# Orthogonal Nonnegative Matrix Factorization for Multi-type Relational Clustering

Ying Liu
Division of Computer Science, Mathematics and Sciences
College of Professional Studies, St. John's University
Queens, NY 11439

Chengcheng Shen
Amazon.com LLC
Seattle, WA 98109

*Abstract*—**Relational clustering with heterogeneous data objects has impact in various important applications, such as web mining, text mining and bioinformatics etc. In this paper, we build a star-structured general model for relational clustering. It is formulated as an orthogonal tri-nonnegative matrix factorization. The model performs matrix approximation among all different data types to look for hidden cluster structure. Under this model, we propose a multiplicative update algorithm to minimize the matrix approximation error for simultaneously clustering of heterogeneous relational objects. The proposed algorithm tries to retain the orthogonality of indicator matrices, which make it easier for result interpretation. We also prove the correctness and convergence of the algorithm under the proposed iterative update rules. Experiments demonstrate the effectiveness of the proposed algorithm and the ability to co-cluster different data objects.**

***Keywords-relational clustering; co-clustering; nonnegative matrix factorization; clustering***

## I. Introduction

Clustering divides data objects into groups or clusters of similar objects. It achieves simplification by representing complex data objects by a few clusters such that data objects within the same cluster are similar while data objects in different clusters are dissimilar. It is also called unsupervised learning in machine learning. Existing algorithms include k-means [24], maximum likelihood estimation [14] and spectral clustering [2, 33]. Most of the conventional algorithms require the data objects to be homogeneous. For example, graph partitioning can be viewed as single-type relational data clustering on a graph affinity matrix, which has homogeneous relations. Relational data consist of objects (representing people, places, and things) connected by links (representing persistent relationships among objects). However, inter-related heterogeneous data objects have practical importance in a wide variety of applications such as text mining [9, 19], web-log mining [39], market-basket data analysis [9, 19], and biological microarray data analysis [21]. In such scenarios, using previous methods to cluster each type of objects independently may not work well since the similarities among one type of objects sometimes depend on the other type of objects, thus traditional clustering methods cannot pass the hidden relational information along the relation chain by considering only one type of object at a time.

Algorithms for clustering on bi-type relational data (bi-clustering) has been proposed by Dhillon et al.[9]. In this algorithm, authors model inter-relationship as a bipartite graph and seek to find the minimal normalized cut in the graph with spectral relaxation. Some information-theory based algorithms have also been proposed. [15] uses an agglomerative hard clustering version of the Information bottleneck method [35] to cluster documents and then words. Dhillon et al. [10] propose an information theoretic co-clustering algorithm to monotonically increase the preserved mutual information by intertwining both the row and column clustering at all stages. Later, a more generalized co-clustering framework based on Bregman divergence is presented by Banerjee et al.[4]. Except those, approximation algorithms [34, 1] are also proposed for co-clustering problems. [31] presents a hierarchical Bayesian model for simultaneously clustering documents and terms, where each document is modeled as a random mixture of document topics and each topic is a distribution over some segments of the text. Another soft co-clustering algorithm [32] is also proposed which is able to work with any regular exponential family distribution and corresponding Bregman divergences.

In many real applications, relationships among multiple data objects usually involve more than two types of data objects, such as transcription factor-gene-tissue specification [16], query-webpage-user [37] and category-document- term [26], etc. Some research efforts have been dedicated to generalize bi-clustering to more than two types of data objects. In [37], authors propose an approach called Recom (Reinforcement Clustering of Multi-type Interrelated data objects) to iteratively improve the cluster quality of interrelated data objects through a reinforcement clustering process. Gao et al. [17] formulate a semi-definite programming algorithm to partition a k-partite graph. A spectral relational clustering (SRC) [26] algorithm iteratively embeds each type of data objects into low dimensional spaces and benefits from the interactions among the hidden structures of different types of data objects. Long et al. [25] propose a family of algorithms to identify the hidden structures by approximation of a k-partite graph under a broad range of distortion measures. A probabilistic framework [28] is proposed for relational clustering to unify various clustering tasks including attributes-based clustering and co-clustering. [8] extends and generalizes information-theoretic framework for high-order relational clustering. In [3], authors propose a multi-way relational clustering by accurately approximating the set of tensors corresponding to the various relations. [5] uses tensor to model the multi-type relationships existing in heterogeneous datasets and then derive a clique expansion algorithm to find the solution for normalized hypergraph cut.

Nonnegative matrix factorization (NMF) can be traced back to 1970s and Paatero's work [29]. Lee and Seung[22] brought much attention to NMF in data mining and machine learning. NMF has been shown effective in many applications such as pattern recognition, text mining etc. The hidden structure of a data matrix can be explored by factorization [27, 22]. [27] and [23] introduce matrix factorization to co-clustering optimization problem. [27] proposes an EM-like algorithm based on multiplicative rules. [23] proposes a hard clustering algorithm for binary data. NMF bi-clustering algorithms are also proposed by [30] and [6] in unsupervised and semi-supervised settings respectively. [30] presents a general framework for co-clustering large datasets utilizing sampling based matrix decomposition methods. A theoretic study shows the equivalence between NMF and kernel K-means/spectral clustering [11].

Overall, the research on relational data clustering has attracted substantial attention, but there are still limited efforts on general relational clustering with NMF. In this paper, we attempt to extend orthogonal NMF and derive an algorithm for general multi-type relational clustering based on inter-relational matrices.

