Clusters Within Clusters: SVD and Counterterrorism

D.B. Skillicorn
skill@cs.queensu.ca

March 2003
External Technical Report
ISSN-0836-0227-
2003-463

Department of Computing and Information Science
Queen's University
Kingston, Ontario, Canada K7L 3N6

Document prepared March 13, 2003

Abstract

We argue that one important aspect of terrorism detection is the ability to detect small-scale, local correlations against a background of large-scale, diffuse correlations. Singular value decomposition (SVD) maps variation, and hence correlation, into proximity in low-dimensional spaces. We show, using artificial datasets whose plausibility we argue for, that SVD is effective at detecting local correlation in this setting.

The figures in this paper can be understood in black and white but are designed to be seen in colour.

Clusters Within Clusters: SVD and Counterterrorism

D.B. Skillicorn
skill@cs.queensu.ca

Abstract: We argue that one important aspect of terrorism detection is the ability to detect small-scale, local correlations against a background of large-scale, diffuse correlations. Singular value decomposition (SVD) maps variation, and hence correlation, into proximity in low-dimensional spaces. We show, using artificial datasets whose plausibility we argue for, that SVD is effective at detecting local correlation in this setting.

1 Introduction

Detecting terrorism can be posed as an unsupervised data mining problem in which the goal is to separate individuals into two classes, threats and non-threats. However, it is unusual because the members of one class (the threats) are actively trying to look as similar to members of the other class as possible. Without information about the particular data mining algorithm in use, the best strategy for doing this is to arrange for their attribute values to be modal.

This has two implications for data mining in terrorism detection: attributes should be such that it is hard to manipulate their values; and the data mining algorithms used should rely on the *relationships* between attributes, rather than simply their values. If attributes are chosen appropriately, then the activities of terrorists and terrorist groups may be visible as unexpected correlation, both among themselves and between the terrorists and their target. However, this correlation must be detected against a background of widespread diffuse correlation in the population at large.

Singular value decomposition is a useful tool for detecting unusual correlation because it transforms variation into proximity. Both distance measures and visual inspection can detect proximity far more easily than they can detect correlation directly.

We present preliminary results using artificial datasets. There is little experience to guide the form of such datasets, but we argue that the ones we use are at least plausible.

Results are encouraging, in the sense that SVD has high detection accuracy with reasonably low false positive rates. We are unable, so far, to specify an ideal algorithm schema for applying SVD, but we show that several strategies are effective.

2 Goals and assumptions

It seems implausible, given our present data mining technologies and understanding of the problems of counterterrorism, that data mining will be able to be deployed as a frontline tool against terrorism (at least in the immediate future). However, a useful role for data mining is as a filter, making it economic to select a manageable subset of individuals for further scrutiny using traditional intelligence techniques. In this view, the benefit of data mining is primarily to improve the effectiveness of other counterterrorism methodologies.

We assume, for the sake of concreteness, that we are dealing with datasets whose rows describe individuals and whose columns are attributes of those individuals. For example, datasets might contain information about which cities an individual has visited, or which flights he has taken.

Goals.

The untargeted case. Given a dataset, find clusters whose correlations are stronger than average, and select its members. If a group of terrorists are detectable as a cluster within the background of other clusters, then their target may also be detectable as part of the same cluster.
The targeted case. Given a dataset and a target, find clusters around the target whose correlations are stronger than expected. Any individual target might be expected to be part of multiple clusters representing his or her interests and collaborations. Hence, a cluster around them is likely to include parts of other clusters in which they are involved. This diffuse pattern should show an unusual concentration if there is a tightly-knit group that is focused on the target.

In the targeted case, the target is of the same kind as the objects described by the rows of the dataset, individuals in this case. In the untargeted case, other kinds of targets are possible.

Assumptions. We wish to construct a detection model that will select some of these individuals for further scrutiny. We make the following assumptions:

- The potential consequences of failing to detect a terrorist are so great that a fairly high level of false positives is acceptable.
- Terrorists act in groups, so individual false negatives are acceptable provided that at least one member of a group is detected.

Attributes. The attributes in such datasets can be usefully divided into two kinds:

- *Incidental* attributes that describe properties and actions that are believed to be potentially correlated to terrorism. These may be static, such as country of citizenship, gender, income and so on; or based on actions such as purchasing particular kinds of plane tickets.
- *Intrinsic* attributes that describe properties and, more commonly, actions that are necessary to carry out a terrorist action, for example carrying out surveillance on a target site.

Both kinds of attributes have values that are shared by terrorists and the general population. The problem with incidental attributes is that if terrorists can learn the values of these attributes that trigger the detection model (and they can), then they can arrange to appear innocent. This leaves the detection model with a 100% false positive rate, which is the worst possible outcome.

The mechanism by which terrorists can learn the relationship of attributes to the detection model is by *probing*, the so-called Carnival Booth algorithm [4]. Terrorists arrange to be considered by the detection model while behaving innocently. Those who do not trigger the model can be reasonably certain that they will not trigger it again on subsequent, less innocent occasions. The use of incidental attributes is a major weakness of airline passenger profiling systems.

Models based on incidental attributes can be made more robust by adding uncertainty into the selection mechanism. This can be done by wrapping the detection model in a layer that obscures its precise functioning, for example, by randomly selecting some individuals who do not trigger the detection model and treating them as if they did. It can also be done by using families of detection models based on different thresholds for individual attributes (i.e. different discretizations of continuous data) or on different sets of attributes. These techniques all break the assumption that a person who has not been selected by the detection model on one occasion will not be selected on another occasion. However, these techniques all add expense and complexity; and using families of detection models risks one model failing to detect a threat that another model would have, which may be politically unacceptable if an incident takes place.

Intrinsic attributes are inherently better because terrorists are *forced* to have certain values for them. Of course, some of the general public will also share these values; but such attributes allow the set of individuals to be separated into those who are not terrorists and those who might be. As we have seen, incidental attributes do not do this reliably. Moreover, the set of individuals who can be eliminated will tend always to be much larger than the set of possible threats who remain.

The use of intrinsic attributes forces terrorists to come under scrutiny. Their only strategy then is to conceal themselves among that part of the population who share the same attribute values – but this becomes harder and harder as the number of attributes increases.

The power of intrinsic attributes can be seen in the aftermath of a terrorist action. Once such an action has taken place, the set of relevant attributes is clear – to plant a bomb in a certain place requires being in that place, for example. And once the correct set of attributes is known, terrorists are often detected very quickly. (This is also the basis of much police work – an alibi is a value for a very specific attribute which eliminates many possible perpetrators.)

