# Temporal analysis of semantic graphs using ASALSAN*

Brett W. Bader
*Sandia National Laboratories*
*Albuquerque, NM, USA*
*bwbader@sandia.gov*

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

Tamara G. Kolda
*Sandia National Laboratories*
*Livermore, CA, USA*
*tgkolda@sandia.gov*

## Abstract

*ASALSAN is a new algorithm for computing three-way DEDICOM, which is a linear algebra model for analyzing intrinsically asymmetric relationships, such as trade among nations or the exchange of emails among individuals, that incorporates a third mode of the data, such as time. ASALSAN is unique because it enables computing the three-way DEDICOM model on large, sparse data. A nonnegative version of ASALSAN is described as well. 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 international trade data and the Enron email corpus to uncover latent components and their transient behavior. The mixture of roles assigned to individuals by ASALSAN 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.*

## 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 introduces ASALSAN, a new algorithm for computing DEDICOM (DEcomposition into DIrectional COMponents) [17] to provide information on latent components in data and the pattern of asymmetric (i.e., directed) relationships among these components.

In this paper, we consider three-way DEDICOM for interpreting directed semantic graphs (i.e., graphs with labeled edges) arising from international trade data and email communications at Enron. Two contributions of this paper are that we provide a new algorithm to compute three-way DEDICOM for large-scale data, and we also present a nonnegative version of ASALSAN.

In the general case, 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 two-way DEDICOM model applied to $\mathbf{X}$ is an approximation

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

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.

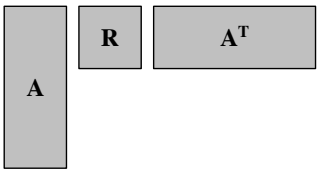


**Figure 1. Two-way DEDICOM model.**

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

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

where $\mathbf{X}_k$ is the $k$th 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 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

*This research was sponsored by the United States Department of Energy and by Sandia National Laboratory, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE–AC04–94AL85000.

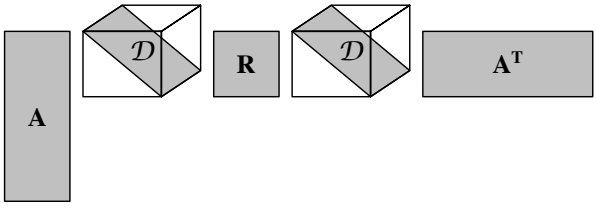


**Figure 2. Three-way DEDICOM model.**

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 (e.g., 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 employee might have characteristics that cause their pattern of email exchanges to look like a mixture of two different idealized roles, such as an executive and a lawyer.

To help uncover latent components and temporal patterns in data, this paper seeks to analyze semantic graphs where each edge is labeled by the time period. This representation has also been called a time graph [28]. We arrange the data in a three-way array such that each slice corresponds to an adjacency matrix of a particular time period.

We present two applications. First, we study a small example of international trade data. Then we investigate the Enron email corpus that was made public by the U.S. Federal Energy Regulatory Commission (FERC) during its investigation of the Enron corporation.

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 present the ASALSAN algorithm for computing three-way DEDICOM models. Section 4 describes our two applications, and we end with conclusions in section 5.

## 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 [17]. 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 [18].

There has been some research of the model and associated applications [19] followed by a number of papers analyzing algorithms for computing the DEDICOM model [25], including variations such as constrained DEDICOM [24, 33] and three-way DEDICOM [23]. Most of the applications of DEDICOM in the literature have focused on two-way data, and there is very little research in the three-way 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 [20]. DEDICOM has never been applied to large-scale sparse data.

The use of multi-way models is relatively new in the context of data mining. Sun et al. [37, 38] use a method they call "dynamic tensor analysis" to look for patterns in multi-way data with a time dimension. In [1], various multiway analyses of (user $\times$ key word $\times$ time) data are used to separate different streams of conversation in chatroom data. Sun et al. [39] 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 [27, 26] a PARAFAC decomposition [16] (also known as CANDECOMP [8]) is applied to (web page $\times$ web page $\times$ anchor text) data, forming a sparse, three-way array representing the web graph with anchor-text-labeled edges. In text analysis, Bader, Berry, and Browne [2] apply standard and nonnegative PARAFAC to a term $\times$ author $\times$ time array of email messages to automatically detect conversations, including topics, participants, and temporal activity.

