Detecting Related Message Traffic

D.B. Skillicorn
School of Computing
Queen's University
skill@cs.queensu.ca

Abstract

Governments routinely intercept messages as part of counterterrorism efforts. We consider the problem of identifying and associating messages between members of a threat group when the content is apparently innocuous and senders and/or receivers are not identifiable as particular people. We show that clusters of related messages can be identified when they use words in correlated ways (which all conversations do) *and* the words are used with the 'wrong' frequency. The proposed technique therefore complements the use of a watch list of words, since the greater the awareness that particular words should not be used, the greater the use of inappropriate words that will reveal the existence of related groups of messages.

1 Introduction

A great deal of email and voice communication comes under surveillance as part of the counterterrorism effort in many countries. Data mining can play an obvious role in this surveillance effort.

Individual messages of interest can be detected using a watch list of words that are potentially significant. For example, the Echelon system [1] intercepts satellite-detectable messages whose sender and/or receiver is outside the UKUSA intelligence sharing agreement.

Messages that are related to each other can sometimes be detected because of properties of the communication mechanism, for example:

- The identity of the sender or receiver is known, for example when landline or cell phones are used.
- The location of the sender or receiver is known, for example when radio is used.

In such situations, techniques drawn from social network analysis [2, 6] can also be used to connect messages [8], either in a virtual space of relationships or in physical space.

In this paper we consider the problem of detecting related groups of messages, for example messages among the members of a terrorist cell, when the senders and receivers cannot be directly established. This typically happens because communication is done via email accounts (which can be created so easily that it is hard to associate them with either owners or locations); via web pages (where identifying the content creator is difficult, but discovering the receivers is even harder because they are hidden inside a large number of innocent visitors to the web site); or via cell phones when the ownership of the handsets cannot be established (e.g. stolen phones or switched SIMs).

Such related sets of messages are likely to show correlated use of words; after all the messages are about something. However, many innocent sets of messages will also show similar correlated word use. Groups that are trying to avoid detection are under an added constraint: they must avoid using words that might appear on a watch list, and they can have only a rough idea of the contents of such a list. This will tend to alter their word usage, both consciously and unconsciously. They will replace words they do not wish to use with others, but the replacement words will be used in ways that do not match their natural frequency.

The contribution of this paper is to show that sets of messages that use unusual words in correlated ways can be detected using matrix decompositions. Neither ordinary conversations (using correlated words with their typical frequencies) nor

individual messages with unusual word frequencies are detected, so that typical sources of false positives do not create a problem. The ability to classify groups of messages as related expands the range of analysis that can be applied to intercepted message traffic.

In Section 2 we discuss some of the linguistic properties that are relevant to this problem. In Section 3 we introduce the matrix decompositions that we use to tackle the problem. In Section 4 we outline the structure of the datasets we use. In Section 5, we describe the experiments performed and their results. Finally, in Section 6 we draw some conclusions.

2 Patterns in conversation and email

Speech has certain characteristics that distinguish it from prose, primarily because it is produced in real-time and cannot be edited. It is widely believed by linguists that email falls in a middle ground between speech and prose, with many of the characteristics of the former – email tends to be constructed on the fly and with little, if any, editing. In what follows, we will assume that both speech and email have similar word-use structure because of their informal nature and rapid construction.

Four properties of informal language are relevant:

1. The frequency distribution of words is Zipf; furthermore the distribution of individual classes of words, such as nouns and verbs, is also Zipf. There are a number of competing explanations for why this property holds for human languages: there are simply more possible long words than short ones and this alone is enough to account for the observed distribution [7]. Whatever, the explanation, it is widely agreed that a Zipf distribution for words is a shallow (although useful) property of languages.

2. If adjacent (or almost adjacent) co-occurrence of words is considered as a relation, then English at least exhibits the small world property – almost all words are reachable from a given word along a path of length not much greater than three [5]. This suggests that words can be considered as forming a sphere, in which the core is almost fully-connected, and successive layers have rich interconnection both to the inner layers and to the other words in their own layer. Considering occurrences of word pairs or sequences seems unlikely to provide much more information than considering single word occurrences.

3. The process of sentence formation in an individual is highly individualized, notwithstanding the previous two points. Authorship studies show that elements of individual style persist even when deliberate attempts are made to conceal them.

4. Language production is largely an unconscious process, so that many aspects of utterances are hard, perhaps impossible, for the speaker to change.

Knowing that conversations and emails are under surveillance, a strategy for avoiding detection, or better still consideration, is to avoid words that might be on a watch list, and to use words in ways that are consistent with their normal usage, in particular to use rare words infrequently and common words frequently. This is not an easy thing to do in real-time since we do not use rare words rarely because we have a model of the Zipf distribution in our heads, but because of largely unconscious language production systems. Trying to consciously adjust the behavior of these unconscious systems is unlikely to work and likely to leave traces that may be detectable.