## II. PROBLEM FORMULATION

An undirected graph consists of a data set of homogeneous *s* points and a set of edges measuring the similarities between points. $W$ is an adjacency matrix for the graph with $w_{ij} \geq 0$ denoting edge weight (similarity) between points *i* and *j*. It has already been shown [19, 2] that for all partitions *E* into *R* clusters, the *R*-way normalized cut is equal to minimize $R - tr\left(\hat{P}^T\left(D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\right)\hat{P}\right)$, where $tr$ denotes matrix trace operator, $D$ is a diagonal matrix with $D_{ii} = \sum_j W_{ij}, 1 \leq i \leq s$, $P \in Rs \times R$ is any matrix, with the following restrictions: (1) the columns of $D^{-1/2}\hat{P}$ are piecewise constant with respect to the clusters $E$. (2) $\hat{P}$ has orthonormal column ($\hat{P}^T\hat{P} = I$). $\hat{P}$ is actually a scaled indicator matrix for all *s* points. A spectral relaxation can be used to obtain the clustering by computing the eigenstructure of $D^{-1/2}WD^{-1/2}$. In a bipartite graph, there are two kinds of heterogeneous data points *X* and *Y*. We can define a new similarity matrix as follows:

$$W = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{2.1}$$

where $A$ is interrelational matrix and $a_{ij} \geq 0$ denotes edge weight (similarity) between points $x_i$ and $y_j$. We see that the bipartite graph is actually a special case of a regular graph. By applying the above result for *R*-way normalized cut, we are able to co-cluster heterogeneous data points X and Y simultaneously.

In this paper, we consider multiple heterogeneous data objects, whose relationships form a star structure. In particular, given $N + 1$ sets of data objects, $X = \{x_1, x_2, \ldots, x_m\}$, $Y_i = \{y_i^1, y_i^2, \ldots, y_i^{n_i}\}$, where $m = |X|$, $n_i = |Y_i|$, $1 \leq i \leq N$ and $|.|$ represents the number of members in a set. There exists relation between each pair of $X$ and $Y_i$ denoted by $R_i \in R^{m \times n}$, where an element $R_i^{pq}$ denotes the relation between $x^p$ and $y_i^q$. Figure 1 shows an example of start-structured multi-type relational data. The model can be considered as a bipartite, tri-partite or *k*-partite graph if we take the central data type $X$ and different number of data objects $Y_i$. The data in Figure 1 can be denoted by four relational matrices $R_1$, $R_2$, $R_3$ and $R_4$. We are interested in simultaneously clustering $X$ into $k$ and $Y_i$ into $k_i$ disjoint clusters. We call it general multi-type relational data clustering.

The objective of spectral clustering is to minimize the normalized cut and maximize $tr\left(\hat{P}^T\left(D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\right)\hat{P}\right)$. Ding et al. [11] showed that that spectral clustering is equivalent to NMF with orthogonality restriction based on the normalized adjacency matrix.

THEOREM 1. *Minimization of normalized cut is equivalent to NMF with the orthogonality restriction.*

*Proof.* Normalized cut minimization is to maximize:

$$\max_{\hat{P}^T\hat{P}=I,\hat{P}\geq 0} tr(\hat{P}^T(D^{-1/2}WD^{-1/2})\hat{P})$$

Let $\widehat{W} = D^{-1/2}WD^{-1/2}$, it can be written as:

$$\begin{aligned} &\underset{\hat{P}^T\hat{P}=I,\hat{P}\geq 0}{\operatorname{argmin}} -2tr(\hat{P}^T\widetilde{W}\hat{P}) \\ &= \operatorname{argmin}_{\hat{P}^T\hat{P}=I,\hat{P}\geq 0}\left\|\widetilde{W}\right\|^2 - 2tr(\hat{P}^T\widetilde{W}\hat{P}) + \left\|\hat{P}^T\hat{P}\right\|^2 \\ &= \operatorname{argmin}_{\hat{P}^T\hat{P}=I,\hat{P}\geq 0}\left\|\widetilde{W} - \hat{P}\hat{P}^T\right\|^2 \end{aligned}$$

This completes the proof.

Further, in the case of bipartite graph, we can extend theorem 1 to have the following theorem for inter-relational matrix A.

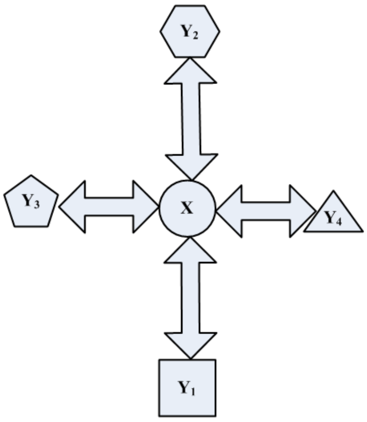


Figure 1: Star-structured multi-type relational data

THEOREM 2. *In the bipartite graph modeled in Eq. (2.1), let* $\hat{X}$ *and* $\hat{Y}$ *be scaled cluster indicator matrices for objects in X and Y.* $D_X$ *and* $D_Y$ *are diagonal matrices, the diagonal elements are the row sums of A and* $A^T$ *respectively. Then Theorem 1 can be transformed to the following orthogonal NMF:*

$$\min_{\hat{X}\geq 0,\hat{Y}\geq 0}\left\|D_X^{-1/2}AD_Y^{-1/2}-\hat{X}\hat{Y}^T\right\|^2 \qquad \hat{X}^T\hat{X}=I,\hat{Y}^T\hat{Y}=I$$

*Proof.*

$$\widetilde{W} = D_X^{-1/2}WD_Y^{-1/2}$$

$$= \begin{bmatrix} 0 & D_X^{-1/2}AD_Y^{-1/2} \\ D_Y^{-1/2}AD_X^{-1/2} & 0 \end{bmatrix}$$

