

Multi-way Clustering on Relation Graphs

Arindam Banerjee*

Sugato Basu†

Srujana Merugu‡

Abstract

A number of real-world domains such as social networks and e-commerce involve heterogeneous data that describes relations between multiple classes of entities. Understanding the natural structure of this type of heterogeneous relational data is essential both for exploratory analysis and for performing various predictive modeling tasks. In this paper, we propose a principled multi-way clustering framework for relational data, wherein different types of entities are simultaneously clustered based not only on their intrinsic attribute values, but also on the multiple relations between the entities. To achieve this, we introduce a relation graph model that describes all the known relations between the different entity classes, in which each relation between a given set of entity classes is represented in the form of multi-modal tensor over an appropriate domain. Our multi-way clustering formulation is driven by the objective of capturing the maximal “information” in the original relation graph, i.e., accurately approximating the set of tensors corresponding to the various relations. This formulation is applicable to all Bregman divergences (a broad family of loss functions that includes squared Euclidean distance, KL-divergence), and also permits analysis of mixed data types using convex combinations of appropriate Bregman loss functions. Furthermore, we present a large family of structurally different multi-way clustering schemes that preserve various linear summary statistics of the original data. We accomplish the above generalizations by extending a recently proposed key theoretical result, namely the minimum Bregman information principle [1], to the relation graph setting. We also describe an efficient multi-way clustering algorithm based on alternate minimization that generalizes a number of other recently proposed clustering methods. Empirical results on datasets obtained from real-world domains (e.g., movie recommendations, newsgroup articles) demonstrate the generality and efficacy of our framework.

1 Introduction

In recent years, there has been a lot of interest in probabilistic relational learning due to a plethora of real-world applications that involve modeling the relations between multiple types of entities, e.g., social networks, e-commerce. Often, the data available in these domains is sparse, high dimensional, incomplete, and noisy, which makes modeling difficult, e.g., movie ratings data in movie recommendation engines such as Yahoo! Movies, offer descriptions on product search web sites such as Froogle. Understanding the latent structure of this type of heterogeneous relational data is important for exploratory analysis and as pre-processing for subsequent predictive modeling tasks. In case of homogeneous data,

this is typically done using clustering techniques [14] that discover the “latent” groups of objects based on attribute similarity. To model the relationships between a pair of entity classes, several structurally different co-clustering techniques [9, 1, 7], which involve simultaneous clustering of the two entity classes represented as the rows and columns of a data matrix, have been proposed. Recently, these techniques were extended to multi-way clustering formulations [15, 4] involving multiple entity classes with pair-wise relations for certain specific structural configurations.

In this paper, we propose a principled multi-way clustering framework for relational data wherein different classes of entities are simultaneously clustered based not only on the intrinsic attribute values of the entities, but also on the multiple relations between the entities. Furthermore, each relation can involve multiple sets of entities (as opposed to pair-wise relations) and the relations can also have attributes. To achieve this, we introduce a relation graph model that describes all the known relations between the different entity classes. In this model, each relation between a given set of entity classes is represented as a multi-dimensional tensor (or data cube) over an appropriate domain, with the dimensions being associated with the various entity classes. Further, each cell in the tensor encodes the relation between a particular set of entities and can either take real values, i.e., the relation has single attribute, or itself is a vector of attributes. This general model is useful for applications in several domains that have multi-type relational datasets. Let us consider one such real-world example.

Example 1 Consider online movie recommendation applications such as Yahoo! Movies. These sites have movie viewer information linked to movie descriptions, which can be represented as several relations – (i) f^1 : ratings for (movie, viewer, actor) tuples corresponding to viewers’ feedback on the performance of actors in different movies; (ii) f^2 : co-occurrence indicators for (movie, actor) pairs specifying which actors acted in which movies; (iii) f^3 : counts for (movie, review words) tuples encoding the movie reviews; (iv) f^4 : values for (viewer, demographic attributes) that specify details such as age, gender, etc., for different viewers. Figure 1.1 shows an illustration of this dataset as a set of tensors (or data cubes) with a few common dimensions,

*Department of CSE, University of Minnesota

†AI Center, SRI International

‡Yahoo! Research

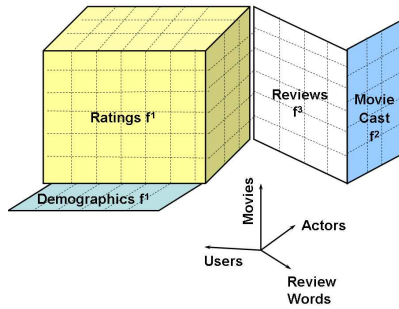


Figure 1.1: Example of multi-type relational data in movie recommendation.

e.g., f^1 and f^4 share the viewer dimension while f^1 and f^2 share both the actor and movie dimensions and f^3 also has the movie dimension. A typical application in this domain would be recommending a movie to a user who has not seen it before. Clustering jointly across the multiple entity classes, (i.e., movies, viewers, actors and words) allows us to capture most of the “information” in the inter-linked datasets in terms of compressed representations comprising of cluster-based statistics. This compression aids considerably in handling sparsity and high-dimensionality issues. This in turn enables us to obtain high quality predictions on the desired unknown ratings by extrapolating from the ratings of users with similar profiles (in terms of both movie preferences, demographic attributes) for similar movies (in terms of user preferences, review descriptions and cast). ■

There are several other domains where the relation graph over tensor representation would be directly applicable. Important examples include - (i) e-commerce applications, e.g., Froogle, which make use of data consisting of product descriptions, customer demographic profiles, transaction data, customer reviews of products and sellers, etc.; (ii) targeted internet advertising, e.g., AdSense, which makes use of a web page content, ad’s textual content, past click-through rates of ads posted on web pages; (iii) social network analysis, e.g., Blogger, which involves modeling user profiles, user generated content, interactions between users, etc.

1.1 Contributions. This paper provides a fairly general multi-way clustering framework for relation graphs, which is driven by the objective of preserving the maximal “information” in the original data. In particular, we make the following key contributions.

1. We introduce a multi-way clustering formulation for relation graphs that is motivated by an approximation point of view, i.e., accurately reconstructing the tensors corresponding to the various relations.

2. Our formulation is applicable to all Bregman divergences (a broad family of loss functions that include squared Euclidean distance, KL-divergence), and also permits analysis of mixed data types using convex combinations of appropriate Bregman loss functions.

3. We present a large family of structurally different multi-way clustering schemes that involve preserving summary statistics that are conditional expectations based on arbitrary partitionings of the tensors associated with the different relations.

4. To achieve the above generalizations, we extend the recently proposed *minimum Bregman information* principle [1], which generalizes both least squares and maximum entropy principles, to the relation graph setting. This also allows us to provide an alternate interpretation of our clustering formulation in terms of minimizing the *loss in Bregman information*.

5. We also propose a highly scalable algorithm based on alternate minimization that is linear in the number of non-zeros in the sparse data, converges to a local optimum of the multi-way clustering objective function, and generalizes several existing algorithms.

1.2 Overview of the Paper. We begin with a review of preliminary definitions in Section 2. Section 3 considers the case of a single relation over multiple entity classes represented as a multi-dimensional tensor, and presents the multi-way clustering formulation and algorithm in detail for this case. Section 4 introduces the notion of a relation graph for representing multiple relations over a set of entity classes, and describes how the multi-way clustering formulation for tensors can be extended to relation graphs associated with a set of tensors. Section 5 provides empirical evidence on the benefits of our multi-way clustering framework on real datasets. Section 6 discusses related work and explains how some of the existing co-clustering and multi-way clustering methods [15, 1, 9, 8] can be derived as special cases of our framework.

2 Preliminaries

In this section, we describe our notation and present some definitions related to Bregman divergences [6], which form a large class of well-behaved loss functions.

