Detecting Unusual and Deceptive Communication in Email

P.S. Keila and D.B. Skillicorn

June 2005
External Technical Report
ISSN-0836-0227-
2005-498

School of Computing
Queen's University
Kingston, Ontario, Canada K7L 3N6

Document prepared June 8, 2005

Abstract

Deception theory suggests that deceptive writing is characterized by reduced frequency of first-person pronouns and exclusive words, and elevated frequency of negative emotion words and action verbs. We apply this model of deception to the Enron email dataset, and then apply singular value decomposition to elicit the correlation structure between emails. This allows us to rank emails by how well they fit the profile of detection. Those emails that are highly ranked using this approach include deceptive emails; other emails that are ranked highly using these frequency counts also indicate organizational dysfunctions such as improper communication of information. Hence this approach can be used as a tool for both external investigation of an organization, and internal management and regulatory compliance.

Detecting Unusual and Deceptive Communication in Email

P.S. Keila and D.B. Skillicorn

Abstract: Deception theory suggests that deceptive writing is characterized by reduced frequency of first-person pronouns and exclusive words, and elevated frequency of negative emotion words and action verbs. We apply this model of deception to the Enron email dataset, and then apply singular value decomposition to elicit the correlation structure between emails. This allows us to rank emails by how well they fit the profile of detection. Those emails that are highly ranked using this approach include deceptive emails; other emails that are ranked highly using these frequency counts also indicate organizational dysfunctions such as improper communication of information. Hence this approach can be used as a tool for both external investigation of an organization, and internal management and regulatory compliance.

1 Introduction

In this paper we consider what can be learned about an organization from the collection of emails to and from its employees. In particular, we are interested in whether fraud, malfeasance, misuse of resources, and other unwanted, and possibly criminal behaviour, can be detected from such data.

Organizations need to be aware of what employees are doing both as a management issue (wasting resources, wasting time, unwittingly having a PC that acts as a zombie) and a criminal issue (committing fraud). Legislation such as the United States Sarbanes-Oxley law makes managers responsible for "adequate internal control structure and procedures for financial reporting" creating, among other things, the expectation that they can detect employees who are committing fraud against the company.

Email is an important vehicle for communication in most companies, both among employees, and between employees and the outside world. It is thus one possible source of data from which potential problems can be detected. This naturally raises concerns about privacy – most employees are not comfortable with the idea that their employer is reading their email. It is therefore important that techniques used to address this problem are as unintrusive as possible. This is also in the organization's interest, since it would be extremely expensive to deploy a technique that did, in fact, 'read' employees' email.

Models of deceptive practices, fraud, or collusion assume that word usage changes [8] to reflect factors such as: self-consciousness or guilt about the deception; and simplified content to make consistent repetition easier and to reduce the cognitive burden of generating a false 'story'. The awareness that some kind of surveillance may be in place may also generate an excessive blandness in messages as their senders try to ensure that the messages do not get flagged [9]; this blandness may itself become a signature. It is also likely that messages between coconspirators will have unusual content, since they communicate to discuss what are, in context, unusual actions.

We apply models based on deception theory to the Enron email dataset, a large now-public dataset of emails to, from, and among Enron staff in the three-year period before the collapse of the company. We show that standard models of deception do well at capturing emails that are deceptive. However, other emails flagged by the model are also likely to be of interest to management, if not to law enforcement. The contribution of this work is therefore to develop a technique for ranking emails along a scale that reflects their importance for discovering malfeasance inside an organization.

This approach could clearly be generalized to other problem domains: ranking messages intercepted by governments that might be relevant to counterterrorism or drug smuggling; spam detection; deceptive advertising; and deceptive web sites (for example, financial investment scams). Indeed this approach can be applied to any domain in which a linguistic model for the behaviour of concern can be constructed.

2 Modelling Deception

2.1 Patterns of Deception in Word Usage

The majority of information exchanged on a daily basis is done using rich media (e.g. face-to-face or voice). In a large meta-analysis of cues for deception [4], some of the communication properties and actions that were positively correlated with deception included: pressing lips together, discrepancies, raising the chin, word and phrase repetitions, negative statements and complaints, vocal tension, pupil dilation, and fidgeting. Properties and actions negatively correlated with deception included: talking time, inclusion of details, logical structure, illustrations, verbal immediacy, facial pleasantness, spontaneous corrections, and admitted lack of memory.

