Data Min Knowl Disc (2011) 22:493–521
DOI 10.1007/s10618-010-0181-y

Community discovery using nonnegative matrix factorization

Fei Wang · Tao Li · Xin Wang · Shenghuo Zhu · Chris Ding

Received: 7 May 2009 / Accepted: 16 June 2010 / Published online: 4 July 2010

Abstract Complex networks exist in a wide range of real world systems, such as social networks, technological networks, and biological networks. During the last decades, many researchers have concentrated on exploring some common things contained in those large networks include the small-world property, power-law degree distributions, and network connectivity. In this paper, we will investigate another important issue, community discovery, in network analysis. We choose Nonnegative Matrix Factorization (NMF) as our tool to find the communities because of its powerful interpretability and close relationship between clustering methods. Targeting different types of networks (undirected, directed and compound), we propose three NMF techniques (Symmetric NMF, Asymmetric NMF and Joint NMF). The

Responsible editor: Eamonn Keogh.

F. Wang (✉) · T. Li · X. Wang
School of Computing and Information Sciences, Florida International University, Miami, FL, USA
e-mail: feiwang@cs.fiu.edu; feiwang03@gmail.com

T. Li
e-mail: taoli@cs.fiu.edu

X. Wang
e-mail: xwang009@cs.fiu.edu

Present Address:
F. Wang
IBM T.J. Watson Research Lab, Hawthorne, NY, USA

S. Zhu
NEC Research Lab America at Cupertino, Cupertino, CA, USA

C. Ding
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

correctness and convergence properties of those algorithms are also studied. Finally the experiments on real world networks are presented to show the effectiveness of the proposed methods.

Keywords Community discovery · Nonnegative matrix factorization

1 Introduction

Nowadays, complex networks exist in a wide variety of systems in different areas, such as social networks (Scott 2000; Wasserman and Faust 1994), technological networks (Amaral et al. 2000; Watts and Strogatz 1998), biological networks (Sharan 2005; Watts and Strogatz 1998) and information networks (Albert et al. 1999; Faloutsos et al.). Despite the diverse physical meanings behind those networks, they usually exhibit common topological properties, such as the small-world phenomenon (Barthelemy and Amaral 1999) and the power-law degree distribution (Faloutsos et al.). Besides that, most real world networks demonstrate that the nodes (or units) contained in their certain parts are densely connected to each other (Palla et al. 2005), which are usually called clusters or communities (Girvan and Newman 2002). Efficiently identifying those communities can help us to know the nature of those networks better and facilitate the analysis on those large networks.

Generally, a network can be represented as a graph, where the graph nodes stand for the units in the network, and the graph edges denote the unit relationships. A typical network with 4 inside communities is illustrated in Fig. 1, where nodes with different colors belong to different communities, and different communities may share commons units.

During the last decades, many algorithms have been proposed to identify the communities contained in a network. For example, the *k-means* clustering method, the

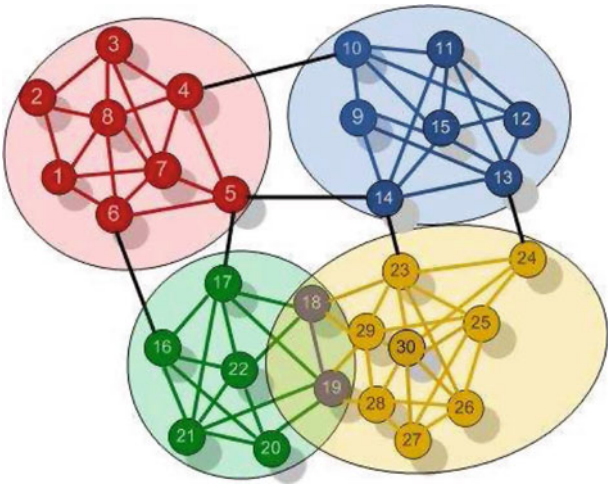


Fig. 1 Illustration of a typical network and its inside communities. Nodes with different colors correspond to different communities, and the numbers on the nodes correspond to their indices

hierarchical agglomerative clustering (HAC) method (Newman 2004a), the modularity optimization method (Newman 2004b) and the probabilistic method based on latent models (Zhang et al. 2007). More recently, the graph based partitioning methods (von Luxburg 2007; Weiss 1999) have aroused considerable interests in machine learning and data mining fields, and these methods have also been successfully applied to community discovery (Flake et al. 2000; Ino et al. 2005; Miao et al. 2008; Ruan and Zhang 2007). The basic idea behind those methods is to treat the whole network as a large graph, then the communities would correspond to the inside subgraphs, which can be identified via graph partitioning methods.

Despite the theoretical and empirical success of graph based methods, they still own some limitations. On one hand, these algorithms usually result in an eigenvalue decomposition problem and the communities can be identified from the resultant eigenvectors. However, it is generally hard to tell the exact physical meanings of those eigenvectors, which is important for explaining the final results when associated with real world applications. On the other hand, it is difficult for them to tackle the overlapping clusters.

As another research topic, Nonnegative Matrix Factorization (NMF) has emerged as a powerful tool for data analysis with enhanced interpretability. It was originally proposed as a method for finding matrix factors with parts-of-whole interpretations (Lee and Seung 1999). Later NMF has been successfully applied to environmetrics (Paatero and Tapper 1994), chemometrics (Xie et al. 1999), information retrieval (Pauca et al. 2004), bioinformatics (Brunet et al. 2004), etc. More recently, Ding et al. pointed out that the NMF based algorithms have close relationships with *kmeans* and graph partitioning methods (Ding et al. 2005, 2006a,b). Moreover, they also show that the results of NMF could be more easily explained (Ding et al. 2008).

In this paper, we propose to apply the NMF based algorithms for community discovery. We consider three types of networks, the undirected network, directed network and compound network, and we develop three different techniques, Symmetric NMF (SNMF), Asymmetric NMF (ANMF) and Joint NMF (JNMF) to identify the hidden communities on different networks. The correctness and convergence properties of these algorithms are also studied. Finally we conduct a set of experiments on real world networks to show the effectiveness of those algorithms. It is worthwhile to highlight several aspects of the proposed algorithms:

- Since all the resultant matrices are nonnegative, our methods own a high interpretability.
- Our method does not force the final resultant clusters to be exclusive, which makes it be capable of dealing with overlapping clusters.
- Some prior knowledge (e.g. side-information) can be easily incorporated into those algorithms (Wang et al. 2008b).

The rest of this paper will be organized as follows. Section 2 will introduce community discovery in undirected networks together with the SNMF algorithm. The details of the ANMF and HNMF algorithms will be introduced in Sects. 3 and 4. The experimental results on real world networks will be introduced in Sect. 5, followed by the conclusions and discussions in Sect. 6.

2 Community discovery in undirected networks

In this section we will consider the problem of community discovery in undirected networks, in which all the relationships between pairwise units are symmetric, i.e., all information carried on the network edges are undirected. In order to present our algorithm more clearly, we first introduce a motivating example.

2.1 A motivating example

An undirected network can be represented as an undirected graph $\mathcal{G}$. Assume it is composed of k communities (clusters) $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of size $p_1, \ldots, p_k$, and the nodes in each cluster $\mathcal{C}_i$ are connecting to each other with the same weight s_i while the nodes in different clusters are disconnected with each other. Then, without loss of generality, we assume that the rows belong to a particular cluster are contiguous, so that all data points belonging to the first cluster appear first and the second cluster next, etc [1]. Then the adjacency matrix of $\mathbf{G}$ can be represented as

$$\mathbf{G} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_k \end{bmatrix}.$$

where $\mathbf{S}_i$ is $p_i \times p_i$ constant matrix with all its elements equal to s_i. Then we can factorize $\mathbf{G} = \mathbf{X}\mathbf{S}\mathbf{X}^\top$ where

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & s_k \end{bmatrix}.$$

In the above factorization, $\mathbf{X}$ provides the cluster membership and the diagonal elements s_i of $\mathbf{S}$ shows the connectivity within each cluster.

Note that in this simple case, the tri-factorization of $\mathbf{G}$ to $\mathbf{X}\mathbf{S}\mathbf{X}^\top$ is equivalent to the solution of perform eigenvalue decomposition on $\mathbf{G}$. In real world applications, the network structure may not be such clear and $\mathbf{G}$ will have non-zero off-diagonal entries. Particularly, when there are overlapping clusters, the columns of $\mathbf{X}$ will not be orthogonal to each other. In this case the tri-factorization of $\mathbf{G}$ will be no longer equivalent to its eigenvalue decomposition as in spectral clustering methods. Then the problem of community discovery in networks can be casted as the following NMF problem:

[1] This can be achieved by multiplying the adjacency matrix $\mathbf{G}$ with a permutation matrix if necessary.

$$\min_{\mathbf{X}\geq 0,\mathbf{S}\geq 0} \ell(\mathbf{G}, \mathbf{XSX}^\top), \tag{1}$$

where $\ell(\mathbf{A}, \mathbf{B})$ is a general loss defined on matrices $\mathbf{A}$, $\mathbf{B}$, among which the *Euclidean Loss*

$$\ell(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{ij} (\mathbf{A}_{ij} - \mathbf{B}_{ij})^2$$

is one of the commonly used loss types. In this paper, we will also make use of such Euclidean loss.

In the undirected case, $\mathbf{G}$ is symmetric, thus $\mathbf{S}$ is also symmetric. In the following we will introduce how to minimize the loss using NMF methods in both cases.

2.2 Symmetric nonnegative matrix factorization

If $\mathbf{G}$ is symmetric, then $\mathbf{S}$ is symmetric. We can then absorb $\mathbf{S}$ into $\mathbf{X}$, *i.e.*, $\hat{\mathbf{X}} = \mathbf{XS}^{1/2}$. Then our problem is to solve the following problem

$$\min_{\hat{\mathbf{X}}\geq 0} \left\| \mathbf{G} - \hat{\mathbf{X}}\hat{\mathbf{X}}^\top \right\|_F^2, \tag{2}$$

According to Wang et al. (2008a), $\hat{\mathbf{X}}$ can be solve by the following multiplicative update rule

$$\hat{\mathbf{X}}_{ik} \longleftarrow \hat{\mathbf{X}}_{ik} \left(\frac{1}{2} + \frac{\left(\mathbf{G}\hat{\mathbf{X}}\right)_{ik}}{\left(2\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{X}}\right)_{ik}} \right). \tag{3}$$

After convergence, the obtained $\hat{\mathbf{X}}^*$ is just the scale partition matrix of the network $\mathbf{G}$ of size $n \times K$, whose i-th row corresponds to the cluster (community) membership of the i-th unit. We can further normalize $\hat{\mathbf{X}}$ to make $\sum_j \hat{\mathbf{X}}_{ij} = 1$, such that $\hat{\mathbf{X}}_{ik}$ corresponds to the posterior probability that the i-th unit belongs to the k-th community.