Notation: Sets such as $\{x_1, \dots, x_n\}$ are enumerated as $\{x_i\}_{i=1}^n$ and an index i running over the set $\{1, \dots, n\}$ is denoted by $[i]_1^n$. Sets are denoted using calligraphic upper case variables, e.g., \mathcal{S} . Random variables and random vectors are denoted using plain upper case letters, e.g., Z with the corresponding lower case letters z denoting the instantiated values. Tensors are denoted using upper case bold letters, e.g., \mathbf{Z} , whereas the corresponding subscripted lower case letters z_u denote the elements.

2.1 Bregman Divergence and Information

Definition 1 Let ϕ be a real-valued strictly convex function defined on the convex set $\mathcal{S} \equiv \text{dom}(\phi)$ ($\subseteq \mathbb{R}^d$), the domain of ϕ , such that ϕ is differentiable on $\text{int}(\mathcal{S})$, the interior of \mathcal{S} . The *Bregman divergence* $d_\phi : \mathcal{S} \times \text{int}(\mathcal{S}) \mapsto \mathbb{R}_+$ is defined as

$$d_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - \langle z_1 - z_2, \nabla\phi(z_2) \rangle,$$

where $\nabla\phi$ is the gradient of ϕ .

Special cases of Bregman divergences include **squared loss**, i.e., $d_\phi(z_1, z_2) = (z_1 - z_2)^2$, which corresponds to $\phi(z) = z^2$, $z \in \mathbb{R}$ and **I-divergence**, i.e., $d_\phi(z_1, z_2) = z_1 \log(z_1/z_2) - (z_1 - z_2)$, which corresponds to $\phi(z) = z \log z - z$, $z \in \mathbb{R}_+$. Given a Bregman divergence and a random variable, the uncertainty in the random variable can be captured in terms of a useful concept called Bregman information [3, 1] defined below.

Definition 2 For any Bregman divergence $d_\phi : \mathcal{S} \times \text{int}(\mathcal{S}) \mapsto \mathbb{R}_+$ and any random variable $Z \sim w(z)$, $z \in \mathcal{Z} \subseteq \mathcal{S}$, the *Bregman information* of Z is defined as the expected Bregman divergence to the expectation, i.e.,

$$I_\phi(Z) = E_w[d_\phi(Z, E_w[Z])] .$$

Intuitively, this quantity is a measure of the “spread” or the “information” in the random variable. Examples of Bregman information include **squared Frobenius norm** (for squared loss) and **negative entropy** (for I-divergence), which respectively correspond to random variables that are uniformly distributed over the entries of a given matrix and the joint probability values of two other random variables [1].

3 Multi-way Clustering on Tensors

In this section, we consider the special case where there are multiple classes of entities connected via a single relation, which can be described by a multi-dimensional tensor. For this case, we develop a multi-way clustering formulation that extends the matrix co-clustering framework presented in [1] to multi-dimensional tensors. Our formulation is driven by the objective of accurately approximating the original tensor using a reconstruction determined solely by the multi-way clustering and certain summary statistics of the original tensor.

3.1 Tensor Model. We begin with a description of our tensor representation. Let U_i , $[i]_1^n$, indicate random variables that take values over n different classes of entities with cardinalities m_i , $[i]_1^n$, respectively. Without loss of generality, we can assume that the support set of U_i corresponds to $\{1, \dots, m_i\}$ for $[i]_1^n$. Any relation f between all the variables U_i can be considered as

a deterministic function on the random vector $U_{all} = (U_1, \dots, U_n)$. Let the range of f be a subset of the convex set $\mathcal{S} = \text{dom}(\phi)$.¹ Since each random variable U_i takes m_i possible values, this relation f can be exactly described by a n -dimensional tensor $\mathbf{Z} \in \mathcal{S}^{m_1 \times \dots \times m_n}$. Further, let $Z = f(U_{all}) = f(U_1, \dots, U_n)$. Then, Z is a U_{all} -measurable random variable taking values in \mathcal{Z} following the joint distribution $p(U_{all}) = p(U_1, \dots, U_n)$. For notational simplicity, let w denote the measure induced on \mathcal{Z} by $p(U_{all})$ so that $p(Z = z_{u_1, \dots, u_n}) = w_{u_1 \dots u_n} = p(u_1, \dots, u_n) = p(u_{all})$.

Example 2 Consider the ratings data in the movie-recommendation problem in Example 1. This data can be viewed as describing a single relation between three classes of entities, i.e., viewers (U_1), actors (U_2), and movies (U_3), and corresponds to a 3-dimensional tensor \mathbf{Z} where z_{u_1, u_2, u_3} is the rating of viewer u_1 for actor u_2 in the movie u_3 . The measure w corresponds to the weights on the ratings and is usually assumed to be uniform over the known values and zero for the missing ones. ■

3.2 Multi-way Clustering Formulation. Let $k_i \leq m_i$, $[i]_1^n$ denote the number of clusters desired for the entity class U_i . Then, a multi-way clustering of these entity classes is defined as the n -tuple $\rho = (\rho_1, \dots, \rho_n)$ where each $\rho_i : \{1, \dots, m_i\} \mapsto \{1, \dots, k_i\}$ denotes a mapping from the entities to their respective clusters. Let \hat{U}_i be a random variable that takes values in $\{1, \dots, k_i\}$ such that $\hat{U}_i = \rho_i(U_i)$. We now seek to characterize the “goodness” of a multi-way clustering ρ in terms of an approximation to the original tensor \mathbf{Z} that depends solely on ρ and certain pre-specified summary statistics that need to be preserved. Let $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}(\rho)$ be such an approximation and let \hat{Z} be a U_{all} -measurable random variable that takes values in this approximate tensor $\hat{\mathbf{Z}}$ following w . Then, the quality of the multi-way clustering ρ can be readily measured in terms of the approximation error or the expected Bregman distortion between the random variables Z and \hat{Z} , i.e.,

$$(3.1) \quad E_w[d_\phi(Z, \hat{Z})] = \sum_{u_{all}} w_{u_{all}} d_\phi(z_{u_{all}}, \hat{z}_{u_{all}}) = d_{\Phi_w}(\mathbf{Z}, \hat{\mathbf{Z}}),$$

where Φ_w is a separable convex function induced on the tensors by the convex function ϕ . The multi-way clustering problem is then to find the optimal ρ that minimizes this expected Bregman distortion.

¹Note that \mathcal{S} can be any arbitrary convex set associated with a well defined convex function ϕ . For example, \mathcal{S} could be $[0, 1]^{d_1} \times \mathbb{R}^{d_2}$, in which case f maps each instantiation of U_{all} to an attribute vector of size $(d_1 + d_2)$ where the first d_1 attributes take values in $[0, 1]$ and the last d_2 attributes are real-valued.

Our formulation clearly depends on how \hat{Z} is characterized, with different choices leading to different clustering formulations. To characterize \hat{Z} , we need to specify (i) what summary statistics are to be preserved, and (ii) how to get an approximation based on the summary statistics. We discuss these aspects in detail in the next two subsections.

3.3 Summary Statistics. We consider summary statistics that are conditional expectations over different partitions of the tensor \mathbf{Z} . Our choice of conditional expectation-based statistics is driven by the fact that the conditional expectation $E[Z|V]$ is the optimal approximation of the original Z with respect to any Bregman divergence among all V -measurable functions [3] for any random variable V . Depending on the variable V , or equivalently, the partitionings of the tensor that one considers, one can obtain different sets of summary statistics. The simplest set of summary statistics is one that consists of the (weighted) multi-way cluster means, i.e., just one conditional expectation $E[Z|\hat{U}_1, \dots, \hat{U}_n]$ corresponding to $V = (\hat{U}_1, \dots, \hat{U}_n)$. Another simple example consists of the marginal averages along any particular dimension of the tensor, i.e., conditional expectations of the form $E[Z|U_i]$ corresponding to $V = U_i$. In general, the summary statistics may comprise of multiple conditional expectations as we now describe.

