Chapter 5

Nonnegative Matrix and Tensor Factorization for Discussion Tracking

Brett W. Bader, Michael W. Berry, and Amy N. Langville

5.1 Introduction

After the filing for Chapter 11 bankruptcy by Enron in December of 2001, an unprecedented amount of information (over 1.5 million electronic mail messages, phone tapes, internal documents) was released into the public domain. Such information served the needs of the Federal Energy Regulatory Commission (FERC) in its investigation against Enron. The emails originally posted on the FERC web site (18) had various integrity problems which required some cleaning as well as the removal of sensitive (private) and irrelevant information. Dr. William Cohen and his research group at Carnegie Mellon University have addressed many of these problems in their release of the Enron Email Sets. The version of the Enron Email Sets[1] dated March 2, 2004 contains 517,431 email messages of 150 Enron email accounts covering a period from December 1979 through February 2004 with the majority of messages spanning the three years: 1999, 2000, and 2001.

The emails in this corpus reflect the day-to-day activities of what was the seventh largest company in the United States at that time. There were, however, certain topics of discussion uniquely linked to Enron activities (5). Enron's development of the Dabhol Power Company (DPC) in the Indian

[1] http://www-2.cs.cmu.edu/~enron

state of Maharashtra (involving years of logistical and political problems) was one such topic. The deregulation of the California energy market and the subsequent rolling blackouts during the summer of 2000 was another topic. The infamous practices of greed, overspeculation, and deceptive accounting, which led to the collapse of Enron in the fourth quarter of 2001, are also documented in the emails. The corpus not only facilitates the study of employee communications within a sizeable network, but it also offers a more detailed view of how large multinational corporations operate on a daily basis.

5.1.1 Extracting Discussions

The goal of this study is to extract meaningful threads of discussion from subsets of the Enron Email Set. The underlying idea is as follows. Suppose we extract a collection of q emails from n authors over a period of p days (or other unit of time). In aggregate, there are a collection of m terms parsed from the q emails. From this data, suppose we create an $m \times n \times p$ term-author-day array[2] $\mathcal{X}$. We then decompose $\mathcal{X}$ using a nonnegative tensor factorization based on PARAFAC to track discussions over time. With some effort, the three-way term-author-day array can be expanded to a four-way term-author-recipient-day array $\mathcal{Y}$ whereby the recipients of the emails (which may or may not be from the list n authors) are also identified. A subsequent nonnegative tensor factorization of $\mathcal{Y}$ would facilitate the tracking of topics through time among different social groups.

In the next section, we provide background information (and related work) on tensor decompositions. Section 5.2 explains the notations used to define these decompositions and algorithms that are given in Section 5.3. Details of the specific Enron subset used in this study are provided in Section 5.4, followed by observations and results obtained from the application of PARAFAC to the subset in Section 9.6. Section 5.6 discusses a visualization approach for identifying clusters in the nonnegative factorizations, which is applied here to the nonnegative matrix factorization. We conclude with a brief discussion of future work in the use of nonnegative tensor factorization for topic/discussion tracking in Section 5.7.

5.1.2 Related Work

For the past forty years, tensor decompositions (38; 19; 11) have been used extensively in a variety of domains, from chemometrics (35) to signal processing (34). PARAFAC is a three-way decomposition that was proposed by Harshman (19) using the name PARAllel FACtors or PARAFAC, while

[2]Note that the array $\mathcal{X}$ is generally sparse due to the word distribution used by each author over time.

Carroll and Chang (11) published the same mathematical model under the name Canonical Decomposition or CANDECOMP. A comprehensive review by Kolda and Bader (22) summarizes these tensor decompositions and provides references for a wide variety of applications using them.

In the context of text analysis and mining, Acar et al. (1) used various tensor decompositions of (user × key word × time) data to separate different streams of conversation in chatroom data. Several web search applications involving tensors relied on query terms or anchor text to provide a third dimension. Sun et al. (36) have used a three-way Tucker decomposition (38) to analyze (user × query term × web page) data for personalized web search. Kolda et al. (23) and Kolda and Bader (21) have used PARAFAC on a (web page × web page × anchor text) sparse, three-way tensor representing the web graph with anchor-text-labeled edges to get hub/authority rankings of pages related to (identified) topics.

Regarding use of nonnegative PARAFAC, Mørup et al. (27) have studied its use for EEG-related applications. They used the associated multiplicative update rule for a least squares and Kulbach-Leibler (KL) divergence implementation of nonnegative PARAFAC, which they called NMWF-LS and NMWF-KL, respectively. FitzGerald et al. (15) and Mørup et al. (26) both used nonnegative PARAFAC for sound source separation and automatic music transcription of stereo signals.

Bader, Berry, and Browne (5) described the first use of a nonnegative PARAFAC algorithm to extract and detect meaningful discussions from email messages. They encoded one year of messages from the Enron Email Set into a sparse term-author-month array and found that the nonnegative decomposition was more easily interpretable through its preservation of data nonnegativity in the results. They showed that Gantt-like charts can be constructed/used to assess the duration, order, and dependencies of focused discussions against the progression of time. This study expands upon that work and demonstrates the first application of a four-way term-author-recipient-day array for the tracking of targeted threads of discussion through time.

5.2 Notation

Three-way and higher multidimensional arrays or tensors are denoted by boldface Euler script letters, e.g., $\mathcal{X}$. An element is denoted by the requisite number of subscripts. For example, element (i, j, k, l) of a fourth-order tensor $\mathcal{X}$ is denoted by x_{ijkl}.

The symbol $\circ$ denotes the tensor outer product,

$$A_1 \circ B_1 = \begin{pmatrix} A_{11}B_{11} & \cdots & A_{11}B_{m1} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} & \cdots & A_{m1}B_{m1} \end{pmatrix}.$$

The symbol $*$ denotes the Hadamard (i.e., elementwise) matrix product,

$$A * B = \begin{pmatrix} A_{11}B_{11} & \cdots & A_{1n}B_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{m1} & \cdots & A_{mn}B_{mn} \end{pmatrix}.$$

And the symbol $\odot$ denotes the Khatri-Rao product (columnwise Kronecker) (35),

$$A \odot B = \begin{pmatrix} A_1 \otimes B_1 & \cdots & A_n \otimes B_n \end{pmatrix},$$

where the symbol $\otimes$ denotes the Kronecker product.

The concept of *matricizing* or *unfolding* is simply a rearrangement of the entries of $\mathcal{X}$ into a matrix. We will follow the notation used in (35), but alternate notations exist. For a four-way array $\mathcal{X}$ of size $m \times n \times p \times q$, the notation $X^{(m \times npq)}$ represents a matrix of size $m \times npq$ in which the n-index runs the fastest over the columns and p the slowest. Many other permutations, such as $X^{(q \times mnp)}$, are possible by changing the row index and the fastest-to-slowest column indices.

The norm of a tensor, $\| \mathcal{X} \|$, is the square root of the sum of squares of all its elements, which is the same as the Frobenius norm of any of the various matricized arrays.

5.3 Tensor Decompositions and Algorithms

While the original PARAFAC algorithm was presented for three-way arrays, it generalizes to higher-order arrays (22). Earlier text analysis work using PARAFAC in (5) focused on the three-way case, but here we present the four-way case because our application also pertains to four-way data.