The task of avoiding consideration is further complicated by the need to talk about actual objects. Replacing the names of these objects by more common words may lead to message profiles that are unusually bland. Replacing the names of these objects by other names (as in simple forms of speech code) produces messages whose noun usage profile is unusual, perhaps including too many infrequent nouns.

It is hard to select anomalous individual messages with confidence, although messages can potentially be ranked by how anomalous they are.

However, *correlated* anomalous messages are both easier to detect and potentially more interesting. In general, a large number of messages may use an infrequent word; it is much less likely that a set of messages will use the same infrequent word. Particular turns of phrase or idioms tend to circulate among members of groups, and these may further tend to create correlations among messages. In other words, authorship, normally considered as a property associated with an individual, can plausibly be associated with a group as well, particularly a group that is already distinctive in its attitudes, goals, and in its milieu.

3 Matrix Decompositions

Matrix decompositions express a matrix, representing a dataset, in a form that reveals aspects of its internal structure. Different matrix decompositions impose different requirements on the structure of the decomposition and so reveal different structures. A typical matrix decomposition allows a matrix A to be expressed as a product

$$A \;=\; C\,F$$

where, if A is $n \times m$, the matrix C is $n \times r$ and F is $r \times m$. Sometimes a third, diagonal $r \times r$ matrix is also part of the decomposition.

There are two natural interpretations of such a decomposition. The first, a *geometric* model, interprets the rows of F as axes in a transformed space, and the rows of C as coordinates in this space. The second interpretation, a *layer* model, sees A as the sum of A_i, where each A_i is the outer product of the ith column of C and the ith row of F (and hence is the same shape as A). The outer product representation is particularly useful for judging the likely discriminative power of a matrix decomposition. If one of the A_i matrices is plotted using a colored representation for the magnitude of its entries and a region of distinct color is visible, then the rows corresponding to this region will be distant from the remaining rows when plotted in space in dimension i (and similarly for the columns). Hence the existence of distinctive regions in layer plots is a kind of shorthand for good clusterings, and potentially for good predictors.

We will use two matrix decompositions:

- Singular value decomposition (SVD) [3] for which

  $$A \;=\; U\,S\,V'$$

  where U and V are orthogonal, and S is diagonal with non-increasing entries.

- Independent Component Analysis (ICA) [4] for which

  $$A \;=\; W\,H$$

where the rows of H are statistically independent.

SVD has the property that the first new axis is aligned along the direction of maximal variation in the data, the second axis along the direction of remaining maximal variation, and so on. Each axis is orthogonal to the others, so the 'factors' corresponding to each axis are linearly independent. The truncated representation for any $k \leq m$ is the most faithful possible in that number of dimensions. A useful property of SVD is that it transforms correlation in the original data into proximity in the transformed space. Fast algorithms for computing the SVD of a sparse matrix (with complexity proportional to r times the number of non-zero entries in A) are known.

A particularly useful property of SVD is that distance of a point from the origin in the transformed space (even when the number of dimensions is reduced by truncation) represents how interesting the point is in the sense of how strongly it is correlated with all of the other points. Hence points far from the origin are most anomalous, while those close to the origin are least anomalous. Both pieces of information can be useful.

ICA is similar to SVD but selects factors (rows of H) that are statistically independent. Typically, these factors do not have a natural ordering on them, as those of SVD do.

4 Datasets

We use artificial, but plausible, datasets for our experiments. We assume that messages have been processed to generate a frequency histogram giving the number of occurrences of each word of some (potentially large) dictionary in each message. Each column of a dataset represents a word

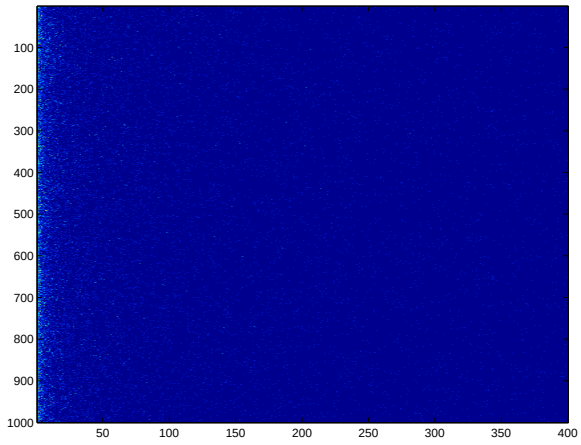


Figure 1: Word frequencies in a typical dataset (lighter colors = greater frequency)

and each row the frequencies of each word in a particular document (representing an email or telephone conversation transcript). The columns are ordered in decreasing order of natural frequency – we ignore which words they actually represent and simply generate the corresponding entries with appropriate frequencies.