Let $\{V_s\}_{s=1}^r$ be a set of r random variables corresponding to different partitionings (or sub- σ -algebras) of the tensor that are determined completely by the random variables U_i and \hat{U}_i , $[i]_1^n$. For example, for a two-dimensional tensor, i.e., a matrix, we may have $V_1 = (\hat{U}_1, U_2)$ and $V_2 = (U_1, \hat{U}_2)$ with $r = 2$. We call such a set a *multi-way clustering basis* and associate it with summary statistics given by the corresponding set of conditional expectations $\{E[Z|V_s]\}_{s=1}^r$. Note that each such set of conditional expectations leads to a different optimal reconstruction for the *same* multi-way clustering, and hence, a structurally different co-clustering scheme. We now describe two specific examples of multi-way clustering bases that will be used in the experiments in Section 5.

Block Multi-way Clustering (BMC). This is the simple case where the basis is the singleton set $\{V_1\} = \{(\hat{U}_1, \dots, \hat{U}_n)\}$. The summary statistics are just the (weighted) multi-way cluster means or $E[Z|\hat{U}_1, \dots, \hat{U}_n]$. In particular, for each multi-way cluster $(\hat{u}_1, \dots, \hat{u}_n)$, there is a single statistic given by $E[Z|\hat{u}_1, \dots, \hat{u}_n]$ so that we have $\prod_{i=1}^n k_i$ parameters to approximate the original tensor \mathbf{Z} .

Bias-Adjusted Multi-way Clustering (BAMC). Another important case is one where the basis consists of $r = n + 1$ random variables $\{V_s\}_{s=1}^{n+1}$ where $V_s = U_s, [s]_1^n$, and $V_{n+1} = (\hat{U}_1, \dots, \hat{U}_n)$. In this case, the summary

statistics not only include the multi-way cluster means $E[Z|\hat{U}_1, \dots, \hat{U}_n]$, but also the (weighted) average values over every possible $(n-1)$ dimensional slice of the tensor, i.e., $\{E[Z|U_i]\}_{i=1}^n$. These additional summary statistics capture the biases of the individual entities, and provide a better approximation. Since each conditional expectation $E[Z|U_i]$ results in m_i values, the total size of the summary statistics is $\prod_{i=1}^n k_i + \sum_{i=1}^n m_i$ related via $\sum_{i=1}^n k_i$ linear dependencies.

Example 3 For the movie recommendation example, one could consider either the BMC/BAMC bases described above or even a more general clustering basis. However, the BAMC scheme is usually a good choice since the summary statistics encode the biases associated with the individual viewers (e.g., level of criticality), actors (e.g., quality of acting) and movies (e.g., quality of plot/direction) in terms of the average viewer rating $E[Z|U_1]$, average actor rating $E[Z|U_2]$ and average movie rating $E[Z|U_3]$ respectively. The coarse structure of the ratings, on the other hand, is captured by the multi-way cluster means $E[Z|\hat{U}_1, \hat{U}_2, \hat{U}_3]$, i.e., the average rating of a viewer cluster for the performance of a cluster of actors in a movie cluster. ■

3.4 Minimum Bregman Information Principle.

Given a multi-way clustering ρ and a clustering basis $\{V_s\}_{s=1}^r$, we now seek to characterize the best approximation \hat{Z} to the original Z . We begin by considering a special class of approximations \mathcal{S}_A such that every $Z' \in \mathcal{S}_A$ preserves the conditional expectations associated with the specified multi-way clustering basis, i.e.,

$$(3.2) \quad \mathcal{S}_A = \{Z' | E[Z|V_s] = E[Z'|V_s], [s]_1^r\}.$$

To find the best approximation in the set \mathcal{S}_A , we invoke the *minimum Bregman information* principle, which was first proposed in [1], and which generalizes the well-known maximum entropy and least squares principles. The MBI principle posits that the best approximation \hat{Z} is the random variable $\hat{Z}_A \in \mathcal{S}_A$ that has the minimum Bregman information, i.e.,

$$(3.3) \quad \hat{Z}_A = \underset{Z' \in \mathcal{S}_A}{\operatorname{argmin}} I_\phi(Z').$$

Intuitively, the “best” approximation given certain information is one that does not make any extra assumptions over the available information. Mathematically, under certain definition of optimality, the notion of “no extra assumptions” translates to *minimum Bregman information* while the “available information” corresponds to the linear constraints associated with the conditional expectation statistics.

The following theorem characterizes the solution to the MBI problem (3.3).

Theorem 1² For any random variable Z , Bregman divergence d_ϕ , multi-way clustering ρ , and clustering basis $\{V_s\}_{s=1}^r$, the problem (3.3) reduces to a convex optimization problem with a unique solution

$$\hat{Z}_A = h_\phi(\Lambda^*, U_{\text{all}}, \rho(U_{\text{all}})) ,$$

where h_ϕ is uniquely determined function and $\Lambda^* = (\Lambda_{V_1}^*, \dots, \Lambda_{V_r}^*)$ are the optimal Lagrange multipliers with respect to the linear constraints $E[Z'|V_s] = E[Z|V_s]$, $[s]_1^r$.

Though the MBI problem (3.3) has a unique solution \hat{Z}_A , in general, the solution cannot be expressed in closed form as a function of the summary statistics, except for certain special bases such as BMC and BAMC, which are discussed below.

Block Multi-way Clustering. The MBI solution \hat{Z}_A in this case is the conditional expectation $E[Z|\hat{U}_1, \dots, \hat{U}_n]$ itself for *all* Bregman divergences, i.e., each entry $z_{u_1, \dots, u_n} = z_{u_{\text{all}}}$ in the original tensor \mathbf{Z} is approximated by the average value across the corresponding the multi-way cluster, i.e., $E[Z|\hat{u}_1, \dots, \hat{u}_n]$.

Bias-Adjusted Multi-way Clustering. For this basis, the MBI solution has a closed form only for specific choices of Bregman divergences. In particular, for squared loss, the MBI solution is given by

$$\hat{Z}_A = \sum_{i=1}^n (E[Z|U_i] - E[Z|\hat{U}_i]) + E[Z|\hat{U}_1, \dots, \hat{U}_n],$$

whereas for I-divergence, the MBI solution is given by

$$\hat{Z}_A = \frac{\prod_{i=1}^n E[Z|U_i] E[Z|\hat{U}_1, \dots, \hat{U}_n]}{\prod_{i=1}^n E[Z|\hat{U}_i]} .$$

In other words, the entry z_{u_1, \dots, u_n} is approximated by additive/multiplicative combinations of the average ratings of the entities u_1, \dots, u_n as well as those of the corresponding clusters $\hat{u}_1, \dots, \hat{u}_n$ and also the average rating across the corresponding multi-way cluster.

A natural question to ask is: Why is the MBI solution \hat{Z}_A a good approximation to the original Z ? First, the approximation is based solely on the multi-way clustering and the specified summary statistics derived from the original Z . Second, the approximation \hat{Z}_A actually preserves all the statistics under consideration. Lastly and more importantly, it can be shown that the MBI solution is the optimal approximation to the original Z among a large class of reconstructions, as the following result shows.