We have $\hat{P}^T = \left[\hat{X}^T\hat{Y}^T\right]$, then

$$min\left\|\widetilde{W}-\hat{P}\hat{P}^T\right\|^2$$

$$= min\left\|\begin{bmatrix} 0 & D_X^{-1/2}AD_Y^{-1/2} \\ D_Y^{-1/2}AD_X^{-1/2} & 0 \end{bmatrix} - \begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}\left[\hat{X}^T\hat{Y}^T\right]\right\|$$

It is obvious to see that the equation above is equivalent to:

$$\min_{\hat{X}\geq 0,\hat{Y}\geq 0}\left\|D_X^{-1/2}AD_Y^{-1/2}-\hat{X}\hat{Y}^T\right\|^2 \qquad \hat{X}^T\hat{X}=I,\hat{Y}^T\hat{Y}=I$$

Theorem 2 detaches the heterogeneous objects *X* and *Y* and allows us to treat different objects separately.

Consider general multi-type relational data clustering. We choose to formulate it as a sum of *N X – Y*ᵢ clustering subproblems, and *X* should have same partition in all subproblems. If we put in inter-relational matrix $R_i$, based on theorem 2, we can write general multi-type relational data clustering as:

$$\min_{\hat{X}\geq 0,\hat{Y}\geq 0}\sum_{i=1}^{N} w_i\left\|D_{X_i}^{-\frac{1}{2}}R_iD_{Y_i}^{-\frac{1}{2}}-\hat{X}\hat{Y}_i^T\right\|^2 \qquad (2.2)$$

$$\hat{X}^T\hat{X}=I,\hat{Y}^T\hat{X}=I$$

where $w_i \geq 0$, is a weighting parameter for each $X - Y_i$ graph, $\hat{X}$ and $\hat{Y}_i$ are scaled indicator matrices for objects in *X* and $Y_i$ , $D_{X_i}$and $D_{Y_i}$ are diagonal matrices and diagonal elements are row sum of $R_i$ and $R_i^T$ respectively.

Eq. (2.2) is NMF with double orthogonality, which is very restrictive and gives poor matrix approximation. We introduce an extra factor $S_i$ to absorb the different scales and also allow some degree of freedom such that we can have different number of clusters in *X* and $Y_i$.

Let $Q_i = D_{X_i}^{-1/2}R_iD_{Y_i}^{-1/2}$, Eq. (2.2) can be rewritten as a bi-orthogonal tri-factorization NMF. We want to minimize the following objective function:

$$J = \sum_{i=1}^{N} w_i\left\|Q_i-\hat{X}S_i\hat{Y}_i^T\right\|^2 \qquad (2.2)$$

$$\hat{X}^T\hat{X}=I,\hat{Y}_i^T\hat{Y}_i=I,\hat{X}\geq 0,\hat{Y}_i\geq 0,\hat{S}_i\geq 0,w_i\geq 0$$

## III. PROOF OF ALGORITHMIC CONVERGENCE AND CORRECTNESS

In this section, motivated by [13], we propose an algorithm to co-cluster multiple relational objects and prove the correctness and convergence of the algorithm. The algorithm, called Orthogonal NMF Relational Clustering (ONRC), is summarized in Table 1.

TABLE I ORTHOGONAL NMF RELATIONAL CLUSTERING

**Orthogonal NMF Relational Clustering**

```
1.  Input: Number of clusters $k$, $k_i$ for central object $X$ and $Y_i$
    Relational matrices $R_i$ and graph weight $w_i$ for each $X$-$Y_i$ graph
2.  From matrices $Q_i = D_{X_i}^{-1/2}R_iD_{Y_i}^{-1/2}$
3.  Initialize $\hat{X}, \hat{Y}_i$ by K-means, then set
    $\hat{X} \leftarrow \hat{X} + 0.2$
    $\hat{Y}_i \leftarrow \hat{Y}_i + 0.2$
    $S_i = \hat{X}^TQ_i\hat{Y}_i$
4.  repeat
      update $\hat{X}$
      $\hat{X}_{jk} \leftarrow \hat{X}_{jk}\frac{(\sum_{i=1}^{N} w_i(Q_i\hat{Y}_iS_i^T))_{jk}}{(\sum_{i=1}^{N} w_i(\hat{X}\hat{X}^TQ_i\hat{Y}_iS_i^T))_{jk}}$        (3.4)
      for i=1 to N do
        update $\hat{Y}_i, S_i$
        $S_{ijk} \leftarrow S_{ijk}\frac{(\hat{X}^TQ_i\hat{Y}_i)_{jk}}{(\hat{X}^T\hat{X}S_i\hat{Y}_i^T\hat{Y}_i)_{jk}}$        (3.5)
        $\hat{Y}_{ijk} \leftarrow \hat{Y}_{ijk}\frac{(Q_i^T\hat{X}S_i)_{jk}}{(\hat{Y}_i\hat{Y}_i^TQ_i^T\hat{X}S_i)_{jk}}$        (3.6)
      end for
    until convergence
5.  Cluster membership analysis
    for each object $x_i$ in $X$
      $x_i \leftarrow \text{argmax}_j \hat{X}_{ij}$
    for each object $y_i^j$ in $Y_i$
      $y_i^j \leftarrow \text{argmax}_k \hat{Y}_{ijk}$
```

### A. *Correctness*

The update rules in Eqs. (3.4) - (3.6) ensures the convergence of the algorithm and the solutions satisfy the KKT complementarily condition.

