
Computational & Mathematical Organization Theory, 11, 183–199, 2005

Structure in the Enron Email Dataset

P.S. KEILA
D.B. SKILLICORN
School of Computing, Queen's University, Kingston, Canada, K7L 3N6
email: keila@cs.queensu.ca
email: skill@cs.queensu.ca

Abstract

We investigate the structures present in the Enron email dataset using singular value decomposition and semidiscrete decomposition. Using word frequency profiles, we show that messages fall into two distinct groups, whose extrema are characterized by short messages and rare words versus long messages and common words. It is surprising that length of message and word use pattern should be related in this way. We also investigate relationships among individuals based on their patterns of word use in email. We show that word use is correlated to function within the organization, as expected. Lastly, we show that relative changes to individuals' word usage over time can be used to identify key players in major company events.

Keywords: matrix decompositions, singular value decomposition, word frequency, organizational structure, organizational role, data mining

1. Introduction

Many countries intercept communication and analyze messages as an intelligence technique. The largest such system is Echelon (European Parliament Temporary Committee on the ECHELON Interception System, 2001), run jointly by the U.S., Canada, U.K, Australia, and New Zealand. The standard publicly-acknowledged analysis of intercepted data is to search messages for keywords, discard those messages that do not contain keywords, and pass those that do to analysts for further processing. Increasingly organizations also examine email to protect themselves against corporate malfeasance. An interesting question is what else can be learned from such messages; for example, can connections between otherwise innocuous messages reveal links between their senders and/or receivers (Skillicorn, 2005).

The Enron email dataset provides real-world data that is arguably of the same kind as data from Echelon intercepts—a set of messages about a wide range of topics, from a large group of people who do not form a closed set. Further, individuals at Enron were involved in several apparently criminal activities. Hence, like Echelon data, there are probably patterns of unusual communication within the dataset. Understanding the characteristics and structure of both normal and abnormal (collusive) emails therefore provides information about how such data might be better analyzed in an intelligence setting. Enron, like most

Presented at the Workshop on Link Analysis, Counterterrorism and Security at the SIAM International Conference on Data Mining, 2005.

large corporations, also contains employees fulfilling many roles, and communicating with each other within a hierarchy and with outside individuals, both corporate and government.

Linguistically, email has been considered to occupy a middle ground between written material, which is typically well-organized, and uses more formal grammatical style and word choices; and speech, which is produced in real-time and characterized by sentence fragments and informal word choices. Although the potential for editing email exists, anecdotal evidence suggests that this rarely happens; on the other hand, email does not usually contain the spoken artifacts of pausing (ums etc.).

We examine the structure of the Enron email dataset, looking for what it can tell us about how email is constructed and used, and also for what it can tell us about individuals, and how they use email to communicate. The contribution of this paper is threefold:

- We discover the basic characteristics of emails as human utterances, something that has not been possible until now because of the lack of a significant corpus;
- We show that individuals generate emails whose word use patterns reflect the individual's roles and relationships in organizations;
- We show that changes within an organization are reflected in changing word use patterns in email.

2. Related Work

Previous attention has been paid to email with two main goals: spam detection, and email topic classification. Spam detection tends to rely on local properties of email: the use of particular words, and more generally the occurrence of unlikely combinations of words. This has been increasingly unsuccessful, as spam email has increasingly used symbol substitution (readable to humans) which makes most of its content seem not to be words at all.

Email topic classification attempts to assist users by automatically classifying their email into different folders by topic. Some examples are O'Brien and Vogel (2004), Cohen (1996), Simon and Xenos (2004) and Lloyd and Spruill (2001). This work has been moderately successful when the topics are known in advance, but perform much less adequately in an unsupervised setting (but see the other papers in this Issue).

An attempt to find connections between people based on patterns in their email can be found in McArthur and Bruza (2003). A network approach to linking an individual's communication to their position within an organization can be found in Diesner and Carley (2005). Unlike our approach, in which the word profile of an email is analyzed, this approach analyzes communication patterns among employees.

3. Matrix Decompositions

We will use two matrix decompositions, *Singular Value Decomposition* (SVD) (Golub and van Loan, 1996), and *SemiDiscrete Decomposition* (SDD) (Kolda and O'Leary, 1999, 1998). Both decompose a matrix, A, with n rows and m columns into the form

$$A = CWF$$

where C is $n \times k$, W is a $k \times k$ diagonal matrix whose entries indicate the importance of each dimension, and F is $k \times m$.

There are several useful ways to interpret such a decomposition. The *factor* interpretation regards the k rows of F as representing underlying or latent factors (and hence better explanations of the data) while the rows of C describe how to mix these factors together to get the observed values in A. The *geometric* interpretation regards the k rows of F as representing axes in some transformed space, and the rows of C as coordinates in this (k-dimensional) space. The *layer* interpretation relies on the fact that A is the sum of k outer product matrices, A_i, where each A_i is the product of the ith column of C and the ith row of F (and the ith diagonal element of W). All of these interpretations can be helpful in interpreting a dataset.

Singular value decomposition is usually interpreted using the factor model (in the social sciences) and the geometric model (in the sciences). An SVD for the matrix A is

$$A = USV'$$

where U and V are orthogonal, the diagonal of S is non-increasing, and $k \leq m$. The usefulness of SVD comes primarily from the fact that the columns of V are orthogonal and hence represent *independent* factors, or axes. The first k columns of U can be interpreted as the coordinates of a point corresponding to each row of A in a k-dimensional space, and that this is the most faithful representation of the relationships in the original data in this number of dimensions.

