2011 11th IEEE International Conference on Data Mining

Nonnegative Matrix Tri-Factorization Based High-Order Co-Clustering and Its Fast Implementation

Hua Wang, Feiping Nie, Heng Huang, Chris Ding
Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
{*huawangcs, feipingnie*}*@gmail.com*, {*heng, chqding*}*@uta.edu*

***Abstract*—The fast growth of Internet and modern technologies has brought data involving objects of multiple types that are related to each other, called as *Multi-Type Relational data*. Traditional clustering methods for single-type data rarely work well on them, which calls for new clustering techniques, called as *high-order co-clustering (HOCC)*, to deal with the multiple types of data at the same time. A major challenge in developing HOCC methods is how to effectively make use of all available information contained in a multi-type relational data set, including both *inter-type* and *intra-type* relationships. Meanwhile, because many real world data sets are often of large sizes, clustering methods with computationally efficient solution algorithms are of great practical interest. In this paper, we first present a general HOCC framework, named as Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF), for simultaneous clustering of multi-type relational data. The proposed O-NMTF approach employs Nonnegative Matrix Tri-Factorization (NMTF) to simultaneously cluster different types of data using the inter-type relationships, and incorporate intra-type information through manifold regularization, where, different from existing works, we emphasize the importance of the orthogonalities of the factor matrices of NMTF. Based on O-NMTF, we further develop a novel Fast Nonnegative Matrix Tri-Factorization (F-NMTF) approach to deal with large-scale data. Instead of constraining the factor matrices of NMTF to be nonnegative as in existing methods, F-NMTF constrains them to be cluster indicator matrices, a special type of nonnegative matrices. As a result, the optimization problem of the proposed method can be decoupled, which results in subproblems of much smaller sizes requiring much less matrix multiplications, such that our new algorithm scales well to real world data of large sizes. Extensive experimental evaluations have demonstrated the effectiveness of our new approaches.**

***Keywords*-High-Order Co-Clustering, Multi-Type Relational Data, Nonnegative Matrix Tri-Factorization, Cluster Indicator Matrix**

I. Introduction

Most traditional clustering algorithms concentrate on dealing with *homogeneous* data, in which all the objects of interest are of one single type. Recently, the rapid progress of modern technologies, especially for those related to Internet, has brought new data much richer in structure, involving objects of multiple types that are related to each other. For example, in a Web search system as illustrated in Figure 1, we have four different types of data entities including words, Web pages, search queries and Web users. Each of these four types of data objects has their own attributes. Meanwhile different types of data are also interrelated to each other in various ways, *e.g.*, Web pages and words are related via co-occurrences. Such data sets are often called as *Multi-Type Relational data* [1], [2].

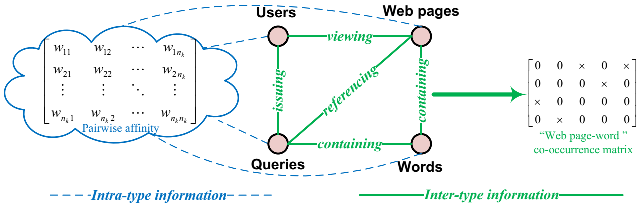


Figure 1. A Web search system is a typical example of multi-type relational data, involving (1) *Inter-type relationships* (green solid lines): the relations between objects in different data types. (2) *Intra-type relationships* (blue dashed lines): the relations between different objects in a same data type.

Different from traditional homogeneous data of one single type, multi-type relational data contains more information with richer structures. Typically, we consider the following two forms of information of a multi-type relational data set:

- **Inter-type relationships** characterize the relations between data objects from different types, such as the co-occurrences between data objects from different types as illustrated by the green solid lines in Figure 1.
- **Intra-type relationships** characterize the native relations between objects within one data type, *e.g.*, as illustrated by the blue dashed lines in Figure 1, the internet hyperlinks among Web pages, the pairwise affinities among users that are induced from user attributes, *etc*.

The rich structures of multi-type relational data provide a potential opportunity to improve the clustering accuracy, which, however, also present a new challenge on how to effectively use all available information contained in a multi-type relational data set. Similar to co-clustering for two-type relational data that makes use of the interrelatedness between the two types of data, simultaneous clustering of multi-type relational data aims to exploit both inter-type and intra-type information, which is called as *high-order co-clustering (HOCC)*. In this paper, we tackle this new, yet important, problem, where we also take into account large-scale data for practical use.

Given the inter-type relationships and the intra-type in-

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.109

formation, we first present a simple but general HOCC framework, named as Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF), for simultaneous clustering on multi-type relational data. In the proposed O-NMTF approach, we use Nonnegative Matrix Tri-Factorization (NMTF) [3] to simultaneously cluster different types of data upon the inter-type relation matrices. Meanwhile, the optional intra-type information for different types of data in form of pairwise affinity is incorporated as manifold regularization to NMTF, where, different from existing manifold regularized NMTF methods, we emphasize the importance of the orthogonality on the factor matrices.

Because existing solution algorithms to NMTF problems, as well as ours to the proposed O-NMTF approach, are usually computationally prohibitive due to involving intensive multiplications on matrices of large sizes, instead of constraining the factor matrices of NMTF to be nonnegative, we further propose a novel Fast Nonnegative Matrix Tri-Factorization (F-NMTF) approach to constrain them to be cluster indicator matrices, a special type of nonnegative matrices. With this new constraint, the optimization problem can be decoupled into a number of subproblems of much smaller sizes, which require much less matrix multiplications. Consequently, our new algorithm is computationally efficient, which makes it of particular use in clustering large-scale multi-type relational data in real world applications.

We summarize our contributions as following.

1) We present a simple, yet effective, framework to tackle the complicated problem of high-order co-clustering of multi-type relational data, which aims to better utilize both inter-type and intra-type relationships of a multi-type relational data set.
2) Different from existing manifold regularized Nonnegative Matrix Factorization (NMF) methods [2], [4], [5], we emphasize the importance of the orthogonalities of the factor matrices, both theoretically and empirically.
3) Instead of enforcing traditional nonnegative constraints on the factor matrices of NMTF, we constrain them to be cluster indicator matrices, a special type of nonnegative matrices. As a result, the optimization problem of the proposed F-NMTF approach can be decoupled into subproblems with much smaller sizes, and the decoupled subproblems involve much less matrix multiplications. Therefore, our approach is computationally efficient and scales well to large-scale real world data. Different from our earlier publication [6] that deals with asymmetric NMTF on rectangle input matrices, in this paper we deal with symmetric square input matrix, which is harder to solve due to the fourth-order term of the factor matrices in the objective.

Notations. Throughout this paper, we denote matrices as uppercase characters and vectors as boldface lowercase characters. The i-th row and j-th column of the matrix M are denoted as $\mathbf{m}_{i\cdot}$ and $\mathbf{m}_{\cdot j}$ respectively. We denote the Frobenius norm and the trace of a matrix as $\|\cdot\|$ and $\mathbf{tr}(\cdot)$ respectively. $M(i,j)$ denotes the (i,j)-th entry of the matrix M, and $\mathbf{v}(i)$ denotes the i-th entry of the vector $\mathbf{v}$.

