Clustering and Topic Discovery of Multiway Data via Joint-NCMTF

Benjamin Cobb, Ricardo Velasquez, Richard Vuduc, Haesun Park
{bcobb33, rvelaquez3}@gatech.edu, {richie, hpark}@cc.gatech.edu
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, GA

***Abstract*—Nonnegative Matrix Factorization (NMF) and Nonnegative Coupled Matrix Tensor Factorization (NCMTF) are Constrained Low-Rank Approximation (CLRA) models which have found use in many applications. In particular, NMF and its variants have been shown to produce high-quality soft clustering and topic modeling results with the property that each clustering assignment relates to a corresponding topic; thereby providing insight into the nature of each item in a given cluster. However, NMF and its variants are unable to process heterogeneous data represented as one or more coupled tenors. Similarly, there do not exist tensorized methods which fully preserve the aforementioned desirable clustering and topic modeling properties of NMF. This paper develops a higher order analog of Joint-NMF, Joint Nonnegative Coupled Matrix Tensor Factorization (Joint-NCMTF), capable of factorizing heterogeneous tensor datasets whilst fully preserving these NMF properties. To accomplish this, we develop higher-order analogs of the entire NMF process, including crucial pre and post-processing steps. By incorporating additional dimensions of information present in datasets posed as coupled higher-order tensors, our proposed Joint-NCMTF method yields higher quality clustering and topic modeling results than methods which incorporate less information. We empirically demonstrate the effectiveness of our proposed method on multiple synthetic and two real-world topic modeling tasks.**

***Index Terms*—Numerical analysis, approximation methods, text analysis, clustering methods**

I. Introduction

Topic modeling and clustering are both increasingly important data analysis tasks. Applications across many disciplines rely on them to discover latent information in data such as in social networks [27], document analysis [31], and psychometric studies [1]. Constrained Low-Rank Approximations (CLRA) are a popular class models for addressing these tasks, one of the most well known being Nonnegative Matrix Factorization (NMF) [21]. NMF and its variants such as Joint Nonnegative Matrix Factorization (Joint-NMF) [12] impose nonnegativity on the factors such that cluster assignments can be directly extracted without the need for downstream post-processing clustering algorithms. The resulting cluster assignments have the desirable the property that each cluster assignment relates to a corresponding topic; thereby providing insight into the nature of each item in a given cluster.

However, matrix based methods such as NMF are inherently constrained in the type of datasets they can process and are unable to handle datasets posed as one or more higher-order tensors, e.g. Figure 1. This restriction limits matrix based methods from incorporating potentially useful multiway information in datasets which can be formulated as tensors. For this reason, significant amounts of effort have been invested in developing tensor based methods [2] [11] [15] [26] [18] [9] [5] [37] based upon low-rank approximations such as Nonnegativve CANDECOMP-PARAFAC (NCP) and Nonnegative Coupled Matrix Tensor Factorization (NCMTF), which are capable of handling these datasets.

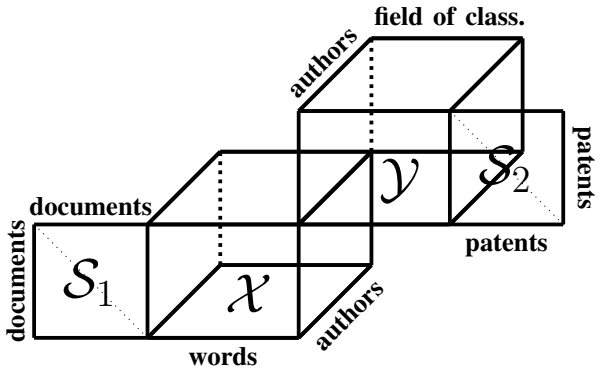


Fig. 1: Example Coupled Matrix Tensor dataset

We propose a method based upon NCMTF which can be viewed as a higher-order analog of Joint-NMF capable of handling datasets posed as coupled higher-order tensors, e.g. Figure 1, whilst preserving the desirable topic modeling capabilities of NMF. By incorporating additional dimensions of information present in these datasets, our proposed Joint Nonnegative Coupled Matrix Tensor Factorization (Joint-NCMTF) method yields higher-quality clustering and topic modeling results than factorization methods which incorporate less information. To our knowledge, we are the first to investigate factoring heterogeneous datasets posed as coupled tensor and matrices such that clusters in conjunction with topic models can be directly extracted from the factors similar to NMF. Our proposed methodology leverages several novel observations, including important data processing steps, to yield high-quality clustering assignments. We demonstrate the effectiveness of our proposed method on synthetic and real-world datasets in terms of commonly used quantitative metrics as well as qualitatively on real-world topic modeling tasks.

II. Preliminaries

A. Related Work

There are many algorithms for computing matrix and tensor decompositions due to the ubiquitousness of datasets posed in these forms. Two such algorithms for CMTF are the Alternating Least Squares (CMTF-ALS) and the gradient-based "all-at-once" methods proposed in [2]. The CMTF-ALS algorithm iteratively solves for each factor matrix whilst holding the other factor matrices constant in a Block Coordinate Descent (BCD) framework. In this work, we adapt the CMTF-ALS algorithm to enforce nonnegativity and symmetry constraints.

Several methods for analyzing multi-way datasets represented by coupled matrices and tensors are available. A framework for solving the multi-way clustering problem on coupled matrix tensor problems is proposed in [6], which they refer to as relation graphs, based upon nonnegative alternating minimization of KL-divergence and minimum Bregman divergence. A fast method for computing a Tucker decomposition based CMTF is developed in [11] that yields a higher Gap statistic [35] than Tucker when applying K-Means to the factors. A K-Means based algorithm is developed in [28] for co-clustering sparsity constrained multi-linear decomposition.

Distributed CMTF implementations based on MadReduce [18] and Hadoop [9] frameworks have been proposed for topic modeling problems. Distributed and gradient based methods for Joint-NMF are presented in [13]. Similar methodologies developed for nonnegative tensor decompositions [29] exist for clustering. Tensor based methods are used by many practitioners to solve many real-world problems [15] [3] [4].

