Noname manuscript No.
(will be inserted by the editor)

Hybrid Clustering based on Content and Connection Structure using Joint Nonnegative Matrix Factorization

Rundong Du · **Barry Drake** · **Haesun Park**

Received: date / Accepted: date

Abstract A hybrid method called JointNMF is presented which is applied to latent information discovery from data sets that contain both text content and connection structure information. The new method jointly optimizes an integrated objective function, which is a combination of two components: the Nonnegative Matrix Factorization (NMF) objective function for handling text content and the Symmetric NMF (SymNMF) objective function for handling network structure information. An effective algorithm for the joint NMF objective function is proposed so that the efficient method of block coordinate descent (BCD) framework can be utilized. The proposed hybrid method simultaneously discovers content associations and related latent connections without any need for postprocessing of additional clustering. It is shown that the proposed method can also be applied when the text content is associated with hypergraph edges. An additional capability of the JointNMF is prediction of unknown network information which is illustrated using several real world problems such as citation recommendations of papers and leader detection in organizations. The proposed method can also be applied to general data expressed with both feature space vectors and pairwise similarities and can be extended to the case with multiple feature spaces or multiple similarity measures.

This work was supported in part by the National Science Foundation (NSF) grant IIS-1348152 and Defense Advanced Research Projects Agency (DARPA) XDATA program grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or DARPA.

R. Du
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA
E-mail: rdu@gatech.edu

B. Drake
Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, GA 30318, USA
E-mail: barry.drake@gtri.gatech.edu

H. Park
School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0765, USA
E-mail: hpark@cc.gatech.edu

Our experimental results illustrate multiple advantages of the proposed hybrid method when both content and connection structure information is available in the data for obtaining higher quality clustering results and discovery of new information such as unknown link prediction.

Keywords Joint nonnegative matrix factorization · Symmetric NMF · constrained low rank approximation · content clustering · graph clustering · hybrid content and connection structure analysis

1 Introduction

Constrained low rank approximation (CLRA) such as Nonnegative matrix factorization (NMF) has played an important role in data analytics, providing a foundational framework for formulating key analytics tasks such as text clustering, graph clustering, and recommendation system [15–17, 14] problems. In this paper, we propose a joint NMF algorithm which jointly optimizes the standard NMF for content clustering and Symmetric NMF (SymNMF) for graph clustering. Detailed discussions of NMF and SymNMF can be found in [12, 13] and [17], respectively. The goal is to simultaneously cluster data sets that contain both content and connection structure, utilizing both information sources, to obtain higher quality clustering results. This type of fusion can be done at the data level (early fusion) or at the result level (late fusion). An advantage of NMF and SymNMF is that both are formulated using one framework of CLRA, and therefore, we can naturally design a joint objective function to obtain the objective function level fusion as we illustrate in a later section.

Numerous data sets contain both text content and connection structure. For example, in a data set of research papers or patents, papers or patents have text content where the citations or co-author relationships define the connection structure; in a data set of emails, email messages have text content and the sender-recipient relations define a hypergraph structure where one email may have multiple recipients. When the connection structure is represented as edges in a graph, in the former case the text content is associated with graph nodes while in the latter case the text content is associated with hypergraph edges. A hybrid clustering method is designed to utilize both content and connection structure information, thus taking advantage of the full information provided in the data.

Many methodologies exist for data clustering. However, our framework using CLRA offers multiple advantages. The proposed method is simple to implement based on an existing numerical routines to solve a nonnegativitiy constrained least squares (NLS) and widely applicable. Also the proposed method can provide valuable insights about the data when there is not enough knowledge about the underlying data model or when one desires only a quick glance at results. In fact, in the area of text and graph clustering, CLRA methods (NMF and SymNMF) have been demonstrated to have superior performance in terms of speed and accuracy [15–17]; The two CLRA based methods, NMF for content clustering and SymNMF for graph clustering, have the same underlying matrix factorization framework, and, can be merged at the objective function level and the result is easily interpretable as clustering result without requiring an additional step for clustering unlike in many of the spectral methods.

In this paper we discuss data with associated text content and connection structure. In addition to text content, other types of information may also be associated with connection structure, such as images and attributes that appear in structured data like a person's age and gender, etc. Our hybrid clustering method can naturally extend to other content information as long as the raw data can be encoded as nonnegative vectors.

This paper is organized as follows: We start with the basic situation where the text content is associated with connection structure, i.e.,graph nodes, and extend the idea to the case where a hypergraph is the correct connection representation and the text content is associated with hypergraph edges (Section 2). We then summarize some related work in Section 3. We have conducted extensive experiments using patent citation data sets and two other types of data sets to show the effectiveness of our method (Section 4). In addition to demonstrating improvements of clustering quality, we list several potential applications of our hybrid clustering approach, including citation recommendation on a paper data set and the application of our hypergraph extension to an Enron email data set (Section 5). Discussions and conclusions can be found in Section 6.

2 Hybrid Clustering via Joint NMF

We have designed fast, scalable algorithms for some variants of NMF for key data analytics problems [13,15,4]. Currently one of the fastest algorithms for hierarchical and flat (non-hierarchical) topic modeling and clustering that also produce consistently high quality solutions are HierNMF2 and FlatNMF2, which are available in our open source software package in C++ called SmallK (`http://smallk.github.io/`). SmallK also includes Python drivers, pysmallk, that allow seamless integration of SmallK into existing Python applications.