Unfortunately, electronic communication does not include many of the most useful non-verbal cues. As individuals increase their usage of electronic forms of communication, there has been research into detecting deception in these new forms of communication. In terms of production, email lies between speech, which is produced in real time; and formal written communications, which can be edited before transmission. Although emails can, in principle, be edited, there is considerable anecdotal evidence that this is seldom done. In terms of preferred media for deception, email scores very low, apparently because senders are aware, at some level, that email has a potentially long lifetime, allowing inconsistencies to be discovered. Media such as the telephone, or even instant messaging are preferred to email when deception is the goal [5].

Models of deception assume that deception leaves a linguistic footprint, both because language production is fundamentally a subconscious process, and because the cognitive demands of deception cause performance deficits in other areas. Research groups [1, 3, 7, 10] have studied ways of automating the process of textual deception detection in a supervised environment. Their approaches compare documents in which it is known that the author is being deceptive to ones in which it is known that the author is telling the truth. Differences in word usage between the two sets of documents suggest a word-use model that may signal deception. There has also been some work aimed at developing a classification, and so predictive, model of deception [11], although prediction performance is still relatively poor.

We use a subset of the words used by Pennebaker *et al.* derived from their Linguistic Inquiry and Word Count (LIWC) program [8]. Work done by various researchers [7, 12] suggests that individuals who are trying to deceive generally use fewer first-person pronouns and exclusive words, and more negative emotion words and action verbs.

Fewer first-person pronouns may indicate authors' attempts to "dissociate" themselves from their words. Fewer exclusive words indicate a less cognitively complex 'story' that is easier to create and to remember consistently. An increased frequency of action verbs may be an artifact of reducing the number of exclusive words, or the result of attempts to distract from the lack of subtlety by including plenty of action. The increased frequency of negative emotion words indicate some degree of self-respect dissonance about the fact of the deception.

2.2 The Enron Email Dataset

The Enron email dataset is the first large-scale collection of real world email released into the public domain. The Federal Energy Regulatory Commission (FERC) originally posted a collection of emails from ex-Enron employees on the Internet in May of 2002 as part of their legal actions against the company. Each email contains the email address of the sender and receiver(s), date, time, subject, body, and text. Attachments were not made available. The dataset used in our research comes from Cohen at Carnegie Mellon University. This version of the dataset contains 517,431 emails from the mail folders of 150 ex-Enron employees including many of the top executives, such as Kenneth Lay (ex-Chairman and CEO) and Jeffrey Skilling (ex-CEO). Though the vast majority of the communication is completely innocent (and boring), the emails of a number of top executives who are currently being prosecuted are in the dataset, Hence, it is reasonable to believe that evidence of deception exists within the dataset.

2.3 Matrix Decompositions

We use Singular Value Decomposition (SVD) as the primary analysis technique. The singular value decomposition of a matrix A is

$$A \; = \; USV'$$

where the dash indicates the transpose. If A is $n \times m$ and has rank r, then U is $n \times r$, S is an $r \times r$ diagonal matrix with decreasing entries $\sigma_1, \sigma_2, \ldots, \sigma_r$ (the singular values), and V is $r \times m$. In addition, both U and V are orthogonal, so that $UU' = I$ and $VV' = I$. In most practical datasets, $r = m$.

Under a geometric interpretation, we will regard the k rows of V as representing axes in some transformed space, and the rows of U as coordinates in this (k-dimensional) space. This k-dimensional representation 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 highly correlated objects will have a large and positive dot product. Two negatively correlated objects will have a large and negative dot product. Uncorrelated objects will have a dot product close to zero. This property of a SVD is useful because there are two ways for two objects to have a dot product close to zero. First, if the respective vectors are orthogonal, then the dot product by definition will be zero. However, when m is less than n (as in most cases) there are 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 for their norms to be small. Thus, vectors that are largely uncorrelated must have small magnitudes and the corresponding objects are placed close to the origin. Second, if an object is strongly correlated with most of the other objects, and most of the other objects are close to the origin, it will also be close to the origin. Highly correlated objects tell us little new information about the data. Points close to the origin (those that are correlated with nothing or everything) can thus considered 'uninteresting', while points corresponding to interesting objects are located far from the origin (potentially in different directions indicating different clusters of such objects).