Several CMTF works impose nonnegativity or sparsity constraints on the factorization, e.g. [28], [15], and [20]. Nonnegative factors can be interpreted as soft or hard clustering assignments. However, we emphasize that all existing NCP based NCMTF work, to the best of our knowledge, does not explicitly leverage the nonnegativity to do clustering along a single mode. As such, existing NCMTF work does not investigate what additional steps or constraints are necessary in addition to nonnegativity to yield high quality cluster assignments. There are several steps in addition to enforcing nonnegativity which we do in this work to yield factors from which high-quality clustering assignments can be extracted. Foremost amongst these are higher-order analogs of the NMF pre and post-processing normalization steps.

Further reference on tensor decompositions and formulations can be found in [22]. For a survey on comparisons between matrix and tensor subspace clustering methods which do not utilize nonnegativity, we refer the reader to [37]. Various CMTF and tensor-based clustering approaches appear in: [17] [24] [30] [33] [7] [8] [26]. BCD methods in the context of tensor and matrix decompositions are discussed in [19]. Finally, we utilize the MatLab Tensor Toolbox [10] for all tensor operations in our algorithms.

B. Notation

We denote tensors by boldface typescript letters (e.g. $\mathcal{X}$), matrices by uppercase letters (e.g. $\mathbf{M}$), vectors by boldface lowercase letters (e.g. $\mathbf{v}$), scalars by lowercase letters (e.g. a), a matrix norm by $||\cdot||$, and Frobenius norm by $||\cdot||_F$. We use MatLab indexing notation to index into matrices. We use MatLab built-in and MatLab Tensor Toolbox functions such as $normr, diag, vecnorm, graph, degree, max, maxk$, and $collapse$ in Algorithm 1 [16] [10]. We use standard notation for matrix and tensor operations throughout the paper. Let $\mathcal{X}$ be an N-order tensor. $\mathcal{X}_{:,i_2,\ldots,i_N}$, $\mathcal{X}_{:,:,i_3,\cdots,i_N}$, and $\mathcal{X}_{(1)}$ denote a mode-1 fiber, a slice along mode-1 and mode-2, and a mode-1 matricization, respectively. The notation generalizes to any mode-n tensor operations used in this paper. For matrix operations, $\circ$, $\otimes$, $\odot$, and $*$ denote the outer, Kronecker, Kathri-Rao, and Hadamar products between two matrices, respectively. $\mathcal{X} \approx [\![\mathbf{A}_1, \ldots, \mathbf{A}_N]\!] = \sum_{i=1}^{r} \mathbf{A}_1(:,i) \circ \cdots \circ \mathbf{A}_3(:,i)$ denotes the rank-r CP decomposition of $\mathcal{X}$.

III. Joint Nonnegative Coupled Matrix Tensor Factorization

A. Modeling

Joint-NCMTF can be applied to a dataset consisting of any number of tensors coupled along one or more modes, with an arbitrary number of coupled matrices along each mode. Each Joint-NCMTF will yield a set of factor matrices corresponding to each mode of the tensors, with a coupled matrix sharing a factor. A low-rank approximation of the original can be derived by summing the higher-order outer products of the factor matrices. Each order-n tensor $\mathcal{X}^i$, including the coupled matrices, can be factorized into a rank-r CP factorization $\mathcal{X}^i \approx [\![\mathbf{A}_1^i, \ldots, \mathbf{A}_n^i]\!] = \sum_{j=1}^{r} \mathbf{A}_1^i(:,j) \circ \cdots \circ \mathbf{A}_n^i(:,j)$.

For each coupling between two tensors $\mathcal{X}^i$ and $\mathcal{X}^j$ respectively along modes k and l, we enforce that the factor matrices along the corresponding modes are equivalent, that is $\mathbf{A}_k^i = \mathbf{A}_l^j$. Thus, the total number of factor matrices will be $F - C$ where F is the total sum of all order counts for each tensor, including coupled matrices, in the dataset and C is the number of couplings.

To simplify the analysis, in this work we only consider the case wherein the tensor is of third-order with a single nonsymmetric matrix and a single symmetric matrix coupled along the first mode. However, the framework proposed as part of this work is capable of handling the more general case.

B. Joint-NCMTF

Consider a third-order tensor, $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, coupled along its first mode with a feature matrix, $\mathbf{M}_1 \in \mathbb{R}^{I_1 \times F_1}$, and symmetric data-data matrix, $\mathbf{S}_1 \in \mathbb{R}^{I_1 \times I_1}$. In this case, the NCMTF problem can be represented by the objective function:

$$\min_{[\mathbf{A}_1, \hat{\mathbf{A}}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{V}_1] \geqslant 0} ||\mathcal{X} - [\![\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]||_F^2 \tag{1}$$

$$+ \alpha_1 ||\mathbf{M}_1 - \mathbf{A}_1 \mathbf{V}_1^T||_F^2 \tag{2}$$

$$+ \alpha_2 ||\mathbf{S}_1 - \hat{\mathbf{A}}_1 \mathbf{A}_1^T||_F^2 \tag{3}$$

$$+ \beta ||\hat{\mathbf{A}}_1 - \mathbf{A}_1||_F^2 \tag{4}$$

Where $\mathbf{V}_1 \in \mathbb{R}^{F_1 \times r}, \mathbf{A}_1 \in \mathbb{R}^{I_1 \times r}, \hat{\mathbf{A}}_1 \in \mathbb{R}^{I_1 \times r}, \mathbf{A}_2 \in \mathbb{R}^{I_2 \times r}$, and $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times r}$ represent the factors resulting from the decomposition of $\mathcal{X}$ in the framework of this model, and r denotes the desired rank. For a graphical representation of the problem, refer to Figure 1.

Symmetry can result from the mutual interactions between data items and is commonly represented as an undirected graph adjacency matrix denoted by $\mathbf{S}_1$ in 3. This symmetry can contain key insights into the low-rank structure of a dataset [23] [12]. To incorporate this, we developed a higher-order-analog of Joint-Nonnegative Matrix Factorization (Joint-NMF) [12] via a symmetry regularization surrogate variable in 4.