Suppose we are given a tensor $\mathcal{X}$ of size $m \times n \times p \times q$ and a desired approximation rank r. The goal is to decompose $\mathcal{X}$ as a sum of vector outer products as shown in Figure 5.1 for the three-way case. It is convenient to group all r vectors together in factor matrices A, B, C, and D, each having r columns. The following mathematical expressions of this model use different

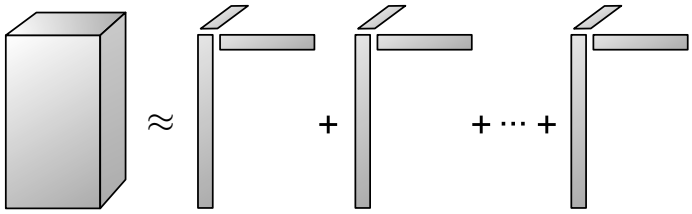

FIGURE 5.1: PARAFAC provides a three-way decomposition with some similarity to the singular value decomposition.

notations but are equivalent:

$$\begin{aligned} x_{ijkl} &\approx \sum_{t=1}^{r} A_{it} B_{jt} C_{kt} D_{lt}, \\ \mathcal{X} &\approx \sum_{t=1}^{r} A_t \circ B_t \circ C_t \circ D_t, \\ X^{(m \times npq)} &\approx A(D \odot C \odot B)^T. \end{aligned} \tag{5.1}$$

Without loss of generality, we typically normalize all columns of the factor matrices to have unit length and store the accumulated weight (i.e., like a singular value) in a vector λ:

$$\mathcal{X} \approx \sum_{t=1}^{r} \lambda_t (A_t \circ B_t \circ C_t \circ D_t).$$

It is common practice to order the final solution so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. In the discussion that follows, we describe a general algorithm for a four-way model without λ because this normalization can be performed in a post-processing step.

Our goal is to find the best fitting matrices A, B, C, and D in the minimization problem:

$$\min_{A,B,C,D} \left\| \mathcal{X} - \sum_{t=1}^{r} A_t \circ B_t \circ C_t \circ D_t \right\|^2. \tag{5.2}$$

The factor matrices are not required to be orthogonal and, in fact, are usually not in most practical applications. Under mild conditions, PARAFAC provides a unique solution that is invariant to factor rotation (19).

Given a value $r > 0$ (loosely corresponding to the number of distinct topics or conversations in our data), PARAFAC finds matrices $A \in \mathbb{R}^{m \times r}$, $B \in \mathbb{R}^{n \times r}$, $C \in \mathbb{R}^{p \times r}$, and $D \in \mathbb{R}^{q \times r}$ to yield Equation (5.1). Each group $\{A_j, B_j, C_j, D_j\}$, for $j = 1, \ldots, r$, defines scores for a set of terms, authors, recipients, and time for a particular conversation in our email collection; the value λ_r after normalization defines the weight of the conversation. (Without loss of generality, we assume the columns of our matrices are normalized to

have unit length.) The scales in D indicate the activity of each conversation topic over time.

5.3.1 PARAFAC-ALS

A common approach to solving Equation (5.2) is an alternating least squares (ALS) algorithm (19; 13; 37), due to its simplicity and ability to handle constraints. At each inner iteration, we compute an entire factor matrix while holding all the others fixed.

Starting with random initializations for A, B, C, and D, we update these quantities in an alternating fashion using the method of normal equations. The minimization problem involving A in Equation (5.2) can be rewritten in *matrix* form as a least squares problem (13):

$$\min_A \left\| X^{(m \times npq)} - AZ \right\|^2, \tag{5.3}$$

where $Z = (D \odot C \odot B)^T$.

The least squares solution for Equation (5.3) involves the pseudo-inverse of Z:

$$A = X^{(m \times npq)} Z^{\dagger}.$$

Conveniently, the pseudo-inverse of Z may be computed in a special way that avoids computing $Z^T Z$ with an explicit Z (35), so the solution to Equation (5.3) is given by:

$$A = X^{(m \times np)}(D \odot C \odot B)(B^T B * C^T C * D^T D)^{-1}.$$

Furthermore, if $\mathcal{X}$ is sparse, then the product $X^{(m \times npq)}(D \odot C \odot B)$ may be computed efficiently (3) without explicitly forming $D \odot C \odot B$. Thus, computing A essentially reduces to several matrix inner products, sparse tensor-matrix multiplication of B, C, and D into $\mathcal{X}$, and inverting an $R \times R$ matrix.

Analogous least-squares steps may be used to update B, C, and D.

5.3.2 Nonnegative Tensor Factorization

When analyzing nonnegative data, such as scaled term frequencies, it is desirable for the decompositions to retain the nonnegative characteristics of the original data and thereby facilitate easier interpretation (24). Just as with matrix factorization, it is possible to impose nonnegativity constraints on tensor factorizations.

Several authors have considered nonnegative tensor factorizations (NTF), and the resulting methods can be categorized into four classes of algorithms:

1. Least squares updates where all negative values are truncated to zero (10),

2. Nonnegative least squares (10; 16),

3. Paatero's penalty function approach (29; 28), and

4. Lee-and-Seung-style (24) multiplicative updates (39; 32; 20).

The first class is not recommended because one does not obtain least squares estimates, meaning that the residual error may increase. Hence, when employing such a technique in an iterative, multi-way algorithm such as PARAFAC-ALS, the algorithm may actually diverge (10). The three remaining classes of algorithms have better convergence properties, and nonnegative least-squares approaches solve a bound-constrained linear least squares problem. Paatero's PMF3 algorithm (28) uses a logarithmic penalty function and solves for all modes simultaneously using a Gauss-Newton approach, which enjoys fast convergence but is slower on larger problems. The multiplicative update is appealing because it is simple and fast to program, scales well with very large datasets, but it can be slow to converge.

With the exception of Paatero's PMF3, each approach harkens back to PARAFAC-ALS except that the factor matrices are updated differently. Each method generally relies on the fact that the residual norm of the various matrix formulations of the PARAFAC model are equal:

$$\begin{aligned}
&||X^{(m\times npq)} - A(D \odot C \odot B)^T||_F = \\
&||X^{(n\times pqm)} - B(A \odot D \odot C)^T||_F = \\
&||X^{(p\times qmn)} - C(B \odot A \odot D)^T||_F = \\
&||X^{(q\times mnp)} - D(C \odot B \odot A)^T||_F.
\end{aligned}$$

Each of these matrix systems may be treated as a separate nonnegative factorization problem using the techniques mentioned previously and solved in an alternating fashion.

For example, Friedlander and Hatz (16) solve each subproblem as a bound constrained linear least-squares problem. They impose sparseness constraints by regularizing the nonnegative tensor factorization with an l_1-norm penalty function. While this function is nondifferentiable, it effectively removes small values yet keeps large entries. While the solution of the standard problem is unbounded (due to the indeterminacy of scale), regularizing the problem has the added benefit of keeping the solution bounded.