The ijth entry of such a dataset is generated by sampling from a Poisson distribution with mean $f * 1/j + 1$, where f is a parameter that allows the overall frequencies to be altered, and the $1/j + 1$ term decreases the probability of the occurrence of a word depending which column represents it (so that inherently infrequent words, supposed to be represented by the later columns, are unlikely to appear in any given document). This approximates the Zipf distribution.

Figure 1 shows the distribution of word frequencies in a dataset with 1000 documents and 400 words. With $f = 3$ such a dataset has about 16000 non-zero entries (4% sparse) and each document contains about 20 distinct words. This dataset is a reasonable representation of, say, the nouns in a collection of 1000 messages.

5 Experiments

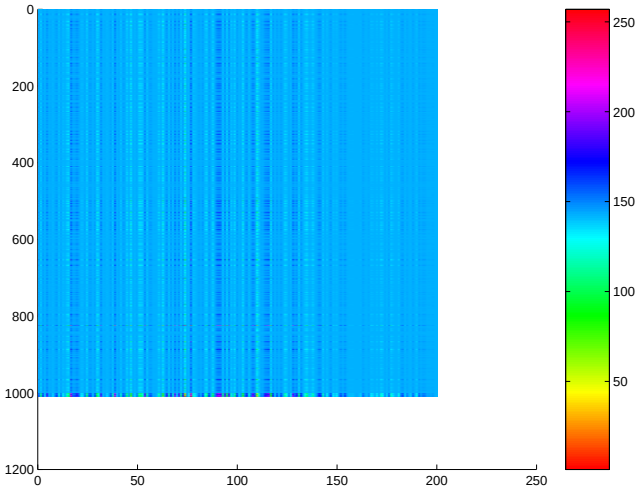
We begin with a dataset constructed as above, with a further 10 similar rows added to it (representing messages that we wish to detect). None of the techniques we use rely on the ordering of the rows so, without loss of generality, we can make these the last ten rows. In all of these matrix decompositions, we discard the first 200 columns of the dataset since the (mostly spurious) correlations among common words obscure the more interesting structure of the less-frequent words.

The first dataset illustrates a set of messages with correlated use of unusual words. A block of size 10 rows by 6 columns with uniformly random 1's and 0's is generated and added to the dataset at rows 1001 to 1010 and columns 301 to 306. Each message therefore has an overlap of 2–3 words with every other message in the group. The natural frequency of occurrences of these words is around 1% so the use in these messages is well above the background usage. This dataset represents a typical scenario in which several objects are being discussed in messages, but different, less frequent words, are being used in place of the object names.

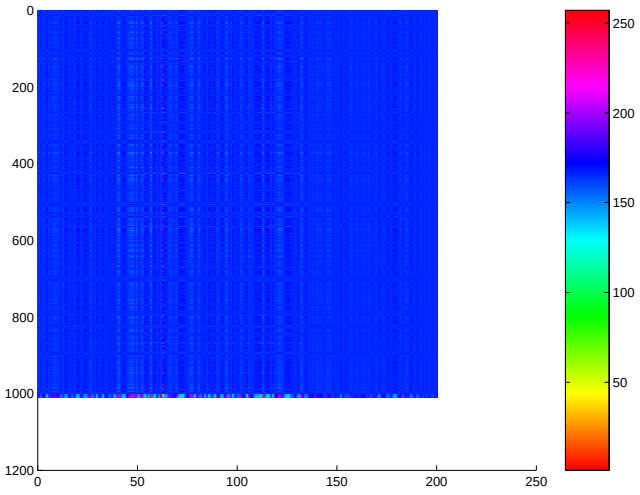
Figure 2 shows the first three layer matrices from the SVD of this dataset. Each of these matrices has the same shape as the dataset, and the colors represent the magnitudes of each entry. The band of different coloring is visible across the bottom of the first two figures. These bands show that the corresponding points (rows 1001–1010) are far from the others in these two dimensions. The bands would not, of course, be as visible if the matrix were not arranged in this way, so the layer representation only acts as a visual shorthand for the quality of the separation in this artificial setting. However, the distances of the points corresponding to these rows from the others do not depend on the order of the rows of the dataset, so the separation is always visible in a 3-dimensional plot.

Figure 3 shows such a plot using the first 3 columns of the U matrix from the decomposition. The messages in the correlated group are marked with red circles. It is easy to see how they are separated from the other messages in the first and second dimensions.