Directly forcing $\hat{\mathbf{A}}_1 = \mathbf{A}_1$ makes (3) difficult to solve via ANLS, necessitating relaxing the problem by removing this strict constraint and instead constraining $\hat{\mathbf{A}}_1 \approx \mathbf{A}_1$ via regularization in (4) [23]. In this case, β is the regularization coefficient. The larger β, the more penalized the objective function for differences between $\hat{\mathbf{A}}_1$ and $\mathbf{A}_1$.

It is possible to incorporate nonsymmetric information into the dataset by coupling a matrix $\mathbf{M}_1$ along the first mode of $\mathcal{X}$, as shown in (2). Observe that this can be extended to accommodate several nonsymmetric matrices by concatenating them and setting $\mathbf{M}_1$ equal to the result.

We can iteratively solve the NNLS subproblems via an active-set based method [21] in a five-block BCD scheme, updating $\mathbf{V}_1, \mathbf{A}_2, \mathbf{A}_3, \hat{\mathbf{A}}_1, \mathbf{A}_1$ as shown in lines 7 to 13 of Algorithm 1. This is based on the CMTF-ALS algorithm presented in [2] and which we refer to as Joint-NCMTF.

Observe that the NCMTF model is equivalent to setting $\alpha_2 = \beta = 0$ (eqs. (1) to (4)) in the Joint-NCMTF formulation.

C. Auto-Coupling

Increasing the order of the dataset formulation inherently sparsifies the problem. This can result in sparse factors with entire rows of zeros from which it is not possible to extract clustering assignments. To address this, we added a constraint to penalize the zero rows in the factors. We propose a method we refer to as "Auto-Coupling", which couples the tensor formulation to itself reduced in all but two modes, i.e. coupling the matrix formulation to the tensor formulation. In our experiments, to compute the Auto-Coupled matrix we let $\mathbf{M}_1 \in \mathbb{R}^{I_1 \times I_2}$ be $\mathcal{X}$ reduced along the third mode and $\mathbf{V}_1 \in \mathbb{R}^{I^2 \times r}$ be the factor matrix corresponding to $\mathbf{M}_1$. We scale the weight of Auto-Coupling using the parameter α_1.

IV. Multiway Data Analysis via NCMTF

CMTF has previously been proposed as a dimensionality reduction technique for use in conjunction with existing clustering methods like K-Means [2] [11] [14]. However, to our knowledge, little work has been done to extract cluster assignments from a single NCMTF factor without the use of downstream clustering algorithms. In this work we propose a method for directly extracting clustering results from a NCMTF factor. This preserves a desirable topic modeling property of NMF, wherein each clustering assignment relates to a corresponding topic; thereby providing insight into the nature of each item in a given cluster.

A. NCMTF Normalization and Clustering

Generally when applying NMF or its variants to the task of clustering a matrix of m data items and n features, $\mathbf{X} \in \mathbb{R}^{m \times n}$, the rows of $\mathbf{X}$ are normalized to unit Frobenius norm. This ensures that each data item is weighted equally in the factorization and that the data items' clustering is done based upon the proportion of a data item's features as opposed to the magnitude of the features.

In Joint-NMF [12], the coupled symmetric adjacency matrix $\mathbf{S}$ is scaled by: Let $\mathbf{A}$ be the unnormalized graph adjacency matrix and $\mathbf{D} \in \mathbb{R}^{m \times m}$ be the diagonal matrix with the degrees of each vertex along the diagonal. Based upon this, let $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. We applied this type of normalization to all coupled symmetric adjacency matrices in our experiments.

In the tensor case, instead of normalizing rows, we normalize the tensor slices along the clustering mode to unit Frobenius norm. For example, let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. In the case of clustering along the first mode, we set each slice $\mathcal{X}_{i::} = \frac{\mathcal{X}_{i::}}{||\mathcal{X}_{i::}||_F}$. This is a higher order analog of NMF normalization which we observed improves cluster quality.

Similarly, we normalize the rows corresponding to data item features of all nonsymmetric coupled matrices to unit norm. We then normalize the coupled feature matrices weighted by hyperparameter coefficients and tensor matricized along the clustering mode together. Intuitively, this is akin to projecting all data items onto a unit hypersphere. The pseudocode of the aforementioned pre-processing normalization steps can be seen in lines 1 to 6 of Algorithm 1.

The rows of the nonnegative CP factors $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ computed as part of eqs. (1) to (4) can be implicitly considered as soft clustering assignments corresponding to each of the r columns. These can be used to extract hard clustering assignments by selecting the largest element per data item row of the scaled factor matrix along the mode of interest, as seen in line 16 of Algorithm 1. Note that we select cluster assignments from the factor matrix along a single given mode. This is not to be confused with co-clustering or multi-way clustering as is done in work such as [6].

When making cluster assignments, we normalize the factor columns along the nonclustering modes and shift the resulting column weights to the clustering mode's factor. For example, when clustering along the first mode we set $\mathbf{A}_1 = \mathbf{A}_1\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$, where $\mathbf{D}_i$ is the diagonal matrix containing the column weights of the mode-i factor matrix. This is a higher-order analog of a post-processing normalization step used in NMF clustering. The pseudocode of this rescaling process can be seen in lines 14 and 15 of Algorithm 1.

B. Topic Modeling via NCMTF

We extract topic modeling results from the NCMTF factors similar to NMF. Each column of a factor matrix along a given mode corresponds to a "topic" of that mode. We select the x largest elements from the factor matrix column corresponding to that topic as seen in lines 17 to 19 of Algorithm 1. Repeating this along each mode for a given "topic" (factor column index) results in the top elements for a "topic" which when taken

together provide insight into the nature of that topic. Note that as NCMTF yields additional factors relative to NMF the resulting topic modeling is more informative.

Algorithm 1: Joint-NCMTF with Auto-Coupling

Input : $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}, \mathbf{S}_1 \in \mathbb{R}^{I_2 \times I_2}, \mathbf{M_1} \in \mathbb{R}^{I_2 \times J}, \alpha_1, \alpha_2, \beta.$ k, number of clusters and topic models. $topn$, number of top features.

