Community Detection in Multi-relational Social Networks

Zhiang Wu[1], Wenpeng Yin[1], Jie Cao[1,⋆], Guandong Xu[2], and Alfredo Cuzzocrea[3]

[1] Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China
[2] Advanced Analytics Institute, University of Technology, Sydney, Australia
[3] Institute of High Performance Computing and Networking, Italian National Research Council, Italy
{zawuster,mr.yinwenpeng}@gmail.com, Jie.Cao@njue.edu.cn, Guandong.Xu@uts.edu.au, cuzzocrea@icar.cnr.it

Abstract. Multi-relational networks are ubiquitous in many fields such as bibliography, twitter, and healthcare. There have been many studies in the literature targeting at discovering communities from social networks. However, most of them have focused on single-relational networks. A hint of methods detected communities from multi-relational networks by converting them to single-relational networks first. Nevertheless, they commonly assumed different relations were independent from each other, which is obviously unreal to real-life cases. In this paper, we attempt to address this challenge by introducing a novel co-ranking framework, named *MutuRank*. It makes full use of the mutual influence between relations and actors to transform the multi-relational network to the single-relational network. We then present GMM-NK (Gaussian Mixture Model with Neighbor Knowledge) based on local consistency principle to enhance the performance of spectral clustering process in discovering overlapping communities. Experimental results on both synthetic and real-world data demonstrate the effectiveness of the proposed method.

Keywords: Social Networks, Community Detection, Multi-relational Network, MutuRank, Gaussian Mixture Model.

1 Introduction

Community detection has become a fundamental yet difficult task ever since the network science came into vogue. What is the nature of network communities? So far, there is no standard answer to this question [23]. In general, actors in a same community tend to interact with each other more frequently than with those outside the community. The communities are also called *groups*, *clusters*, *cohesive subgroups* or *modules* in different research fields [21].

⋆ Corresponding author.

X. Lin et al. (Eds.): WISE 2013, Part II, LNCS 8181, pp. 43–56, 2013.

Although a large body of research efforts have been devoted to community detection [4,8,21], most of the existing methods are designed for *single-relational* networks. This kind of network is composed of a set of nodes (i.e., objects) connected by a set of edges (i.e., links) which represent relationships of a single type. Whereas, in many real-world situations, objects are usually associated with each other in multiple aspects. For example, in Twitter, users could be followers/followees of others, could retweet tweets of others, could produce topic relevant tweets with others and etc. Considering further the relationships among scholars in DBLP, we could treat co-authorship, citation and venue as distinct relation types between scholars.

In some literatures [19,20], multi-relational networks (a.k.a. *heterogeneous* or *multi-mode* networks) often contain more than one typed entities. However, the meaningful communities are still defined on the same typed entities. So, in this paper, we limit our scope to the multi-relational network containing multiple typed relations but one typed entities. Especially to deserve to be mentioned, the multi-relational network considered in this paper is a special case of the multi-mode network when only one typed entities are considered.

To date, general methods handling a multi-relational network, such as [1,14,20] and etc., first try to convert it to a single-relational network, and then employ existing methods for community detection. However, such approaches usually suppose that multiple typed relations are independent, which is not the case in real situations. Taking a look at the scholar circle, two researchers may have relevant research interest, co-author several papers, co-operate several projects, and even publish papers in the same conferences. How can we ignore an intuition that scholars with similar academic backgrounds are more likely to publish papers in same venues? How can we neglect the tendency that persons who co-direct a project are very likely to co-author some literature?

Motivated by that, in this work, we propose a novel co-ranking framework, *MutuRank*, to determine the weights of various relation types and objects simultaneously. The essential part of this framework lies in that it makes full use of the mutual influence between those relations and actors, with the aim of deriving equilibrium/stationary probability distributions as evaluation scores for actors and relations, respectively. To be specific, the mutual-feedback here means that: (i) the importance of a relation depends on the probability distribution of actors and the importance of other relations, i.e., a relation, selected by high-weight actors with high probabilities, deserves high weight itself; (ii) the importance of an actor depends on the probability distribution of relations and its neighbors' importance, i.e., an actor, linked by high-weight actors with strong and high-weight relations, deserves high-weight. Here, *strong relation* implies the intense closeness of two objects under that relation, and *high-weight relation* indicates that the relation type itself is very important. The probability distributions derived from such mutual-feedback are able to convey the intrinsic status of relations and objects more accurately.