Terrorists can only try to conceal their forced actions among those of many others. This is difficult for two reasons:

- A terrorist group is forced to make coordinated actions, and such actions are potentially visible as correlations in the data. For example, if they meet to plan, then they are located at the same place at the same time.

- A terrorist group is forced to carry out actions that are correlated with their target, and these actions are also potentially visible in the data. For example, they may travel the same route as the target but earlier in time.

These properties of datasets available for counterterrorism suggest that the problem is not closely related to outlier detection because terrorists try, as far as possible, to take on modal values for attributes. However, if intrinsic attributes are used, terrorist groups cannot avoid correlations both among themselves and with their targets. It is these correlations, which reveal themselves as locally dense regions within appropriate representations of the dataset, that data mining must search for – and which suggest the title of this paper. Some evidence for this is provided by Krebs [12], who analyzed the connections among the group involved in the destruction of the World Trade Center. He showed that the members of the group were indeed tightly correlated. Of particular note is that a single meeting among a subset of them reduced the mean distance between members by 40% from its value given their relationships alone. Such is the power of intrinsic action attributes.

An immediate concern is that datasets describing any human population will be full of correlated subsets, and it might prove impossible to detect the correlations due to terrorism against such a noisy background. Consider the dentists of Pittsburgh[1]. We might expect that they would appear as a correlated group – they come from similar (educated) socioeconomic groups, they live in similar settings, and they travel to similar conferences and conventions. However, as we consider more aspects of their lives, these correlations begin to be diluted by others: they travel to differing parts of the country for family occasions, their children insist on holidays in different places, and they have different hobbies. The terrorists of Pittsburgh (should there be any) might also appear strongly correlated by a few attributes, but this correlation is much less likely to dilute as further attributes are considered.

More formally, the reasons why correlation in terrorist groups might be visible against a background of widespread correlation are these:

[1] Apologies to both dentists and Pittsburgh for this example.

1. Most individuals are part of a fairly large number of subgroups with whom they are correlated – enough that the strength of membership in each one is quite small.

   Consider the folk theorem about six degrees of separation, the contention that a chain of acquaintances of length less than or equal to six can be built between any two people in some large population (originally the population of the U.S. in Milgram's original work, now often claimed for the total world population). If a given individual is acquainted with (say) a individuals, then each of these a individuals must be acquainted with a fairly large number of others outside the original set of a or else the powers do not increase quickly enough (since $a^6 \approx$ the large population).

   This result contradicts our intuition that an individual's social circle tends to be small. The resolution (see, for example, [14]) is that such small social circles are bridged by rare, but not too rare, 'long-distance' connections.

   Acquaintanceship is a reasonable, although not perfect, surrogate for correlation in the kind of datasets we are interested in – we would not be surprised that acquaintances would turn out to be fairly well correlated in large datasets – they live in similar places and have similar lifestyles, including travel arrangements. What is less obvious is that the 'long distance' connections in acquaintanceship are likely to produce strong correlations as well – for an acquaintanceship survives only if its members have 'something in common'. Hence the implication of six degrees of separation (and the existence of short paths in acquaintanceship graphs) is that correlation smears rapidly across subgroups because of the richness of cross-connections of common interests and behavior.

2. We might expect terrorists to be substantially less connected by correlation than most people because they have a much narrower focus. Informally, we might suspect that terrorists don't buy life insurance, don't take holidays, don't buy lottery tickets, and don't have children in Little League. We quote from Krebs [12, p49], relying on previous work on the social network structures of criminals: "Conspirators don't form many new ties outside of the network and often minimize the activation of existing ties inside the network".

These properties provide some assurance that a signature for terrorist actions exists in datasets that are sufficiently large and diverse. Note that, in this context, high dimensionality is a benefit because it acts to smear the background correlation in the population at large.

3 Data Generation Models

Since, for obvious reasons, real datasets containing terrorist actions are not available, the quality of detection models will have to be evaluated using artificial datasets. This immediately raises the question of what kinds of datasets are plausible and, of course, any choice is open to criticism.

Intrinsic attributes can be divided into those related to actions and those related to state. We now consider the properties of each.

For action attributes, an immediate issue is how to handle their temporal nature. They could be coded with time signatures attached and temporal data mining techniques used – but I am not aware of any present data mining technology powerful enough to detect temporal subsequences when different parts of them are carried out by different individuals (this is an interesting problem, though). It seems simpler, and perhaps more robust, to handle temporal properties by creating attributes for actions covering a period of time. For example, if visits to New York are an action of interest, then these can be converted into attributes as visits per month: January visits, February

visits, and so on. It is also sensible to use overlapping time periods (creating partly correlated attributes) to avoid sensitivity to boundary choices.

Attributes representing actions will also be:

- Sparse, because only a small fraction of the total population of individuals will carry out any given task (e.g. only a small fraction of the U.S. travelling public visit San Francisco in a given month).
- Have a frequency distribution whose mode is close to 1 and which decreases quickly (e.g. those people who visit San Francisco in a given month mostly visit only once [2]).

Such attributes can plausibly be generated by first introducing a high level of sparseness and then generating the nonzero values using a Poisson distribution with mean close to 1.

State attributes will have much flatter distributions. For example, the locations of residences of members of a terrorist group around a target might be expected to conform to a normal distribution because of the pressures for closeness to the target, counterbalanced by the pressure to remain far from each other. State attributes will also tend to be dense (everyone has to live somewhere).

It is, of course, arguable that other distributions are appropriate; for example, human intervention often creates distributions with heavy tails because humans deal quickly with the easy cases, leaving only harder ones.

4 Singular Value Decomposition

Singular Value Decomposition (SVD) [7] is a well-known technique for reducing the dimensionality of data.

Suppose that a dataset is represented as a matrix A with n rows (corresponding to individuals) and m columns (corresponding to their attributes). Then the matrix A can be expressed as

$$A = USV'$$

where U is an $n \times m$ orthogonal matrix, S is an $m \times m$ diagonal matrix whose r non-negative entries (where A has rank r) are in decreasing order, and V is an $m \times m$ orthogonal matrix. The superscript dash indicates matrix transpose. The diagonal entries of S are called the *singular values* of the matrix A.

One way to understand SVD is as an axis transformation to new orthogonal axes (represented by V), with stretching in each dimension specified by the values on the diagonal of S. The rows of U give the coordinates of each original row in the coordinate system of the new axes.

The useful property of SVD is that this transformation is such that the maximal variation among objects is captured in the first dimension, as much of the remaining variation as possible in the second dimension, and so on. Hence, truncating the matrices so that U_k is $n \times k$, S_k is $k \times k$ and V_k is $m \times k$ gives a representation for the dataset in a lower-dimensional space. Moreover, such a representation is the best possible with respect to both the Frobenius and L_2 norms.