Output: $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times k}, \hat{\mathbf{A}}_1 \in \mathbb{R}^{I_1 \times k}, \mathbf{A}_2 \in \mathbb{R}^{I_2 \times k}, \mathbf{A}_3 \in \mathbb{R}^{I_3 \times k}, \mathbf{V}_1 \in \mathbb{R}^{F_1 \times k}$

Pre-Processing Step:

/* normalize data item features */

$\mathcal{X}_{(1)} = normr(\mathcal{X}_{(1)})$

/* enforce auto-coupling */

$\mathbf{M}_1 = \left[normr(\mathbf{M}_1) | normr(collapse(\mathcal{X}, 3))\right]$

/* normalize data item features */

$\left[\sqrt{\alpha_1}\mathbf{M}_1 | \mathcal{X}_{(1)}\right] = normr(\left[\sqrt{\alpha_1}\mathbf{M}_1 | \mathcal{X}_{(1)}\right])$

/* normalize adjacency matrix */

$\mathbf{S}(\mathbf{S} > 0) = 1$

$\mathbf{D} = degree(graph(\mathbf{S}))$

$\mathbf{S} = \mathbf{D}^{\frac{-1}{2}}\mathbf{S}\mathbf{D}^{\frac{-1}{2}}$

Core Algorithm:

while *not converged* **do**

$\min_{\mathbf{V}_1 \geqslant 0} \|\sqrt{\alpha_1}\mathbf{M}_1 - \sqrt{\alpha_1}\mathbf{A}_1\mathbf{V}_1^T\|_F^2$

$\min_{\mathbf{A}_2 \geqslant 0} \|\mathcal{X}_{(2)} - \mathbf{A}_2(\mathbf{A}_3 \odot \mathbf{A}_1)^T\|_F^2$

$\min_{\mathbf{A}_3 \geqslant 0} \|\mathcal{X}_{(3)} - \mathbf{A}_3(\mathbf{A}_2 \odot \mathbf{A}_1)^T\|_F^2$

$$\min_{\hat{\mathbf{A}}_1 \geqslant 0} \left\| \begin{bmatrix} \sqrt{\alpha_2}\mathbf{A}_1 \\ \sqrt{\beta}I_1 \end{bmatrix} \hat{\mathbf{A}}_1^T - \begin{bmatrix} \sqrt{\alpha_2}\mathbf{S}_1^T \\ \sqrt{\beta}\mathbf{A}_1^T \end{bmatrix} \right\|_F^2$$

$$\min_{\mathbf{A}_1 \geqslant 0} \left\| \begin{bmatrix} \mathbf{A}_3 \odot \mathbf{A}_2 \\ \sqrt{\alpha_1}\mathbf{V}_1 \\ \sqrt{\alpha_2}\hat{\mathbf{A}}_1 \\ \sqrt{\beta}\mathbf{I}_1 \end{bmatrix} \mathbf{A}_1^T - \begin{bmatrix} \mathcal{X}_{(1)}^T \\ \sqrt{\alpha_1}\mathbf{M}_1^T \\ \sqrt{\alpha_2}\mathbf{S}_1^T \\ \sqrt{\beta}\hat{\mathbf{A}}_1^T \end{bmatrix} \right\|_F^2$$

end

Post-Processing Step:

/* scale clustering factor */

$\mathbf{D} = diag(vecnorm(\begin{bmatrix} \mathbf{A}_3 \odot \mathbf{A}_2 \\ \sqrt{\alpha_1}\mathbf{V}_1 \end{bmatrix}))$

$\mathbf{A_1} = \mathbf{A_1} * \mathbf{D}$

/* extract cluster assignments */

$[-, cluster\text{-}assignments] = max(\mathbf{A}_1, [], 2)$

/* extract topic modeling results */

$[-, document\text{-}index] = maxk(\mathbf{A}_1, topn)$

$[-, word\text{-}index] = maxk(\mathbf{A}_2, topn)$

$[-, author\text{-}index] = maxk(\mathbf{A}_3, topn)$

V. Experiments

A. Evaluation Metrics

Pairwise F1-Score (PWF1): The F1-score is an external metric commonly used to evaluate the quality of cluster results. We use a pairwise variant of the F1-score originally developed to measure the performance of Joint-NMF [12] referred to as the Pairwise F1-score. We define the Pairwise F1-score as follows. Each of the $\frac{n(n-1)}{2}$ pairs of data items fall into one of four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as defined in [12]. Based upon this, the Pairwise F1-Score is defined as follows:

$$PWF1 = \frac{2\#TP}{2\#TP + \#FN + \#FP} \tag{5}$$

We refer the reader to [25] for a comprehensive overview of other internal clustering metrics.

Silhouette Index (S-Index): The Silhouette Index is a commonly used internal metric for evaluating cluster separation based upon pairwise distances of inner and outer cluster assignments. Each point is assigned a "Silhouette" score:

$$\frac{b(i) - a(i)}{max(a(i), b(i))},$$

where $a(i)$ is the average distance of a point i to all other points in its assigned cluster, whilst $b(i)$ is the average distance of point i to points in the next closest cluster. Its range is between -1 to 1, with 1 representing perfect cluster separation and -1 representing the opposite thereof. We take each point's Silhouette score and average them together to yield a single score, which we refer to as "S-Index" from here on.

B. Datasets

1) Synthetic Datasets: To experiment with the performance of the proposed approaches for a variety of problem structures, we generate synthetic datasets with different noise levels and types. Each generated synthetic dataset consists of a third-order tensor and a coupled symmetric adjacency matrix along the first mode. We denote the resulting tensor and matrix respectively as $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and $\mathbf{S} \in \mathbb{R}^{I_1 \times I_1}$. We form each synthetic dataset from 3 factor matrices $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times r}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times r}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times r}$. Each row of $\mathbf{A}_1$ and $\mathbf{A}_3$ are random one-hot vectors, i.e. all elements equal to zero except one element equal to 1. This represents the ground truth cluster assignment of each data item. $\mathbf{A}_2$ is a random column normalized matrix. We then set $\mathcal{X} = [\![\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$ and $\mathbf{S} = \mathbf{A}_1\mathbf{A}_1^T$.