Theorem 2 Given a random variable Z , Bregman divergence d_ϕ , multi-way clustering ρ and clustering basis $\{V_s\}_{s=1}^r$, let \mathcal{S}_B be the set of generalized additive

functions based on natural parameterizations (denoted by $g_\phi(\cdot)$) of the summary statistics, i.e.,

$$\mathcal{S}_B = \left\{ Z'' \mid Z'' = g_\phi^{(-1)} \left(\sum_{s=1}^r q_s(g_\phi(E[Z|V_s])) \right) \right\} ,$$

where $q_s, [s]_1^r$ are arbitrary functions. Then, the MBI solution \hat{Z}_A in (3.3) is the unique minimizer of the expected distortion with respect to Z in \mathcal{S}_B , i.e.,

$$\hat{Z}_A = \underset{Z'' \in \mathcal{S}_B}{\operatorname{argmin}} E[d_\phi(Z, Z'')] .$$

Note that for the simple BMC case, the summary statistics consist of a single conditional expectation so that \mathcal{S}_B includes *all deterministic functions* of the summary statistics, and hence, the MBI solution is the optimal reconstruction in this entire set. In the general case, the optimality property is limited to the specific, but large class of approximations described in Theorem 2. In particular, for squared error, the natural parameterization g_ϕ is the identity mapping itself so that \mathcal{S}_B consists of all generalized additive models, while in case of I-divergence, g_ϕ corresponds to log transformation and \mathcal{S}_B is the set of all generalized multiplicative models.

Choosing the MBI solution as the best approximation, i.e., $\hat{Z} = \hat{Z}_A$ also leads to the following result, which shows that the expected distortion between the original Z and the approximation \hat{Z} is exactly equal to the loss in Bregman information, thus, providing an alternative characterization of the problem formulation in terms of minimizing the loss in Bregman information.

Theorem 3 For any random variable Z , Bregman divergence d_ϕ and MBI solution \hat{Z} as defined in (3.3),

$$E[d_\phi(Z, \hat{Z})] = I_\phi(Z) - I_\phi(\hat{Z}) .$$

The multi-way clustering problem for tensors, can therefore be posed as one of finding the optimal multi-way clustering ρ^* that solves the minimization problem:

$$(3.4) \quad \min_{\rho} E[d_\phi(Z, \hat{Z})] = \min_{\rho} [I_\phi(Z) - I_\phi(\hat{Z})] .$$

3.5 Algorithm. We now propose an alternate minimization scheme for optimizing the multi-way clustering objective function in (3.4) that is applicable to all Bregman loss functions and multi-way clustering bases. Our algorithm considers each dimension in turn, finds the optimal clustering with respect to that dimension keeping everything else fixed, and recomputes the MBI solution, and this process is repeated till convergence.

²Please see [18] for more details on the theorems and proofs.

Since the cluster assignment step along each dimension is going to have similar form, we focus on the i^{th} dimension with index U_i and current clustering ρ_i . Let U_{-i} be the index and ρ_{-i} denote the clustering over all the other dimensions. From Theorem 1, the MBI solution in this case is given by $\hat{Z} = h_\phi(\Lambda, U_{\text{all}}, \rho(U_{\text{all}})) = h_\phi(\Lambda, (U_i, U_{-i}), (\rho_i(U_i), \rho_{-i}(U_{-i})))$ where Λ are the optimal Lagrange multipliers w.r.t ρ . Now, for any candidate clustering $\tilde{\rho}_i$ over the i^{th} dimension, we consider a new reconstruction $\tilde{Z} = h_\phi(\Lambda, (U_i, U_{-i}), (\tilde{\rho}_i(U_i), \rho_{-i}(U_{-i})))$. Then, from the definition of expectation, it follows that

$$E[d_\phi(Z, \tilde{Z})] = \sum_{u_i=1}^{m_i} w_{u_i} E_{U_{-i}|u_i}[d_\phi(Z, \tilde{Z}(u_i, \tilde{\rho}_i(u_i)))]$$

where $\tilde{Z}(u_i, \tilde{\rho}_i(u_i)) = h_\phi(\Lambda, (u_i, U_{-i}), (\tilde{\rho}_i(u_i), \rho_{-i}(U_{-i})))$. Thus, the objective function can be decomposed into a weighted average over m_i terms, each of which depends on a single u_i and its assignment $\tilde{\rho}_i(u_i)$. Hence, the optimal clustering ρ_i^{new} can be obtained by computing the optimal cluster assignments for each u_i as in step B of Algorithm 2.

Once the cluster assignments are updated, the objective function decreases, but the reconstruction \tilde{Z}^{new} based on Λ and the new cluster assignments $\rho^{\text{new}} = (\rho_i^{\text{new}}, \rho_{-i})$ need not be the MBI solution. As the following theorem shows, we can, in fact, obtain a better approximation to the original Z by computing the MBI solution \hat{Z}^{new} with respect to ρ^{new} .

Theorem 4 *For any random variable Z , Bregman divergence d_ϕ and multi-way clustering ρ^{new} , the MBI reconstruction \hat{Z}^{new} has less expected distortion with respect to Z than any reconstruction \tilde{Z}^{new} based on non-optimal Lagrange multipliers, i.e.,*

$$E[d_\phi(Z, \hat{Z}^{\text{new}})] \leq E[d_\phi(Z, \tilde{Z}^{\text{new}})] .$$

To draw analogy with the KMeans algorithms, the computation of the MBI solution is equivalent to computing the means of every cluster. Putting together the cluster assignment steps and the computation of MBI solution, we have a simple and elegant algorithm (Algorithm 1) for multi-way clustering of tensors, which is conceptually similar to the KMeans algorithm. Since both step B and C decrease the objective function, Algorithm 1 is guaranteed to converge to a local optimum.

Theorem 5 *The multi-way tensor clustering algorithm (Algorithm 1) monotonically decreases the multi-way clustering objective function in (3.4) and converges to a locally optimal solution.*

Algorithm 1 Multi-way Tensor Clustering

Input: Tensor $\mathbf{Z} \subseteq S^{m_1 \times \dots \times m_n}$, probability measure w , Bregman divergence $d_\phi : S \times \text{int}(S) \mapsto \mathbb{R}_+$, num. of clusters (k_1, \dots, k_n) , co-clustering basis $\{V_s\}_{s=1}^r$.

Output: Multi-way clustering ρ^* that (locally) optimizes the objective function in (3.4).

Method:

Initialize with an arbitrary clustering ρ .

repeat

Step A: Pick a dimension i , ($1 \leq i \leq n$) to update.

Step B: Compute cluster assignments ρ_i

$\rho_i(u_i) = \underset{\hat{u}_i: [\hat{u}_i]_1^{k_i}}{\text{argmin}} E_{U_{-i}|u_i}[d_\phi(Z, h_\phi(\Lambda, (u_i, U_{-i}), (\hat{u}_i, \rho_{-i}(U_{-i}))))]$

Step C: Compute the MBI solution and optimal Lagrange multipliers Λ for ρ using (3.3).

until convergence

return ρ

For some important choices of Bregman divergences and clustering bases (BAMC/BMC), the MBI solution in Step C can be computed in closed form, so that Algorithm 1 requires a computational effort that is only linear per iteration in the size of the data (non-zeros in the tensor) and is hence, very scalable. It must be noted here that Algorithm 1 is a meta algorithm that can be instantiated for any choice of Bregman divergence and clustering basis, since the MBI solution can always be computed numerically using iterative projection algorithms [6]. In practice, we can also use multiple restarts or local search heuristics [8] to improve the locally optimal solution.

4 Multi-way Clustering on Relation Graphs

In this section, we introduce our *relation graph* model, which captures multiple relations between a specified set of variables, where each relation corresponds to a tensor over possibly different domains. Then, we extend the multi-way clustering formulation in Section 3 for this setting by defining a suitable convex function over these relation graphs. More specifically, we use this convex function to define the Bregman information of a relation graph and thereby, characterize the optimal reconstruction of the relation graph via MBI principle. The multi-way clustering problem is then posed in terms of determining the clustering that finds the optimal reconstruction of the original relation graph. We also describe an efficient alternate minimization-based algorithm (MRGC) for multi-way relation graph clustering.

4.1 Relation Graph Model. We begin with a description of the multi-relational setting. As in Section 3, let U_i , $[i]_1^n$ indicate random variables that take values over n different classes of entities indexed by $\{1, \dots, m_i\}$, $[i]_1^n$ respectively and let $U_{\text{all}} = \{U_i\}_{i=1}^n$.