The history of tensor decompositions in general goes back at least forty years [40, 16, 8], and they have been used in increasing frequency in other domains, especially chemometrics [36]. Early work on algebraic analysis of asymmetric structure includes, e.g., [15, 10, 17], and significant proposals continue to appear over time, including very recently [31]. Unfortunately, space limitations have forced us to restrict discussion to the most direct precursors of our current work.

### 2.2 Social network analysis

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.

Recently, there has been research analyzing the social networks detectable in the Enron email corpus. Diesner and Carley [13] 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 apparently formed a tight clique.

Chapanond et al. [9] analyzed the Enron corpus for structures within the organization. They used graph theoretical and spectral analysis techniques to identify communities.

McCallum et al. [30] 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 Enron corpus to learn discussion topics based on the directed interactions and relationships between people and their communications.

## 3 Models and algorithms

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 $i$th entry of $\mathbf{a}$ is denoted by $a_i$. Matrices are denoted by boldface capital letters, e.g., $\mathbf{A}$. The $j$th column of $\mathbf{A}$ is denoted by $\mathbf{a}_j$ and element $(i, j)$ by $a_{ij}$.

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

The symbol $\otimes$ denotes the matrix 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.

### 3.1 ASALSAN for three-way DEDICOM

Three-way DEDICOM is similar to the two-way model (1) in that the asymmetry relationships are in a matrix $\mathbf{R}$, but in addition 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 quasi-multilinear models called PARATUCK2 [21], 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 way factor analysis is extended to three ways by PARAFAC [16] and confers the same kind of special uniqueness property. With a unique solution, the factors are plausibly a valid description with greater reason to believe that they have more explanatory meaning than a rotated two-way solution, using, e.g., VARIMAX rotation [22].

To fit the three-way DEDICOM model, one must solve the following minimization problem

$$\min_{\mathbf{A},\mathbf{R},\mathcal{D}} f(\mathbf{A}, \mathbf{R}, \mathcal{D}) \tag{3}$$

where

$$f(\mathbf{A}, \mathbf{R}, \mathcal{D}) = \sum_{k=1}^{m} \left\| \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^\mathsf{T} \right\|_F^2 \tag{4}$$

and $\mathbf{A}$ 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 for two-way DEDICOM.

There are few algorithms for solving (3); in addition, these algorithms are not efficient with large, sparse arrays. Kiers [23] has presented an alternating least squares (ALS) algorithm for three-way DEDICOM. To update $\mathbf{A}$, Kiers minimizes (4) over the columns of $\mathbf{A}$, updating each column as a separate ALS subproblem. Each subproblem to compute one column of $\mathbf{A}$ involves a full eigendecomposition of a dense $n \times n$ matrix, which makes this procedure prohibitively expensive for large, sparse $\mathcal{X}$. To update $\mathcal{D}$, Kiers solves for each element of $\mathcal{D}$ with an ALS procedure. $\mathbf{R}$ is estimated by a least-squares update, which we use in our procedure.

Here we propose an alternating algorithm, which we call ASALSAN (for Alternating Simultaneous Approximation, Least Squares, and Newton), and adapt it for use on larger applications using a compression technique. Our approach for updating $\mathbf{A}$ and $\mathcal{D}$ is an improvement because it is capable of dealing with large, sparse arrays.

To begin, we either start with random initializations for $\mathbf{A}$ and $\mathbf{R}$ and set $\mathbf{D}_k = \mathbf{I}$. Or we set $\mathbf{D}_k = \mathbf{I}$, initialize $\mathbf{A}$ from an eigendecomposition of $\sum_{k=1}^{m}(\mathbf{X}_k + \mathbf{X}_k^\mathsf{T})$ and use them to compute an initial $\mathbf{R}$ as below. Then we update $\mathbf{A}, \mathbf{R}$, and $\mathcal{D}$ in an alternating fashion as follows.