We introduced 5 types of noise into $\mathcal{X}$ and $\mathbf{S}$. Coupled symmetric matrix connection removal (sym-rm-noise): Symmetrically remove x percent of entries from $\mathbf{S}$, where x is the noise level. Coupled symmetric matrix perturbation (sym-noise): Symmetrically add $(1-x^2)$ nonzero entries to $\mathbf{S}$, where x is the noise level. All nonzero entries of $\mathbf{S}$ are then set to 1. This type of noise simulates noisy adjacency connections between data items in $\mathbf{S}$. Block perturbation (perturb-block-noise): Adds uniform noise to the nonzero elements of $\mathcal{X}$. Does not increase the density of the problem. Noise level scaled to same norm as $\mathcal{X}$ and then multiplied by noise level. Dense perturbation (perturb-dense-noise): Adds uniform noise to each entry of $\mathcal{X}$. Makes problem dense. Noise level scaled to same norm as $\mathcal{X}$ and then multiplied by noise level. Sparsification (sparsify-noise): Sparsifies problem by zeroing x percent of entries from $\mathcal{X}$, where x is the noise level.

Unless stated otherwise, all synthetic experiments have a true factor rank, i.e. the number of columns in each factor, $r = 6$ and set sym-noise=.05, sym-rm-noise=.9, perturb-block-noise=0, perturb-dense-noise=.5, and sparsify-noise=.9.

Model Parameters						
Method	$\mathcal{X}$	$\mathbf{M}_1$	$\mathbf{S}_1$	α_1	α_2	β
K-Means	$I_1 \times I_2$	-	-	-	-	-
NMF	$I_1 \times I_2$	-	-	0	0	0
Joint-NMF	$I_1 \times I_2$	-	$I_1 \times I_1$	0	1	1
NMF-MOD	$I_1 \times I_2$	$I_1 \times I_3$	-	1	0	0
CP	$I_1 \times I_2 \times I_3$	-	-	0	0	0
Joint-NCMTF	$I_1 \times I_2 \times I_3$	$I_1 \times F_1$	$I_1 \times I_1$	1	1	1

TABLE I: Problem formulation comparison of surveyed methods. The algorithms for Joint-NCMTF and NMF-MOD are proposed as part of this work. Note that NMF-MOD can be considered as fitting into the framework proposed by [32]. All methods, except K-Means, fit within the Joint-NCMTF framework. In subsequent experiments, $F_1 = I_2$ when auto-coupling tensor based methods.

2) AMiner: Real-world experiments were run on subsets of the ArnetMiner (AMiner) [34] academic graph. A set of keywords such as ['mechanical engineer', 'economy', 'politic'] were selected. All documents without these substrings in either the *keyword* or *fos* (field of study) were filtered out. All documents without *title*, *abstract*, *author*, or *venue* ids were filtered out. Standard preprocessing was performed, dropping all words above or below a certain frequency along with stop word removal to form Term Frequency-Inverse Document Frequency (TF-IDF) vectors. Stemming was not performed. We then selected the top x documents in terms of citation relationship count, pruned all authors with less than 3 documents, and used the resulting corpus to populate $(document \times word \times author)$ tensor and corresponding $(document \times document)$ citation relationship adjacency matrix.

3) PatentsView: Real-world experiments were run on subsets of the PatentsView dataset [36] which consists of over 12 million patents with information such as title, abstract, inventors, assignee, and citation. Each patent belongs to one of 7 different categories and one of 38 different subcategories. Each patent has at least one *assignee*, which is the entity to which the patent was assigned to. Generally each assignee was a corporation. We filtered out the top patents for the selected subcategory-assignee pair in terms of citation relationships. We then applied similar preprocessing steps as on the Aminer dataset to the titles and abstracts of the selected patents and used the results to form TF-IDF vectors. The resulting TF-IDF vectors were then used to form a $(patent \times words \times inventor)$ tensor and corresponding $(patent \times patent)$ citation relationship adjacency matrix.

C. Quantitative Results

We compare the proposed Joint-NCMTF algorithm against several popular state-of-the-art baselines listed in Table I. Unless otherwise stated, all experiments were run with $\alpha_1 = \frac{||\mathcal{X}||_F^2}{16||\mathbf{M}_1||_F^2}$, $\alpha_2 = \frac{||\mathcal{X}||_F^2}{||\mathbf{S}||_F^2}$ and $\beta = 0.25$. The performance of Joint-NCMTF is sensitive to the setting of α_2 and β. Rigorous hyperparameter tuning is recommended to yield the best performance of Joint-NCMTF. Deriving default α_2 and β with rigorous justification is an avenue of future research.

Figures 2 to 4 show the comparisons between the surveyed methods for the aforementioned datasets and metrics. The center marker of all line plots is the average over 5 random initializations, whilst the lower and upper bars each respectively denote the minimum and maximum. To account for factor sparsity, in Figures 3 and 4 we select the top 25% of data items in terms of clustering factor column weight to evaluate the clustering metrics. In the instance that a method assigns less than 25% of data items to clusters we set the score corresponding to the method to zero.

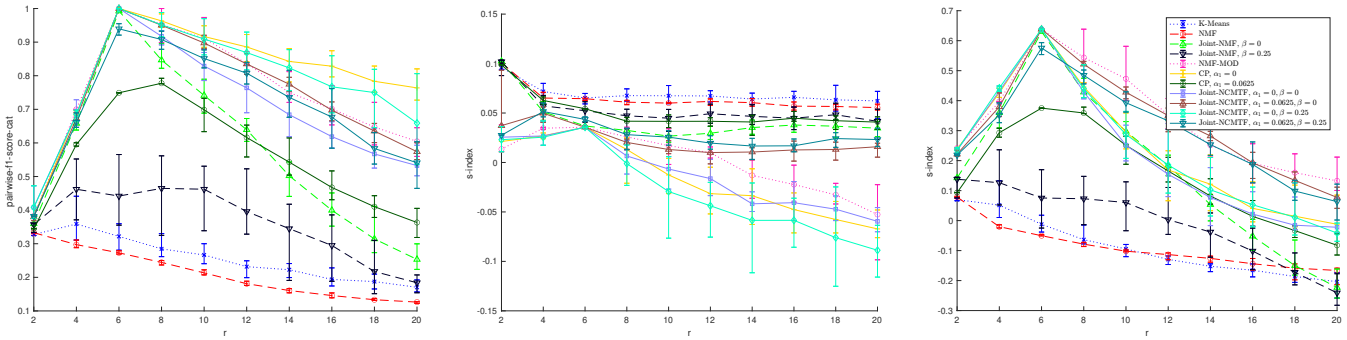
Several consistent trends between the surveyed methods were observed across the experiments as seen in, Figures 2 to 4. The tensor based methods consistently outperform the matrix based methods in terms of the external PWF1 metric, particularly in the case of overfactoring. This indicates that the tensor based methods perform significantly better at ground truth retrieval than the matrix based methods.