2.3 An illustrative example

Before we go into the details of applying NMF based methods for community discovery on other types of networks, first let's see an illustrative example here on how SNMF works on undirected networks. The network structure is shown in Fig. 1, and we construct the adjacency matrix $\mathbf{G} \in \mathbb{R}^{n\times n} (n = 30)$ as $\mathbf{G}_{ij} = 1$ if there is an edge connecting node i and node j; and $\mathbf{G}_{ij} = 0$ otherwise. The scaled partition matrix $\hat{\mathbf{X}}$ is randomly initialized. The final results of applying Eq. 3 to update $\hat{\mathbf{X}}$ till convergence are shown in Fig. 2, where the more the color of the (i, j)-th block tends to red, the more probable that unit i belongs to community j. From the figures we can

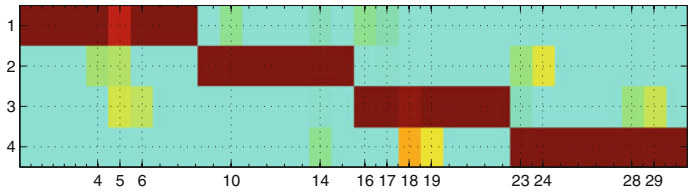


Fig. 2 Results of applying SNMF to discover the communities in the network shown in Fig. 1, in which the x-axis corresponds to the data indices, and the y-axis represents the community categories

clearly see that our SNMF can successfully discover the community structure in the network. Moreover, for the overlapping region of two communities (e.g. node 18 and 19), SNMF can also successfully detect them and assign possibilities to them which indicating the extent they belonging to each cluster.

3 Community discovery in directed graph

Another type of frequently used network is the directed network, i.e., the edges contained in the network are all directed, which makes the adjacency matrix $\mathbf{A}$ asymmetric since $\mathbf{A}_{ij} \neq \mathbf{A}_{ji}$. A simple directed network with the same topology structure as Fig. 1 is shown in Fig. 3. In the following we will introduce an Asymmetric Nonnegative Matrix Factorization (ANMF) approach to detect communities in a directed network.

3.1 Asymmetric nonnegative matrix factorization

As stated above, the adjacency matrix $\mathbf{A}$ is asymmetric in the directed case. Now consider the model under Euclidean distance.

$$\begin{aligned} \min_{\mathbf{X},\mathbf{S}} \quad & \ell(\mathbf{A}, \mathbf{X}\mathbf{S}\mathbf{X}^\top), \\ \text{s.t.} \quad & \mathbf{X} \in \mathbb{R}_+^{n\times q}, \mathbf{S} \in \mathbb{R}_+^{q\times q} \end{aligned} \tag{4}$$

where $\ell(\mathbf{X}, \mathbf{Y}) \stackrel{\text{def}}{=} \|\mathbf{X} - \mathbf{Y}\|_F^2$. Note that since $\mathbf{A}$ is asymmetric, $\mathbf{S}$ is also asymmetric. We do not enforce the normalization constraint on $\mathbf{X}$ at this moment. We can further normalize $\mathbf{X}$ it by transferring a diagonal matrix between $\mathbf{X}$ and $\mathbf{S}$ as

$$\mathbf{X}\mathbf{S}\mathbf{X}^\top = (\mathbf{X}\mathbf{D}^{-1})(\mathbf{D}\mathbf{S}\mathbf{D}^\top)(\mathbf{X}\mathbf{D}^{-1})^\top \tag{5}$$

Then we have the following theorem.

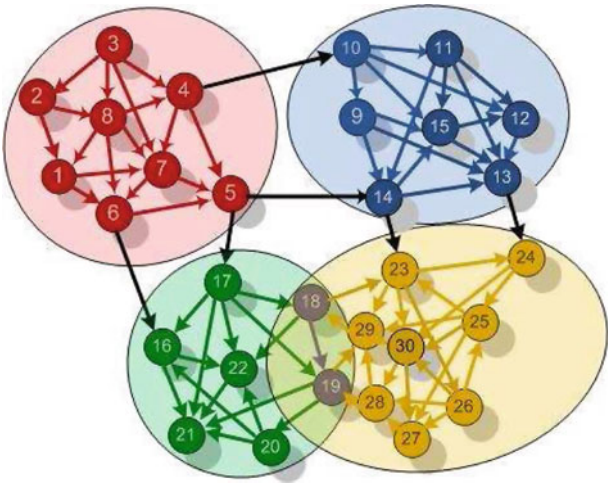


Fig. 3 A simple directed network

Theorem 1 *The loss* $\ell(\mathbf{A}, \mathbf{XSX}^\top)$ *is nonincreasing under the alternative update rules:*

$$\mathbf{X}_{ik} \leftarrow \mathbf{X}_{ik} \left(\frac{\left[\mathbf{A}^\top \mathbf{XS} + \mathbf{AXS}^\top\right]_{ik}}{\left[\mathbf{XSX}^\top \mathbf{XS}^\top + \mathbf{XS}^\top \mathbf{X}^\top \mathbf{XS}\right]_{ik}} \right)^{\frac{1}{4}}, \tag{6}$$

$$\mathbf{S}_{kl} \leftarrow \mathbf{S}_{kl} \frac{\left[\mathbf{X}^\top \mathbf{AX}\right]_{kl}}{\left[\mathbf{X}^\top \mathbf{XSX}^\top \mathbf{X}\right]_{kl}}. \tag{7}$$

The loss is invariant under these updates if and only if $\mathbf{X}$ *and* $\mathbf{S}$ *are at a stationary point of the loss with the constraints.*

Proof See Appendix I.

3.2 An illustrative example

We also give an intuitive example here to illustrate how ANMF works on directed networks. The topology of the network is shown in Fig. 3, which is the same as in Fig. 1 except for the directed edges. We first constructed the $n \times n (n = 30)$ adjacent matrix $\mathbf{A}$ as $\mathbf{A}_{ij} = 1$ is there is an edge pointing from $\mathbf{x}_i$ to $\mathbf{x}_j$; and $\mathbf{A}_{ij} = 0$ otherwise. Matrices $\mathbf{X}$ and $\mathbf{S}$ are randomly initialized. The final iteration result of $\mathbf{X}$ after convergence is shown in Fig. 4, from which we can see that our algorithm can correctly discover the community structure of the network.

4 Community discovery in compound networks

In the previous sections we have introduced how to make use of the NMF based methods for community discovery in one network, and the network can be undirected or

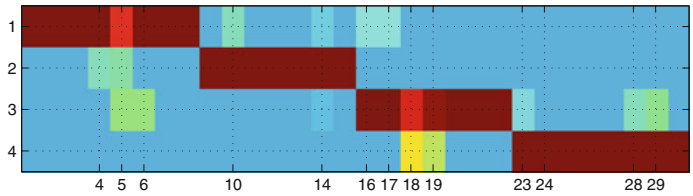


Fig. 4 Results of applying ANMF to discover the communities in the network shown in Fig. 1, in which the x-axis corresponds to the data indices, and the y-axis represents the community categories

directed. However, in real world applications, we may face with multiple networks for heterogeneous data analysis. For example, in an automatic movie recommendation system, we are given at least three networks: (i) the user network which shows the relationships among users with common watching interests; (ii) the movie network which shows the content relationship among movies (e.g., they are on the same topic or belong to the same category); and (iii) the user-movie network shows the ratings that the users give to those movies. We can analyze these networks to answer the following problems:

- Hidden Links: Two users do not have any links in the user network, but they have similar opinions on the watched movies (i.e., give similar ratings to those movies). Similarly, two movies do not have any links in the movie network, but they are watched by the same user or by different user with similar ratings.
- Hidden Clusters: This can be seen as extensions of hidden links. In fact, we want to identify user clusters or movie clusters based on all the information available (e.g., user network, movie network, and user-movie network).
- Boundary Spanners: The users that are familiar with many users (via user network) while they do not watch similar movies or give similar ratings as their friends can be thought as boundary spanners. The boundary spanners are well-positioned to be innovators, since they have access to ideas and information flow into other clusters.

Our aim is to develop a versatile model to formally analyze data associated with multiple networks. For simplicity, we only consider the data sets with two types of entities, and we assume all the networks are undirected. The algorithms for analyzing data associated with more networks or directed networks can be similarly derived, but much more complicated. First let's introduce some notations. We use $\mathbf{U}$ to denote the user-user matrix, $\mathbf{D}$ to denote the movie-movie matrix, and $\mathbf{M}$ to denote the user-movie matrix. We are looking for a latent matrix $\mathbf{X}$, which reflects some "intrinsic" relationships between the users and the movies, such that the following three objectives are minimized simultaneously: $\|\mathbf{M}-\mathbf{X}\|$, $\|\mathbf{U}-\mathbf{X}\mathbf{X}^\top\|$, $\|\mathbf{D}-\mathbf{X}^\top\mathbf{X}\|$. In such a way, we are able to better recover the relationships between user and movie, and the community structures of user and movie. In the following we will formalize this problem mathematically.

4.1 Joint nonnegative matrix factorization

As stated above, the problem we aim to solve is

$$\begin{aligned} \min \quad & \ell(\mathbf{X}, \mathbf{M}, \mathbf{U}, \mathbf{D}) \\ s.t. \quad & \mathbf{X} \in \mathbb{R}_{+}^{n \times m} \end{aligned}$$

where $\ell(\mathbf{X}, \mathbf{M}, \mathbf{U}, \mathbf{D}) \stackrel{\text{def}}{=} \|\mathbf{M} - \mathbf{X}\|^2 + \alpha\|\mathbf{U} - \mathbf{X}\mathbf{X}^\top\|^2 + \beta\|\mathbf{D} - \mathbf{X}^\top\mathbf{X}\|^2$, and $\alpha > 0$, $\beta > 0$ are constants to tradeoff the importance between different terms. Then we have the following theorem.

Theorem 2 *The loss* $\ell(\mathbf{X}, \mathbf{M}, \mathbf{U}, \mathbf{D})$ *is nonincreasing under the alternative update rule*

$$\mathbf{X}_{ij} \leftarrow \mathbf{X}_{ij} \left(\frac{\left[\mathbf{M} + 2\alpha\mathbf{U}\mathbf{X} + \mathbf{X}\hat{\mathbf{D}}\right]_{ij}}{2(\alpha + \beta)\left[\mathbf{X}\mathbf{X}^\top\mathbf{X}\right]_{ij}} \right)^{\frac{1}{4}} \tag{8}$$

where $\hat{\mathbf{D}} = 2\beta\mathbf{D} - \mathbf{I}$. *To guarantee the nonnegativeness of* $\mathbf{X}$*, we should set the similarities between pairwise movies to be larger than* $1/2\beta$.

