SAND2007-0640C

Discussion Tracking in Enron Email Using PARAFAC *

Brett W. Bader† Michael W. Berry‡ Murray Browne‡

January 18, 2007

Abstract

In this study, we apply a non-negative tensor factorization algorithm to extract and detect meaningful discussions from electronic mail messages for a period of one year. For the publicly released Enron electronic mail collection, we encode a sparse term-author-month array for subsequent three-way factorization using the the PARAllel FACtors (or PARAFAC) three-way decomposition first proposed by Harshman. Using non-negative tensors, we preserve natural data non-negativity and avoid subtractive basis vector and encoding interactions present in techniques such as principal component analysis. Results in thread detection and interpretation are discussed in the context of published Enron business practices and activities, and benchmarks addressing the computational complexity of our approach are provided. The resulting tensor factorizations can be used to produce Gantt-like charts that can be used to assess the duration, order, and dependencies of focused discussions against the progression of time.

1 Introduction

When Enron closed its doors on December 2, 2001 and filed for Chapter 11 bankruptcy it began a turn of events that released an unprecedented amount of information (over 1.5 million electronic mail messages, phone tapes, internal documents) into the public domain. This information was the cornerstone of the Federal Energy Regulatory Commission's (FERC) investigation against the global energy corporation. The original set of emails was posted on FERC's web site [12], but it suffered document integrity problems and attempts were made to improve the quality of the data and remove sensitive and irrelevant private information. Dr. William Cohen of Carnegie Mellon University oversaw the distribution of this improved corpus — known as the Enron Email Sets. The latest version of the Enron Email Sets[1] (dated – March 2, 2004) contains 517,431 email messages of 150 Enron employees covering a period from December 1979 through February 2004 with the majority of messages spanning the three years: 1999, 2000, and 2001.

For the most part, the emails reflect the day-to-day activities of America's seventh largest company, but certain distinct topics of discussion are linked to Enron. One involved Enron's development of the Dabhol Power Company (DPC) in the Indian state of Maharashtra, an endeavor awash in years of logistical and political problems. Another topic was the deregulation of the California energy market, which led to rolling blackouts during the summer of 2000 — a situation that Enron (and other energy companies) took advantage of financially. Eventually a combination of greed, over speculation, and deceptive accounting practices snowballed into an abrupt collapse in the fourth quarter of 2001, which again are reflected in the emails. The Enron Email Sets provides a unique opportunity not only to study the mechanics of a sizeable email network, but it also offers a glimpse of the machinations of how huge global corporations operate on a day-to-day basis.

In this research, we seek to extract meaningful threads of discussion from a subset of the Enron Email Set. The idea underlying our thread extraction is as follows. Suppose we have a collection of q emails from n authors over a period of p months. In aggregate, there are a collection of m terms parsed from the q emails. From this data, we create an $m \times n \times p$ term-author-month array[2] $\mathcal{X}$. We then decompose $\mathcal{X}$ using PARAFAC or a non-negative tensor factorization to track discussions over time.

In the next section we provide background information on tensor decompositions and related work. Section 3 provides a formal discussion of the notations used to define these decompositions and algorithms that are given in Section 4. Details of the the specific Enron sub-

*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.

†Sandia National Laboratories, Albuquerque, NM 87185

‡Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450

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

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

set used in this study are provided in Section 5, and our observations and results in applying PARAFAC to this subset of emails follow in Section 6. Finally, a brief summary of future work in the use of non-negative tensor factorization for discussion tracking is given in Section 7.

2 Related work

Tensor decompositions date back forty years [27, 13, 7], and they have been used extensively in a variety of domains, from chemometrics [24] to signal processing [23]. PARAFAC is a three-way decomposition that was proposed by Harshman [13] using the name Parallel Factors or PARAFAC. At the same time, Carroll and Chang [7] published the same mathematical model, which they call Canonical Decomposition or CANDECOMP.

The use of multidimensional models is relatively new in the context of text analysis. Acar et al. [1] use 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. [25] apply a 3-way Tucker decomposition [27] to the analysis of (user × query term × web page) data in order to personalize web search. Kolda et al. [15] and Kolda and Bader [14] use 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 an identified topic.

3 Notation

Multidimensional arrays and tensors are denoted by boldface Euler script letters, e.g., $\mathcal{X}$. Element (i, j, k) of a third-order tensor $\mathcal{X}$ is denoted by x_{ijk}.

