SANDIA REPORT

SAND2006-2161
Unlimited Release
Printed April 2006

Temporal Analysis of Social Networks using Three-way DEDICOM

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.doe.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/ordering.htm

SAND2006-2161
Unlimited Release
Printed April 2006

Temporal Analysis of Social Networks using Three-way DEDICOM

Brett W. Bader
Applied Computational Methods Department
Sandia National Laboratories
Albuquerque, NM 87185–0316
bwbader@sandia.gov

Richard A. Harshman
University of Western Ontario
Department of Psychology
London, Ontario, Canada
harshman@uwo.ca

Tamara G. Kolda
Computational Science and Mathematics Science Research Department
Sandia National Laboratories
Livermore, CA 94551–9159
tgkolda@sandia.gov

Abstract

DEDICOM is an algebraic model for analyzing intrinsically asymmetric relationships, such as the balance of trade among nations or the flow of information among organizations or individuals. It provides information on latent components in the data that can be regarded as "properties" or "aspects" of the objects, and it finds a few patterns that can be combined to describe many relationships among these components. When we apply this technique to adjacency matrices arising from directed graphs, we obtain a smaller graph that gives an idealized description of its patterns. Three-way DEDICOM is a higher-order extension of the model that has certain uniqueness properties. It allows for a third mode of the data, such as time, and permits the analysis of semantic graphs. We present an improved algorithm for computing three-way DEDICOM on sparse data and demonstrate it by applying it to the adjacency tensor of a semantic graph with time-labeled edges. Our application uses the Enron email corpus, from which we construct a semantic graph corresponding to email exchanges among Enron personnel over a series of 44 months. Meaningful patterns are recovered in which the representation of asymmetries adds insight into the social networks at Enron.

Contents

Appendix

Figures

Tables

Temporal Analysis of Social Networks using Three-way DEDICOM

1 Introduction

Social network analysis is the study of theoretical models for interactions among people and/or organizations. An emerging challenge in social network analysis is to analyze temporal changes in network structure to help explain the evolving nature of friendships, collaborations, or organizational structure. This paper seeks to analyze those evolving patterns by studying semantic graphs where the edges have labels (in our case, the edges are labeled with a timestamp). These representations have also been called time graphs [27] or a time series of graphs [31].

This paper takes a family of models called DEDICOM (DEcomposition into DIrectional COMponents) from the psychometrics literature [12, 14] and uses them to interpret these graphs. DEDICOM is an algebraic model for analyzing asymmetric data in data analysis. It provides information on latent components in the data that can be regarded as "properties" or "aspects" of the objects, and it provides patterns of asymmetric (and symmetric) relationships among these components. We apply this technique to adjacency matrices arising from directed graphs for an idealized description of such a graph. Our present goal is to extend this research to the analysis of semantic graphs, especially communication graphs for analyzing social networks. DEDICOM has broader uses, and we also suggest potential applications in other areas, such as web analysis, web traffic, and bibliometrics.

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[1] applied to $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{ARA}^T + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix of loadings or "weights" for the n vertices on $p < n$ dimensions, $\mathbf{R} \in \mathbb{R}^{p \times p}$ is a matrix that captures the asymmetric relationships on these latent dimensions of $\mathbf{A}$, and $\mathbf{E}$ is a matrix containing the unexplained error. A simplified interpretation of DEDICOM is that it takes a large matrix and shrinks it into a condensed, idealized summary in the $\mathbf{R}$ matrix.

In DEDICOM, individual objects (i.e., rows of $\mathbf{A}$—in social networks, individual people) can have substantial weights on more than one of the latent components or patterns (i.e., have more than one substantial value per row of $\mathbf{A}$). 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 (e.g., an executive and a lawyer). Thus, DEDICOM components are perhaps better regarded as "properties" or "aspects" of the objects (e.g., roles of the individuals). Conversely, one might say that DEDICOM identifies types of correspondence patterns that have distinctive properties, and these are then linked to the individuals that exhibit mixtures of these patterns in their particular history of messages.

Individuals can be complex, and DEDICOM does not place them in a particular cluster, but instead the patterns are teased out by fitting observed patterns via mixtures of a few idealized patterns. DEDICOM shows how aspects or properties of the objects (how much a given person is like each of the idealized employee types) and aspects of a potential correspondent, will influence the patterns of communication between them (from each to the other).

The DEDICOM model can be extended to three-way data. For example, one may include time

[1] The dual-domain model has a different matrix on the left and right of $\mathbf{R}$.

as a third mode in the data. Here we consider the simplest three-way DEDICOM model[2]:

$$\mathbf{X}_i = \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T + \mathbf{E}_i \quad \text{for } i = 1, \ldots, m, \tag{2}$$

where $\mathbf{X}_i$ is the ith frontal slice of the data tensor $\mathcal{X} \in \mathbb{R}^{n \times n \times m}$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix of loadings, $\mathbf{D}_i$ is the ith frontal slice of the tensor $\mathcal{D} \in \mathbb{R}^{p \times p \times m}$, and $\mathbf{R} \in \mathbb{R}^{p \times p}$ is the asymmetry matrix. Each $\mathbf{D}_i$ matrix is a diagonal matrix that gives the weights of the columns of $\mathbf{A}$ for each level in the third mode (e.g., year of transaction). The matrix $\mathbf{R}$ captures the aggregate trends over time and, when multiplied on the left and right by the slices of tensor $\mathcal{D}$, within a particular year as well.

Three-way DEDICOM is a part of a broader family of multilinear models called PARATUCK2 [19], 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 [11] 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.

Our primary application will 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, and posted to the web, by the U.S. Federal Energy Regulatory Commission (FERC) during its investigation of Enron.

The Enron email corpus is appealing to researchers because it is a large collection of emails from real people that covers a period of 3.5 years. When FERC opened an investigation into Enron's alleged price manipulation in western energy markets, they collected more than half a million email messages belonging to 158 employees at Enron. FERC released this information to the public, and academic researchers have since cleaned and refined the database. Our research uses a simplified version of the database that was posted to the web by Priebe et al. [32]. It consists of 184 email addresses of former Enron employees and the emails passed among them over the same 3.5 years.

In the sections that follow, 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. In section 5, we apply the algorithms to the Enron communication network.

