4408			0.2427		
Topology+Prior	MLNMF	2%	5%	8%	2%	5%	8%
		0.3159	0.3239	0.3451	0.2664	0.278	0.3081
Topology+Prior+Content	WSCDSM	**0.5254**	**0.7522**	**0.8083**	**0.3532**	**0.4297**	**0.4435**

Table 3. Comparative results in terms of AC, and the best results are in bold.

Information used	Method	Cora			Citeseer		
Only topology	SNMF	0.4173			0.2539		
Only content	SMR	0.3002			0.2111		
Topology+Content	SCI	0.4169			0.3442		
	NEMBP	0.5757			0.4951		
Topology+Prior	MLNMF	2%	5%	8%	2%	5%	8%
		0.4088	0.4106	0.4387	0.4109	0.4233	0.4598
Topology+Prior+Content	WSCDSM	0.5373	**0.7692**	**0.7906**	0.4761	**0.5136**	**0.5444**

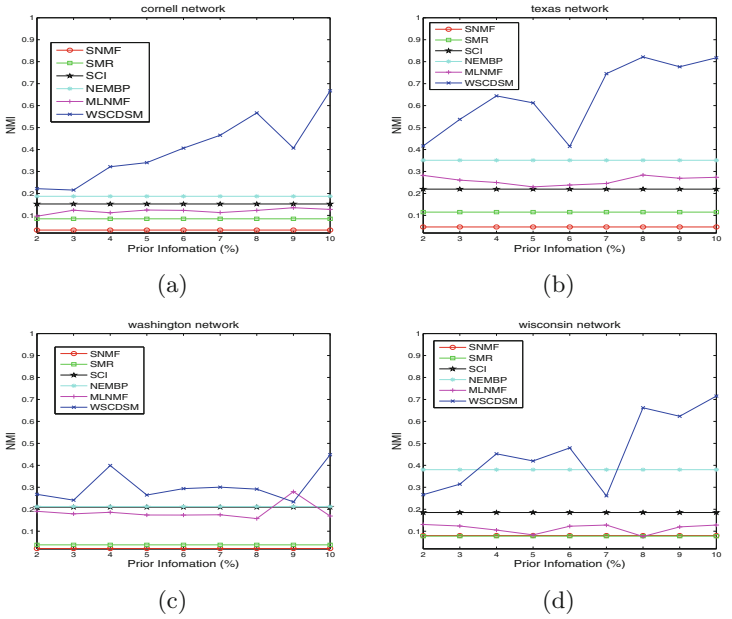

Fig. 1. Comparative results in terms of NMI on (a) Cornell network; (b) Texas network; (c) Washington network; (d) Wisconsin network.

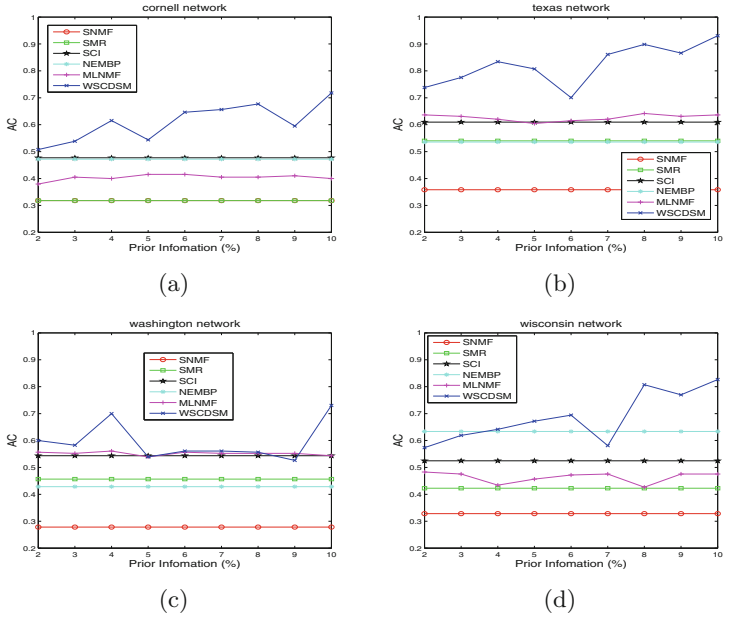

Fig. 2. Comparative results in terms of AC on (a) Cornell network; (b) Texas network; (c) Washington network; (d) Wisconsin network.