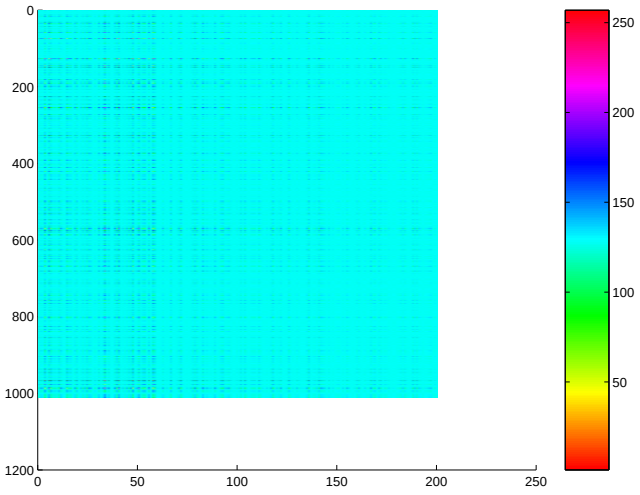
Figure 4 shows the lower right-hand corner of the truncated correlation matrix obtained by truncating each of the matrices in the SVD to $k = 3$, remultiplying to generate a matrix of the same shape as A, and then computing the correlation matrix of this new product. Such a



(a) SVD outer product at level 1



(b) SVD outer product at level 2



(c) SVD outer product at level 3

Figure 2: Layers of the SVD for a dataset containing correlated unusual word usage

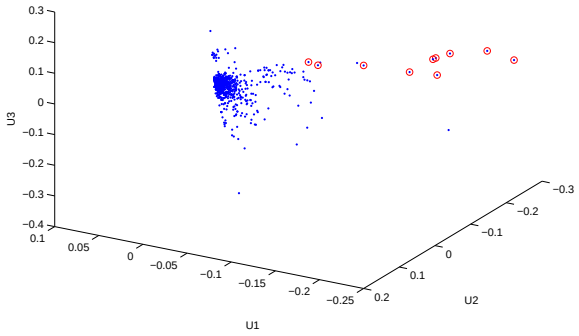


Figure 3: The 3-dimensional plot of messages for a dataset containing correlated unusual word usage

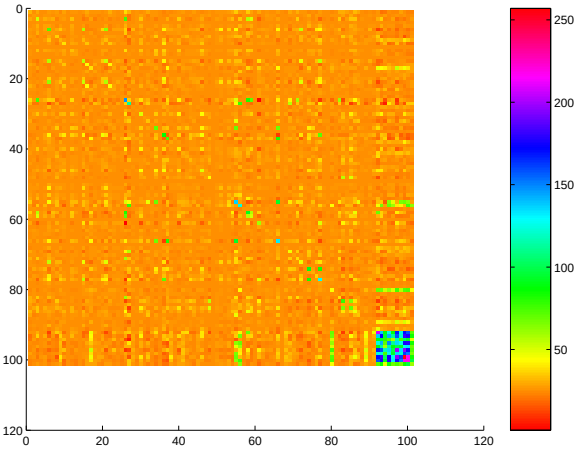


Figure 4: The truncated correlation matrix for a dataset containing correlated unusual word usage

correlation matrix captures not only the direct correlations (which would be visible in AA') but also the indirect correlations (via the use of the truncated SVD). The group of messages is clearly visible in this matrix.

Figure 5 shows the first layer matrix from the ICA of this dataset. The correlation within the message group (and its relationship to words 301-306 which correspond to columns 101–106) are clearly visible.

Figure 6 shows how the values in the layer matrix are reflected by unusual positions in the plot of points using the first 3 columns of the W matrix. Figure 7 illustrates the entire W matrix,

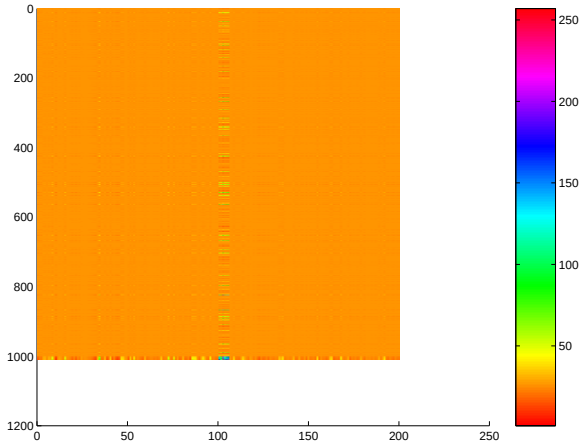


Figure 5: Layer 1 of the ICA for a dataset containing correlated unusual word use

