SANDIA REPORT

SAND2006-7744
Unlimited Release
Printed December 2006

Supersedes SAND2006-2161
Dated June 2006

Pattern Analysis of Directed Graphs Using DEDICOM: An Application to Enron Email

Brett W. Bader, Richard A. Harshman, and Tamara G. Kolda

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2006-7744
Unlimited Release
Printed December 2006

Supersedes SAND2006-2161
dated June 2006

Pattern Analysis of Directed Graphs Using DEDICOM: An Application to Enron Email

Brett W. Bader
Applied Computational Methods Department
P.O. Box 5800
Sandia National Laboratories
Albuquerque, NM 87185-1318

Richard A. Harshman
Department of Psychology
University of Western Ontario
London, Ontario, Canada N6A 5C2

Tamara G. Kolda
Computational Science and Mathematics Research Department
Sandia National Laboratories
P.O. Box 969
Livermore, CA 94550-9159

Abstract

DEDICOM is a linear algebra model for analyzing intrinsically asymmetric relationships, such as trade among nations or the exchange of emails among individuals. DEDICOM decomposes a complex pattern of observed relations among objects into a sum of simpler patterns of inferred relations among latent components of the objects. Three-way DEDICOM is a higher-order extension of the model that incorporates a third mode of the data, such as time, giving it stronger uniqueness properties and consequently enhancing interpretability of

solutions. In this paper, we present algorithms for computing these decompositions on large, sparse data as well as a variant for computing an asymmetric nonnegative factorization. When we apply these techniques to adjacency arrays arising from directed graphs with edges labeled by time, we obtain a smaller graph on latent semantic dimensions and gain additional information about their changing relationships over time. We demonstrate these techniques on the Enron email corpus to learn about the social networks and their transient behavior. The mixture of roles assigned to individuals by DEDICOM showed strong correspondence with known job classifications and revealed the patterns of communication between these roles. Changes in the communication pattern over time, e.g., between top executives and the legal department, were also apparent in the solutions.

Contents

Figures

Tables

1 Introduction

Often it is useful to distill a large amount of data down to a manageable size to facilitate interpretation, and our goal is to do this by uncovering latent profiles and their asymmetric interrelationships. Existing data-analytic models and methods do not generally let one seek out and describe patterns of asymmetric relationships in a dataset. This paper shows how a family of models called DEDICOM (DEcomposition into DIrectional COMponents) from the psychometrics literature [13] can provide information on latent components in data and the pattern of asymmetric (i.e., directed) relationships among these components.

In this paper, DEDICOM is used to interpret directed semantic graphs (i.e., graphs with labeled edges) arising from email communications at Enron. A contribution of this paper is that we provide algorithmic improvements that enable DEDICOM to be applied to large-scale data. We also present a nonnegative factorization of DEDICOM for capturing asymmetric interrelationships.

In the general case, we consider a directed graph with n vertices whose square adjacency matrix $\mathbf{X}$ contains a nonzero entry x_{ij} for each edge (i, j). The single-domain DEDICOM model applied to $\mathbf{X}$ is an approximation

$$\mathbf{X} \approx \mathbf{A}\mathbf{R}\mathbf{A}^T, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix of loadings or "weights" for the n vertices on $p < n$ dimensions and $\mathbf{R} \in \mathbb{R}^{p \times p}$ is a matrix that captures the asymmetric relationships on these latent dimensions of $\mathbf{A}$; see Figure 1. The dual-domain DEDICOM model has a different matrix on the left and right of $\mathbf{R}$ but is not considered in this paper.

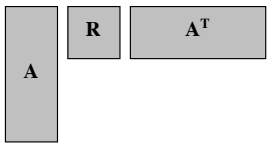

Figure 1. Two-way DEDICOM model.

The DEDICOM model can be extended to three-way data, and here we use time as the third mode. If our graph has m discrete time edge labels, then we can construct an adjacency matrix $\mathbf{X}_k$ for each edge type, $k = 1 \ldots m$, and store them as an array $\mathcal{X} \in \mathbb{R}^{n \times n \times m}$. The three-way DEDICOM model for $\mathcal{X}$ is

$$\mathbf{X}_k \approx \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^T \quad \text{for } k = 1, \ldots, m, \tag{2}$$

where $\mathbf{X}_k$ is the kth adjacency matrix in $\mathcal{X}$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix of loadings, $\mathbf{D}_k$ is a diagonal matrix that gives the weights of the columns of $\mathbf{A}$ for each level in the

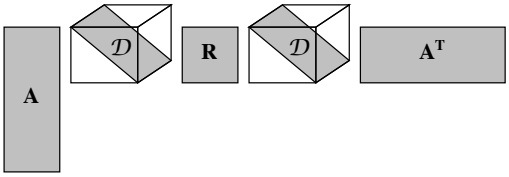

Figure 2. Three-way DEDICOM model.

third mode, and $\mathbf{R} \in \mathbb{R}^{p \times p}$ is the asymmetry matrix; see Figure 2. The matrix $\mathbf{R}$ captures the aggregate trends over time and, when multiplied on the left and right by $\mathbf{D}_k$, within a particular time period as well. The array $\mathcal{D}$ is the collection of matrices $\mathbf{D}_k$. In variations of this model, the scaling array $\mathcal{D}$ and/or loadings matrix $\mathbf{A}$ may be different on the left and right of $\mathbf{R}$.

A simplified interpretation of DEDICOM is that it takes a large array and condenses the interrelationships into an idealized summary in the $\mathbf{R}$ matrix. Rows of $\mathbf{A}$ correspond to nodes (i.e., individual people in a social network) and can have substantial weights in more than one of the latent components, which can be regarded as roles. For example, an individual at a company might have characteristics that cause their pattern of email exchanges to look like a mixture of two different idealized patterns, such as an executive and a lawyer.

We focus on email graphs, which are useful for social network analysis. Specifically, we consider the email corpus of the Enron corporation that was made public by the U.S. Federal Energy Regulatory Commission (FERC) during its investigation of Enron. Our research uses a simplified version of the database by Priebe et al. [29] that includes only email exchanges among 184 email addresses.

To help explain the temporal patterns in social networks, this paper seeks to analyze semantic graphs where the communications are discretized by time. In the Enron data, we arrange the data in a three-way array such that each slice corresponds to a particular month and year. This representation can be interpreted as a graph where each edge is labeled by the time period. This representation has also been called a time graph [25] or a time series of graphs [28].

The paper is organized as follows. In section 2, we discuss work connected with the Enron corpus and past research on the DEDICOM models. In section 3, we describe new algorithms for computing the two- and three-way DEDICOM models. Section 4 describes the Enron corpus we used. We apply the algorithms to the Enron corpus in section 5 and discuss some conclusions in section 6.

2 Related Work

We mention relevant work from the psychometrics community for further background on DEDICOM and multi-way models. We also outline related work in social network analysis, mostly pertaining to the Enron data.

2.1 DEDICOM and multi-way models.

The DEDICOM family of models was first introduced in [13]. One of the earliest applications of DEDICOM studied the asymmetries in telephone calls among cities. Later, it was developed as a tool for analyzing asymmetric relationships that arise in marketing research [14]. While research on the model continued in the psychometrics and multilinear algebra communities, its influence did not spread far outside of these communities.