1. Updating $\mathbf{A}$: We write a model that approximately 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:

$$\begin{aligned}\begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_1^\mathsf{T} & \cdots & \mathbf{X}_m & \mathbf{X}_m^\mathsf{T}\end{pmatrix} = \\ \mathbf{A}\begin{pmatrix}\mathbf{D}_1\mathbf{R}\mathbf{D}_1 & \mathbf{D}_1\mathbf{R}^\mathsf{T}\mathbf{D}_1 & \cdots & \mathbf{D}_m\mathbf{R}\mathbf{D}_m & \mathbf{D}_m\mathbf{R}^\mathsf{T}\mathbf{D}_m\end{pmatrix} \\ \begin{pmatrix}\mathbf{I}_{2m} \otimes \mathbf{A}^\mathsf{T}\end{pmatrix}.\end{aligned} \tag{5}$$

   Here $\mathbf{I}_{2m}$ is the identity matrix of size $2m \times 2m$. We approximate this nonlinear problem as a linear least squares problem by holding the $\mathbf{A}$ matrix on the right constant and computing the least squares solution for the $\mathbf{A}$ on the left using the method of normal equations,

which simplifies to

$$\mathbf{A} \leftarrow \left[\sum_{k=1}^{m} \left(\mathbf{X}_k \mathbf{A}\mathbf{D}_k \mathbf{R}^\mathsf{T}\mathbf{D}_k + \mathbf{X}_k^\mathsf{T}\mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\right)\right] \left[\sum_{k=1}^{m}(\mathbf{B}_k + \mathbf{C}_k)\right]^{-1} \quad (6)$$

where

$$\mathbf{B}_k \equiv \mathbf{D}_k\mathbf{R}\mathbf{D}_k(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{D}_k\mathbf{R}^\mathsf{T}\mathbf{D}_k, \quad (7)$$

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

This equation updates all columns of $\mathbf{A}$ simultaneously and avoids the costly eigendecomposition of Kiers' method. While the update for $\mathbf{A}$ is not guaranteed to decrease $f(\mathbf{A}, \mathbf{R}, \mathcal{D})$, the approximation works well in practice, especially close to a solution. We believe this is the case because the DEDICOM model and its transpose are considered simultaneously in the stacked representation of (5). Hence, its squared residual norm is equal to twice the value of $f$ in (4). The least squares update (6)–(8) is a good approximation when $\mathbf{A}$ is not changing much, so it improves the residual norm of (5), and the new objective value $f$ is at least as good as before. In our experience, it decreased $f$ with every iteration except for the first when starting from a random initial guess.

2. Updating $\mathbf{R}$: We use the closed form solution for $\mathbf{R}$ from Kiers [23]. It involves vectorizing $\mathcal{X}$ and $\mathbf{R}$ and stacking them in a manner such that the objective function in (3) changes to

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

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

$$\text{Vec}(\mathbf{R}) = \left(\sum_{k=1}^{m}(\mathbf{D}_k\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{D}_k) \otimes (\mathbf{D}_k\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{D}_k)\right)^{-1} \sum_{k=1}^{m} \text{Vec}(\mathbf{D}_k\mathbf{A}^\mathsf{T}\mathbf{X}_k\mathbf{A}\mathbf{D}_k). \quad (9)$$

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 for large-scale data. Because this solution for $\mathbf{R}$ minimizes (4) while holding $\mathbf{A}$ and $\mathcal{D}$ constant, it decreases $f(\mathbf{A}, \mathbf{R}, \mathcal{D})$.

3. Updating $\mathcal{D}$: We improve upon the alternating, elementwise minimization of Kiers [23] by considering a simultaneous, full-scale minimization with respect to the diagonal elements for each slice $\mathbf{D}_k$:

$$\min_{\mathbf{D}_k} \left\| \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^\mathsf{T} \right\|_F^2. \quad (10)$$

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 gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ of (10) are provided in an earlier technical report of this work [4]. 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., [12]. Non-negativity constraints on $\mathcal{D}$ are handled easily in this framework.

By alternating over each slice $\mathbf{D}_k$ and holding $\mathbf{A}$ and $\mathbf{R}$ constant, this solution for $\mathcal{D}$ decreases $f$.