SVD has often been used for dimensionality reduction in data mining. When m is large, Euclidean distance between objects, represented as points in m-dimensional space is badly behaved. Choosing some smaller value for k allows a faithful representation in which Euclidean distance is practical as a similarity metric. When $k = 2$ or 3, visualization is also possible.

[2] The Zipf distribution is a plausible distribution for attributes such as these.

Another way to understand SVD is the following: suppose that points corresponding to both rows and columns are plotted in the same k-dimensional space. Then each point corresponding to a row is at the weighted median of the positions of the points corresponding to the columns and, simultaneously, each point corresponding to a column is at the weighted median of the positions of the points corresponding to the rows. Hence SVD can be viewed as translating correlation or similarity into proximity. Unfortunately, only positive correlation is taken into account by SVD, so that rows that are strongly negatively correlated will not be placed close together in space.

SVD measures variation with respect to the origin, so it is usual to transform the matrix A so that the attributes have zero mean. If this is not done, the first singular vector represents the vector from the origin to the center of the data, and this information is not usually particularly useful. For example, when A is the adjacency matrix of a graph, it is the second singular vector which describes the partitioned structure (if any) of the graph.

While SVD is a workhorse of data manipulation, it has number of subtle properties that are not well-known. We will use four of them.

Fact 1: The singular value decomposition of a matrix is insensitive to the addition (or subtraction) of independent zero-mean random variables with bounded variance [1]. This property has been used to speed up the computation of SVD by sampling or by quantizing the values of the matrix. In counterterrorism, the effect we are looking for is so small and the results so important that neither of these is attractive. However, the fact does explain why SVD is good at detecting clusters within clusters – the outer cluster representing the majority of the data has zero mean (by normalization) and so, by the *fuzzy central limit theorem*, increasingly resembles a normal distribution as the number of ordinary individuals (and the number of attributes) increases.

Fact 2: SVD is a numerical technique, and so the magnitudes of the attribute values matter. However, multiplying the attribute values of a row of A by a scalar larger than 1 has the effect of moving the corresponding point further from the origin. Because the positions of all of the other points depend, indirectly, on their correlations with the scaled point, via their mutual interactions with the attributes, points that are correlated with the scaled point are pulled towards it. When there is little structure in the low-dimensional representation of a dataset, this scaling technique can be used to find the individuals who are (positively) correlated with a given individual. In practice, this often makes it easier to see a cluster that would otherwise be hidden inside another in a visualization.

Fact 3: Although SVD translates only positive correlation to proximity, negative correlation information can be extracted from the SVD indirectly. Let A_k be the product

$$A_k \;=\; U_k S_k V_k'$$

The matrix $C = A_k A_k'$ can be understood as a kind of correlation matrix in which some kinds of correlation have been discarded (those arising from dimensions $k+1$ and higher) while some higher-order correlation information has been included [10, 11]. In other words, entries in C are non-zero even when the corresponding entry of AA' was zero (that is, even when there is no direct correlation between a pair of individuals).

The connection between the sign of entries in this matrix and correlation was noticed empirically by Konstotathis and Pottenger [11]. It is also related to a well-known technique for partitioning graphs using spectral methods [2, 9]. The following explanation shows why the magnitude of the

entries of C can be regarded as correlations (both positive and negative) between individuals. Consider the ijth entry of C. This entry arises as a sum of values, each of which is the product of a column of U, an entry of S, and a row of V. Such an entry is negative when individuals i and j are on opposite sides of the origin in one of the dimensions. A sum of such values represents an average 'reflection' in the origin in all k dimensions.

When the ijth entry of C is negative, we can conclude that individuals i and j are negatively correlated. In fact, we can go further – when entry C_{ij} is smaller than C_{ii} we can conclude that the correlation between individuals i and j is weak.

Fact 4: The decomposition depends on all the data used, both normal and anomalous. The precise geometry of the detection boundary of SVD is hard to predict without performing the decomposition, and impossible without knowledge of the dataset. Hence, a terrorist group cannot reverse engineer the transformation to determine how they will appear, even knowing that SVD is being used. In particular, SVD is resistant to probing attacks since any attempt to probe cannot control for the innocent individuals considered at the same time.

5 Algorithms

There are a number of algorithmic tools based on SVD that can be combined in various ways. Our results do not indicate a clear optimal strategy for using SVD, but they do reveal several effective tactics.

If we start with a high-dimensional dataset (e.g. $m = 30$), then we can apply SVD, truncate U to two or three columns and plot the corresponding rows. Their positions are the best low-dimensional representation of the original data.

Zero-mean normally distributed data appears as a spherical cluster centered on the origin even in low dimension. Any correlated set of individuals tends to appear as a cluster further from the origin along the first singular vector (and sometimes the second). Hence, a frequency plot of the first column of U may sometimes reveal a possible target and associated terrorists. Visual inspection of the plot can also be revealing.

When a target is known, there are several further options for selecting individuals as potential threats:

- Project points onto the vector from the origin to the target point in the transformed two- or three-dimensional space, and classify either the individuals whose points are further from the origin than the target, or are close to the target as threats.
- Classify the t closest neighbors of the target in two- or three- dimensional space as threats.
- Classify the individuals whose points fall in a cone from the origin centered at the target as threats (i.e. the cosine similarity used in latent semantic indexing [5]).

In all of these techniques, the target can be selected after the SVD has been computed – hence only one SVD is required.

Correlation information can be used in three ways:

- The selection mechanisms described above can be used in the plots based only on the points correlated with a particular target. This requires only replotting and not recomputation of the SVD. However, as we will see, there is little to be gained from this.

- Individuals who are not correlated with the target can be successively removed, and the SVD repeated on the resulting smaller dataset. This typically reduces the dataset size by 75% or more, but the contraction at each repetition becomes smaller because the remaining individuals all have fairly strong correlation with the target. All SVDs after the first have to be recomputed for each target because the winnowed datasets are target dependent.

- The sizes of the datasets remaining after each round themselves provide information. If the target does not have an unusual correlation with other individuals then the contracted datasets shrink in size quite slowly. When there is an unusual correlation of other individuals with the target, the contracted datasets tend to become smaller quite rapidly.

6 Experiments

In the experiments that follow, the part of the matrix A representing normal individuals will consist of 1000 rows and 30 columns. The 30 columns represent a set of attributes about each individual – we assume that these are intrinsic attributes and that a threat is forced to correlate with a target in the values of at least some of these attributes. Each dataset has a small number of additional rows added to represent a terrorist group. The results presented are qualitative, partly because there are too many free parameters to make an exhaustive analysis straightforward, and partly because there is not yet agreement about what structures in datasets are plausible. However, the results are for the first random dataset of each kind generated – no selection of datasets to provide better than average results was made. Many of our experiments were more clear cut than the examples reported here.