First we assume that the text content is associated with the graph nodes (e.g. paper/patents with citations). We assume that a data set's text information is represented in a nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$ and the graph structure is represented in a nonnegative symmetric matrix $S \in \mathbb{R}_+^{n\times n}$, where m is the number of features, the columns of X represent the n data items, the (i, j)-th element of S represents a relationship such as similarity between the i-th and j-th data items, and $\mathbb{R}_+$ denotes the real nonnegative numbers. Then the NMF formulation for text clustering/topic modeling [14] is

$$\min_{W\geq 0, H\geq 0} \|X - WH\|_F \tag{1}$$

and the SymNMF formulation for graph clustering [16,17] is

$$\min_{H\geq 0} \|S - H^T H\|_F \tag{2}$$

where $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$, and a given integer k, which is typically much smaller than m or n, represents the reduced dimension, i.e., number of clusters [12]. In (1), each column of W, subject to some scaling, is regarded as the representative of each cluster or a topic in the document collection. The matrix H can be seen as a low rank

(rank k) representation of the data points since each data item in X can be explained by an additive linear combination of the representative columns in W, i.e., the columns of H are approximative coordinates of data items in X with columns of W as basis vectors. Similarly, in (2), H is a low rank representation of the nodes in the graph. Such a low rank approximation also gives us k clusters, since $H_{i,j}$ can be seen as a measurement of strength that the j-th data item belongs to the i-th cluster. Therefore, each column of H gives the soft clustering assignment information. By taking the row index with the maximum value in each column vector of H as the cluster index of each data item, one can also perform hard clustering [12,13].

The hybrid clustering method we propose finds a low rank representation that simultaneously represents the text content and the graph structure of the data items by jointly optimizing the combined NMF and SymNMF objective functions:

$$\min_{W\geq 0,H\geq 0} \alpha_1||X-WH||_F^2+\alpha_2||S-H^TH||_F^2. \tag{3}$$

where $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ are the weighting parameters. By adjusting the parameters α_i, we can emphasize one over the other. In the extreme case, some α_i can be set to zero: e.g. when $\alpha_2 = 0$ in the above, we are only concerned with the content, when $\alpha_1 = 0$, we only pay attention to the structural information and ignore the content. Excluding these special cases, we can assume $\alpha_1 = 1$ without loss of generality and Eqn. (3) becomes

$$\min_{W\geq 0,H\geq 0} ||X-WH||_F^2+\alpha||S-H^TH||_F^2. \tag{4}$$

with $\alpha \geq 0$ as the weighting parameter.

Now we extend our method to hypergraphs where the text content is associated with hypergraph nodes. Once this is done, it would be natural to extend our method further to the cases where text is associated with graph or hypergraph edges due to the duality that exists between edges and nodes of a hypergraph and the fact that a graph can be treated as a special case of a hypergraph.

A hypergraph $\mathscr{H}$ is a pair $\mathscr{H} = (\mathscr{V},\mathscr{E})$, where $\mathscr{V} = \{v_1,\ldots,v_m\}$ is the set of vertices and $\mathscr{E} = \{e_1,\ldots,e_n : e_i \subset \mathscr{V}\}$ is the set of hyperedges. Unlike a graph edge, a hypergraph edge e_i may connect more than two vertices in the graph. Such a hypergraph $\mathscr{H}$ can be represented by an incidence matrix $M = (m_{ij}) \in \mathbb{R}^{m\times n}$, where

$$m_{ij} = \begin{cases} 1, & v_i \in e_j; \\ 0, & \text{otherwise.} \end{cases}$$

The dual hypergraph $\mathscr{H}^*$ is the hypergraph corresponding to the incidence matrix M^T.

Assume there's a k-way partition of the vertices $(\mathscr{V}_1,\ldots,\mathscr{V}_k)$ where $\mathscr{V}_1 \cup \cdots \cup \mathscr{V}_k = \mathscr{V}$ and $\mathscr{V}_i \cap \mathscr{V}_j = \emptyset$ for all $1 \leq i \neq j \leq k$. Define the matrix $H = (h_{ij}) \in \mathbb{R}^{k\times n}$ as

$$h_{ij} = \frac{[v_j \in \mathscr{V}_i]}{\sqrt{d_v(j)}\left(\sum_{v_l\in\mathscr{V}_i}\frac{1}{d_v(l)}\right)^{1/2}} \tag{5}$$

which is a normalized partition indicator matrix where

$$[v_j \in \mathcal{V}_i] = \begin{cases} 1, & v_j \in \mathcal{V}_i; \\ 0, & \text{otherwise.} \end{cases}$$

and $d_v(l) = \sum_{j=1}^{n} m_{lj}$ is the degree of vertex v_l. It is shown in [33] that the following optimization problem

$$\max_{H} \operatorname{tr} HSH^T \tag{6}$$

is equivalent to minimizing the hypergraph normalized cut as defined in [33], where

$$S = D_v^{-1/2} M D_e^{-1} M^T D_v^{-1/2} \tag{7}$$

is symmetric, $D_v = \operatorname{diag}(d_v(1), \ldots, d_v(m))$, $D_e = \operatorname{diag}(d_e(1), \ldots, d_e(n))$, and $d_e(l) = \sum_{i=1}^{m} m_{il}$ is the degree of edge e_l. Following the same argument as in [16], it can be shown that (6) is equivalent to $\min_H \|S - H^T H\|_F^2$ and by relaxing constraint (5) to $H \geq 0$, we obtain the objective function of SymNMF. Therefore, in the case of a hypergraph, we can use the matrix S defined in Eqn. (7) as the similarity matrix in Eqn. (4).

There are many ways to find a solution for the objective function (4). Theoretically, a Newton-like algorithm can be developed to directly solve (4). However, as pointed out in [17], a Newton-like algorithm can not utilize the sparsity of X and S for speeding up because the matrices $X - WH$ and $S - H^T H$ need to be computed explicitly and thus the sparsity will be destroyed. On the other hand, an alternating nonnegative least square (ANLS) algorithm can be sped up with sparsity. To apply an ANLS-like algorithm that can utilize the sparse nature of text documents and associated networks, we propose reformulating (4) in the following form with a penalty term

$$\min_{W,H,\tilde{H} \geq 0} ||X - WH||_F^2 + \alpha ||S - \tilde{H}^T H||_F^2 + \beta \|\tilde{H} - H\|_F^2. \tag{8}$$

where $\tilde{H} \in \mathbb{R}_+^{k \times n}$ and $\beta \geq 0$ is the regularization parameter. This reformulation is motivated from our earlier work to generate an algorithm that is based on the block coordinate descent (BCD) scheme so that each sub-problem in the BCD is a nonnegativity constrained least squares (NLS) problem for which we have developed a highly efficient algorithm and optimized open-source software [2]. Then Eqn. (8) can be solved using a 3-block coordinate descent (BCD) scheme, i.e. minimize the objective function with respect to W, $\tilde{H}$ and H in turn. Specifically, we solve the following three subproblems in turn:

$$\min_{W \geq 0} \|H^T W^T - X^T\|_F^2 \tag{9}$$

$$\min_{\tilde{H} \geq 0} \left\| \begin{bmatrix} \sqrt{\alpha} H^T \\ \sqrt{\beta} I_k \end{bmatrix} \tilde{H} - \begin{bmatrix} \sqrt{\alpha} S \\ \sqrt{\beta} H \end{bmatrix} \right\|_F^2 \tag{10}$$

$$\min_{H \geq 0} \left\| \begin{bmatrix} W \\ \sqrt{\alpha} \tilde{H}^T \\ \sqrt{\beta} I_k \end{bmatrix} H - \begin{bmatrix} X \\ \sqrt{\alpha} S \\ \sqrt{\beta} \tilde{H} \end{bmatrix} \right\|_F^2 \tag{11}$$

where each subproblem is simply a nonnegative least squares problem (NLS), which is convex. Thus, an active-set-based algorithm can find the optimal solution in a finite number of operations and ensures that the solution is in the feasible region. This avoids the case of nearly linear dependent vectors, which has profound implications for real-world applications such as chemical detection where false negatives and false positives can increase dramatically in the presence of rank deficiency [7]. The above three block BCD algorithm converges to a stationary point according to Bertsekas' theorem [1]. The identity submatrices I_k in the above equations make the problem better conditioned than the subproblems in the standard NMF that uses two block BCD alternating updating W and H. We solve each NLS problem using the block principal pivoting (BPP) algorithm [13]. Theoretically, to force H to be identical to $\tilde{H}$, the value of the parameter β has to be infinity. This problem has been studied extensively and we use a scheme similar to that proposed in [30]. It should be pointed out that in [13] it is shown that algorithms based on the BCD framework have guaranteed convergence to a stationary point, whereas, popular and easy to implement algorithms such as Multiplicative Updating (MU) may not converge. In addition, extensive experiments show that the BPP method is faster and more accurate than MU.

3 Related Work

The use of joint matrix factorization for clustering can also be seen in [19,27,11], all of which consider clustering using information from different sources. [27] is also a method for hybrid clustering of connection structure and content data. Their formulation (2JointMF) is $\min_{W_0,W_1,H} \|A - W_0H\|_F^2 + \|X - W_1H\|_F^2 + \lambda \sum_{j=1}^{n} \|H(:,j)\|_1^2 + \eta(\|W_0\|_F^2 + \|W_1\|_F^2)$, with the constraints $H \geq 0$ and $0 \leq W_0H \leq 1$, where A is the adjacency matrix of the graph (can be an asymmetric matrix representing a directed graph) and X is the feature-data matrix. Our JointNMF is different from 2JointMF in the following ways: (1) JointNMF has nonnegative constraints on all matrix factors while 2JointMF has nonnegative constraint on H only. The nonnegative constraints on all factors usually lead to better interpretations of the result. For example, the W factor in our formulation can be interpreted as topic vectors due to its nonnegativity. (2) 2JointMF factors the graph matrix in a way similar to factoring a general feature matrix, while the symmetric factorization from JointNMF acknowledges the symmetric similarity relation encoded in a graph and also relates itself with minimizing normalized cut. (3) The constraint $0 \leq W_0H \leq 1$ makes 2JointMF computationally much harder. The algorithm in [19] also jointly minimizes several NMF objectives. However, they do not consider graph information and therefore SymNMF was not in their formulation. The objective function in [11] looks the same as ours, however the matrices in the formulas have different meanings and their formulation is only used in the context of graph clustering.

Some other methods for hybrid clustering (of graph and node content) can be summarized as the following categories: (1) Generative models [5,9,3,10,20,23]. These algorithms learn a latent cluster indicator for each node, based on which all the content and links are generated. Such latent cluster indicator could be a vector that measures how likely a node belongs to each cluster (for soft clustering), or a single variable that

assigns a node to a specific cluster (hard clustering). (2) Discriminative models [32]. The authors of [32] argue that generative models fail to consider additional factors that could affect the community memberships and isolate the content that is irrelevant to community memberships. They propose a discriminative algorithm, PCL-DC, to overcome these two shortcomings. (3) Topic modeling with network regularization [22,26]. These methods start with the objective function of a topic modeling method and add the graph related part as a regularization term. (4) Augmenting the graph with content information [24]. This method reduces the hybrid clustering problem to a graph clustering problem by augmenting the graph with new vertices and edges that reflect the text content. (5) Entropy based [6]. This method jointly minimizes the entropy of document clusters and maximizes modularity of graph clusters. (6) cluster ensembles [25] This algorithm assembles the result of a graph clustering algorithm and the result of a document clustering algorithm into a combined result (late fusion). (7) Cluster selection [8]. This method selects a graph clustering algorithm when the graph has clear structure information and selects content only algorithms when the graph has ambiguous structure information.

4 Clustering US Patent, BlogCatalog and Flickr Data

All experiments were performed on a server with two Intel(R) Xeon(R) CPU E5-2680 v3 CPUs and 377GB memory.

The main data set used for the experiments is the US patent claim and citation data from PatentsView[1]. Some advantages of using US patents as a data source are: (1) the openness, centralized management and availability of relatively structured data format makes the patent data easier to obtain and process; (2) the abundance of the patent database ensures enough samples that can be studied; (3) patents were carefully assigned with classification labels, and such labels were examined by patent examiners; therefore the classification information can be used as a relatively reliable ground truth.

We use the Cooperative Patent Classification (CPC) system, where each classification label has the scheme illustrated in Figure 1. We select 13 CPC classes (A22,

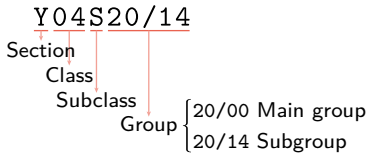

Fig. 1 An example classification label in the CPC scheme

A42, B06, B09, B68, C06, C13, C14, C40, D02, D10, F22, Y04) and use patents under each class to construct 13 different data sets. For each data set, we first construct the term-document matrix representing the patent claims and the graph adjacency

[1] http://www.patentsview.org

Table 1 Some statistics of US patent data sets.