The algorithm stops when the it ceases to make improvements to $\frac{f(\mathbf{A},\mathbf{R},\mathcal{D})}{\|\mathcal{X}\|_F^2}$ according to some threshold value or when it reaches a maximum number of iterations. While the update for $\mathbf{A}$ is not guaranteed to decrease (4), our experience has been that each cycle of updates for $\mathbf{A}, \mathbf{R}$, and $\mathcal{D}$ always improved (4) and converged to a stable function value.

ASALSAN 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" [21]. Thus, when the systematic structure in the data is reasonably well approximated by the DEDICOM model and the components are adequately distinct (e.g., not collinear in either $\mathbf{A}$ or $\mathcal{D}$), the uniqueness property increases the probability of correspondence between the recovered patterns and the original empirical source patterns.

When dealing with large arrays, ASALSAN uses a compression technique. 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,

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

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 (10) is the same as

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

except that $\mathbf{Q}^\mathsf{T}\mathbf{X}_k\mathbf{Q}$ 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 (9) and (10) above.

The dominant costs of ASALSAN per iteration are linear in the number of nonzeros of $\mathbf{X}_k$ and/or $\mathcal{O}(p^2 n)$ and come from the following steps: $\mathbf{A}^\mathsf{T}\mathbf{A}$, QR factorization of $\mathbf{A}$, $\mathbf{X}_k\mathbf{A}\mathbf{R}^\mathsf{T}$, $\mathbf{X}_k^\mathsf{T}\mathbf{A}\mathbf{R}$, and $\mathbf{Q}^\mathsf{T}\mathbf{X}_k\mathbf{Q}$. In contrast, the dominant costs in Kiers' ALS algorithm [23] come from updating $\mathbf{A}$ with $p$ diagonalizations of a dense $n \times n$ matrix, costing $\mathcal{O}(pn^3)$.

### 3.2 Nonnegative ASALSAN

Because we often deal with nonnegative data in $\mathcal{X}$, it sometimes helps to examine decompositions that retain the nonnegative characteristics of the original data. So we have modified ASALSAN to compute a three-way DEDICOM model with non-negativity constraints on $\mathbf{A}$, $\mathbf{R}$, and $\mathcal{D}$. We call this algorithm NN-ASALSAN, for "nonnegative" ASALSAN. Modifications to the updates of both $\mathbf{A}$ and $\mathbf{R}$ are made as follows: we replace the least squares solution with the multiplicative update introduced in [29] as implemented in [2]. Specifically, we modify the step to solve for the $\mathbf{A}$ appearing on the left in (5):

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

where $\mathbf{B}_k$ and $\mathbf{C}_k$ are the same as in (7)-(8) above and $\epsilon$ is a small number like $10^{-9}$. The solution for $\mathbf{R}$ is given by:

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

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

An algorithm for a nonnegative two-way DEDICOM model follows directly from NN-ASALSAN 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.

**Table 1. Time in seconds per iteration (average number of iterations) on both data sets.**

| Algorithm | World trade | | Enron | |
|---|---|---|---|---|
| ASALSAN | 0.069 | (50) | 0.85 | (184) |
| NN-ASALSAN | 0.083 | (47) | 1.0 | (74) |
| Kiers [23] | 0.022 | (67) | 22.3 | (400+) |

## 4 Experimental results

We consider two applications: a small example using the international trade data used previously in [20] and the larger email graph of the Enron corporation that was made public during the federal investigation.

ASALSAN was written in MATLAB, using the Tensor Toolbox [5, 6, 7], and Kiers' algorithm [23] was compiled Pascal code obtained from the author. All tests were performed on a dual 3GHz Pentium Xeon desktop computer with 2GB of RAM.

Table 1 shows the timings per iteration and average number of iterations to satisfy a tolerance of $10^{-5}$ (World trade) or $10^{-7}$ (Enron) in the change of fit for the three algorithms (using the same stopping criteria). We suspect the performance gap on the world trade example is due to more overhead in our MATLAB code relative to Kiers' compiled executable. Due to the poor asymptotic scalability of Kiers' algorithm on the larger Enron data, its running time is much slower than ASALSAN. Processing larger data sets, the discrepancy will grow even larger.