Alternatively, Welling and Weber (39), and subsequently others (32; 20; 15; 27), update A using the multiplicative update introduced in (24) while

holding B, C, and D fixed, and so on:

$$A_{i\rho} \leftarrow A_{i\rho} \frac{(X^{(m\times npq)}Z)_{i\rho}}{(AZ^TZ)_{i\rho}+\epsilon}, \quad Z = (D \odot C \odot B)$$

$$B_{j\rho} \leftarrow B_{j\rho} \frac{(X^{(n\times pqm)}Z)_{j\rho}}{(BZ^TZ)_{j\rho}+\epsilon}, \quad Z = (A \odot D \odot C)$$

$$C_{k\rho} \leftarrow C_{k\rho} \frac{(X^{(p\times qmn)}Z)_{k\rho}}{(CZ^TZ)_{k\rho}+\epsilon}, \quad Z = (B \odot A \odot D)$$

$$D_{l\rho} \leftarrow D_{l\rho} \frac{(X^{(q\times mnp)}Z)_{l\rho}}{(CZ^TZ)_{l\rho}+\epsilon}, \quad Z = (C \odot B \odot A).$$

Here ϵ is a small number like 10^{-9} that adds stability to the calculation and guards against introducing a negative number from numerical underflow. Because our data is large, this is the approach that we use.

As was mentioned previously, $\mathcal{X}$ is sparse, which facilitates a simpler computation in the procedure above. The matrix Z from each step should not be formed explicitly because it would be a large, dense matrix. Instead, the product of a matricized $\mathcal{X}$ with Z should be computed specially, exploiting the inherent Kronecker product structure in Z so that only the required elements in Z need to be computed and multiplied with the nonzero elements of $\mathcal{X}$. See (3) for details.

5.4 Enron Subset

The original collection of Enron emails used in this study (and in the NTF discussed in (5)) is available online (12). Although this collection comprises 517,431 emails extracted from 150 different mail directories, we use the Enron email subset (or graph) prepared by Priebe et al. (30) that consists of messages among 184 Enron email addresses plus thirteen more that have been identified in (6) as interesting. We considered messages only in 2001, which resulted in a total of 53, 733 messages over 12 months (messages were sent on a total of 357 days).

As discussed in (5), the lack of information on the former Enron employees has hampered the performance evaluation of any model of the Enron Email Set. Having access to a corporate directory or organizational chart of Enron at the time of these emails (at least for the year 2001) would greatly help test the validity of results (via PARAFAC or any other model). Other researchers using the Enron Email Set have had this same problem. Hopefully, in time, more historical information will be available. Illustrations of the true/false

positive rates of NTF-based classification on a different dataset are discussed in (5).

The Priebe dataset (30) provided partial information on the 184 employees of the small Enron network, which appears to be based largely on information collected by Shetty and Adibi (33). Most of the employees' position and business unit data is provided. Additional employee information was collected from the email messages themselves and from relevant information posted on the FERC website (14). To further help our assessment of results, we searched for corroborating information of the preexisting data or for new identification information, such as title, business unit, or manager. Table 5.1 lists eleven of the most notable authors (and their titles) whose emails have been tracked (5).

TABLE 5.1: Eleven of the 197 email authors represented in the term-author-time array $\mathcal{X}$.

Name	**Email Account** (@enron.com)	**Title**
Richard Sanders	b..sanders	VP Enron Wholesale Services
Greg Whalley	greg.whalley	President
Jeff Dasovich	jeff.dasovich	Employee Government Relationship Executive
Jeffery Skilling	jeff.skilling	CEO
Steven Kean	j..kean	VP and Chief of Staff
John Lavorato	john.lavorato	CEO Enron America
Kenneth Lay	kenneth.lay	CEO
Louise Kitchen	louise.kitchen	President Enron Online
Mark Haedicke	mark.haedicke	Managing Director Legal Department
Richard Shapiro	richard.shapiro	VP Regulatory Affairs
Vince Kaminski	vince.kaminski	Manager Risk Management Head, Enron Energy Services

Aliasing of email addresses was used by some of the 197 authors in the year 2001), namely different email accounts of the form employee_id@enron.com were used by the same employee. For example, sample aliases of Vince Kaminski, one of the eleven notable authors in Table 5.1, include j.kaminski, j..kaminski, and vince.kaminski.

5.4.1 Term Weighting Techniques

In this study, we considered two datasets: three-way term-author-day and four-way term-author-recipient-day data. The three-way data correspond to a sparse array $\mathcal{X}$ of size $69157 \times 197 \times 357$ with $1,770,233$ nonzeros. The

69,157 terms were parsed from the 53,733 messages using a master dictionary of 121,393 terms created by the General Text Parser (GTP) software environment (in C++) maintained at the University of Tennessee (17). This larger set of terms was previously obtained when GTP was used to parse 289,695 of the 517,431 emails defining the Cohen distribution at CMU (see Section 7.1). To be accepted into the dictionary, a term had to occur in more than one email and more than 10 times among the 289,695 emails.

The four-way data correspond to a sparse array $\mathcal{Y}$ of size $39573 \times 197 \times 197 \times 357$ with 639,179 nonzeros. The 39,573 terms were parsed from the email messages in the same manner as for the three-way data. There are fewer terms because we are restricting the set of messages to be only those between the same 197 individuals. In the three-way set, there are more messages because many are sent to individuals outside of the set of 197.

We scaled the nonzero entries of $\mathcal{X}$ and $\mathcal{Y}$ according to a weighted frequency:

$$x_{ijk} = w_{ijk} g_i a_j,$$
$$y_{ijkl} = w_{ijkl} g_i a_j r_k,$$

where w_{ijkl} is the local weight for term i sent to recipient k by author j in day l, g_i is the global weight for term i, a_j is an author normalization factor, and r_k is a recipient normalization factor. While some scaling and normalization are necessary to properly balance the arrays, many schemes are possible.

For the three-way data, we used the scaling from a previous study in (5) for consistency. Let f_{ijk} be the number of times term i is written by author j in day k, and define $h_{ij} = \frac{\sum_k f_{ijk}}{\sum_{jk} f_{ijk}}$. The specific components of each nonzero are listed below:

Log local weight	$w_{ijk} = \log(1 + f_{ijk})$
Entropy global weight	$g_i = 1 + \sum_{j=1}^{n} \frac{h_{ij} \log h_{ij}}{\log n}$
Author normalization	$a_j = \frac{1}{\sqrt{\sum_{i,k} (w_{ijk} g_i)}}$

For the four-way data, we followed a different scheme. Let f_{ijkl} be the number of times term i is sent to recipient k by author j in day l. Define the entropy of term i by

$$e_i = -\sum_{j,k,l} f_{ijkl} \log f_{ijkl}.$$

The specific components of each nonzero are listed below:

Log local weight	$w_{ijkl} = \log(1 + f_{ijkl})$
Entropy global weight	$g_i = 1 - \frac{e_i}{\max_i e_i}$
Author normalization	$a_j = \frac{1}{\sqrt{\sum_{i,k} (w_{ijkl} g_i)^2}}$
Recipient normalization	$r_k = \frac{1}{\sqrt{\sum_{i,k} (w_{ijkl} g_i a_j)^2}}$

These weights are adapted from the well-known log-entropy weighting scheme (8) used on term-by-document matrices. The log local weight scales the raw term frequencies to diminish the importance of high frequency terms. The entropy global weight attempts to emulate an entropy calculation of the terms over all messages in the collection to help discriminate important terms from frequent, less important terms. The author and recipient normalizations help to correct imbalances in the number of messages sent from and received by each individual. Without some type of normalization, discussions involving prolific authors and/or popular recipients would tend to dominate the results.