2 Related work

To our knowledge, this is the first application of DEDICOM to social network analysis and data mining, in general. We mention some relevant work from the psychometrics community for further background on the DEDICOM model. We also outline some of the more recent work in social network analysis, mostly pertaining to the Enron data, that is relevant to ours.

2.1 DEDICOM and Multi-way models

The DEDICOM family of models was first introduced in [12]. 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 [15]. While research on the model continued in the psychometrics and multilinear algebra communities, its influence did not spread far outside of these communities.

Since its introduction, there has been some research of the model and associated applications

[2]In variations of this model, the scaling tensor $\mathcal{D}$ and/or $\mathbf{A}$ matrix may be different on the left and right of $\mathbf{R}$.

[17, 28] followed by a number of papers analyzing algorithms for computing the DEDICOM model [24], including variations such as constrained DEDICOM [23, 33] and three-way DEDICOM [22].

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 [18]. Lundy et al. [28] presented an application of three-way DEDICOM to skew-symmetric data for paired preference ratings of treatments for chronic back pain. They used Paatero's Multilinear Engine program [30] to fit the three-way model.

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. [38] apply a three-way Tucker decomposition [40] 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 [26, 25] a PARAFAC decomposition [11] (also known as the Canonical Decomposition or CANDECOMP decomposition [4]) on a (web page $\times$ web page $\times$ anchor text) was used on 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 [40, 11, 4], and they have been used extensively in other domains ranging from chemometrics [37] to image analysis [41].

2.2 Enron data and social network analysis

The Enron corpus contains a large amount of information that a variety of researchers have investigated. Research on the 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 [35] 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 the 151 employees of Enron by assigning a social contact if at least 5 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 [8] 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 had been been more diverse with respect to the employees' positions except among the top executives, who had formed a tight clique.

Chapanond, Krishnamoorthy, and Yener [5] 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 [29] 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 [3] 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 [21] 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 [31] 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 [34] 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 Algorithms

In this section we discuss the algorithms for computing the decompositions in (1) and (2).

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}. Tensors (i.e., multi-way arrays) 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 ith frontal slice of $\boldsymbol{\mathcal{X}}$ is denoted by $\mathbf{X}_i$. The symbol $\otimes$ denotes the Kronecker product; for example, $\mathbf{x} = \mathbf{a} \otimes \mathbf{b}$ means $x_\ell = a_i b_j$ with $\ell = j + (i-1)(J)$ for all $1 \le i \le I$, $1 \le j \le J$. The symbol $*$ denotes the Hadamard (i.e., elementwise) matrix product.

3.1 Two-way DEDICOM

In the 2-way case, a square matrix $\mathbf{X}$ is decomposed according to (1). The goal is to find the best-fitting matrices $\mathbf{A}$ and $\mathbf{R}$ such that $\| \mathbf{E} \|_F$ is minimized. We will ignore $\mathbf{E}$ for the rest of this discussion and focus on the best approximation

$$\mathbf{X} \approx \mathbf{A}\mathbf{R}\mathbf{A}^T.$$

We wish to solve the following 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 give rise to other benefits of 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 [20] such that the variance across columns of $\mathbf{A}$ is maximized.

A further practice in some problems 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). This is commonly handled by estimating the diagonal values from the current approximation $\mathbf{A}\mathbf{R}\mathbf{A}^T$ and including them in $\mathbf{X}$.

The original algorithm proposed in [15] used the singular value decomposition (SVD) of $\mathbf{X}$ to provide an approximate minimizer of (7). Subsequent research has been aimed at computing the

global minimizer. The algorithm in [13] was one of the first least squares algorithms proposed. It was based on treating the left-hand $\mathbf{A}$ and the right-hand $\mathbf{A}$ as if they were independent, without guarantee that they are equal upon convergence. Subsequent attempts ensured that they would be equal (e.g., by adding a term to the objective function). Unfortunately, these algorithms are either not guaranteed to converge or are intractable for larger datasets.

In the following subsections, we briefly mention previous algorithms and then discuss our alternating least squares algorithm. In all cases, the principle challenge lies with finding $\mathbf{A}$ because computing $\mathbf{R}$ is a simple least squares problem.

3.1.1 Generalized Takane Algorithm

In [39], Takane proposed an algorithm for computing a DEDICOM model. His method appeared to be efficient on many practical problems, but it was not robust on all cases. Moreover, no general convergence properties were known for this method. As has been noted by Harshman and Kiers [16], Takane's algorithm tends to be very efficient but does not always converge monotonically.

In 1990, Kiers et al. [24] revisited Takane's algorithm and proposed a modification and generalization, from which Takane's algorithm now follows as a special case. The algorithm proceeds by minimizing the loss function

$$\sigma(\mathbf{A}, \mathbf{R}) = \left\| \mathbf{X} - \mathbf{A}\mathbf{R}\mathbf{A}^T \right\|_F^2 \tag{4}$$

subject to the constraint that $\mathbf{A}$ is orthonormal. Because $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, the minimum of σ with respect to $\mathbf{R}$ and fixed $\mathbf{A}$ is given by $\mathbf{R} = \mathbf{A}^T\mathbf{X}\mathbf{A}$. Minimizing (4) with respect to $\mathbf{A}$, for fixed $\mathbf{R}$, is equivalent to maximizing

$$f(\mathbf{A}) = tr(\mathbf{A}^T\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{A}) \tag{5}$$

subject to the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. In [24] it is shown that an algorithm maximizing $f(\mathbf{A})$ subject to $\mathbf{A}^T\mathbf{A} = I$ is readily given. They show that by updating matrix $\mathbf{A}$ as the Gram-Schmidt orthonormalized version of $(\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}^T\mathbf{A} + \mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}\mathbf{A} + 2\alpha\mathbf{A})$ for any α larger than the largest eigenvalue of the symmetric part of $(-\mathbf{X} \otimes \mathbf{A}^T\mathbf{X}\mathbf{A})$ then $f(\mathbf{A})$ always increases. Takane's original algorithm follows from this update rule when α is chosen equal to zero. It is worth noting that as α increases, the update will increasingly resemble its predecessor, and the rate of convergence might slow down. Thus, to exploit the efficiency of Takane's algorithm, it was proposed to compute an update for $\mathbf{A}$ (call it $\tilde{\mathbf{A}}$) using $\alpha = 0$ and then evaluate $f(\tilde{\mathbf{A}})$. If $f(\tilde{\mathbf{A}}) \leq f(\mathbf{A})$, then one computes the update for $\mathbf{A}$ based on the generalized algorithm. The procedure continues until $f(\tilde{\mathbf{A}}) \approx f(\mathbf{A})$.