We denote $\mathbb{R}$ as the real number set, $\mathbb{R}_+$ as the nonnegative real number set, and Ψ as the *cluster indicator matrix* set. An indicator matrix $G \in \Psi^{n \times c}$ is a special type of nonnegative matrix: all the entries of $\mathbf{g}_{i\cdot}$ $(1 \leq i \leq n)$ are equal to 0 except for one and only one entry equal to 1, indicating the cluster membership of the corresponding data point, *i.e.*, $\mathbf{g}_{i\cdot} \in \{0,1\}^c$ and $\sum_j \mathbf{g}_{i\cdot}(j) = 1$.

II. Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF) for High-Order Co-Clustering

In this section, we first briefly review co-clustering of two-type relational data using NMTF, from which we gradually develop the proposed O-NMTF approach for high-order co-clustering of multi-type relational data. Our goal is to employ both inter-type relationships and intra-type relationships of the input data via a compact, yet effective, framework.

Problem formalization. Given a data set with K types of data objects $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K\}$, where $\mathcal{X}_k = \{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_{n_k}^k\}$ represents n_k data objects of the k-th type. Suppose that we have a set of *inter-type* relationship matrices $\{R_{kl} \in \mathbb{R}^{n_k \times n_l}\}$ between different types of data objects and $R_{lk} = R_{kl}^T$, where $R_{kl}(i,j)$ measures how closely $\mathbf{x}_i^k$ is related to $\mathbf{x}_j^l$. Besides, we also have *intra-type* information for each type of the data, *e.g.*, for the k-th type of data we have pairwise affinities $W_k \in \mathbb{R}^{n_k \times n_k}$ between the the data objects in $\mathcal{X}_k$. Our goal is to learn from R_{kl} $(1 \leq k, l \leq K)$ and W_k $(1 \leq k \leq K)$ a model that is able to simultaneously partition the data objects in $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K$ into $c_1, c_2, \ldots, c_K$ disjoint clusters respectively. We denote $n = \sum_k n_k$ and $c = \sum_k c_k$.

A. A Brief Review of Co-Clustering via NMTF

The simplest multi-type relational data involves only two types of objects, which widely appear in real world applications, *e.g.*, *words* and *documents* in document analysis, *users* and *items* in collaborative filtering, *experimental conditions* and *genes* in microarray data analysis. Instead of being independent, the clustering tasks of different types of objects are often closely related. As a result, *co-clustering* methods (also called as *bi-clustering* in some research papers), which simultaneously cluster the both types of data by leveraging their interrelatedness, have been proposed [3], [5], [7]–[9]. Among these methods, NMTF based co-clustering methods have attracted increased attention in recent years due to their mathematical elegance and promising empirical results.

Motivated by the close connection between NMF and K-means clustering [3], [10], Ding *et al.* [3] proposed to use NMTF to simultaneously cluster the rows and columns of a nonnegative input relationship matrix R_{12} by decomposing it into three nonnegative factor matrices, which minimizes:

$$\begin{aligned} J_1 = \left\| R_{12} - G_1 S_{12} G_2^T \right\|^2 , \\ s.t. \quad G_1 \geq 0,\ G_2 \geq 0,\ S_{12} \geq 0 \ , \end{aligned} \tag{1}$$

where $G_1 \in \mathbb{R}_+^{n_1 \times c_1}$ and $G_2 \in \mathbb{R}_+^{n_2 \times c_2}$ are continuous and act as the "soft" cluster indications [10] for $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, and $S_{12} \in \mathbb{R}_+^{c_1 \times c_2}$ absorbs the different scales of R_{12}, G_1 and G_2. The original NMF problem [11] requires R_{12} to be nonnegative. In co-clustering scenarios, however, this constraint (thereby the nonnegativity constraint on S_{12}) can be relaxed [12] to achieve additional flexibility, which leads to the semi-NMTF problem that minimizes:

$$J_2 = \| R_{12} - G_1 S_{12} G_2^T \|^2, \quad s.t. \quad G_1 \geq 0, G_2 \geq 0 \ . \tag{2}$$

Simultaneous clustering of data objects in $\mathcal{X}_1$ and $\mathcal{X}_2$ is hence achieved by solving Eq. (2). Because the rows of the resulted $G_k\,(k \in \{1,2\})$ (with normalization) can be interpreted as the posterior probability for clustering on $\mathcal{X}_k\,(k \in \{1,2\})$ [10], a possible way to obtain the cluster label of $\mathbf{x}_i^k$ is to use the following rule [3], [10], [12], [13]:

$$l\left(\mathbf{x}_i^k\right) = \arg\max_j G_k\left(i, j\right) \ . \tag{3}$$

B. Objective of O-NMTF

A natural generalization of the co-clustering objective in Eq. (2) to simultaneously cluster multi-type relational data, called as *high-order co-clustering* [1], [14]–[17], is to solve the following optimization problem [1], [16], [17]:

$$\begin{aligned} \min\ J_3 = \textstyle\sum_{0<k<l\leq K} \left\| R_{kl} - G_k S_{kl} G_l^T \right\|^2 , \\ s.t. \quad G_k \geq 0,\ \forall\, 0 < k \leq K \ . \end{aligned} \tag{4}$$

In spite of the clear intuition of the above formulation, the generalization of existing NMTF algorithms to solve Eq. (4), however, is not straightforward. Motivated by [7] that deals with bipartite graphs, we consider to solve an equivalent Symmetric NMTF (S-NMTF) problem.

We first introduce the following useful lemma.

Lemma 1: The optimization problem in Eq. (2) can be equivalently solved by the following S-NMTF problem:

$$\min\ J_4 = \| R - GSG^T \|^2, \quad s.t. \quad G \geq 0 \ , \tag{5}$$

in which

$$\begin{aligned} R = \begin{bmatrix} 0^{n_1 \times n_1} & R_{12}^{n_1 \times n_2} \\ R_{21}^{n_2 \times n_1} & 0^{n_2 \times n_2} \end{bmatrix}, \ G = \begin{bmatrix} G_1^{n_1 \times c_1} & 0^{n_1 \times c_2} \\ 0^{n_2 \times c_1} & G_2^{n_2 \times c_2} \end{bmatrix}, \\ S = \begin{bmatrix} 0^{c_1 \times c_1} & S_{12}^{c_1 \times c_2} \\ S_{21}^{c_2 \times c_1} & 0^{c_2 \times c_2} \end{bmatrix} \ , \end{aligned} \tag{6}$$

where the superscripts denote the matrix sizes, and $R_{21} = R_{12}^T$, $S_{21} = S_{12}^T$. $0^{n_1 \times n_1}$ is a matrix with all zero entries.

Proof: Upon the definitions of R, G and S, we derive:

$$\begin{aligned} \left\| R - GSG^T \right\|^2 &= \left\| \begin{bmatrix} 0 & R_{12} \\ R_{12}^T & 0 \end{bmatrix} - \begin{bmatrix} 0 & G_1 S_{12} G_2^T \\ G_2 S_{12}^T G_1^T & 0 \end{bmatrix} \right\|^2 \\ &= 2 \left\| R_{12} - G_1 S_{12} G_2^T \right\|^2 \ , \end{aligned}$$

which proves the lemma. ■

Based upon Lemma 1, we have the following theorem.