Class	#Patents	#Citations	#Groups
A22	4976	28746	230
A42	4213	29285	134
B06	2938	11549	82
B09	3522	17302	38
B68	790	2433	93
C06	3347	17562	141
C13	1010	3717	87
C14	583	1125	69
C40	3748	28854	41
D02	3170	11216	158
D10	2548	8486	154
F22	3040	7977	359
Y04	3242	21518	76

Table 2 Some statistics of BlogCatalog and Flickr data sets.

Data	#Nodes	#Edges	#Tags	#Groud truth clusters
BlogCatalog	31228	782584	5387	60
Flickr	32576	2749800	77234	170

matrix representing the patent citation relations. Our algorithm requires a symmetric adjacency matrix and therefore we treat the citation graph as undirected by ignoring the directions. We then clean the data by removing terms that appear very infrequently and documents that are too short or duplicated, and extract the largest connected components of the graph. Finally, we apply tf-idf to the term-document matrix, normalize its columns to have unit 2-norm, obtaining the matrix X, and let S be $D^{-1/2}AD^{-1/2}$, where $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix, $D = \mathrm{diag}(d_1,\ldots,d_n)$ and $d_i = \sum_{j=1}^{n} A_{ij}$ is the degree of vertex i. We use CPC groups as ground truth clusters. Some statistics about these data sets (after cleaning) are listed in Table 1.

To verify our algorithm on other types of data, we also use the BlogCatalog data set from [28] and the Flickr data set from [29]. These data sets have users as graph nodes and represent user commenting and friendship relations as graph edges. The content comes from user generated keywords/tags that are used to describe their blog articles (BlogCatalog) or photos (Flickr), which is different from traditional text content. The ground truth clusters of BlogCatalog data set are defined by categories of each blog and the ones for the Flickr data set are defined by user groups. We apply the same preprocessing as for the US patent data sets. Some statistics regarding these two data sets (after preprocessing) are listed in Table 2.

We now define the measures for the evaluation of the clustering results. Assume we computed k clusters $B_1,\ldots,B_k$ and the ground truth has k' clusters $G_1,\ldots,G_{k'}$. We compute the confusion matrix $C = (c_{ij})_{k\times k'}$, where $c_{ij} = |A_i \cap B_j|$. Then we define the *average F_1 score* [31] as

$$F_1 = \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k}\max_j F_1(A_i,B_j) + \frac{1}{k'}\sum_{j=1}^{k'}\max_i F_1(B_j,A_i)\right)$$

Table 3 Type of predictions

In prediction	In ground truth	Type
c-connected	c-connected	TP
c-disconnected	c-disconnected	TN
c-connected	c-disconnected	FP
c-disconnected	c-connected	FN

where

$$F_1(A_i,B_j) = F_1(B_j,A_i) = \frac{2c_{ij}}{|A_i|+|B_j|}$$

This score measures how well an algorithm can recover the ground truth clusters. We also use another measure called *rand index* [21], which measures how well an algorithm can predict the connections among data items. Assume there are n data items in total. For each of the $n(n-1)/2$ pairs of data items, we say the two items are *c-connected* if they belong to the same cluster, otherwise we call them *c-disconnected* (prefix c is added to distinguish from connectivity in graph theory). Clustering results can also be treated as a prediction of c-connectivity of each pair of data items. A prediction regarding one pair of data items can have four cases of true positive (TP), true negative (TN), false positive (FP) or false negative (FN) according to the rules listed in Table 3. Then the rand index can be defined as

$$RI = \frac{\#TP+\#TN}{\#TP+\#TN+\#FN+\#FP} = \frac{\#TP+\#TN}{n(n-1)/2}$$

We compare our algorithm with NMF and SymNMF, which have leading performance in text clustering and graph clustering, respectively. For hybrid clustering, we choose PCL-DC [32] to compare with based on its popularity and source code availability. Although we mentioned many other algorithms in Section 3, we found that for other algorithms, either the code is not available or the code is available but we encountered runtime errors during experimental tests. Both JointNMF and PCL-DC have parameters to set. For JointNMF, we let the default parameter be $\alpha = \|X\|_F^2/\|S\|_F^2$, meaning half-half balance between graph clustering and text clustering, and set $\beta = \alpha\|S\|_{max}$, where $\|S\|_{max}$ is the maximum absolute value of elements in S. The authors of PCL-DC do not provide a method to specify its regularization parameter λ. Therefore, it is important to first study how the parameter change will affect the algorithm performance. It is found that for $\lambda < 1$, PCL-DC sometimes becomes extremely slow, such that it may take weeks to run over all the data sets (estimated based on sampling run). Therefore, λ is varied within $[1,20]$. In Figure 2, we show how the average F1 score changes when λ varies in that range for the first four data sets listed in Table 1. The code of PCL-DC[2] provides two models (popularity link model and productivity link model), which we label as PCL-DC-1 and PCL-DC-2, respectively. The performance change of JointNMF when its parameter α varies in the same range is also studied. We observe that the PCL-DC is either worse than JointNMF or very sensitive to the parameters, and it is concluded that when λ exceeds