3.1.2 New Alternating Least Squares Algorithm

Here we propose an alternating least squares algorithm and adapt it for use on larger applications. While the modified Takane method for finding $\mathbf{A}$ would work for our case, we describe this new approach to motivate the three-way algorithm.

We start with some 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 $\mathbf{X}^T$ simultaneously by stacking the data side by side:

$$\begin{pmatrix} \mathbf{X} & \mathbf{X}^T \end{pmatrix} = \mathbf{A} \left(\begin{pmatrix} \mathbf{R} & \mathbf{R}^T \end{pmatrix} \begin{pmatrix} \mathbf{A}^T & 0 \\ 0 & \mathbf{A}^T \end{pmatrix} \right)$$

The least squares update for $\mathbf{A}$ is found by postmultiplying both sides by the pseudo-inverse of the rest of the model:

$$\mathbf{A}_{new} \leftarrow \begin{pmatrix} \mathbf{X} & \mathbf{X}^T \end{pmatrix} \left(\begin{pmatrix} \mathbf{R} & \mathbf{R}^T \end{pmatrix} \begin{pmatrix} \mathbf{A}^T & 0 \\ 0 & \mathbf{A}^T \end{pmatrix} \right)^{\dagger}$$

For large applications, we obviously wish to avoid forming any large matrices. Hence, we may change this update to

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

The most expensive operations are the matrix products $\mathbf{X}\mathbf{A}$, $\mathbf{X}^T\mathbf{A}$, and $\mathbf{A}^T\mathbf{A}$. 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 $\mathcal{O}(p^2n)$.

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 $\mathbf{A}$:

$$\mathbf{R}_{new} = \mathbf{A}^\dagger\mathbf{X}(\mathbf{A}^T)^\dagger. \tag{6}$$

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.

3.2 Three-way DEDICOM

The three-way DEDICOM model (2) is a part of the family of models called PARATUCK2 [19], which have some uniqueness properties. 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 tensor $\mathcal{D}$) on either side that apply weights to the columns of $\mathbf{A}$. In variations of this model, the scaling tensors on the left and right of $\mathbf{R}$ may be different.

The algorithm for three-way DEDICOM is more complicated than the standard DEDICOM model because the $\mathbf{A}$ and $\mathbf{R}$ matrices apply across all levels of $\mathcal{X}$. We wish to solve the following minimization problem

$$\min_{\mathbf{A},\mathbf{R},\mathcal{D}} \sum_{i=1}^{m} \left\| \mathbf{X}_i - \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T \right\|_F^2. \tag{7}$$

There are few algorithms for computing the three-way DEDICOM model. Unfortunately, these algorithms are intractable for larger datasets.

3.2.1 Kiers' Three-way DEDICOM Algorithm

Kiers [22] has presented an alternating least squares algorithm for three-way DEDICOM (as well as the related PARAFAC2 model, which is the same as (2) except that the $\mathbf{R}$ matrix is symmetric). His procedure minimizes (7) over the three parameter sets $\mathbf{A}$, $\mathbf{R}$, and $\mathcal{D}$ in an alternating fashion. The basic steps are as follows:

1. Updating $\mathbf{A}$. Kiers minimizes (7) over the columns of $\mathbf{A}$, updating each column with its own minimization subproblem. Further details may be found in [22].
2. Updating $\mathbf{R}$. A closed form solution exists for this minimization problem. It involves vectorizing $\mathcal{X}$ and $\mathbf{R}$ and stacking them in a manner such that the objective function changes to

$$f(\mathbf{R}) = \left\| \begin{pmatrix} \mathrm{Vec}(\mathbf{X}_1) \\ \vdots \\ \mathrm{Vec}(\mathbf{X}_m) \end{pmatrix} - \begin{pmatrix} \mathbf{A}\mathbf{D}_1 \otimes \mathbf{A}\mathbf{D}_1 \\ \vdots \\ \mathbf{A}\mathbf{D}_m \otimes \mathbf{A}\mathbf{D}_m \end{pmatrix} \mathrm{Vec}(\mathbf{R}) \right\| \tag{8}$$

   Minimizing (8) over $\mathrm{Vec}(\mathbf{R})$ is a multiple regression problem, and its solution is

$$\mathrm{Vec}(\mathbf{R}) = \left(\sum_{i=1}^{m} (\mathbf{D}_i\mathbf{A}^T\mathbf{A}\mathbf{D}_i) \otimes (\mathbf{D}_i\mathbf{A}^T\mathbf{A}\mathbf{D}_i) \right)^{-1} \sum_{i=1}^{m} \mathrm{Vec}(\mathbf{D}_i\mathbf{A}^T\mathbf{X}_i\mathbf{A}\mathbf{D}_i). \tag{9}$$

3. Updating $\mathcal{D}$. A closed form solution does not appear to exist. Kiers opts for an alternating least squares solution that solves this problem elementwise. Further details may be found in [22].

The updates of $\mathbf{A}$ and $\mathcal{D}$ are slow due to the columnwise and elementwise minimizations.

3.2.2 New three-way DEDICOM Algorithm

Here we propose an alternating least squares algorithm and adapt it for use on larger applications. Our approach seeks improvements over the Kiers method for updating $\mathbf{A}$ and $\mathcal{D}$.

Once again, we start with some initial estimates 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 $\mathcal{X}$ by stacking the data side by side:

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

   Here $\mathbf{I}_{2m}$ is the identity matrix of size $2m \times 2m$. The least squares update for $\mathbf{A}$ is the matrix of $\mathbf{X}_i$ slices multiplied on the right by the pseudo-inverse of the matrix

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

   This computation simplifies to

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

   where

$$\begin{aligned}\mathbf{B}_i &\equiv \mathbf{D}_i\mathbf{R}\mathbf{D}_i(\mathbf{A}^T\mathbf{A})\mathbf{D}_i\mathbf{R}^T\mathbf{D}_i,\\ \mathbf{C}_i &\equiv \mathbf{D}_i\mathbf{R}^T\mathbf{D}_i(\mathbf{A}^T\mathbf{A})\mathbf{D}_i\mathbf{R}\mathbf{D}_i.\end{aligned}$$

   This least squares problem updates all columns of $\mathbf{A}$ simultaneously and is an improvement over a columnwise update of $\mathbf{A}$.