The correlation between two objects is proportional to the dot product between their positions regarded as vectors from the origin. Two objects that are highly correlated have a dot product (the cosine of the angle between the two vectors) that is large and positive. Two objects that are highly negatively correlated have a dot product that is large and negative. Two objects that are uncorrelated have dot product close to zero. This property is useful because there are two ways for a dot product to be close to zero. The obvious way is for the vectors concerned to be orthogonal. However, when m is less than n (as it typically is) there are many fewer directions in which vectors can point orthogonally than there are vectors. Hence if most vectors are uncorrelated, they must still have small dot products but cannot all be orthogonal. The only alternative is that their values must be small. Hence vectors that are largely uncorrelated must have small magnitudes, and the corresponding objects are placed close to the origin in the transformed space. Hence, in the transformed space from an SVD, the points corresponding to objects that are 'uninteresting' (they are correlated either with nothing or with everything) are found close to the origin, while points corresponding to interesting objects are located far from the origin (potentially in different direction indicating different clusters of such objects).

The SemiDiscrete Decomposition (SDD) of a matrix A is

$$A = XDY$$

where the entries of X and Y come from the set $\{-1, 0, +1\}$, D is a diagonal matrix, and k can have any value, not necessarily less than m.

The natural interpretation of SDD is a layer one (McConnell and Skillicorn, 2002). Each A_i corresponds to a column of X and a row of Y, weighted by an entry from D. The product of x_i and y_i is a stencil representing a 'bump' (where the product has a $+1$) and corresponding 'ditch' (where the product has a -1). The corresponding value of D gives the height of the bump and ditch at each level. Hence an SDD expresses a matrix as the sum of bumps, with the most significant bumps appearing first. Because the choice of the sequence of bumps depends on both their area (how many locations in the matrix they cover) and their height, altering the scale of A will change the resulting SDD. In particular, taking the signed square of each value in the matrix will give greater emphasis to the heights of bumps and hence select outlying regions of the dataset earlier. Conversely, taking the signed square root of each value in the matrix will tend to find large homogeneous regions earlier.

SDD generates a ternary, unsupervised hierarchical classification of the samples, based on the values in each successive column of the X matrix. Consider the first column of X. Those samples for which this column has the value $+1$ can be grouped; those samples for which this column has the value -1 are, in a sense, similar but opposite; and those samples for which this column has the value 0 are unclassified at this level. This can be repeated for columns 2, 3, and so on, to produce a classification tree. Neither SVD nor SDD exploit the order of rows and columns in the data matrix, so they do not start with any advantage over more conventional data-mining techniques.

4. Structure from Word Usage

Most emails contain few words from a possibly very large vocabulary, so a document-word (email-word) matrix is extremely sparse. Although SVD could be performed on such matrices using sparse matrix techniques such as Lanczos methods, we choose instead to analyze matrices whose rows correspond to emails and whose columns correspond to word frequency rank. Each word is assigned its rank in the overall email dataset so, for example, the word "time" might be the second most frequent word. Each row of the matrix then contains, in its first column, the rank assigned to the most frequent word *in that email*; in its second column, the rank assigned to the second most frequent word in that email, and so on. Two emails are similar in this representation if they have similar word usage profiles *in descending order of frequency*; in other words, the similarity metric is more discriminating than one based only on a bag-of-words similarity metric.

4.1. Basic Structure

An SVD analysis of the entire email dataset is shown in figure 1, based on 494,833 messages using 160,203 distinct words (no stemming has been applied).

The most obvious and striking feature of this plot is that it results in a 'butterfly' shape, that is the emails separate into two clusters that grow increasingly different with distance from the origin. This separation is quite surprising; as far as we are aware previous analysis of email datasets has revealed separation by topic, but not such a strong structural separation.

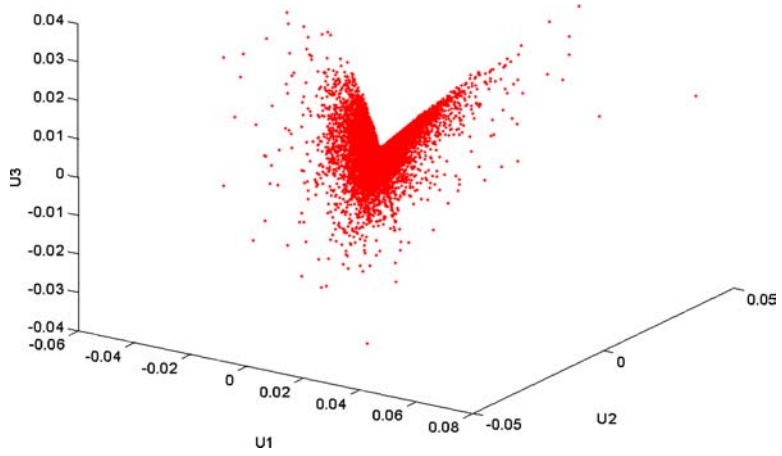

Figure 1. SVD plot of entire email set of 494,833 messages. Note the strong bifurcation.

This structure remains more or less fixed as the set of nouns is reduced, indicating that it is not an artifact of particular choice of nouns under consideration.

To explore the structure of the dataset more deeply, we reduce the number of words under consideration by removing those we believed made the least contribution to interesting structure. We use the BNC corpus (BNC, 2004), which is a frequency-ranked list of words in both spoken and written English, to assist. We first remove words that appear in the Enron dataset but not in the BNC corpus. This removes almost all of the strings that are not real words (artifacts of email processing and also of postprocessing of the dataset); and also almost all of the proper names and acronyms. We also remove words that were very frequent (appeared more than 1000 times in the dataset) and very infrequent (appeared fewer than 20 times in the dataset). Reducing the set of words removes some emails entirely. Figure 2 shows the SVD plot for this reduced dataset. As expected, the 'less interesting' emails are the ones that disappear, and a secondary structure begins to appear. The two 'wings' reduce to borders, and there are marked extensions that extend into the page on the left wing and out of the page on the right—in other words, the overall shape becomes a spiral. We reduced the word set further by retaining only words whose frequency of use in the email dataset is greater than their frequency of use in English (as recorded in the BNC corpus). This restricts attention to the 7424 words that Enron people use to communicate amongst themselves more than the general population. We call this *Enronspeak*, the normal patterns of utterance within the organization. This further reduces the number of email messages, producing the SVD plot shown in figure 3. The spiral shape is now very pronounced.