We then combine the probability distributions of relations linearly to produce a single-relational network just as some typical literature did. Although

any of the existing methods can be used to partition the network into crisp communities, this paper focus on discovering overlapping communities by which most real networks are characterized [11,22]. For instance, a scientist might belong to an academic community as well as some personal life communities (e.g., school, hobby, family). In this paper, we propose a novel soft clustering method named *GMM-NK* (Gaussian Mixture Model with Neighbor Knowledge), which originates from this inspiration: the probability of an object belonging to a community could be derived not only via its own gaussian value, but also from the probabilities of its near neighbors belonging to the same community. This idea is also in accordance with *the local consistency principle* of machine learning in which very related objects should have similar category attributes.

Finally, we perform experiments on simulated synthetic data as well as real-world DBLP dataset. Experiments results on both datasets validate the good performance of our proposed community detection algorithm.

The remainder of this paper is organized as follows. In Section 2, we present the related work. In Section 3, we introduce *MutuRank* for identification of relation distribution. In Section 4, we show how to mine communities using soft clustering method. Experimental results will be given in Section 5. We finally conclude this paper in Section 6.

2 Related Work

Multi-relational social network has attracted much attention in the few years. A great many studies attempted to integrate latent [18] or explicit [14,3,1] heterogeneous relations to form a single-relational network. In almost all of these studies, different relations are considered to be independent from each other. However, from a case study [16] conducted on a heterogeneous network consisting of about 300,000 game players with six different relations including three positive and three negative ones, a conclusion was drawn that positive links were highly reciprocal while negative links were not. It is somewhat consistent with our motivation that various relations interact with each other. Our *MutuRank* is greatly inspired by the *MultiRank* framework presented in [9]. *MultiRank* employed a PageRank-like [10,24] random walk model to co-rank objects and relations in a multi-relational network. Whereas, *MultiRank* assumed that the probability for a node to select a relation is not only independent of the node's importance, but also independent of the network structure. This assumption is not strictly rational, and is remedied by our *MutuRank* model.

Community detection has been extensively studied [4]. So far, most of the existing methods can be classified into two main categories, in terms of whether or not explicit optimization objectives are being used. The methods with explicit optimization objectives typically consider the global topology of a network, and aim to optimize a criterion defined over a network partition. Some methods along this line include the Kernighan-Lin algorithm [7], stochastic block models [5], modularity optimization [8], and traditional clustering techniques [15] such as K-means, multi-dimensional scaling (MDS), and spectral clustering. On

the other hand, the methods without using explicit optimization objectives discover communities based on predefined assumptions or heuristic rules. For example, Clique Percolation Method (CPM) [11] is based on the concept of k-clique, and a k-clique community is then defined as the union of all "adjacent" k-cliques, which by definition share k-1 nodes. To sum up, most of the methods using explicit optimization objectives often lead to crisp partitions, while our GMM-NK employs the spectral clustering to carry out soft community detection.

It is worthy of mentioning some work having the same target with this paper, i.e., discovering communities from multi-relational networks. Cai et al. [1] proposed a regression-based algorithm to learn the optimal relation weights, and then utilized threshold cut as the optimization objective for community detection. Tang et al. [20] proposed several integration strategies including network integration, utility integration, feature integration and partition integration for transforming the multi-relational network to single-relational network. However, they failed to consider the mutual influence between relations and actors.

3 Identification of Relation Distribution

In this section, we aim to introduce our *MutuRank* algorithm to identify the relation distribution, and thus show how to transform the multi-relational network into a single-relational graph.

Let's begin by introducing some notation conventions. Let $\mathcal{N}$ denote the node set with n elements, and use i or j to represent the index of a random node, hence, $1 \leq i, j \leq n$. Let $\mathcal{R}$ denote the relation set with totally m relation types $\{k|1 \leq k \leq m\}$ where k is the relation index. Further, let R be the real field, and represent the affinity tensor as $\mathcal{S} = (s_{i,j,k})$ where $s_{i,j,k}$ $(s_{i,j,k} \in R)$ denotes the relation strength between nodes i and j under the k-th relation type. In most cases, *relation strength* is also treated as a kind of *similarity*. Additionally, vectors $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ and $\mathbf{q} = (q_1, q_2, \cdots, q_m)$ denote the probability distributions of nodes and relation types, respectively.

3.1 Co-ranking Nodes and Relations in Multi-relational Network

As mentioned earlier, the rationale of our *MutuRank* framework is that the weight of a node is affected not only by its neighbors' weights, but also by the strength of various link types which are also endowed with different weights. The more important neighbors and more strong links with high importance, the more important a node will be. Accordingly, given a node i, the probability for information transiting from i's neighbor j to i is as follows:

$$\mathrm{Prob}_1(i|j) = \frac{\sum_k q_k \cdot s_{i,j,k}}{\sum_l \sum_k q_k \cdot s_{l,j,k}}, \tag{1}$$

where q_k means the current weight of the k-th relation, and $\mathrm{Prob}_1(\cdot|\cdot)$ represents the probability of a node selecting another node, just as the the random walk

in PageRank. Above equation demonstrates that the transition probability between two nodes does not keep unchanged, which is different with the traditional PageRank. It is easy to understand that if the probability distribution of relations is consistent with the arrangement of link strengthes between two nodes, this pair of nodes may have relatively higher transition probability.