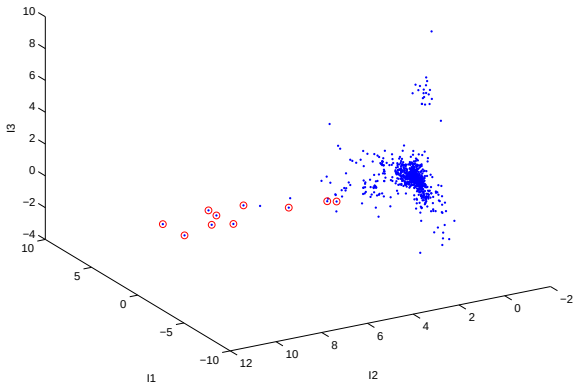


Figure 6: The 3-dimensional plot from ICA of a dataset containing correlated unusual word usage

showing how the message group can be seen as a block of unusual magnitude at the bottom of the first column.

We now show that both correlation and unusual frequency are required in order to form detectable groups of related messages. We first add to the base dataset a block of correlated messages whose frequencies are natural. To do this, we generate a block of 5 rows by 6 columns and place non-zero entries in it with frequencies appropriate to columns 301–306 of the base dataset. We then insert this block twice, at rows 1001–1005 and 1006–1010. Because there is only approximately a 1% chance of a word of this frequency being used in a message, such blocks may contain very few

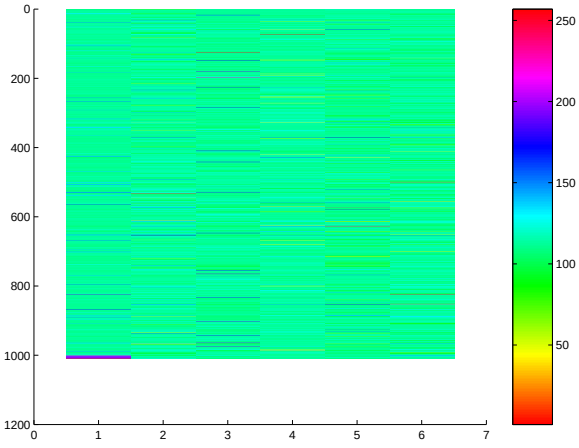


Figure 7: The W matrix from ICA for a dataset containing correlated unusual word usage

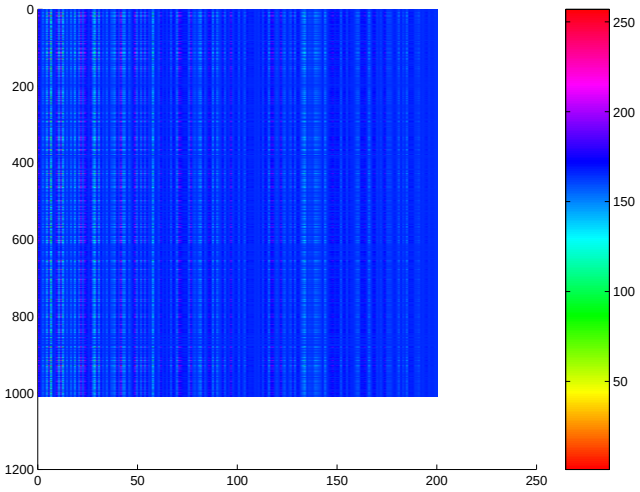
1's. However, even when $f = 30$ is used, so that there are a significant number of 1's in the repeated block, no structure is seen.

Figure 8 shows that there is no obvious structure related to rows 1001–1010. Figure 9 shows that the points corresponding to rows 1001–1010 are not separated from the main mass of points. Figure 10 shows that there is little correlation among the rows of this group of messages.

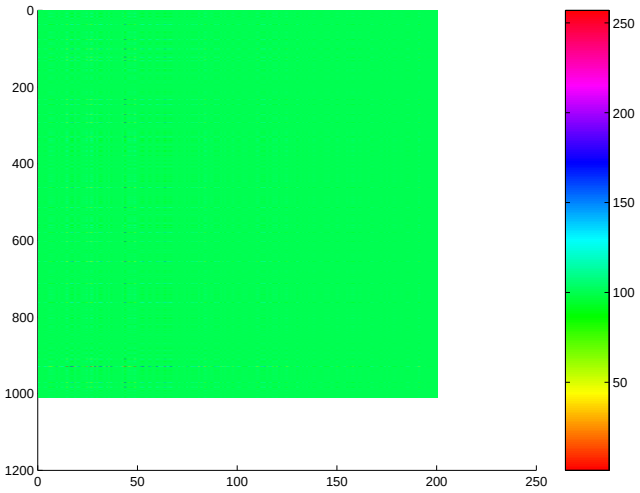
Figure 11 shows that ICA does not see any structure related to this group in the first 3 dimensions, and Figure 12 shows that there is no structure at deeper levels either.