2. Updating $\mathbf{R}$: We use the closed form solution for $\mathbf{R}$ from Kiers [22], which is listed in (9) above. Provided that the number of latent dimensions is not large (specifically that p^2 is not large), then Kiers' step for updating $\mathbf{R}$ will suffice.

3. Updating $\mathcal{D}$: We improve upon the elementwise minimization of Kiers [22] by considering a full-scale minimization with respect to the diagonal elements for each slice $\mathbf{D}_i$:

$$\min_{\mathbf{D}_i}\left\|\,\mathbf{X}_i - \mathbf{A}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\mathbf{A}^T\,\right\|_F^2. \tag{10}$$

   Because there are only p variables for each of the m slices, Newton's method applied to (10) is not expensive and offers fast quadratic convergence. The pieces needed are the gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ of (10), which are provided in Appendix A. 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 calculation; see [7] for further information.

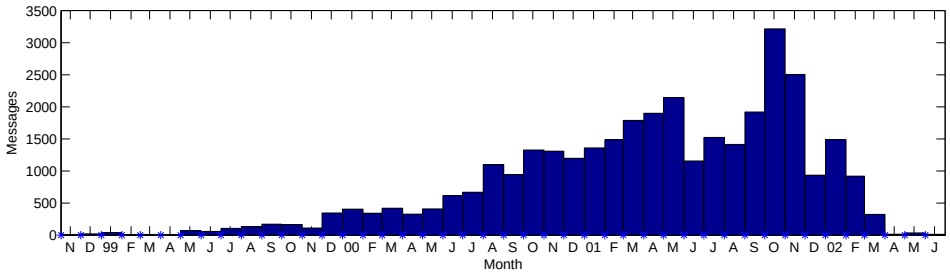

Figure 1. Number of emails per month in the Enron email graph.

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 $\mathbf{D}_i$ matrices were generated randomly to relate the patterns. When these $\mathcal{X}$ tensors 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.

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 $\mathbf{A}$,

$$\mathbf{A} = \mathbf{Q}\tilde{\mathbf{A}}, \tag{11}$$

using, for example, a compact QR decomposition. 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 (10) is the same as

$$\min_{\mathbf{D}_i} \left\| \mathbf{Q}^T\mathbf{X}_i\mathbf{Q} - \tilde{\mathbf{A}}\mathbf{D}_i\mathbf{R}\mathbf{D}_i\tilde{\mathbf{A}}^T \right\|_F^2, \tag{12}$$

except that $\mathbf{Q}^T\mathbf{X}_i\mathbf{Q}$ and $\tilde{\mathbf{A}}$ are both of size $p \times p$. We use these smaller matrices in place of $\mathbf{X}_i$ and $\mathbf{A}$, respectively, in the updates of both $\mathbf{R}$ and $\mathcal{D}$ in (9) and (10) above.

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

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 [6] 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. [32] that consists of messages solely among 184 Enron email addresses. The timestamps of these messages are included in the raw data. Unfortunately, some of the dates are clearly wrong because they refer to times too far in the past for email (1979). We neglected these and considered messages only in the interval 13-Nov-1998 through 21-Jun-2002. This resulted in a total of 34,427 messages over 44 months. Figure 1 shows a histogram of the messages in our graph.

We constructed an email graph and labeled the edges using the timestamps categorized by month and year. Our final graph had 184 email addresses and 44 time periods, which resulted in a sparse tensor $\mathcal{X}$ of size $184 \times 184 \times 44$ with 9838 nonzeros. We scaled the entries so that

$$x_{ijk} = \begin{cases} \log_2(w_k) + 1 & \text{if } i \text{ emailed } j \text{ during month } k, \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{for } 1 \leq i, j \leq n = 184, \quad 1 \leq k \leq m = 44, \quad (13)$$

where w_k is the number of unique messages from address i to address j during month k. This simple weighting reduces the biasing from prolific emailers. Other weightings are possible as well.

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 and slowly made available.

The Priebe data set [32] 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 [36]. 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. We searched for corroborating information of the preexisting data or for new identification information, such as title, business unit, or manager to help analyze our results. We also collected some relevant information posted on the FERC website [9].

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 [2].

Table 1 shows the $\mathbf{A}$ and $\mathbf{R}$ matrices for a single decomposition ($p = 3$) of the two-way DEDICOM model. 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.

DEDICOM does not actually identify clusters, except in special circumstances when such clusters happen to exist in the data as we are partially seeing in the Enron data. The components or patterns of asymmetric relationships that it identifies have loadings in $\mathbf{A}$ that are continuously-valued, like factor loadings, rather than discrete cluster membership assignments.

Here, DEDICOM describes the employees by the different latent dimensions. The first factor ($\mathbf{a}_1$) describes an executive role that fits many of the top executives. The second factor ($\mathbf{a}_2$) describes a legal role, and the third factor ($\mathbf{a}_3$) describes a pipeline employee.

The $\mathbf{R}$ matrices show that most of the communication is among employees that share the same role, as evidenced by the large diagonal values in $\mathbf{R}$. We do see some asymmetric communication. The entries in the lower triangular portion are typically larger than the corresponding transpose entry in the upper triangular. This suggests that slightly more communication "flows up" the management chain than "down."