A practice in some applications of DEDICOM is to ignore the diagonal entries of each $\mathbf{X}_k$ in the minimization of (3). For both applications, this makes sense because we wish to ignore self-loops (i.e., no self-trade or sending email to yourself). We use an imputation technique of estimating the diagonal values from the current approximation $\mathbf{A}\mathbf{D}_k\mathbf{R}\mathbf{D}_k\mathbf{A}^\mathsf{T}$ at each iteration and including them in $\mathbf{X}_k$.

### 4.1 World trade

For a simple algorithmic comparison, we tested the international trade data of [20]. The data consists of import/export data among 18 nations in Europe, North America, and the Pacific Rim from 1981 to 1990. A semantic graph of this data corresponds to a dense adjacency array $\mathcal{X}$ of size $18 \times 18 \times 10$.

We computed a three component ($p = 3$) model using ASALSAN and Kiers' algorithm, and the same minimizers may be found among the results of both algorithms. We also used NN-ASALSAN to compute a new fully-nonnegative version of the DEDICOM model ($\mathbf{A}, \mathbf{R}, \mathcal{D} \geq 0$). Because the nonnegative results are more easily interpreted and are new, we report just these results.

| | #1 | #2 | #3 |
|---|---|---|---|
| #1 North America | 4589 | 187 | 178 |
| #2 Europe | 126 | 896 | 89 |
| #3 Japan | 60 | 168 | 37 |

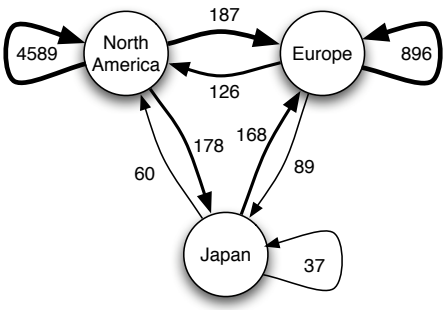


**Figure 3. World trade: $\mathbf{R}$ matrix from NN-ASALSAN and associated graph showing aggregate trade patterns.**

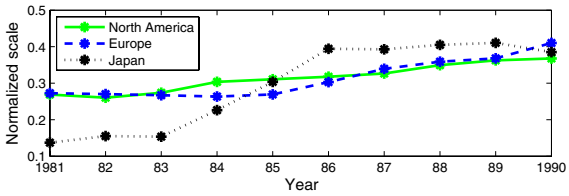


**Figure 4. Scales in $\mathcal{D}$ for trade regions indicate the level of commerce over time.**

The $\mathbf{A}$ matrix identifies nations that tend to have the same patterns of trade. In our analysis, the three latent dimensions correspond mostly to geographical regions. The first component identifies North American countries, dominated by the US and Canada. The second component contains the European countries lead by Germany, France, the Netherlands, Italy, Belgium, and the UK. The third component is dominated by Japan but also includes small participation from the UK and Italy. Given world geography and modes of shipment, these three latent groupings make sense.

The aggregate trade patterns over the ten years among these three regions is summarized in the $\mathbf{R}$ matrix and its corresponding directed graph in Figure 3. From the self-loops, we can see a large amount of trade within North America (between the US and Canada) and within Europe. Trade imbalances are also evident by the asymmetry of $\mathbf{R}$. For example, during this time period, Japan exports more to Europe than it imports from Europe.

The scales in $\mathcal{D}$ indicate the strength of each region's world commerce over time. Figure 4 shows these scales over the ten years. All curves are trending up due to economic expansion during this time. Of particular interest is Japan's rapidly ascending curve, which we believe is due to its economic expansion following the recession in the early 1980's.

## 4.2 Enron email

The whole Enron email collection is available online [11] 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. We considered messages only in the interval 13-Nov-1998 through 21-Jun-2002, which resulted in an email graph of 34,427 messages over 44 months. Our final graph corresponds to a sparse adjacency array $\mathcal{X}$ of size $184 \times 184 \times 44$ with 9838 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. Other researchers using the Enron corpus have had this same problem, and information on some participants has been collected and 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 [35]. It provides most employees' position and business unit. To facilitate a better analysis of our results, we collected extra information on the participants from the email messages themselves and found some relevant information posted on the FERC website [14]. 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 ASALSAN is able to align employ-