Proof See Appendix II.

4.2 An illustrative example

In this section we will give a concrete example to illustrate the utility of our method. Specifically, suppose that we are given the following 5 references.

- Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
- W. Bruce Croft, Jamie Callan: A Language Modeling Approach to Metadata for Cross-Database Linkage and Search. Proceedings of the 2004 annual national conference on Digital government research, 2004, pp. 1–2.
- John D. Lafferty, Chengxiang Zhai: Document Language Models, Query Models, and Risk Minimization for Information Retrieval. SIGIR, 2001, pp. 111–119.
- Deng Cai, Qiaozhu Mei, Jiawei Han, Chengxiang Zhai: Modeling hidden topics on document manifold. Proceeding of the 17th ACM conference on Information and knowledge management, 2008, pp. 911–920.
- Jiawei Han, Yongjian Fu: Mining Multiple-Level Association Rules in Large Databases. IEEE Trans. Knowl. Data Eng. (1999) 11(5):798–804.

Then we can construct a 5×5 author-author similarity matrix $\mathbf{U}$ according to their co-author relationships (i.e., $U_{ij} = 1$ if author i and author j have co-authored for one paper, otherwise $U_{ij} = 0$. $\mathbf{U}$ is symmetric). Moreover, we also construct a dictionary (Large, Database, Association, Rule, Language, Model, Information Retrieval,

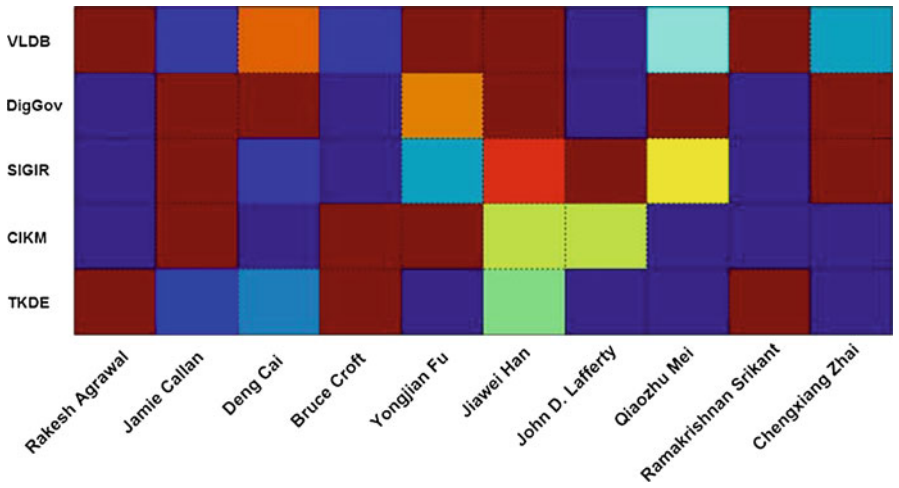


Fig. 5 The inferred hidden relationships between the authors and the papers

Mine) from the paper titles by collecting them and removing the stop words, so we can also construct a paper-paper similarity matrix by computing the cosine similarity of pairwise paper titles. We also have a given author-paper relationship matrix, which indicates who write the paper. We set $\alpha = \beta = 1$, and the predicted hidden author-paper relationship is shown in Fig. 1, where the red color indicates strong relationships, while blue color suggests weak relationships. From the figure we can discover that:

- The authors have strong relationships with their own papers, which is consistent with the prior information.
- The co-authorship can be revealed. For example, Lafferty and Chengxiang have co-authored the SIGIR paper, and Chengxiang also wrote the CIKM paper, then Lafferty also has a strong relationship with the CIKM paper.
- The paper similarity can be revealed. For example, Rakesh wrote the VLDB paper on association rule mining, then he also has a strong relationship with the TKDE paper which also talked about association rule mining.

Therefore we can see that our algorithm can integrates the author-author relationships, paper-paper relationships and author-paper relationships together, and it really finds out some "hidden relationships" between authors and papers. (Fig. 5).

5 Experiments

In this section we will present a set of experiments on real world data sets to validate the effectiveness of our NMF based algorithms for community discovery.

Table 1 The results of applying SNMF to the NIPS co-author network

Group 1	Subutai Ahmad, Volker Tresp, R. Hofmann Ralph Neuneier, H.G. Zimmermann
Group 2	T. J. Sejnowski, A. Pouget, P. Viola, J. R. Movellan, G. Tesauro, K. Doya, N. N. Schraudolph, P. Dayan, G. E. Hinton
Group 3	J. C. Platt, S. Nowlan, J. Shawe-Taylor Nello Cristianini
Group 4	Satinder P. Singh , M. Kearns, Andy Barto, R. S. Sutton, David Cohn
Group 5	John Moody, Todd Leen, Y.Kabashima David Saad
Group 6	Y. Bengio,J. Denker, Y. LeCun, I. Guyon, H. P. Graf, L. Bottou, S. Solla
Group 7	C. Williams, D. Barber H. Hertz, M. Opper
Group 8	C. Koch, A. Moore, R. Goodman B. Mel, H. Seung, D. Lee
Group 9	M. Jordan, Z. Ghahramani, C. Bishop T. Jaakkola, L. Saul P. Smyth, D. Wolpert
Group 10	A. Smola, B. Schölkopf, V. Vapnik P. Barlett, R. Meir, R. Williamson

5.1 Undirected network

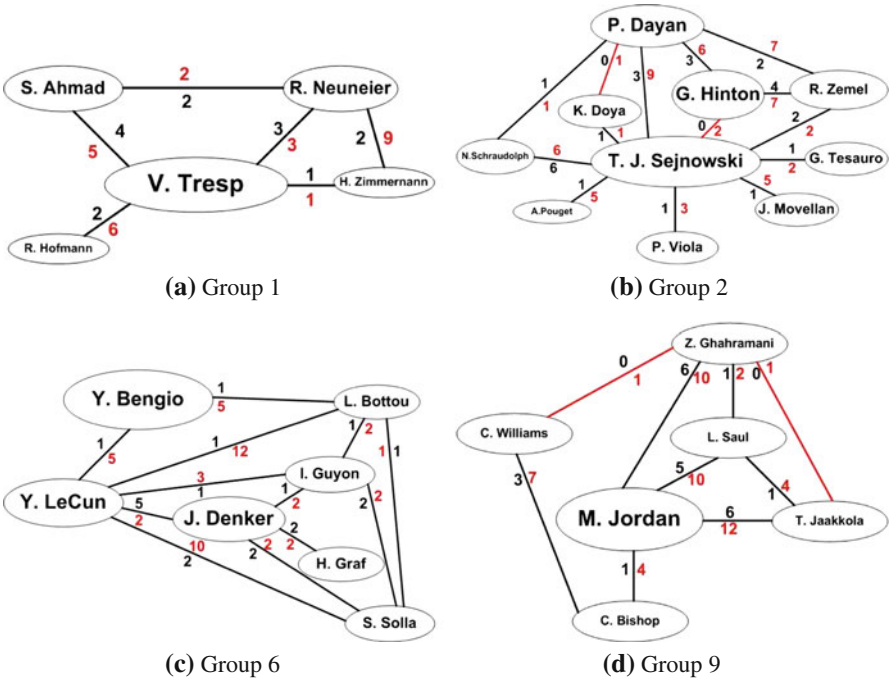
5.1.1 An illustrative example

We use the *NIPS Conference Papers Vol. 0–12 Data Set*[2] in our experiments. It contains all the paper information of NIPS proceedings from volume 0 to volume 12. There are totally 2037 authors with 1740 papers. We use the co-author data to construct the undirected network. The (i, j)-th element of such adjacency matrix $\mathbf{G}$ of such co-author graph is equal to the number of co-authored NIPS papers of the i-th and the j-th authors. So the size of $\mathbf{G}$ is 2037×2037.

We first treat $\mathbf{G}$ as the similarity matrix between authors and apply K-means algorithm to initialize the scaled partition matrix $\hat{\mathbf{X}}$ in Eq. 2 and then apply Eq. 3 to update it until convergence[3] After we getting $\hat{\mathbf{X}}$, we can first row-normalize it to $\sum_j \hat{\mathbf{X}}_{ij} = 1$, and finally the community that the i-th author belonging to can be determined as $l_i = argmax_j \hat{\mathbf{X}}_{ij}$. We set the number of communities to be 10 manually, and some representative authors in each community are illustrated in Table 1.

[2] Available at http://www.cs.toronto.edu/~roweis/data.html.

[3] Here *convergence* is defined as that the change of $\hat{\mathbf{X}}$ in two successive iteration steps are no larger than a specified threshold ε.



(a) Group 1 **(b)** Group 2

(c) Group 6 **(d)** Group 9

Fig. 6 The co-author network structures on some of the discovered communities. The nodes represent different authors, and the node size indicates the paper number of the corresponding author. The edges connecting pairwise authors indicate that the two authors have co-authored at least one NIPS paper, and the black numbers on the edges correspond to the number of NIPS papers the two authors co-authored, and the red numbers correspond to their co-authored papers on the DBLP record. The red lines denote that the two authors have co-authored no NIPS papers, but they have other co-authored papers according to the DBLP

To better demonstrate the quality of the results, we draw the co-author structures for some of the communities, which are shown in Fig. 6. From the figure we can discover some typical characteristics of these structures:

- Generally there are one or several "superstars" in each community, for example, *V. Tresp, T. J. Sejnowski, Y. Lecunn* and *M. Jordan*. The whole community can be effectively consolidated through the wide social relationships of these superstars.
- The authors that have co-authored NIPS papers also tend to co-author papers on other conferences/journals. This can be observed by the red numbers on the network edges (the number of co-authored papers according to the DBLP record), which are usually larger than the associated black numbers (the number of co-authored NIPS papers).
- The authors in the same community that have never co-authored any NIPS paper may tend to cooperate for other papers, which can be observed by the red lines in the figures.

Table 2 The basic information of WebKB

School	Course	Dept.	Faculty	Other	Project	Staff	Student	Total
Cornell	44	1	34	581	18	21	128	827
Texas	36	1	46	561	20	2	148	814
Washington	77	1	30	907	18	10	123	1166
Wisconsin	85	0	38	894	25	12	156	1210

Besides, we can also infer how active the authors are in the NIPS community according to their label entropies. The label entropy of the i-th author can be computed as

$$le(i) = -\sum_j \hat{\mathbf{X}}_{ij} \log \hat{\mathbf{X}}_{ij}$$

According to our experiments, the authors with the highest label entropy are *V. Vapnik, M. Jordan* and *G. Hinton*, which are all big names in today's machine learning community.