As a point of reference, we compute the singular value decomposition $\mathbf{X} = U\Sigma V^T$. Table 1

	DEDICOM Solution			SVD (left) Solution			SVD (right) Solution		
Employee	1	2	3	1	2	3	1	2	3
J. Lavorato - CEO, Enron America	**0.41**	0.07	0.04	**0.30**	-0.07	**-0.21**	**0.31**	-0.09	-0.07
L. Kitchen - President, Enron Online	**0.26**	**0.21**	0.04	**0.31**	0.07	-0.05	**0.29**	0.02	0.04
M. Grigsby - Director, West Desk Gas Trading	**0.22**	-0.01	-0.01	**0.16**	-0.09	**-0.33**	**0.14**	-0.06	**-0.20**
D. Delainey - CEO, ENA and Enron Energy Services	**0.20**	0.06	0.06	**0.20**	-0.05	-0.00	**0.20**	-0.05	0.03
G. Whalley - President,	**0.17**	0.05	0.04	0.08	-0.02	-0.02	**0.24**	-0.07	0.02
L. Taylor - Executive Assistant to Greg Whalley,	**0.17**	0.06	0.03	**0.24**	-0.05	-0.08	0.09	-0.01	-0.02
S. Beck - COO,	**0.16**	0.04	0.02	**0.17**	-0.05	-0.08	**0.11**	-0.03	-0.01
J. Arnold - VP, Financial Enron Online	**0.15**	0.06	-0.01	**0.13**	-0.02	-0.09	**0.14**	0.04	-0.09
S. Neal - VP, East Desk Gas Trading	**0.15**	0.04	-0.02	**0.12**	0.00	**-0.14**	**0.12**	-0.00	**-0.12**
J. Shankman - President, Enron Global Markets	**0.14**	0.03	0.02	**0.12**	-0.03	-0.02	**0.13**	-0.04	-0.03
R. Shapiro - VP, Regulatory Affairs	**0.14**	0.06	**0.11**	**0.16**	-0.05	0.07	**0.18**	-0.07	**0.10**
R. Buy - Manager, Chief Risk Management Officer	**0.13**	0.03	0.03	0.08	-0.03	0.01	**0.15**	-0.05	-0.00
S. Kean - VP, Chief of Staff	**0.12**	0.06	**0.12**	**0.17**	-0.08	0.09	**0.15**	-0.04	0.09
J. Steffes - VP, Government Affairs	**0.12**	0.08	**0.11**	**0.19**	-0.02	0.06	**0.13**	-0.06	0.08
M. Lenhart - Analyst, West Desk Gas Trading	**0.10**	-0.03	0.00	0.06	-0.07	**-0.17**	0.06	-0.06	**-0.16**
J. Williamson - Executive Assistant,	0.10	0.02	0.05	**0.18**	-0.10	0.05	0.02	-0.01	0.01
K. Keiser - Employee, Gas	0.10	-0.02	-0.02	0.07	-0.05	**-0.21**	0.05	-0.02	**-0.12**
J. Reitmeyer - Associate, Eastern Rockies Natural Gas Trader	0.08	-0.02	-0.01	0.03	-0.03	-0.07	0.06	-0.05	**-0.16**
T. Jones - Employee, Financial Trading Group (ENA Legal)	**-0.12**	**0.38**	-0.02	**0.17**	**0.36**	**0.13**	0.10	**0.24**	**0.10**
M. Taylor - Manager, Financial Trading Group ENA Legal	-0.10	**0.35**	-0.01	**0.13**	**0.27**	**0.13**	**0.13**	**0.26**	**0.12**
S. Shackleton - Employee, ENA Legal	**-0.13**	**0.31**	-0.02	0.08	**0.26**	**0.10**	0.08	**0.26**	**0.10**
S. Panus - Senior Legal Specialist, ENA Legal	**-0.11**	**0.26**	-0.02	0.09	**0.27**	0.10	0.05	**0.20**	0.08
M. Heard - Senior Legal Specialist, ENA Legal	-0.10	**0.24**	-0.02	0.06	**0.20**	0.09	0.08	**0.22**	0.09
E. Sager - VP and Asst Legal Counsel, ENA Legal	-0.01	**0.24**	0.02	**0.12**	**0.13**	0.10	**0.15**	**0.21**	**0.12**
S. Bailey - Legal Assistant, ENA Legal	**-0.11**	**0.23**	-0.02	0.04	**0.17**	0.07	0.06	**0.24**	0.10
J. Hodge - Asst General Counsel, ENA Legal	-0.04	**0.21**	-0.00	0.06	**0.13**	0.06	**0.14**	**0.22**	0.09
M. Haedicke - Managing Director, ENA Legal	0.05	**0.18**	0.03	**0.16**	**0.13**	**0.10**	**0.14**	0.07	0.06
K. Mann - Lawyer,	-0.05	**0.16**	0.00	0.07	**0.17**	0.08	0.05	**0.13**	0.06
D. Perlingiere - Legal Specialist, ENA Legal	-0.05	**0.15**	-0.01	0.06	**0.17**	0.02	0.05	**0.14**	0.04
S. Corman - VP, Regulatory Affairs	-0.04	-0.01	**0.33**	0.08	**-0.18**	**0.22**	0.07	**-0.18**	**0.21**
K. Watson - Employee, Transwestern Pipeline Company (ETS)	-0.08	-0.03	**0.32**	0.03	**-0.16**	**0.19**	0.04	**-0.18**	**0.22**
L. Donoho - Employee, Transwestern Pipeline Company (ETS)	-0.08	-0.03	**0.30**	0.03	**-0.16**	**0.18**	0.03	**-0.17**	**0.20**
D. Fossum - VP, Transwestern Pipeline Company (ETS)?	-0.06	-0.00	**0.30**	0.07	**-0.18**	**0.23**	0.05	**-0.13**	**0.16**
M. Lokay - Admin. Asst., Transwestern Pipeline Company (ETS)	-0.07	-0.02	**0.28**	0.03	**-0.14**	**0.17**	0.04	**-0.17**	**0.20**
K. Hyatt - Director, Asset Development TW Pipeline Co. (ETS)	-0.06	-0.02	**0.25**	0.03	**-0.13**	**0.17**	0.04	**-0.14**	**0.17**
R. Hayslett - VP, Also CFO and Treasurer	-0.04	-0.01	**0.23**	0.04	**-0.13**	**0.16**	0.05	**-0.14**	**0.16**
L. Blair - Employee, Northern Natural Gas Pipeline (ETS)	-0.06	-0.02	**0.22**	0.02	**-0.13**	**0.16**	0.02	**-0.11**	**0.13**
T. Geaccone - Manager, (ETS)	-0.05	-0.02	**0.21**	0.03	**-0.13**	**0.15**	0.02	**-0.12**	**0.14**
S. Scott - Employee, Transwestern Pipeline Company (ETS)	0.02	-0.01	**0.20**	0.08	**-0.15**	**0.10**	0.08	**-0.12**	0.03
J. Dasovich - Employee, Government Relationship Executive	**0.13**	0.05	**0.19**	**0.25**	**-0.15**	**0.13**	**0.11**	-0.08	0.09
R matrix / singular values	70.3	11.6	6.7	86.3			86.3		
	15.4	68.2	5.0		54.1			54.1	
	9.9	6.7	59.5			52.6			52.6

Table 1. Two-way DEDICOM and SVD results on the Enron email graph. The top 10 entries from all reported columns of **A**, **U**, and **V** are listed in the table. Entries exceeding a threshold of 0.10 (before rounding) are highlighted.