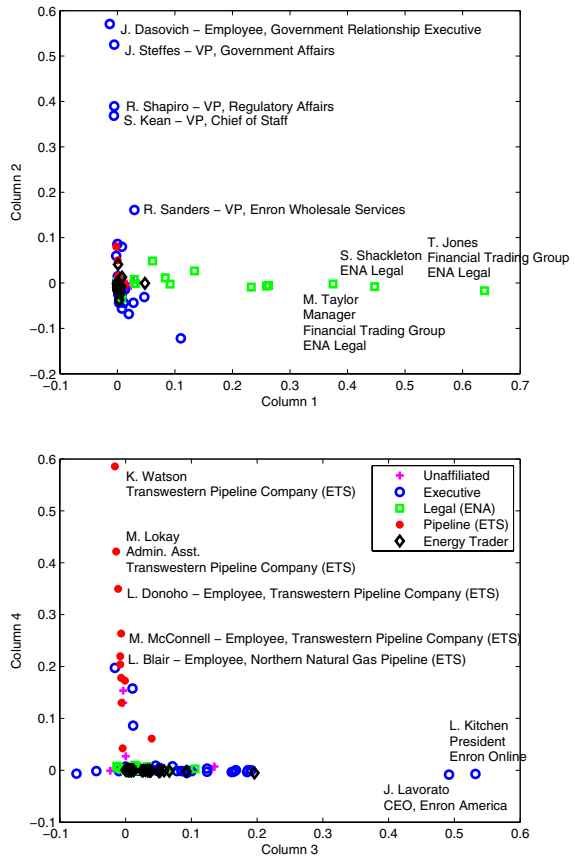


**Figure 5. Scatter plots of the first and second columns of $\mathbf{A}$ (top) and third and fourth columns of $\mathbf{A}$ (bottom) .**

ees according to their business unit and identify many of these dual roles. Next we summarize our findings of using ASALSAN and NN-ASALSAN to analyze the Enron email network.

We computed a four-component ($p = 4$) decomposition of the adjacency array $\mathcal{X}$ using ASALSAN. This is a difficult optimization problem, and 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 5 plots the four columns of the $\mathbf{A}$ matrix. 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 than is found in two-way DEDICOM [4].

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 the fourth role is the pipeline employees. The energy traders 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.

| | #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 |

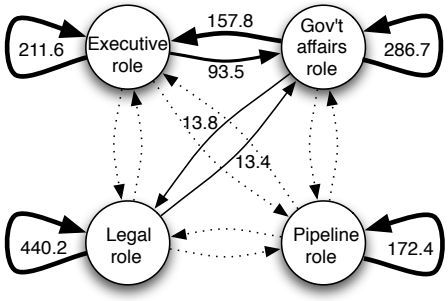


**Figure 6. Enron corpus: $\mathbf{R}$ matrix and associated graph showing aggregate communication patterns.**

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 6. 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 the pipeline employees interacted almost exclusively with themselves. We interpret the negative off-diagonal elements as having less communication than one would expect from a typical null hypothesis, which suggests 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

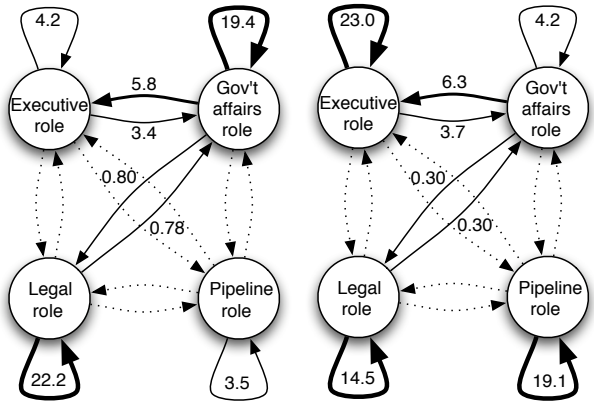


**Figure 8. Graphs of $\mathrm{D}_k\mathrm{R}\mathrm{D}_k$ showing communication patterns for $k =$ October 2000 (pre-crisis, left) and $k =$ October 2001 (during crisis, right).**

| | #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 |

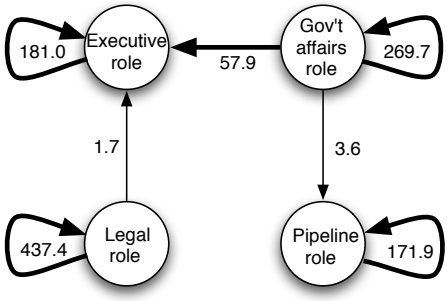


**Figure 9. Enron corpus: $\mathrm{R}$ matrix from NN-ASALSAN and associated graph showing aggregate communication patterns.**

October 2000 through October 2001, after which there is a drop-off. 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 [13].