Scaling in different ways can influence the analysis. Our scaling of the four-way data in $\mathcal{Y}$ does a decent job of balancing authors, recipients, and time. We find single spikes and some multiple spike groups, plus multiple authors communicating with multiple recipients in several cases. Other schemes may be used to focus more on single authors, recipients, or days.

5.5 Observations and Results

In this section, we summarize our findings of applying NTF on the three- and four-way versions of the Enron email collection. Our algorithms were written in MATLAB, using sparse extensions of the Tensor Toolbox (2; 3; 4). All tests were performed on a dual 3GHz Pentium Xeon desktop computer with 2GB of RAM.

5.5.1 Nonnegative Tensor Decomposition

We computed a 25-component ($r = 25$) nonnegative decomposition of the term-author-day array $\mathcal{X}$. One iteration took about 26 seconds, and the average run required about 17 iterations to satisfy a tolerance of 10^{-4} in the relative change of fit. We chose the smallest minimizer from among ten runs from random starting points, and the relative norm of the difference was 0.9561.

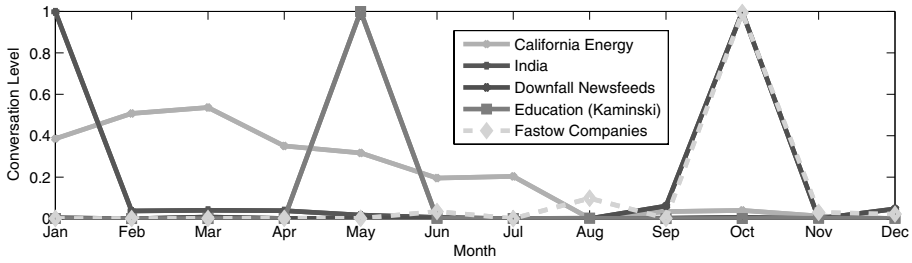

FIGURE 5.2 (SEE COLOR INSERT FOLLOWING PAGE 130.): Five discussion topics identified in the three-way analysis over months.

We also computed a 25-component ($r = 25$) nonnegative decomposition of the term-author-recipient-day array $\mathcal{Y}$. One iteration required just under 16 seconds, and between 8 and 12 iterations would satisfy a tolerance of 10^{-4} in the relative change of fit. We chose the smallest minimizer from among ten runs from random starting points, and the relative norm of the difference was 0.9716.

5.5.2 Analysis of Three-Way Tensor

PARAFAC can be used to identify and track discussions over time in each triad $\{A_j, B_j, C_j\}$, for $j = 1, \ldots, r$. A discussion or thread is associated with the topic and primary participants identified in the columns of A and B, respectively, and the corresponding column of C provides a profile over time, showing the relative activity of that discussion over 12 months or over 357 days.[3] As demonstrated in (5), discussions can be visualized as a histogram (or Gantt chart) of the monthly activity for each discussion identified by the classical and nonnegative PARAFAC models, respectively. Here, we comment on both the monthly and daily discussions that were uncovered by both models.

Qualitatively, the results of the nonnegative decomposition and the standard three-way PARAFAC were very similar. The major difference lies in the ability to interpret the results. In the 25 discussion groups tracked by PARAFAC, only six of the groups had any discernible meaning based on known Enron activities (25). In comparison, the nonnegative PARAFAC model revealed eight group discussions that could be interpreted. Figure 5.2 shows the temporal activity of some of these discussions.

The topics generated by the nonnegative PARAFAC model certainly reflected known events of the year 2001. In the first quarter of that year, Enron was still dealing with the fallout of the 2000 California energy crisis. Discussions about the Federal and California state governments' investigation of the California situation were observed as well as Enron's attempted development

[3] Eight days of the year 2001 involved no discussions for the 197 author subset used.

of the Dabhol Power Company (DPC) in the Indian State of Maharashtra. Whereas the company's efforts in India had been ongoing for several years, emails of the first six months of 2001 reflected several of the day-to-day dealings with that situation.

By October of 2001, Enron was in serious financial trouble. A merger with the Dynegy energy company fell through and forced Enron to file for Chapter 11 bankruptcy. Many of the emails in the months of October and November were newsfeeds from various organizations that were being routed through the company. As it was reported that Chief Financial Officer Andy Fastow was heavily involved with the deceptive accounting practices,[4] it is not surprising that a topic we labelled *Fastow companies* emerged. Predictably, a *college Football* topic emerged in late fall as well. One of the surprise topics uncovered was an education-related topic due in large part to the interests and responsibilities of Vince Kaminski, head of research. Kaminski taught a class at Rice University in Houston in the Spring of 2001, and was the focal point of emails about internships, class assignments, and resume evaluation (5).

Since only eight of the 25 topics had any discernible meaning, it would seem apparent that a significant amount of *noise* or undefined content can still permeate a term-author-month array. In some instances, there are indicators of a possible thread of some kind (not necessarily directly related to Enron), but a closer inspection of those emails reveals no identifiable topic of discussion.

The daily results reported in (5) provided a similar interpretation as the monthly results but at a finer resolution. In general, one observed four different types of discussions: (*i*) discussions centered largely on one or a few days, (*ii*) continual activity, represented as multiple weekly spikes in activity throughout the year, (*iii*) continual activity with lulls, where a period of calm separates bursts of discussion, and (*iv*) a series of weekly spikes of activity usually spanning three or more months.

Of the 25 discussion groups mined with the PARAFAC model, roughly half were of the first type. Examples include a flood of emails about the possible Dynegy/Enron merger (November 11 and 12th), a topic on January 7th in which Enron employees (Kean, Hughes, and Ambler) were discussing India based on an article published by Reuters and another media report, and a discussion centered on the August 27 U.S. Court of Appeals ruling on section 126 of an Environment Protection Agency code.

The nonnegative PARAFAC model identified temporal patterns similar to those of PARAFAC with a majority being a series of weekly activity spikes spanning three or more months. Roughly one third were single spikes patterns, and just two discussions are somewhat bimodal with a lull. A few of the more interesting (single spike) discussion groups extracted by the nonnegative model included a flurry of emails on August 22 in response to an email

[4]Setting up bogus companies to improve Enron's bottom line, for example.

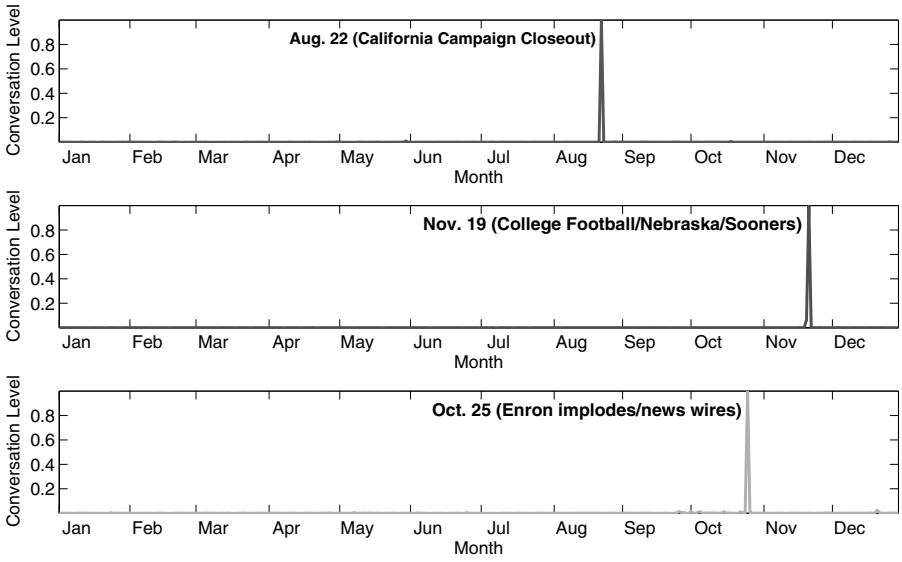

FIGURE 5.3: Three discussion topics identified in the three-way analysis over days.