Theorem 1: It is equivalent to solve Eq. (4) and to solve the following problem:

$$\min\ J_5 = \| R - GSG^T \|, \quad s.t. \quad G \geq 0 \ , \tag{7}$$

in which

$$\begin{aligned} R &= \begin{bmatrix} 0^{n_1 \times n_1} & R_{12}^{n_1 \times n_2} & \cdots & R_{1K}^{n_1 \times n_K} \\ R_{21}^{n_2 \times n_1} & 0^{n_2 \times n_2} & \cdots & R_{2K}^{n_2 \times n_K} \\ \vdots & \vdots & \ddots & \vdots \\ R_{K1}^{n_K \times n_1} & R_{K2}^{n_K \times n_2} & \cdots & 0^{n_K \times n_K} \end{bmatrix}, \\ G &= \begin{bmatrix} G_1^{n_1 \times c_1} & 0^{n_1 \times c_2} & \cdots & 0^{n_1 \times c_K} \\ 0^{n_2 \times c_1} & G_2^{n_2 \times c_2} & \cdots & 0^{n_2 \times c_K} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_K \times c_1} & 0^{n_K \times c_2} & \cdots & G_K^{n_K \times c_K} \end{bmatrix}, \\ S &= \begin{bmatrix} 0^{c_1 \times c_1} & S_{12}^{c_1 \times c_2} & \cdots & S_{1K}^{c_1 \times c_K} \\ S_{21}^{c_2 \times c_1} & 0^{c_2 \times c_2} & \cdots & S_{2K}^{c_2 \times c_K} \\ \vdots & \vdots & \ddots & \vdots \\ S_{K1}^{c_K \times c_1} & S_{K2}^{c_K \times c_2} & \cdots & 0^{c_K \times c_K} \end{bmatrix}, \end{aligned} \tag{8}$$

where $R_{ji} = R_{ij}^T$ and $S_{ij} = S_{ji}^T$.

The proof of Theorem 1 can be easily obtained by generalizing the proof of Lemma 1. We skip it due to space limit.

Theorem 1 presents a general framework via S-NMTF for high-order co-clustering of multi-type relational data using the inter-type relationship matrices, which can be further developed to incorporate the intra-type information via the manifold regularization as following [4], [5]:

$$\begin{aligned} \min\ J_6 = \left\| R - GSG^T \right\|^2 + \lambda\, \mathbf{tr}\left(G^T L G \right) , \\ s.t.\ G \geq 0,\ G^T D G = I \ , \end{aligned} \tag{9}$$

where $\lambda > 0$ is a tradeoff parameter that balances the relative importance of the two terms, $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ is the normalized graph Laplacian and

$$W = \begin{bmatrix} W_1^{n_1 \times n_1} & 0^{n_1 \times n_2} & \cdots & 0^{n_1 \times n_K} \\ 0^{n_2 \times n_1} & W_2^{n_2 \times n_2} & \cdots & 0^{n_2 \times n_K} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_K \times n_1} & 0^{n_K \times n_2} & \cdots & W_K^{n_K \times n_K} \end{bmatrix} \tag{10}$$

and D is a diagonal matrix with $D\left(i, i\right) = \sum_j W\left(i, j\right)$. Here $W_k \in \mathbb{R}^{n_k \times n_k}\,(1 \leq k \leq K)$ is the pairwise affinity matrix between the data objects of the k-th type.

Through Eq. (9), as one of our contribution, the complicated problem of high-order co-clustering of multi-type relational data is formulated in a simple form as a regularized NMTF problem.

C. Importance of the Orthogonality Constraint in Manifold Regularization

In Eq. (9), we emphasize the orthogonal constraints $G^T DG = I$ on the factor matrix G, which do not appear in related works [4], [5], though orthogonality plays an significant role in NMF [3], [10], [12]. A rigorous orthogonality among columns of G, *i.e.*, $G^T G = I$, enforces sparsity on G [10], [12]. Therefore it provides a *hard* clustering [10], [12] where each data object belongs to only one cluster (due to orthogonality and nonnegativity, each row of G has only one nonzero element). Without orthogonality, however, each row of G could have more nonzero elements. Thus NMF provides a *soft* clustering [10], [12]. The orthogonality constraints $G^T DG = I$ in Eq. (9) is functioning similar to $G^T G = I$ yet enjoys additional benefits due to its additional role in the regularization.

We note that the manifold regularization term in Eq. (9) indeed is the optimization objective of Laplacian based spectral clustering such as normalized cut [18] method. Without the orthogonality constraint $G^T DG = I$, the optimal solution G^* will contains identical columns. This can be seen as follows. We first rewrite

$$\mathbf{tr}\left[G^T \left(D - W\right) G\right] = \textstyle\sum_{j=1}^{c} \mathbf{g}_{\cdot} j^T (D - W) \mathbf{g}_{\cdot} j, \qquad (11)$$

where $\mathbf{g}_{\cdot} j$ is the j-th column of G, and $c = \sum_k c_k$. Without the orthogonality constraint, different columns become independent of each other, and thus reach the same minimum with the same solution, *i.e.*, $\mathbf{g}_{\cdot 1}{}^* = \cdots = \mathbf{g}_{\cdot c}{}^* = \mathbf{e}$, where $\mathbf{e} = [1, \ldots, 1]^T$. This degenerate solution of G is equivalent to randomly assigning labels to data objects, thus the clustering performance is degraded.

We also notice that in the formulation of Eq. (7), the orthogonality is not necessary: without the orthogonality constraint, the optimal solution G^* will have different columns. The main reason for this desirable feature [10] is due to the matrix approximation nature of Eq. (7), as opposed to the trace minimization of Eq. (9).

To summarize, we have shown that (A) orthogonality is not a necessary constraint for Eq. (7), whereas (B) it is an indispensable constraint for the manifold regularized objectives such as Eq. (9).

D. Algorithm to Solve O-NMTF

The computational algorithm of the proposed O-NMTF approach is listed in Algorithm 1. Upon solution, the cluster labels are obtained from the resulted G_k using Eq. (3).

The main challenge to derive Algorithm 1 is the fourth-order matrix polynomial incurred by the symmetric usage of G in Eq. (9). Existing works [3], [10], [16], [19] tackle this difficulty in an intuitive way: solve the problem as a standard NMF problem with two different factor matrices and set them as same in the solution. In this paper, we use a new matrix inequality presented in Lemma 4, which was proposed in our earlier publication in [2], to tackle this difficulty in a principled way.

Algorithm 1: Algorithm to solve F-NMTF in Eq. (22)

```
Data: Relationship matrices: ${R_{ij}}_{1 \le i < j \le K}$
Pairwise affinity matrices: ${W_k}_{1 \le k \le K}$
1. Construct $R, G, S$ as in Eq. (8), and $W$ as in Eq. (10).
2. Initialize $G$ as in [3].
repeat
    3. Compute $S = \left(G^T G\right)^{-1} G^T RG \left(G^T G\right)^{-1}$.
    4. Update $G_{ij} \leftarrow G_{ij} \left[ \frac{(RGS + \lambda WG)_{ij}}{\left(GSG^T GS + DG\Lambda\right)_{ij}} \right]^{\frac{1}{4}}$ where
    $\Lambda = G^T RGS - G^T GSG^T GS + \lambda G^T WG$.
until Converges
Result: Cluster indicator matrices: ${G_k}_{1 \le k \le K}$
```

Correctness of the Algorithm. The following theorem guarantees the correctness of Algorithm 1.