There has been some research of the model and associated applications [15] followed by a number of papers analyzing algorithms for computing the DEDICOM model [22], including variations such as constrained DEDICOM [21, 30] and three-way DEDICOM [20]. Most of the applications of DEDICOM in the literature have focused on two-way (matrix) data, and there is even less research in the three-way (tensor) case. One of the first applications involving three-way data provided asymmetric measures of world trade (import-export matrices) among a set of nations considered over a period of 10 years [16].

The use of multi-way models is relatively new in the context of data mining and has appeared recently in some web applications. Sun et al. [35] apply a three-way Tucker decomposition [37] to the analysis of (user $\times$ query term $\times$ web page) data in order to personalize web search. In [1], various tensor decompositions of (user $\times$ key word $\times$ time) data are used to separate different streams of conversation in chatroom data. In [24, 23] a PARAFAC decomposition [12] (also known as CANDECOMP [6]) is applied to (web page $\times$ web page $\times$ anchor text) data, forming a sparse, three-way tensor representing the web graph with anchor-text-labeled edges. This was the first use of PARAFAC for analyzing semantic graphs as well as the first instance of applying PARAFAC to sparse data. The history of tensor decompositions in general goes back forty years [37, 12, 6], and they have been used extensively in other domains, especially chemometrics [34].

2.2 Enron data and social network analysis.

Research on the Enron corpus falls into several broad areas, including social network analysis, graph theoretics, and natural language processing. Initial studies on the database itself focused on the statistical and graph theoretic properties. Shetty and

Adibi [32] constructed a MySQL database of the corpus to facilitate a statistical analysis of the data. Their results show the distribution of emails per user, sent emails per user, and emails over time. They also derived a social network involving 151 employees of Enron by assigning a social contact if at least five bi-directional messages connect two employees and categorizing each employee by their management level.

More recently, there has been research on the database from a network analytic perspective. This includes analyzing the social networks detectable in the email graph. Diesner and Carley [10] show that the communication network was denser, more centralized, and more connected during the crisis than during normal times. Their analysis also shows that during the crisis, communication among Enron's employees was more likely to be exchanged between employees in different positions, except among the top executives, who had formed a tight clique.

Chapanond, Krishnamoorthy, and Yener [7] analyzed the Enron corpus for structures within the organization. They used both graph theoretical and spectral analysis techniques to identify communities.

McCallum, Corrada-Emmanuel, and Wang [27] proposed the Author-Recipient-Topic (ART) model for social network analysis. ART is a Bayesian network for social network analysis that builds on Latent Dirichlet Allocation and the Author-Topic model. They use ART on the email corpus to learn discussion topics based on the directed interactions and relationships between people and their communications.

Berry and Browne [5] apply nonnegative matrix factorizations to discover concepts and topics in the Enron corpus. They discuss results of topic detection and message clustering in the context of published Enron business practices and activities.

Keila and Skillicorn [19] have investigated structures in the Enron corpus using singular value decomposition and semidiscrete decomposition. They present relationships among individuals based on their patterns of word use in email and word frequency profiles. They present a case that word use among those with alleged criminal activity may be "slightly distinctive."

Priebe, Conroy, Marchette, and Park [28] introduced a theory of scan statistics on graphs and applied them to the problem of anomaly detection using a time series of Enron email graphs.

Sarkar and Moore [31] proposed a method for the dynamic analysis of social networks. They embed an evolving friendship graph in p dimensional space using multidimensional scaling and allow entities to move in this space over time.

3 DEDICOM Models and Algorithms

3.1 Notation.

We use the following notation. Scalars are denoted by lowercase letters, e.g., a. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{a}$. The ith entry of $\mathbf{a}$ is denoted by a_i. Matrices are denoted by boldface capital letters, e.g., $\mathbf{A}$. The jth column of $\mathbf{A}$ is denoted by $\mathbf{a}_j$ and element (i, j) by a_{ij}.

Multi-way arrays and tensors are denoted by boldface Euler script letters, e.g., $\boldsymbol{\mathcal{X}}$. Element (i, j, k) of a third-order tensor $\boldsymbol{\mathcal{X}}$ is denoted by x_{ijk}, and the kth slice of $\boldsymbol{\mathcal{X}}$ is denoted by $\mathbf{X}_k$ (i.e., a matrix formed by holding the last index of $\boldsymbol{\mathcal{X}}$ fixed at k).

The symbol $\otimes$ denotes the Kronecker product, and the symbol $*$ denotes the Hadamard (i.e., elementwise) matrix product. The Frobenius norm of a matrix, $\| \mathbf{Y} \|_F$, is the square root of the sum of squares of all its elements.

3.2 Two-way DEDICOM.

In the two-way case, a square matrix $\mathbf{X}$ is decomposed according to (1). The goal is to find the best-fitting matrices $\mathbf{A}$ and $\mathbf{R}$ in the minimization problem

$$\min_{\mathbf{A},\mathbf{R}} \left\| \mathbf{X} - \mathbf{A}\mathbf{R}\mathbf{A}^T \right\|_F^2 \tag{3}$$

subject to $\mathbf{A}$ having orthogonal columns.

There are two indeterminacies of scale and rotation that need to be addressed. First, the columns of $\mathbf{A}$ may be scaled in a number of ways without affecting the solution. We scale them to have unit length in the 2-norm. Other choices discussed in [14] provide other benefits for interpreting the results. Second, the matrix $\mathbf{A}$ can be transformed with any nonsingular matrix $\mathbf{Q}$ with no loss of fit to the data because $\mathbf{A}\mathbf{R}\mathbf{A}^T = (\mathbf{A}\mathbf{Q})(\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-T})(\mathbf{A}\mathbf{Q})^T$. Thus, the solution obtained in $\mathbf{A}$ is not unique. Nevertheless, it is standard practice to apply some accepted rotation to "fix" $\mathbf{A}$. We will adopt VARIMAX rotation [18] such that the variance across columns of $\mathbf{A}$ is maximized.

A further practice in some applications is to ignore the diagonal entries of $\mathbf{X}$ in the residual calculation. For our case, this makes sense because we wish to ignore self-loops (i.e., sending email to yourself). We use the common technique of estimating the diagonal values from the current approximation $\mathbf{A}\mathbf{R}\mathbf{A}^T$ at each iteration and including them in $\mathbf{X}$.

The principle challenge in computing the model lies with finding $\mathbf{A}$ because computing $\mathbf{R}$ is a simple least squares problem. The original algorithm proposed in [14]

used the singular value decomposition (SVD) of $\mathbf{X}$ to provide an approximate minimizer of (3). Subsequent research has been aimed at computing an exact minimizer; see [2] for a brief history. Unfortunately, these algorithms are not adaptable large-scale data. In 1990, Kiers et al. [22] proposed a modification and generalization of Takane's algorithm [36] that had a convergence guarantee.