with subject line *California Campaign Closeout*. In essence, Richard Shapiro praised a subset of employees who worked on California-related projects and many responded to his acknowledgement. A second discussion group identified by terms such as *college football*, *Nebraska*, *Sooners*, *bowl*, *Cougars*, and *Tennessee* was initiated by M. Motley on November 20. Finally, a third group (involving many news wire stories) described Enron's pending implosion around October 25 and 26. PARAFAC also found this topic but two days earlier—we speculate that the difference is due to the random initialization of both the PARAFAC and nonnegative PARAFAC models. Figure 5.3 shows the temporal activity of these discussions.

5.5.3 Analysis of Four-Way Tensor

When analyzing the four-way term-author-recipient-day array $\mathcal{Y}$, we observed four types of profiles over time: (*i*) discussions centered largely on one or a few days, resulting in a single spike, (*ii*) continual activity, represented as multiple weekly spikes throughout the year, (*iii*) continual activity with lulls, where a period of calm separates bursts of discussion, and (*iv*) a series of weekly spikes usually spanning three or more months.

In the analysis of the three-way $\mathcal{X}$ data, NTF identified temporal patterns that include these four cases. Roughly one third are single spikes patterns, and just two discussions are of the bimodal type with a lull. Of the 25 groups found in the four-way analysis of $\mathcal{Y}$, roughly half were single spikes. Four were double spikes in time, and nine had sustained activity over many weeks.

Previous research in (5) showed results containing a single spike in time

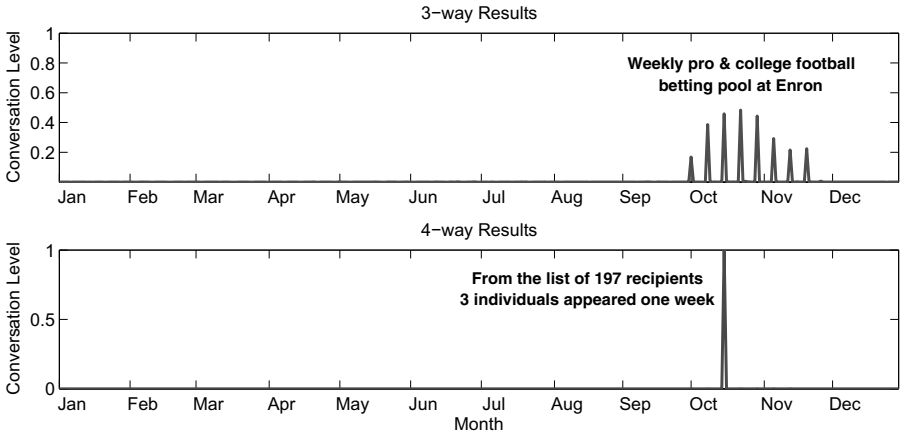

FIGURE 5.4: Weekly betting pool identified in the three-way (top) and four-way (bottom) analyses.

but not any examples that spanned some number of days. Here we present several examples of the latter type and also show what is gained in going from a three-way to four-way analysis.

Figure 5.4 shows a series of email messages announcing the results of a weekly betting pool based on the number of winning teams chosen correctly out of all pro and college football games for the week. Most of the top terms were names, but after a dozen terms more interesting terms, such as *games*, *score*, *picked*, and *prize*, start to appear. Each email lists all of the names entered in that week's pool and their record, which explains why the names appear high in the list of terms for the group.

The unusual feature of this group is that the time profile is so regular. This is because the discussion took place weekly for one day. Results of the betting pool were sent out after the conclusion of all games in the pro and college football schedules.

The four-way analysis identified this discussion but only found a single spike in time. The group showed that the organizer only sent this message to four recipients (out of 197 email addresses) in this case. Presumably the four recipients did not participate in other weeks, and none of the remaining 193 addresses participated in other weeks. If the recipient list were expanded to include others in the betting pool, then the four-way analysis might have picked up other days and recipients as well.

As a second example, Figure 5.5 shows the temporal activity for a discussion involving FERC and its rulings on RTOs. From one of the newsfeeds from `issuealert@scientech.com` on May 4, 2001 there was this description:

> "For background, an RTO is a regional entity that is designed to consolidate control and delivery of electricity across various types of transmission systems within a particular region. The origins of FERC's RTO policy dates back to its December 1999 Order 2000, in which it strongly encouraged all transmission-owning util-

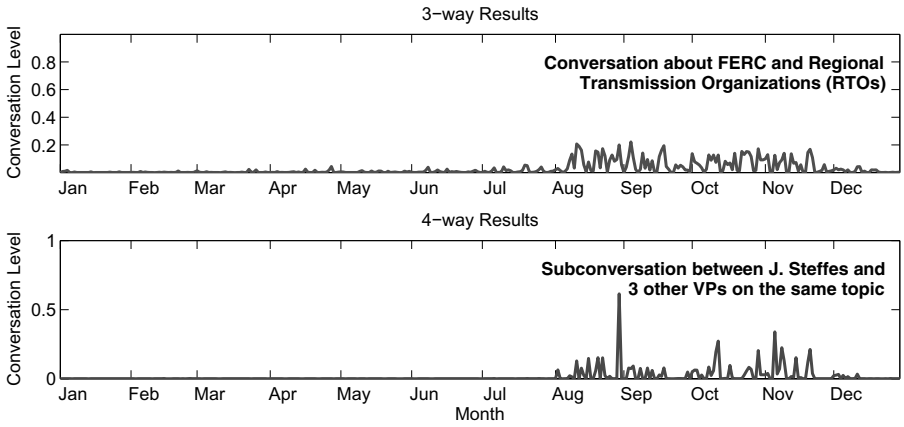

FIGURE 5.5: Long running discussion on FERC's various rulings of RTOs.

> ities to submit plans for joining or forming an RTO by Oct. 15, 2000, with actual membership established by December of this year. FERC is now sorting through the applications that it has received, and its approvals or rejections illuminate certain preferences that some members of the commission hold. Over the last year or two, FERC has engaged in an ongoing debate between its preference for transco (for-profit) models for RTOs, as opposed to independent system operators (non-profit). Chairman Curt Heacutebert has been the most vocal supporter of the transco model, while other commissioners such as William Massey have supported ISOs. However, moving forward, it is becoming increasingly clear that FERC also seems to have other set agendas for how it wants the network of RTOs to operate, including the limit of one entity per region."

S. Novosel sent email with subjects like "Subject: FERC Orders on CA and RTO West." A lot of the discussion in this group is reactions and opinions to FERC rulings. The four-way analysis identified this large conversation with many of the same terms, such as RTO, FERC, market, as well as many of the same names. What distinguishes the four-way analysis from the three-way analysis group is that it is a thread of the larger conversation involving primarily the VP's of government affairs, regulatory affairs, chief of staff and Enron wholesale services. As such the time profile of this subconversation nests within the larger conversation identified in the three-way analysis. What is gained from this four-way analysis is the direction of discussion and the recipients in this social network.