We now consider a dataset where unusual words uses are present but they are not correlated. We generate 10 independent vectors of size 1 by 6 with a uniform distribution of 0's and 1's (as in the first dataset) and then place each of these vectors in non-overlapping columns starting from column 280. The resulting dataset therefore has 10 final rows in which rare words are used with much greater than their natural frequency.

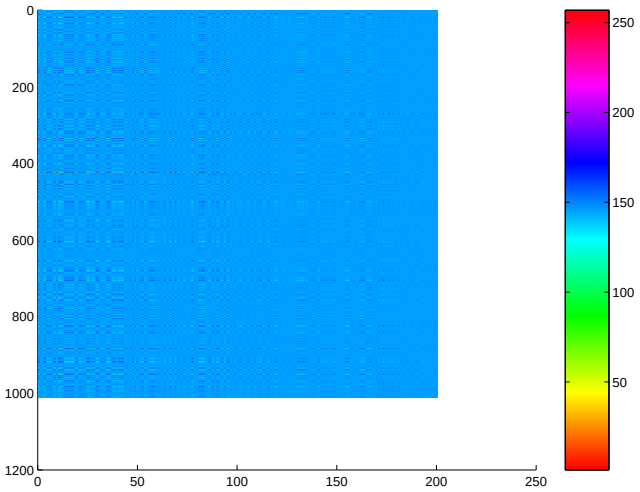
Figure 13 shows that there is no obvious structure as the result of these unusual words usages. As expected Figure 14 shows the points corresponding to these messages scattered all over the plot. Notice, though, that several of these points are far along the $U1$ axis as a result of the unusual word usage they contain. The truncated correlation matrix, shown in Figure 15, also shows the lack of correlation among these rows. Similar results for



(a) SVD outer product at level 1



(b) SVD outer product at level 2



(c) SVD outer product at level 3

Figure 8: Layers of SVD for a dataset with correlation but typical frequencies

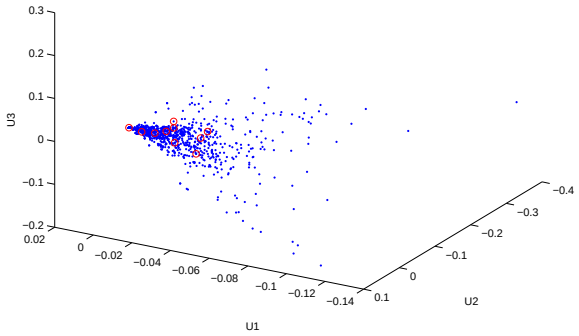


Figure 9: 3-dimensional plot from SVD for a dataset with correlation but typical frequencies

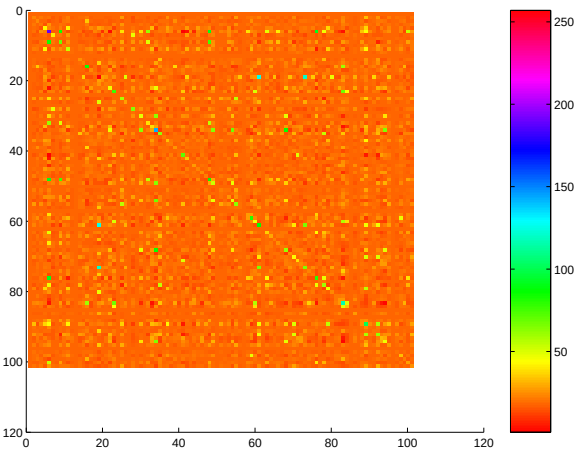


Figure 10: Truncated correlation matrix for a dataset with correlation but typical frequencies

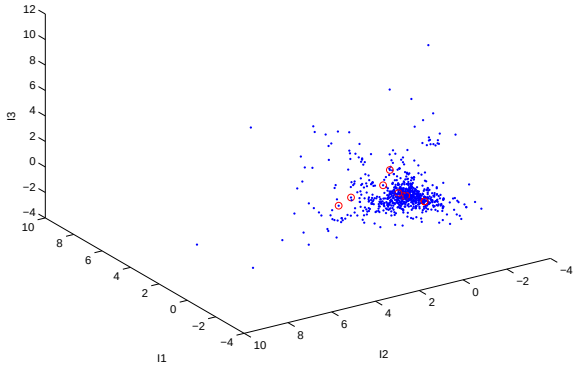


Figure 11: 3-dimensional plot from ICA for a dataset with correlation but typical frequencies