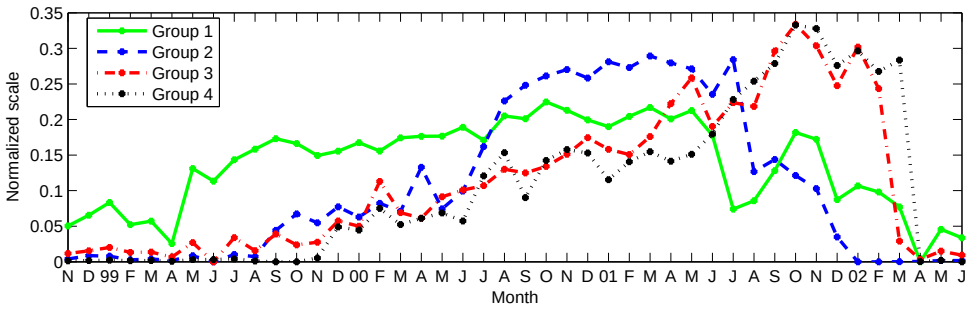

Figure 2. Scales in $\mathcal{D}$ indicate the strength of participation of each group's communication over time.

shows the first three columns of the left singular vectors (**U** matrix) and right singular vectors (**V** matrix). Because **X** is nearly symmetric, the left and right singular vectors are nearly the same. Any differences between **U** and **V** indicate whether the person is more likely to send mail (**U**) or receive mail (**V**).

The SVD solution is somewhat similar to the DEDICOM model. Many of the same people are identified and weighted similarly by DEDICOM and SVD. However, there are many more negative entries in SVD than in DEDICOM. The DEDICOM model also provides directional information between the latent groups in the **R** matrix that the SVD does not show.

Table 2 shows the **A** and **R** matrices for three instances ($p = 2, 3, 4$) of the three-way DEDICOM model. The 2-dimensional solution groups the employees largely from the legal department and those executives dealing with government and regulatory affairs. The 3-dimensional solution adds a another role of top executives, and the 4-dimensional solution includes those from the pipeline business in a fourth role.

The aggregate communication patterns over the 44 months among these 2-4 groups is summarized in the **R** matrix. In the 2-dimensional solution we see that most of the communication is within each group as evidenced by the large diagonal elements and small off-diagonal elements. The 3-dimensional solution shows some communication between the government/regulatory affairs people and other senior VP's (dimensions 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 VP's were mostly recipients of messages while the government/regulatory affairs employees were senders. With the addition of the pipeline employees in the 4-dimensional solution, we see that they interact almost exclusively with themselves due to the small off-diagonal elements in the fourth row and column.

The scales in tensor $\mathcal{D}$ indicate the strength of each group's participation in the communication over time. Figure 2 shows these scales of the 4-dimensional solution on the small Enron network. It is here where one sees the temporal nature of each cluster's communications. The legal department is relatively consistent over the whole time period as shown by the broad hump in the plot. On the other hand, the government/regulatory affairs employees have frequent communications from October 2000 through October 2001, when there is a precipitous dropoff. The VP's and pipeline employees have roughly the same communications pattern, where they have frequent communications after October 2001. We believe these results are consistent with findings in [8].

The matrix **R** captures the aggregate trends over time. To see the trends within a particular year, we take the **R** from the $p = 4$ solution and multiply it on the left and right by the slices of tensor $\mathcal{D}$. For example, Table 3 shows the communication patterns among the four groups in **A** in