Theorem 2: If the update rules of G and S in Algorithm 1 converges, the final solution satisfies the KKT condition.

The proof of Theorem 2 can be obtained following the same way as in [3], [10], [12]. We skip it due to space.

Convergence of the Algorithm. We use the auxiliary function approach [11] to prove the convergence of Algorithm 1.

Lemma 2: [11] $Z(h, h')$ is an auxiliary function of $F(h)$ if the conditions $Z(h, h') \ge F(h)$ and $Z(h, h') = F(h)$ are satisfied. [11] If Z is an auxiliary function for F, then F is non-increasing under the update $h^{(t+1)} = \arg\min_h Z(h, h')$.

Lemma 3: [3] For any matrices $A \in \mathbb{R}_+^{n \times n}$, $B \in \mathbb{R}_+^{k \times k}$, $S \in \mathbb{R}_+^{n \times k}$ and $S' \in \mathbb{R}_+^{n \times k}$, and A and B are symmetric, the following inequality holds:

$$\sum_{ip} \frac{(AS'B)_{ip} S_{ip}^2}{S'_{ip}} \ge \mathbf{tr}\left(S^T ASB\right) \quad . \qquad (12)$$

Lemma 4: [2] For any nonnegative symmetric matrices $A \in \mathbb{R}_+^{k \times k}$ and $B \in \mathbb{R}_+^{k \times k}$, for $H \in \mathbb{R}_+^{n \times k}$ the following inequality holds:

$$\mathbf{tr}\left(HAH^T HBH^T\right) \le \sum_{ik} \left(\tfrac{H' AH'^T H' B + H' BH'^T H' A}{2}\right)_{ik} \frac{H_{ik}^4}{H_{ik}'^3} \; . \qquad (13)$$

Now we prove the convergence of Algorithm 1.

Theorem 3: Let

$$J(G) = \mathbf{tr}\left(-2RGSG^T + GSG^T GSG^T - 2\lambda G^T WG + 2\Lambda G^T DG\right),$$

then the following function

$$Z\left(G,G'\right) =$$
$$-\,2\sum_{ijkl} G'_{ji}S_{jk}G'_{kl}R_{li}\left(1+\log\frac{G_{ji}G_{kl}}{G'_{ij}G'_{kl}}\right)+\sum_{ij}\left(G'SG'^{T}G'S\right)_{ij}\frac{G_{ij}^{4}}{G_{ij}'^{3}}$$
$$-\,2\lambda\sum_{ijk} G'_{ji}W_{jk}G'_{ki}\left(1+\log\frac{G_{ji}G_{ki}}{G'_{ji}G'_{ki}}\right)+\sum_{ij}\left(DG'\Lambda\right)_{ij}\frac{G_{ij}^{4}+G_{ij}'^{4}}{G_{ij}'^{3}}$$

is an auxiliary function of $J\left(G\right)$. Furthermore, it is a convex function in G and its global minimum is

$$G_{ik} = G_{ik}\left[\left(RGS+\lambda WG\right)_{ik}/\left(GSG^{T}GS+DG\Lambda\right)_{ik}\right]^{\frac{1}{4}}. \tag{14}$$

Proof: By applying Lemma 4, we have

$$\mathbf{tr}\left(GSG^{T}GSG^{T}\right) \le \sum_{ij}\left(G'SG'^{T}G'S\right)_{ij}\frac{G_{ij}^{4}}{G_{ij}'^{3}}.$$

Because of Lemma 3 and the inequality of $2ab < a^2 + b^2$, we have

$$\Lambda G^{T}DG \le \sum_{ij}\left(DG'\Lambda\right)_{ij}\frac{G_{ij}^{2}}{G'_{ij}} \le \sum_{ij}\left(DG'\Lambda\right)_{ij}\frac{G_{ij}^{4}+G_{ij}'^{4}}{G_{ij}'^{3}}$$

Because $z \le 1 + \log z, \forall\ z > 0$, we have

$$\mathbf{tr}\left(RGSG^{T}\right) \ge \sum_{ijkl} G'_{ji}S_{jk}G'_{kl}R_{li}\left(1+\log\frac{G_{ji}G_{kl}}{G'_{ij}G'_{kl}}\right).$$

$$\mathbf{tr}\left(G^{T}WG\right) \ge \sum_{ijk} G'_{ji}W_{jk}G'_{ki}\left(1+\log\frac{G_{ji}G_{ki}}{G'_{ji}G'_{ki}}\right).$$

Summing over these bounds, we get $Z\left(G,G'\right)$ that clearly satisfies (1) $Z\left(G,G'\right) \ge J\left(G\right)$ and (2) $Z\left(G,G\right) = J\left(G\right)$.

Following the same derivations as in [3], [5], [12], [16], we obtain the Hessian matrix of $Z\left(G,G\right)$, which is positive definite (we skip the derivations due to space limit). Thus $Z\left(G,G\right)$ is a convex function of G. We obtain the global minimum of $Z\left(G,G\right)$ by setting $\partial Z\left(G,G\right)/\partial G_{ij} = 0$ and solving for G, from which we can get Eq. (14). This completes the proof of Theorem 3. ■

Theorem 4: Updating G and S using the rules in Algorithm 1 monotonically decreases $J\left(G\right)$ in Theorem 3.

Proof: According to Lemma 2 and Theorem 3, we can get that $J\left(G^{0}\right) = Z\left(G^{0},G^{0}\right) \ge Z\left(G^{1},G^{0}\right) \ge J\left(G^{1}\right)\ldots$. Thus $J\left(G\right)$ is monotonically decreasing. ■

Because $J\left(G\right)$ in Eq. (22) is obviously lower bounded by 0, Theorem 3–4 guarantee the convergence of Algorithm 1.

Now the remainder is to determine the Lagrangian multiplier Λ. From the KKT condition, summing over i, we obtain $\Lambda_{kk} = \left(G^{T}RGS - G^{T}GSG^{T}GS + \lambda G^{T}WG\right)_{kk}$. This gives the Lagrangian multipliers the value on the diagonal. For non-diagonal elements, we use the Lagrangian without nonnegativity constraints [3]. Thus, we get

$$\Lambda = G^{T}RGS - G^{T}GSG^{T}GS + \lambda G^{T}WG\ . \tag{15}$$

III. Fast Nonnegative Matrix Tri-Factorization

Despite its mathematical elegance, J_6 in Eq. (9) suffers from two problems that impede its practical use. First, instead of being cluster indicator matrix, G is relaxed to be continuous, which makes the immediate outputs of Eq. (9) are not cluster labels. Thus, an additional post-processing step (*e.g.*, using Eq. (3)) is required, which could lead to non-unique solutions [3]. Second, and more important, same as in existing works [3], [5], [10], [12], Algorithm 1 employs the alternately iterative method, in each iteration step of which intensive matrix multiplications are involved. As a result, it is infeasible to apply such algorithms to large-scale real world data due to the expensive computational cost.