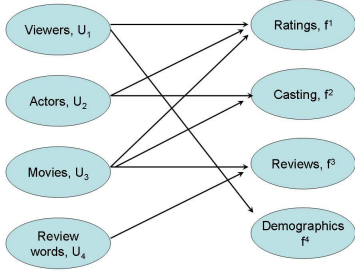


Figure 4.2: Relation graph for movie recommendation.

The tensor formulation in Section 3 considers the case where there is a single relation connecting the entity classes U_i , $[i]_1^n$. In a multi-relational setting, there exist multiple such relations between the variables in \mathcal{U}_{all} . Let f^j , $[j]_1^l$ denote these relations, each of which corresponds to a set of variables $\mathcal{U}^{f^j} \subseteq \mathcal{U}_{all}$, and also a tensor \mathbf{Z}^j whose entries take values in an appropriate domain. Note that it is possible that a relation f^j corresponds to a singleton set, i.e., $\mathcal{U}^{f^j} = \{U_i\}$ for some i , ($1 \leq i \leq n$), in which case f^j just maps the entities denoted by U_i to their intrinsic attributes. Let U^{f^j} denote the random vector corresponding to the variables in \mathcal{U}^{f^j} . Then, we can define a new random variable $Z^j = f^j(U^{f^j})$, which takes values among the elements of the tensor \mathbf{Z}^j following the probability distribution $p^j(U^{f^j})$ associated with the relation f^j .³ The dependencies between the various entity classes U_i , $[i]_1^n$, relations f^j , $[j]_1^l$, data tensors \mathbf{Z}^j , $[j]_1^l$ and random variables Z^j , $[j]_1^l$ can be conveniently represented in the form of relation graph defined below.

Definition 3 A *relation graph* \mathbf{G} is a directed $(\mathcal{U}_{all} + \mathcal{F}, \mathcal{E})$ -bipartite graph where $\mathcal{U}_{all} = \{U_i\}_{i=1}^n$ is a set of random variables and $\mathcal{F} = \{f^j\}_{j=1}^l$ is a set of relations. Each relation f^j is a deterministic mapping defined over the random vector U^{f^j} determined by all the variables $U \in \mathcal{U}^{f^j} \subseteq \mathcal{U}_{all}$ that are linked to f^j by an edge $e_{U, f^j} \in \mathcal{E}$. Further, each relation node $f^j \in \mathcal{F}$ is associated with a random variable $Z^j = f^j(U^{f^j})$, whose support set is the tensor \mathbf{Z}^j .

Example 4 Figure 4.1 shows the relation graph for the scenario described in Example 1. First, we observe that there are four classes of entities corresponding to viewers (U_1), actors (U_2), movies (U_3) and review words (U_4). We also have four relations between these entity classes:

³Note that the probability distributions are conditioned on the relations and do not have to be consistent with respect to each other.

(1) f^1 : viewer U_1 's rating of actor U_2 in movie U_3 , (2) f^2 : actor U_2 's participation in movie U_3 , (3) f^3 : word U_4 's occurrence in description of movie U_3 , and (4) f^4 : viewer U_1 's attributes such as gender and age. Each of these relations f^j corresponds to a data tensor \mathbf{Z}^j . Note that we view \mathbf{Z}^4 to be a tensor with a single axis and elements corresponding to (gender, age) pairs which can be embedded in $\mathcal{S} = [0, 1] \times \mathbb{R}_{++}$. ■

4.2 Multi-way Clustering Formulation. We define a multi-way clustering $\rho = (\rho_1, \dots, \rho_n)$ for (k_1, \dots, k_n) clusters along each dimension and the corresponding random variables $\hat{U}_i = \rho_i(U_i)$, $[i]_1^n$, exactly as in Section 3. To characterize the “goodness” of a multi-way clustering, we first observe that the relation graph \mathbf{G} with l relations is described by the l data tensors \mathbf{Z}^j or equivalently, the random variables Z^j . Hence, a natural formulation would be in terms of approximating each of these tensors based on the multi-way clustering. However, it is often not possible to have a single loss function to measure the distortion across all the data tensors \mathbf{Z}^j , as they might take values in different domains, e.g., in the movie recommender system, the entries in \mathbf{Z}^1 take real values in the range $[-10, 10]$ while those in \mathbf{Z}^2 take only binary values.

To handle this, we measure the approximation error in terms of a weighted combination of the suitable Bregman loss functions applied to each of the tensors. More specifically, let ϕ^j be a suitable convex function defined over the entries in \mathbf{Z}^j and let ν^j denote the weight associated with \mathbf{Z}^j . In practice, the weights ν^j , $[j]_1^l$ can be chosen to be proportional to the number of observations on the corresponding relations or the relevance of these relations for a specific prediction problem of interest. Let $\hat{\mathbf{Z}}^j$ be a reconstruction of the original tensor \mathbf{Z}^j , such that $\hat{\mathbf{Z}}^j$ depends only upon the clustering of the random variables in U^{f^j} , and a pre-specified set of summary statistics derived from the multi-way clustering and \mathbf{Z}^j . Further, let \hat{Z}^j be a random variable that takes values in the approximate tensor $\hat{\mathbf{Z}}^j$ following the distribution $p^j(U^{f^j})$. The “goodness” of the multi-way clustering can now be measured in terms of the weighted expected distortion between the random vectors (Z^1, \dots, Z^l) and $(\hat{Z}^1, \dots, \hat{Z}^l)$, i.e.,

$$(4.5) \quad \sum_{j=1}^l \nu^j E_{p^j} [d_{\phi^j}(Z^j, \hat{Z}^j)].$$

One can arrive at a more elegant interpretation of the above cost function by observing that the random variables \hat{U}_i , $[i]_1^n$, and \hat{Z}^j , $[j]_1^l$ define a new relation graph $\hat{\mathbf{G}}$ which has identical structure as the original relation graph \mathbf{G} , but is an approximation in terms of the data values. Let $G = (Z^1, \dots, Z^l)$ be the random

vector that captures all information in the relation graph \mathbf{G} and similarly, let $\hat{G} = \hat{G}(\boldsymbol{\rho}) = (\hat{Z}^1, \dots, \hat{Z}^l)$. The cost function in (4.5) can now be expressed as $E[d_\phi(G, \hat{G})]$ where

$$(4.6) \quad \phi(G) = \sum_{j=1}^l \nu^j \phi^j(Z^j).$$

The multi-way clustering problem is to find the $\boldsymbol{\rho}$ that minimizes the expected Bregman distortion $E[d_\phi(G, \hat{G})]$.

Example 5 In the movie recommendation example, since \mathbf{Z}^1 corresponds to ratings in $[-10, 10]$, it is reasonable to have $\phi^{(1)}$ as the squared error. For the tensors \mathbf{Z}^2 and \mathbf{Z}^3 consisting of co-occurrence values, it is more appropriate to have $\phi^{(2)}$ and $\phi^{(3)}$ as I-divergence. Similarly, $\phi^{(4)}$ can be chosen as an appropriate convex function over a 2-tuple of binary and real values. ■

Similar to the tensor setting, the choice of \hat{G} is critical to the clustering formulation. Further, as before, it can be fully described in terms of the summary statistics that need to be preserved and the reconstruction procedure based on the MBI principle.

4.3 Summary Statistics. As in the case of tensor formulation, for each relation f^j , we consider summary statistics that are conditional expectations of the random variables Z^j with respect to random variables $\{V_s^j\}_{s=1}^{r_j}$ that correspond to the different partitions of \mathbf{Z}^j . The complete clustering basis in this case is the union of all these sets of random variables, i.e., $\{\{V_s^j\}_{s=1}^{r_j}\}_{j=1}^l$. Further, since the random variables Z^j , $[j]_1^l$ correspond to different relations, one can consider different set of conditional expectations for each of these random variables, as illustrated in the following example.