The reason for the strong bifurcation of emails is not clear. In general, the left hand 'wing' consists of messages with few distinct nouns; the emails near the origin are messages with a moderate number of distinct nouns, and the right hand 'wing' consists of messages with many distinct nouns. Recall that distance from the origin is a surrogate for interestingness, at least with respect to correlation structure. This spiral shape shows that there are three

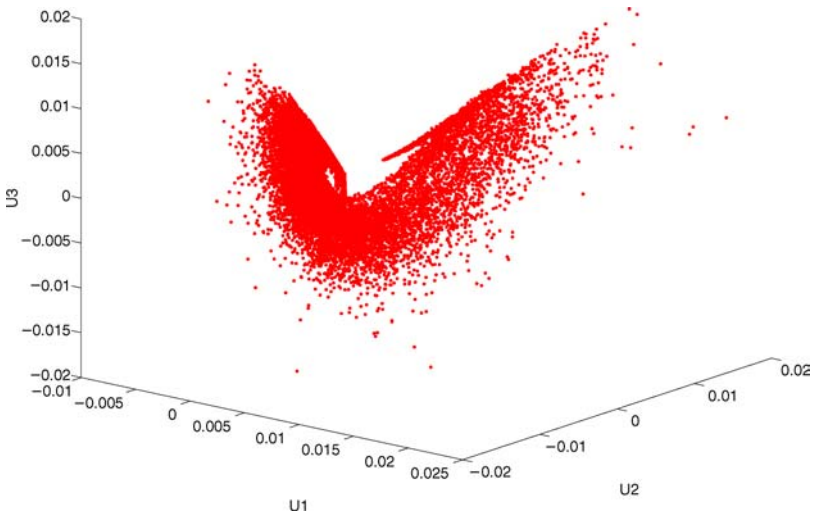

Figure 2. SVD plot of 350,248 emails, when the word set is reduced by (a) removing all words that appear in the Enron emails but not in the BNC corpus, and (b) removing all words with frequency greater than 1000 or less than 20.

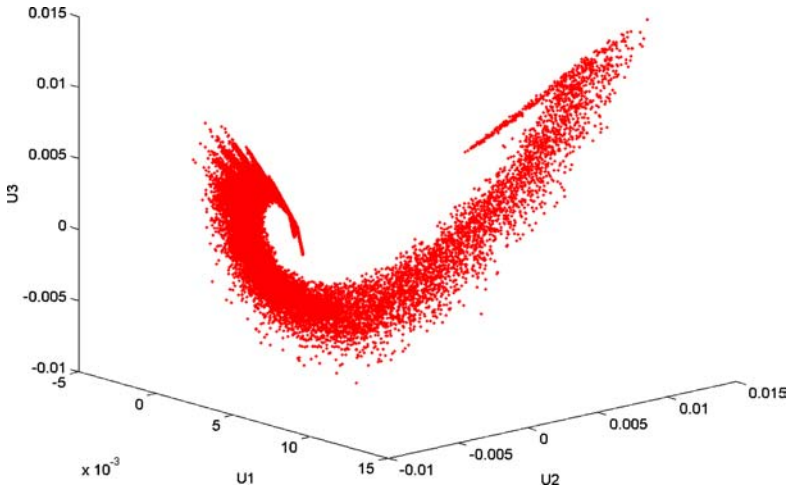

Figure 3. SVD plot of 289,695 emails, when the word set is reduced further by removing words whose frequency is greater in Enron email than in the BNC corpus (Enronspeak)—a set of 7424 words. The left hand end of the spiral goes into the page, while the right hand end comes out of the page.

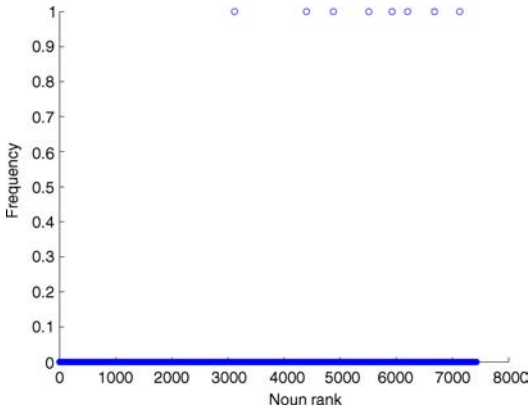

Figure 4. Noun frequency distribution for a typical extremal message on the left wing.

ways for an email to be uninteresting:

1. It contains very few distinct words (the sharp spike at the back of the left hand wing, which ends up quite near the origin);
2. It is of moderate size and uses words in ordinary ways (the region near the origin);
3. It is very long, and contains so many different nouns that it correlates with many of the other emails (the sharp spike at the front of the right hand wing which also ends up quite near the origin).

The remaining extremal emails are those that have the most interesting correlational structure. Words on the right wing use more nouns altogether, and so have greater opportunities for interesting correlation, whereas nouns on the left wing use few nouns and so have fewer opportunities. Hence the butterfly structure is quite asymmetric, with the right wing much larger and further from the origin than the left. Figure 4 shows the word frequency profile for a typical extremal message on the left wing. Figure 5 shows the word frequency profile for an extremal message on the right wing.

Extremal emails on the left wing can be characterized as: having been composed by a single author, short (in Enronspeak, although potentially containing many ordinary words), and tending to use each noun only once. Extremal emails on the right wing can be characterized as: coming from outside Enron, either digests with many different topics (sports updates, general news) or emails that reference many proper names, long (containing 100–350 Enronspeak nouns), and having more typical word frequency (Zipf-like) profiles.

Figures 6 and 7 show the way in which other properties correlate with position in the SVD plot. Figure 6 shows that message length correlates well with position along the spiral. Figure 7 shows that infrequent words are much more likely to occur at the left hand end, and frequent words to occur at the right hand end. Hence, message length is, at least to some extent, inversely correlated with rareness of words used.