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,

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

The symbol $\otimes$ denotes the Kronecker product,

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

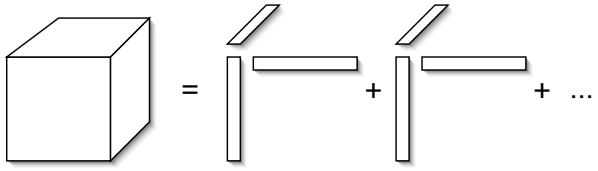

Figure 1: PARAFAC provides a 3-way decomposition with some similarity to the singular value decomposition.

And the symbol $\odot$ denotes the Khatri-Rao product (columnwise Kronecker) [24],

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

The concept of *matricizing* or *unfolding* is simply a rearrangement of the entries of $\mathcal{X}$ into a matrix. Although different notations exist, we are following the notation used in [24]. For a three-dimensional array $\mathcal{X}$ of size $m \times n \times p$, the notation $X^{(m \times np)}$ represents a matrix of size $m \times np$ in which the n-index runs the fastest over the columns and p the slowest. Other permutations, such as $X^{(p \times nm)}$, are possible by changing the row index and the fast/slow column indices.

The norm of a tensor, $\| \mathcal{X} \|$, is the same as the Frobenius norm of the matricized array, i.e., the square root of the sum of squares of all its elements.

4 Tensor decompositions and algorithms

Suppose we are given a tensor $\mathcal{X}$ of size $m \times n \times p$ and a desired approximation rank r. The goal is to decompose $\mathcal{X}$ as a sum of vector outer products as shown in Figure 1. It is convenient to group all r vectors together in factor matrices A, B, C each having r columns. The following mathematical expressions of this model use different notations but are equivalent:

$$\begin{aligned} x_{ijk} &\approx \sum_{l=1}^{r} A_{il}B_{jl}C_{kl}, \\ \mathcal{X} &\approx \sum_{l=1}^{r} A_l \circ B_l \circ C_l, \\ X^{(m \times np)} &\approx A(C \odot B)^T. \end{aligned} \tag{4.1}$$

PARAFAC may apply to general N-way data, but because our application only pertains to three-way data, we are only considering the specific three-way problem at this time.

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:

$$\mathcal{X} \approx \sum_{l=1}^{r} \lambda_l (A_l \circ B_l \circ C_l)$$

Moreover, we typically re-order the final solution so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. In the following subsections, we describe general algorithms for the 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, and C in the minimization problem:

$$\min_{A,B,C} \left\| \mathcal{X} - \sum_{l=1}^{r} A_l \circ B_l \circ C_l \right\|^2 . \tag{4.2}$$

It is important to note that the factor matrices are not required to be orthogonal. Under mild conditions, PARAFAC provides a unique solution that is invariant to factor rotation [13]. Hence, the factors are plausibly a valid description of the data with greater reason to believe that they have more explanatory meaning than a "nice" rotated two-way solution.

Given a value $r > 0$ (loosely corresponding to the number of distinct topics or conversations in our data), the tensor decomposition algorithms find matrices $A \in \mathbb{R}^{m \times r}$, $B \in \mathbb{R}^{n \times r}$, and $C \in \mathbb{R}^{p \times r}$, to yield Equation (4.1). Each triad $\{A_j, B_j, C_j\}$, for $j = 1, \ldots, r$, defines scores for a set of terms, authors, and months 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 C indicate the activity of each conversation topic over time.

4.1 PARAFAC-ALS. A common approach to solving Equation (4.2) is an alternating least squares (ALS) algorithm [13, 9, 26], 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, and C, we update these quantities in an alternating fashion using the method of normal equations. The minimization problem involving A in Equation (4.2) can be rewritten in *matrix* form as a least squares problem [9]:

$$\min_{A} \left\| X^{(m \times np)} - AZ \right\|^2 , \tag{4.3}$$

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

The least squares solution for Equation (4.3) involves the pseudo-inverse of Z:

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

Conveniently, the pseudo-inverse of Z may be computed in a special way that avoids computing $Z^T Z$ with an explict Z [24], so the solution to Equation (4.3) is given by:

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

Furthermore, the product $X^{(m \times np)}(C \odot B)$ may be computed efficiently if $\mathcal{X}$ is sparse [14] by not forming the Khatri-Rao product $C \odot B$. Thus, computing A essentially reduces to several matrix inner products, sparse tensor-matrix multiplication of B and C into $\mathcal{X}$, and inverting an $R \times R$ matrix.