For the internal S-Index, we evaluated all methods on two different validation matrices respectively consisting of the tensor $\mathcal{X}$ collapsed along modes 3 and 2, i.e. $collapse(\mathcal{X}, 3) \in \mathbb{R}^{I_1 \times I_2}$ and $collapse(\mathcal{X}, 2) \in \mathbb{R}^{I_1 \times I_3}$. For the real-world Aminer and PatentsView datasets these resulted in $(documents \times words)$ and $(documents \times authors)$ validation matrices. Note that as the $(documents \times words)$ matrix is the $\mathcal{X}$ matrix used by the matrix based methods, we assume that the matrix based methods will yield "good" clusters with regards to internal metrics evaluated on this validation matrix. Conversely, the matrix based approaches, with the exception of NMF-MOD, do not explicitly incorporate the author information. As a result we expect the matrix based methods to yield comparatively worse clusters with regards to internal metrics evaluated on the $(documents \times authors)$ validation matrix. Figures 2b, 2c, 3b, 3c, 4d and 4e provide empirical evidence supporting these hypotheses.

We evaluate the impact of auto-coupling on the clustering results by comparing the tensor based methods with and without the auto-coupling parameter set to 0. In all experiments, we auto-coupled the $collapse(\mathcal{X}, 3)$ matrix to the tensor based formulations. We generally observed that auto-coupling caused the tensor based methods to perform more similarly to the matrix based methods.

D. Qualitative Results

Table II shows sample topics yielded by NMF and Joint-NCMTF on ['mechanical engineer', 'economy', 'politic'] Aminer subset for $r = 16$. Based upon the top documents and words, both topics clearly pertain to wall climbing robots and presumably contain documents from the 'mechanical engineer' keyword. Table III shows sample topics yielded by NMF and Joint-NCMTF on ['Transportation','Electrical Devices','Electrical Lighting','Information Storage','Motors & Engines + Parts','Computer Hardware & Software'] subcategory PatentsView dataset for r = 12 with no assignee information. The top documents and vocabulary indicate that the

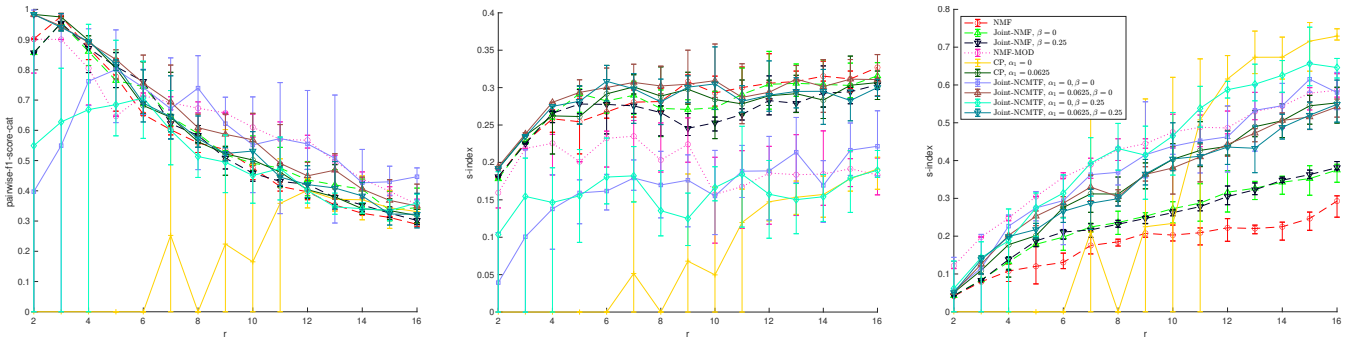


(a) PWF1 for ground truth cluster assignments

(b) S-Index on $collapse(\mathcal{X}, 3)$

(c) S-Index on $collapse(\mathcal{X}, 2)$

Fig. 2: External and Internal metrics for all points in synthetic dataset with sym-noise=.05, sym-rm-noise=.9, perturb-block-noise=0, perturb-dense-noise=.5, and sparsify-noise=.9. True r-value is $r = 6$.



(a) PWF1 for ground truth keywords

(b) S-Index on $collapse(\mathcal{X}, 3)$, i.e. $documents \times words$

(c) S-Index on $collapse(\mathcal{X}, 2)$, i.e. $documents \times authors$

Fig. 3: External and internal metrics for top 25% of data items of [‘mechanical engineer’, ‘economy’, ‘politic’] Aminer subset.

Documents
A wall-climbing robot without any active suction mechanisms
A miniature ceiling walking robot with flat tacky elastomeric footpads
Development of Wall Climbing Robot System by Using Impeller Type Adhesion Mechanism
Geckobot: a Gecko Inspired Climbing Robot using Elastomer Adhesives
Mobility of an in-pipe robot with screw drive mechanism inside curved pipes

Words	**Authors**
robot	- - - -
mechanism	- - - -
wall	- - - -
climbing	- - - -
control	- - - -