5.1.2 Qualitative comparisons

We further compare the performance of the proposed Symmetric Nonnegative Matrix Factorization (SNMF) method with other competitive algorithms. The data set we used here is the WebKB data set[4], which consists of about 6000 web pages from computer science departments of four schools (Cornell, Texas, Washington, and Wisconsin). The web pages are classified into seven categories. The numbers of pages in each category are shown in Table 2.

Besides the proposed SNMF algorithm, we also implement the Symmetric Tri-Nonnegative Matrix Factorization (STriNMF) (Ding et al. 2006b) algorithm and the Symmetric Convex Coding (SCC) (Long et al. 2007). For all the algorithms, we use the same document relationship matrix which is just the inner-product matrix between the document vectors. We randomly initialize the matrices updated and proceed multiplicative updates over 400 iterations. We calculate the clustering accuracy (Wang et al. 2008b), normalized mutual information (Strehl et al. 2002)[5] and the objective function loss they aim to minimize after each iteration step. Finally we average all these values over 100 different initializations and report the performances in Figs. 7, 8 and 9. From these figures we can clearly see that our proposed SNMF algorithm, although only requires update one matrix at each iteration (the other two algorithms need to update two matrices at each iteration), can achieve better clustering results (in terms of clustering accuracy and NMI) and low function loss. Moreover, we also

[4] http://www.cs.cmu.edu/~WebKB/. The data set we used in our experiments can be downloaded from http://www.nec-labs.com/~zsh/files/link-fact-data.zip.

[5] Once we obtained the relaxed cluster indicator matrix ($\hat{\mathbf{X}}$ in SNMF, $\mathbf{H}$ in STriNMF (Ding et al. 2006b), and $\mathbf{B}$ in SCC (Long et al. 2007)), the cluster assignment of a specific data point is just the index of its corresponding row. For example, in SNMF, once we got $\hat{\mathbf{X}}$, the cluster assignment for $\mathbf{x}_i$ is just $arg\max_j \hat{\mathbf{X}}_{ij}$.

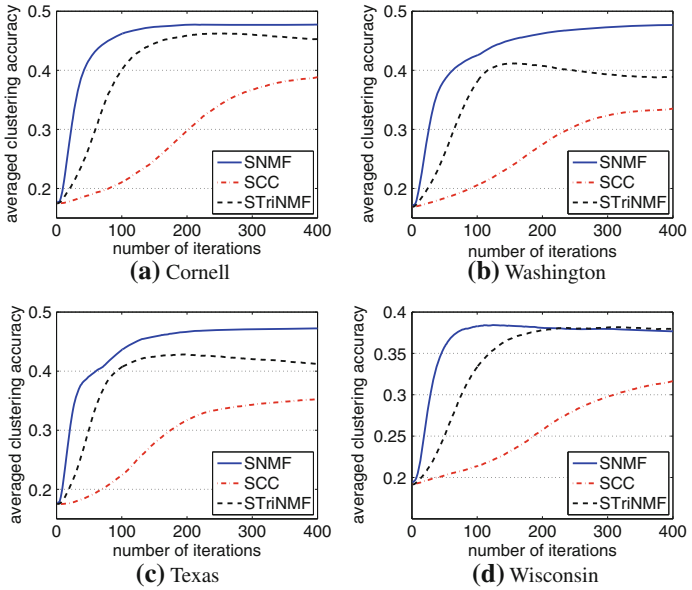


Fig. 7 Averaged clustering accuracy results for different matrix factorization based methods. The x-axis is the number of iterations, and the y-axis is the clustering accuracy averaged over 100 independent runs

record the final averaged clustering performance over 400 iterations with STriNMF and SCC, and spectral clustering (where we use the document inner-product matrix as the similarity matrix) and Kernel K-means (with linear inner-product kernel, and the results are also averaged over 100 independent runs with random initializations). The results are shown in Tables 3 and 4. From the tables we can clearly observe the superiority of our proposed SNMF algorithm.

5.2 Directed network

5.2.1 An illustrative example

We use the Enron email data set(Priebe et al. 2005) to evaluate the effectiveness of our ANMF method on directed networks. The data set contains the email communication records among 184 users in a period of 189 weeks, from 1998 to 2002.

We construct an 184x184 adjacency matrix $\mathbf{A}$ for such directed email graph, whose (i, j)-th element is computed as the number of emails that user i sent to user j during that period. The matrices $\mathbf{X}$ and $\mathbf{S}$ are initialized randomly, and then iteratively updated using Eqs. 6 and 7 until convergence. The number of communities is set to 4.

Table 5 shows the communities of the Enron data set discovered by the ANMF algorithm, where we give some representative employees of each community together with their jobs. After checking the working departments of these employees, we find that the people in the same community usually work in the same department. This is quite reasonable since people tend to send email to their colleagues in the same group.

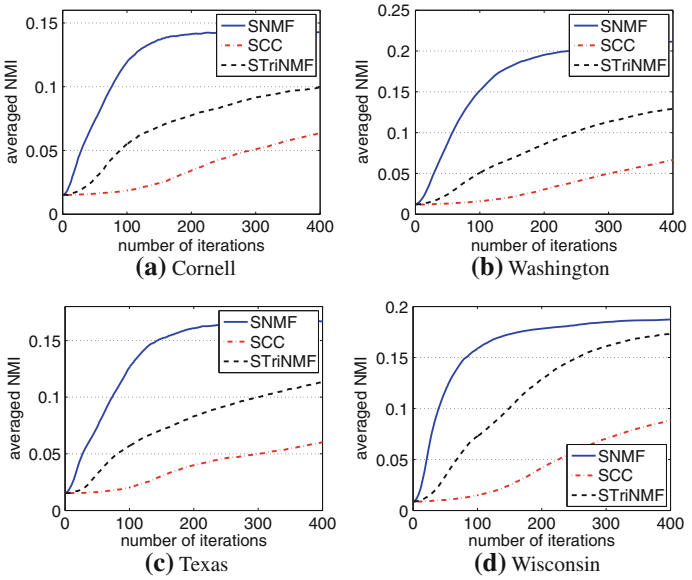


Fig. 8 Averaged normalized mutual information results for different matrix factorization based methods. The x-axis is the number of iterations, and the y-axis is the NMI value averaged over 100 independent runs

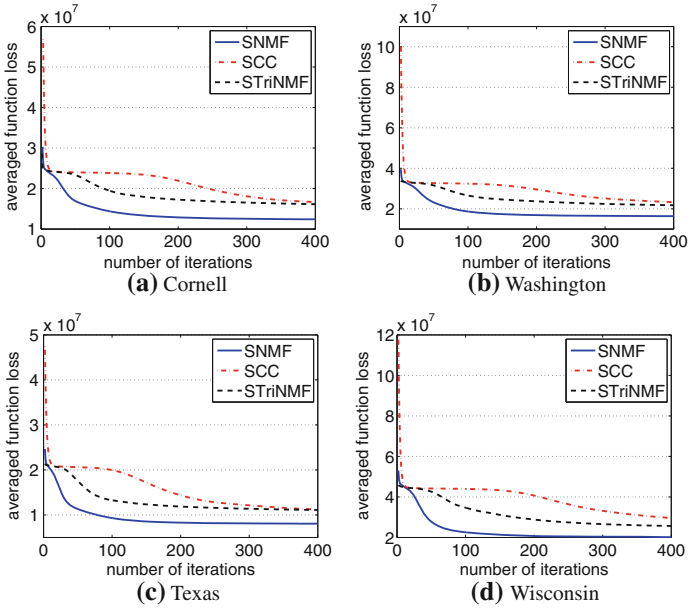


Fig. 9 Averaged function loss variations for different matrix factorization based methods. The x-axis is the number of iterations, and the y-axis is the function loss averaged over 100 independent runs

Table 3 Averaged clustering accuracy comparison for symmetric algorithms on WebKB data set

	Cornell	Washington	Texas	Wisconsin
SNMF	**0.4773**	**0.4766**	**0.4646**	0.3768
STriNMF	0.4525	0.3888	0.4048	**0.3796**
SCC	0.3880	0.3352	0.3478	0.3164
Spectral clustering	0.4012	0.3790	0.3729	0.3230
Kernel K-means	0.3658	0.3239	0.3374	0.3145

Table 4 Averaged NMI comparison for symmetric algorithms on WebKB data set

	Cornell	Washington	Texas	Wisconsin
SNMF	**0.1427**	**0.2114**	**0.1734**	**0.1872**
STriNMF	0.0994	0.1293	0.1166	0.1731
SCC	0.0637	0.0665	0.0605	0.0889
Spectral clustering	0.0877	0.0897	0.0817	0.1221
Kernel K-means	0.0523	0.0603	0.0623	0.0851

To further validate the correctness of the results, we draw the email sending structure within each community in Fig. 10, from which we can clearly see that the people in the same community contact very closely to each other.

5.2.2 Qualitative results

We also utilize the WebKB data set (with the contained webpage link information) to test the effectiveness of the proposed Asymmetric Nonnegative Matrix Factorization (ANMF) method, where $\mathbf{X}$ and $\mathbf{S}$ are randomly initialized, and we record the clustering accuracy, NMI and objective function loss after each iteration with a total of 400 iterations. The results are shown in Figs. 11, 12 and 13, where all the curves are averaged over 100 independent runs with different initializations. We also compare the final results of our algorithm with Directed Spectral Clustering (DSC) (Zhou et al. 2005). The experimental results are shown in Tables 6 and 7. From the table we can see that our algorithm can generally outperform DSC.

5.3 Compound network

In this section, we will apply the JNMF method introduced in Sect. 4 to *Collaborative Filtering* (Yu and Tresp 2005). Collaborative filtering exploits the correlations between item ratings across a set of users. It's goal is to predict the ratings of a testing user on new items given his/her historical ratings on other items and also the ratings given by other like-minded users on all items.