While the modified Takane method [22] for finding $\mathbf{A}$ could work with large-scale data, we describe an alternating least squares (ALS) algorithm to motivate our new three-way algorithm, for which there is no good alternative for large-scale problems. We do not yet have a proof of convergence, but we have observed good performance in practice.

3.2.1 Alternate ALS algorithm.

We start with random initial estimates for $\mathbf{A}$ and $\mathbf{R}$ and write a model that solves for $\mathbf{A}$ on both the left and right simultaneously. We consider a model for $\mathbf{X}$ and $\mathbf{X}^T$ together by stacking the data side by side:

$$\begin{pmatrix}\mathbf{X} & \mathbf{X}^T\end{pmatrix} = \mathbf{A}\left[\begin{pmatrix}\mathbf{R} & \mathbf{R}^T\end{pmatrix}\begin{pmatrix}\mathbf{A}^T & 0\\ 0 & \mathbf{A}^T\end{pmatrix}\right]. \tag{4}$$

The least squares update for $\mathbf{A}$ is found by fixing the $\mathbf{A}$ in the right-hand matrix and solving for the leftmost $\mathbf{A}$. For large applications, we wish to avoid forming any large matrices or computing the pseudo-inverse of a large rectangular matrix. Hence, we may find the least squares solution more easily by using the normal equations and simplifying terms:

$$\mathbf{A} \leftarrow \left(\mathbf{X}\mathbf{A}\mathbf{R}^T + \mathbf{X}^T\mathbf{A}\mathbf{R}\right)\left(\mathbf{R}\mathbf{A}^T\mathbf{A}\mathbf{R}^T + \mathbf{R}^T\mathbf{A}^T\mathbf{A}\mathbf{R}\right)^{-1}.$$

Using the most recent approximation for $\mathbf{A}$, we can compute a least squares estimate of $\mathbf{R}$ by multiplying both sides of $\mathbf{X}$ by the pseudo-inverse of $\mathbf{A}$, using the normal equations:

$$\mathbf{R} \leftarrow (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}. \tag{5}$$

Thus, with these two update rules, one may alternatingly solve for $\mathbf{A}$ and $\mathbf{R}$ to arrive at a DEDICOM model that best fits the data in a least-squares sense.

The most expensive operations are the matrix products $\mathbf{X}\mathbf{A}$, $\mathbf{X}^T\mathbf{A}$, and $\mathbf{A}^T\mathbf{A}$. The inverted matrix is $p \times p$ so it is inconsequential provided that $p \ll n$. If $\mathbf{X}$ is sparse, then the computations involving $\mathbf{X}$ are proportional to the number of nonzeros in $\mathbf{X}$, and $\mathbf{A}^T\mathbf{A}$ is $\mathcal{O}(p^2 n)$.

3.3 Three-way DEDICOM.

Three-way DEDICOM is similar to the two-way model in that the asymmetry relationships are in a matrix $\mathbf{R}$, but there are diagonal scaling matrices (represented as frontal slices of array $\mathcal{D}$) on either side that apply weights to the columns of $\mathbf{A}$. In variations of this model, the scaling arrays on the left and right of $\mathbf{R}$ may be different.

Three-way DEDICOM is a part of a broader family of multilinear models called PARATUCK2 [17], which can empirically determine a unique best fitting axis orientation in $\mathbf{A}$ without the need for a separate factor rotation. This corresponds to the extension of factor analysis to three ways by PARAFAC [12] and the same kind of special uniqueness property that emerges there. With a unique solution, the factors are plausibly a valid description with greater reason to believe that they have more explanatory meaning than a "nice" rotated two-way solution.

We wish to solve the following minimization problem

$$\min_{\mathbf{A},\mathbf{R},\mathcal{D}} \sum_{k=1}^{m} \left\| \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^T \right\|_F^2, \tag{6}$$

where $\mathbf{A}$ is not required to be orthogonal. Because the $\mathbf{A}$ and $\mathbf{R}$ matrices apply across all frontal slices of $\mathcal{X}$, algorithms are more complicated than the standard DEDICOM model.

There are few algorithms for solving (6); in addition, these algorithms are not efficient with large, sparse datasets. Kiers [20] has presented an ALS algorithm for three-way DEDICOM. To update $\mathbf{A}$, Kiers minimizes (6) over the columns of $\mathbf{A}$, updating each column with its own ALS subproblem. Each subproblem to compute a column of $\mathbf{A}$ involves an eigendecomposition of a dense $n \times n$ matrix, which makes this procedure prohibitively expensive for large-scale data. Determining whether this procedure can be adapted for large-scale data is a topic of future study. To update $\mathcal{D}$, Kiers solves for each element of $\mathcal{D}$ with an ALS procedure. The update of $\mathbf{A}$ and the elementwise update of $\mathcal{D}$ are also not adapted for large-scale applications.

3.3.1 New ALS algorithm.

Here we propose an alternating least squares algorithm and adapt it for use on larger applications. Our approach offers improvements over Kiers' method for updating $\mathbf{A}$ and $\mathcal{D}$ as well as a compression technique for dealing with large-scale data.

Once again, we start with random initializations for $\mathbf{A}$, $\mathbf{R}$, and now $\mathcal{D}$. We update these quantities in an alternating fashion as follows.

1. Updating $\mathbf{A}$: We write a model that solves for $\mathbf{A}$ on both the left and the right and for all frontal slices of $\mathcal{D}$ simultaneously. We consider all frontal slices of

$\mathfrak{X}$ by stacking the data side by side:

$$\begin{aligned}\begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_1^T & \cdots & \mathbf{X}_m & \mathbf{X}_m^T\end{pmatrix} = \\ \mathbf{A}\begin{pmatrix}\mathbf{D}_1\mathbf{R}\mathbf{D}_1 & \mathbf{D}_1\mathbf{R}^T\mathbf{D}_1 & \cdots \\ & \mathbf{D}_m\mathbf{R}\mathbf{D}_m & \mathbf{D}_m\mathbf{R}^T\mathbf{D}_m\end{pmatrix}\left(\mathbf{I}_{2m}\otimes\mathbf{A}^T\right). \quad (7)\end{aligned}$$

Here $\mathbf{I}_{2m}$ is the identity matrix of size $2m \times 2m$. We compute the least squares solution for the $\mathbf{A}$ on the left using the method of normal equations, which simplifies to

$$\mathbf{A} \leftarrow \left[\sum_{k=1}^{m}\left(\mathbf{X}_k\mathbf{A}\mathbf{D}_k\mathbf{R}^T\mathbf{D}_k + \mathbf{X}_k^T\mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\right)\right] \left[\sum_{k=1}^{m}(\mathbf{B}_k + \mathbf{C}_k)\right]^{-1}$$

where

$$\mathbf{B}_k \equiv \mathbf{D}_k\mathbf{R}\mathbf{D}_k(\mathbf{A}^T\mathbf{A})\mathbf{D}_k\mathbf{R}^T\mathbf{D}_k, \quad (8)$$
$$\mathbf{C}_k \equiv \mathbf{D}_k\mathbf{R}^T\mathbf{D}_k(\mathbf{A}^T\mathbf{A})\mathbf{D}_k\mathbf{R}\mathbf{D}_k. \quad (9)$$

This least squares problem updates all columns of $\mathbf{A}$ simultaneously and avoids the costly eigendecomposition of Kiers' method.