3 Analysis of Emails

There are a number of processed versions of the Enron email dataset available in the public domain. The version we used in this work has 289,695 emails, and was created by our research group at Queen's. We construct two matrices from the email dataset.

The first matrix captures the full word-use pattern of each email. The BNC corpus [2] provides a frequency-ranked list of nouns in English. We construct a matrix in which each row corresponds to an email, and contains the (English) rank of each of the words in the message, arranged in decreasing order. "Time" is the most common noun in English, so every email that contains the word "time" will have a 1 in the first column. "Quantum" is the 3652nd most common word in English, so every email that contains this word will have 3652 somewhere in its row, but the position will depend on which *other* words it contains.

The second matrix has one row for each email but only four columns, each corresponding to a class of words associated with deception. These four columns count the frequency of:

- first-person pronouns (I, me, my, etc.);
- exclusive words (but, except, without, etc.);
- negative emotion words (hate, anger, greed, etc.);
- action verbs (go, carry, run, etc.);

Of the 289,695 emails in the dataset, only 265,836 have at least one occurrence of at least one cue.

4 Results for the email-rank matrix

A plot of the first three columns of the U matrix of an SVD of the email-rank matrix is shown in Figure 1, with one dot for each of 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. Deeper analysis, reported in [6], shows that the left 'wing' consists of short messages using relatively rare words; and the right 'wing' consists of long messages using relatively common words. Recall that distance from the origin is a surrogate for interesting correlation structure among messages. Hence the interesting messages are those that lie along the extremal edges of each 'wing'. This strong bifurcation is unexpected and we have not yet fully understood the implications of its presence. However, the same structure appears in artificial data generated with typical Zipf frequency distributions so it is not an artifact of this particular dataset.

5 Results for the email-deception cue matrix

A plot of the first three columns of the U matrix of an SVD of the email-deception cues matrix is shown in Figure 2.

The extremal messages are of three distinct kinds, shown by the three points of the broadly triangular structure. The top left corner of the plot corresponds to emails with a large number of exclusive words. Emails in this region tend to be emotionally-charged messages to coworkers, family and friends. The top right corner corresponds to emails with many first person pronouns.

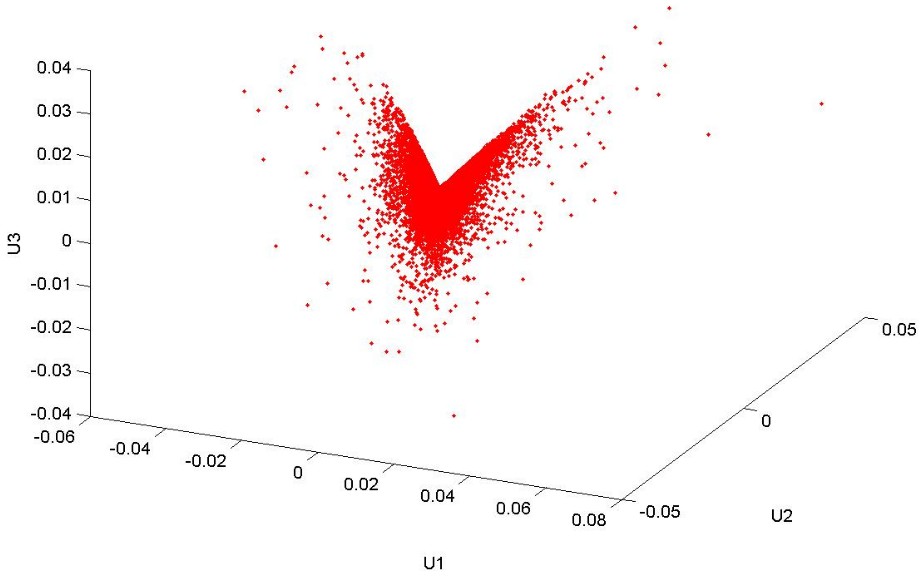

Figure 1: SVD plot of entire email set of 494,833 messages. Note the strong bifurcation.