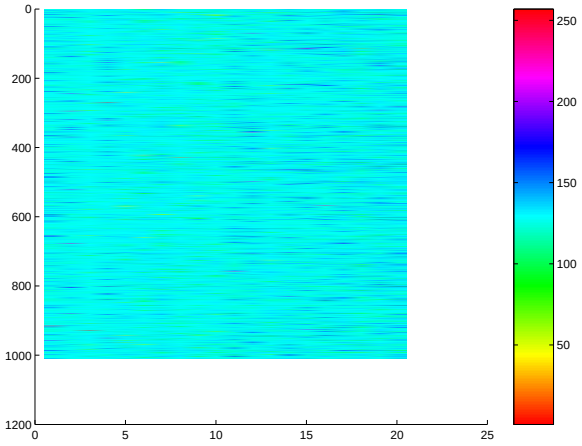


Figure 12: The W matrix from ICA for a dataset with correlation but typical frequencies

ICA can be seen in Figures 16 and 17.

Although we have shown results only for a particular base dataset and particular modifications to it, the results shown here are typical of similar datasets. Although as a matter of practicality, the datasets are small, they are not unreasonable as examples of email or phone conversations collected over a short period of time.

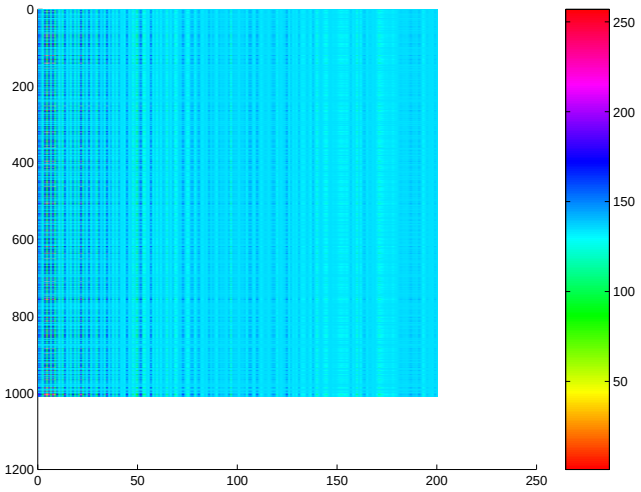
Notice that only a few dimensions of the decomposition are needed to give good results. Hence the complexity of these matrix decompositions is quite practical since the data matrices are sparse – the complexity is effectively linear in the number of messages considered.

6 Conclusion

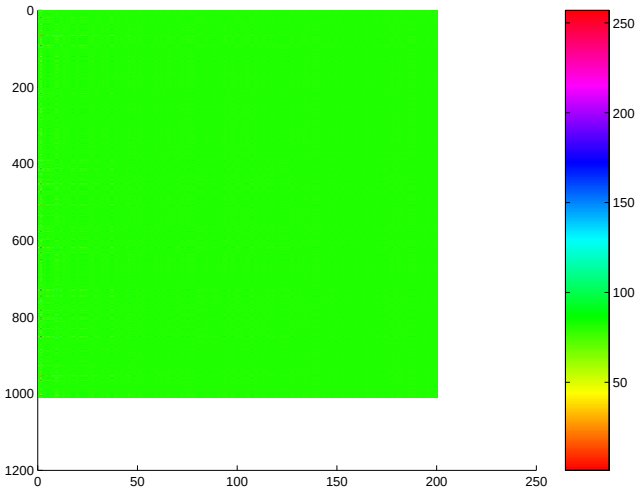
We have presented preliminary results showing that groups of related messages can sometimes be identified by internal evidence alone, that is without information about senders and receivers. Identifying such groups requires that the messages use words in a correlated way and that the words are used with the 'wrong' frequencies. We suggest that groups whose purpose is malign are likely to use words with the 'wrong' frequency because of their awareness that certain words will trigger suspicion and scrutiny. Their attempts to substitute for such words, either using speech code or on-the-fly, are likely to produce words that are either too common or too unusual, and hence to produce the message profiles that our technique will detect. We have shown that the use of correlated words alone does not trigger detection by our technique – which is just as well since most collections of messages will contain conversations with such correlations. We also show that unusual word use alone does not trigger detection by our technique – so eccentrics and non-native speakers do not cause false positives either.

References

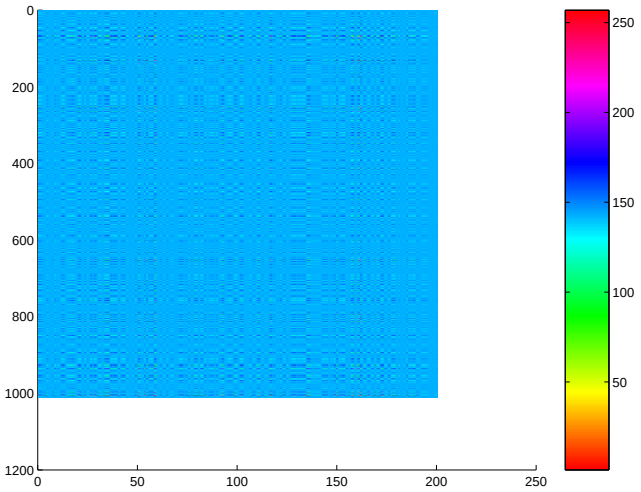
[1] European Parliament Temporary Committee on the ECHELON Interception System. Final report on the existence of a global system for the interception of private and commercial communications (echelon interception system), 2001.