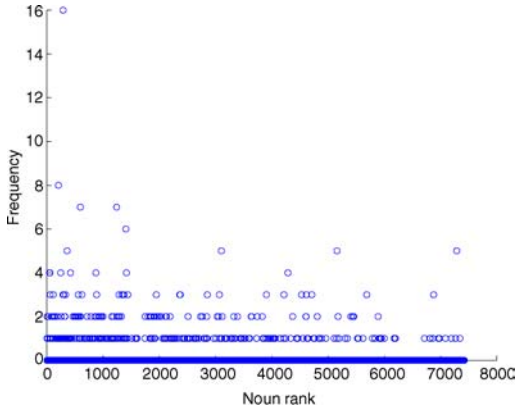

Figure 5. Noun frequency distribution for a typical extremal message on the right wing.

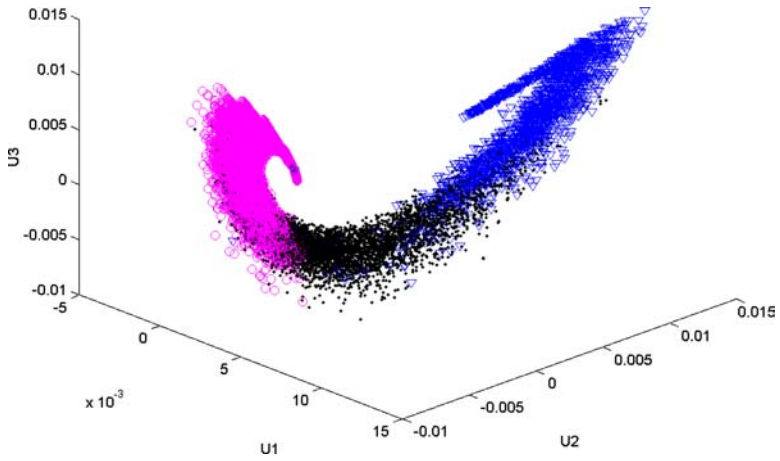

Figure 6. SVD plot labelled by message length (magenta: <20 nouns; black: <70 nouns.

Figure 8 shows the relationship between emails and their senders. The Corporate Policy Committee (CPC) consisted of 15 influential executives at Enron. These executives included the CEO, Chairman, Vice-Chairman, CFO, CAO, a number of heads from different Enron divisions, and an in-house lawyer. One member from this committee has since committed suicide, four have been charged and found guilty of various accounting and securities frauds, and three have been indicted. The figure shows the distribution of emails for those members of the committee whose emails remain in the dataset. Kean was responsible for circulated summaries of references to Enron in the media, and this explains his unusual email profile and relationships.

Figure 9 shows that the interestingness of an email (measured by distance from the origin) peaks for messages with about 220 total nouns, dropping to an asymptote for longer

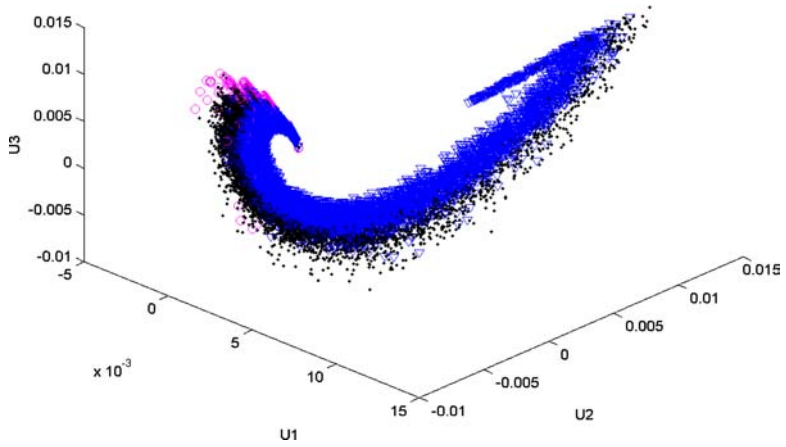

Figure 7. SVD plot labelled by average noun frequency rank, that is by rarity (magenta: >14,000; black: >8000).

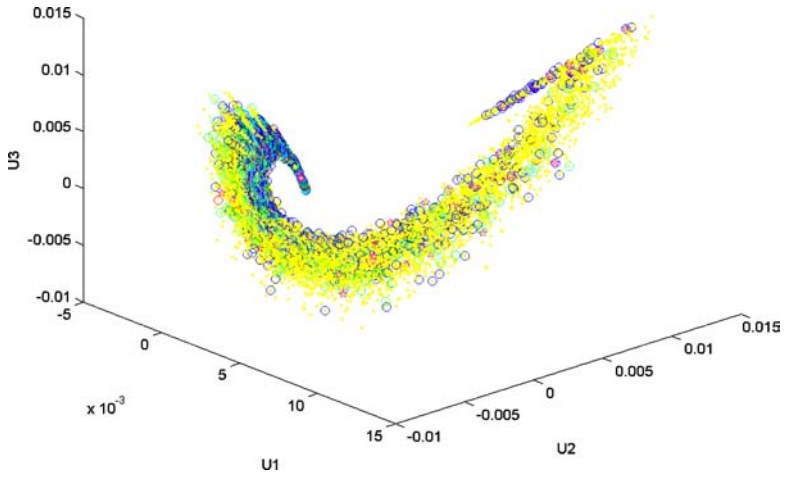

Figure 8. SVD plot labelled by email senders from the CPC. Magenta circle: Delaney; black circle: Derrick; red circle: Horton; blue circle: Kean; green circle: Lay; cyan circle: Skilling; magenta star: Whalley.

messages. This is surprising, since a message that contain this many Enronspeak words can contain several thousand words.