Furthermore, based on the probability distribution of nodes, the probability of the node i selecting the k-th relation is adjusted as:

$$\text{Prob}_2(k|i) = \frac{q_k \cdot \sum_j s_{i,j,k}}{\sum_k q_k \cdot \sum_j s_{i,j,k}}, \tag{2}$$

where $\text{Prob}_2(\cdot|\cdot)$ represents the probability of a node selecting a relation type. It is worth mentioning that the *MultiRank* [9] has a similar objective with us. However, *MultiRank* relaxed the joint probability of a node and a relation to be two mutually independent probabilities, which leads to the probability of a node selecting a relation has nothing to do with the similarity structure of the whole network. Let's illustrate it by a simple example: we assume the k-th relation has gotten a very high weight in previous iterations, while node i has very low similarities with all of its neighbors. This situation might probably happen when almost all of node pairs, except node i and its neighbors, have higher mutual similarities under relation k. *MultiRank* still assumed the probabilities of all nodes selecting relation k in the next iteration to be equal, which is obviously unfair to the node i. So, we believe that node i should reduce its probability of selecting relation k based on its true similarities to its neighbors.

Let $\mathbf{p}^* = (p_1^*, p_2^*, \cdots, p_n^*)$ and $\mathbf{q}^* = (q_1^*, q_2^*, \cdots, q_m^*)$ be the prior distributions of nodes and relations, respectively. To conclude, we have following iterative equations for computing the ranking scores of nodes and relations simultaneously:

$$p_i^{t+1} = \sum_j p_j^t \cdot \text{Prob}_1^t(i|j) + \alpha \cdot p_i^* \quad (1 \le i \le n), \tag{3}$$

$$q_k^{t+1} = \sum_i p_i^t \cdot \text{Prob}_2^t(k|i) + \beta \cdot q_k^* \quad (1 \le k \le m), \tag{4}$$

where t is the times of iteration, and α, β are two parameters to balance the knowledge coming from network structure and the prior knowledge.

3.2 Theoretical Analysis

In this section, we prove the existence and uniqueness of stationary probability distributions $\mathbf{p}$ and $\mathbf{q}$ so that they can be used to co-rank the nodes and relation types effectively.

First, we show why our iterative algorithm in Eqs.(3) and (4) will converge. Let $\Omega_n = \{\mathbf{p} = (p_1, p_2, \cdots, p_n) \in R^n | p_i \ge 0, 1 \le i \le n, \sum_{i=1}^n p_i = 1\}$ and $\Omega_m = \{\mathbf{q} = (q_1, q_2, \cdots, q_m) \in R^m | q_k \ge 0, 1 \le k \le m, \sum_{k=1}^m q_k = 1\}$. We also set $\Omega = \{[\mathbf{p}, \mathbf{q}] \in R^{n+m} | \mathbf{p} \in \Omega_n, \mathbf{q} \in \Omega_m\}$. We notice that Ω_n, Ω_m and Ω are

closed convex sets. We call $\mathbf{p}$ and $\mathbf{q}$ to be positive if all their entries are greater than 0. For convenience, we represent Eqs.(3) and (4) as following simpler form:

$$\mathbf{p}^{t+1} = f_1(\mathbf{p}^t, \mathbf{q}^t), \tag{5}$$

$$\mathbf{q}^{t+1} = f_2(\mathbf{p}^t, \mathbf{q}^t). \tag{6}$$

We then have the following theorem:

Theorem 1. *For any* $\mathbf{p}^t \in \Omega_n$ *and* $\mathbf{q}^t \in \Omega_m$, *then* $f_1(\mathbf{p}^t, \mathbf{q}^t) \in \Omega_n$ *and* $f_2(\mathbf{p}^t, \mathbf{q}^t) \in \Omega_m$

Proof. With Eq. (3), it is easy to prove that if $\mathbf{p}^t$, $\mathbf{q}^t$ and $\mathbf{p}^*$ are probability distributions, $\mathbf{p}^{t+1}$ is also a probability distribution. Similarly, if $\mathbf{p}^t$, $\mathbf{q}^t$ and $\mathbf{q}^*$ are probability distributions, $\mathbf{q}^{t+1}$ is also a probability distribution according to Eq. (4). □

On the basis of Theorem 1, we next attempt to show the existence of positive solution for MutuRank. But before that, it is necessary for us to know the connectivity among the objects and the relations within that multi-relational network. We first give a definition:

Definition 1 (Irreducibility). $\mathcal{S} = (s_{i,j,k})$ *is called irreducible if* $s_{i,j,k}$ *(n-by-n matrices) for fixed* k *(*$1 \leq k \leq m$*) are irreducible.*

Here, $\mathcal{S}$'s irreducibility suggests that two objects can be connected via some relations. *Irreducibility* is a reasonable assumption that we will use in following discussion. It was also adopted by literature [10] in the PageRank matrix for calculating PageRank values. Further, such an assumption contributes to following conclusion:

Theorem 2. *If* $\mathcal{S} = (s_{i,j,k})$ *is irreducible, then there exist* $\bar{\mathbf{p}} \in \Omega_n$ *and* $\bar{\mathbf{q}} \in \Omega_m$ *s.t.,* $\bar{\mathbf{p}} = f_1(\bar{\mathbf{p}}, \bar{\mathbf{q}})$ *and* $\bar{\mathbf{q}} = f_2(\bar{\mathbf{p}}, \bar{\mathbf{q}})$, *and* $\bar{\mathbf{p}} > \mathbf{0}$, $\bar{\mathbf{q}} > \mathbf{0}$.

Proof. This problem can be addressed as a fixed point problem. Suppose a mapping $F : \Omega \to \Omega$ as follows:

$$F([\mathbf{p}, \mathbf{q}]) = [f_1(\mathbf{p}, \mathbf{q}), f_2(\mathbf{p}, \mathbf{q})]. \tag{7}$$

It is clear that $F(\cdot)$ is well-defined (i.e., when $[\mathbf{p}, \mathbf{q}] \in \Omega$, $F([\mathbf{p}, \mathbf{q}]) \in \Omega$) and continuous. According to the Brouwer Fixed Point Theorem, there exists $[\bar{\mathbf{p}}, \bar{\mathbf{q}}] \in \Omega$ such that $F([\bar{\mathbf{p}}, \bar{\mathbf{q}}]) = [\bar{\mathbf{p}}, \bar{\mathbf{q}}]$, i.e., $f_1(\bar{\mathbf{p}}, \bar{\mathbf{q}}) = \bar{\mathbf{p}}$ and $f_2(\bar{\mathbf{p}}, \bar{\mathbf{q}}) = \bar{\mathbf{q}}$. □

Now, we will have a deep discussion that both $\bar{\mathbf{p}}$ and $\bar{\mathbf{q}}$ are positive. Let us re-write the Eq. (4) as:

$$q_k^{t+1} = q_k^t \cdot \sum_i p_i^t \cdot \frac{\sum_j s_{i,j,k}}{\sum_k q_k^t \cdot \sum_j s_{i,j,k}} + \beta \cdot \mathbf{q}^*. \tag{8}$$

If $q_k^T = 0$ (i.e., after the $T-$th iteration, the importance of the k-th relation type becomes 0), then, $\sum_j s_{i,j,k} = 0$. Note that $\sum_j s_{i,j,k}$ means the sum of the similarities between node i and all its neighbors. Apparently, $\sum_j s_{i,j,k} = 0$ indicates that node i is an isolated node under the k-th relation, which is in contradiction with the irreducibility of $\mathcal{S}$. Similarly, given the Equation (3), if $p_i^T = 0$, we can easily find that the resulting situation also violates the irreducibility of $\mathcal{S}$. Due to the space limit, we do not provide any detailed analysis.

Finally, we give the conditions under which our algorithm will converge to a unique solution. Literature [6] has guaranteed the uniqueness of the fixed point in the Brouwer Fixed Point Theorem with following prerequisites: (i) for each point in the domain boundary of the mapping, it is not a fixed point; (ii) 1 is not an eigenvalue of the Jacobian matrix of the mapping. As for the first condition, we have shown, in Theorem 2, that all the fixed points of $F(\cdot)$ are positive when $\mathcal{S}$ is irreducible, i.e., they do not lie on the boundary $\partial\Omega$ of Ω. As regards the second condition, we have following conclusion:

Theorem 3. *If* 1 *is not the eigenvalue of the Jacobian matrix of mapping* F *for all* $[\mathbf{p}, \mathbf{q}] \in \Omega/\partial\Omega$*, the probability distributions in Theorem 2 are unique.*

Note that Theorem 3 presents only a condition for the solution's uniqueness. How to prove or guarantee 1 is not the eigenvalue of a function's Jacobian matrix remains an open problem. Nevertheless, it does not affect *MutuRank*'s convergence and real-world application.

Up to now, we have elaborated how to calculate the relation distribution via our *MutuRank* algorithm. Based on the results of this step, we keep nodes unchanged and merge those pairwise relations linearly to construct a single-relational graph. Next, we utilize spectral clustering to perform community detection on our graph data.