Table 5 The results of applying ANMF to the Enron email network

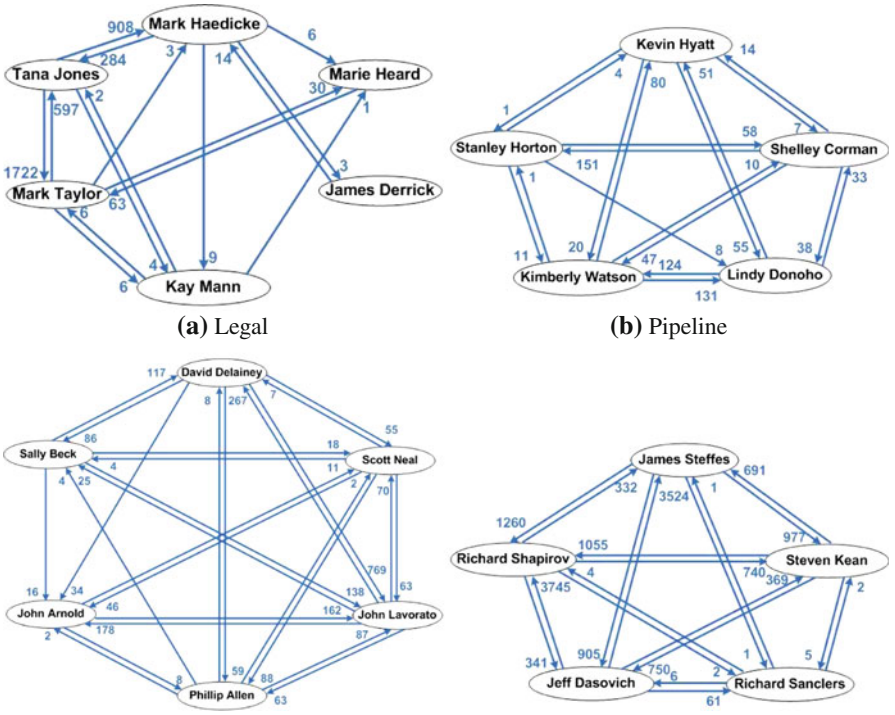
Role	Name	Job
Legal	Mark Haedicke	Managing Director
	Tana Jones	Employee
	Marie Heard	Sr. Specialist
	Mark Taylor	Employee
	Kay Mann	Lawyer
	James Derrick	Lawyer
Pipeline	Kevin Hyatt	Director
	Stanley Horton	President
	Shelley Corman	V.P.
	Kimberly Watson	Employee
	Lindy Donoho	Employee
Trading	David Delainey	CEO
Executive	Sally Beck	COO
	John Arnold	V.P.
	Phillip Alen	Manager
	John Lavorato	CEO
	Scott Neal	V.P.
Government	James Steffes	V.P.
Affairs	Richard Shapiro	V.P.
Services	Jeff Dasovich	Employee
	Richard Sanders	V.P.
	Steven Kean	V. P.

We use the MovieLens [6] data set in our experiments, where we extracted a subset of 500 users with more than 40 ratings and 500 movie items. The basic characteristics of those data sets are summarized in Table 8.

In our experiments, we select 200, 300, 400 movies along with all the users from both data sets and apply them to evaluate our method. We assume that for all the users we only know 5, 10, 20 ratings that he/she gives. The remaining ratings are used for testing. Note that the known ratings are randomly selected and each experiment is carried out 20 times. To show the superiority of our method, we also conduct a set of competitive approaches including:

- Pearson Correlation Coefficient Based Approach (PCC). The implementations is the same as in (Resnick et al. 1994).
- Aspect Model Based Approach (AM). The implementations is the same as in (Hofmann and Puzicha 1999).

[6] http://www.grouplens.org/.



(a) Legal **(b)** Pipeline

(c) Trading/Excutive **(d)** Government Affairs/Services

Fig. 10 The email network structures of the discovered communities. The nodes represent different employees, and the edge directions points from the email senders to the receivers, and the number near the arrows indicate how many emails are sent

- Personality Diagnosis (PD). The implementations is the same as in (Pennock et al. 2000).
- Nonnegative Matrix Factorization Based Approach (NMF). The implementation is based on (Chen et al. 2007).
- Item Graph Based Approach (IG). The implementation is the same as in (Wang et al. 2006).

For our JNMF method, we also explore the user and movie information provided by the data set to construct the user similarity matrix **U** and movie similarity matrix **D**. For each user, the *MovieLens* data set provides his/her age, gender, occupation and zip code. We use the first three attributes to compute user-user similarities (where all the three information are discretized such that each user is just a three-dimensional vector, and the user-user similarities can me computed by vector dot-product). The movies are encoded as 0–1 vectors according to its topic and genre, and the movie-movie similarities can also computed by vector dot-product. We randomly initialize **X** and use Eq. 8 to update it till convergence. In our experiments, we first normalize all the entries in the rating matrix **M**, user matrix **U** and movie matrix **D** to be [0, 1], and when the final results come out, we then re-normalize them into scale 1 to 5. α and β are all set to 0.1 manually, and the mean absolute error (MAE) is used to evaluate the results of those algorithms, which is defined as

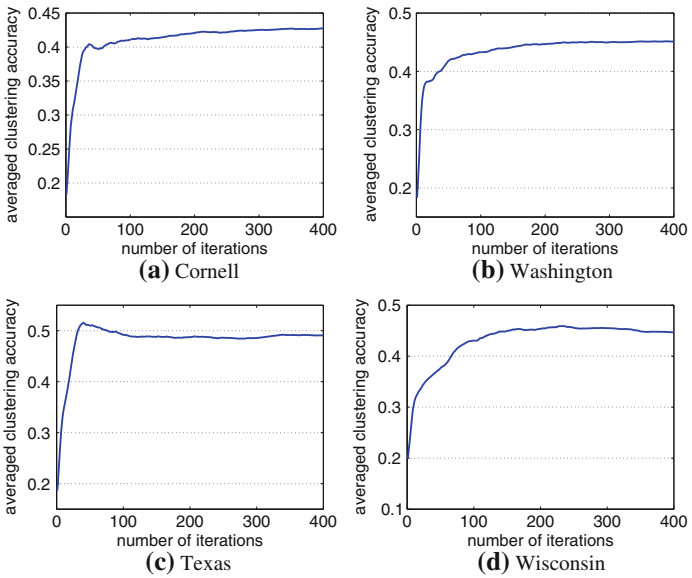


Fig. 11 Averaged clustering accuracy results for asymmetric nonnegative matrix factorization. The x-axis is the number of iterations, and the y-axis is the clustering accuracy averaged over 100 independent runs

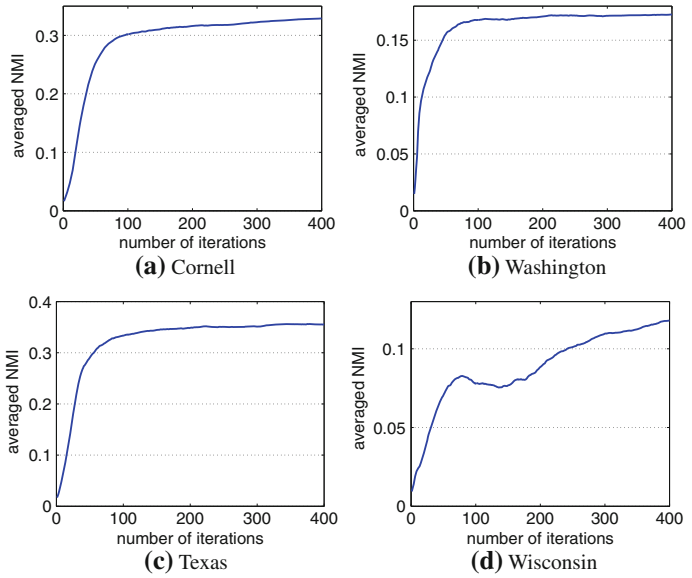


Fig. 12 Averaged normalized mutual information results for asymmetric nonnegative matrix factorization. The x-axis is the number of iterations, and the y-axis is the NMI value averaged over 100 independent runs

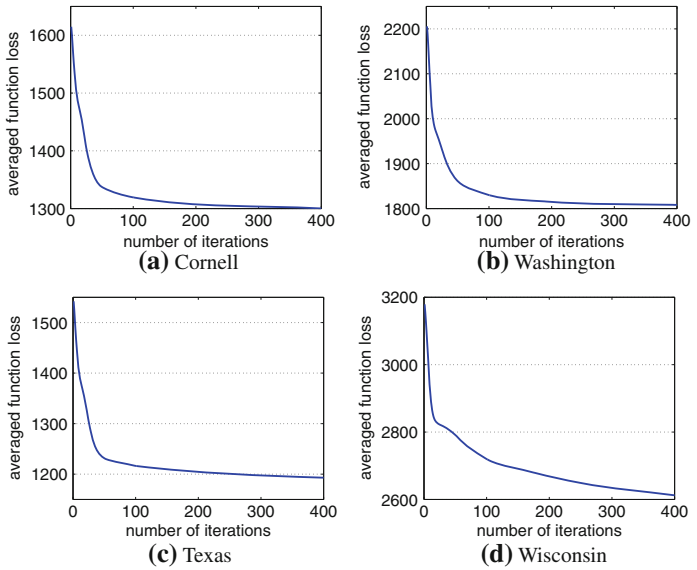


Fig. 13 Averaged function loss variations for asymmetric nonnegative matrix factorization. The x-axis is the number of iterations, and the y-axis is the function loss averaged over 100 independent runs

Table 6 Averaged clustering accuracy comparison for symmetric algorithms on WebKB data set

	Cornell	Washington	Texas	Wisconsin
ANMF	0.4274	**0.4509**	**0.4863**	**0.4466**
DSC	**0.4439**	0.4023	0.4558	0.4153

Table 7 Averaged NMI comparison for symmetric algorithms on WebKB data set

	Cornell	Washington	Texas	Wisconsin
ANMF	0.3287	**0.1725**	**0.3360**	**0.1179**
DSC	**0.3450**	0.1408	0.2984	0.0923

$$MAE = \frac{\sum_{u \in \mathcal{U}} |R_u(t_j) - \hat{R}_u(t_j)|}{|\mathcal{U}|} \tag{9}$$

where $\mathcal{U}$ is the set of users and $|\mathcal{U}|$ denotes its size, $R_u(t_j)$ denotes the true rating that user u gives to movie t_j, and $\hat{R}_u(t_j)$ denotes the estimated rating that user u gives to t_j.

The prediction results on the training sets of MovieLens is summarized in Table 9. Note that all the MAE values in both tables are averaged over 20 independent runs. From these tables we can clearly observe the superiority of our method.