Emails in this region tend to be about non-business recreational activities, for example a fantasy football league. Points towards the bottom portion of the graph correspond to emails with many action verbs. Given that we expect deceptive messages to be less cognitively complex and thus low in exclusive words and first person pronouns, while being high in action verbs, we expect deceptive messages to congregate towards the bottom half of the plot. There *are* messages that can be labeled deceptive in nature (contract negotiations, employees discussing confidential information with one another) in this area of the plot, but there are many more messages on a variety of other topics.

Figure 3 shows the same plot, with the points labelled according to who sent each email. This figure shows that emails from Enron employees are, on average, further from the origin than those originating outside the organization.

We now apply a deception model to this data. Recall that deceptive messages are expected to contain: reduced frequency of personal pronouns (dissociation from the content); reduced frequency of exclusive words (to reduce cognitive load); increased frequency of negative emotion words (sublimated guilt); and increased frequency of action verbs (distraction). Since SVD is a numerical technique, increased magnitude of attribute values increases the importance of those values. Since we want *low* values of first-person pronouns and exclusive words to be most significant, we must alter the raw frequency counts. For the exclusive word frequencies, we simply subtract the raw frequency from the greatest frequency seen in the data. For the first-person pronouns, greater care is needed. First, we observe that overall use of first-person pronouns in the dataset is rare, probably because they are considered inappropriate in certain kinds of business emails. As a result,

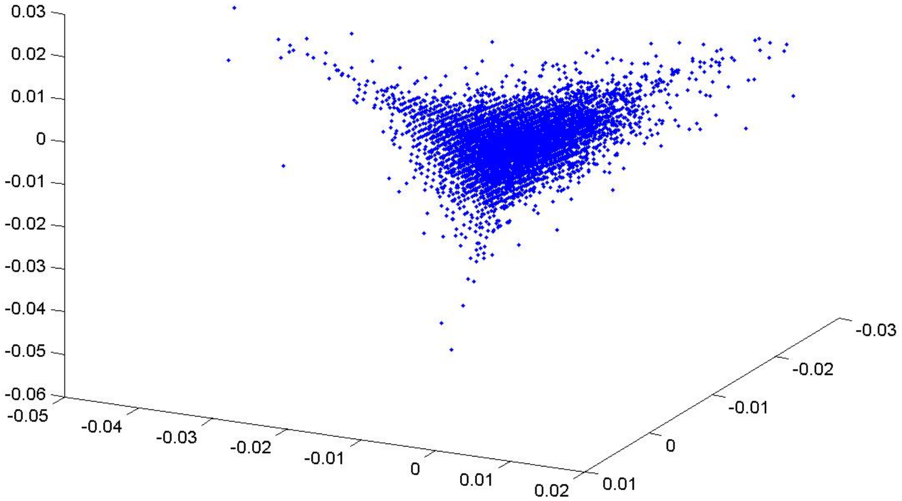

Figure 2: SVD plot of the email set of 265,836 emails using word frequencies in four categories related to deception as the attributes

the characteristic signature of reduced first-person pronouns in other settings seems likely to be present in emails that contain 5-15 occurrences of a first-person pronoun (some emails contain 50 or more occurrences). Hence we alter first-person pronoun data to select a frequency pyramid centred around frequency 10, and set frequencies outside the range 5-15 to zero.

Figure 4 shows an SVD plot of this altered dataset, in which we expect emails that match the deception signature to appear far from the origin. Points are labelled by the sender of the email using the same key as in Figure 3.