(a) NMF

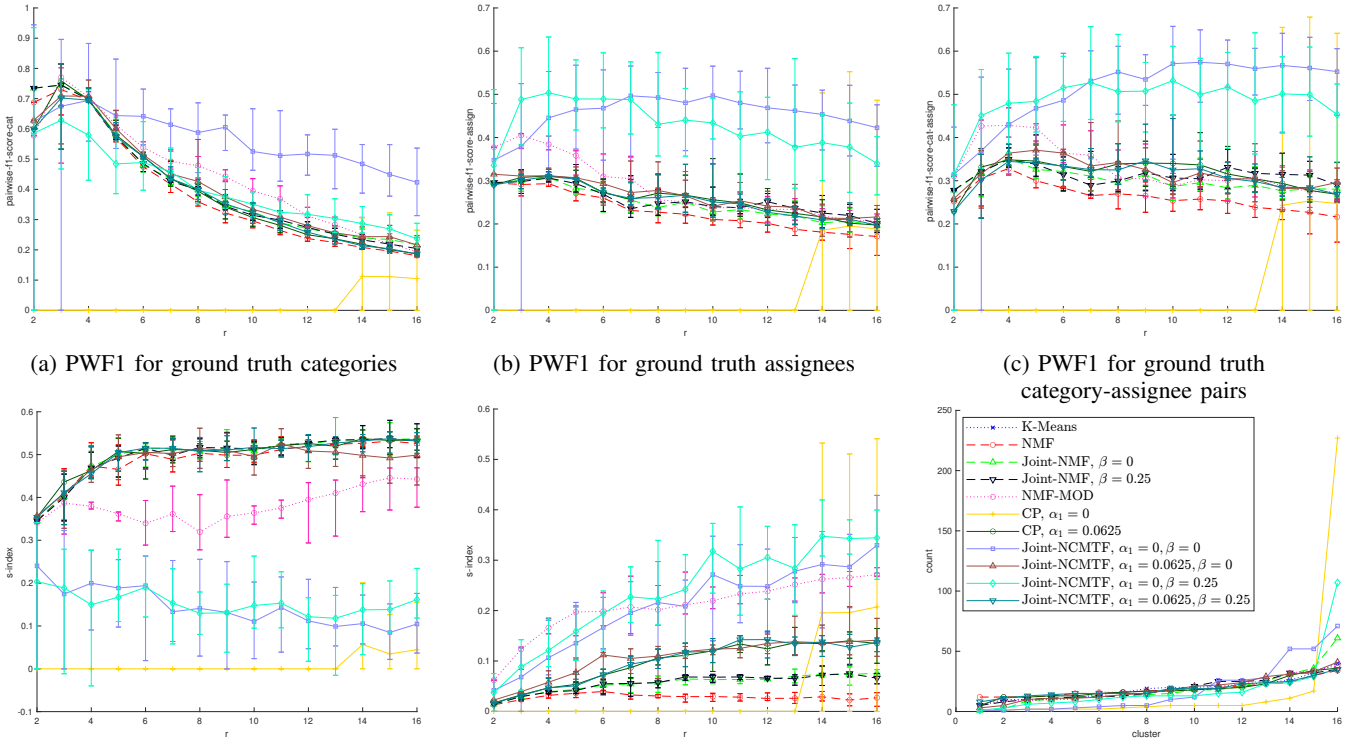
Documents
A miniature ceiling walking robot with flat tacky elastomeric footpads
Tankbot: a miniature, peeling based climber on rough and smooth surfaces
Geckobot: a Gecko Inspired Climbing Robot using Elastomer Adhesives
Waalbot: An Agile Small-Scale Wall Climbing Robot Utilizing Pressure Sensitive Adhesives
Rotating Magnetic Miniature Swimming Robots With Multiple Flexible Flagella

Words	**Authors**
robot	Metin Sitti
surfaces	Ozgur Unver
climbing	Shugen Ma
smooth	Houxiang Zhang
design	Guanghua Zong

(b) Joint-NCMTF

TABLE II: Sample topics yielded by NMF and Joint-CMTF on [‘mechanical engineer’, ‘economy’, ‘politic’] Aminer dataset for $r = 16$. For this topic, we manually verified that all top authors were mechanical engineers and/or roboticists.

cluster corresponding to this topic pertains to LED lighting and presumably belongs to the ‘Electrical Lighting’ subcategory. Each of the topics found by both NMF and Joint-CMTF were highly interpretable and appeared to belong to one of the 6 subcategories used to form the dataset. For both datasets we manually verified that all top authors were associated with the inferred topic. NMF and Joint-CMTF yielded similar topics for all sampled datasets. Joint-NCMTF provides an additional dimension of insight to the topic relative to NMF by finding top authors in addition to top documents and words. This additional dimension of information is a benefit of the tensor based formulation.



(a) PWF1 for ground truth categories

(b) PWF1 for ground truth assignees

(c) PWF1 for ground truth category-assignee pairs

(d) S-Index on $collapse(\mathcal{X}, 3)$, i.e. $documents \times words$

(e) S-Index on $collapse(\mathcal{X}, 2)$, i.e. $documents \times authors$

(f) Cluster sizes for all data items

Fig. 4: External and internal metric scores for top 25% of PatentsView subset data items (Figures 4a to 4e) along with the cluster sizes for all data items (Figure 4f). Patentsview subset formed from top 25 patents in terms of citation relationships for (Transportation) and (Motors & Engines + Parts) categories selected from 6 largest automobile producers. The proposed Joint-NCMTF without auto-coupling performs best of all surveyed methods for detecting ground truth cluster assignments as seen in Figures 4a to 4c whilst Joint-NCMTF with auto-coupling performs more similarly to the matrix based methods. The tensor based methods without auto-coupling yield a disproportionately size cluster as seen in Figure 4f. This is caused by the extreme sparsity of the tensor formulation resulting in factor matrix rows consisting of all zeros. We by default assign the data item corresponding to a zero row to the first cluster. Note that auto-coupling alleviates this thereby stabilizing the Joint-NCMTF results.

Documents
LED light with thermoelectric generator
Systems and methods for generating and modulating illumination conditions
Continuity maintaining biasing member
Multicolored LED lighting method and apparatus
Methods and apparatus for providing power to lighting devices

Words	**Authors**
light	- - - -
housing	- - - -
fluorescent	- - - -
ledbased	- - - -
lighting	- - - -

(a) NMF

Documents
Systems and methods for generating and modulating illumination conditions
Electric shock resistant L.E.D. based light
Methods and apparatus for controlling devices in a networked lighting system
LED-based light having rapidly oscillating LEDs
LED lighting apparatus with swivel connection

Words	**Authors**
light	Trevor Ehret
housing	Craig Mackiewicz
fluorescent	Christopher P. Natoli
ledbased	John Ivey
fixture	David L. Simon

(b) Joint-NCMTF

TABLE III: Sample topics yielded by NMF and Joint-NCMTF on ['Transportation','Electrical Devices','Electrical Lighting','Information Storage','Motors & Engines + Parts','Computer Hardware & Software'] PatentsView subset dataset for $r = 12$ with no assignee information and no stemming. Note that both topics appear to pertain to LED lighting. NMF and Joint-NCMTF yielded similar topics. Joint-NCMTF provides an additional dimension of insight to the topic relative to NMF by finding top authors in addition to top documents and words. We manually verified that for this topic all top authors were associated with electrical and/or lighting based patents.