In order to tackle these difficulties to work with large-scale data, instead of solving the relaxed clustering problems in Eq. (9), we solve the original clustering problem. Specifically, we constrain the factor matrices of NMTF to be cluster indicator matrices and minimize the following objective:

$$\min\ J_7 = \left\|R - GSG^{T}\right\|^{2} + \lambda\,\mathbf{tr}\left(G^{T}LG\right),\ s.t.\ G \in \Psi\ , \tag{16}$$

where R, G, S are defined as in Eq. (8), and W is defined as in Eq. (10). We call Eq. (16) as the proposed Fast Nonnegative Matrix Tri-Factorization (F-NMTF) approach.

Note that, the orthogonal constraints on G are removed in Eq. (16), because their purposes (unique solution and labeling approximation) are automatically accomplished by the new constraints. Surprisingly, with the new constraint on the factor matrix G to be cluster indicator matrix, though more stringent, as shown theoretically shortly in the rest of this section and empirically later in Section V, the computational speed of our new approach can be significantly improved.

A. An Efficient Algorithm to Solve the F-NMTF Problem

Although the solution to the optimization problem in Eq. (9), where G is continuous, is well studied in literature, minimizing J_7 in Eq. (16) that uses cluster indicator matrices is hard to solve in general, because it is a combinatorial optimization problem. In our earlier publication [6], we have discussed how to use clustering indicator matrix to implement fast NMTF, which, however, only considers rectangle NMTF involving only quadratic terms of the factor matrices in the objective. For NMTF objective for symmetric matrix as in Eq. (16), where the first term involves a fourth-order term of the variable G, it is not trivial to decouple Eq. (16) by taking advantage of the nature of cluster indicator matrices as did in [6]. Thus we first simplify the problem.

Simplification of the first term of J_7. Because R by definition is symmetric, we denote its eigen-decomposition[1] as $R = P_R\Sigma_R P_R^T$, where $\Sigma_R \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal elements as the r leading eigenvalues of R,

[1] In most practical cases, R is also semi-definite positive, which makes the eigen-decomposition feasible.

and P_R is the corresponding eigenvector matrix. Then we have $R = A_R Q_R \left(A_R Q_R\right)^T$, where $A_R = P_R \Sigma_R^{\frac{1}{2}} \in \mathbb{R}^{n\times r}$, and $Q_R \in \mathbb{R}^{r\times r}$ is an arbitrary orthonormal matrix such that $Q_R^T Q_R = I$.

Similarly, given the eigen-decomposition $S = P_S \Sigma_S P_S^T$, we denote $A_S = P_S \Sigma_S^{\frac{1}{2}}$. Then we have $S = A_S A_S^T$. Thus minimizing the first term of J_7 can be written as

$$\begin{aligned} &\min \left\| A_R Q_R \left(A_R Q_R\right)^T - G A_S \left(G A_S\right)^T \right\|^2 , \\ &\quad s.t. \quad G \in \Psi,\ Q_R^T Q_R = I \ . \end{aligned} \tag{17}$$

If matrix GA_S approximates matrix $A_R Q_R$, then matrix $GA_S \left(GA_S\right)^T$ approximates matrix $A_R Q_R \left(A_R Q_R\right)^T$. Thus solving the problem in Eq. (17) can be reasonably transformed to solve the following problem:

$$\min \left\| G A_S - A_R Q_R \right\|^2 \ s.t.\ G \in \Psi,\ Q_R^T Q_R = I \ . \tag{18}$$

Simplification of the second term of J_7. Because $\mathbf{tr}\left(G^T G\right) = n$ is a constant, minimizing the second term of J_7 is equivalent to the following problem:

$$\max \ \mathbf{tr}\left(G^T \widetilde{W} G\right) \quad s.t. \quad G \in \Psi \ , \tag{19}$$

where $\widetilde{W} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Eq. (19) is further equivalent to

$$\min \left\| G G^T - \widetilde{W} \right\|^2 \quad s.t. \quad G \in \Psi \ . \tag{20}$$

Again, because $\widetilde{W}$ is symmetric, we denote its eigen-decomposition as $\widetilde{W} = P_W \Sigma_W P_W^T$. Thus, we have $\widetilde{W} = A_W Q_W \left(A_W Q_W\right)^T$, where $A_W = P_W \Sigma_W^{\frac{1}{2}}$ and Q_W is an arbitrary orthonormal matrix. Following the same idea as above, we can reasonably transform Eq. (20) as:

$$\min \left\| G - A_W Q_W \right\|^2 \ s.t.\ G \in \Psi,\ Q_W^T Q_W = I \ . \tag{21}$$

Combining Eq. (18) and Eq. (21), the original F-NMTF problem in Eq. (16) is transformed to the following problem:

$$\begin{aligned} &\min J_8 \left(G, A_S, Q_R, Q_W\right) = \\ &\qquad \left\| G A_S - A_R Q_R \right\|^2 + \lambda \left\| G - A_W Q_W \right\|^2 \\ &\quad s.t. \quad G \in \Psi,\ Q_R^T Q_R = I,\ Q_W^T Q_W = I \ . \end{aligned} \tag{22}$$

Optimization algorithm to minimize J_8. We first present the following useful theorem.

Theorem 5: [6] Given a general optimization problem:

$$\min \left\| B - AQ \right\|^2 , \quad s.t. \quad Q^T Q = I \ , \tag{23}$$

when A and B are fixed, the optimum Q is given by $Q = UV^T$, where $H = A^T B$ and the Singular Value Decomposition (SVD) of H is given by $H = U \Lambda V^T$.

Now we alternatively optimize the four variables of J_8.

First, when G is fixed, the problem in Eq. (22) is decoupled to the following two problems:

$$\min \left\| G A_S - A_R Q_R \right\|^2 \quad s.t. \quad Q_R^T Q_R = I \ , \tag{24}$$

$$\min \left\| G - A_W Q_W \right\|^2 \quad s.t. \quad Q_W^T Q_W = I \ . \tag{25}$$

When A_S is fixed, applying Theorem 5 to Eq. (24), $Q_R = U_R V_R$ where U_R and V_R are obtained by SVD on $A_R^T G A_S$. When Q_R are fixed, A_S can be obtained by solving the linear equation $GA_S = A_R Q_R$, which has been well studied in literature and can be efficiently solved.

Following the same way above, applying Theorem 5 to Eq. (25), we obtain $Q_W = U_W V_W$ where U_W and V_W are obtained by SVD on $A_W^T G$.

Now that A_S, Q_R, Q_W are solved, we fix them to solve G. Because G is a cluster indicator matrix, let $C^T = A_S Q_S$, $D^T = A_R Q_R$ and $E = A_W Q_W$, Eq. (22) is decoupled to the following simpler problem for each $i \left(1 \le i \le n\right)$:

$$\min \left\| \mathbf{d}_{\cdot i} - C \mathbf{g}_{i\cdot}^T \right\|^2 + \lambda \left\| \mathbf{g}_{i\cdot} - \mathbf{e}_{i\cdot} \right\|^2 \quad s.t. \quad G \in \Psi \ . \tag{26}$$

Suppose $\mathbf{g}_{i\cdot}$ corresponds to the data point of the k-th type. The solution of G is obtained by:

$$G\left(i, j\right) = \begin{cases} 1 & j = j^* \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{27}$$

where

$$j^* = \underset{\left(\sum_{k'=1}^{k-1} c_{k'}\right)+1 \le \hat{k} \le \left(\sum_{k'=1}^{k} c_{k'}\right)}{\operatorname{argmin}} \left(\left\| \mathbf{d}_{\cdot i} - \mathbf{c}_{\cdot \hat{k}} \right\|^2 - 2\lambda E\left(i, \hat{k}\right) \right) \ . \tag{28}$$

The above procedures are summarized in Algorithm 2. Due to the nature of alternating optimization, Algorithm 2 is guaranteed to converge to a local minima. A careful look at Algorithm 2 show that the sizes of the matrices on which we need to perform SVD are small, which is the pre-specified low matrix ranks. Thus step 3 and 5 can be efficiently computed. In addition, step 4 solves a linear equation, and step 6 enumerates vector norms which is definitely faster than matrix multiplication. In summary, as long as the pre-specified ranks of A_R, A_S and A_W are not high, the computation speed of the algorithm will be fast.