2. Updating $\mathbf{R}$: We use the closed form solution for $\mathbf{R}$ from Kiers [20]. It involves vectorizing $\mathfrak{X}$ and $\mathbf{R}$ and stacking them in a manner such that the objective function in (6) changes to

$$f(\mathbf{R}) = \left\| \begin{pmatrix}\mathrm{Vec}(\mathbf{X}_1)\\ \vdots \\ \mathrm{Vec}(\mathbf{X}_m)\end{pmatrix} - \begin{pmatrix}\mathbf{A}\mathbf{D}_1\otimes\mathbf{A}\mathbf{D}_1\\ \vdots \\ \mathbf{A}\mathbf{D}_m\otimes\mathbf{A}\mathbf{D}_m\end{pmatrix}\mathrm{Vec}(\mathbf{R})\right\|.$$

Minimizing $f(\mathbf{R})$ over $\mathrm{Vec}(\mathbf{R})$ is a multiple regression problem, and its solution is

$$\begin{aligned}\mathrm{Vec}(\mathbf{R}) &= \left(\sum_{k=1}^{m}(\mathbf{D}_k\mathbf{A}^T\mathbf{A}\mathbf{D}_k)\otimes(\mathbf{D}_k\mathbf{A}^T\mathbf{A}\mathbf{D}_k)\right)^{-1} \\ &\quad \sum_{k=1}^{m}\mathrm{Vec}(\mathbf{D}_k\mathbf{A}^T\mathbf{X}_k\mathbf{A}\mathbf{D}_k). \quad (10)\end{aligned}$$

Provided that the number of latent dimensions is not large (specifically that p^2 is not large), then this step for updating $\mathbf{R}$ will suffice.

3. Updating $\mathcal{D}$: We improve upon the elementwise minimization of Kiers [20] by considering a full-scale minimization with respect to the diagonal elements for each slice $\mathbf{D}_k$:

$$\min_{\mathbf{D}_k} \left\| \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^T \right\|_F^2 . \tag{11}$$

Because there are only p variables for each of the m slices, Newton's method applied to (11) is not expensive and offers fast quadratic convergence. The gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ of (11) are provided in an earlier technical report of this work [2]. Extra conditions are needed to ensure that the Newton step is a descent direction, and we use a modified Cholesky decomposition of $\mathbf{H}$ to find the matrix $\mathbf{H}+\lambda\mathbf{I}$ that is safely positive definite for the Newton step calculation; see, e.g., [9].

A comment for large data sets is in order. The steps for updating $\mathbf{R}$ and $\mathcal{D}$ can be expensive if n is large. However, we may simplify the complexity by projecting the data in $\mathcal{X}$ onto a basis of $\mathbf{A}$ and working in this space. Specifically, we find an orthonormal basis $\mathbf{Q} \in \mathbb{R}^{n\times p}$ of matrix $\mathbf{A}$ using, e.g., a compact QR decomposition,

$$\mathbf{A} = \mathbf{Q}\tilde{\mathbf{A}}, \tag{12}$$

where $\tilde{\mathbf{A}}$ is upper triangular. Then we use $\mathbf{Q}$ to project $\mathcal{X}$ onto the basis of $\mathbf{A}$. By the orthogonality of $\mathbf{Q}$, the minimization problem of (11) is the same as

$$\min_{\mathbf{D}_k} \left\| \mathbf{Q}^T\mathbf{X}_k\mathbf{Q} - \tilde{\mathbf{A}}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\tilde{\mathbf{A}}^T \right\|_F^2 , \tag{13}$$

except that $\mathbf{Q}^T\mathbf{X}_k\mathbf{Q}$ and $\tilde{\mathbf{A}}$ are both of size $p \times p$. We use these smaller matrices in place of $\mathbf{X}_k$ and $\mathbf{A}$, respectively, in the updates of both $\mathbf{R}$ and $\mathcal{D}$ in (10) and (11) above.

We do not have yet a proof of convergence for this new algorithm, but this algorithm was tested on synthetic data constructed to contain known structure. Arrays of up to size $50 \times 50 \times 45$ were constructed using $p = 2$ to 6 latent components in $\mathbf{A}$. An asymmetric $\mathbf{R}$ matrix and diagonal $\mathbf{D}_k$ matrices were generated randomly to relate the patterns. When these $\mathcal{X}$ arrays were analyzed from a number of random starting positions, the global optimum was found among a number of minimizers. The global optimum always revealed the original patterns used to create the data, up to permutation of column order and multiplication of columns by scaling constants.

Note that the accurate recovery of built-in structure occurred without rotation, and was more exact than would typically be obtained by rotation methods. This is because the three-way solution is essentially fully identified without side conditions, i.e., is "essentially unique" [17]. Thus, when the systematic structure in the data is reasonably well approximated by the DEDICOM model, the uniqueness property increases the probable correspondence between the recovered patterns and the original empirical source patterns.

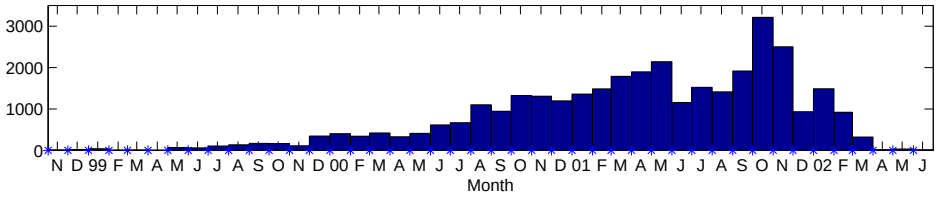

Figure 3. Number of emails per month in our Enron email graph.

The dominant costs of this ALS algorithm are linear in the number of nonzeros of $\mathbf{X}_k$ and/or $\mathcal{O}(p^2 n)$ and come from the following steps: $\mathbf{A}^T\mathbf{A}$, QR factorization of $\mathbf{A}$, $\mathbf{X}_k\mathbf{A}\mathbf{R}^T$, $\mathbf{X}_k^T\mathbf{A}\mathbf{R}$, and $\mathbf{Q}^T\mathbf{X}_k\mathbf{Q}$.