Example 6 For the movie recommendation example, we have four relations and corresponding random variables Z^1, \dots, Z^4 . Following Example 3, we observe that for the first relation involving viewer ratings, it would be appropriate to choose the bias-adjusted multi-way clustering (BAMC) basis. Similarly for the relations f^2, f^3 based on co-occurrence counts, one could chose either the BAMC or BMC scheme depending on the prominence of the entity-bias whereas for f_4 , compression can be obtained by only choosing the BMC scheme. ■

4.4 MBI Principle for Relation Graphs. We now focus on obtaining the approximation $\hat{G} = (\hat{Z}^1, \dots, \hat{Z}^l)$, given a fixed multi-way clustering a clustering basis, by following a similar strategy as in the case of the tensor formulation.

First, we characterize the Bregman information of the random vector $G = (Z^1, \dots, Z^l)$ associated with a

relation graph \mathbf{G} using the convex function ϕ in (4.6), i.e.,

$$\begin{aligned} I_\phi(G) &= E[d_\phi(G, E[G])] \\ &= \sum_{j=1}^l \nu^j E_{p_j} [d_{\phi^j}(Z^j, E_{p_j}[Z^j])] = \sum_{j=1}^l \nu^j I_{\phi^j}(Z^j). \end{aligned}$$

This allows to invoke the minimum Bregman information (MBI) principle and pick the “best” approximation as the one that has the minimum Bregman information, subject to the linear constraints arising from the summary statistics to be preserved, i.e.,

$$(4.7) \quad \hat{G}_A \equiv \operatorname{argmin}_{G' \in \mathcal{S}_A} I_\phi(G')$$

where $G' = (Z'^1, \dots, Z'^l)$ and \mathcal{S}_A is given by

$$\mathcal{S}_A = \{G' | E_{p_j}[Z'^j | V_s^j] = E_{p_j}[Z^j | V_s^j], \forall [s]_1^{r_j}, \forall [j]_1^l\}.$$

Due to the separability of the convex function ϕ and the resulting Bregman information, it can be shown that the MBI solution \hat{G}_A can in fact be readily expressed in terms of the MBI solutions corresponding to the component random variables Z^j , $[j]_1^l$.

Theorem 6 For any relation graph \mathbf{G} , Bregman divergence d_ϕ and multi-way clustering $\boldsymbol{\rho}$, and clustering basis $\{\{V_s^j\}_{s=1}^{r_j}\}_{j=1}^l$, the MBI solution $\hat{G}_A = (\hat{Z}_A^1, \dots, \hat{Z}_A^l)$, where \hat{G}_A is as in (4.7) and \hat{Z}_A^j is the MBI solution for the tensor \mathbf{Z}^j with respect to the basis $\{V_s^j\}_{s=1}^{r_j}$ and clustering $\boldsymbol{\rho}$, as defined in (3.3).

Using the above decomposition result and Theorem 1, one can uniquely determine the MBI solution \hat{G}_A . This reconstruction not only preserves all the specified summary statistics, but also results in the minimum expected Bregman distortion with respect to the original G among a large class of possible reconstructions. Further, the expected Bregman distortion between the original G and the MBI approximation \hat{G}_A can also be expressed as the loss in Bregman information due to clustering, i.e.,

$$(4.8) \quad E[d_\phi(G, \hat{G}_A)] = I_\phi(G) - I_\phi(\hat{G}_A).$$

Hence forth, we define $\hat{G} \equiv \hat{G}_A$ so that the multi-way clustering problem can be posed as that of finding the optimal multi-way clustering $\boldsymbol{\rho}^*$ that solves the following minimization problem,

$$(4.9) \quad \min_{\boldsymbol{\rho}} E[d_\phi(G, \hat{G})] = \min_{\boldsymbol{\rho}} I_\phi(G) - I_\phi(\hat{G}_A).$$

4.5 MRGC Algorithm. In order to solve the multi-way clustering problem for relation graphs, we adopt a similar alternate minimization strategy as in the case of

tensor clustering. Similar to Section 3, multi-way relation graph clustering (MRGC) involves an iterative process where the cluster assignments of each dimension are updated followed by the computation of the MBI solution. The only difference is that the optimal cluster assignments and the MBI computation depend on multiple tensors associated with different relations.

First, we consider the cluster assignment step for the i^{th} dimension. For any relation f^j , let $U_{-i}^{f^j}$ denote the random variable over the dimensions other than i and ρ_{-i}^j the corresponding clustering. For any multi-way clustering ρ , the MBI solution $\hat{G} = (\hat{Z}^1, \dots, \hat{Z}^l)$ where \hat{Z}^j is the MBI solution associated with the tensor Z^j and is determined by Theorem 1. Now, for any candidate clustering $\tilde{\rho}_i$ along the i^{th} dimension, we can consider a reconstruction \tilde{G} consisting of relation-wise reconstructions \tilde{Z}^j similar to the one described in Section 3.4. In particular, $\tilde{Z}^j = \hat{Z}^j$ when the relation f^j does not involve the i^{th} dimension, i.e., $U_i \notin \mathcal{U}^{f^j}$.

As in Section 3.4, the expected distortion $E[d_\phi(G, \tilde{G})]$ can be expressed as a sum over m_i terms, each of which depends on a single u_i and its assignment $\tilde{\rho}_i(u_i)$. Hence, the new clustering ρ_i^{new} can be obtained by computing the optimal cluster assignments for each u_i (Step B of Algorithm 2). The new cluster assignment $\rho^{\text{new}} = (\rho_i^{\text{new}}, \rho_{-i})$ has a lower expected distortion, but the resulting reconstruction \tilde{G}^{new} is not the MBI solution. From Theorems 6 and 4, it can be shown that [18] the MBI solution \hat{G}^{new} corresponding to ρ^{new} is a better approximation than \tilde{G}^{new} . Hence, we recompute the MBI solution \hat{G}^{new} , which because of Theorem 6 only involves recomputing the MBI solutions of the relations that involve U_i . Algorithm 2 shows the main steps and is guaranteed to monotonically decrease the multi-way clustering objective function till it reaches a local optimum [18]. For special cases where the multi-way clustering basis and the Bregman loss functions (ϕ^1, \dots, ϕ^j) are such that the MBI solutions for all relations have a closed form, the multi-way relation graph clustering (MRGC) algorithm only requires linear computational time per iteration and is highly scalable. Further as noted earlier, Algorithm 2 can be instantiated for any relation graph setting with a well defined set of Bregman loss functions and multi-way clustering basis.

5 Experimental Results

In this section, we provide experimental results that highlight the flexibility and effectiveness of our multi-way clustering framework. First, we describe experiments on synthetic data for studying the dependence of the clustering quality and the resulting approximation on the choice of Bregman divergence and multi-way clustering basis. Then, we present results on real world datasets for document and movie categorization tasks to

Algorithm 2 Multi-way Relation Graph Clustering

Input: Relation graph $\mathbf{G} = (\mathbf{Z}^1, \dots, \mathbf{Z}^j)$ based on relations f_j , $[j]_1^l$, associated weights ν^j , $[j]_1^l$, probability measures w^j , $[j]_1^l$, Bregman divergences $d_{\phi_j} : S \times \text{int}(S) \mapsto \mathbb{R}_+$, num. of clusters (k_1, \dots, k_n) , multi-way clustering basis $\{\{V_s^j\}_{s=1}^{r_j}\}_{j=1}^l$.

Output: Multi-way clustering ρ^* that (locally) optimize the objective function in (4.9).

Method:

Initialize with an arbitrary clustering ρ .

repeat

Step A: Pick a dimension i , ($1 \leq i \leq n$) to update.