4 Community Detection in Single-Relational Network

In this part, we exploit the widely-used spectral clustering framework to conduct clustering in single-relational graph. The rationale of spectral clustering is as follows: first represents graph nodes with vectors by means of some matrix operations over the graph's affinity matrix, then calls certain basic clustering algorithm, such as K-means or GMM (Gaussian Mixture Model), to do clustering. GMM should be more suitable than K-means in clustering objects in social network because most entities are unlikely interested in only one community.

In many real-world social networks, actors tend to exert influence to their friends. For instance, in a scientific coauthorship network, if most of a researcher's friends have much interest in data mining, it is reasonable to infer the researcher himself/herself is very likely interested in data mining too. When it turns to applying GMM to cluster objects in social networks, we believe that the probability of an object belonging to a community is decided not only by the value of Gaussian function of the object itself, but also according to the probabilities of the object's neighbors belonging to that community. Indeed, some literatures [12,13]

aimed to assign a node with the label that most of its neighbors have, so that a consensus on a label will finally form a community. Hence, we attempt to modify the traditional GMM algorithm to be a novel model named GMM-NK(Gaussian Mixture Model with Neighbor Knowledge).

Similar with GMM, GMM-NK is also a linear superposition of Gaussian components for the purpose of providing a richer class of density models than the single Gaussian. Clustering based on GMM-NK is probabilistic in nature and aims at maximizing the likelihood function with regard to the parameters (comprising the means and covariances of the components and the mixing coefficients).

Consider n data points $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ in d-dimensional space, the probability density of x_i can be defined as follows:

$$p(x_i|\pi, \mu, \Sigma) = \sum_{z=1}^{c} \pi_z \cdot N^*(x_i; \mu_z, \Sigma_z), \tag{9}$$

where c is the component number, π_z is the prior probability of the z-th Gaussian component. $N^*(x_i; \mu_z, \Sigma_z)$ is defined as:

$$N^*(x_i; \mu_z, \Sigma_z) = \gamma \cdot N(x_i; \mu_z, \Sigma_z) + (1-\gamma) \cdot \sum_{j:j\neq i} s_{i,j} \cdot N(x_j; \mu_z, \Sigma_z), \tag{10}$$

where $0 < \gamma \leq 1$ is a parameter for controlling the impacts from neighbors, and $s_{i,j}$ is the similarity between node x_i and x_j. $\gamma = 1$ means that we do not consider the influence of neighbors and this algorithm reduces to the original GMM. Let $N(x_i; \mu_z, \Sigma_z)$ denote the standard Gaussian function, i.e.,

$$N(x_i; \mu_z, \Sigma_z) = \frac{\exp\{-\frac{1}{2}(x_i - \mu_z)^T \sum_z^{-1}(x_i - \mu_z)\}}{((2\pi)^d |\Sigma_z|)^{\frac{1}{2}}}. \tag{11}$$

Similar with the basic GMM, the log of the likelihood function is then given by:

$$\begin{aligned} lnP(\mathcal{X}|\pi, \mu, \Sigma) &= ln \prod_{i=1}^{n} P(x_i|\pi, \mu, \Sigma) = ln \prod_{i=1}^{n} \sum_{z=1}^{c} \pi_z \cdot N^*(x_i; \mu_z, \Sigma_z) \\ &= \sum_{i=1}^{n} ln\{\sum_{z=1}^{c} \pi_z \cdot N^*(x_i; \mu_z, \Sigma_z)\}. \end{aligned} \tag{12}$$

An elegant and powerful method for finding maximum likelihood solutions for models with latent variables is Expectation Maximization algorithm (EM). EM is an iterative algorithm in which each iteration contains an E-step and a M-step. In the E-step, we compute the probability of the z-th Gaussian component given the data point x_i using the current parameter values:

$$p(z|x_i, \pi, \mu, \Sigma) = \frac{\pi_z \cdot N^*(x_i; \mu_z, \Sigma_z)}{\sum_{j=1}^{c} \pi_j \cdot N^*(x_i; \mu_j, \Sigma_j)}. \tag{13}$$