In plots of two- or three-dimensional space, points corresponding to normal individuals are shown as (blue) dots, the target is shown as a (red) star, and the points corresponding to terrorist as (blue) squares.

Experiment 1. We begin with a dataset in which the points corresponding to ordinary individuals are generated distributed normally around the origin with variance 1. A terrorist group of size 10 is generated distributed normally with variance 1 around one of the normal individuals. Figure 1 shows how SVD can detect a small cluster against a background cluster. Here we assume that no target is specified beforehand – the labelling confirms the fairly clear presence of a small cluster to the left of the main cluster.

Figure 2 plots only those points that are correlated with the target. This is, of course, artificial since we are assuming that we do not know the target. However, it illustrates what selection for correlation is doing – it removes many points but does not help much with identifying the terrorist cluster because the points that are removed are far from the target in the transformed space.

Experiment 2. In the previous experiment, the terrorist cluster had the same variance as the base cluster. Hence, it is likely that points from the terrorist cluster will be overrepresented among points far from the origin. We now show that this is not the reason for the quality of the SVD plot by repeating the experiment with the variance of the terrorist cluster at 0.5. We now expect points from the terrorist cluster to remain inside the background cluster on average.

Figure 3 shows that the presence of an outlier cluster in the transformed space is as clear as it was before. Figure 4 plots only those points correlated with the target and, as before, the separation of the target cluster is a little clearer.

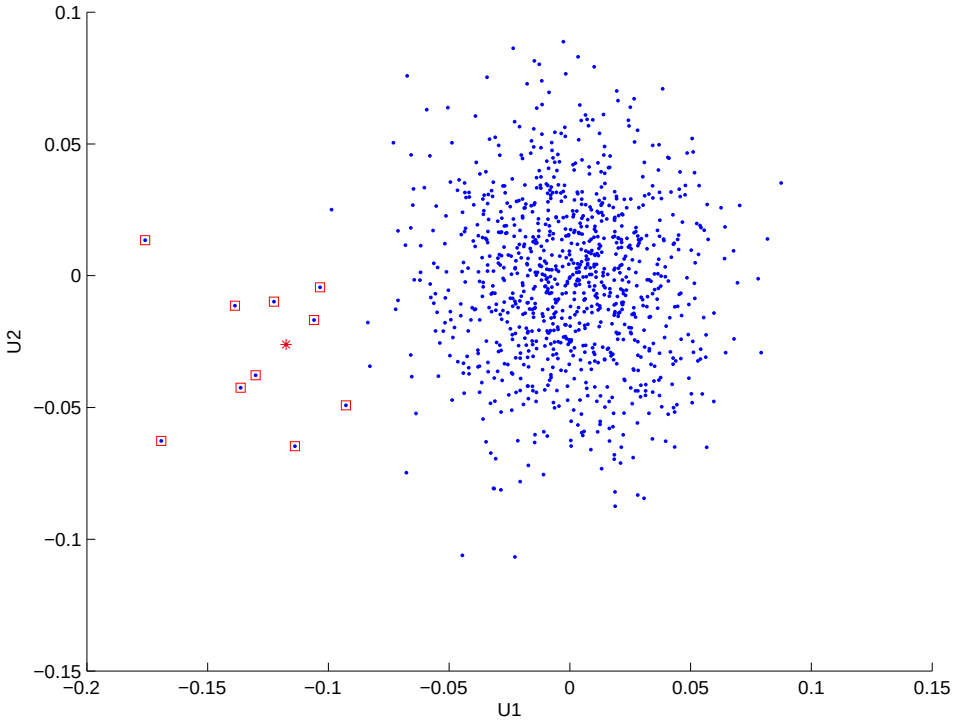

Figure 1: Plot of first two dimensions of the transformed space. Individuals normally distributed with mean 0 and variance 1, one individual randomly chosen as target, terrorists normally distributed around that individual in 30 dimensions with variance 1.

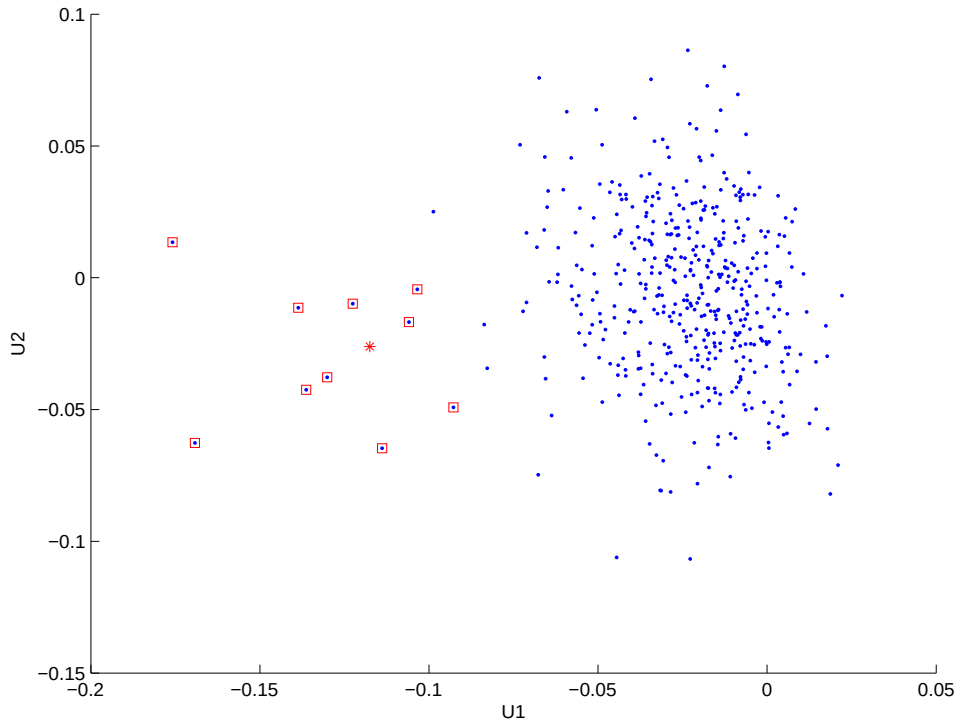

Figure 2: Same plot as in Figure 1 showing only individuals correlated with the target. Note the change of scale on the axes.