Step B: Compute cluster assignments ρ_i

$$\rho_i(u_i) = \underset{\hat{u}_i: [\hat{u}_i]_1^{k_i}}{\text{argmin}} \sum_{j: U_i \in \mathcal{U}^{f^j}} \nu^j w_{u_i}^j E_{U_{-i}|u_i} [d_\phi(Z^j, \tilde{Z}^j(u_i, \hat{u}_i))]$$

$$\text{where } \tilde{Z}^j(u_i, \hat{u}_i) = h_\phi(\Lambda^j, (u_i, U_{-i}^{f^j}), (\hat{u}_i, \rho_{-i}^j(U_{-i}^{f^j})))$$

Step C: Compute MBI solution \hat{G} for ρ using (4.7).

until convergence

return ρ

Configuration	Squared Error	nMI
A	55.9 ± 7.7	0.742 ± 0.026
B	71.3 ± 8.9	0.664 ± 0.051
C	190.9 ± 21.2	0.528 ± 0.053

Table 5.1: Performance using different Bregman divergences and clustering bases with $k_1 = k_2 = k_3 = 5$. nMI was averaged over U_1, U_2 and U_3 and $\nu^{(1)} = \nu^{(2)}$.

demonstrate the benefits of tensor and multi-relational clustering over 2-dimensional co-clustering. We also describe a case study on performing collaborative filtering for movie recommendations via a multi-way clustering approach.

5.1 Choice of Bregman Divergence and Clustering Basis. As mentioned earlier, the MRGC algorithm is applicable for any choice Bregman divergence and also for any valid clustering basis (including, but not limited to BAMC and BMC). When the choice of the Bregman divergence and the clustering basis capture the natural structure of the data, one can obtain a highly accurate approximation of the original data. To study this dependence, we generated 10 sets of two 50×50 matrices \mathbf{Z}^1 and \mathbf{Z}^2 (shown in Figure 5.3), which correspond to relations between U_1, U_2 and U_1, U_3 respectively, where U_1, U_2, U_3 are three classes of entities. The matrices \mathbf{Z}^1 and \mathbf{Z}^2 were obtained from structured Gaussian mixture models [18] corresponding to the *BAMC* and *BMC* scheme respectively with $k_1 = k_2 = k_3 = 5$. The bijection result between Bregman divergences and regular exponential families [3] suggests that the squared loss as the appropriate Bregman divergence in this case. To validate this, we performed multi-way clustering using the following three configurations - (A) squared error and BAMC for \mathbf{Z}^1 /BMC for \mathbf{Z}^2 , (B) I-divergence and BAMC for \mathbf{Z}^1 /BMC for \mathbf{Z}^2 , (C) squared error and BMC

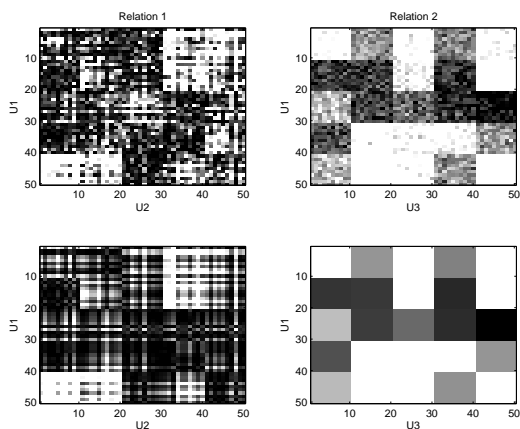


Figure 5.3: Multi-way clustering-based reconstruction with configuration A (bottom row) of relational data consisting of two 50x50 matrices (top row).

for both \mathbf{Z}^1 and \mathbf{Z}^2 . Table 5.1 shows the approximation error/clustering quality (as measured by normalized mutual information (nMI) [20]) with respect to the original matrices, clearly demonstrating that the right choice of Bregman divergence and clustering basis (i.e., configuration A) results in better performance. Figure 5.3 also shows the reconstruction for this case. These results suggest that one could use domain knowledge to identify a suitable generative distribution (e.g., Gaussian, Poisson) and summary statistics (e.g., entity class bias) and instantiate the multi-way clustering algorithm with the appropriate Bregman divergence and clustering basis.

5.2 Benefit of Tensor Structure: Newsgroup Topics. In this experiment, we demonstrate the improvement in clustering performance by considering the tensor structure. We consider a subset of the well-known 20 Newsgroups[9] data. Following the methodology of [16], we create a subset of 100 documents sampled randomly from 4 newsgroups – sci.space, sci.electronics, talk.politics.guns, talk.politics.mideast. For each article, we extracted information about the author, the newsgroup, and the words in the post. By aggregating across all the 400 documents in this collection, we created a tensor with 3 dimensions: author, newsgroup category, and words used. The goal of this experiment is to get good topics from word clusters. Note that in this dataset, the 4 newsgroups categories form a natural partitioning into 2 clusters: one related to science (sci.space and sci.electronics) and the other to politics (talk.politics.guns and talk.politics.mideast). So we clustered this tensor using BMC with I-divergence to get 2 newsgroup clusters, 20 author clusters, and 20 word clusters. We compare the performance with matrix co-

Tensor		Matrix	
Topic 1	Topic 2	Topic1	Topic2
arab	nasa	arab	rocket
land	earth	space	land
jewish	space	turkish	nasa
israel	rocket	israeli	earth
policy	international	israel	jewish
	program		president
	million		

Table 5.2: Two sample word clusters each from tensor co-clustering and matrix co-clustering, representing topics about Arab-Israel politics and NASA’s space program.

clustering using I-divergence, by considering only the word distributions in every newsgroup (no author information).

Both the algorithms were able to find the clustering of the newsgroup categories. However, the word clusters or topics obtained for the tensor structure were much better than those in the matrix structure. As seen in Table 5.2, the tensor structure enables the algorithm to find coherent topic words, whereas the topics obtained with the matrix structure have some noise words in them, e.g., “space” in topic about Arab-Israel politics, “jewish” and “president” in the topic about NASA’s space program.

5.3 Benefit of Relation Graph: Movie Categories. This experiment illustrates the utility of clustering over the relation graph structure. We consider the IMDB-EachMovie dataset [2], which has movies with descriptions from the IMDB dataset (actors, genre, etc.) and corresponding user ratings from the EachMovie dataset. We create a subset of 100 movies on which 39031 users have provided ratings on a scale of 1-5. For each movie, we consider the actors and the genre labels from IMDB. Note that movies can be associated with multiple genre labels. Altogether, the 100 movies have 19 different genre labels, and there are 1177 actors acting in them.

We created a relation graph comprising of two matrices hinged along the movie dimension: (movie, user ratings) and (movie, actor). Using BAMC along the user ratings matrix and BMC along the actor matrix, we partitioned the movies into 20 clusters. The quality of the clusters were evaluated using the pair-wise F-measure metric [2], which is a useful extrinsic cluster evaluation measure when data points have multiple class labels. For the dataset considered, the pairwise F-measure score was 0.4. As a baseline, we considered matrix co-clustering over the (movie, user ratings) matrix, using BAMC. The pairwise F-measure was a substantially lower value, 0.13, illustrating that using the relation graph structure over the multiple relations of the movie data gives better clustering results.

Algo	MRGC	COCLUST	SVD	PLSA	CORR
MAE	0.723	0.743	0.754	0.739	0.813

Table 5.3: Mean absolute error (MAE) on MovieLens for collaborative filtering approaches. For MRGC ($\nu^{genre} = 0.2$) and COCLUST, #clusters on each dimension (movie, user, genre) = 5, I-divergence loss is used. Rank of SVD and #topics in PLSA = 5, #neighbors in the correlation method = 50. Std. devn. in all results is below 0.02.