Algorithm 2: Algorithm to solve Eq. (22).

```
Input: The inter-type relationship matrix $R$ defined in Eq. (8) and the
       intra-type affinity matrix $W$ defined in Eq. (10).
1. Initialize $G \in \Psi^{n\times c}$ with arbitrary cluster indicator matrix.
2. Compute $A_R$ and $A_W$ from $R$ and $W$ respectively.
repeat
    3. Compute $Q_R = U_R V_R$ where $U_R$ and $V_R$ are obtained by
    SVD on $A_R^T G A_S$.
    4. Compute $A_S$ by solving the linear equation $GA_S = A_R Q_R$.
    5. Compute $Q_W = U_W V_W$ where $U_W$ and $V_W$ are obtained by
    SVD on $A_W^T G$.
    6. Compute $G$ by Eq. (27).
until Converges
Output: Cluster indicator matrix $G$.
```

IV. Related Works

In this section, we review a few related works, and examine their connections to the proposed approaches.

Co-clustering of two-type relational data. Co-clustering has received increased attention in recent years due to the ubiquity of two-type relational real world data. The original NMF method decomposes a nonnegative input matrix into two nonnegative factor matrices, which is shown to be closely related to simultaneous clustering on the rows and columns of the input matrix [10]. However, due to the stringent constraints and limited freedom, NMF can only work with nonnegative input data, typically with unsatisfactory performance. Ding *et al.* [3] investigated tri-factorization to introduce an additional factor matrix S to absorb the different scales among the input data matrix and the two side factor matrices. They [12] also relaxed the nonnegativity constraints on the input matrix, thereby expanded the applicability of NMF methods. Lately, manifold regularization is leveraged to incorporate pairwise affinity among data objects [4], [5], which, however, dropped the orthogonality constraints on factor matrices. As one of our contribution, we impose them into our objectives as in Eq. (9) and emphasize their importance, theoretically and empirically.

Besides, there also exist many co-clustering using mechanisms other than NMF, such as [7], [8].

High-order co-clustering of multi-type relational data. Due to the progress of modern technologies, especially for those related to Internet, multi-type relational data have appeared in many real world applications, which arouses considerable research interests for simultaneous clustering on them. Latent semantic analysis [14] and spectral clustering [1] that deal with homogenous data were extended to handle multi-type relational data, which, however, only employ inter-type relationships. Two recent works used NMF to deal with multi-type relational data. Wang *et al.* [16] incorporated weakly supervised constraints into NMF objective to exploit human labeling information in clustering. Chen *et al.* [17] further developed NMF to deal with star-structured multi-type relational data, which, however, is not able to work with general multi-type relational data like the proposed approaches.

All above mentioned NMF based (high-order) co-clustering methods use the traditional nonnegative constraints on the factor matrices, which, as discussed earlier, lead to solution algorithms with intensive multiplications of large matrices. In contrast, the proposed F-NMTF method using cluster indicator matrix could decouple the original optimization problem into a number of smaller subproblems requiring much less matrix multiplications, which make it more suitable for practical use.

V. Empirical Studies

In this section we experimentally evaluate the proposed O-NMTF and F-NMTF approaches. Because our approaches present a general clustering framework applicable to a variety of data types with different conditions, we evaluate it in (simultaneous) clustering tasks for single-type, two-type, and multi-type relational data with different experimental settings.

Table I
Data sets used in Section V-A and Section V-B.

	AT&T	Yale	Newsgroup4	WebKB4	WebKB4	RCV
# data	400	165	3970	8280	2340	193844
# classes	40	15	4	7	20	1979

The two proposed approaches have a tradeoff parameter λ in Eq. (9) and Eq. (22). We will investigate it with some details in Section V-A. In F-NMTF approach, we have an additional parameter, *i.e.*, the pre-specified ranks for low-rank matrix approximations of SVDs in steps 3 and 5 of Algorithm 2. We set it as $\min(c, \operatorname{rank}(M))$, where $c = \sum_{k=1}^{K} c_k$ as defined before and M is the input matrix of SVD in steps 3 and 5 of Algorithm 2.

In order to exploit as much information as possible and make performance comparison more reasonable, we also incorporate weakly supervised labeling information, including both must-links and cannot-links for a give type of data, into the pairwise affinity matrix W in Eq. (16) following [20].

To evaluate the clustering performance, we adopt two standard measures widely used in literature [4]: clustering accuracy and normalized mutual information (NMI).

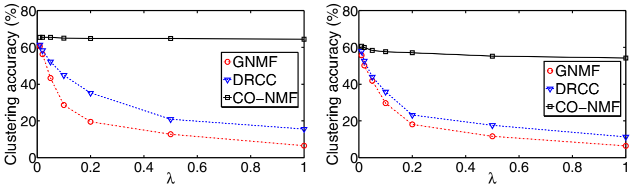
A. Study on Regularization Parameter

In Section II-B, we theoretically point out the importance of the orthogonal constraints on the factor matrices when NMF objective involves manifold regularization, such as our objective in Eq. (9). Therefore we first evaluate its impact on the regularization parameter λ in Eq. (9).

We use AT&T and Yale face databases in our experiments. The data are summarized in Table I. Following standard experimental conventions using face data, we resize all the face images to 32×32.

Experimental settings. In unsupervised clustering on homogeneous data, we need two inputs from a data set: data features and pairwise affinities. The former is obtained using $R = \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix}$ from the feature matrix of the testing data sets. The latter are computed by constructing neighborhood graphs following [5], where the neighborhood size is set 10, same as [5]. The cluster numbers of the data sets are set as the ground truth.

We compare our O-NMTF approach against two related clustering methods that combine NMF and Laplacian regularization: (1) graph regularized NMF (GNMF) [4] method and (2) dual regularized co-clustering (DRCC) [5] method. These two methods are largely same, except that the former uses two-factor factorization and imposes Laplacian regularization on one side factor, while the latter uses the three-factor factorization and imposes Laplacian regularizations on the both side factors. Following [5], two graphs are built



(a) AT&T data set.

(b) Yale data set.

Figure 2. Clustering accuracy of compared methods on homogenous data with respect to the regularization parameter λ.

on both feature side and data point side for DRCC method, and its parameters are set as $\lambda = \mu$ which is same as in [5]. Note that, when using Laplacian regularization, neither of these two methods considers the orthogonalities on the factor matrices, which, however, as analyzed in Section II-B, plays a significant role in NMF based optimization objective involving manifold regularization.