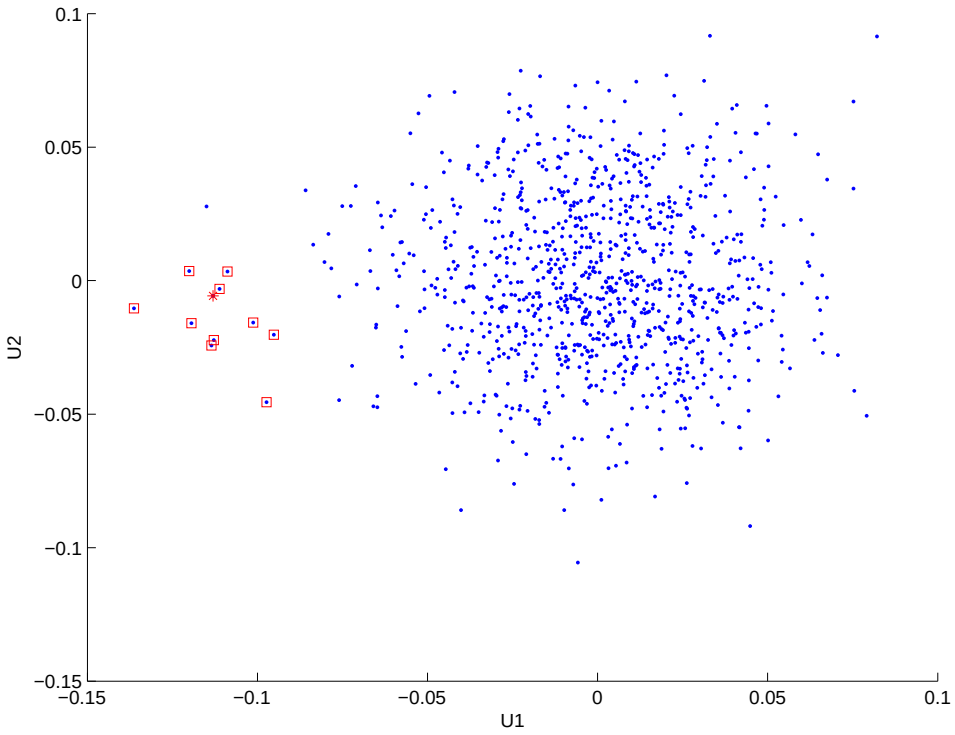

Figure 3: Individuals normally distributed with mean 0 and variance 1, one individual randomly chosen as target, terrorists normally distributed around it with variance 0.5.

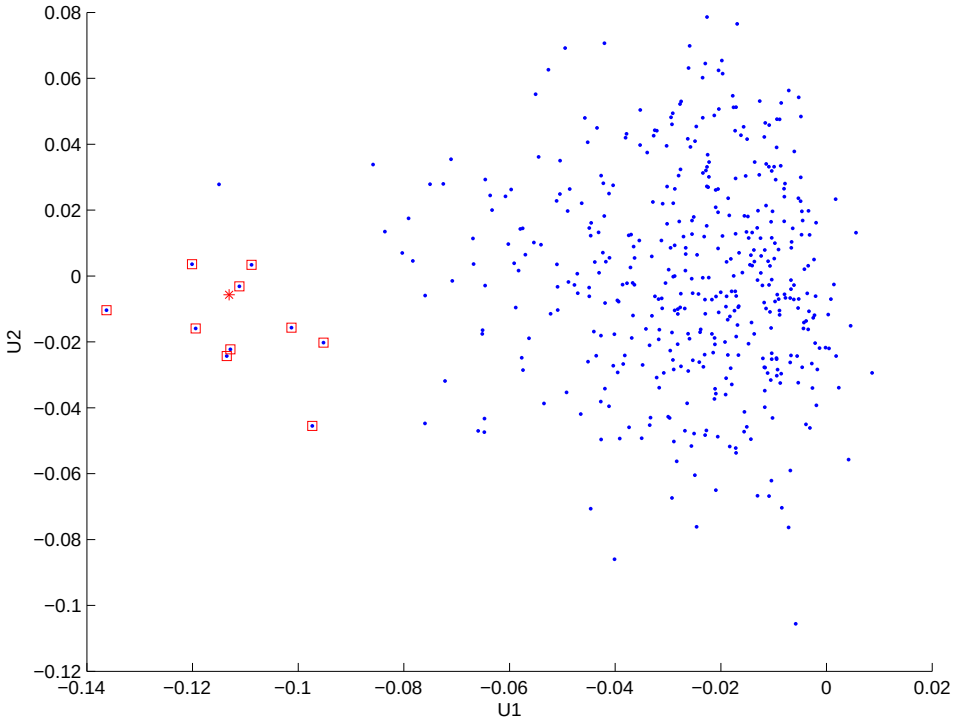

Figure 4: Same plot showing only individuals correlated with the target.

In these datasets, SVD discovers the terrorist cluster without knowing the target because, by Fact 1, the base cluster has little linear structure. The first singular value is overwhelmingly likely to point towards the median of the smaller embedded cluster around the target even if its points are entirely inside the larger cluster.

This is arguably an easy dataset, but not entirely trivial because the fuzzy central limit theorem suggests that, given enough data, and given that normalization takes place after the data is collected, we can expect that many parts of a dataset should look as if they were generated by a normal distribution.

Experiment 3. We now consider a dataset with following structure: 100 points are generated, normally distributed around 0 with variance 1. 100 clusters of 10 points are generated, normally distributed with variance 1 with centers at each of the original points. A terrorist cluster of size 10, normally distributed with variance 1 is generated around a random one of the second level points. So rather than a single background cluster around zero, we have a large set of background clusters with many different centers.

Figure 5 shows the resulting plot. It is clear that the terrorist cluster cannot be distinguished from the background without prior knowledge of the target, and only imperfectly then. Even when the uncorrelated points are removed from the plot (Figure 6), the terrorist cluster is not cleanly separated either by projection along a line to the target or by proximity to the target. However, Figure 7 shows that the target cluster is fairly well identified using cosine similarity to the target (we have shown a cluster that would include the entire terrorist group, but many cones with smaller angles would also detect several members of the group).

Experiment 4. Figure 8 shows the plot for the dataset of Experiment 3 when the row corresponding to the target is scaled by a factor of 1.2. Figure 9 shows only the points correlated with the target. There is very little difference between these plots and those where the target row is unweighted.

However, we now repeat the SVD using only those 422 rows of the original matrix that are correlated with the target. The results are shown in Figures 10 and 11 (with only the points that are still correlated with the target plotted). Both projection onto a vector from the origin to the target, and proximity now begin to discover the terrorist cluster.

Figures 12 and 13 show what happens after a third round of SVD on the 362 points correlated with the target on the previous round. Both projection onto a vector and proximity continue to improve their predictive performance, and now clearly identify the terrorist group. Notice the flattening of the number of uncorrelated points being removed at each stage.