[2] `https://homepage.cs.uiowa.edu/~tyng/codes/community_detection.zip`

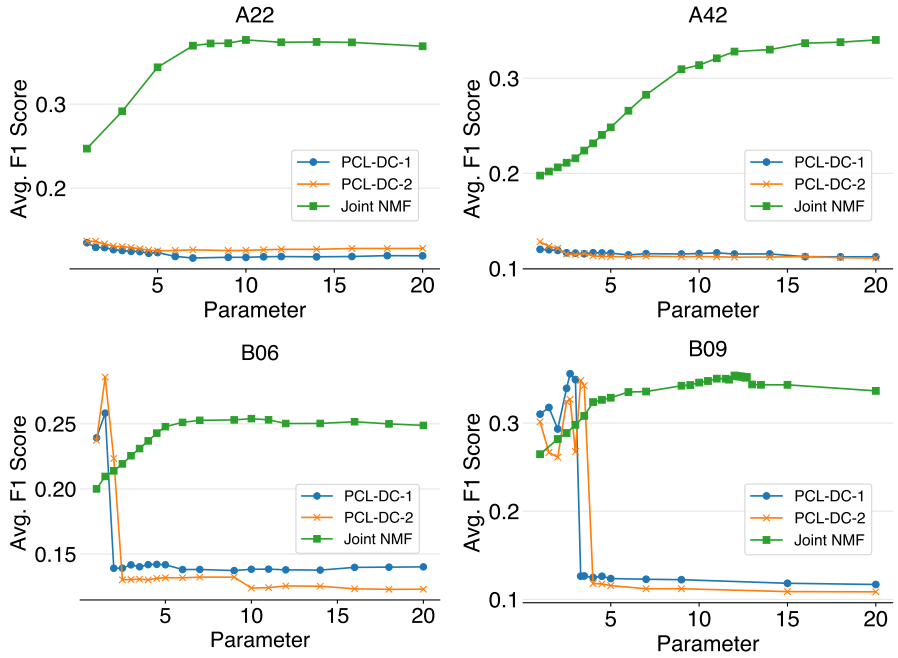

Fig. 2 Parameter sensitivity of PCL-DC and JointNMF. The parameter of PCL-DC is λ and the parameter of JointNMF is α.

Table 4 Comparison of average F1 scores

Class	JointNMF	NMF	SymNMF	PCL-DC-1	PCL-DC-2
A22	**0.3730**	0.2293	0.3457	0.1351	0.1369
A42	**0.3215**	0.1779	0.3199	0.1201	0.1280
B06	**0.2502**	0.1905	0.2307	0.2393	0.2373
B09	**0.3336**	0.2449	0.2690	0.3101	0.3014
B68	0.3806	0.3044	0.3730	**0.4034**	0.3671
C06	**0.2257**	0.1830	0.2004	0.1156	0.1158
C13	**0.2990**	0.2664	0.2953	0.2616	0.2224
C14	0.3584	0.3232	**0.3603**	0.2692	0.2659
C40	0.1939	0.1709	0.1673	0.1951	**0.1981**
D02	**0.2990**	0.2131	0.2683	0.1756	0.2268
D10	**0.3046**	0.2452	0.2783	0.1612	0.2999
F22	**0.3006**	0.2211	0.2926	0.1533	0.1388
Y04	0.2489	0.2029	0.2019	**0.2599**	0.2596
blogcatalog	0.2038	0.2150	0.0750	**0.2754**	0.2754
flickr	0.1545	0.0748	**0.1660**	0.0855	0.0855

a certain threshold (depending on the data), there is a large drop in clustering quality. Therefore, to have a tolerable run time while having a fair clustering quality, $\lambda = 1$ is chosen for the comparison experiments. The results of the comparison are listed in Table 4 to Table 6, where each value is the average over 10 runs.

Table 5 Comparison of rand index

Class	JointNMF	NMF	SymNMF	PCL-DC-1	PCL-DC-2
A22	**0.9785**	0.9768	0.9772	0.9274	0.9489
A42	**0.9650**	0.9633	0.9647	0.9225	0.9318
B06	**0.9368**	0.9357	0.9024	0.8775	0.8815
B09	**0.8497**	0.8387	0.7600	0.8464	0.8333
B68	0.9496	0.9423	**0.9508**	0.9272	0.8897
C06	0.9175	0.9150	**0.9182**	0.8969	0.8967
C13	0.8918	0.8873	**0.8927**	0.8598	0.8485
C14	**0.9086**	0.9036	0.9071	0.8233	0.7934
C40	0.6575	0.6507	**0.6820**	0.6593	0.6692
D02	**0.9612**	0.9594	0.9578	0.8922	0.8831
D10	**0.9080**	0.9048	0.9075	0.8676	0.8771
F22	0.9811	0.9797	**0.9816**	0.9554	0.9549
Y04	**0.8879**	0.8853	0.8697	0.8668	0.8622
blogcatalog	0.7572	**0.7652**	0.6173	0.7259	0.7259
flickr	0.0560	0.0409	**0.0782**	0.0620	0.0620

Table 6 Comparison of running time (seconds)

Class	JointNMF	NMF	SymNMF	PCL-DC-1	PCL-DC-2
A22	769.4	304.4	219.2	**55.6**	57.5
A42	311.9	161.9	163.1	**24.3**	24.8
B06	193.8	115.8	**59.8**	444.5	1800.8
B09	145.6	109.6	**48.2**	406.6	588.8
B68	48.2	60.8	**7.6**	288.3	439.0
C06	489.8	269.0	160.6	21.1	**20.9**
C13	70.9	76.1	**8.8**	421.5	377.2
C14	29.5	25.0	**4.7**	220.6	83.8
C40	240.8	127.8	**54.3**	394.0	597.3
D02	534.5	238.5	**117.3**	1623.5	831.8
D10	280.8	155.4	95.9	**14.7**	1728.4
F22	1294.1	404.4	267.2	38.4	**36.7**
Y04	291.9	125.8	**103.8**	1568.3	987.6
blogcatalog	401.3	**222.8**	1515.6	4463.4	4522.4
flickr	12455.6	2437.9	3504.5	**1181.3**	1236.0

Using these patent data sets, from our experiments it can be observed that: (1) JointNMF usually has the best average F1 scores, and its average F1 score is almost always better than that of NMF or SymNMF alone; (2) JointNMF and SymNMF have the best rand index; (3) SymNMF is usually the fasted algorithm. On BlogCatalog and Flickr data sets, which have different kinds of content and graph edges, the performance varies depending on the data. However, the performance of JointNMF is comparable to the best method with the exception of run time on the Flickr data set. In conclusion, for patent data sets, based on content and citations, JointNMF produces better quality solutions for clustering; for prediction of pairwise connection, both JointNMF and SymNMF perform well; speed-wise, JointNMF is not the fastest, but is comparable to other methods. On other types of data, the performance of each method varies, and JointNMF generates comparable results. The JointNMF method has other advantages: its parameter has explicit meanings (weight between text and

graph), the clustering quality is not very sensitive to the parameter setting, and its default parameter works very well.