5.4 Augmented Collaborative Filtering Case Study.

The main task in collaborative filtering is to predict the preference of a given user for an item using known preferences of the other users. In this study, we explore the possibility of obtaining better predictions using additional information via multi-way clustering. We used a subset of the MovieLens dataset (movie-lens.umn.edu) consisting of 456 users, 600 movies (13158 ratings in the range [1-5]) as well as the movie-genre (19 genres) information. Our prediction approach involved performing multi-way clustering on a subset of the user-movie ratings as well as the movie-genre memberships in order to obtain a grouping of the users, movies and the genres. In particular, we used BAMC scheme for the ratings and BMC for the genre memberships with I-divergence as the loss function. The multi-way clusters as well as the summary statistics were then used to make predictions on the test set. To take into account missing values while clustering, we assumed a uniform measure for known values and zero measure for the missing ones as in [1]. We also tried varying the relative importance of the genre memberships and ratings by settings $\nu^{genre} = [0, 0.2, 0.4, 0.6, 0.8]$ on training sets with different levels of sparsity in the ratings. We find that the genre information tends to become more useful as the sparsity increases, in particular the optimal ν^{genre} (w.r.t test error) for 20%, 40%, 60% and 80% ratings was found to 0.6, 0.4, 0.2 and 0.2 respectively.

Table 5.3 shows the average mean absolute error on a 80%-20% train-test split of (456 users \times 500 movies) for the multi-way clustering approach as well as other collaborative filtering approaches such as SVD [19], PLSA [13], Pearson correlation and also co-clustering [1]. From the table, we note that the multi-way clustering provides slightly better accuracy than other methods because of the additional genre information. This becomes even more evident when we consider the prediction accuracy on ratings corresponding to the 100 movies that are not included in the clustering. In the multi-way clustering approach, the “unseen” movie is assigned to a cluster based solely on genre and the rating is predicted to be the user’s average rating for that movie cluster resulting in (MAE = 0.809), whereas in all the other approaches, the prediction for a “unseen” movie is just the user’s overall average rating that results in much higher error (MAE = 0.883).

6 Related Work

In this section, we briefly describe how the multi-way relation graph clustering (MRGC) model is related to other clustering models as well as graphical models that have been proposed in the literature.

6.1 Other Multi-way Clustering Models. Long et al. [16] recently proposed the collective factorization of related matrices (CFRM) model for joint clustering on multi-type relational data. The CFRM model is, however, restricted to only pairwise relations between entity types. As in the case of MRGC, the CFRM formulation is also motivated from an approximation point of view, but the distortion is measured using squared loss. Further, the optimal approximation is obtained via non-negative factorization of relevant matrices, unlike MRGC where the approximation is determined in terms of the conditional expectations via the MBI principle. It can be shown that the CFRM model corresponds to a real relaxation of a special case of MRGC model for the block clustering formulation (BMC) and squared loss. Though the spectral clustering algorithm is not scalable to large datasets, it can give better results than alternate minimization in some cases.

A follow-up paper by Long et al. [15] proposed a relational summary network (RSN) model for clustering over k -partite graphs using Bregman divergences. The RSN model considers only pairwise interactions, and in fact, follows as a special case of the MRGC model where all the relations are described by matrices and the summary statistics correspond to the block clustering (BMC) scheme.

Bekkerman et al. [4] introduce a model of multi-way distributional clustering (MDC) based on pairwise interactions between variables, which maximizes an objective function equal to the sum of weighted pairwise mutual information between the clustered random variables. The MDC formulation can be shown to be a special case of MRGC assuming only pairwise relations, summary statistics corresponding to the BAMC scheme and an I-divergence distortion measure. However, MDC uses an interleaved top-down and bottom-up algorithm, which can be more effective at performing correction for local optimization than the alternate minimization MRGC algorithm. Multivariate information bottleneck [10] is another important work that corresponds to a soft clustering formulation similar to MRGC for a specific class of loss functions based on I-divergence.

In addition to the clustering models discussed above, there are several other “multi-way” clustering techniques, that share the same nomenclature, but address slightly unrelated problems, some of which are discussed in [15, 16].

6.2 Matrix Co-clustering and KMeans. As mentioned in Section 3, our tensor clustering formulation is a direct generalization of Bregman matrix co-clustering [1]. In this work, Banerjee et al. [1] demonstrate how Bregman matrix co-clustering generalizes a variety of partitional co-clustering algorithms: information theoretic co-clustering [9], minimum sum-square residue clustering [8], fully-automatic cross-associations [7], as well as clustering algorithms such as KMeans. Hence, these methods can be considered as special cases of the MRGC model as well. There are several related co-clustering formulations that have been proposed in the literature — [17] contains an extensive survey on various models and their applications.

6.3 Probabilistic Graphical Models. The MRGC model is also closely related to probabilistic graphical models. More specifically, the bijection between exponential families and Bregman divergences [3] suggests that it is possible to have a probabilistic view of the multi-way clustering problem as maximum likelihood estimation over a structured mixture of exponential distributions. Based on this connection, it can be shown that the MRGC formulation is associated with a family of graphical models over the latent cluster variables where the dependencies are determined by the choice of clustering basis. This is similar in principle to the recent work on Bayesian models, e.g., LDA [5], which formulate the clustering problem in terms of a graphical model with latent variables. Several papers have proposed various specific model structures for clustering in particular domains [21]. The MRGC formulation, in particular, has several commonalities with relational clustering methods based on Probabilistic Relational Models (PRM) [11] and Probabilistic Entity-Relationship (PER) [12] models, which support probabilistic inference over relations.

7 Conclusions

In summary, our current work significantly expands the applicability of clustering techniques by providing a broad multi-way clustering framework with flexibility along two directions —(a) simultaneously clustering of multiple classes of entities, (b) extension to multiple relations via a relation-graph model. Furthermore, our formulation is applicable to all Bregman loss functions and also allows a large family of structurally different multi-way clustering schemes based on the summary statistics that need to be preserved. The proposed MRGC algorithm is also highly scalable and suitable for handling high dimensional, sparse data involving mixed data types. Our formulation can also be shown to have a natural interpretation in terms of probabilistic generative model, which in turn allows extensions to active, online and semi-supervised learning.

Acknowledgments We would like to thank Inderjit Dhillon, Joydeep Ghosh and the SRI iLink team for useful discussions. This work was supported partly by DARPA, Contract #NBCHD030010, Order #T310.

References

- [1] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. Technical report, University of Texas at Austin, 2004.
- [2] A. Banerjee, C. Krumpelman, S. Basu, R. Mooney, and J. Ghosh. Model-based overlapping clustering. In *KDD*, 2005.
- [3] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *JMLR*, 6:1705–1749, 2005.
- [4] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *ICML*, 2005.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- [6] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1998.
- [7] D. Chakrabarti, S. Papadimitriou, D. S. Modha, and C. Faloutsos. Fully automatic cross-associations. In *KDD*, 2004.
- [8] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*, 2004.
- [9] I. Dhillon, S. Mallela, and D. Modha. Information-theoretic co-clustering. In *KDD*, 2003.
- [10] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *UAI*, 2001.
- [11] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, 1999.
- [12] D. Heckerman, C. Meek, and D. Koller. Probabilistic entity-relationship models, PRMs and plate models. In *SRL Workshop, ICML*, 2004.
- [13] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115, 2004.
- [14] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.
- [15] B. Long, X. Wu, Z. Zhang, and P. S. Yu. Unsupervised learning on k-partite graphs. In *KDD*, 2006.
- [16] B. Long, Z. Zhang, X. Wu, and P. S. Yu. Spectral clustering for multi-type relational data. In *ICML*, 2006.
- [17] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Trans. Computational Biology and Bioinformatics*, 1(1):24–45, 2004.
- [18] S. Merugu, A. Banerjee, and S. Basu. Multi-way clustering on relation graphs. Technical report, Yahoo! Research, October 2006.
- [19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender systems—a case study. In *WebKDD Workshop, KDD*, 2000.
- [20] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop of AI for Web Search, AAAI*, 2000.
- [21] X. Wang, N. Mohanty, and A. McCallum. Groups and topic discovery from relations and text. In *LinkKDD Workshop, KDD*, 2005.