Experiment 5. Using the dataset of Experiment 3, we now multiply the row corresponding to the target by 4. Figure 14 shows what happens – the target point moves far from the origin, but it also tends to pull the correlated points towards it, and so away from the main cluster. Both proximity and proximity on the projection onto the vector from origin to target are effective at finding the terrorist cluster.

Experiment 6. In the previous experiments (Experiments 3–5), the terrorist cluster was still distinguished because it was the only cluster at the 'third' level. We now generate a dataset with 100 points, normally distributed around 0 with variance 1. 100 clusters of 10 points are generated, normally distributed with variance 1 around each of the original points, and 20 clusters of size 10

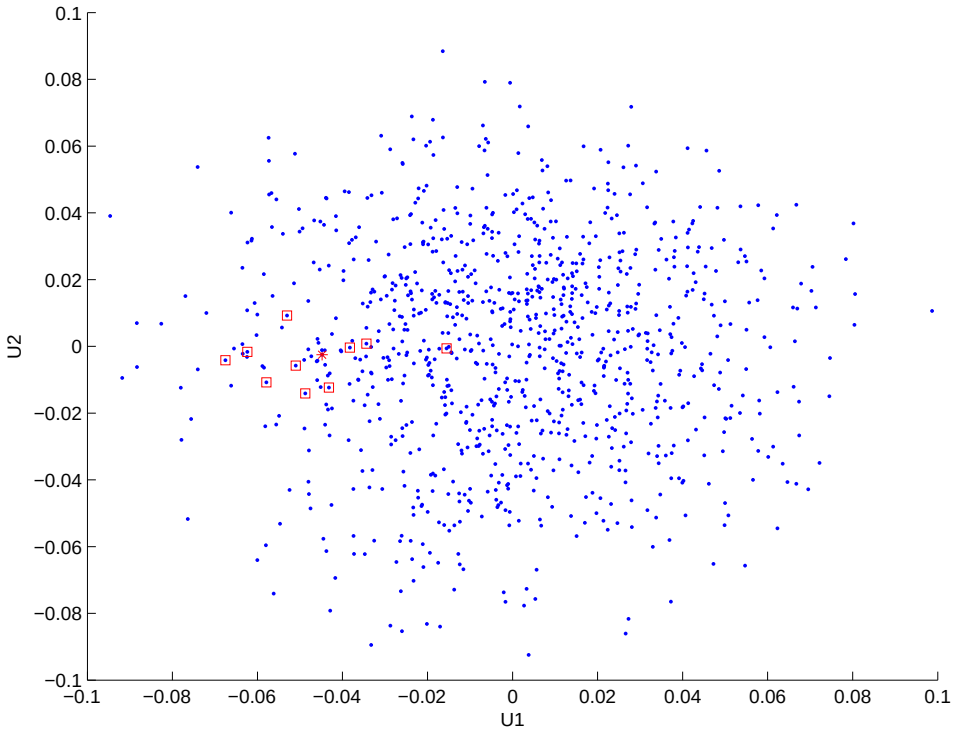

Figure 5: 10 normal clusters with variance 1 with centers drawn from a normal distribution with mean 0 and variance 1; terrorist cluster normally distributed around a randomly chosen individual with variance 1.

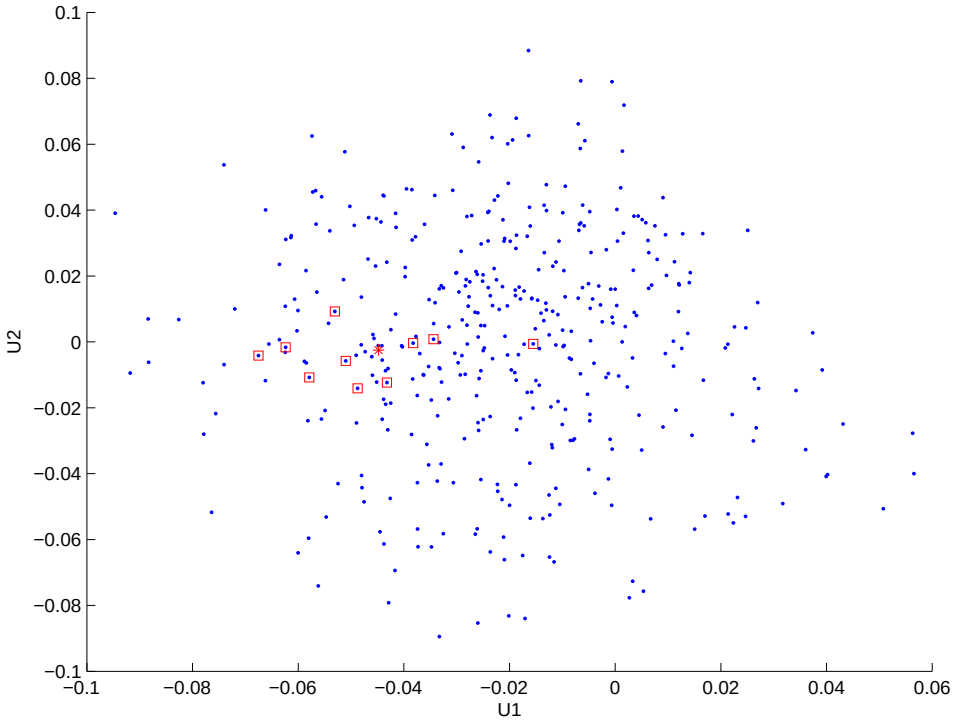

Figure 6: Same plot showing the 422 individuals correlated with the target.