Table 8 Characteristics of the eachmovie data set

	MovieLens
Number of users	500
Number of items	1000
Average ratio of rated items/users	87.7
Density of data	8.77%
Scale of ratings	5

Table 9 Results comparison of different methods on the MovieLens training data set

Training set	Algs	Given5	Given10	Given20
	PCC	0.935	0.914	0.887
	AM	0.951	0.920	0.893
	PD	0.916	0.890	0.868
ML_200	NMF	0.915	0.889	0.872
	IG	0.872	0.850	0.822
	JNMF	**0.827**	**0.810**	**0.788**
	PCC	0.946	0.918	0.892
	AM	0.984	0.957	0.933
	PD	0.925	0.906	0.884
ML_300	NMF	0.922	0.897	0.878
	IG	0.882	0.863	0.826
	JNMF	**0.852**	**0.830**	**0.809**
	PCC	0.975	0.953	0.912
	AM	1.203	1.034	0.973
	PD	0.933	0.910	0.897
ML_400	NMF	0.948	0.921	0.904
	IG	0.897	0.874	0.832
	JNMF	**0.863**	**0.845**	**0.811**

6 Conclusions and discussions

In this paper, we propose how to apply nonnegative matrix factorization based methods to solve the community discovery problem. We have proposed three concrete algorithms, Symmetric NMF, Asymmetric NMF and Joint NMF to work on undirected networks, directed networks and compound networks, and we also proved the correctness and convergence of those algorithms. Finally the experiments on real world network data sets are presented to show the effectiveness of those methods.

Despite the good aspects, there are still some limitations of the proposed algorithms that are worthy of working on as our future works. (1). The proposed methods strongly depends on the quality of the provided $\mathbf{G}$ (for SNMF), $\mathbf{A}$ (for ANMF) or $\mathbf{U}$, $\mathbf{M}$, $\mathbf{D}$ (for JNMF) matrices. If these matrices cannot well reveal the genuine data relationships, then the algorithm will not produce good results. In the future we will try to design

a recursive scheme to learn the data relationship matrix and the cluster assignment matrix together; (2). The power of 1/4 in ANMF and JNMF will make the algorithms converge somewhat slowly. In the future we will try to construct some other auxiliary functions that can lead to better update rules.

Appendix I: Proof of Theorem 1

Correctness: First we will prove the correctness of the updating rules in Eqs. 6 and 7, i.e., we will show that if they converged solutions, then the final solution would satisfy the KKT condition. By introducing the Lagrangian multipliers $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for the nonnegativity of $\mathbf{X}$ and $\mathbf{S}$, we can construct the Lagrangian function of Eq. 4 as

$$L = \left\|\mathbf{A} - \mathbf{XSX}^\top\right\|^2 - tr(\boldsymbol{\beta}_1\mathbf{X}^\top) - tr(\boldsymbol{\beta}_2\mathbf{S}^\top)$$

Then we have

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{X}} &= 2\left(\mathbf{XSX}^\top\mathbf{XS}^\top + \mathbf{XS}^\top\mathbf{X}^\top\mathbf{XS} - \mathbf{AXS}^\top - \mathbf{A}^\top\mathbf{XS}\right) - \boldsymbol{\beta}_1\\ \frac{\partial L}{\partial \mathbf{S}} &= 2\left(\mathbf{X}^\top\mathbf{XSX}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{AX}\right) - \boldsymbol{\beta}_2\end{aligned}$$

Let $\frac{\partial L}{\partial \mathbf{X}} = 0$ and $\frac{\partial L}{\partial \mathbf{S}} = 0$, and follow the KKT complementary slackness condition, we have

$$\begin{aligned}\boldsymbol{\beta}_{1_{ij}}\mathbf{X}_{ij} &= 2\left(\mathbf{XSX}^\top\mathbf{XS}^\top + \mathbf{XS}^\top\mathbf{X}^\top\mathbf{XS} - \mathbf{AXS}^\top - \mathbf{A}^\top\mathbf{XS}\right)_{ij}\mathbf{X}_{ij} = 0\\ \boldsymbol{\beta}_{2_{ij}}\mathbf{S}_{ij} &= 2\left(\mathbf{X}^\top\mathbf{XSX}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{AX}\right)_{ij}\mathbf{S}_{ij} = 0\end{aligned}$$

Then we can see that the updating rules Eqs. 6 and 7 satisfy the above KKT conditions. Moreover, since matrices $\mathbf{A}, \mathbf{S}, \mathbf{X}$ are all nonnegative during the updating process, so the final $\mathbf{X}$ and $\mathbf{S}$ would also be nonnegative. Therefore we prove the correctness of our algorithm.

Convergence: Now we prove the convergence of our algorithm. Following (Lee and Seung 2000), we will use the *auxiliary function* approach to achieve this goal. Fixing $\mathbf{S}$, since

$$\begin{aligned}\ell(\mathbf{A}, \mathbf{XSX}^\top) &= \mathrm{tr}\left(\mathbf{XSX}^\top\mathbf{XS}^\top\mathbf{X}^\top\right) - 2\mathrm{tr}\left(\mathbf{A}^\top\mathbf{XSX}^\top\right) + \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)\\ &\le \frac{1}{2}\mathrm{tr}\left(\mathbf{PS}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}^\top\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{PS}^\top\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}\right)\\ &\quad -2\mathrm{tr}\left(\mathbf{A}^\top\mathbf{XSX}^\top\right) + \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right) \quad \text{(by Lemma 6)}\\ &\le \frac{1}{2}\mathrm{tr}\left(\mathbf{RS}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}^\top\widetilde{\mathbf{X}}^\top\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{RS}^\top\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}\widetilde{\mathbf{X}}^\top\right)\\ &\quad -2\mathrm{tr}\left(\mathbf{A}^\top\widetilde{\mathbf{X}}\mathbf{SZ}^\top\right) - 2\mathrm{tr}\left(\mathbf{A}^\top\mathbf{Z}\mathbf{S}\widetilde{\mathbf{X}}^\top\right)\end{aligned}$$

$$
\begin{aligned}
&-2\mathrm{tr}\left(\mathbf{A}^\top \widetilde{\mathbf{X}}\mathbf{S}\widetilde{\mathbf{X}}^\top\right) + \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right) \quad \text{(by Lemmas 7 and 3)}\\
&\stackrel{\text{def}}{=} \mathcal{Q}(\mathbf{X}, \widetilde{\mathbf{X}})
\end{aligned}
$$

where $\mathbf{P}_{kl} = \left[\mathbf{X}^\top\mathbf{X}\right]^2_{kl} \Big/ \left[\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right]_{kl}$, $\mathbf{R}_{ik} = [\mathbf{X}]^4_{ik} \Big/ \left[\widetilde{\mathbf{X}}\right]^3_{kl}$, and $\mathbf{Z}_{ij} = \widetilde{\mathbf{X}}_{ij} \ln(\mathbf{X}_{ij}/\widetilde{\mathbf{X}}_{ij})$. Thus $\mathcal{Q}(\mathbf{X}, \widetilde{\mathbf{X}})$ satisfied the conditions of being an auxiliary function for $\mathbf{X}$. Then let $\{\mathbf{X}^{(t)}\}$ be the series of matrices obtained from the iterations of Eqs. 6 and 7, where the superscript (t) denotes the iteration number. Now let's define

$$
\mathbf{X}^{(t+1)} = arg\min_{\mathbf{X}} \mathcal{Q}(\mathbf{X}, \mathbf{X}^{(t)})
$$

By the construction of $\mathcal{Q}$, we have

$$
\mathcal{Q}(\mathbf{X}^{(t)}, \mathbf{X}^{(t)}) \geq \mathcal{Q}(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}) \geq \mathcal{Q}(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t+1)})
$$

Then the loss $\ell\left(\mathbf{G}, \mathbf{X}^{(t)}\mathbf{S}\left(\mathbf{X}^{(t)}\right)^\top\right) = \mathcal{Q}(\mathbf{X}^{(t)}, \mathbf{X}^{(t)})$ is monotonically decreasing. Let $\mathcal{L}(\mathbf{X}) = \mathcal{Q}(\mathbf{X}, \widetilde{\mathbf{X}})$. Then we can find the solution for $\min_{\mathbf{X}} \mathcal{Q}(\mathbf{X}, \widetilde{\mathbf{X}})$ by the following Karush-Kuhn-Tucker condition

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{X}_{ik}} = 2\frac{\mathbf{X}^3_{ik}}{\widetilde{\mathbf{X}}^3_{ik}}\left[\widetilde{\mathbf{X}}\mathbf{S}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}^\top + \widetilde{\mathbf{X}}\mathbf{S}^\top\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\mathbf{S}\right]_{ik}\\
-2\frac{\widetilde{\mathbf{X}}_{ik}}{\mathbf{X}_{ik}}\left[\mathbf{A}^\top\widetilde{\mathbf{X}}\mathbf{S} + \mathbf{A}\widetilde{\mathbf{X}}\mathbf{S}^\top\right]_{ik} = 0
\end{aligned}
$$

So we have the update rule for $\mathbf{X}$ as Eq. 6.

Fixing $\mathbf{X}$. If $\mathbf{S}_{kl}$ is zero, $\mathbf{S}_{kl}$ is a fixed point and satisfies KKT condition. Now we assume $\mathbf{S}_{kl}$ is positive.

$$
\begin{aligned}
\ell(\mathbf{A}, \mathbf{X}\mathbf{S}\mathbf{X}^\top) &= \mathrm{tr}\left(\mathbf{X}\mathbf{S}\mathbf{X}^\top\mathbf{X}\mathbf{S}^\top\mathbf{X}^\top\right) - 2\mathrm{tr}\left(\mathbf{A}^\top\mathbf{X}\mathbf{S}\mathbf{X}^\top\right) + \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)\\
&\leq \mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\widetilde{\mathbf{S}}\mathbf{X}^\top\mathbf{X}\mathbf{T}^\top\right) - 2\mathrm{tr}\left(\mathbf{A}^\top\mathbf{X}\mathbf{S}\mathbf{X}^\top\right) + \mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)\\
&\stackrel{\text{def}}{=} \mathcal{Q}(\mathbf{S}, \widetilde{\mathbf{S}})
\end{aligned}
$$

where $\mathbf{T}_{kl} = \mathbf{S}^2_{kl}/\widetilde{\mathbf{S}}_{kl}$. Let $\mathcal{L}(\mathbf{S}) = \mathcal{Q}(\mathbf{S}, \widetilde{\mathbf{S}})$ be the auxiliary function of $\mathbf{S}$, then the Karush-Kuhn-Tucker conditions is

$$
\frac{\partial \mathcal{L}}{\partial \mathbf{S}_{kl}} = 2\frac{\mathbf{S}_{kl}}{\widetilde{\mathbf{S}}_{kl}}\left[\mathbf{X}^\top\mathbf{X}\widetilde{\mathbf{S}}\mathbf{X}^\top\mathbf{X}\right]_{kl} - 2\left[\mathbf{X}^\top\mathbf{A}\mathbf{X}\right]_{kl} = 0.
$$

We have the update rule for $\mathbf{S}$ as Eq. 7.