5. Authors and Emails

We now consider the matrix whose objects are individuals and whose columns are word frequencies, aggregated over all of their emails in the dataset. Hence each row captures a characteristic word use pattern for an individual. More interestingly, correlation in word use

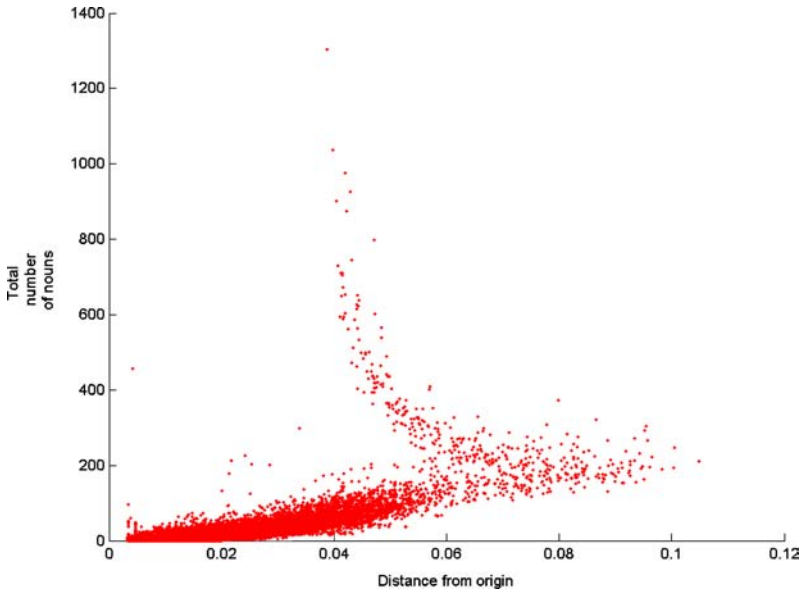

Figure 9. Plot of interest (i.e. distance from the origin in an SVD plot) versus total number of nouns in the message.

patterns determines position in an SVD plot, so that individuals with similar patterns will be placed close together. We might expect that individuals with similar job responsibilities and similar rank might use words in similar ways, both because of writing style, and because of similarity in typical subject matter. Further details of participants and their situation within Enron can be found in Shetty and Adibi (2004).

Figure 10 shows an SVD plot with a point for each individual in the dataset. The basic structure is a T-shape, with Vice-presidents along one arm towards the bottom right, and traders and other managers towards the bottom left. Core figures in the company tend to appear close to the center.

We can further restrict our attention to the individuals whose distance from the origin in the SVD plot is greater than the median distance. This leaves 30 individuals, including most of those with a significant role in the organization. Figure 11 shows the SVD plot of the 30 most interesting individuals.

Figure 12 shows the same plot, but with the points labelled by their SDD classification. Note how the (unsupervised) clustering properly distinguishes the functional properties of these individuals. Note also that the SDD labelling agrees, in general, with the positional similarities from SVD.

We can also add weights to certain rows and columns in the raw data. This has the effect of moving them away from the origin, and hence making them seem more important—but it also tends to cause correlated objects or attributes to follow them. We experiment with this by increasing the weight on words used by Lay and Skilling by a factor of 1.4. The result is shown in figure 13. The effect is to begin to partition the entire set of words into

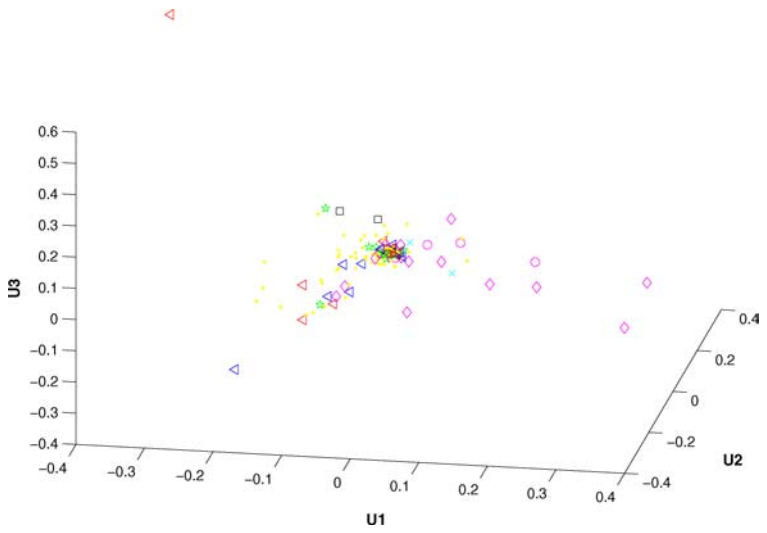

Figure 10. Relationships among 150 individuals based on similarity of email word use. Magenta: VP (diamond), President (circle); Black: CEO; Green: Director; Blue: Trader; Red: Manager; Cyan: Lawyer; Yellow: Unknown/Other. In this and subsequent figures, a set of 1713 words used by no more than 15 people are used.

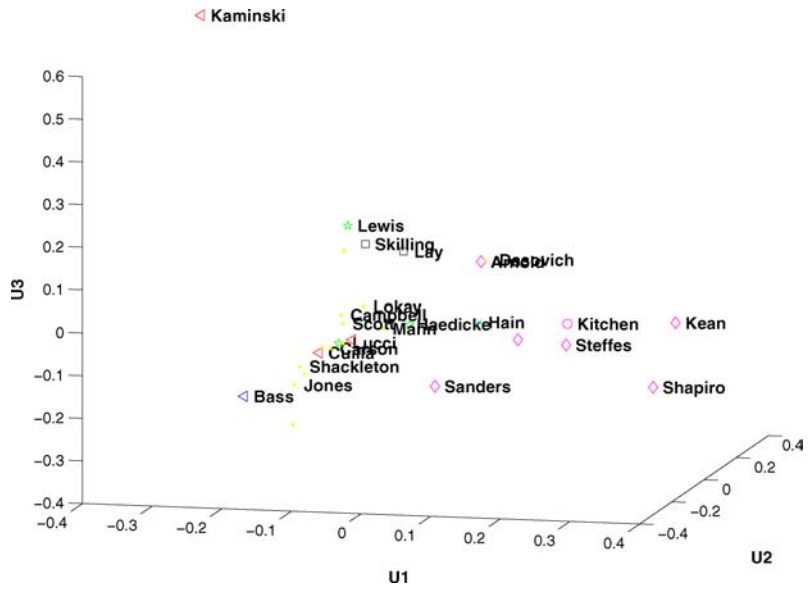

Figure 11. Relationships among 30 most interesting individuals. Labelling as in figure 10.