VI. Discussion and Future Work

Joint-NCMTF is a powerful data analysis tool capable of handling datasets posed as coupled matrices and tensors whilst preserving the desirable clustering and topic modeling properties of NMF. Avenues of future research include deriving gradient methods, analyzing trade-offs of tensor formulations, implementing distributed variants, and providing rigorous justification of default hyperparameters for Joint-NCMTF.

Acknowledgments

This work was supported by the National Science Foundation (Grant No. OAC-2106738). We would like to thank Gudrun Cobb and Cassey Waldo for their extensive proofreading and editing as well as the anonymous reviewers for their invaluable comments which helped improve this work.

References

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan Yousef. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131, 10 2022.

[2] Evrim Acar, Tamara G. Kolda, and Daniel M. Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations, 2011.

[3] Miguel Ramos de Araujo, Pedro Manuel Pinto Ribeiro, and Christos Faloutsos. Tensorcast: Forecasting with context using coupled tensors (best paper award). In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 71–80, 2017.

[4] Sanaz Bahargam and Evangelos E. Papalexakis. Constrained coupled matrix-tensor factorization and its application in pattern and topic detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 91–94, 2018.

[5] Thirunavukarasu Balasubramaniam, Richi Nayak, Khanh Luong, and Md Abul Bashar. Identifying covid-19 misinformation tweets and learning their spatio-temporal topic dynamics using nonnegative coupled matrix tensor factorization. *Social Network Analysis and Mining*, 11, 12 2021.

[6] Arindam Banerjee, Sugato Basu, and Srujana Merugu. Multi-way clustering on relation graphs. In *SDM*, 2007.

[7] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S. Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 509–514, New York, NY, USA, 2004. Association for Computing Machinery.

[8] Arindam Banerjee, Chase Krumpelman, Joydeep Ghosh, Sugato Basu, and Raymond J. Mooney. Model-based overlapping clustering. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 532–537, New York, NY, USA, 2005. Association for Computing Machinery.

[9] Alex Beutel, Partha Pratim Talukdar, Abhimanu Kumar, Christos Faloutsos, Evangelos E. Papalexakis, and Eric P. Xing. *F¡sc¿lexi¡/sc¿F¡sc¿a¡/sc¿CT: Scalable Flexible Factorization of Coupled Tensors on Hadoop*, pages 109–117.

[10] Tamara G. Kolda Brett W. Bader et al. Tensor toolbox for matlab.

[11] Dongjin Choi, Jun-Gi Jang, and U Kang. S3cmtf: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization. *PLOS ONE*, 14(6):1–20, 06 2019.

[12] Rundong Du, Barry Drake, and Haesun Park. Hybrid clustering based on content and connection structure using joint nonnegative matrix factorization. *J. of Global Optimization*, 74(4):861–877, aug 2019.

[13] Srinivas Eswar, Benjamin Cobb, Koby Hayashi, Ramakrishnan Kannan, Grey Ballard, Richard Vuduc, and Haesun Park. Distributed-memory parallel jointnmf. In *Proceedings of the 37th International Conference on Supercomputing*, ICS '23, page 301–312, New York, NY, USA, 2023. Association for Computing Machinery.

[14] Shashank Gupta, Raghuveer Thirukovalluru, Manjira Sinha, and Sandya Mannarswamy. Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 278–281, 2018.

[15] Jette Henderson, Joyce C. Ho, Abel N. Kho, Joshua C. Denny, Bradley A. Malin, Jimeng Sun, and Joydeep Ghosh. Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 214–223, 2017.

[16] The MathWorks Inc. Matlab (r2024a), 2024.

[17] Stefanie Jegelka, Suvrit Sra, and Arindam Banerjee. Approximation algorithms for tensor clustering. In *International Conference on Algorithmic Learning Theory*, 2009.

[18] ByungSoo Jeon, Inah Jeon, Lee Sael, and U Kang. Scout: Scalable coupled matrix-tensor factorization - algorithm and discoveries. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 811–822, 2016.

[19] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *J. of Global Optimization*, 58(2):285–319, feb 2014.

[20] Jingu Kim and Haesun Park. Sparse nonnegative matrix factorization for clustering. 2008.

[21] Jingu Kim and Haesun Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM J. Sci. Comput.*, 33(6):3261–3281, nov 2011.

[22] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[23] Da Kuang, Chris Ding, and Haesun Park. *Symmetric Nonnegative Matrix Factorization for Graph Clustering*, pages 106–117.

[24] Chaoyue Liu and Mikhail Belkin. Clustering with bregman divergences: an asymptotic analysis. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[25] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, 2010.

[26] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[27] Andrew Mccallum, Andrés Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. pages 786–791, 01 2005.

[28] Evangelos E. Papalexakis, Nicholas D. Sidiropoulos, and Rasmus Bro. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE Transactions on Signal Processing*, 61(2):493–506, 2013.

[29] Namyong Park, Byungsoo Jeon, Jungwoo Lee, and U Kang. Bigtensor: Mining billion-scale tensor made easy. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 2457–2460, New York, NY, USA, 2016. Association for Computing Machinery.

[30] Amnon Shashua, Ron Zass, and Tamir Hazan. Multi-way clustering using super-symmetric non-negative tensor factorization. In *European Conference on Computer Vision*, 2006.

[31] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. pages 1105–1114, 04 2018.

[32] Ajit P. Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 650–658, New York, NY, USA, 2008. Association for Computing Machinery.

[33] Er Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 03 2001.

[34] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 990–998, New York, NY, USA, 2008. Association for Computing Machinery.

[35] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[36] Andrew Toole, Christina Jones, and Sarvothaman Madhavan. Patentsview: An open data platform to advance science and technology policy. *SSRN Electronic Journal*, 01 2021.

[37] Alaeddine Zahir, Khalide Jbilou, and Ahmed Ratnani. High-dimensional multi-view clustering methods, 2023.