We introduce Lagrangian multipliers $\mu_0$, $\mu_i$, $\mu_{i+N}$, $\lambda_0$, $\lambda_i$ and minimize the following Lagrangian function from Eq.(2.3):

$$L(\hat{X}, \hat{Y}_i, S_i, \lambda_0, \lambda_i, \mu_0, \mu_i, \dots, \mu_{i+N})$$

$$= \sum_{i=1}^{N} w_i \left\| Q_i - \hat{X} S_i \hat{Y}_i^T \right\|^2 \qquad (3.7)$$

$$-tr\left(\mu_0 \hat{X}^T\right) - tr \sum_{i=1}^{N} (\mu_i S_i^T) - tr \sum_{i=1}^{N} (\mu_{i+N} \hat{Y}_i^T)$$

$$+tr\left(\lambda_0\left(\hat{X}^T \hat{X}^T - I\right)\right) + tr \sum_{i=1}^{N} (\lambda_i\left(\hat{Y}_i^T \hat{Y}_i^T - I\right))$$

Based on KKT complementarily conditions $\frac{\partial L}{\partial \hat{X}} = 0$, $\frac{\partial L}{\partial S_i} = 0$, and $\frac{\partial L}{\partial \hat{Y}_i} = 0$, we can get the following equations,

$$\sum_{i=1}^{N} w_i\left(-2Q_i Y_i S_i^T + 2\hat{X} S_i Y_i^T Y_i S_i^T\right) + 2\hat{X}\lambda_0 - \mu_0 = 0 \qquad (3.8)$$

$$w_i\left(-2\hat{X}^T Q_i \hat{Y}_i + 2\hat{X}^T \hat{X} S_i \hat{Y}_i^T \hat{Y}_i\right) - \mu_i = 0$$

$$w_i\left(-2Q_i^T \hat{X} S_i + 2\hat{Y}_i\left(\hat{X} S_i\right)^T\left(\hat{X} S_i\right)\right) + 2\hat{Y}_i \lambda_i - \mu_{i+N} = 0$$

KKT conditions give

$$\mu_0 \circ \hat{X} = 0 \qquad \mu_i \circ S_i = 0 \qquad \mu_{i+N} \circ \hat{Y}_i = 0 \qquad (3.9)$$

where ° is the Hadamard product operator. Substitute Eq. (3.8) into Eq. (3.9), we have the following equations:

$$\left(\sum_{i=1}^{N} w_i\left(-2Q_i Y_i S_i^T + 2\hat{X} S_i Y_i^T Y_i S_i^T\right) + 2\hat{X}\lambda_0\right) \circ \hat{X} = 0 \quad (3.10)$$

$$w_i\left(-2\hat{X}^T Q_i \hat{Y}_i + 2\hat{X}^T \hat{X} S_i \hat{Y}_i^T \hat{Y}_i\right) \circ S_i = 0$$

$$\left(w_i\left(-2Q_i^T \hat{X} S_i + 2\hat{Y}_i\left(\hat{X} S_i\right)^T\left(\hat{X} S_i\right)\right) + 2\hat{Y}_i \lambda_i\right) \circ \hat{Y}_i = 0$$

$\lambda_0$ and $\lambda_i$ can also be computed [13],

$$\lambda_0 = \sum_{i=1}^{N} w_i\left(\hat{X}^T Q_i \hat{Y}_i S_i^T - \hat{Y}_i^T \hat{Y}_i S_i^T\right) \qquad (3.11)$$

$$\lambda_i = w_i\left(\hat{Y}_i^T Q_i^T \hat{X} S_i - \left(\hat{X} S_i\right)^T\left(\hat{X} S_i\right)\right)$$

Based on Eqs (3.10, 3.11), we can derive the updating rules of Eqs. (3.4 – 3.6). We can prove that Eq (3.10) are satisfied if *X*, *S<sub></sub>*$S_i$ and $\hat{Y}_i$ are locally minimized.

### B. *Covergence*

In this section, we prove hat $\sum_{i=1}^{N} w_i \left\| Q_i - \hat{X} S_i \hat{Y}_i^T \right\|^2$ is decreasing monotonically under the update rules Eqs. (3.4 – 3.6).

A function $G(\hat{X}^{(t+1)}, \hat{X}^t)$ is called an auxiliary function of $L(\hat{X}^{(t+1)})$ if it satisfies $G(\hat{X}^{(t+1)}, \hat{X}^t) \geq L(\hat{X}^{(t+1)})$ and $G(\hat{X}^{(t+1)}, \hat{X}^t) = L(\hat{X}^{(t+1)})$ for any $\hat{X}^{(t+1)}$ and $\hat{X}^t$ when $S_i$ and $\hat{Y}_i$ are fixed. If we define

$$\hat{X}^{(t+1)} = argmin\, G(\hat{X}^{(t+1)}, \hat{X}^{(t)})$$

then we have

$$L(\hat{X}^{(t)} = G(\hat{X}^{(t)}, \hat{X}^{(t)}) \geq G(\hat{X}^{(t+1)}, \hat{X}^{(t)}) \geq L(\hat{X}^{(t+1)})$$

then $L(\hat{X}^{(t)})$ is monotonically decreasing or nonincreasing. We have

$$L\left(\hat{X}^{(t+1)}\right) = tr(\sum_{i=1}^{N} w_i Q_i Q_i^T)$$
$$- tr(2\sum_{i=1}^{N} w_i \hat{X}^{(t+1)^T} Q_i \hat{Y}_i S_i^T)$$
$$+ tr((\sum_{i=1}^{N} w_i S_i \hat{Y}_i^T \hat{Y}_i S_i^T + \lambda_0) \hat{X}^{(t+1)^T} \hat{X}^{(t+1)})$$

We can show that the following function is an auxiliary function of $L(\hat{X}^{(t+1)})$.