3.3.2 Nonnegative algorithm.

We also considered a DEDICOM model with non-negativity constraints on $\mathbf{A}$, $\mathbf{R}$, and $\mathcal{D}$. Because we are dealing with nonnegative data in $\mathcal{X}$, it often helps to examine decompositions that retain the nonnegative characteristics of the original data. Modifications to the least squares steps above are needed, and we use the multiplicative update introduced in [26] as implemented in [5]. Specifically, we modify the step to solve for the $\mathbf{A}$ appearing on the left in (7):

$$a_{ic} \leftarrow a_{ic} \frac{\left[\sum_{k=1}^{m} \left(\mathbf{X}_k\mathbf{A}\mathbf{D}_k\mathbf{R}^T\mathbf{D}_k + \mathbf{X}_k^T\mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\right)\right]_{ic}}{\left[\mathbf{A}\sum_{k=1}^{m}(\mathbf{B}_k + \mathbf{C}_k)\right]_{ic} + \epsilon}.$$

where $\mathbf{B}_k$ and $\mathbf{C}_k$ are the same as in (8)-(9) above and ϵ is a small number like 10^{-9}. The solution for $\mathbf{R}$ is given by:

$$\mathrm{Vec}(\mathbf{R})_i \leftarrow \frac{\mathrm{Vec}(\mathbf{R})_i \left[\sum_{k=1}^{m} \mathrm{Vec}(\mathbf{D}_k\mathbf{A}^T\mathbf{X}_k\mathbf{A}\mathbf{D}_k)\right]_i}{\left[\sum_{k=1}^{m}(\mathbf{D}_k\mathbf{A}^T\mathbf{A}\mathbf{D}_k) \otimes (\mathbf{D}_k\mathbf{A}^T\mathbf{A}\mathbf{D}_k)\mathrm{Vec}(\mathbf{R})\right]_i + \epsilon}.$$

We used the same procedure for updating $\mathcal{D}$ as above, using the same nonnegativity constraints.

A nonnegative two-way DEDICOM algorithm follows directly from this algorithm when one considers a matrix $\mathbf{X}$ as an array $\mathcal{X}$ having a single slice ($m = 1$) and the $\mathcal{D}$ array is just the identity matrix.

4 Enron Corpus

For a relevant application, we consider the email graph of the Enron corporation that was made public during the federal investigation.

The whole collection is available online [8] and contains 517,431 emails stored in the mail directories of 150 users. We use a smaller graph of the Enron email corpus prepared by Priebe et al. [29] that consists of messages solely among 184 Enron email addresses. We considered messages only in the interval 13-Nov-1998 through 21-Jun-2002, which resulted in a total of 34,427 messages over 44 months. Figure 3 shows a histogram of the number of messages in our data for each month.

We constructed an email graph and labeled the edges by month and year of the message. Our final graph corresponds to a sparse adjacency array $\mathcal{X}$ of size $184 \times 184 \times 44$ with 9838 nonzeros and, when ignoring edge labels, to a sparse adjacency matrix $\mathbf{X}$ of size 184×184 with 3960 nonzeros. We scaled the nonzeros entries by $\log_2(w) + 1$, where w is the number of messages. This simple weighting reduces the biasing from prolific emailers; other weightings produced similar results.

An obvious difficulty in dealing with the Enron corpus is the lack of information regarding the former employees. Without access to a corporate directory or organizational chart at Enron at the time of these emails, it is difficult to ascertain the validity of our results and assess the performance of the DEDICOM model. Other researchers using the Enron corpus have had this same problem, and information on the participants has been collected slowly and made available.

The Priebe data set [29] provided partial information on the 184 employees of the small Enron network, which appears to be based largely on information collected by Shetty and Adibi [33]. It provides most employees' position and business unit. To facilitate a better analysis of the DEDICOM results, we collected extra information on the participants from the email messages themselves and found some relevant information posted on the FERC website [11]. We searched for corroborating information of the preexisting data or for new identification information, such as title, business unit, or manager to help assess our results.

We labeled each of the 184 individuals according to the following five categories: executive (56), legal (15), pipeline (13), energy trader (29), and unaffiliated (71). Executives were considered as director level and higher. Legal employees were from the legal department in Enron North America (ENA). Pipeline employees were mainly those from the Transwestern Pipeline Company, a division of Enron Transportation Services (ETS). Energy traders were those individuals who traded gas or electricity in energy markets. The unaffiliated category were those employees for whom we had very little information and were largely unknown. The executive label took precedence over any of the others (e.g., the VP of Legal would be an "executive"). We will see that DEDICOM is able to align employees according to their business

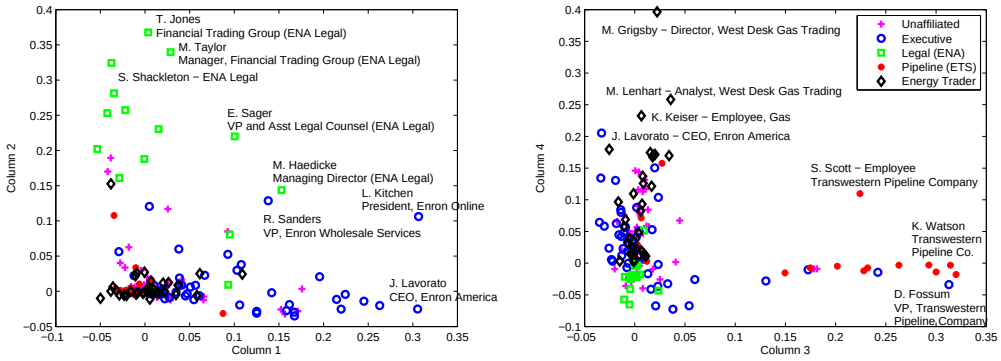

Figure 4. Two-way DEDICOM: plots of the first and second columns of **A** (left plot) and third and fourth columns of **A** (right plot) .

unit and identify many of these dual roles.

5 Experimental Results

In this section we summarize our findings of applying two-way and three-way DEDICOM on the Enron email network. Our algorithms were written in MATLAB, using sparse extensions of the Tensor Toolbox [3, 4]. All tests were performed on a dual 3GHz Pentium Xeon desktop computer with 2GB of RAM.

5.1 Two-way DEDICOM.

We computed a four-component ($p = 4$) decomposition of $\mathbf{X}$ using two-way DEDICOM. An iteration of two-way DEDICOM took about 0.06 seconds, and each run was 6–100 iterations. The apparent global minimizer was found in roughly 1 out of 4 runs from random initializations, and the relative norm of the difference was 0.768 (excluding diagonal).

Figure 4 shows the four columns of the $\mathbf{A}$ matrix in two plots: $\mathbf{a}_2$ versus $\mathbf{a}_1$ and $\mathbf{a}_4$ versus $\mathbf{a}_3$. DEDICOM clearly separates the employees by their roles within the organization without knowledge of the labels we have assigned the employees. The clustering of blue circles (executives) in the left plot of Figure 4 shows that the first factor ($\mathbf{a}_1$) describes an "executive" role that fits many of the top executives, and the second factor ($\mathbf{a}_2$) describes a "legal" role by the clustering of green squares. Some employees are described by both of these roles because they lie somewhere inbetween the two axes. For example, M. Haedicke was the managing director of ENA Legal, and M. Sanders was the VP of Enron Wholesale Services. Both of these individuals fall on a 45° line between the two roles, indicating that they shared a legal and executive role in their job, based on their communications.

In the right plot, the third factor ($\mathbf{a}_3$) describes employees in the pipeline business, and the fourth factor ($\mathbf{a}_4$) describes employees who are energy traders. We will call them "pipeline" and "trading" roles, respectively. One pipeline employee (S. Scott) stands out as having a partial role as an energy trader, which is information that is not apparent from his job title.