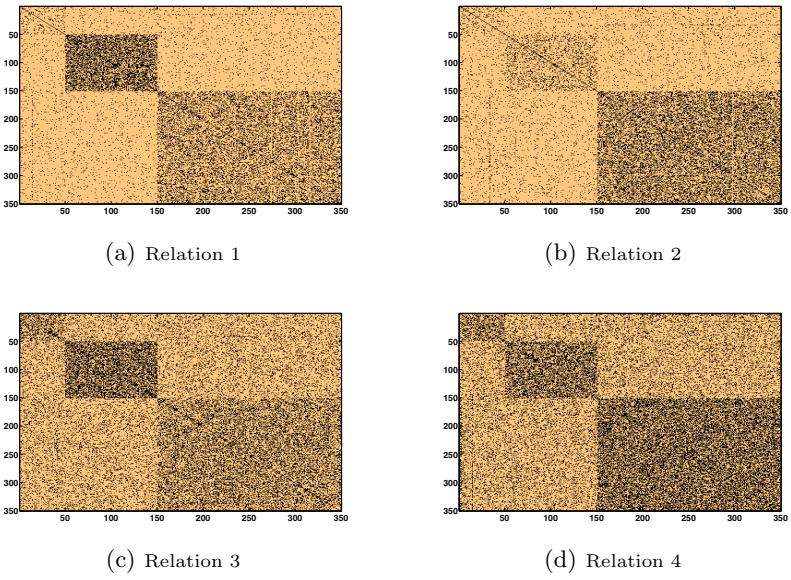
In the M-step, we re-estimate the parameters using the current responsibilities, as follows:

$$\mu_z^{new} = \frac{1}{n_z} \sum_{i=1}^{n} x_i \cdot p(z|x_i, \pi, \mu, \Sigma), \tag{14}$$

$$\Sigma_z^{new} = \frac{1}{n_z} \sum_{i=1}^{n} p(z|x_i, \pi, \mu, \Sigma)(x_i - \mu_z^{new})(x_i - \mu_z^{new})^T, \tag{15}$$

$$\pi_z^{new} = \frac{n_z}{n}, \tag{16}$$

where $n_z = \sum_{i=1}^{n} p(z|x_i, \pi, \mu, \Sigma)$. The EM algorithm runs iteratively until the log likelihood reaches (approximate) convergence. In experiments, we simply run GMM-NK until the increment of log likelihood is less than 10^{-7} and pick the one with largest log likelihood as the best estimate of the underlying communities.



(a) Relation 1 (b) Relation 2

(c) Relation 3 (d) Relation 4

Fig. 1. The four relation networks on the synthetic dataset

5 Experimental Validation

In this section, we demonstrate the effectiveness of our *MutuRank* and *GMM-NK* algorithms for community detection in multi-relational social networks. For the sake of simplicity, $\alpha = \beta = 0.5$ in Eqs.(3) and (4), and $\gamma = 0.65$ in Eq.(10) are the default settings in these experiments.

5.1 Experiments on Synthetic Dataset

Generally, real-world corpus does not provide the ground truth information about the membership of component objects. So, we start with a synthetic dataset to illustrate some good properties of our framework. This dataset is generated by a synthetic data simulator coded in MATLAB [20]. The synthetic network contains 350 nodes which roughly form 3 communities containing 50, 100, 200 nodes, respectively. Furthermore, 4 different relations are constructed to look at the clustering structure in different angles, as shown in Fig. 1.

Since a priori community memberships (a.k.a. ground truth) is known, we then adopt commonly used *normalized mutual information* (NMI) [2] as the evaluation measure. Let L and G denote the label vector obtained by community detection methods and the ground truth. NMI is defined as:

$$NMI(L;G) = \frac{I(L;G)}{\sqrt{H(L)H(G)}}, \tag{17}$$

where $I(L;G)$ is the mutual information of two variables.

Apart from our proposed GMM-NK, basic GMM and K-means algorithms are both implemented as baselines. Fig. 2 shows the overall comparison on the synthetic dataset.

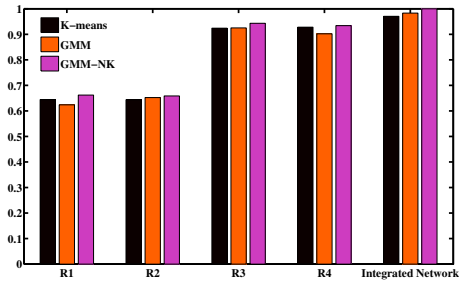


Fig. 2. Performance comparison on the synthetic dataset

Note that aggregated bars on R_1 to R_4 correspond to the results on 4 different relations, and "Integrated Network" corresponds to the results on the single relation network obtained by our $MutuRank$. As can be seen, NMI on the integrated network outperforms any of single relations, regardless of community detection algorithms. That implies that $MutuRank$ can incorporate multi-relations to form a good quality single-relation network, in which community structures emerge more clearly than that on any original single-relation graphs. Furthermore, we observe that our GMM-NK indeed shows better performance than both GMM and K-means, and combining $MutuRank$ with GMM-NK achieves perfect clustering, i.e., $NMI \approx 1$ on integrated network of GMM-NK.

5.2 Experiments on DBLP Dataset

We now proceed to provide experimental results on the real-world dataset, i.e., the DBLP dataset. We first discuss data collection and graph construction, and then report and analyze experimental findings.