$$G\left(\hat{X}^{(t+1)}, \hat{X}^{(t)}\right)$$
$$= \sum_{i=1}^{N} w_i \|Q_i\|^2$$
$$- 2\sum_{jk}(\sum_{i=1}^{N} w_i \hat{X}^{(t+1)^T} Q_i \hat{Y}_i S_i^T)_{jk}$$
$$+ \sum_{jk} \frac{\left[\hat{X}^{(t)}(\sum_{i=1}^{N} w_i S_i \hat{Y}_i^T S_i^T + \lambda_0)\right]_{jk} \hat{X}_{jk}^{2(t+1)}}{\hat{X}_{jk}^{(t)}}$$

It is obvious that when $\hat{X}^{(t)} = \hat{X}^{(t+1)}$ the equality holds $(\hat{X}^{(t+1)}, \hat{X}^t) = L(\hat{X}^{(t+1)})$ . Second, we can show that the inequality $G(\hat{X}^{(t+1)}, \hat{X}^t) \geq L(\hat{X}^{(t+1)})$ holds. The third term in $G(\hat{X}^{(t+1)}, \hat{X}^t)$ is always greater than or equal ti the third term in $L(\hat{X}^{(t+1)})$. If we take the gradient and set it to sero, we can obtain the minimum of $G(\hat{X}^{(t+1)}, \hat{X}^t)$.

$$\frac{\partial G(\hat{X}^{(t+1)}, \hat{X}^{(t)})}{\partial \hat{X}_{jk}^{(t+1)}} = -2\sum_{jk}(\sum_{i=1}^{N} w_i Q_i \hat{Y}_i S_i^T)_{jk}$$
$$+ 2\sum_{jk} \frac{\left[\hat{X}^{(t)}(\sum_{i=1}^{N} w_i S_i \hat{Y}_i^T S_i^T + \lambda_0)\right]_{jk} \hat{X}_{jk}^{(t+1)}}{\hat{X}_{jk}^{(t)}}$$
$$= 0$$

If we substitute $\lambda_0$ in Eq. (3.11) into the above equation, we can derive the update rule of Eq. (3.4) for *X*. Thus we have $\hat{X}^{(t+1)} = argmin\, G(\hat{X}^{(t+1)}, \hat{X}^t)$. Therefore, under this update rule, $L(\hat{X}^{(t)})$ decreases monotonically when $S_i$ and $Y_i$ are fixed.

Similarly, we can also prove that the update rules of Eqs. (3.5, 3.6) are also monotonically decrease. Obviously, since Eq. (2.3) is bound from below, the algorithm will converge.

### C. *Initialization*

We may use K-means to initialize. The rules can be defined as follows:

(1) $\hat{X}$ : run K-means on rows of any of $Q_i$ with cluster number k, then we set $\hat{X}$← $\hat{X}$ + 0.2.

(2) $\hat{Y}_i$: run K-means on columns of $Q_i$ with cluster number $k_i$, then we set $\hat{Y}_i$← $\hat{Y}_i$+ 0.2.

(3) For each $S_i$, set the derivative $\frac{\partial J}{\partial S_i} = 0$ in Eq. (2.3), we obtain $S_i = \hat{X}^T Q_i \hat{Y}_i$.

## IV. DISCUSSION

In this section, we discuss the relationships between ONRC, spectral graph clustering and other related multi-type relational data clustering algorithms.

### A. Graph Clustering

Graph clustering is an important problem in many fields. Existing algorithms are based on different definitions of graph cut, such as minmax cut [12], ratio cut [20], NCut [33] etc. Spectral graph clustering is equivalent to NMF on a pairwise similarity matrix [11]. ONRC is based on this idea, and further extract the relation information between different data objects by matrix approximation on inter-relational matrices. Dhillon et al. [9] propose bipartite spectral graph partitioning to co-cluster relational data. The algorithm is converted to a singular value decomposition. In ONRC, the matrix factorization for each bipartite graph is $\|Q_i - XS_iY_i^T\|^2$, , if we restrict $S_i$ to be diagonal, the matrix factorization is equivalent to SVD. Because of the diagonality, bipartite spectral clustering can only have same number of clusters in different type of objects. In ONRC, the restriction is removed so that ONRC could have different number of clusters in different types of data objects. Gao et al. [17] propose a consistent bipartite graph co-partitioning (CBGC) algorithm for heterogeneous data co-clustering. CBGC is a tripartite graph partition algorithm and tries to find a consistent partition in the common data objects shared by both bipartite graphs. The algorithm is based on graph clustering theory and seeks to minimize sum of normalized cuts of two bipartite graphs. In this sense, ONRC is equivalent to CBGC with $S_i$ be identity matrices, but CBGC formulates graph cut minimization as an SDP problem which is known for high computation cost.

### B. Other relational ClusteringAlgorithms

Several relational data clustering algorithms have been proposed, such as SRC [26], relational summary network (RSN) [25], mixed membership relational clustering (MMRC) [28] etc. SRC has an objective function:

$$L = \sum_{1 \le i \le j \le m} w_a^{ij} \left\|R^{ij} - C^i A^{ij} (C^j)^T\right\|^2$$

where $C^i$ is indicator matrix. This formula is similar to Eq.(2.3). SRC uses spectral decomposition by taking the leading eigenvectors to minimize the objective function, while ONRC uses orthogonal NMF by multiplicative update, but the computation of eigenvectors is always expensive. SRC has to use K-means or other post-processing methods to extract cluster structure, which increases both computation cost and clustering errors.