Experimental results. For each clustering method with each different value of λ, we repeat the experiment for 100 times with randomly initialized G. The average clustering accuracy over the 100 trials are reported. Figure 2 shows the results of the three compared methods on AT&T and Yale data set. It can be seen that, all the three methods performs reasonably well at small $\lambda = 0.01$, and our O-NMTF approach is slightly better. As λ increases gradually to $\lambda = 1$, the clustering performances of both GNMF and DRCC methods drop very quickly, while the performance of our O-NMTF approach still remains stably. This is because the Laplacian term becomes dominant when λ is big. For the two non-orthogonal methods, in the solution of G^*, different columns become very similar — they are close to the degenerate solution of the Laplacian: $G^* \approx [\mathbf{e}, \ldots, \mathbf{e}]$, because $\mathbf{e}$ is the eigenvector corresponding to smallest eigenvalue of the Laplacian. Thus the cluster assignment at this case is rather random, resulting low accuracy.

A general picture of the importance of the orthogonal constraints is as the following. When λ is small ($\lambda \simeq 0.01$), the solution and performance of the orthogonal NMF formulation like our approach is similar to its non-orthogonal counterpart as GNMF and DRCC. When λ is medium or large ($\lambda \simeq 1$), the solution to orthogonal formulation gives better and more stable accuracy; in this case, different columns of the optimal G^* of its non-orthogonal counterpart are very similar as explained above.

Upon the above results, we set $\lambda = 0.01$ in the sequel.

B. Co-Clustering of Two-Type Related Data

Because two-type relational data is the most fundamental multi-type relational data, we evaluate the proposed methods in co-clustering tasks. Following [5], we use the four benchmark data sets as summarized in Table I. The RCV1 data set has very big sample size and feature size [21]. In order to run the experiments on contemporary computers, following previous studies [21], we remove the keywords (features) appearing less than 100 times in the corpus, which results in 1979 (out of 47236) keywords in our experiments.

Experimental settings. In semi-supervised clustering of two-type relational data, we need three inputs from a data set: the relationship matrix between the two types of data, the pairwise affinity matrices for each type of data, and the weakly supervised constraints. We obtain the relationship matrices directly from the testing data sets. We construct neighborhood graphs from the both sides of the relationship matrix following [5] to obtain the pairwise affinity matrices, where the neighborhood size is set as 10. Instead of constraining the data from both data point side and feature side, we only pick up constraints from the data point side, same as in most real world applications. Following [16], we generate the weakly supervised constraints as follows: for each constraint, we randomly pick up one pair of data points from the input data sets (the labels of which are available for evaluation purpose but unavailable for clustering). If the labels of this pair of points are the same, we generate a must-link. If the labels are different, a cannot-link is generated. We pick up 50 constraints for each data set. In all the experiments, the results are averaged over 100 trials to eliminate the difference caused by constraints and the perturbation caused by initialization.

We compare the proposed approaches against two related clustering methods: (1) graph regularized NMF (GNMF) [4] method and (2) dual regularized co-clustering (DRCC) [5] method. Note that, these two methods only incorporate unsupervised pairwise affinity into NMF, but do not involve supervision information. Therefore, we also compare our approach against the following two methods: (3) penalized matrix factorization (PMF) [16] method and (4) constrained K-means (CKmeans) [22] method. Both of them conduct clustering upon inter-type relationships and weakly supervised constraints.

Experimental results. The clustering performance comparisons on the four experimental data sets are reported in Table II. The two proposed approaches consistently outperforms the other methods, sometimes very significantly, which confirms their effectiveness in co-clustering of two-type relational data. In addition, the proposed O-NMTF approach is generally better than F-NMTF approach, which is consistent with their mathematical formulations in that the factor matrices in the former are continuous and have better data representation power.

We also report the run time of the compared methods. All our experiments are performed on a Dell PowerEdge 2900 server, which has two quad-core Intel Xeon 5300 sequence CPU processors at 3.0 GHz and 48G bytes memory. From the results in Table III we can see that the F-NMTF method is only slower than CKMeans method, which, however, has much worse clustering performance. These results demonstrate that F-NMTF method is not only better in terms of

Table II
CLUSTERING ACCURACY AND NMI COMPARISONS ON THE FOUR TWO-TYPE RELATIONAL DATA SETS.

	Methods	Newsgroup4	WebKB4	WebACE	RCV1
Accuracy	GNMF	0.889	0.731	0.513	0.202
	DRCC	0.931	0.738	0.568	0.210
	PMF	0.923	0.727	0.566	0.214
	CKMeans	0.643	0.594	0.477	0.185
	O-NMTF	**0.949**	**0.781**	**0.615**	**0.251**
	F-NMTF	0.933	0.751	0.591	0.247
NMI	GNMF	0.716	0.462	0.618	0.313
	DRCC	0.782	0.491	0.629	0.316
	PMF	0.758	0.488	0.602	0.309
	CKMeans	0.612	0.375	0.491	0.268
	O-NMTF	**0.801**	**0.557**	**0.652**	**0.339**
	F-NMTF	0.785	0.521	0.630	0.327

Table III
RUN TIME (IN SECONDS) OF THE COMPARED METHODS.

Methods	Newsgroup4	WebKB4	WebACE	RCV1
GNMF	1.07×10^4	4.41×10^4	5.36×10^3	1.48×10^6
DRCC	1.81×10^4	6.33×10^4	8.61×10^3	2.18×10^6
PMF	2.15×10^4	7.53×10^4	1.13×10^4	2.96×10^6
CKMeans	6.57×10^3	2.06×10^4	2.15×10^3	8.51×10^5
O-NMTF	1.51×10^4	4.52×10^4	6.01×10^4	1.68×10^6
F-NMTF	$\mathbf{7.64 \times 10^3}$	$\mathbf{2.76 \times 10^4}$	$\mathbf{3.78 \times 10^3}$	$\mathbf{1.04 \times 10^6}$

clustering performance measured by accuracy and NMI, but also faster than state-of-the-art (co-)clustering methods.

In summary, the proposed F-NMTF approach has competitive clustering performance to the proposed O-NMTF approach, but with much faster computational speed. Both of the two proposed approaches have satisfactory clustering performance on all the four benchmark data sets.

C. High-Order Co-Clustering of Multi-Type Related Data

Finally, we evaluate the proposed approaches for simultaneous clustering of multi-related data by using both inter-type and intra-type information, which is the ultimate goal of this work.

Data set. We use a data set sampled from the Bulletin Board Systems (BBS) in [23]. In a BBS system, the users first register IDs. Using their IDs, the users can read messages published by other users and leave their own messages. The whole system consists of many discussion fields, each of which contains many boards with similar themes. The boards are named to reflect the contents of the articles in them [23]. Once an ID posts a new article (initial article) on one board, the others can show their opinions by replying the initial article using reply articles. The initial article and reply articles constitute a topic. Each board contains many topics. Each topic connects several IDs through articles.

We use two subsets of the BBS data in [23]. In each data set, several boards are sampled from several discussion fields. In each board, 80 topics are sampled randomly. The names of the fields and boards that we use are listed in Table IV. The user IDs related to these topics and boards are found out. Then the tensor is constructed by the co-occurrence of these three data types.

Table IV
TWO SUBSETS OF DATA SAMPLED FROM A BBS DATA SET [23].