Data Collection. According to the rankings of authoritative conferences in computer science[1], we crawled publication information of both "Top Tier" and "Second Tier" conferences of 13 different fields from the DBLP web site. Table 1 shows the legends of these 13 different fields. Note that all publication periods are from 2000 to 2010. In total, the extracted DBLP dataset contains 97 conferences, 185,490 researchers and 105,264 publications.

Table 1. The Legends of 13 Fields

Categories	Explanations	Categories	Explanations
DB	Databases	DP	Distributed and Parallel Computing
DM	Data Mining	GV	Graphics, Vision and HCI
AI	Artificial Intelligence	MM	Multimedia
NL	Natural Language Processing	NC	Networks, Communications and Performance
ED	Computer Education	SE	Security and Privacy
IR	Information Retrieval	OS	Operating Systems/Simulations
W3	Web and Information Systems		

Graph Construction. We treat researchers as nodes in network, and treat fields as distinct relation types. Then, for each relation k ($1 \leq k \leq 13$), we construct a *relation-graph* G_k where the link strength of two researchers i and j is calculated according to their publications in that field. More exactly, suppose the numbers of their publications in relation k are $p_{i,k}$ and $p_{j,k}$, respectively. Then their relevance can be determined by Eq. 18.

$$s_{i,j,k} = e^{-(\frac{2(p_{i,k}-p_{j,k})}{p_{i,k}+p_{j,k}})^2}. \tag{18}$$

Baselines. In whole, as for relation distribution, we have three choices, i.e., (1) *Uniform*: setting to uniform distribution; or (2) *MultiRank*: using the MultiRank [9] algorithm; or (3) *MutuRank*: using our *MutuRank* algorithm. With respect to spectral clustering, it could be implemented on the basis of K-means, GMM, or our *GMM-NK*. Hence, we combine these choices to obtain 9 baselines for experimental comparison.

Experiment Results. First, we give the relation distribution derived from *MutuRank* algorithm, as shown in Fig. 3. Relation types "AI" and "DB" enjoy apparent dominance, while "ED" is the most unimportant one. It results from the tendency that researchers in "AI" and "DB" areas not only have huge

[1] http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html

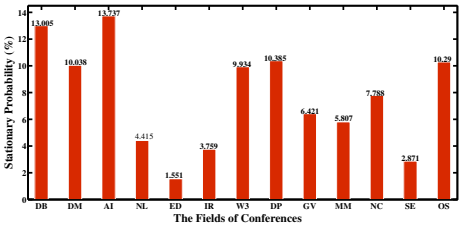


Fig. 3. The stationary relation distribution

Table 2. Performance Comparison on DBLP Dataset

NMI	K-means	GMM	GMM-NK
Uniform	0.618	0.622	0.687
MultiRank	0.757	0.762	0.776
MutuRank	0.824	0.871	**0.913**

amount, but also own rich publications. Contrarily, "ED" and "SE" only attract a few scholars to dedicate. Understandably, the more the high-weight edges in a dimension, the more important the relation type.

Second, we investigate the overall performance of afore-mentioned 9 baselines on DBLP dataset. Though K-means generates crisp communities, two soft clustering methods GMM and GMM-NK output the probabilities that every researcher belonging to all communities. It is straightforward to assign the community label to each researcher by choosing the highest probability. To facilitate the evaluation, we need to assign the ground-truth label to each researcher, and thus we select the field in which the researcher has the most publications as his/her ground-truth label. Table 2 demonstrates the comparison results in terms of NMI. Above statistics suggest the apparent regularity of performance trends. On the whole, once the clustering method is fixed, NMI increases with the algorithms deriving relation distribution, according to order "Uniform→ MultiRank→MutuRank". The similar phenomenon also happens to the order "K-means→GMM→GMM-NK", when the algorithm of computing relation distribution is fixed. The results shown in Table 2 are sufficient to validate the effectiveness of both our *MutuRank* in capturing relation distribution, and our GMM-NK on community detection.

Last but not the least, we look inside the membership probabilities calculated by GMM-NK. We sampled five famous scholars and showed their membership probabilities of 13 fields, as shown in Fig. 4. Note that the probability values below 0.001 are set to 0. These results well matched the "Research Interest" of these five scholars given by AMiner (`http://arnetminer.org`) [17]. For instance, `Jiawei Han` has always been very active on "DB" and "DM", and `Chengxiang Zhai` concentrates on the "IR" field.