RSN tries to approximate a k-partite graph to a hidden relation summary network by minimizing

$$\sum_{1 \le i \le j \le m} D(A^{ij}, C^i B^{ij} (C^j)^T)$$

where $C^i$ is an indicator matrix, D is a distance function. The distance function is generalized as Bregman divergence. RSN can be considered as a generalized K-means on k-partite graphs with various Bregman divergences. If we choose D as euclidean distance, the objective function of ONRC is equivalent to that of RSN with normalized relational matrices and orthogonality.

MMRC is based on a probabilistic model for relational clustering. Membership vector of each object is assigned based on parameters which denote the probability the object associates with each latent class. MMRC seeks to maximize a log-likelihood function which is an estimation of the parameters for latent variables. It has been shown that the edge-cut based graph clustering is equivalent to MMRC model under normal distribution with the diagonal constraint on parameter matrix[28]. In hard version of MMRC, by omitting the soft membership parameters, the maximization of a log-likelihood function of hard clustering on a heterogeneous relation matrix is equivalent to minimize $D(R, (C^i)^T \gamma^{C^j}$ , where D is a distance function, is relational representative matrix.

## V. EXPERIMENTAL EVALUATION

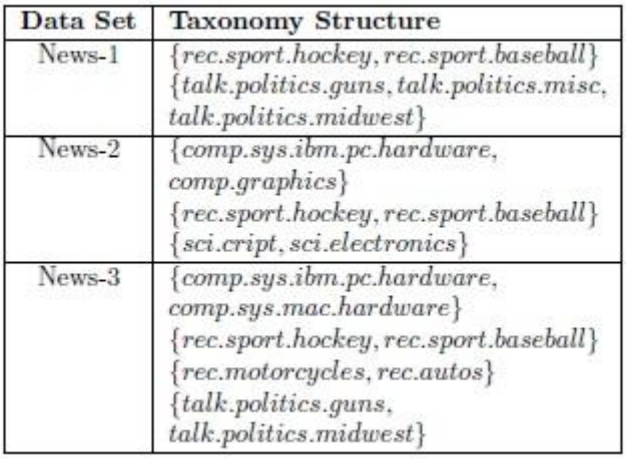
### A. Tri-type Relational Data Clustering

In this section, we evaluate the effectiveness of the ONRC algorithm on tri-type star-structured data as shown in Figure 1. The data sets used in the experiments are from the 20-Newsgroup data (http://people.csail.mit.edu/˜jrennie/20Newsgroups). We use text classification package Rainbow (http://www.cs.cmu.edu/˜mccallum/bow/rainbow/) to preprocess the data by removing stop words and file headers and selecting words with more than 5 counts. The document-word matrix is based on tf-idf. The document-category matrix *R* was built as follows. The rows correspond to categories, and columns to documents. $R_{ij}$ indicates the relation between category $C_i$ and document $d_j$. If $d_j$ belongs to *k* categories $C_1$, $C_2$, . . . ,$C_k$, the weights $R_{1j}$ ,$R_{2j}$ , . . . ,$R_{kj}$ are set to 1/k. All other elements of this column are set to 0. Three data sets, News-1, News-2 and News-3 are listed in Table II. The documents in each data set are generated by sampling 100 documents from each category.

The number of clusters for documents and categories are 2, 3 and 4 for News-1, News-2 and News-3, respectively. For the number of word clusters, we adopt the real number of categories, 5, 6 and 8 for three datasets. To check for document-word co-clustering, we pick News-2 and use 3 as number of clusters for both documents and words. We simply use equal weight $w_i$ for each $R_i$ and set $w_i$ = 1. If we set weight for one $X - Y_i$ be 0, then it will reduce tri-type to bi-type, which is essentially a bipartite clustering.

We put ONRC, SRC, RSN and spectral co-clustering [18, 9] in comparison. The spectral clustering can only have same numbers of row and column clusters, we choose the number of document clusters in experiments. We use the Normalized Mutual Information (NMI) and cluster error rate to evaluate the clustering quality.

TABLE II TAXONOMY STRUCTURE FOR THREE DATA SETS

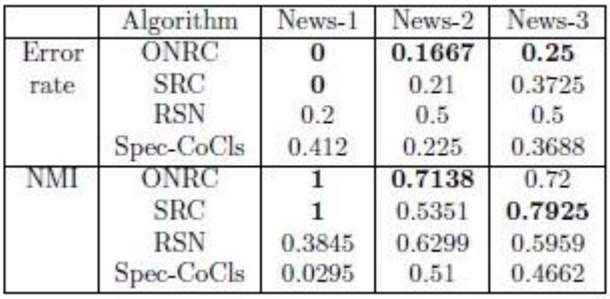


| Data Set | Taxonomy Structure |
|---|---|
| News-1 | {*rec.sport.hockey*, *rec.sport.baseball*}<br>{*talk.politics.guns*, *talk.politics.misc*, *talk.politics.midwest*} |
| News-2 | {*comp.sys.ibm.pc.hardware*, *comp.graphics*}<br>{*rec.sport.hockey*, *rec.sport.baseball*}<br>{*sci.cript*, *sci.electronics*} |
| News-3 | {*comp.sys.ibm.pc.hardware*, *comp.sys.mac.hardware*}<br>{*rec.sport.hockey*, *rec.sport.baseball*}<br>{*rec.motorcycles*, *rec.autos*}<br>{*talk.politics.guns*, *talk.politics.midwest*} |

## *B. Performance*

Table III shows error rates and NMI scores of the three algorithms on all the data sets. Best values are in bold. The comparison shows that ONRC is an effective algorithm to identify the cluster structure of star-structured multi-type relational data. CBGC can generate similar cluster quality as ONRC on News-1, but the running time is much longer when we use SDPT3 solver [36], so we didn't continue to test. It's been shown that SRC runs slower than tri-NMF [7]. We see that ONRC is an efficient way for multi-relational clustering.