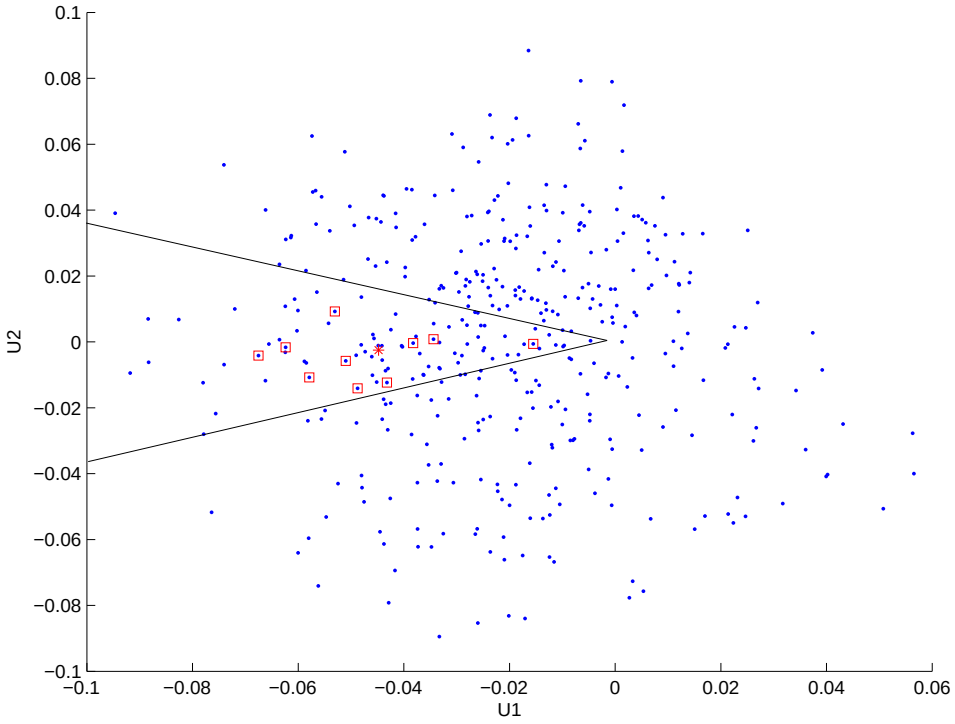

Figure 7: Same plot showing low false positive rate using a cone aimed at the target as a selector.

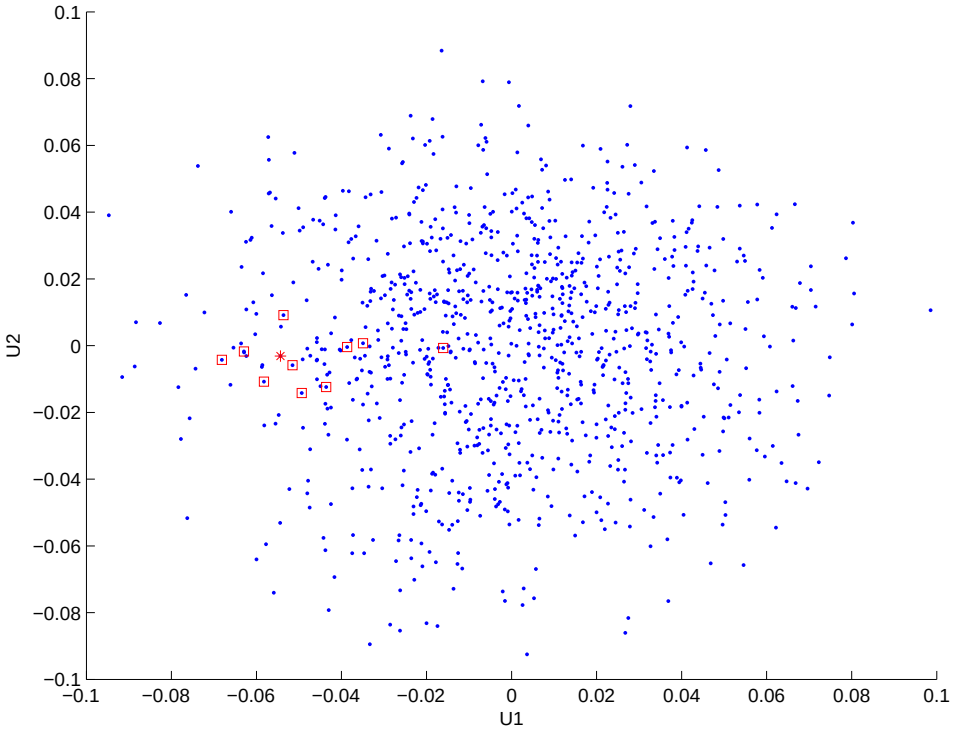

Figure 8: Same dataset as for Experiment 3, with the target row scaled by 1.2.

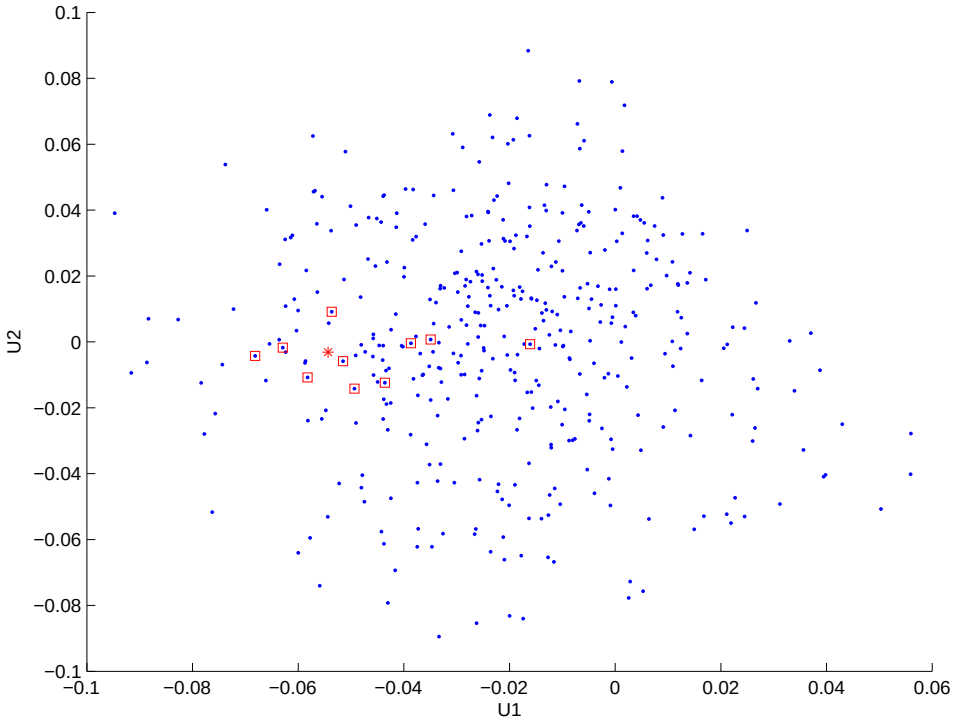

Figure 9: Same plot showing only the 422 individuals correlated with the target.