The third example in Figure 5.6 is a group identified in the four-way analysis that was not previously identified in any three-way analysis. This email exchange involves the forwarding of the Texas A&M school fight song wav file from E. Bass to four others in the list of 197 recipients. It is reasonable to suggest that perhaps these folks were A&M alumni. Alternatively, the sender

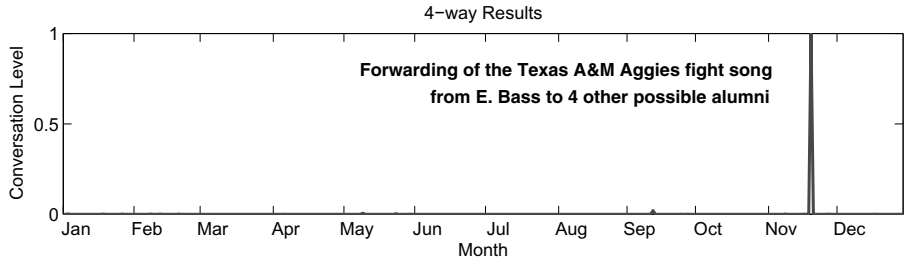

FIGURE 5.6: Forwarding of Texas A&M school fight song.

may be an alum and the four recipients went to a football game and asked "what is everyone singing?" Exposing that type of social interaction is an advantage for four-way analysis over the three-way analysis without recipients.

5.6 Visualizing Results of the NMF Clustering

The previous sections demonstrate the value of three-way and four-way tensor decompositions. Yet it is either very cumbersome or often impossible to visualize these higher-dimensional tensors. Figures 5.4–5.6 are attempts at visualizing the information provided by the tensors, yet they are somewhat limited in scope. As an alternative, in this section, we resort to the standard two-way (or matrix) decomposition to help us visualize some of the patterns uncovered by the three-way and higher decompositions. In general, one can always easily visualize any two dimensions of an n-way tensor decomposition by considering the matrix associated with those dimensions as created by the tensor decomposition. In this spirit, we discuss a tool for visualizing clusters in two-way factors.

It is well known (9) that the nonnegative matrix factorization (NMF) can be used to cluster items in a collection. For instance, if the data matrix is a term-by-document matrix X, which has been factored with the NMF as $X = AB$, then the rows of $\underline{A}$ can be used to cluster terms, while the columns of $\underline{B}$ can be used to cluster documents. As a result, terms and documents are, in some sense, clustered independently. There are two main types of clustering: hard clustering and soft clustering. Hard clustering means that items (in this case, terms and documents) can belong to only one cluster, whereas in soft clustering items are allowed to belong to multiple clusters, perhaps with varying weights for these multiple assignments. If hard clustering is employed, then cluster assignment is easy. Term i belongs to cluster j if $A(i, j)$ is the maximum element in the i^{th} row of A. Similarly, document k belongs to cluster l if $B(l, k)$ is the maximum element in the k^{th} column of B.

Once cluster assignments are available (by either hard or soft clustering), a

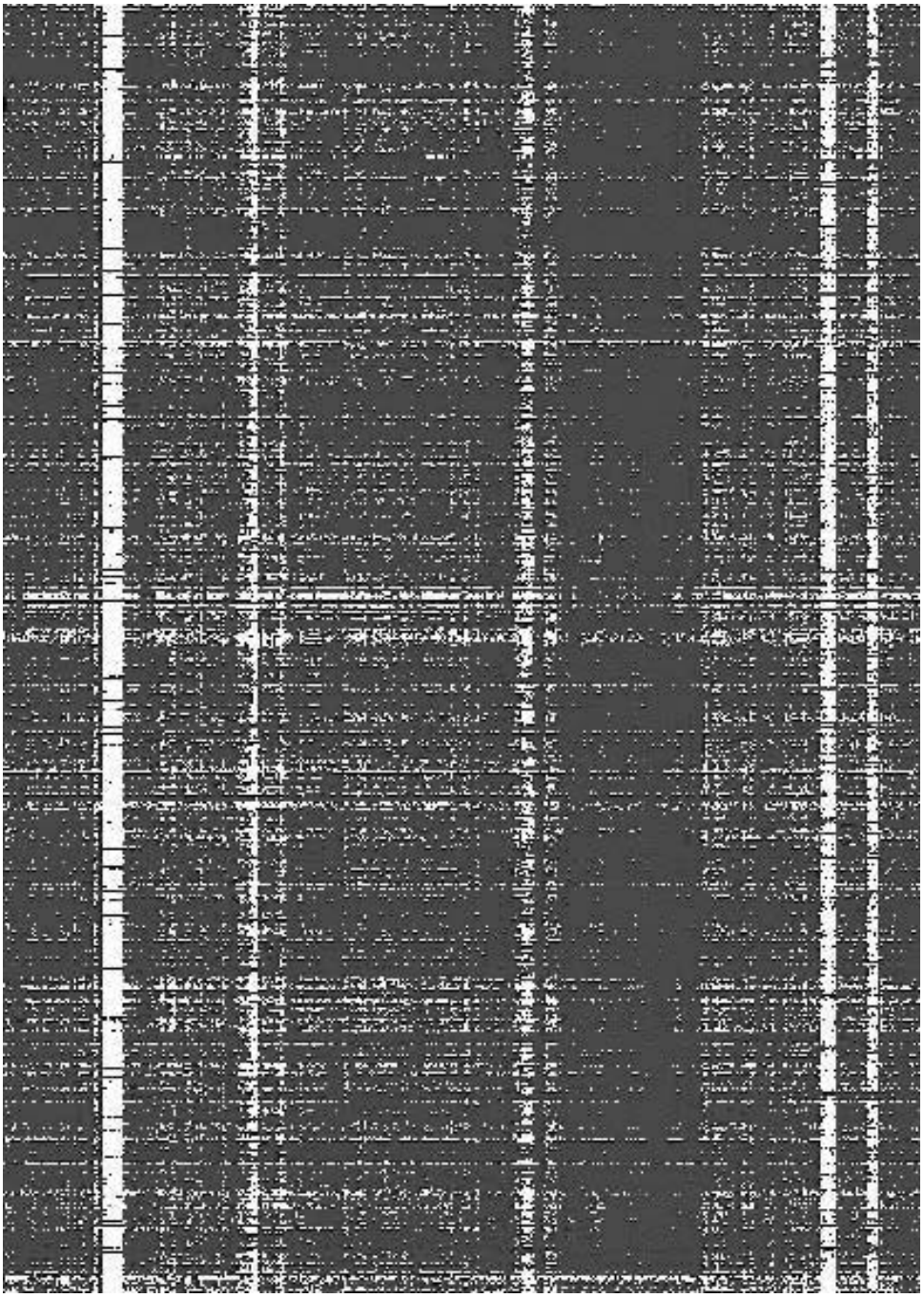

FIGURE 5.7 (SEE COLOR INSERT FOLLOWING PAGE 130.): Pixel plot of the raw Enron term-by-email matrix.

very useful next step is to display the clustering results visually. We demonstrate the value of this by considering once again the Enron email dataset described in Section 5.4. The raw term-by-email matrix for this dataset appears to have no structure, as shown in the pixel plot of Figure 5.7. Each nonzero entry in the raw matrix is represented by a pixel, and the magnitude of the entry is captured by the intensity of the pixel.