To see the communication patterns within a particular year, we multiply $\mathbf{R}$ on the left and right by the slices of array $\mathcal{D}$. For example, Figure 8 shows the communication patterns among the four roles in $\mathbf{A}$ in October, 2000 and October, 2001. These two time periods were analyzed in [13] 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.

Here, we comment on the results for different values of $p$. Proceeding from lower- to higher-component solutions, ASALSAN 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 has a different temporal communication pattern. The 6-component model adds the energy traders.

Next, we computed a four-component ($p = 4$) nonnegative decomposition ($\mathbf{A}, \mathbf{R}, \mathcal{D} \geq 0$) of the adjacency array $\mathcal{X}$ using NN-ASALSAN. We chose the smallest minimizer from among 40 runs from random starting points, and the relative residual norm was 0.885 (excluding diagonal).

Qualitatively, the scatter plots of the columns of $\mathbf{A}$ are similar to Figure 5 and are not shown here. The scales in $\mathcal{D}$, 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 $\mathbf{R}$ matrix is more easily interpreted. Figure 9 shows the $\mathbf{R}$ matrix and its corresponding graph. It is clear from this graph that communication generally "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 nonnegative model yields similar roles as identified by ASALSAN.

According to the DEDICOM model, the $i$th row of $\mathbf{A}$ can be considered as scores of how strongly the $i$th employee is associated with each role. In other words, $a_{ij}$ is the strength of the association between employee $i$ and role $j$. Next we quantify the accuracy of these assignments.

We had independently labeled the four latent roles in $\mathbf{A}$ as executive, legal, trading, and pipeline. For each employee for which we obtained a true label (we did not consider the "unaffiliated" employees), Table 2 compares this label against the prediction made by ASALSAN. It should be noted that ASALSAN identified a "government affairs" role that did not directly correspond to the job titles we had. Since there was no "trader" role identified, we omit those employees from the tables for the three-way models. We computed the percentage of each true job type that was correctly predicted by the top one or two predictions from ASALSAN.

Note that while several employees had dual roles, we had

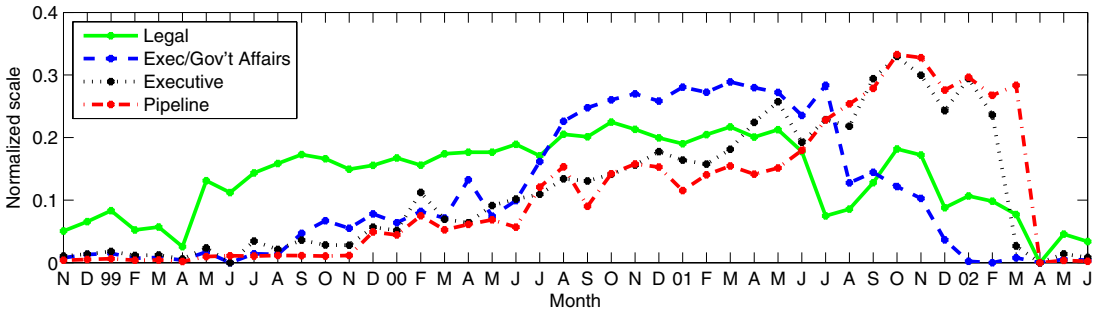


**Figure 7. Scales in $\mathcal{D}$ indicate the strength of participation of each role's communication over time.**

**Table 2. Percent of employees matching their actual business unit and job title label based on their primary and primary/secondary latent role assignments.**

| True label | Highest score | 1st and 2nd highest score |
|---|---|---|
| **ASALSAN** | | |
| Executive | 75% | 95% |
| Legal | 73% | 80% |
| Pipeline | 62% | 77% |
| Overall | 73% | 89% |
| **NN-ASALSAN** | | |
| Executive | 73% | 93% |
| Legal | 73% | 87% |
| Pipeline | 62% | 85% |
| Overall | 71% | 90% |

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, ASALSAN 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 89–90% if the two highest scoring roles are considered.