A key feature of the DEDICOM model that is not present in other models, such as PCA or MDS, is that the $\mathbf{R}$ matrix provides asymmetric relationships (i.e., directional information) between the latent roles in $\mathbf{A}$. The large adjacency matrix $\mathbf{X}$, showing nonsymmetric relations among employees at Enron, related by flows of email, is condensed into a smaller matrix $\mathbf{R}$ giving the same kind of asymmetric relations but among "types" or abstract idealized individuals. In this case, the relations among elements in $\mathbf{R}$ are exchanges of email. The latent components are patterns of the same kind of flow as among the surface objects, just abstracted into a "higher level" summary of patterns.

Figure 5 shows the $\mathbf{R}$ matrix and its corresponding graph representation for the

	#1	#2	#3	#4
#1 (Executive)	70.1	11.1	4.8	17.4
#2 (Legal)	10.8	55.9	0.6	4.8
#3 (Pipeline)	6.8	1.3	52.2	0.9
#4 (Trading)	6.4	2.4	1.7	51.8

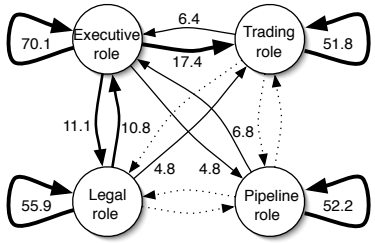

Figure 5. Two-way DEDICOM: **R** matrix and associated graph showing aggregate communication patterns. Entries less than 4 are represented as a dotted line, and dominant communications have thicker edges.

four-component solution. The **R** matrix shows that most of the communication is among employees who share the same role, as evidenced by the large diagonal values in **R**. We do see some asymmetric communication. The entries in the **R** matrix are mostly symmetric, with the exception of roles 1 and 4. The large value of $r_{1,4} = 17.4$ and smaller value $r_{4,1} = 6.4$ suggests that more information is flowing from the executive management to the energy traders than in the reverse direction.

5.2 Three-way DEDICOM.

We computed a four-component ($p = 4$) decomposition of the adjacency array $\mathcal{X}$ using three-way DEDICOM. An iteration of three-way DEDICOM took about 0.85 seconds, requiring an average of 184 iterations to satisfy a tolerance of 10^{-7} in the change of fit. This was a more difficult optimization problem, and an apparent global minimizer was not revealed. We chose the smallest minimizer from among 40 runs starting from random initializations. The relative norm of the difference was 0.885 (excluding diagonal).

Figure 6 plots the four columns of the **A** matrix. The three-way solution has less scatter in **A** than two-way DEDICOM, and the employees tend to line up on a single latent dimension corresponding to their role. This is due to the fact that each latent dimension in three-way DEDICOM is associated with a profile over time, so the roles it identifies tend to be more specific with less dual participation.

The first column is the legal role, and the second column identifies executives who deal with government and regulatory affairs. The third role is the top executives, and

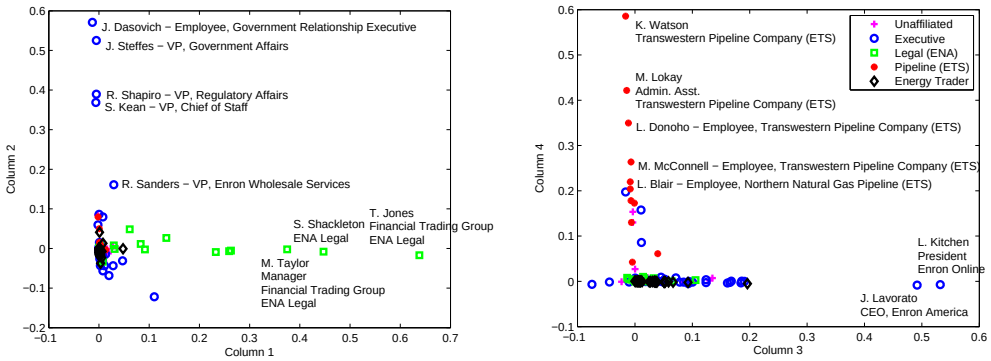

Figure 6. Three-way DEDICOM: scatter plots of the first and second columns of $\mathbf{A}$ (left plot) and third and fourth columns of $\mathbf{A}$ (right plot) .

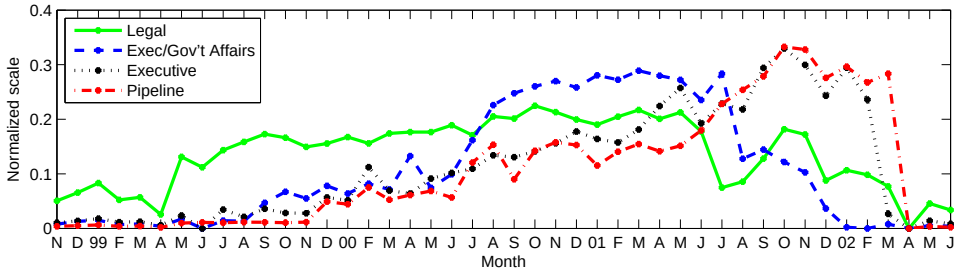

Figure 7. Scales in $\mathfrak{D}$ indicate the strength of participation of each role's communication over time.

the fourth role is the pipeline employees. The energy traders, who were identified in the two-way DEDICOM model, are missing from this analysis but are included partially in the third role (and higher dimensional solutions; see below). The government affairs node is a subgroup of the executive role and has different temporal communications, which is why it is identified as a separate role.

The aggregate communication patterns over the 44 months among these four roles is summarized in the $\mathbf{R}$ matrix and its corresponding directed graph in Figure 8. Most of the communication is within each role as evidenced by the large magnitude diagonal elements and small off-diagonal elements. There is some communication between the government/regulatory affairs executives and other senior executives (roles 2 and 3, respectively). However, the communication is substantially asymmetric in that the $r_{2,3}$ element is larger than $r_{3,2}$. This indicates that the top executives were mostly recipients of messages while the government/regulatory affairs executives were senders. The small off-diagonal elements in the fourth row and column indicate that

	#1	#2	#3	#4
#1 (Legal)	440.2	13.4	-7.9	-5.6
#2 (Exec/Gov't Affairs)	13.8	286.7	157.8	0.4
#3 (Executive)	-23.6	93.5	211.6	-4.8
#4 (Pipeline)	-4.8	-5.9	-6.5	172.4

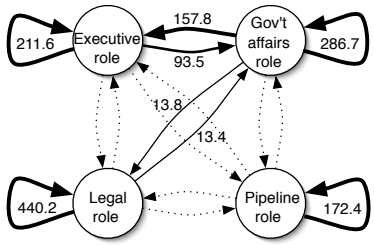

Figure 8. Three-way DEDICOM: $\mathbf{R}$ matrix and associated graph showing aggregate communication patterns. Entries less than 4 are represented as a dotted line, and dominant communications have thicker edges.

the pipeline employees interacted almost exclusively with themselves. We interpret the negative off-diagonal elements as having less communication than would expect from a typical null hypothesis, which indicates that the executive role avoided communicating with the legal role.