	2-Dimensional Solution		3-Dimensional Solution			4-Dimensional Solution			
Employee	1	2	1	2	3	1	2	3	4
T. Jones - Employee, Financial Trading Group (ENA Legal)	**0.64**	-0.02	**0.64**	-0.02	0.01	**0.64**	-0.01	0.02	-0.00
S. Shackleton - Employee, ENA Legal	**0.45**	-0.02	**0.45**	-0.01	-0.02	**0.45**	-0.00	-0.01	-0.00
M. Taylor - Manager, Financial Trading Group ENA Legal	**0.38**	0.00	**0.37**	-0.01	0.01	**0.37**	0.01	0.02	-0.00
S. Bailey - Legal Assistant, ENA Legal	**0.26**	-0.01	**0.26**	-0.01	-0.01	**0.26**	-0.00	-0.01	-0.00
S. Panus - Senior Legal Specialist, ENA Legal	**0.26**	-0.01	**0.26**	-0.01	-0.01	**0.26**	-0.00	-0.00	-0.00
M. Heard - Senior Legal Specialist, ENA Legal	**0.23**	-0.01	**0.23**	-0.01	0.00	**0.23**	-0.00	0.00	-0.00
J. Hodge - Asst General Counsel, ENA Legal	**0.13**	0.03	**0.13**	0.03	0.00	**0.13**	0.03	0.01	-0.00
L. Kitchen - President, Enron Online	**0.10**	**0.08**	**0.11**	**-0.13**	**0.53**	**0.11**	**-0.09**	**0.53**	0.00
S. Dickson - Employee, ENA Legal	**0.09**	-0.00	**0.09**	-0.00	0.00	**0.09**	-0.00	0.00	-0.00
E. Sager - VP and Asst Legal Counsel, ENA Legal	**0.08**	0.04	**0.08**	0.01	**0.06**	**0.08**	0.02	**0.07**	-0.00
J. Dasovich - Employee, Government Relationship Executive	-0.01	**0.58**	-0.02	**0.57**	0.04	-0.01	**0.58**	0.06	0.01
J. Steffes - VP, Government Affairs	-0.00	**0.49**	-0.01	**0.52**	**-0.08**	0.00	**0.53**	**-0.06**	-0.01
R. Shapiro - VP, Regulatory Affairs	-0.01	**0.43**	-0.01	**0.39**	**0.09**	-0.00	**0.40**	**0.10**	-0.00
S. Kean - VP, Chief of Staff	-0.01	**0.35**	-0.01	**0.37**	-0.05	-0.00	**0.37**	-0.04	-0.00
R. Sanders - VP, Enron Wholesale Services	0.03	**0.16**	0.03	**0.16**	-0.01	0.03	**0.16**	-0.01	-0.00
D. Delainey - CEO, ENA and Enron Energy Services	0.01	**0.12**	0.01	**0.08**	**0.08**	0.01	**0.09**	**0.09**	-0.00
S. Corman - VP, Regulatory Affairs	-0.00	**0.08**	-0.00	**0.08**	-0.01	-0.00	**0.08**	-0.00	**0.20**
M. Carson - Employee, Corporate and Environmental Policy	-0.00	**0.07**	-0.00	**0.09**	-0.02	-0.00	**0.08**	-0.02	-0.00
S. Scott - Employee, Transwestern Pipeline Company (ETS)	-0.00	**0.08**	-0.00	**0.08**	-0.00	-0.00	**0.08**	-0.00	0.04
J. Lavorato - CEO, Enron America	0.02	**0.12**	0.02	**-0.08**	**0.49**	0.02	-0.04	**0.49**	0.00
M. Grigsby - Director, West Desk Gas Trading	0.00	0.04	0.00	-0.04	**0.20**	0.00	-0.03	**0.20**	-0.00
G. Whalley - President,	0.01	0.06	0.01	-0.03	**0.19**	0.01	-0.01	**0.19**	0.00
J. Steffes - VP, Government Affairs	0.00	0.04	0.00	-0.04	**0.19**	0.00	-0.02	**0.18**	0.00
K. Presto - VP, East Power Trading	0.01	0.01	0.01	-0.06	**0.19**	0.01	-0.05	**0.18**	0.00
S. Beck - COO,	0.01	0.02	0.01	-0.05	**0.17**	0.01	-0.03	**0.17**	0.00
B. Tycholiz - VP, Marketing	0.01	0.04	0.01	-0.03	**0.17**	0.01	-0.02	**0.16**	0.00
J. Arnold - VP, Financial Enron Online	0.03	0.02	0.03	-0.05	**0.16**	0.03	-0.04	**0.16**	-0.00
J. Williamson - Executive Assistant,	0.00	0.02	0.00	-0.03	**0.14**	0.00	-0.02	**0.14**	0.01
K. Watson - Employee, Transwestern Pipeline Company (ETS)	-0.00	0.01	-0.00	0.00	0.01	-0.00	-0.00	0.01	**0.59**
M. Lokay - Admin. Asst., Transwestern Pipeline Company (ETS)	-0.00	0.01	-0.00	0.01	0.01	-0.00	0.01	0.01	**0.42**
L. Donoho - Employee, Transwestern Pipeline Company (ETS)	-0.00	0.01	-0.00	0.01	0.01	-0.00	0.01	0.01	**0.35**
M. McConnell - Employee, Transwestern Pipeline Company (ETS)	0.00	0.00	0.00	-0.00	0.01	0.00	-0.00	0.01	**0.26**
L. Blair - Employee, Northern Natural Gas Pipeline (ETS)	-0.00	0.01	-0.00	0.01	0.00	-0.00	0.00	0.00	**0.22**
K. Hyatt - Director, Asset Development TW Pipeline Business (ETS)	-0.00	0.02	-0.00	0.02	0.00	-0.00	0.01	0.00	**0.20**
D. Schoolcraft - Employee, Gas Control (ETS)	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	**0.18**
T. Geaccone - Manager, (ETS)	0.00	0.00	0.00	-0.00	0.01	0.00	-0.00	0.01	**0.17**
R. Hayslett - VP, Also CFO and Treasurer	0.00	0.01	0.00	-0.00	0.02	0.00	-0.00	0.02	**0.16**
R matrix	438.3	12.1	440.3	18.6	-0.9	440.2	1.6	-15.0	0.4
	15.3	291.9	19.7	292.5	168.4	1.6	278.3	135.4	1.6
			-17.0	104.1	216.4	-29.3	70.7	201.6	-6.2
						1.4	-4.6	-7.5	172.3

Table 2. Three-way DEDICOM results on the Enron email graph for three different decompositions, $p = 2, 3, 4$. The top 10 entries from all reported columns of $\mathbf{A}$ are listed in the table. Entries exceeding a threshold of 0.06 are highlighted.

	$\mathbf{D}_t\mathbf{R}\mathbf{D}_t$			
October 2000	**22.2**	0.1	-0.5	0.0
	0.1	**19.0**	**4.7**	0.1
	-0.9	**2.5**	**3.6**	-0.1
	0.0	-0.2	-0.1	**3.5**
October 2001	**14.5**	0.0	-0.9	0.0
	0.0	**4.1**	**5.5**	0.1
	-1.8	**2.9**	**22.5**	-0.7
	0.1	-0.2	-0.8	**19.1**

Table 3. $\mathbf{D}_t\mathbf{R}\mathbf{D}_t$ matrices showing communication patterns for October, 2000 and October, 2001.

October, 2000 and October, 2001. These two time periods were analyzed in [8] and correspond to times before and during the crisis at Enron. From the diagonal entries in these matrices we see that the intra-group communication in groups 1 and 2 decreases over this time period while it increases in groups 3 and 4. The inter-group communication between the VP's does not change appreciably.

6 Conclusions and Discussion

We have shown in the Enron email graphs that the two-way and three-way DEDICOM models identify groups of employees that share some idealized role or attribute. When each row of $\mathbf{A}$ contains only one substantial loading, the employees belong to a single group, which is a sort of configuration (in factor analysis) that Harris and Kaiser [10] called "simple cluster" configuration. When this kind of solution is obtained, the model did indeed identify distinct clusters of individuals. A more common intermediate result might be that it identifies some people who were pretty much purely of a certain type and other people who had mixed characteristics. For example, a given person might "load" on both an executive and a lawyer component or aspect, 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 groups of the same and different type. They show how a particular person's combination of roles or attributes influences the pattern of messages he/she exchanges with particular other employees given the other employee's roles or attributes. The $\mathbf{R}$ matrix is asymmetric and offers an idealized version of a directed graph involving the components identified in $\mathbf{A}$.

In addition, three-way DEDICOM shows the associated communication patterns over time in the tensor $\mathcal{D}$. The scales in each $\mathbf{D}_t$ show the strength of participation of a particular group for time period t.