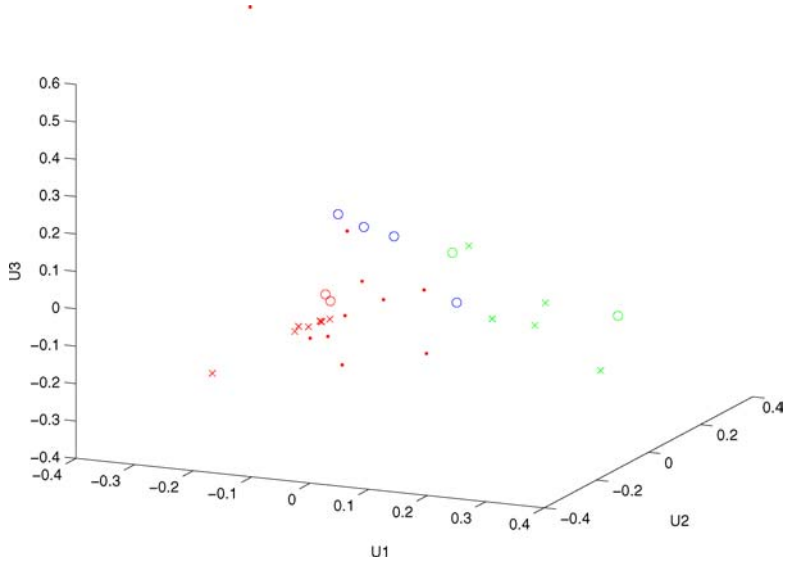

Figure 12. Relationships among 30 most interesting individuals, labelled (and so clustered) automatically by SDD classification.

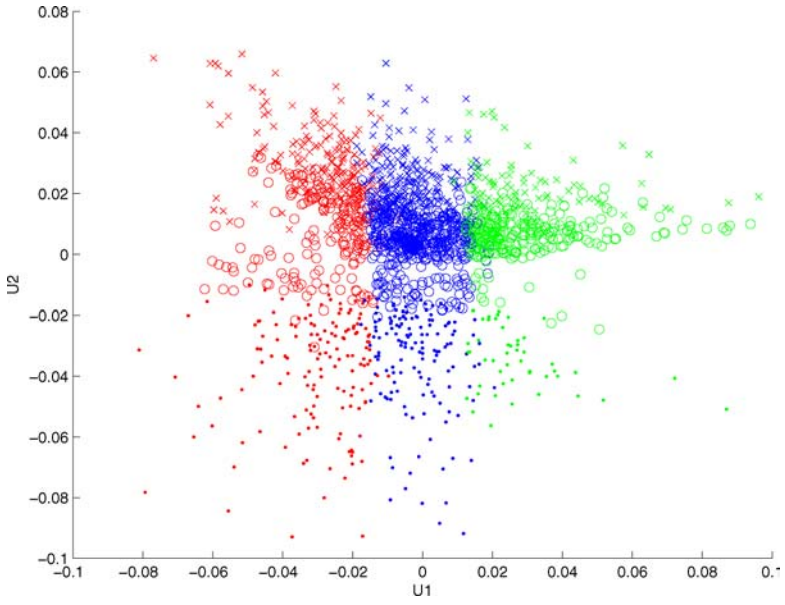

Figure 13. SDD labelled plot of words, weighting emails from Lay and Skilling by 1.4.

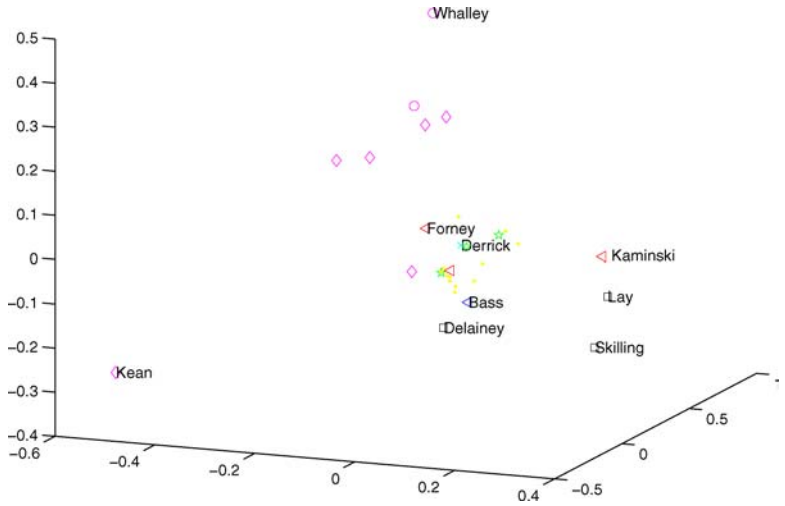

Figure 14. SVD plot of individuals when words used by Lay and Skilling are weighted by 1.4. Lay and Skilling move closer together, but Bass and Delainey; and Forney and Derrick move closer to each other, not to Lay and Skilling.

two clusters, one perhaps corresponding to the language of senior executives, and the other to the language of ordinary organization members.

Figure 14 plots the positions of individuals by word use, when the words used by Lay and Skilling are increased in weight by a factor of 1.4. The effect of this is to make their emails more 'interesting', moving them from the origin; but emails that are correlated with them will also tend to be 'pulled' further from the origin as well. Hence upweighting particular rows of the matrix (or indeed columns) allows us to examine relationships in more subtle ways. Several other pairs of individuals move into closer proximity compared to figure 11. This suggests that their word use patterns are influenced, in some indirect way, by those of Lay and Skilling; but in a way which makes the pairs similar to each other, not necessarily to Lay and Skilling. This may reflect discussions between pairs about topics affected by Lay and Skilling (for example, Lay and Skilling's hobbyhorses) or more subtle changes in word use reflecting off-line meetings in which they were all involved, or even attempts to sound more managerial.