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

4.2 Nonnegative Tensor Factorization. We also considered a PARAFAC model with non-negativity constraints on the factor matrices. Because we are dealing with non-negative data in $\mathcal{X}$, it often helps to examine decompositions that retain the non-negative characteristics of the original data. Modifications to the ALS algorithm are needed, and we use the multiplicative update introduced in [16] and adapted for tensor decompositions by Mørup [18, 19]. We also incorporate the addition of ϵ for stability as was done in [6]. Overall, the approach is similar to PARAFAC-ALS except that the factor matrices are updated differently.

First, we note that residual norm of the various formulations of the PARAFAC model are equal:

$$\begin{aligned} &||X^{(m \times np)} - A(C \odot B)^T||_F = \\ &||X^{(n \times mp)} - B(C \odot A)^T||_F = \\ &||X^{(p \times mn)} - C(B \odot A)^T||_F. \end{aligned}$$

Each of these matrix systems is treated as a NMF problem and solved in an alternating fashion. That is, we solve for A using the multiplicative update rule holding B and C fixed, and so on:

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

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

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

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.

As was mentioned previously, $\mathcal{X}$ is sparse, which facilitates a simpler computation in the procedure above. Each matricized version of $\mathcal{X}$ (which has the same

nonzeros but reshaped) is a sparse matrix. 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 [14] for details.

5 Enron Subset

For a relevant application, we consider the email corpus of the Enron corporation that was made public during the federal investigation. The whole collection is available online [8] and contains 517,431 emails stored in the mail directories of 150 users. We use a smaller graph of the Enron email corpus prepared by Priebe et al. [20] 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.

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 of Enron at the time of these emails, it is difficult to ascertain the validity of our results and assess the performance of the DEDICOM model. Other researchers using the Enron corpus have had this same problem, and information on the participants has been collected slowly and made available.

The Priebe data set [20] 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 [22]. It provides most employees' position and business unit. To facilitate a better analysis of the DEDICOM results, we collected extra information on the participants from the email messages themselves and found some relevant information posted on the FERC website [10]. To help assess our results, we searched for corroborating information of the preexisting data or for new identification information, such as title, business unit, or manager. Table 1 lists eleven of the most notable authors (and their titles) whose emails were tracked in this study.

Of the 197 authors whose emails were tracked (in the year 2001), there were a few cases of aliasing. That is, different email accounts of the form `employee_idenron.com` were used by the same employee. A few sample aliases from the eleven notable authors in Table 1 are: Vince Kaminski (`j.kaminski, j..kaminski, vince.kaminski`) and David Delaney (`david.delainey, w..delainey`).

5.1 Term Weighting.

Our data corresponds to a sparse adjacency array $\mathcal{X}$ of size $69157 \times 197 \times 12$ with 1,042,202 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 [11]. 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 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.

Unique to previous parsings of Enron subsets by GTP (see [5, 21, 4]), a much larger *stoplist* of unimportant words was used to filter out the content-rich 69,157 terms for the $\mathcal{X}$ array. This stoplist of 47,154 words was human-generated by careful screening of the master (GTP-generated) dictionary for words with no specific reference to an Enron-related person or activity.

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

$$x_{ijk} = l_{ijk} g_i a_j \tag{5.4}$$

where l_{ijk} is the local weight for term i written by author j in month k, g_i is the global weight for term i, and a_j is an author normalization factor.

Let f_{ijk} be the number of times term i is written by author j in month k, and define $h_{ij} = \frac{\sum_k f_{ijk}}{\sum_{jk} f_{ijk}}$. The specific components of each nonzero are listed below:

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

6 Observations and Results

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

6.1 PARAFAC.

We computed a 25-component ($r = 25$) decomposition of the term-author-month array $\mathcal{X}$ using PARAFAC. One ALS iteration of PARAFAC took about 22.5 seconds, requiring an average of 27 iterations to satisfy a tolerance of 10^{-4} in the change of fit. We chose the smallest minimizer from among 10

Table 1: Eleven of the 197 email authors represented in the term-author-month 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

runs starting from random initializations. The relative norm of the difference was 0.8904.