The scales in $\mathcal{D}$ indicate the strength of each role's participation in the communication over time. Figure 7 shows these scales of the four-component model. It is here where one sees the temporal nature of each cluster's communications. The legal department has relatively sustained communication over the whole time period as shown by the broad hump in the plot. On the other hand, the government affairs executives have frequent communications from October 2000 through October 2001, after which there is a dropoff. The top executives and pipeline employees have similar communications pattern, where they have frequent communications after October 2001. We believe these results are consistent with findings in [10].

To see the communication patterns within a particular year, we multiply $\mathbf{R}$ on the left and right by the slices of tensor $\mathcal{D}$. For example, Figure 9 shows the communication patterns among the four roles in $\mathbf{A}$ in October, 2000 and October, 2001. These two time periods were analyzed in [10] and correspond to times before and during the crisis at Enron. We see that the intra-role communication in the government affairs and legal roles decreases over this time period while it increases in the executive and pipeline roles, precisely those being investigated.

The communication patterns in Figure 9 do not exhibit any additional asymmetry than the original $\mathbf{R}$ matrix because the multiplication by $\mathbf{D}_k$ on the left and right is a symmetric update. To see differences in asymmetry, one may compute a DEDICOM model with two or more $\mathbf{R}$ matrices covering the time periods of interest while still

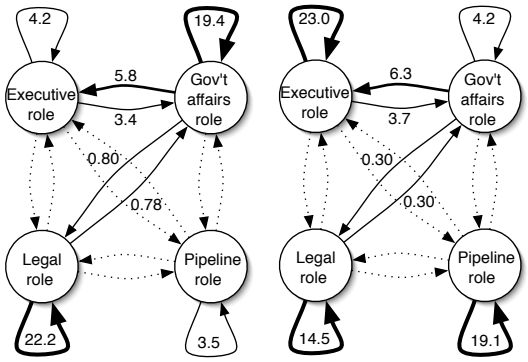

Figure 9. Graphs of $\mathbf{D}_k\mathbf{R}\mathbf{D}_k$ showing communication patterns for $k =$ October 2000 (pre-crisis, left) and $k =$ October 2001 (during crisis, right).

retaining the same $\mathbf{A}$ matrix over the whole data set. For example, we may be interested in three $\mathbf{R}$ matrices for the Enron data: before, during, and after the crisis.

Last, we comment on the DEDICOM results for different values of p. Proceeding from lower- to higher-component solutions, DEDICOM partitions the employees into increasing specific roles, so we can establish a loose hierarchical clustering of the employees. For example, the first four dimensional solutions are represented by the four-component model described above: The 2-component model groups the employees largely from the legal department and those executives dealing with government and regulatory affairs. The 3-component model adds another role of top executives, and the 4-component model includes those from the pipeline business as a fourth role. The 5-component model includes another executive role that is similar to the government and regulatory affairs role, but it has a different temporal communication pattern. The 6-component model adds the energy traders seen in the two-DEDICOM model.

5.3 Non-negative three-way DEDICOM.

We computed a four-component ($p = 4$) nonnegative decomposition of the adjacency array $\mathcal{X}$. An iteration took about 1 second, and a run required an average of 74 iterations to satisfy a tolerance of 10^{-7} in the relative change of fit. We chose the smallest minimizer from among 40 runs from random starting points, and the relative norm of the difference was 0.885 (excluding diagonal).

Qualitatively, the results of the non-negative decomposition are the same as the standard three-way results above. The scatter plots of the columns of $\mathbf{A}$ are sim-

	#1	#2	#3	#4
#1 (Legal)	437.4	0	1.7	0
#2 (Exec/Gov't Affairs)	0	269.7	57.9	3.6
#3 (Executive)	0	0	181.0	0
#4 (Pipeline)	0	0	0	171.9

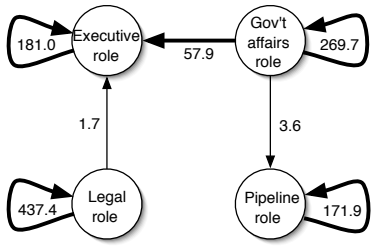

Figure 10. Non-negative three-way DEDICOM: **R** matrix and associated graph showing aggregate communication patterns.

ilar to Figure 6 and are not shown here. The scales in $\mathfrak{D}$ indicating the strength of participation of each role's communication over time are also nearly identical to Figure 7.

The benefit of the non-negativity constraints is that the **R** matrix is more easily interpreted. Figure 10 shows the corresponding **R** matrix and its graph ignoring self-links. It is clear from the asymmetry of the matrix that more communication "flows up" the management chain to the top executives. Also, the government affairs executives are passing information to the pipeline employees.

Higher component solutions of the non-negative model yields similar roles as identified in the standard three-way DEDICOM model.

5.4 Classification results.

In the DEDICOM model, the ith row of **A** can be considered as scores of how strongly the ith employee is associated with each role. In other words, a_{ij} is the strength of the association between employee i and role j. Here we quantify how accurate the assignments are. For two-way DEDICOM, we labeled the four roles as executive, legal, trading, and pipeline. For each employee for which we have a true label, we compared this against the prediction made by DEDICOM. As shown in Table 1, the executive role scored highest for 45% of the true executives and in the top two for a total of 84%. We did not consider the "unaffiliated" employees.

In the three-way model, DEDICOM identified a "government affairs" role, which did not directly correspond to the job titles we had. Since there is no longer a "trader"

True label	Highest score	1st and 2nd highest score
Two-way DEDICOM		
Executive	45%	84%
Legal	67%	87%
Pipeline	77%	77%
Trader	86%	97%
Overall	62%	87%
Three-way DEDICOM		
Executive	75%	95%
Legal	73%	80%
Pipeline	62%	77%
Overall	73%	89%
Nonnegative three-way DEDICOM		
Executive	73%	93%
Legal	73%	87%
Pipeline	62%	85%
Overall	71%	90%

Table 1. Percent of employees matching their actual business unit and job title label based on their primary and primary/secondary latent role assignments by DEDICOM.

role, we omit those employees from the tables for the three-way models. Once again, we computed the percentage of each true job type that was correctly predicted by top or top two predictions from DEDICOM.

Note that many people had dual roles and so we arbitrarily labeled VPs and directors as "executives" irrespective of their business unit. Of course, it is then the case that some executives instead load on their business unit. For example, DEDICOM may identify the VP of Legal in a "legal" role, but our label is "executive." However, in most cases, the other role (e.g., "executive") is then picked up as the next highest scoring role, resulting in an overall classification of 87–90% if the two highest scoring roles are considered.

6 Conclusions and Discussion

We have shown in the Enron email graphs that the two-way and three-way DEDICOM models identify roles of employees that share some idealized role or attribute. The components or patterns of asymmetric relationships that DEDICOM identifies have loadings in $\mathbf{A}$ that are continuously valued, like factor loadings, rather than discrete cluster membership assignments.

When each row of $\mathbf{A}$ contains only one substantial loading, the employees belong to a single role. Our analysis identified some people who were purely in a single role and other people who had mixed roles. For example, a given person might "load" on both an executive and a lawyer component, and thus show email exchanges resembling each of these two roles to some extent.