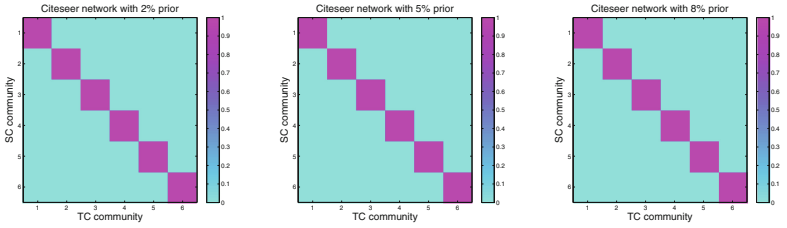

Fig. 3. The coupling relationship between TC and SC community structure on Citeseer network. (a) 2% prior used; (b) 5% prior used; (c) 8% prior used.

Based on the above results, we believe that the higher detection accuracy of TC community structure, the better matching of TC communities and semantic information. Due to the limited space, here we only take Citeseer network for an example to present the better matching between TC and SC community structure, as shown in Fig. 3. We find that each community has different semantic explanations with each other, and the semantic matching is robust to the increase of prior information.

4.2 The Matching Between Semantic and Communities

The *Lsat.fm* system contains 1892 users, and each user has 11,946 dimensional contents, including a list of most-listened-musical to artists and tag assignments, i.e. [user, tag, artist] tuples. Due to the *Lsat.fm* network has no ground truth with respect to the community label of node, we use *Louvain* method [3] as did in Wang *et al.* [17], but we set the number of communities to be 31, and the corresponding community structure is regarded as the ground truth.

The coupling relationship and semantic information of some communities are presented in Fig. 4. From the Fig. 4(a), we find that our WSCDSM framework can match most TC communities with one specific semantic topic, and only several TC communities have two or three semantic topics. Besides, there are also few communities that they have no semantic topic, which demonstrates the content information of such communities maybe background words. Figure 4(b) depicts a community of only one topic related to Britney Spears, a legend and amazing singer in Louisiana, USA. Her music often has characteristics of "pop", "dance", "rnb" and "electronic". An example community of two topics are shown in Fig. 4(c), this community is composed by a group fans who like "rock" and "heavy metal" two styles of music, and among which the style of "rock" music contains hard rock, classic rock and progressive rock. A community of three topics are illustrated in Fig. 4(d), which is characterized by three types of music including "synthpop", "new wave" and "electronic". For these three types music, Depeche Mode, a representative band, is very popular and active in British. Based on the above analysis, we find that our WSCDSM framework can relatively accurately match the community structure and semantic information.

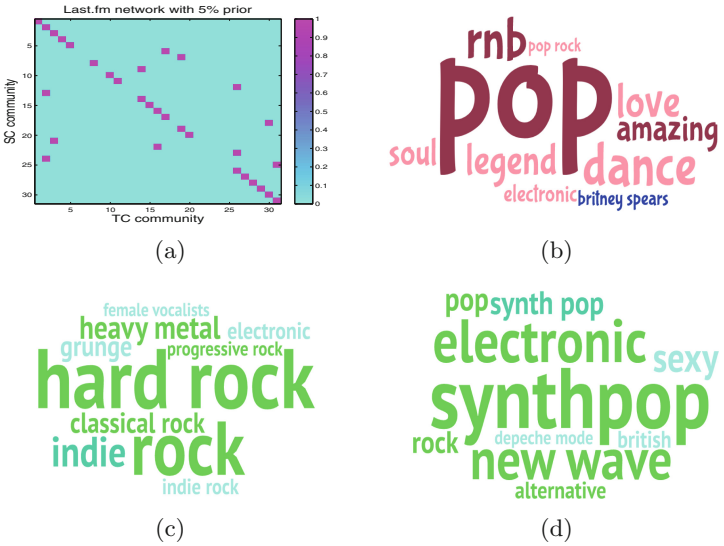

Fig. 4. The matching relationship between SC and TC community, as well as some examples of community interpretation on *Lsat.fm* network. (a) coupling relationship; community with (b) one topic; (c) two topics; (d) three topics.