6.2 Non-negative Tensor Decomposition. We computed a 25-component ($r = 25$) non-negative decomposition of the term-author-month array $\mathcal{X}$. One iteration took about 22 seconds, and most runs required less than 50 iterations to satisfy a tolerance of 10^{-4} in the relative change of fit. We chose the smallest minimizer from among 10 runs from random starting points, and the relative norm of the difference was 0.8931.

6.3 Analysis of Results. PARAFAC is able to identify and track discussions over time in each triad $\{A_j, B_j, C_j\}$, for $j = 1, \ldots, r$. A discussion 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. Figures 2 and 3 present a histogram (or Gantt chart) of the monthly activity for each discussion identified by the classical and non-negative PARAFAC models, respectively.

Qualitatively, the results of the non-negative decomposition are the same as the standard three-way PARAFAC results. The difference between the two models was in the ability to interpret the results. In the 25 discussion groups depicted in Figure 2 only six of the groups have any discernible meaning based on our knowledge of the events surrounding Enron [17]. In comparison, the non-negative PARAFAC model revealed eight group discussions that could be interpreted.

The topics generated by the non-negative PARAFAC model do reflect the events of the year 2001, a tumultuous one for the global energy corporation to say the least. In the first quarter of the year, the company 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 showed up in emails during this time frame. Another ongoing and ever-present topic was Enron's attempted development of the Dabhol Power Company (DPC) in the Indian State of Maharashtra. The company's efforts in India had been ongoing for several years, and the emails of the early half of 2001 reflect some of the day-to-day dealings with the less-than-profitable situation in India.

By October of 2001, Enron was in serious financial trouble, and when a merger with the Dynegy energy company fell through, Enron was forced to file for Chapter 11 bankruptcy. Many of the emails of this time frame (more specifically in October and November) were newsfeeds from various news organizations that were being passed around Enron. Because it was learned at this time that Chief Financial Officer Andy Fastow was heavily involved with the deceptive accounting practices (by setting up sham companies to boost Enron's bottom line), it is not surprising a thread on this topic (*Fastow companies*) emerged. Predictably, the *College Football* topic emerges in late fall as well. One of the surprise topics uncovered was the *Education* topic, which reflects the interests and responsibilities of Vince Kaminski, head of research. Kaminski taught a class at nearby Rice University in Houston in the Spring of 2001, and was the focal point of emails about internships, class assignments and resume evaluation.

The fact that only eight of the 25 topics had any discernable meaning reflects the nature of topic

detection. A significant amount of *noise* or undefined content may permeate the term-author-month arrays. Sometimes, as shown by the thickness of the gray bar in Figure 3, there are indicators of a possible thread of some kind (not necessarily directly related to Enron), but further inspection of those emails reveals no identifiable topic of discussion.

7 Future Work

As demonstrated by this study, non-negative 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 decompositions is an important advancement in email surveillance research. Previously reported clusters of Enron emails using non-negative matrix factorization (i.e., two-way decompositions) [4, 5, 21] were unable to extract discussions such as the *Education* thread mentioned in Section 6.2 or sequence the discussion of the company's downfall by source (newfeeds versus employee-generated). The optimal segmentation of *time* as the third dimension for email clustering may be problematic. Grouping emails by month may not be sufficient for some applications and so more research in the cost-benefit tradeoffs of finer time segmentation (e.g., grouping by weeks, days, or even minutes) is needed. Determining the optimal tensor rank r for models such as PARAFAC is certainly another important research topic. Term weighting in three-way arrays is also an area that greatly influences the quality of results but is not yet well understood.

References

[1] E. Acar, S. A. Çamtepe, M. S. Krishnamoorthy, and B. Yener. Modeling and multiway analysis of chatroom tensors. In *ISI 2005: IEEE International Conference on Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 256–268. Springer Verlag, 2005.

[2] B. W. Bader and T. G. Kolda. Efficient MATLAB computations with sparse and factored tensors. Technical Report SAND2006-7592, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California, Dec. 2006.

[3] B. W. Bader and T. G. Kolda. Matlab tensor toolbox, version 2.1. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`, December 2006.

[4] M. Berry and M. Browne. Email Surveillance Using Nonnegative Matrix Factorization. *Computational & Mathematical Organization Theory*, 11:249–264, 2005.

[5] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons. Algorithms and Applications for Approximate Nonnegative Matrix Factorization. *Computational Statistics & Data Analysis*, 2007. to appear.