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.

Furthermore, 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 perhaps a web/citation graph, where edges convey authority among vertices. A third mode enters when the 2-way data are categorized by time, demographic, click number, or some other feature of the data.

Finally, we suggest a few extensions to the DEDICOM model and its application in data mining

that we intend to pursue. First, constrained DEDICOM [23] 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 **A** factors themselves so that the columns of **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 **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 [23] offered an algorithm for handling different subspace constraints on **A**. More recently, Rocci [33] proposed a new algorithm for fitting any constrained DEDICOM model.

Second, a nonnegative factorization of DEDICOM, where **A** and/or **R** are nonnegative, would preserve the non-negativity of the data, which could be desirable in some domains and applications.

Finally, DEDICOM has been applied to skew-symmetric data [17] 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: IEEE International Conference on Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 256–268. Springer Verlag, 2005.

[2] B. W. Bader and T. G. Kolda. MATLAB tensor classes for fast algorithm prototyping. Technical Report SAND2004-5187, Sandia National Laboratories, Albuquerque, NM 87185 and Livermore, CA 94550, Oct. 2004. Submitted to ACM T. Math. Software.

[3] M. W. Berry and M. Browne. Email surveillance using nonnegative matrix factorization. In *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

[4] 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.

[5] A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of Enron email data. In *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

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

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

[8] J. Diesner and K. M. Carley. Exploration of communication networks from the Enron email corpus. In *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

[9] Federal Energy Regulatory Commision. Ferc: Information released in Enron investigation. `http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp`.

[10] C. W. Harris and H. F. Kaiser. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29(4):347–362, 1964.

[11] 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.

[12] 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`.

[13] R. A. Harshman. Alternating least squares estimation for the single domain DEDICOM model, 1981. Unpublished technical memorandum, Bell Laboratories, Murray Hill, NJ `http://publish.uwo.ca/~harshman/asym1981.pdf`.

[14] R. A. Harshman. DEDICOM: A family of models generalizing factor analysis and multidimensional scaling for decomposition of asymmetric relationships. Unpublished manuscript, University of Western Ontario, 1982.

[15] 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.

[16] R. A. Harshman and H. A. L. Kiers. Algorithms for DEDICOM analysis of asymmetric data. In *European Meeting of the Psychometric Society*, Enschede, 1987.

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

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

[19] 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, March 1996.

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

[21] P. S. Keila and D. B. Skillicorn. Structure in the Enron email dataset. In *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

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

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

[24] 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.

[25] 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.

[26] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM 2005: Proceedings of the 5th IEEE International Conference on Data Mining*, pages 242–249. IEEE Computer Society, 2005.

[27] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. 12th Internat. World Wide Web Conference*, pages 568–576, 2003.

[28] M. E. Lundy, R. A. Harshman, P. Paatero, and L. Swartzman. Application of the 3-way DEDICOM model to skew-symmetric data for paired preference ratings of treatments for chronic back pain. Presentation, June 2003. TRICAP2003, Lexington, Kentucky `http://publish.uwo.ca/~harshman/tricap03.pdf`.

[29] 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 *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

[30] P. Paatero. The multilinear engine–a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *J. Computat. Graphical Stat.*, 8:854–888, 1999.

[31] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on Enron graphs. In *Workshop on Link Analysis, Counterterrorism and Security, SIAM Conf. on Data Mining*, Newport Beach, CA, 2005.

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

[33] R. Rocci. A general algorithm to fit constrained DEDICOM models. *Statistical Methods and Applications*, 13:139–150, 2004.

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

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

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

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

[38] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: a novel approach to personalized Web search. In *WWW 2005: Proceedings of the 14th international conference on World Wide Web*, pages 382–390, 2005.

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

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

[41] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *ECCV 2002: 7th European Conference on Computer Vision*, volume 2350 of *Lecture Notes in Computer Science*, pages 447–460. Springer-Verlag, 2002.

A Hessian and gradient calculation

The gradient $\mathbf{g}$ of (10) is given by

$$g_k = \frac{\partial f}{\partial d_k} = \sum_{i,j} \left[2(x_{ij} - \sum_{p,q} a_{ip} d_p r_{pq} d_q a_{jq})(- \sum_{p} a_{ip} d_p r_{pk} a_{jk} - \sum_{q} a_{ik} r_{kq} d_q a_{jq}) \right]. \tag{14}$$

This may be written more simply in matrix form for each element of $\mathbf{g}$:

$$g_k = - \sum_{i,j} \left[2(\mathbf{X} - \mathbf{ADRDA}^T) * (\mathbf{ADr}_k \mathbf{a}_k^T + \mathbf{a}_k \mathbf{r}_{k,:} \mathbf{DA}^T) \right]_{i,j}, \tag{15}$$

where we use the notation $\mathbf{r}_{k,:}$ to refer to the kth row of $\mathbf{R}$.

The Hessian $\mathbf{H}$ of (10) is given by

$$\begin{aligned} h_{st} = \frac{\partial^2 f}{\partial d_s \partial d_t} = \sum_{i,j} 2(x_{ij} - \sum_{p,q} a_{ip} d_p r_{pq} d_q a_{jq})(-a_{is} r_{st} a_{jt} - a_{it} r_{ts} a_{js}) + \\ 2(- \sum_{p} a_{ip} d_p r_p s a_{js} - \sum_{q} a_{is} r_{sq} d_q a_{jq}) \\ (- \sum_{p} a_{ip} d_p r_p t a_{jt} - \sum_{q} a_{ip} r_{pq} d_q a_{jq}). \end{aligned} \tag{16}$$

This may be written more simply in matrix form for each element of H:

$$\begin{aligned} h_{st} = -2 \sum_{i,j} \Big[(\mathbf{X} - \mathbf{ADRDA}^T) * (\mathbf{a}_s r_{st} \mathbf{a}_t^T + \mathbf{a}_t r_{ts} \mathbf{a}_s^T) \\ - (\mathbf{ADr}_s \mathbf{a}_s^T + \mathbf{a}_s \mathbf{r}_{s:} \mathbf{DA}^T) * (\mathbf{ADr}_t \mathbf{a}_t^T + \mathbf{a}_t \mathbf{r}_{t:} \mathbf{DA}^T) \Big]_{i,j}. \end{aligned} \tag{17}$$