Note that the origin is in the left-hand top corner. Hence the most unusual emails are those towards the right, at the extreme edges of the two 'layer' clusters. These regions do include primarily emails that are deceptive (we do not include examples here because they are long and often profane). Points in the lower of these two clusters match the standard deception model most closely: emails have reduced frequencies of first-person pronouns and exclusive words and increased frequencies of negative emotion words and action verbs. Points in the higher of the two clusters are similar, but do not exhibit increased frequency of negative emotion words. Note that the extremal emails in this cluster originate from within Enron. Given that there was no stigma associated with many of Enron's activities within the company, we can speculate that these points are characteristic of 'Enron-style deception'. The cluster in the top left corresponds to messages with a strong action verb component, but little else – they are often to-do lists or lists of accomplishments. There is no straightforward way to validate the ranking, in the sense that deceptive messages might not be

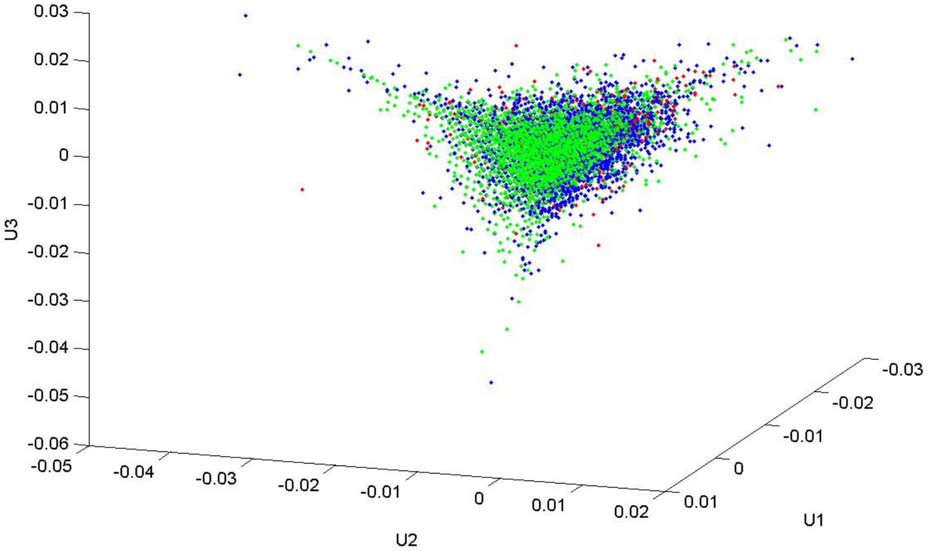

Figure 3: SVD plot of the email set of 265,836 emails labelled by sender (red: emails sent by executives; blue: emails set by other employees; green: emails sent by external parties)

highly ranked and this would be hard to discover. However, in a set of half a million emails, finding *any* email message worth reading suggests that the model is doing some useful selection.

Figure 5 shows the same SVD plot, but with each point labelled according to which attributes have high values using the following key:

- magenta – emails with high frequencies of negative emotion words and actions verbs;
- cyan – emails with low frequencies of first-person pronouns;
- yellow – emails with low frequencies of exclusive words.

Emails of a particular kind can now easily be selected from such a plot, based on their colour coding and their distance from the origin. The corresponding emails can then be retrieved and analyzed by appropriate individuals.

Nothing in the analysis so far has depended on the identities of senders or receivers, so the vast majority of emails undergo only the most limited scrutiny (extracting certain kinds of words and counting them). Only those emails that appear to be of significant interest undergo further analysis; and the boundary between the kind of widespread analysis we have discussed so far, and a more intrusive examination in which individuals are identified provides a place and a process for regulation. For example, individuals' privacy can be protected by requiring an explicit permission for identities to be revealed, and an unchangeable logging of the fact that this happened.

This approach can also be applied to the emails of a single employee to help select this individual's most unusual messages, for example when a particular individual is suspected of criminal