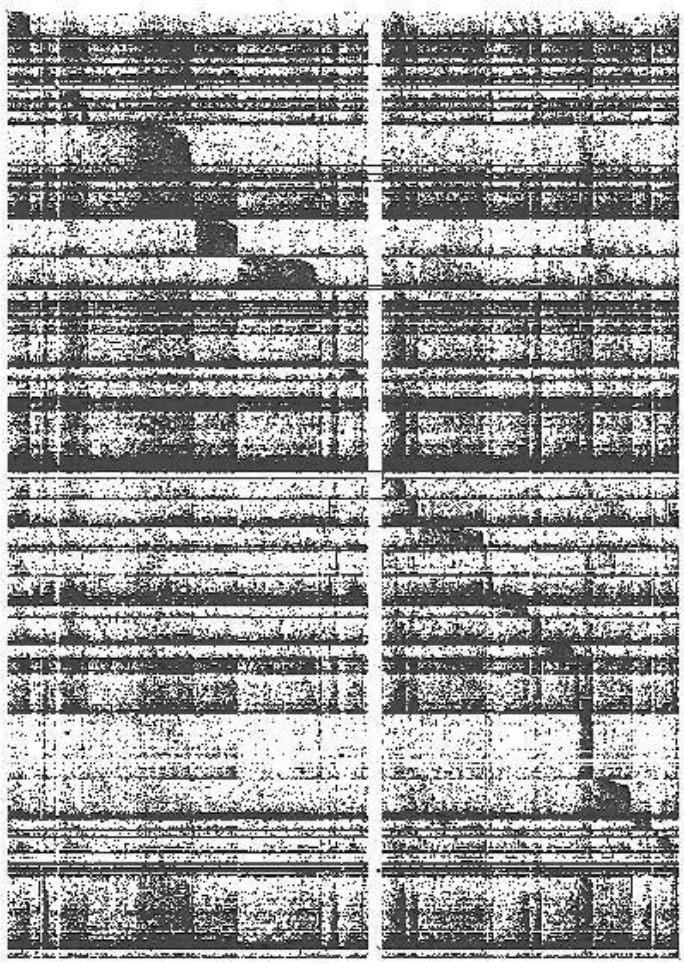

FIGURE 5.8 (SEE COLOR INSERT FOLLOWING PAGE 130.): Pixel plot of the reordered Enron term-by-email matrix.

Figure 5.8 is simply a reordered version of the raw Enron term-by-email matrix using $r = 50$ (the number of columns of A and rows of B). Both the terms and the documents were reordered according to the hard cluster assignments produced by the NMF. The nice block structure of the reordered matrix reveals the hidden clusters. For instance, a dense block means that a set of documents frequently used the same set of terms. Contrasting Figure 5.7 with Figure 5.8 reveals just how much structure was hidden in the dataset.

While the visualization of Figure 5.8, which was created with the NMF, is valuable to practitioners, an even more valuable tool allows the practitioner to more deeply examine clusters of interest and perhaps attach a meaning to the cluster. This is possible with the help of the vismatrix tool[5] created by David Gleich.

[5] http://www.stanford.edu/~dgleich/programs/vismatrix

This tool has a mouseover feature that enables a user to hold the mouse over any pixel in the matrix (reordered or otherwise) and determine which term and which document the pixel corresponds to. Figure 5.9 is a screenshot from the vismatrix tool.

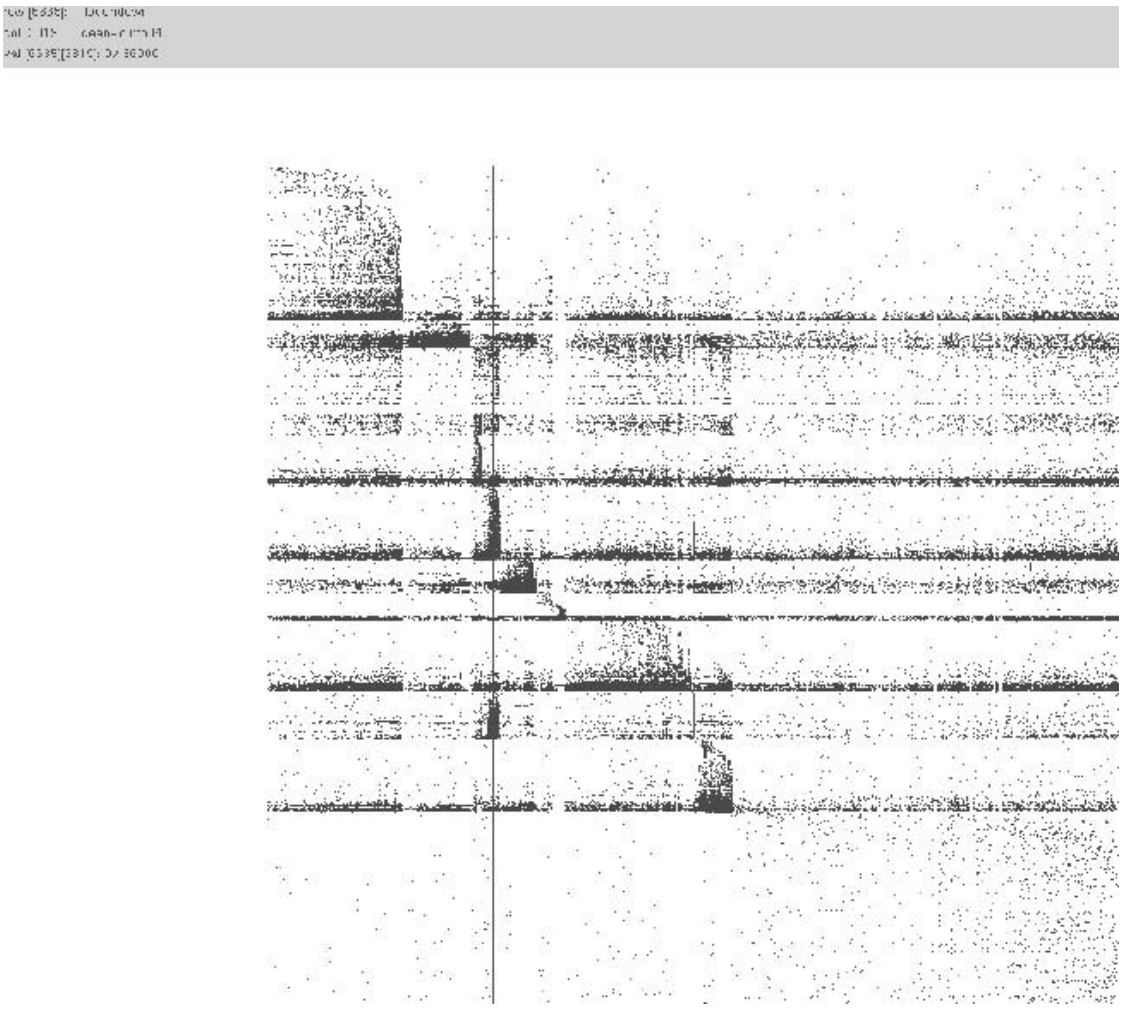

FIGURE 5.9 (SEE COLOR INSERT FOLLOWING PAGE 130.): Pixel plot of the reordered Enron term-by-document matrix with term and document labels.