Therefore, let $\ell(\mathbf{X}, \mathbf{S}) = \ell(\mathbf{G}, \mathbf{X}\mathbf{S}\mathbf{X}^\top)$, then we have

$$
\ell(\mathbf{X}^{(0)}, \mathbf{S}^{(0)}) \geq \ell(\mathbf{X}^{(1)}, \mathbf{S}^{(0)}) \geq \ell(\mathbf{X}^{(1)}, \mathbf{S}^{(1)}) \geq \ldots
$$

So $\ell(\mathbf{X}, \mathbf{S})$ is monotonically decreasing. Since $\ell(\mathbf{X}, \mathbf{S})$ is obviously bounded below, the theorem is proved. □

Appendix II

First we decompose

$$\begin{aligned} J &= \|\mathbf{M} - \mathbf{X}\|^2 + \alpha\|\mathbf{U} - \mathbf{X}\mathbf{X}^\top\|^2 + \beta\|\mathbf{D} - \mathbf{X}^\top\mathbf{X}\|^2 \\ &= \mathrm{tr}\left(\mathbf{M}^\top\mathbf{M} + \alpha\mathbf{U}^\top\mathbf{U} + \beta\mathbf{D}^\top\mathbf{D}\right) \\ &\quad +(\alpha+\beta)\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\right) \\ &\quad -\mathrm{tr}\left(2\mathbf{M}^\top\mathbf{X} + 2\alpha\mathbf{U}\mathbf{X}\mathbf{X}^\top + (2\beta\mathbf{D} - \mathbf{I})\mathbf{X}^\top\mathbf{X}\right) \end{aligned}$$

Let $\hat{\mathbf{D}} = 2\beta\mathbf{D} - \mathbf{I}$, then we have
Convergence: By introducing the Lagrangian multiplier $\boldsymbol{\gamma}$ for the nonnegativity of $\mathbf{X}$, we can construct the Lagrangian function of J as

$$L = J - \mathrm{tr}\left(\boldsymbol{\gamma}\mathbf{X}^\top\right)$$

Then

$$\frac{\partial L}{\partial \mathbf{X}} = 4(\alpha+\beta)(\mathbf{X}\mathbf{X}^\top\mathbf{X}) - (2\mathbf{M} + 4\alpha\mathbf{U}\mathbf{X} + 2\mathbf{X}\hat{\mathbf{D}}) - \boldsymbol{\gamma}$$

Let $\frac{\partial L}{\partial \mathbf{X}} = 0$, and follow the KKT complementary slackness condition, we have

$$\begin{aligned} \gamma_{ij}\mathbf{X}_{ij} &= \left[4(\alpha+\beta)(\mathbf{X}\mathbf{X}^\top\mathbf{X}) - (2\mathbf{M} + 4\alpha\mathbf{U}\mathbf{X} + 2\mathbf{X}\hat{\mathbf{D}})\right]_{ij}\mathbf{X}_{ij} \\ &= 0 \end{aligned}$$

Then we can see that the update rule Eq. 8 satisfies the above condition.
Correctness: From Lemmas 5 and 6, we have

$$\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}\right) \le \mathrm{tr}\left(\mathbf{P}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right) \le \mathrm{tr}\left(\mathbf{R}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)$$

where $\mathbf{P}_{ij} = \left[\mathbf{X}^\top\mathbf{X}\right]_{ij}^2 / \left[\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right]_{ij}$, $\mathbf{R}_{ij} = [\mathbf{X}]_{ij}^4 / \left[\widetilde{\mathbf{X}}\right]_{ij}^3$. On the other hand, from Lemmas 2 and 3,

$$\begin{aligned} \mathrm{tr}\left(\mathbf{M}^\top\mathbf{X}\right) &\ge \mathrm{tr}\left(\mathbf{M}^\top\mathbf{Z}\right) + \mathrm{tr}\left(\mathbf{M}^\top\widetilde{\mathbf{X}}\right) \\ \mathrm{tr}\left(\mathbf{U}\mathbf{X}\mathbf{X}^\top\right) &\ge \mathrm{tr}\left(\widetilde{\mathbf{X}}^\top\mathbf{U}\mathbf{Z}\right) + \mathrm{tr}\left(\mathbf{Z}^\top\mathbf{U}\widetilde{\mathbf{X}}\right) + \mathrm{tr}\left(\widetilde{\mathbf{X}}^\top\mathbf{U}\widetilde{\mathbf{X}}\right) \\ \mathrm{tr}\left(\hat{\mathbf{D}}\mathbf{X}^\top\mathbf{X}\right) &\ge \mathrm{tr}\left(\hat{\mathbf{D}}\widetilde{\mathbf{X}}^\top\mathbf{Z}\right) + \mathrm{tr}\left(\hat{\mathbf{D}}\mathbf{Z}^\top\widetilde{\mathbf{X}}\right) + \mathrm{tr}\left(\hat{\mathbf{D}}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right) \end{aligned}$$

where $\mathbf{Z}_{ij} = \widetilde{\mathbf{X}}_{ij} \ln\left(\mathbf{X}_{ij}/\widetilde{\mathbf{X}}_{ij}\right)$. Then we have

$$\begin{aligned} J \;\leq\; & \mathrm{tr}\left(\mathbf{M}^\top\mathbf{M} + \alpha\mathbf{U}^\top\mathbf{U} + \beta\mathbf{D}^\top\mathbf{D}\right) \\ & + (\alpha+\beta)\mathrm{tr}\left(\mathbf{R}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right) - 2\mathrm{tr}\left(\mathbf{M}^\top\widetilde{\mathbf{X}} + \mathbf{M}^\top\mathbf{Z}\right) \\ & + 2\alpha\mathrm{tr}\left(\widetilde{\mathbf{X}}^\top\mathbf{U}^\top\mathbf{Z} + \mathbf{Z}^\top\mathbf{U}^\top\widetilde{\mathbf{X}} + \widetilde{\mathbf{X}}^\top\mathbf{U}^\top\widetilde{\mathbf{X}}\right) \\ & - \mathrm{tr}\left(\hat{\mathbf{D}}\widetilde{\mathbf{X}}^\top\mathbf{Z} + \hat{\mathbf{D}}\mathbf{Z}^\top\hat{\mathbf{X}} + \hat{\mathbf{D}}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right) \\ \stackrel{\mathrm{def}}{=}\; & \mathcal{Q}(\mathbf{X}, \widetilde{\mathbf{X}}) \end{aligned}$$

which can be served as the auxiliary function of $\mathbf{X}$. To find its local minimum, we can write the Karush-Kuhn-Tucker condition it should satisfy as

$$\frac{\partial\mathcal{Q}}{\partial\mathbf{X}_{ij}} = 4(\alpha+\beta)\frac{\mathbf{X}_{ij}^3}{\widetilde{\mathbf{X}}_{ij}^3}\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)_{ij} - \frac{\widetilde{\mathbf{X}}_{ij}}{\mathbf{X}_{ij}}\left(2\mathbf{M} + 4\alpha\mathbf{U}\widetilde{\mathbf{X}} + 2\widetilde{\mathbf{X}}\hat{\mathbf{D}}\right) \tag{10}$$

Therefore we get

$$\mathbf{X}_{ij} \leftarrow \mathbf{X}_{ij}\left(\frac{\left[\mathbf{M} + 2\alpha\mathbf{U}\mathbf{X} + \mathbf{X}\hat{\mathbf{D}}\right]_{ij}}{2(\alpha+\beta)\left[\mathbf{X}\mathbf{X}^\top\mathbf{X}\right]_{ij}}\right)^{\frac{1}{4}} \tag{11}$$

Appendix: Lemmas

Lemma 1 *The following inequality holds,*

$$-\ln\sum_i x_i \;\leq\; -\sum_i \alpha_i \ln\frac{x_i}{\alpha_i}$$

The equality holds when $\alpha_i = x_i/(\sum_i x_i)$.

Proof Because of the convexity of logarithm, we have

$$-\ln\sum_i \alpha_i z_i \;\leq\; -\sum_i \alpha_i \ln z_i,$$

where $z_i > 0$ and $\sum_i \alpha_i = 1$. The equality holds when z_i's are equal to each other. Plugging $z_i = x_i/\alpha_i$, we obtain the lemma. □

Lemma 2 *Matrices* $\mathbf{A}$ *and* $\mathbf{X}$ *are nonnegative.* $\widetilde{\mathbf{X}}$ *is positive. We have*

$$-\mathrm{tr}\left(\mathbf{A}^\top\mathbf{X}\right) \leq -\mathrm{tr}\left(\mathbf{A}^\top\mathbf{L}\right) - \mathrm{tr}\left(A^\top\widetilde{\mathbf{X}}\right)$$

where $\mathbf{L}_{ij} = \widetilde{\mathbf{X}}_{ij} \ln \mathbf{X}_{ij} / \widetilde{\mathbf{X}}_{ij}$. *The equality holds when* $\widetilde{\mathbf{X}} = \mathbf{X}$.

Proof By Lemma 1, we have that

$$\begin{aligned} -\mathrm{tr}\left(\mathbf{A}^\top \mathbf{X}\right) &= -vec(\mathbf{A})^\top vec(\mathbf{X}) \\ &\leq -vec(\mathbf{A})^\top vec(\mathbf{L}) - vec(\mathbf{A})^\top vec(\widetilde{\mathbf{X}}) \\ &= -\mathrm{tr}\left(\mathbf{A}^\top \mathbf{L}\right) - \mathrm{tr}\left(\mathbf{A}^\top \widetilde{\mathbf{X}}\right) \end{aligned}$$

□

Lemma 3 *Vector* $\mathbf{b}$ *and* $\mathbf{x}$ *are nonnegative,* $\widetilde{\mathbf{x}}$ *is nonnegative. Then*

$$-\mathbf{b}^\top \mathbf{x} \ln \mathbf{b}^\top \mathbf{x} \leq -\mathbf{b}^\top \widetilde{\mathbf{x}} \ln \mathbf{b}^\top \mathbf{x} + \mathbf{b}^\top (\mathbf{x} - \widetilde{\mathbf{x}})$$

The equality holds when $\widetilde{\mathbf{x}} = \mathbf{x}$.

Proof Since $\xi \ln(\xi/\zeta) - \xi + \zeta \geq 0$. □

Lemma 4 *Matrices* $\mathbf{A}, \mathbf{B}, \mathbf{X}$ *are nonnegative. Matrix* $\widetilde{\mathbf{X}}$ *is positive. We have*

$$\begin{aligned} -\mathrm{tr}\left(\mathbf{B}\mathbf{X}^\top \mathbf{A}\mathbf{X}\right) \leq & -\mathrm{tr}\left(\mathbf{B}\widetilde{\mathbf{X}}^\top \mathbf{A}\mathbf{Z}\right) - \mathrm{tr}\left(\mathbf{B}\mathbf{Z}^\top \mathbf{A}\widetilde{\mathbf{X}}\right) \\ & -\mathrm{tr}\left(\mathbf{B}\widetilde{\mathbf{X}}^\top \mathbf{A}\widetilde{\mathbf{X}}\right) \end{aligned} \tag{12}$$

where $\mathbf{Z}_{ij} = \widetilde{\mathbf{X}}_{ij} . \ln \mathbf{X}_{ij} / \widetilde{\mathbf{X}}_{ij}$. *The equality holds when* $\widetilde{\mathbf{X}} = \mathbf{X}$.