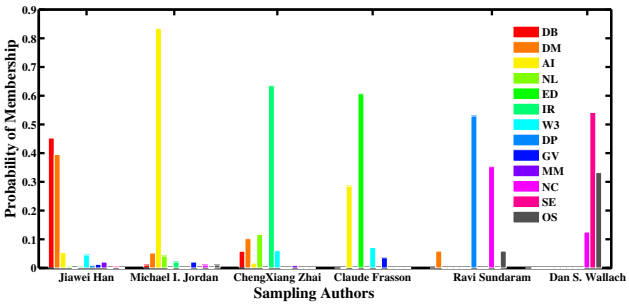


Fig. 4. The membership probabilities of sampling authors

6 Conclusions

In this paper, we addressed the issue of community detection in multi-relational network. Similar with some literature, we also attempted to first convert the original multi-relational network into a single-relational graph which has been studied more extensively. In this phase, the proposed *MutuRank* model enabled us to acquire relation distribution with the consideration of interdependency between relations and objects. Subsequently, we utilized spectral clustering on the basis of a novel GMM-NK algorithm to detect communities in the single-relational graph. Experimental results on both synthetic and real-world data demonstrate the effectiveness of the proposed method.

Acknowledgments. This research was partially supported by National Natural Science Foundation of China under Grants 71072172 and 61103229, National Key Technologies R&D Program of China under Grants 2013BAH16F01 and 2013BAH16F04, Industry Projects in Jiangsu S&T Pillar Program under Grant BE2012185, Key Project of Natural Science Research in Jiangsu Provincial Colleges and Universities under Grant 12KJA520001, and the Natural Science Foundation of Jiangsu Province of China under Grant BK2012863, and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Cai, D., Shao, Z., He, X., Yan, X., Han, J.: Community mining from multi-relational networks. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 445–452. Springer, Heidelberg (2005)
2. Cao, J., Wu, Z., Wu, J., Xiong, H.: SAIL: Summation-based incremental learning for information-theoretic text clustering. IEEE Transactions on Systems, Man, and Cybernetics–Part B 43(2), 570–584 (2013)
3. Dai, B.T., Chua, F.C.T., Lim, E.P., Faloutsos, C.: Structural analysis in multi-relational social networks. In: Proc. of the International SIAM Conference on Data Mining (SDM 2012), 451–462 (2012)

4. Fortunato, S.: Community detection in graphs. Physics Reports 486, 75–174 (2010)
5. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. Physical Review E 83, 016107 (2011)
6. Kellogg, R.: Uniqueness in the schauder fixed point theorem. Proc. of the American Mathematical Society 60, 207–210 (1976)
7. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Systems Technical Journal 49, 291–307 (1970)
8. Newman, M.: Fast algorithm for detecting community structure in networks. Physical Review E 69(6), 066113 (2004)
9. Ng, M., Li, X., Ye, Y.: Multirank: co-ranking for objects and relations in multi-relational data. In: Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1217–1225 (2011)
10. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web (1999)
11. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
12. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76(3), 036106 (2007)
13. Rahimian, F., Payberah, A.H., Girdzijauskas, S., Jelasity, M., Haridi, S.: Ja-be-ja: A distributed algorithm for balanced graph partitioning. Technical Report, Swedish Institute of Computer Science (2013)
14. Rodriguez, M., Shinavier, J.: Exposing multi-relational networks to single-relational network analysis algorithms. Journal of Informetrics 4(1), 29–41 (2010)
15. Slater, P.B.: Established clustering procedures for network analysis. Technical Report, arXiv:0806.4168 (June 2008)
16. Szell, M., Lambiotte, R., Thurner, S.: Multirelational organization of large-scale social networks in an online world. Proc. of the National Academy of Sciences 107(31), 13636–13641 (2010)
17. Tang, J., Yao, L., Zhang, D., Zhang, J.: A combination approach to web user profiling. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(1), 2 (2010)
18. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 817–826 (2009)
19. Tang, L., Liu, H., Zhang, J.: Identifying evolving groups in dynamic multimode networks. IEEE Transactions on Knowledge and Data Engineering 24(1), 72–85 (2012)
20. Tang, L., Wang, X., Liu, H.: Community detection via heterogeneous interaction analysis. Data Mining Knowledge Discovery 25(1), 1–33 (2012)
21. Tang, L., Liu, H.: Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery 2(1), 1–137 (2010)
22. Wei, F., Qian, W., Wang, C., Zhou, A.: Detecting overlapping community structures in networks. World Wide Web Journal 12(2), 235–261 (2009)
23. Yang, B., Liu, J., Feng, J.: On the spectral characterization and scalable mining of network communities. IEEE Transactions on Knowledge and Data Engineering 24(2), 326–337 (2012)
24. Yu, W., Zhang, W., Lin, X., Zhang, Q., Le, J.: A space and time efficient algorithm for simrank computation. World Wide Web Journal 15(3), 327–353 (2012)