Notice the upper lefthand corner contains the word `touchdown`, which represents the term (term ID#6635) being pointed to, and the identifier `dean-cinfo84`, which represents the document ID. This document, document 3819, was email message #84 saved by an Enron employee named Dean in his `cinfo` folder. Scrolling over pixels in a dense block causes the term and document labels to change in this area of the vismatrix tool. The human brain can quickly process many terms at once. As a result, the user can attach a

judgment to the quality of the clustering and can often attach a label as well. For instance, the cluster over which the yellow crosshairs of Figure 5.9 lie also contains the terms (among others) *football*, *longhorn*, *Texas*, *quarterback*, *score*, *redshirt*, *freshmen*, *punt*, and *tackle*, prompting a user to potentially label this cluster *Texas Longhorn Football*.

The vismatrix tool also allows a user to quickly scan document labels as well. Thus, hidden patterns that pertain to the documents can be found. For instance, this Enron dataset contains one small cluster of 12 documents using 447 terms. Figure 5.10 is a close-up[6] of this part of the reordered Enron term-by-email matrix.

FIGURE 5.10 (SEE COLOR INSERT FOLLOWING PAGE 130.): Close-up of one section of pixel plot of the reordered Enron term-by-document matrix.

[6]The vismatrix tool also contains zoom in and zoom out features.

Using the mouse to scroll over this small dense block reveals that the following terms (among others) are assigned to this small cluster: *fortune*, *ceo*, *coo*, *top*, *women*, and *powerful*. These terms and abbreviations, in fact, refer to Louise Kitchen (a top-ranking Enron employee responsible for energy trading and Enron Online) who was named one of the 50 most powerful women in business by Fortune Magazine in 2001. Mousing over this same small but dense block, but focusing on the document labels this time reveals that all 12 of these emails have the label `kitchen-l-americaspress#`, meaning that they were all saved in Louise Kitchen's own private `l-americaspress` folder. So what appeared to be a small possibly interesting cluster, after further inspection thanks to the vismatrix tool, is an "ego cluster," and thus perhaps of only marginal interest.

5.7 Future Work

As demonstrated by this study, nonnegative tensor factorization (implemented by PARAFAC) can be used to extract meaningful discussions from email communications. The ability to assess term-to-author (or term-to-email) associations both semantically and temporally via three-way and four-way decompositions is an important advancement in email surveillance research. Previously reported clusters of Enron emails using nonnegative matrix factorization (i.e., two-way decompositions) (7; 9; 31) were unable to extract discussions such as the *Education* thread mentioned in Section 5.5.1 or sequence the discussion of the company's downfall by source (newfeeds versus employee-generated). The optimal segmentation of *time* as a third (or fourth) dimension for email clustering may be problematic. Grouping or clustering emails by month may not be sufficient for tracking event-driven activities and so more research in the cost-benefit tradeoffs of finer time segmentation (e.g., grouping by weeks, days, or even minutes) is certainly needed. Determining the optimal tensor rank r for models such as PARAFAC is certainly another important research topic. Determining an optimal term weighting scheme for multi-way arrays is also an important task that could greatly influence the quality of results—more research on this topic is especially needed. Finally, the visualization of multi-way arrays (tensors) certainly constitutes an important area of software development that could greatly facilitate both the identification and interpretation of communications.

Acknowledgments

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. The authors would like to thank the anonymous referees for their helpful comments and suggestions on improving the original version.

References

[1] E. Acar, S. A. Çamtepe, M. S. Krishnamoorthy, and Bülent 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. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006.

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

[4] B. W. Bader and T. G. Kolda. Matlab tensor toolbox, version 2.2. http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/, January 2007.

[5] B. W. Bader, M. W. Berry, and M. Browne. Discussion Tracking in Enron Email Using PARAFAC. In M.W. Berry and M. Castellanos, editors, *Survey of Text Mining II: Clustering, Classification, and Retrieval*, pages 147–163. Springer-Verlag, London, 2008.

[6] 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.

[7] M. W. Berry and M. Browne. Email surveillance using nonnegative matrix factorization. *Computational & Mathematical Organization Theory*, 11:249–264, 2005.

[8] M. W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia, PA, second edition, 2005.

[9] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[10] R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm. *J. Chemometr.*, 11(5):393–401, 1997.

[11] 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.

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

[13] N. (Klaas) M. Faber, R. Bro, and P. K. Hopke. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometr. Intell. Lab. Syst.*, 65(1):119–137, January 2003.

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

[15] D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *ISSC 2005: Proceedings of the Irish Signals and Systems Conference*, 2005.

[16] M. P. Friedlander and K. Hatz. Computing nonnegative tensor factorizations. Technical Report TR-2006-21, Department of Computer Science, University of British Columbia, October 2006.

[17] J. T. Giles, L. Wo, and M. W. Berry. GTP (General Text Parser) Software for Text Mining. In H. Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, pages 455–471. CRC Press, Boca Raton, FL, 2003.

[18] T. Grieve. The Decline and Fall of the Enron Empire. *Slate*, October 14 2003. http://www.salon.com/news/feature/2003/10/14/enron/index_np.html.

[19] 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. Available at http://publish.uwo.ca/~harshman/wpppfac0.pdf.

[20] T. Hazan, S. Polak, and A. Shashua. Sparse image coding using a 3D non-negative tensor factorization. In *ICCV 2005: 10th IEEE International Conference on Computer Vision*, volume 1, pages 50–57. IEEE Computer Society, 2005.

[21] 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.

[22] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 2008. to appear.

[23] 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.

[24] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 21 October 1999.

[25] B. Mclean and P. Elkind. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron.* Portfolio, 2003.

[26] M. Mørup, M. N. Schmidt, and L. K. Hansen. Shift invariant sparse coding of image and music data. Technical report, Technical University of Denmark, 2007.

[27] M. Mørup, L. Hansen, J. Parnas, and S. M. Arnfred. Decomposing the time-frequency representation of EEG using nonnegative matrix and multi-way factorization. Available at http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4144/pdf/imm4144.pdf, 2006.

[28] P. Paatero. A weighted non-negative least squares algorithm for three-way "PARAFAC" factor analysis. *Chemometr. Intell. Lab. Syst.*, 38(2):223–242, October 1997.

[29] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[30] 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.

[31] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[32] A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML 2005: Machine Learning, Proceedings of the Twenty-second International Conference*, 2005.

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

[34] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro. Blind PARAFAC receivers for DS-CDMA systems. *IEEE Transactions on Signal Processing*, 48(3):810–823, 2000.

[35] A. Smilde, R. Bro, and P. Geladi. *Multi-Way Analysis: Applications in the Chemical Sciences*. Wiley, West Sussex, England, 2004.

[36] 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. ACM Press, New York, 2005.

[37] G. Tomasi and R. Bro. PARAFAC and missing values. *Chemometr. Intell. Lab. Syst.*, 75(2):163–180, February 2005.

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

[39] M. Welling and M. Weber. Positive tensor factorization. *Pattern Recogn. Lett.*, 22(12):1255–1261, 2001.