Proof Since

$$-\mathrm{tr}\left(\mathbf{B}\mathbf{X}^\top \mathbf{A}\mathbf{Z}\right) = -vec(\mathbf{X})^\top (\mathbf{B}^\top \otimes \mathbf{A}) vec(\mathbf{Z})$$

Then following Lemma 2, we prove this lemma. □

Lemma 5 *Matrix* $\mathbf{A}$ *is nonnegative. Vector* $\mathbf{x}$ *is nonnegative, and* $\widetilde{\mathbf{x}}$ *is positive. We have*

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \frac{1}{2}\mathbf{y}^\top \mathbf{A}\widetilde{\mathbf{x}} + \frac{1}{2}\widetilde{\mathbf{x}}^\top \mathbf{A}\mathbf{y}$$

where $\mathbf{y}_i = \mathbf{x}_i^2 / \widetilde{\mathbf{x}}_i$. *The equality holds when* $\widetilde{\mathbf{x}} = \mathbf{x}$.

Proof Let $q_i = \mathbf{x}_i / \widetilde{\mathbf{x}}_i$, then we have

$$\begin{aligned} & \frac{1}{2}\mathbf{y}^\top \mathbf{A}\widetilde{\mathbf{x}} + \frac{1}{2}\widetilde{\mathbf{x}}^\top \mathbf{A}\mathbf{y} - \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ &= \frac{1}{2}\widetilde{\mathbf{x}}^\top \mathbf{Q}^2 \mathbf{A}\widetilde{\mathbf{x}} + \frac{1}{2}\widetilde{\mathbf{x}}^\top \mathbf{A}\mathbf{Q}^2 \widetilde{\mathbf{x}} - \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ &= \frac{1}{2}\sum_{ij} \mathbf{A}_{ij} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_j (q_i - q_j)^2 \geq 0 \end{aligned}$$

Plugging in $\widetilde{\mathbf{x}} = \mathbf{x}$, we can verify the equality. □

Lemma 6 *Matrices* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ *are nonnegative. Matrix* $\widetilde{\mathbf{X}}$ *is positive. We have*

$$\text{tr}\left(\mathbf{B}\mathbf{X}^\top\mathbf{A}\mathbf{X}\right) \leq \frac{1}{2}\text{tr}\left(\mathbf{B}\mathbf{Y}^\top\mathbf{A}\widetilde{\mathbf{X}} + \mathbf{B}\widetilde{\mathbf{X}}^\top\mathbf{A}\mathbf{Y}\right)$$

where $\mathbf{Y}_{ij} = \mathbf{X}_{ij}^2/\widetilde{\mathbf{X}}_{ij}$. *The equality holds when* $\widetilde{\mathbf{X}} = \mathbf{X}$.

Proof By Lemma 5, we have

$$\begin{aligned}
&\text{tr}\left(\mathbf{B}\mathbf{X}^\top\mathbf{A}\mathbf{X}\right) = vec(\mathbf{X})^\top(\mathbf{B}^\top \otimes \mathbf{A})vec(\mathbf{X}) \\
\leq\ &\frac{1}{2}vec(\mathbf{Y})^\top(\mathbf{B}^\top \otimes \mathbf{A})vec(\widetilde{\mathbf{X}}) + \frac{1}{2}\widetilde{\mathbf{X}}^\top(\mathbf{B}^\top \otimes \mathbf{A})vec(\mathbf{Y}) \\
=\ &\frac{1}{2}\text{tr}\left(\mathbf{B}\mathbf{Y}^\top\mathbf{A}\widetilde{\mathbf{X}} + \mathbf{B}\widetilde{\mathbf{X}}^\top\mathbf{A}\mathbf{Y}\right)
\end{aligned}$$

Plugging in $\widetilde{\mathbf{x}} = \mathbf{x}$, we can verify the equality. □

Lemma 7 *Matrix* $\mathbf{A}$ *is nonnegative symmetric. Matrix* $\mathbf{X}$ *is nonnegative. Matrix* $\widetilde{\mathbf{X}}$ *is positive. We have*

$$\text{tr}\,(\mathbf{P}\mathbf{A}) \leq \text{tr}\left(\mathbf{R}\mathbf{A}\widetilde{\mathbf{X}}^\top\right)$$

where $\mathbf{P}_{kl} = \left[\mathbf{X}^\top\mathbf{X}\right]_{kl}^2 / \left[\mathbf{X}^\top\mathbf{X}\right]_{kl}$ *and* $\mathbf{R}_{ik} = [\mathbf{X}]^4/[\widetilde{\mathbf{X}}]_{kl}^3$. *The equality holds when* $\widetilde{\mathbf{X}} = \mathbf{X}$.

Proof Let $\mathbf{Y}_{ik} = \mathbf{X}_{ik}^2/\widetilde{\mathbf{X}}_{ik}$. Since

$$\left(\sum_i \alpha_i x_i\right)^2 \leq \left(\sum_i \alpha_i\right)\left(\sum_i \alpha_i x_i^2\right),$$

we have

$$\mathbf{P}_{kl} \leq \sum_i \frac{\mathbf{X}_{ik}^2\mathbf{X}_{il}^2}{\widetilde{\mathbf{X}}_{ik}\widetilde{\mathbf{X}}_{il}} = \mathbf{Y}^\top\mathbf{Y} \tag{13}$$

Then, by Lemma 6, we have

$$\text{tr}\,(\mathbf{P}\mathbf{A}) \leq \text{tr}\left(\mathbf{Y}^\top\mathbf{Y}\mathbf{A}\right) = \text{tr}\left(\mathbf{Y}\mathbf{A}\mathbf{Y}^\top\right) \leq \mathbf{R}\mathbf{A}\widetilde{\mathbf{X}}^\top$$

where $\mathbf{R}_{ik} = \mathbf{Y}_{ik}^2/\widetilde{\mathbf{X}}_{ik} = \mathbf{X}_{ik}^4/\widetilde{\mathbf{X}}_{ik}^3$. Plugging in $\widetilde{\mathbf{X}} = \mathbf{X}$, we can verify the equality. □

Acknowledgment The work is partially supported by NSF grants IIS-0546280, CCF-0830659, and DMS-0915110.

References

Albert R, Jeong H, Barabasi A-L (1999) The diameter of the world wide web. Nature 401:130

Amaral LAN, Scala A, Bartheélémy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97:11149–11152

Barthelemy M, Amaral LAN (1999) Small-world networks: evidence for a crossover picture. Phys Rev Lett 82:3180

Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 102(12):4164–4169

Chen G, Wang F, Zhang C (2007) Collaborative filtering using orthogonal nonnegative matrix tri-factorization. In: ICDM workshops on high performance computing, pp 303–308

Ding C, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of SIAM international conference on data mining, pp 606–610

Ding C, Li T, Jordan MI (2006a) Convex and semi-nonnegative matrix factorizations. LBNL Tech Report 60428

Ding C, Li T, Peng W, Park H (2006b) Orthogonal nonnegative matrix tri-factorizations for clustering. In: SIGKDD, pp 126–135

Ding C, Li T, Peng W (2008) On the equivalence between nonnegative matrix factorization and probabilistic latent semantic indexing. Comput Stat Data Anal 52(1):155–173

Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. In: SIGCOMM, pp 251–262

Flake GW, Lawrence S, Giles CL (2000) Efficient identification of web communities. In: SIGKDD, pp 150–160

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

Hofmann T, Puzicha J (1999) Latent class models for collaborative filtering. In: IJCAI, pp 688–693

Ino H, Kudo M, Nakamura A (2005) Partitioning of web graphs by community topology. In: WWW, pp 661–669

Lee DD, Seung HS (1999) Learning the parts of objects by nonnegative matrix factorization. Nature 401:788–791

Lee DD, Seung HS (2000) Algorithms for non-negative matrix factorization, pp 556–562

Long B, Zhang Z, Wu X, Yu PS (2007) Relational clustering by symmetric convex coding. In: Proceedings of the 24th international conference on machine learning, pp 569–576

Miao G, Song Y, Zhang D, Bai H (2008) Parallel spectral clustering algorithm for large-scale community data mining. In: The 17th WWW workshop on social web search and mining (SWSM)

Newman MEJ (2004a) Detecting community structure in networks. Eur Phys J B 38:321–330

Newman MEJ (2004b) Fast algorithm for detecting community structure in very large networks. Phys Rev E 69

Paatero P, Tapper U (1994) Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126

Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818

Pauca VP, Shahnaz F, Berry MW, Plemmons RJ (2004) Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. In: The 4th SIAM international conference on data mining, pp 452–456

Pennock DM, Horvitz E, Lawrence S, Giles CL (2000) Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach. In: UAI

Priebe CE, Conroy JM, Marchette DJ, Park Y (2005) Scan statistics on Enron graphs. In: Proceedings of SIAM international conference on data mining

Resnick P., Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews, pp 175–186

Ruan J, Zhang W (2007) An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In: ICDM

Scott J (2000) Social network analysis: a handbook, 2nd edn. Sage Publications, London

Sharan R et al (2005) From the cover: conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA 102(6):1974–1979

Strehl A, Ghosh J, Cardie C (2002) Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
Wang F, Ma S, Yang L, Li T (2006) Recommendation on item graphs. In: ICDM, pp 1119–1123
Wang D, Li T, Zhu S, Ding CHQ (2008a) Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp 307–314
Wang F, Li T, Zhang C (2008b) Semi-supervised clustering via matrix factorization. In: The 8th SIAM international conference on data mining
Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
Weiss Y (1999) Segmentation using eigenvectors: a unifying view. In: ICCV, pp 975–982
Xie YL, Hopke PK, Paatero P (1999) Positive matrix factorization applied to a curve resolution problem. J Chemometr 12(6):357–364
Yu K, Tresp V (2005) Learning to learn and collaborative filtering. In: NIPS workshop on inductive transfer: 10 years later
Zhou D, Huang J, Schölkopf B (2005) Learning from labeled and unlabeled data on a directed graph. In: Proceedings of the 22nd international conference on machine learning, pp 1036–1043
Zhang H, Giles CL, Foley HC, Yen J (2007) Probabilistic community discovery using hierarchical latent gaussian mixture model. In: AAAI, pp 663–668