5 Other Applications

In this section we present additional applications of our JointNMF framework beyond clustering. We demonstrate our JointNMF on other potential applications such as citation recommendations of papers/patents and activity/leader detection in an organization.

5.1 Citation Recommendation

When applied to papers/patents with citations or web pages with hyperlinks, the formulation (4) can also be understood as finding a basis W for the text space, such that under this basis, the representation (coordinates) of the documents can also reflect their linkage information. Therefore, when we express a new vector $\boldsymbol{x}$ in the text space using the basis W, i.e. finding a vector $\boldsymbol{h}$ that solves the following optimization problem

$$\min_{\boldsymbol{h}\geq 0} \|\boldsymbol{x} - W\boldsymbol{h}\|_2 \tag{12}$$

We can use closeness of $\boldsymbol{h}$ to the column vectors in H to decide how likely the new document represented by $\boldsymbol{h}$ should cite some of the documents in H. For example, one can recommend a new document to cite the i-th original document if the i-th entry of $H^T\boldsymbol{h}$ is larger than certain threshold. Another method is to set the threshold for the cosine similarity between $\boldsymbol{h}$ and column vectors in H. It will be observed that each method has its advantages.

For this task, we use the paper title/abstract and citation data **cit-HepTh** from SNAP[18], which contains 27,770 papers from January 1993 to April 2003 in the hep-th (high energy physics - theory) section of arXiv. Note that this is a different task from clustering and therefore the data preprocessing procedure is a little different: the raw adjacency matrix for S (i.e. $S = A$) is used. The normalized version $D^{-1/2}AD^{-1/2}$ is related to minimizing the normalized cut [16] and therefore good for clustering. Here the raw adjacency matrix is a better indicator of citations, which is used as an input that the algorithm learns from, instead of a basis for clustering.

To evaluate our method, the data is separated into training and test sets by treating papers published earlier than 2003 as the training set and papers published in 2003 as the test set. We use JointNMF to learn a matrix W from the document and citation relations in the training set, and then make predictions of citations for documents in the test set and compare the predictions with the actual citations.

To verify that the W computed by our algorithm indeed reflects the network structure better, we also design several baseline methods. A naive method is to predict citations based on number of words shared by two documents. One method based on NMF is to learn the matrix W used in (12) only by NMF, i.e. $\min_{W\geq 0,H\geq 0} \|X_{train} - WH\|_F$. Another method based on NMF is to directly learn the $\boldsymbol{h}$ vector in (12) using

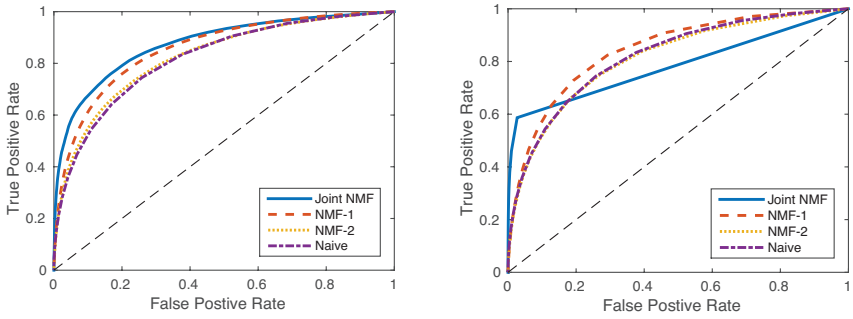

Fig. 3 ROC curves for citation recommendation algorithms applied to paper abstract and citation data. The left uses cosine similarity for the prediction, while the right uses inner product.

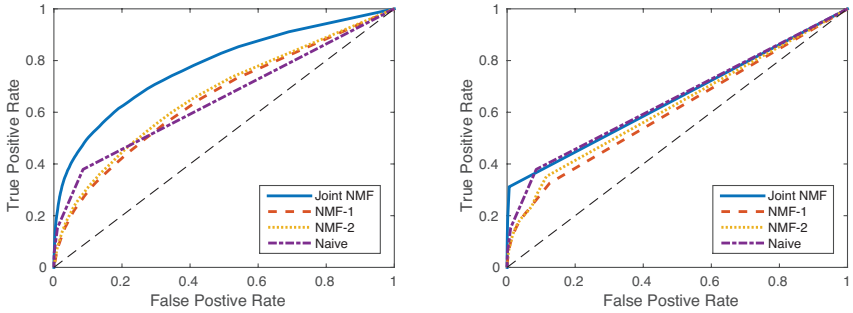

Fig. 4 ROC curves for citation recommendation algorithms applied to paper title and citation data. The left uses cosine similarity for the prediction, while the right uses inner product.

$\min_{W,H,\boldsymbol{h}\geq 0} \|[X_{train},\boldsymbol{x}] - W[H,\boldsymbol{h}]\|_F$. For the two NMF-based methods, the rest of the steps for making predictions are the same as JointNMF, once the matrix W or the vector $\boldsymbol{h}$ is obtained. In this subsection, we denote these two NMF based methods as NMF-1 and NMF-2, respectively.

For both prediction methods (compute $H^T\boldsymbol{h}$, the inner product , or compute cosine similarity scores), a threshold is needed. Instead of evaluating these algorithms with a fixed threshold, we show the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold values. In general, the closer the curve is to the upper left corner of the graph, the better the algorithm results. Or quantitatively, the larger the area under the curve (AUC) is, the better.

Paper abstracts are used to extract text content. The experimental results are shown in Figure 3. Some observations are: when cosine similarity is used, JointNMF makes the overall best predictions, and when inner product is used, at certain threshold values JointNMF can achieve relatively high true positive rate with a very low false positive rate. One can choose which method to use based on requirements.

The experiments are repeated using only paper titles as text content; similar results are observed, as shown in Figure 4. From the results we can observe that even with very little text information (such as paper titles), our method still works well.