6. Changes in Word Usage over Time

We divided the email set into four sections covering periods in the years 1999 to 2001. Our first subset is a collection of all the emails sent and received in the year 1999. Enron's first attempt at manipulating energy prices in California occurred in May of 1999. Although reprimanded for the attack, Enron traders engaged in substantially the same conduct the following spring under the schemes Death Star, Ricochet, Fat Boy, and Get Shorty. Hence there is reason to believe that traders at Enron were devising ways to game the newly

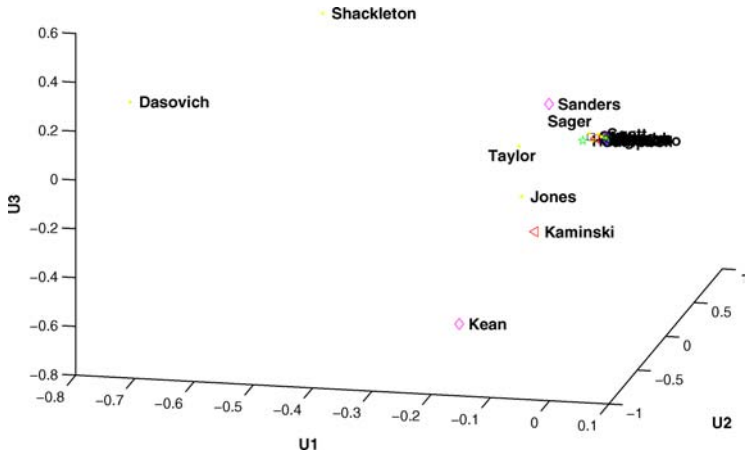

Figure 15. SVD plot of the top 30 most interesting individuals based on word usage—1999.

deregulated energy market in California in the latter half of 1999. The second subset is the collection of all of the emails sent and received in the year 2000. From May to August of 2000, the West Coast trading desk at Enron booked over $200 million in profits, which is roughly four times the profit the desk had made in all of 1999. It was also the first time since the end of World War II that power companies in California were forced to declare rolling blackouts. The third and fourth subsets are the emails sent and received in 2001, divided by the time when Skilling left the company and it began its public fall. For each subset of emails, we use the same set of 1713 nouns used above. We then create a noun-usage profile for each user over each of the 4 time periods. The resulting graphs can be seen in figures 15, 16, 17, and 18.

Overall, the majority of employees start as a tight cluster whose behavior is quite similar. As time goes on, this homogeneity is destroyed and we see individuals moving outwards (becoming more interesting) and moving out in different directions (becoming more diverse). This may be partly due to changing roles within the company, and partly to a changing emotional tenor, almost certainly involving reduced company loyalty, but also probably stress about the future. In each of the four figures, Kaminski, Kean, and Dasovich are far from the origin and hence interesting. Their word use patterns are significantly different from the rest of the company. Knowing the role Kean and Kaminski played in the company, this is not entirely surprising. Dasovich's appearance, however, is less expected. Dasovich was an Enron government relations executive.

Major changes in the company environment can thus be mapped to the individuals who played a significant role in producing these changes, or were affected in a substantial way by them. Hain and Haedicke, members of the Enron general counsel, make an appearance in the first half of 2001 (figure 17). From our limited knowledge of the Enron crisis, it makes sense that in figure 17 Hain and Haedicke move away from the origin; similarly it makes sense that Skilling and Kitchen move away from the origin in figure 18. In late 2000,

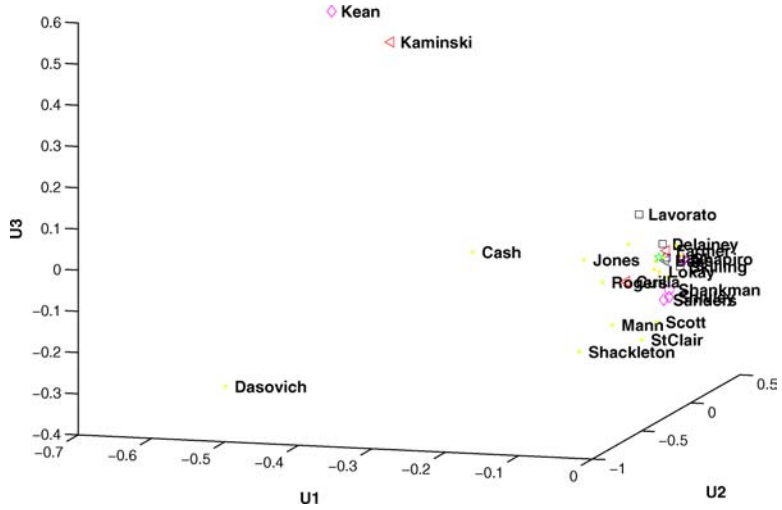

Figure 16. SVD plot of the top 30 most interesting individuals based on word usage—2000.

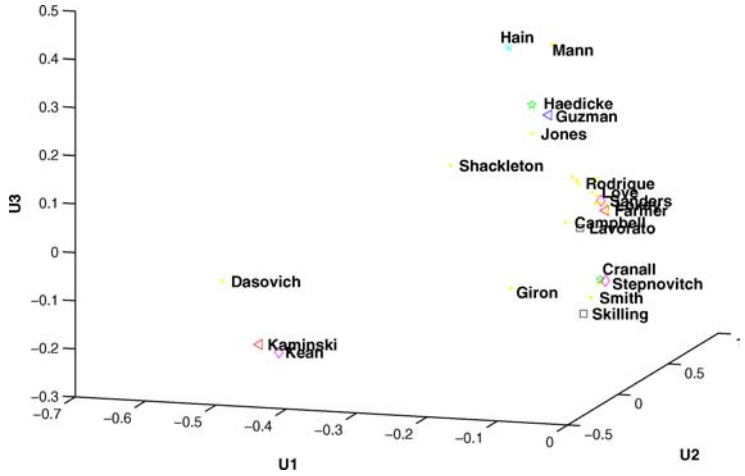

Figure 17. SVD plot of the top 30 most interesting individuals based on word usage—2001 before Skilling's departure.