TABLE III ERROR RATE AND NMI COMPARISONS OF ONRC, SRC AND RSN ALGORITHMS



| | Algorithm | News-1 | News-2 | News-3 |
|---|---|---|---|---|
| Error rate | ONRC | **0** | **0.1667** | **0.25** |
| | SRC | **0** | 0.21 | 0.3725 |
| | RSN | 0.2 | 0.5 | 0.5 |
| | Spec-CoCls | 0.412 | 0.225 | 0.3688 |
| NMI | ONRC | **1** | **0.7138** | 0.72 |
| | SRC | **1** | 0.5351 | **0.7925** |
| | RSN | 0.3845 | 0.6299 | 0.5959 |
| | Spec-CoCls | 0.0295 | 0.51 | 0.4662 |

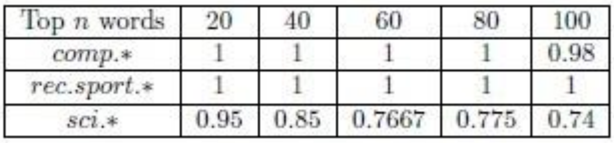
## *C. Document-words Co-clustering*

To check for the co-clustering quality, we calculate the mutual information [38] for each word in each category (*comp, rec.sport and sci*) from News-2. If a category *c* and a term *t* have probabilities *P(c)* and *P(t),* then their mutual information *I(t, c)* is defined to be:

$$I(t,c) = \log_2 \frac{p(t,c)}{p(t) \times p(c)} = \log_2 \frac{p(t \wedge c)}{p(t) \times p(c)}$$

We then sort those words based on the mutual information values. We check if words with higher value of mutual information have been correctly clustered with corresponding category. Table IV shows the percentage of top 100 words correctly co-clustered with corresponding category. We observe high rate of correct co-clustering. Table V lists 15 words with highest mutual information in each category.

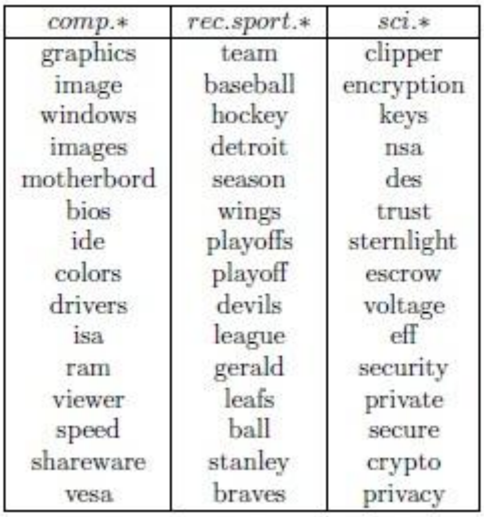
TABLE IV PERCENTAGE OF TOP 100 WORDS FROM NEWS-2 IN MUTUAL INFORMATION CO-CLUSTERED WITH CORRESPONDING CATEGORY



| Top $n$ words | 20 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|
| *comp.** | 1 | 1 | 1 | 1 | 0.98 |
| *rec.sport.** | 1 | 1 | 1 | 1 | 1 |
| *sci.** | 0.95 | 0.85 | 0.7667 | 0.775 | 0.74 |

.

TABLE V TOP 15 WORDS SORTED BY MUTUAL INFORMATION FROM NEWS-2



| *comp.** | *rec.sport.** | *sci.** |
|---|---|---|
| graphics | team | clipper |
| image | baseball | encryption |
| windows | hockey | keys |
| images | detroit | nsa |
| motherbord | season | des |
| bios | wings | trust |
| ide | playoffs | sternlight |
| colors | playoff | escrow |
| drivers | devils | voltage |
| isa | league | eff |
| ram | gerald | security |
| viewer | leafs | private |
| speed | ball | secure |
| shareware | stanley | crypto |
| vesa | braves | privacy |

# VI. CONCLUSIONS

In this paper, we present a model for clustering multi-type relational data based on inter-relational matrices. This model is essentially a matrix factorization. Under this model, we propose an algorithm ONRC to cluster multi-type interrelated data objects simultaneously. Objective function relaxes the strict double orthogonal bi-NMF of graph cut minimization to orthogonal triNMF. ONRC exploits successive updates to minimize the matrix approximation error and also keeps the orthogonality of indicator matrices. The correctness and convergence are also proved in this paper. We observe ONRC performs better than SRC, RSN and spectral co-clustering algorithm. ONRC also reveals the abilities to co-clustering different data objects.

# REFERENCES

[1] A. Anagnostopoulos, A. Dasgupta, and R. Kumar, Approximation algorithms for co-clustering, in Proc. of the 27th ACM SIGMOD-

SIGACT-SIGART symposium on Principles of database systems, 2008, pp. 201–210.

[2] F. R. Bach and M. I. Jordan, Learning spectral clustering, with application to speech separation, Journal of Machine Learning Research, 7 (2006), pp. 1963–2001.

[3] A. Banerjee, S. Basu, and S. Merugu, Multi-way clustering on relation graphs, in Proceedings of the 7th SIAM International Conference on Data Mining, 2007.

[4] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, A generalized maximum entropy approach to bregman co-clustering and matrix approximation, in Proc. of the 10th ACMInternational Conference on Knowledge Discovery and Data Mining, 2004, pp. 509–514.

[5] S. Chen, F. Wang, and C. Zhang, Simultaneous heterogeneous data clustering based on higher order relationships, in Workshop Proc. of the 7th IEEE International Conference on Data Mining, 2007, pp. 387–392.