[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] 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.

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

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

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

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

[12] 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.

[13] 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`.

[14] 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.

[15] 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.

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

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

[18] M. Mørup. Decomposing event related eeg using parallel factor (parafac). Presentation, August 29 2005. Workshop on Tensor Decompositions and Applications, CIRM, Luminy, Marseille, France.

[19] M. M?rup, L. K. Hansen, and S. M. Arnfred. Sparse Higher Order Non-negative Matrix Factorization. *Neural Computation*, 2006. submitted.

[20] 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`.

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

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

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

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

[25] 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.

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

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

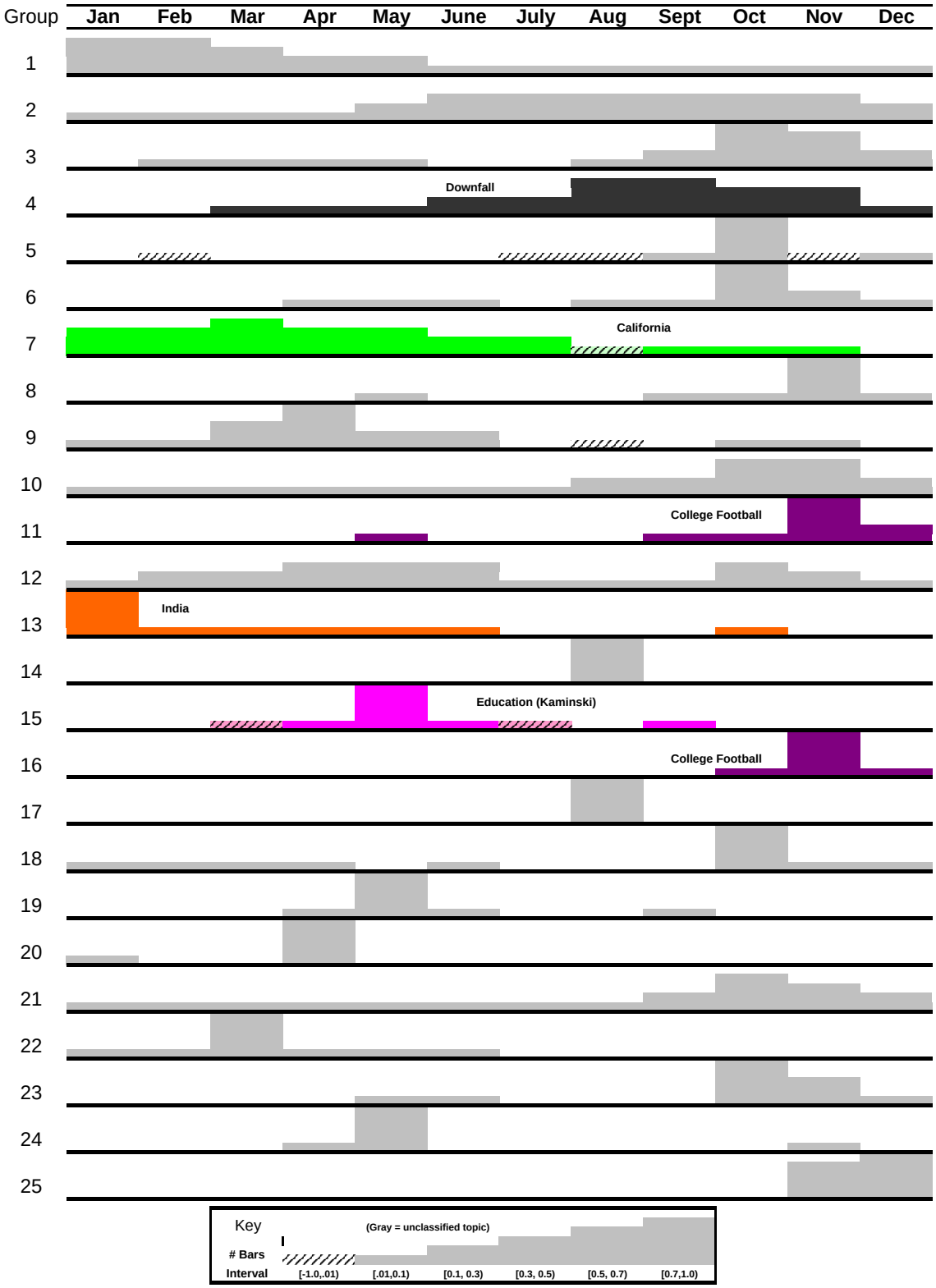

Figure 2: Six distinguishable discussions among the twenty-five extracted by classical PARAFAC. Diagonal shading of cells is used to indicate negative components in the tensor groups.