Table 7 Frequency of number of memberships

#memberships	1	2	3	4	5	6	7	11
#employees	1069	149	45	17	8	7	1	1

5.2 Activity and Leader Detection from Enron Email Data

In an organization where various groups of people work on different subjects and engage in different activities, JointNMF can be used to detect such group structure, reveal the working subject/activities and find administrators/leaders in the organization. We assume that (1) within-group communications (e.g. emails) reflects the subject on which the team is working/activities engaged in and (2) people involved in multiple groups would likely hold a higher position in the organization, since they may be in charge of these groups. Each communication can be seen as a hypergraph edge that connects all people involved in the communication and the communication content is the text associated with the edge. Clustering the text data can distinguish and identify different working subjects/activities and clustering the graph data can divide people into workgroups. JointNMF utilizes both types of data simultaneously and therefore can distinguish different groups of people working on the same subject and different subjects worked on by the same group of people. After clustering, one can count and compare the number of groups/clusters each person belongs to—the more groups a person belongs to, the more likely the person is in a leadership or administrative position.

A subset of Enron email data extracted by a group from UC Berkeley [3], containing 1702 emails is used. First we construct the term document matrix from email content and the hypergraph incidence matrix from email-sender/recipient relations. The hypergraph has Enron employees as vertices and their emails as edges, and a vertex is connected by an edge if and only if the corresponding employee is the sender or a recipient of the corresponding email. After that, we clean the data by removing terms that appear very infrequently and emails that are too short or duplicated, and extracting the largest connected components of the hypergraph. The tf-idf transformation is then applied to the term-document matrix, its columns are normalized to have unit 2-norm, which obtains the matrix X. S is computed using (7) in which M is the incidence matrix of the dual hypergraph. Finally, we apply JointNMF with $\alpha = \|X\|_F^2/\|S\|_F^2$ and $\beta = \alpha\|S\|_{max}$ to find 20 groups of employees. Note that since the dual hypergraph is used, the resulting clusters are clusters of emails rather than clusters of employees. To induce clusters of employees, one simply inserts employees involved in the same cluster of emails into one employee cluster. In this way, we can actually induce overlapping employee clusters from non-overlapping email clusters. It is assumed that an employee has j memberships if the employee belongs to j clusters. The number of memberships is counted for each employee and the frequency of each number is listed in Table 7. Employees that had at least 6 memberships are examined in online news and we found that they all held relatively high positions in Enron. Their names and positions are listed in Table 8. To see the effect of our algorithm on topic modeling, we list some

[3] `http://bailando.sims.berkeley.edu/enron_email.html`

Table 8 Employees that has j memberships ($j \geq 6$) and their positions in Enron

j	Name	Position in Enron
11	Steven Kean	Chief of staff
7	Jeff Dasovich	Governmental affairs executive
6	Susan Mara	California director of Regulatory Affairs
	Richard Shapiro	VP of regulatory affairs
	Paul Kaufman	VP of Government Affairs
	James Steffes	VP of Government Affairs
	Tim Belden	Head of trading
	Richard Sanders	VP of Enron Whole Sale Services
	Joe Hartsoe	VP of Federal Regulatory Affairs

VP: vice president

Table 9 Topic keywords of clusters

#	Keywords
0	ubs, warburg, forecast, confidential, win
1	blackberry, handheld, wireless
2	california, power, confidential, tariff, pursuant
3	caiso, refund, ferc, proceedings
4	burrito, peace, things, price, market, board, california
5	document, fax, tonight, sign, back, attach, thanks
6	wholesale, policy, compliance, receipt, legal, service
7	enron, please, know, attach, meeting, contact, call, any, time
8	london, conference, meeting, next, week
9	handheld, blackberry, wireless, agreement, confidential
10	testify, witness, fault, burden, cut, budget
11	california, electricity, energy, price, market, power, rate, bill
12	recommendation, template, participant, management
13	passcode, please, effective, confidential, change
14	stanford, university, expert, try, best, mail, california
15	account, invoice, trust, fund, transfer
16	expense, report, employee, name , approve, amount
17	folder, info, audit, access, apollo, email, sensitivity, server
18	sent, talk, presentation, thanks, infrastructure, amendment
19	hpl, aep, agreement, compete, deal, arrangement

topic keywords for each cluster in Table 9. It can be observed that some emails are communications about/with other companies and regulatory agencies (0,3,19); some are about administrative tasks or daily work (5,7,8,13,15,16,18); some are about legal issues (6,10); and some are related to the California energy crisis (2,11).

6 Conclusions and Discussions

With a simple CLRA formulation in (4), JointNMF is able to solve a variety of problems. The basic application of JointNMF is to cluster hybrid data with both content and connection structure, where the connection structure can be either a graph or a hypergraph, and the content can be associated with either the hypergraph nodes or

the edges. When X is any nonnegative feature-data matrix and S is a nonnegative data-data similarity matrix, the JointNMF formulation (4) naturally applies without any modification. When there are multiple feature-data matrices $X_1, \ldots, X_p$ and multiple similarity matrices $S_1, \ldots, S_q$, one can extend (4) to

$$\min_{W_i \geq 0, H \geq 0} \sum_{i=1}^{p} \alpha_i ||X_i - W_i H||_F^2 + \sum_{j=1}^{q} \gamma_j ||S_j - H^T H||_F^2$$

JointNMF can also be applied to predict paper/patent citations and detect activities and leaders in an organization.

As a hybrid clustering method, JointNMF, with easy-to-set parameters, successfully improves the cluster quality over content-only and connection-only clustering algorithms. It also outperforms one of the leading hybrid clustering methods in the sense of average F1 score and rand index.

Although the current default parameters ($\alpha = \|X\|_F^2 / \|S\|_F^2$ and $\beta = \alpha \|S\|_{max}$) for JointNMF are usually good enough, it was noticed in our experimental results that these are not optimal. We plan to study this further in future research to better understand these parameter values.

Our next research effort, in addition to those noted above, is to accelerate the JointNMF algorithm using a divide-and-conquer approach, as in [15]. In our experiments, JointNMF also demonstrates excellent potential for predicting paper/patent citations and activities and leaders in an organization. The application of JointNMF to citation recommendation and activity/leader detection will be further explored and more experimental results on additional data sets will serve to compare JointNMF with other algorithms in these two important areas that have many applications in critical domains such as organized crime detection.