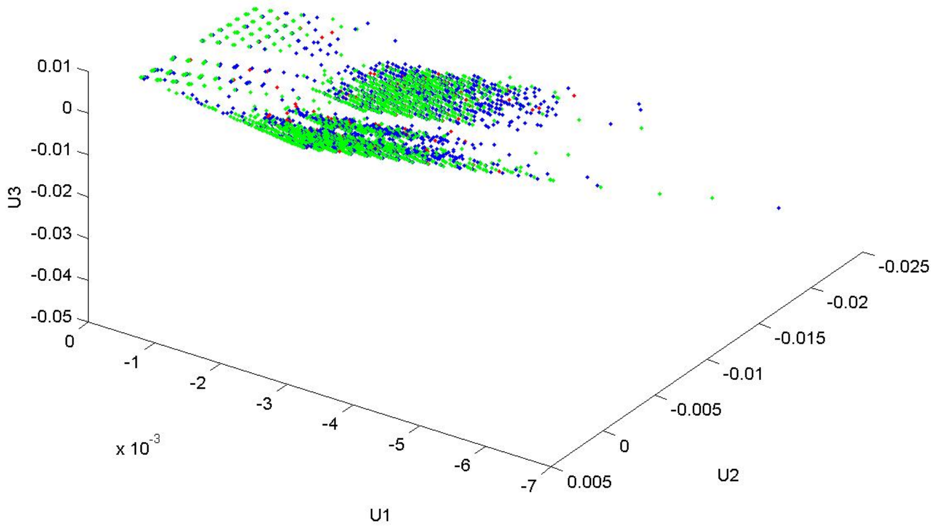

Figure 4: SVD plot of emails using a deception model (labelled by sender)

involvement. Even for a single individual, reading all of his or her incoming and outgoing email is expensive. The techniques described here can be used to select the putatively most interesting emails to be read first. Figure 6 shows the emails sent by Forney, the third Enron executive to be arrested. There are two obvious clusters in this plot, the left one corresponding to deceptive business emails, and the right one corresponding to emotionally-charged emails to family and friends. The most interesting message of each kind is indicated in the Figure. Even counting how many email messages fall beyond a given distance from the origin in a plot of this kind provides a simple test for potential problems with an individual.

6 Conclusions

Conventional models of deception do capture deceptive emails well in the Enron email dataset, although with some adaptation to account for the fact that these emails are (intended to be) written in a business context. Rather than learning a predictive model for deceptive emails, our approach ranks emails by how likely they are to be deceptive. Some appropriate fraction of the most likely emails can then be selected for further analysis depending on the context and the cost of doing so.

We also see that the attributes associated with the deception model capture emails that reflect a variety of potential problems with an organization, for example complaining, conveying information improperly, or spending organizational resources and employee time on non-work-related issues.

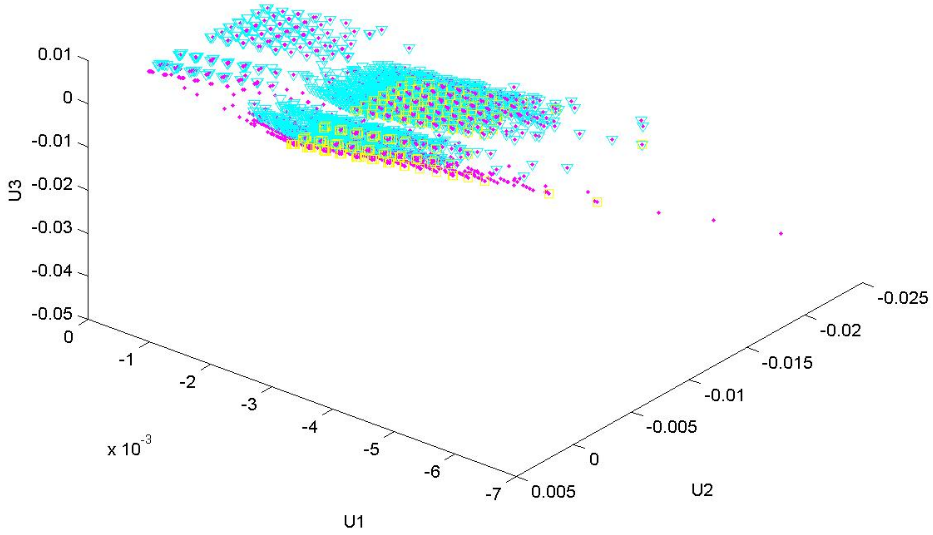

Figure 5: Emails labelled according to which deception markers are most significant

Analysis by such a signature can therefore be useful for detecting both potential organizational dysfunctions and criminal behaviour such as fraud. This approach can therefore be used by law enforcement in the criminal investigation of organizations, by organizations themselves (and their auditors) to ensure that they meet regulatory obligations; and by organizations practising due diligence in, for example, a takeover situation.