5 Conclusion

In this paper, we proposed a unified weakly supervised framework simultaneously for community detection and semantic matching. In our framework, the semi-supervised information is firstly utilized to improve the community accuracy. Then by introducing a coupling matrix, the node content information is used to adjust the TC community structure and simultaneously match semantic interpretation for each community. The results on several real networks demonstrated that, for one thing, integrating with few percentage of must-link prior our framework can improve the accuracy of community detection. For another, under the guidance of coupling matrix, the TC community and SC community structure can realize fully interaction with each other, and further derive a well semantic description for communities.

Acknowledgments. This work was supported by the Major Project of National Social Science Fund(14ZDB153), the major research plan of the National Natural Science Foundation (91746205,91746107,91224009,51438009), and the Natural Science Foundation of China (61772361).

References

1. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
2. White, S., Smyth, P.: A spectral clustering approach to finding communities in graphs. In: Proceedings of the 2005 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp. 274–285 (2005)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), P10008 (2008)
4. Zhang, S., Zhao, H.: Normalized modularity optimization method for community identification with degree adjustment. Phys. Rev. E **88**(5), 052802 (2013)
5. He, D., Liu, D., Jin, D., Zhang, W.: A stochastic model for the detection of heterogeneous link communities in complex networks. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Palo Alto, California, USA, pp. 130–136. AAAI Press (2015)
6. Eaton, E., Mansbach, R.: A spin-glass model for semi-supervised community detection. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 900–906 (2012)
7. Ma, X., Gao, L., Yong, X.: Semi-supervised clustering algorithm for community structure detection in complex networks. Physica A: Stat. Mech. Appl. **389**(1), 187–197 (2010)
8. Yang, L., Cao, X., Jin, D.: A unified semi-supervised community detection framework using latent space graph regularization. IEEE Trans. Cybern. **45**(11), 2585–2598 (2015)
9. Zhang, Z.Y.: Community structure detection in complex networks with partial background information. EPL (europhys. lett.) **101**(4), 48005 (2013)
10. Liu, X., Wang, W., He, D.: Semi-supervised community detection based on nonnegative matrix factorization with node popularity. Inf. Sci. **381**, 304–321 (2017)
11. Zhang, Z.Y., Sun, K.D., Wang, S.Q.: Enhanced community structure detection in complex networks with partial background information. Sci. Rep. **3**, 3241 (2013)
12. Ruan, Y., Fuhry, D., Parthasarathy, S.: Efficient community detection in large networks using content and links. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1089–1098. ACM (2013)
13. Pool, S., Bonchi, F., Leeuwen, M.: Description-driven community detection. ACM Trans. Intell. Syst. Technol. (TIST) **5**(2), 28 (2014)
14. Yang, T., Jin, R., Chi, Y.: Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 927–936. ACM (2009)
15. Pei, Y., Chakraborty, N., Sycara, K.: Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
16. Liu, L., Xu, L., Wangy, Z.: Community detection based on structure and content: a content propagation perspective. In: 2015 IEEE International Conference on Data Mining (ICDM), pp. 271–280. IEEE (2015)
17. Wang, X., Jin, D., Cao, X.: Semantic community identification in large attribute networks. In: AAAI, pp. 265–271 (2016)
18. Wang, F., Li, T., Wang, X.: Community discovery using nonnegative matrix factorization. Data Min. Knowl. Discov. **22**(3), 493–521 (2011)

19. Hu, H., Lin, Z., Feng, J.: Smooth representation clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3834–3841 (2014)
20. He, D., Feng, Z., Jin, D.: Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)