References

1. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
2. Boyd, R., Drake, B., Kuang, D., Park, H.: `http://smallk.github.io/` (2016)
3. Chang, J., Blei, D.M.: HIERARCHICAL RELATIONAL MODELS FOR DOCUMENT NETWORKS. The Annals of Applied Statistics **4**(1), 124–150 (2010)
4. Choo, J., Lee, C., Reddy, C.K., Park, H.: Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Transactions on Visualization and Computer Graphics **19**(12), 1992–2001 (2013). DOI 10.1109/TVCG.2013.212. URL `http://dx.doi.org/10.1109/TVCG.2013.212`
5. Cohn, D.A., Hofmann, T.: The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity. In: T.K. Leen, T.G. Dietterich, V. Tresp (eds.) Advances in Neural Information Processing Systems 13, pp. 430–436. MIT Press (2001)
6. Cruz, J., Bothorel, C., Poulet, F.: Entropy based community detection in augmented social networks. In: 2011 International Conference on Computational Aspects of Social Networks (CASoN), pp. 163–168 (2011). DOI 10.1109/CASON.2011.6085937
7. Drake, B., Kim, J., Mallick, M., Park, H.: Supervised Raman spectra estimation based on nonnegative rank deficient least squares. In: Proceedings 13th International Conference on Information Fusion, Edinburgh, UK (2010)
8. Elhadi, H., Agam, G.: Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods. In: Proceedings of the 7th Workshop on Social Network Mining and Analysis, SNAKDD '13, pp. 10:1–10:7. ACM, New York, NY, USA (2013). DOI 10.1145/2501025.2501034

9. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences **101**(suppl 1), 5220–5227 (2004). DOI 10.1073/pnas.0307760101
10. Gruber, A., Rosen-Zvi, M., Weiss, Y.: Latent Topic Models for Hypertext. In: Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08), pp. 230–239. AUAI Press, Corvallis, Oregon (2008)
11. Jin, D., Gabrys, B., Dang, J.: Combined node and link partitions method for finding overlapping communities in complex networks. Scientific Reports **5** (2015). DOI 10.1038/srep08600
12. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. Journal of Global Optimization **58**(2), 285–319 (2014). DOI 10.1007/s10898-013-0035-4
13. Kim, J., Park, H.: Fast nonnegative matrix factorization: An active-set-like method and comparisons. SIAM Journal on Scientific Computing **33**(6), 3261–3281 (2011)
14. Kuang, D., Choo, J., Park, H.: Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In: M.E. Celebi (ed.) Partitional Clustering Algorithms, pp. 215–243. Springer International Publishing (2015). DOI 10.1007/978-3-319-09259-1_7
15. Kuang, D., Park, H.: Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 739–747. ACM (2013)
16. Kuang, D., Park, H., Ding, C.H.: Symmetric Nonnegative Matrix Factorization for Graph Clustering. In: SDM, vol. 12, pp. 106–117. SIAM (2012)
17. Kuang, D., Yun, S., Park, H.: SymNMF: Nonnegative low-rank approximation of a similarity matrix for graph clustering. Journal of Global Optimization **62**(3), 545–574 (2015). DOI 10.1007/s10898-014-0247-2
18. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data` (2014)
19. Liu, J., Wang, C., Gao, J., Han, J.: Multi-View Clustering via Joint Nonnegative Matrix Factorization. In: Proceedings of the 2013 SIAM International Conference on Data Mining, Proceedings, pp. 252–260. Society for Industrial and Applied Mathematics (2013)
20. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link LDA: Joint Models of Topic and Author Community. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pp. 665–672. ACM, New York, NY, USA (2009). DOI 10.1145/1553374.1553460
21. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
22. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic Modeling with Network Regularization. In: Proceedings of the 17th International Conference on World Wide Web, WWW '08, pp. 101–110. ACM, New York, NY, USA (2008). DOI 10.1145/1367497.1367512
23. Nallapati, R.M., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint Latent Topic Models for Text and Citations. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, pp. 542–550. ACM, New York, NY, USA (2008). DOI 10.1145/1401890.1401957
24. Ruan, Y., Fuhry, D., Parthasarathy, S.: Efficient Community Detection in Large Networks Using Content and Links. In: Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, pp. 1089–1098. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013)
25. Strehl, A., Ghosh, J.: Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. J. Mach. Learn. Res. **3**, 583–617 (2003). DOI 10.1162/153244303321897735
26. Sun, Y., Aggarwal, C.C., Han, J.: Relation Strength-aware Clustering of Heterogeneous Information Networks with Incomplete Attributes. Proc. VLDB Endow. **5**(5), 394–405 (2012). DOI 10.14778/2140436.2140437
27. Tang, J., Wang, X., Liu, H.: Integrating Social Media Data for Community Detection. In: Proceedings of the 2011 International Conference on Modeling and Mining Ubiquitous Social Media, MSM'11, pp. 1–20. Springer-Verlag, Berlin, Heidelberg (2012). DOI 10.1007/978-3-642-33684-3_1
28. Wang, X., Tang, L., Gao, H., Liu, H.: Discovering Overlapping Groups in Social Media. In: 2010 IEEE International Conference on Data Mining, pp. 569–578 (2010). DOI 10.1109/ICDM.2010.48
29. Wang, X., Tang, L., Liu, H., Wang, L.: Learning with multi-resolution overlapping communities. Knowledge and Information Systems **36**(2), 517–535 (2013). DOI 10.1007/s10115-012-0555-0
30. Xu, Y., Yin, W., Wen, Z., Zhang, Y.: An alternating direction algorithm for matrix completion with nonnegative factors. Frontiers of Mathematics in China **7**(2), 365–384 (2012). DOI 10.1007/s11464-012-0194-5. URL `http://dx.doi.org/10.1007/s11464-012-0194-5`

31. Yang, J., Leskovec, J.: Overlapping community detection at scale: A nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 587–596. ACM (2013)
32. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining Link and Content for Community Detection: A Discriminative Approach. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pp. 927–936. ACM, New York, NY, USA (2009). DOI 10.1145/1557019.1557120
33. Zhou, D., Huang, J., Schölkopf, B.: Learning with Hypergraphs: Clustering, Classification, and Embedding. In: B. Schölkopf, J.C. Platt, T. Hoffman (eds.) Advances in Neural Information Processing Systems 19, pp. 1601–1608. MIT Press (2007)