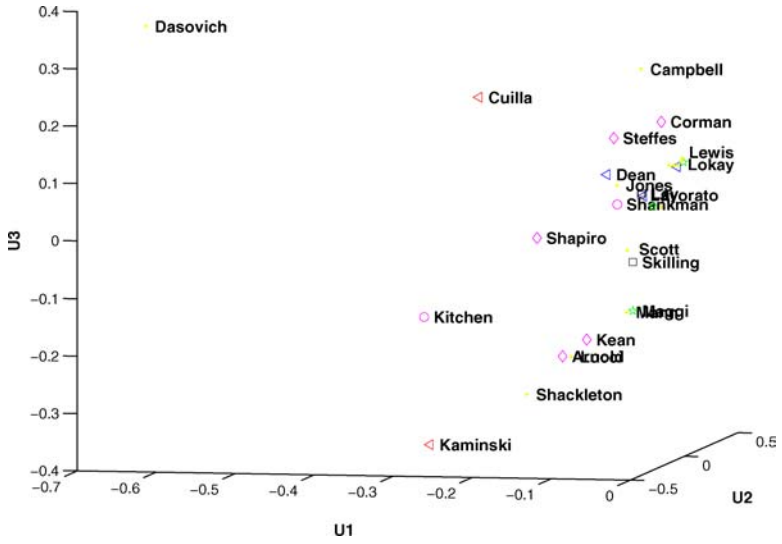

Figure 18. SVD plot of the top 30 most interesting individuals based on word usage—2001 after Skilling's departure.

Haedicke was made fully aware of the activities of the West Coast trading desk and began to think of ways to protect Enron's role in the affair. Jones, an unknown character in the Enron saga, maintains a close correlation to Haedicke. Skilling took his first extended vacation in June of 2001 and formally resigned that August. Knowing that the Enron's trading business came under scrutiny in the later half of 2001 it is not surprising to see that Kitchen, President of Enron Online, appears in figure 18 and, although similar to Kean, seems to play a more significant role.

7. Conclusion

Using matrix decompositions such as singular value decomposition and semidiscrete decomposition, we have explored the structure of a large real-world email corpus. The structure of email messages, using similarity based on word use frequency profiles shows a distinctive butterfly/spiral pattern which we have not been able to fully account for. There appears to be a strong differentiation between short messages using rare (in this context) words, and long messages using more typical words. The characteristic length of the emails with the most interesting correlative structure seems surprisingly long.

We also analyzed the relationships among individuals based on the word use frequency profiles of the emails they send. There is a clear effect of company role on word use patterns—individuals of similar status and role tend to communicate in similar ways. There are also some hints that emphasizing certain words tends to pull together individuals who are not obviously associated in the company environment, but there may be several explanations for this behavior.

It is also clear that word usage patterns change over time, reflecting both the effects of individuals' actions to change the organization, and also the effect of those changes on themselves and others. This provides a way to track organizational activity and identify the key players at each stage.

Acknowledgments

This research was funded by Natural Sciences and Engineering Research Council of Canada.

References

British National Corpus (BNC), (2004), www.natcorp.ox.ac.uk.

Cohen, W.W. (1996), "Learning to Classify English Text with ILP Methods," in L. De Raedt (Eds.), *Advances in Inductive Logic Programming*, IOS Press, pp. 124–143.

Diesner, J. and K. Carley (2005), "Exploration of Communication Networks from the Enron Email Corpus," in *Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining*, pp. 3–14.

European Parliament Temporary Committee on the ECHELON Interception System (2001), "Final Report on the Existence of a Global System for the Interception of Private and Commercial Communications," Echelon Interception System.

Golub, G.H. and C.F. van Loan (1996), *Matrix Computations*, 3rd edn. Johns Hopkins University Press.

Kolda, G. and D.P. O'Leary (1998), "A Semi-Discrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval," *ACM Transactions on Information Systems*, 16, 322–346.

Kolda, T.G. and D.P. O'Leary (1999), "Computation and Uses of the Semidiscrete Matrix Decomposition," *ACM Transactions on Information Processing*.

Lloyd, D. and N. Spruill (2001), "Security Screening and Knowledge Management in the department of defense," in *Federal Conference on Statistical Methodology*.

McArthur, R. and P. Bruza (2003), Discovery of Implicit and Explicit Connections Between People Using Email Utterance," in *Proceedings of the Eighth European Conference of Computer-supported Cooperative Work, Helsinki*, pp. 21–40.

McConnell, S. and D.B. Skillicorn (2002), "Semidiscrete Decomposition: A Bump Hunting Technique," in *Australasian Data Mining Workshop*, pp. 75–82.

O'Brien, C. and C. Vogel (2004), "Exploring the Subject of Email Filtering: Feature Selection in Statistical Filtering."

Shetty, J. and J. Adibi (2004), "The Enron Email Dataset Database Schema and Brief Statistical Report," Technical report, Information Sciences Institute.

Simon, A.F. and M. Xenos (2004), "Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis," *Political Analysis*, 12, 63–75.

Skillicorn, D.B. (2005), "Beyond Keyword Filtering for Message and Conversation Detection," in *IEEE International Conference on Intelligence and Security Informatics (ISI2005)*, Springer-Verlag Lecture Notes in Computer Science LNCS 3495, pp. 231–243.

P.S. Keila is a graduate student in the School of Computing at Queen's University. His research area is data mining in text.

D.B. Skillicorn is a professor in the School of Computing at Queen's University, where he heads the Smart Information Management Laboratory. His research area is data mining using matrix decompositions, particularly applied to complex datasets in areas such as biomedicine, geochemistry, counterterrorism and fraud.