(a) SVD outer product at level 1



(b) SVD outer product at level 2



(c) SVD outer product at level 3

Figure 13: Layers of the SVD for a dataset with unusual word frequencies but not correlation

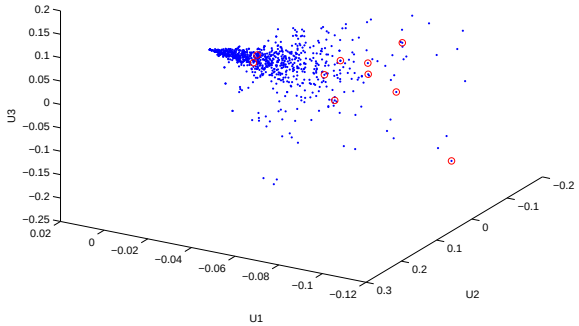


Figure 14: 3-dimensional plot from SVD for a dataset with unusual word frequencies but not correlation

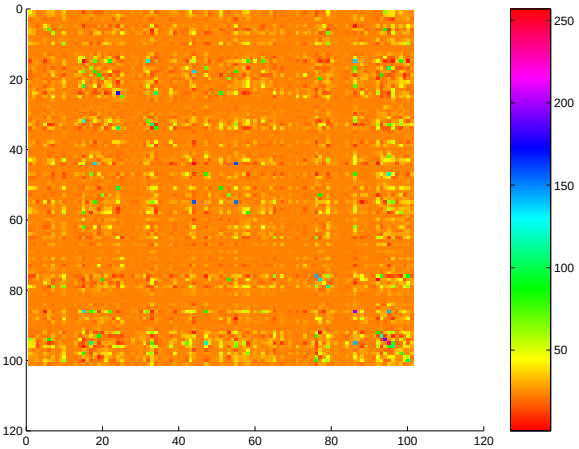


Figure 15: Truncated correlation matrix for a dataset with unusual word frequencies but not correlation

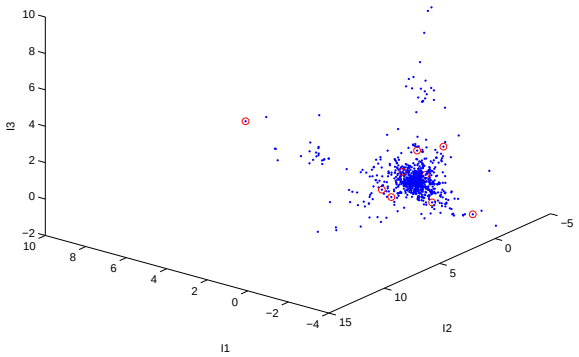


Figure 16: 3-dimensional plot from ICA for a dataset with unusual word frequencies but not correlation

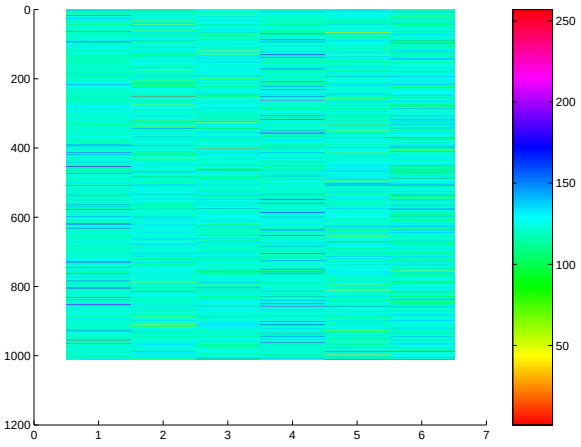


Figure 17: The W matrix from ICA for a dataset with unusual word frequencies but not correlation

[2] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1), 1997.

[3] G.H. Golub and C.F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[4] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[5] R. Ferrer i Cancho and R.V. Solé. The small world of human language. *Proceedings of the Royal Society of London Series B – Biological Sciences*, pages 2261–2265, 2001.

[6] V.E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[7] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEETIT: IEEE Transactions on Information Theory*, 38(6):1842–1845, 1992.

[8] J.R. Tyler, D.M. Wilkinson, and B.A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. HP Labs, 1501 Page Mill Road, Palo Alto CA, 94304, 2003.