## 5 Conclusions and discussion

ASALSAN is a new algorithm for fitting a three-way DEDICOM model (optionally with non-negativity constraints) that scales to large, sparse data. We have shown some of its capabilities in analyzing temporal data in international trade and communications. The matrix $\mathbf{R}$ captures the asymmetry of the original data, offering an idealized version of a directed graph involving the latent components identified in $\mathbf{A}$, and the array $\mathcal{D}$ describes the associated temporal patterns.

ASALSAN may derive useful information from any directed graph. With its capacity to handle large-scale data, new applications include analyzing web traffic between servers over time or a web/citation graph over time.

We suggest two extensions to ASALSAN that we intend to pursue. First, constrained DEDICOM [24] is an extension of DEDICOM that has been suggested in the 1990's and pursued more recently [33]. The idea is to constrain the $\mathbf{A}$ factors themselves so that the columns of $\mathbf{A}$ lie in a prescribed column space to include domain knowledge or incorporate human understanding into the problem. 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. Second, DEDICOM has been applied to skew-symmetric data [19] and has yielded additional insight in asymmetric problems that we believe would be useful in large-scale applications.

## Acknowledgments

We thank Henk Kiers for providing his Pascal code for computing the three-way DEDICOM model. Earlier versions of this work have appeared as technical reports at Sandia National Laboratories [3, 4].

## References

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

[2] B. W. Bader, M. W. Berry, and M. Browne. Discussion tracking in Eron email using PARAFAC. In M. W. Berry and M. Castellanos, editors, *Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition*. Springer. To appear.

[3] B. W. Bader, R. A. Harshman, and T. G. Kolda. Pattern analysis of directed graphs using DEDICOM: An application to enron email. Technical Report SAND2006-7744, Sandia National Laboratories, Dec. 2006.

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

[5] B. W. Bader and T. G. Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Trans. Math. Softw.*, 32(4):635–653, Dec. 2006.

[6] B. W. Bader and T. G. Kolda. Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, 2007. To appear.

[7] B. W. Bader and T. G. Kolda. MATLAB Tensor Toolbox Version 2.2, Jan. 2007. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`.

[8] 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.

[9] A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of Enron email data. In WLACS05 [41].

[10] N. Chino. A graphical technique for representing the asymmetric relationships between objects. *Behaviormetrika*, 5:23–40, 1978.

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

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

[13] J. Diesner and K. M. Carley. Exploration of communication networks from the Enron email corpus. In WLACS05 [41].

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

[15] J. C. Gower. The analysis of asymmetry and orthogonality. In J. B. et al., editor, *Recent Developments in Statistics*, pages 109–123. North Holland, Amsterdam, 1977.

[16] R. A. Harshman. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis. *UCLA working papers in phonetics*, 16:1–84, 1970. Available at `http://publish.uwo.ca/~harshman/wpppfac0.pdf`.

[17] 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`.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] H. A. L. Kiers. Personal communication, 2006.

[24] H. A. L. Kiers and Y. Takane. Constrained DEDICOM. *Psychometrika*, 58(2):339–355, June 1993.

[25] 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.

[26] 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.

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

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

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

[30] 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 [41].

[31] A. Okada and T. Imaizumi. Multidimensional scaling of asymmetric proximities with a dominance point. In R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis*, pages 307–318. Springer Verlag, Berlin, 2007.

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

[33] R. Rocci. A general algorithm to fit constrained DEDICOM models. *Stat. Meth. App.*, 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. Ex employee status report, 2005. `http:www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls`.

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

[37] J. Sun, S. Papadimitriou, and P. S. Yu. Window-based tensor analysis on high-dimensional and multi-aspect streams. In *ICDM06: Proceedings of the 6th IEEE Conference on Data Mining*, pages 1076–1080. IEEE Computer Society, 2006.

[38] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, 2006.

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

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

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