References

[1] D.P. Biros, J. Sakamoto, J.F. George, M. Adkins, J. Kruse, J.K. Burgoon, and J.F. Nunamaker Jr. A quasi-experiment to determine the impact of a computer based deception detection training system: The use of Agent 99 trainer in the us military. In *Proceedings of the 38th Hawaii Intl Conference on Systems Science*, volume 1, 2005.

[2] British National Corpus (BNC), 2004. **www.natcorp.ox.ac.uk**.

[3] J.R. Carlson, J.F. George, J.K. Burgoon, M. Adkins, and C.H. White. Deception in computer mediated communication. *Academy of Management Journal*, 13:5–28, 2004.

[4] B.M. DePaulo, J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychology Bulletin*, 9:74–118, 2003.

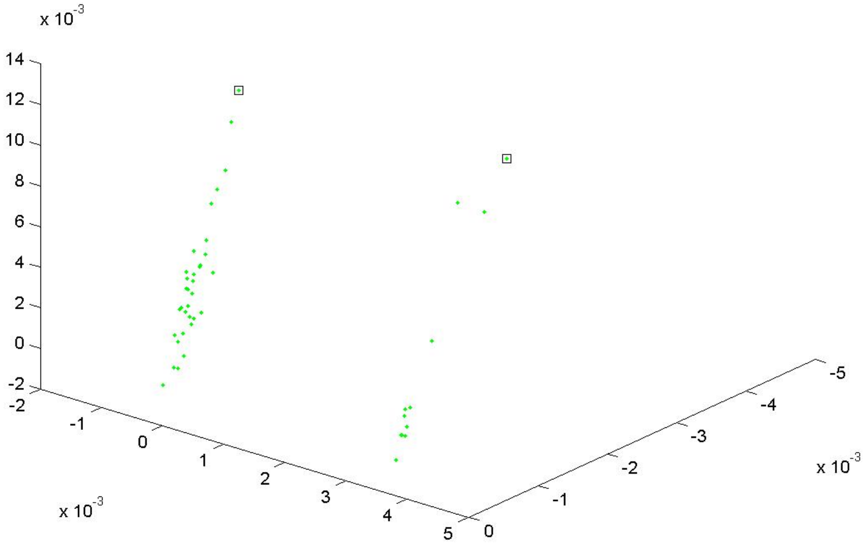

Figure 6: Emails sent by John Forney (senior trader) with his two putatively most interesting emails highlighted

[5] J.T. Hancock, J. Thom-Santelli, and T. Ritchie. Deception and design: The impact of communication technology on lying behavior. In *Computer Human Interaction (CHI2004)*, pages 129–134, April 2004.

[6] P.S. Keila and D.B. Skillicorn. Structure in the Enron email dataset. In *Workshop on Link Analysis, Security and Counterterrorism, SIAM International Conference on Data Mining*, pages 55–64, 2005.

[7] M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. Lying words: Predicting deception from linguistic style. *PSPB*, 29:665–675, 2003.

[8] J.W. Pennebaker, M.E. Francis, and R.J. Booth. Linguistic inquiry and word count (LIWC). Erlbaum Publishers, 2001.

[9] D.B. Skillicorn. Beyond keyword filtering for message and conversation detection. In *IEEE International Conference on Intelligence and Security Informatics (ISI2005)*, pages 231–243. Springer-Verlag Lecture Notes in Computer Science LNCS 3495, May 2005.

[10] L. Zhou, J.K. Burgoon, J.F. Nunamaker Jr, and D. Twitchel. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13:81–106, 2004.

[11] L. Zhou, J.K. Burgoon, D.P. Twitchell, T. Qin, and J.F. Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–165, 2004.

[12] L. Zhou, D.P. Twitchell, T. Qin, J.K. Burgoon, and J.F. Nunamaker Jr. An exploratory study into deception detection in text-based computer mediated communication. In *Proceedings of the 36th Hawaii Intl Conference on Systems Science*, 2003.