Data set 1		Data set 2	
Field name	Board name	Field name	Board name
Comp. Sci.	C++ Builder	Comp. Sci.	Virus
Comp. Sci.	Delphi	Comp. Sci.	Unix
Comp. Sci.	Database	Entertainment	Music
Sports	Basketball	Entertainment	Dance
Sports	Volleyball	Society	Law
Sports	Badminton	Society	Commerce

Experimental settings. In the experiments, there exist three data types: topics ($\mathcal{X}_1$), user IDs ($\mathcal{X}_2$) and boards ($\mathcal{X}_3$). The topic-user matrix (R_{12}) is constructed with the number of articles each user posted in each topics with TF-IDF normalization. The topic-board matrix (R_{13}) is constructed such that if a topic belongs to a board, then the corresponding entry of R_{13} is 1. R_{23} is constructed such that if the user had posted any articles on that board, then the corresponding element of R_{23} is set to 1. Finally the elements of R_{23} are also normalized using TF-IDF scheme.

We only use the pairwise affinity matrices W_1 and W_2 for $\mathcal{X}_1$ and $\mathcal{X}_2$, which are constructed using R_{12} in a same way as in Section V-B. We set $W_3 = I$ to emulate the case in real applications when the unsupervised intra-type information of a given type, *i.e.*, pairwise affinities between data objects of $\mathcal{X}_3$ in the current case, is not available.

Following the settings in [16], we randomly generate 500 constraints on $\mathcal{X}_2$ based upon their registered profiles, 100 constraints on $\mathcal{X}_1$ based upon the boards they belong to, and 10 constraints on $\mathcal{X}_3$ based upon their corresponding fields.

Besides our approach, the results the Spectral relational Clustering (SRC) [1] method, Multiple Latent Semantic Analysis (MLSA) [14] method and PMF method are also included for comparison. All these three methods were devised for simultaneous clustering of multi-related data. However, none of them uses the intra-type information as we do. The evaluation metric used in the current experiments is the F1 score computed using the clustering results on topics, the ground truth of which is set to be the classes corresponding to the field names they belong to.

Experimental results. We repeat the experiments for both PMF method and our approaches for 100 times with randomly initialized G, and average F1 scores are reported. The experimental results are shown in Table V, in which the value of d represent different number of clusters. From the table we can see the clear advantages of the proposed approaches, which again demonstrate the usefulness of the proposed method in the tasks of simultaneous clustering of multi-type relational data.

We also report the run time of the compared methods on the same machine as that used in Section V-B. The results in Table VI show that our F-NMTF method again is much faster the compared methods, which demonstrate its computational efficiency and adds to its practical values.

Table V
THE F1 MEASURE OF THE FOUR COMPARED ALGORITHMS ON THE TESTING DATA SET 1 (TOP) AND TESTING DATA SET 2 (BOTTOM).

	MLSA	SRC	PMF	O-NMTF	F-NMTF
$d = 3$	0.712	0.731	0.795	**0.847**	0.821
$d = 5$	0.756	0.634	0.815	**0.844**	0.830
$d = 7$	0.711	0.621	0.780	**0.819**	0.811
$d = 9$	0.699	0.482	0.734	**0.767**	0.747
$d = 3$	0.761	0.763	0.795	**0.828**	0.819
$d = 5$	0.761	0.730	0.796	**0.821**	0.813
$d = 7$	0.729	0.701	0.794	**0.824**	0.820
$d = 9$	0.683	0.660	0.792	**0.819**	0.809

Table VI
THE RUN TIME (IN SECONDS) OF THE FOUR COMPARED ALGORITHMS ON THE TWO TESTING DATA SETS, WHERE THE CLUSTER NUMBER IS $d = 9$.

	MLSA	SRC	PMF	O-NMTF	F-NMTF
Data set 1	174.3	161.4	232.1	415.2	**115.1**
Data set 2	184.5	169.2	241.6	469.0	**132.3**

VI. Conclusions

In this paper, we presented a general O-NMTF framework for high-order co-clustering of multi-type relational data. Our approach simultaneously clusters different types of data using the inter-type relationships by transforming the original NMTF problem into a symmetric NMTF problem, into which we can also optionally incorporate the intra-type information. Instead of constraining the factor matrices of NMTF to be nonnegative as in existing methods, we further proposed a novel F-NMTF approach to constrain them to be cluster indicator matrices. As a result, the optimization problem of the proposed method can be decoupled into a number of subproblems of smaller sizes requiring much less matrix multiplications, which makes our new algorithm of particular use for large-scale real world data. Extensive empirical studies evaluated various aspects of our approach, and demonstrated the usefulness of the proposed approaches.

Acknowledgment. This research was supported by NSF-IIS 1117965, NSFCCF-0830780, NSF-DMS-0915228, NSF-CCF-0917274.

References

[1] B. Long, Z. Zhang, X. Wu, and Y. P., "Spectral clustering for multi-type relational data," in *ICML*, 2006.

[2] H. Wang, H. Huang, and D. H., "Simultaneous Clustering of Multi-Type Relational Data via Symmetric Nonnegative Matrix Tri-factorization," in *CIKM*, 2011.

[3] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *SIGKDD*, 2006.

[4] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE TPAMI*, 2010.

[5] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *SIGKDD*, 2009.

[6] H. Wang, F. Nie, H. Huang, and F. Makedon, "Fast Nonnegative Matrix Tri-Factorization for Large-Scale Data Co-Clustering," in *IJCAI*, 2011.

[7] I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *SIGKDD*, 2001.

[8] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," in *SIGKDD*, 2004.

[9] H. Ma, W. Zhao, Q. Tan, and Z. Shi, "Orthogonal Nonnegative Matrix Tri-factorization for Semi-supervised Document Co-clustering," *Advances in Knowledge Discovery and Data Mining*, vol. 6119, pp. 189–200, 2010.

[10] C. Ding, X. He, and H. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *SDM*, 2005.

[11] D. Lee and H. Seung, "Algorithms for Non-negative Matrix Factorization," in *NIPS*, 2001.

[12] C. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE TPAMI*, 2010.

[13] H. Wang, H. Huang, F. Nie, and C. Ding, "Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization," in *SIGIR*, 2011.

[14] X. Wang, J. Sun, Z. Chen, and C. Zhai, "Latent semantic analysis for multiple-type interrelated data objects," in *SIGIR*, 2006.

[15] B. Long, Z. Zhang, and P. Yu, "A probabilistic framework for relational clustering," in *SIGKDD*, 2007.

[16] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *SDM*, 2008.

[17] Y. Chen, L. Wang, and M. Dong, "Non-negative matrix factorization for semi-supervised heterogeneous data co-clustering," *IEEE TKDE*, vol. 22, no. 10, pp. 1459–1474, 2010.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[19] T. Li, C. Ding, and M. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *ICDM*, 2007.

[20] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu, "Semi-supervised sparse metric learning using alternating linearization optimization," in *SIGKDD*, 2010.

[21] W. Chen, Y. Song, H. Bai, C. Lin, and E. Chang, "Parallel spectral clustering in distributed systems," *IEEE TPAMI*, 2010.

[22] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML*, 2001.

[23] Z. Kou and C. Zhang, "Reply networks on a bulletin board system," *Physical Review E*, vol. 67, no. 3, p. 36117, 2003.