The entries in matrix $\mathbf{R}$ describe the communication patterns between roles of the same and different type. They show how a particular person's combination of roles influences the pattern of messages he/she exchanges with other employees given their roles. The $\mathbf{R}$ matrix is asymmetric and offers an idealized version of a directed graph involving the latent components identified in $\mathbf{A}$.

In addition, three-way DEDICOM shows the associated communication patterns over time in the array $\mathcal{D}$. The scales in each slice $\mathbf{D}_k$ show the strength of participation of a particular role for time period k.

In the present study, we investigated a semantic graph with edges labeled by time. As an alternative to time, we point out that our semantic graph could have incorporated different types of communication media (e.g., email, phone, and mail communications) instead of time in the third mode. Then an analysis with three-way DEDICOM would represent information about the vertices across all forms of communication (appropriately scaled by slices of $\mathcal{D}$) in the $\mathbf{A}$ and $\mathbf{R}$ matrices.

Moreover, DEDICOM is not limited to the analysis of sociometric and intercommunication data; DEDICOM may derive useful information from any directed graph. New possibilities include analyzing a network of web traffic between servers over time or a web/citation graph, where edges convey authority among vertices. A third mode enters when the two-way data are categorized by time, demographic, click number, or some other feature of the data.

Finally, we suggest two extensions to the DEDICOM model and its application in data mining that we intend to pursue. First, constrained DEDICOM [21] is an extension of DEDICOM that has been suggested in the 90's and pursued more recently. The idea is to put constraints on the $\mathbf{A}$ factors themselves so that the columns of $\mathbf{A}$ lie in a prescribed column space. For example, in the email graph, one might want to impose a constraint on the first column of $\mathbf{A}$ so that it contains only the top executives. Many other variations are possible. This procedure allows for including domain knowledge or incorporating human understanding into the problem. Kiers

and Takane [21] offered an algorithm for handling different subspace constraints on **A**. More recently, Rocci [30] proposed a new algorithm for fitting any constrained DEDICOM model. Second, DEDICOM has been applied to skew-symmetric data [15] and has yielded some benefits. There might be ways to apply this technique to semantic graphs as well.

References

[1] E. Acar, S. A. Çamtepe, M. S. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. In *ISI 2005*, pp. 256–268, 2005.

[2] B. W. Bader, R. A. Harshman, and T. G. Kolda. Temporal analysis of social networks using three-way DEDICOM. Tech. Rep. SAND2006-2161, Sandia National Labs, Albuquerque, NM and Livermore, CA, Apr. 2006.

[3] B. W. Bader and T. G. Kolda. Algorithm xxx: MATLAB tensor classes for fast algorithm prototyping. *ACM Trans. Math. Softw.* To appear.

[4] B. W. Bader and T. G. Kolda. MATLAB Tensor Toolbox Version 2.0, 2006. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`.

[5] M. W. Berry and M. Browne. Email surveillance using nonnegative matrix factorization. In WLACS05 [38].

[6] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 35:283–319, 1970.

[7] A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of Enron email data. In WLACS05 [38].

[8] W. W. Cohen. Enron email dataset. Webpage. `http://www.cs.cmu.edu/~enron/`.

[9] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[10] J. Diesner and K. M. Carley. Exploration of communication networks from the Enron email corpus. In WLACS05 [38].

[11] Federal Energy Regulatory Commision. `http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp`.

[12] R. A. Harshman. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16:1–84, 1970.

[13] R. A. Harshman. Models for analysis of asymmetrical relationships among n objects or stimuli. In *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology*, McMaster University, Hamilton, Ontario, August 1978. `http://publish.uwo.ca/~harshman/asym1978.pdf`.

[14] R. A. Harshman, P. E. Green, Y. Wind, and M. E. Lundy. A model for the analysis of asymmetric data in marketing research. *Marketing Science*, 1(2):205–242, 1982.

[15] R. A. Harshman and M. E. Lundy. Multidimensional analysis of preference structures. In *Telecommunications Demand Modelling: An integrated view*, pp. 185–204. Elsevier Science, 1990.

[16] R. A. Harshman and M. E. Lundy. Three-way DEDICOM: Analyzing multiple matrices of asymmetric relationships. Paper presented at the Annual Meeting of the North American Psychometric Society, 1992.

[17] R. A. Harshman and M. E. Lundy. Uniqueness proof for a family of models sharing features of Tucker's three-mode factor analysis and PARAFAC/CANDECOMP. *Psychometrika*, 61(1):133–154, 1996.

[18] H. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[19] P. S. Keila and D. B. Skillicorn. Structure in the Enron email dataset. In WLACS05 [38].

[20] H. A. L. Kiers. An alternating least squares algorithm for PARAFAC2 and three-way DEDICOM. *Computational Statistics and Data Analysis*, 16:103–118, 1993.

[21] H. A. L. Kiers and Y. Takane. Constrained DEDICOM. *Psychometrika*, 58(2):339–355, June 93.

[22] H. A. L. Kiers, J. M. F. ten Berge, Y. Takane, and J. de Leeuw. A generalization of Takane's algorithm for DEDICOM. *Psychometrika*, 55(1):151–158, 1990.

[23] T. G. Kolda and B. W. Bader. The TOPHITS model for higher-order web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*, 2006. `http://www.cs.rit.edu/~amt/linkanalysis06/accepted/`.

[24] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM 2005*, pp. 242–249, Nov. 2005.

[25] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW 2003*, pp. 568–576, 2003.

[26] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[27] A. McCallum, A. Corrada-Emmanual, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks, with application to Enron and academic email. In WLACS05 [38].

[28] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on Enron graphs. In WLACS05 [38].

[29] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Enron data set, 2006. `http://cis.jhu.edu/~parky/Enron/enron.html`.

[30] R. Rocci. A general algorithm to fit constrained DEDICOM models. *Stat. Meth. App.*, 13:139–150, 2004.

[31] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *SIGKDD Explor. Newsl.*, 7(2):31–40, 2005.

[32] J. Shetty and J. Adibi. The Enron email dataset database schema and brief statistical report, 2005. `http:www.isi.edu/~adibii/Enron/Enron_Dataset_Report.pdf`.

[33] J. Shetty and J. Adibi. Ex employee status report, 2005. `http:www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls`.

[34] A. Smilde, R. Bro, and P. Geladi. *Multi-way analysis: applications in the chemical sciences*. Wiley, 2004.

[35] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: a novel approach to personalized Web search. In *WWW 2005*, pp. 382–390, 2005.

[36] Y. Takane. Diagonal estimation in DEDICOM. In *Proceedings of the 1985 Annual Meeting of the Behaviormetric Society*, pp. 100–101, Sapporo, 1985.

[37] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[38] *Workshop on Link Analysis, Counterterrorism and Security*. `http://www.cs.queensu.ca/home/skill/proceedings/`, 2005.

DISTRIBUTION:

1 MS 0123 Donna Chavez, LDRD Office, 1011

2 MS 0899 Technical Library, 4536

5 MS 1318 Brett Bader, 1416

2 MS 9018 Central Technical Files, 8944

5 MS 9159 Tammy Kolda, 8962

1 Richard Harshman
Professor
Department of Psychology
University of Western Ontario
London, Ontario, CANADA N6A 5C2