[6] Y. Chen, L. Wang, and M. Dong, A matrix-based approach for semi-supervised document co-clustering, in Proc. of the 17th ACM Conference on Information and Knowledge Management, 2008, pp. 1523–1524.

[7] Y. Chen, L. Wang, and M. Dong, Non-negative matrix factorization for semi-supervised heterogeneous data co-clustering, IEEE Trans. Knowl. Data Eng., (in press) (2009).

[8] A. D. Chiaravalloti, G. Greco, A. Guzzo, and L. Pontieri, An information-theoretic framework for high-order co-clustering of heterogeneous objects, in Proc. of the 17th European Conference on Machine Learning, 2006, pp. 598–605.

[9] I. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in Proceedings of the seventh ACM SIGKDD, ACM, August 2001, pp. 269–274.

[10] I. Dhillon, S. Mallela, and D. Modha, Information-theoretic co-clustering, in Proceedings of the Ninth ACM SIGKDD, 2003, pp. 89–98.

[11] C. Ding, X. He, and H. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, in Proc. SIAM Int'l Conf. Data Mining, 2005, pp. 606–610.

[12] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, A min-max cut algorithm for graph partitioning and data clustering, in Proceedings of ICDM, 2001, pp. 107–114.

[13] C. Ding, T. Li, W. Peng, and H. Park, Orthogonal nonnegative matrix trifactorizations for clustering, in Proceedings of the Twelfth ACM SIGKDD, 2006, pp. 126–135.

[14] R. Duda, P. Hart, and D. Stork, Pattern classification, Second Edition. John Wiley and Sons Inc., (2001).

[15] R. El-Yaniv and O. Souroujon, Iterative double clustering for unsupervised and semi-supervised learning, in Proc. of the 12th European Conference on Machine Learning, 2001, pp. 121–132.

[16] L. Everett, L.-S. Wang, and S. Hannenhalli, Dense subgraph computation via stochastic search, Bioinformatics, 22 (2006), pp. e117–e123.

[17] B. Gao, T. Liu, X. Zheng, Q. Cheng, and W. Ma, Consistent bipartite graph co-partitioning for starstructured high-order heterogeneous data co-clustering, in Proceedings of the eleventh ACM SIGKDD, 2005, pp. 41–50.

[18] A. Gottlieb, Spectral coclustering (biclustering) matlab implementation. http://adios.tau.ac.il/SpectralCoClustering/.

[19] M. Gu, H. Zha, C. Ding, X. He, and H. S. J. Xia, Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering, Penn State University Technical Report, (2001).

[20] L. Hagen and A. B. Kahng, New spectral methods for ratio cut partitioning and clustering, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 11 (1992), pp. 1074–1085.

[21] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Research, 13 (2003), pp. 703–716.

[22] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 401 (1999), pp. 788–791.

[23] T. Li, A general model for clustering binary data, in Proc. of the 11th ACM International Conference on Knowledge Discovery and Data Mining, 2005, pp. 188–197.

[24] S. P. Lloyd, Least squares quantization in pcm, Special issue on quantization, IEEE Trans. Inform. Theory, 28 (1982), pp. 129– 137.

[25] B. Long, X. Wu, Z. Zhang, and P. S. Yu, Unsupervised learning on k-partite graphs, in Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining, 2006, pp. 317–326.

[26] B. Long, Z. Zhang, X. Wu, and P. S. Yu, Spectral clustering for multi-type relational data, in Proc. of the 23rd international conference on Machine learning, 2006, pp. 585–592.

[27] B. Long, Z. Zhang, and P. S. Yu, Co-clustering by block value decomposition, in Proc. of the 11th ACM International Conference on Knowledge Discovery and Data Mining, 2005, pp. 635–640.

[28] B. Long, Z. Zhang, and P. S. Yu, A probabilistic framework for relational clustering, in Proc. of the 13th ACM International Conference on Knowledge Discovery and Data Mining, 2007, pp. 470–479.

[29] P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5 (1994), pp. 111–126.

[30] F. Pan, X. Zhang, and W. Wang, A general framework for fast co-clustering on large datasets using matrix decomposition, in Proc. of the 24th International Conference on Data Engineering, 2008, pp. 1337–1339.

[31] M. M. Shafiei and E. E. Milios, Latent dirichlet coclustering, in Proc. of the 6th International Conference on Data Mining, 2006, pp. 542–551.

[32] Model-based overlapping co-clustering, in Proc. of the 4th Workshop on Text Mining, 6th SIAM International Conference on Data Mining, 2006.

[33] J. Shi and J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 888–905.

[34] S. Sra, S. Jegelka, and A. Banerjee, Approximation algorithms for bregman clustering co-clustering and tensor clustering, Technical Report of Max Planck Institute for Biological Cybernetics, No. 117 (2008).

[35] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing, 1999, pp. 368–377.

[36] K. Toh, M. Todd, and R. Tutuncu, A matlab software for semidefinite-quadratic-linear programming. http://www.math.nus.edu.sg/ mattohkc/sdpt3.html.

[37] J. Wang, H. Zeng, Z. Chen, H. Lu, T. Li, and W. Ma, Recom: reinforcement clustering of multi-type interrelated data objects, in SIGIR, 2003, pp. 274–281.

[38] Y. Xu, G. Jones, J.-T. Li, and B.Wang, A study on mutual information-based feature selection for text categorization, Journal of Computational Information Systems, 3 (2007), pp. 1007–1012.

[39] H. Zeng, Z. Chen, and W. Ma, A unified framework for clustering heterogeneous web objects, in Proc. 3rd