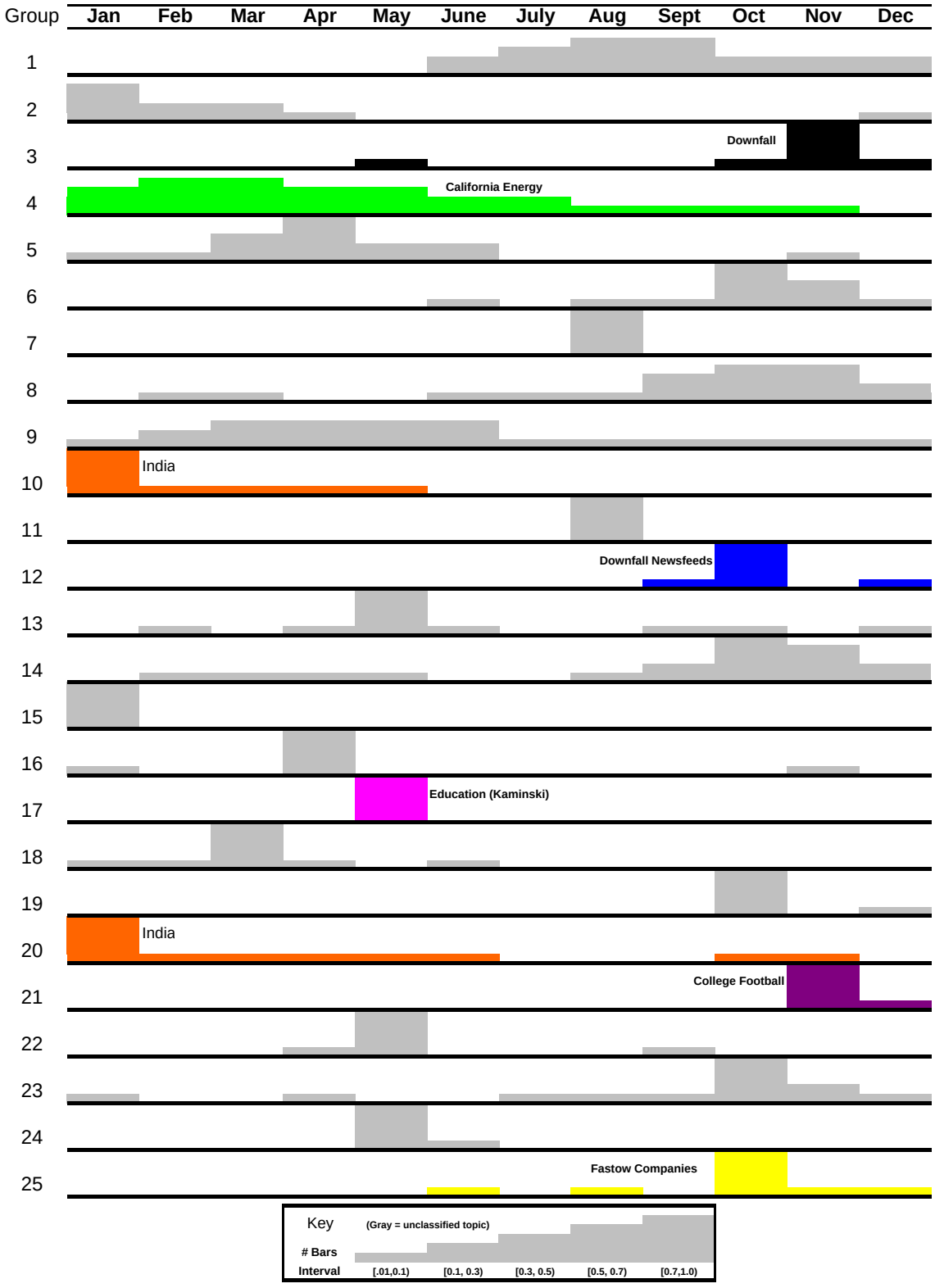

Figure 3: Eight distinguishable discussions among the twenty-five extracted by non-negative PARAFAC.