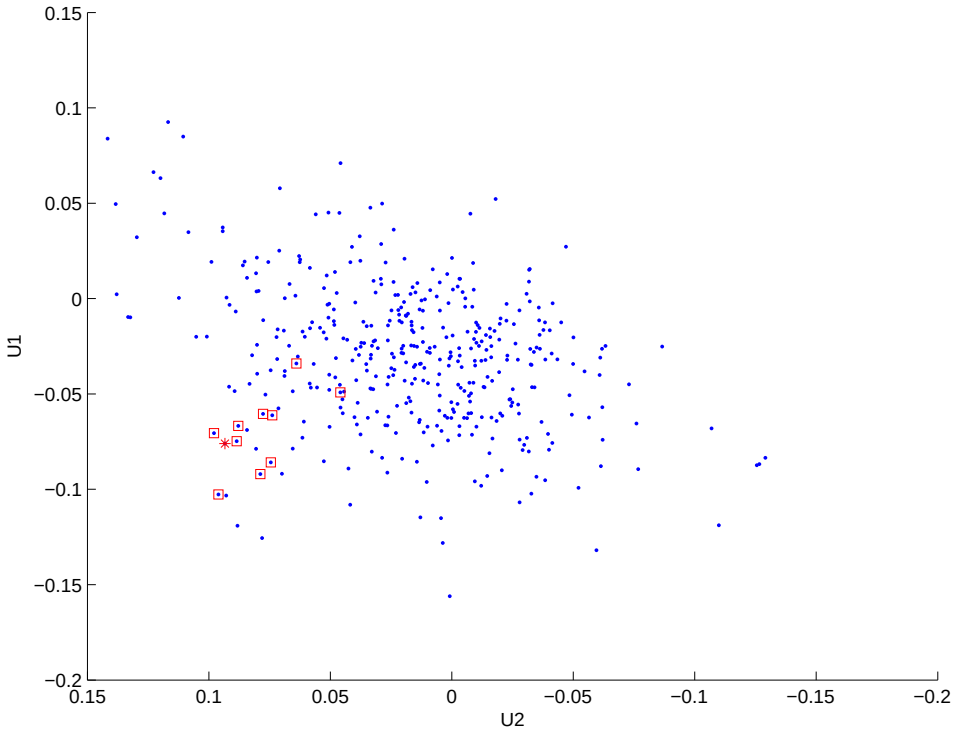

Figure 10: Repeated SVD using only the 422 individuals correlated with the target in the first round.

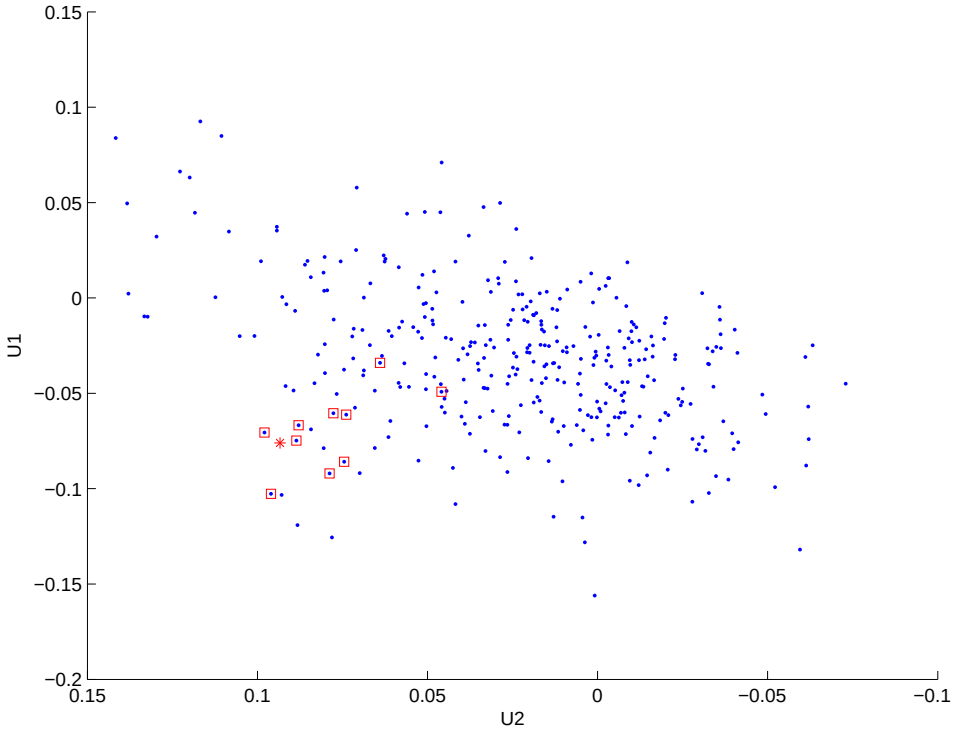

Figure 11: Same plot showing only the 362 individuals still correlated with the target.

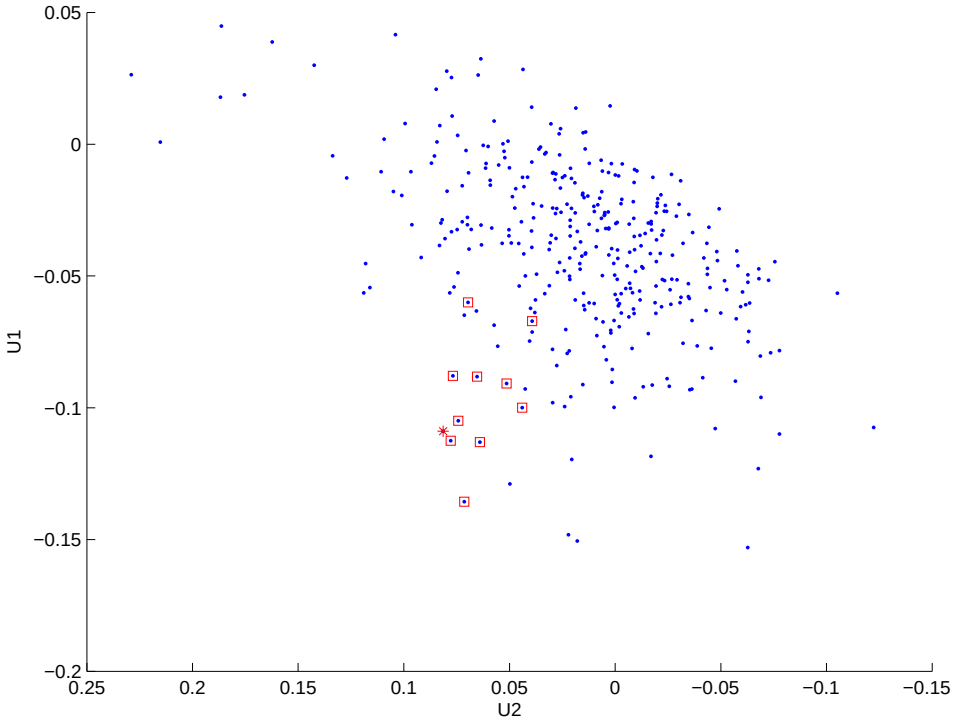

Figure 12: Repeated SVD using only the 362 individuals correlated with the target.

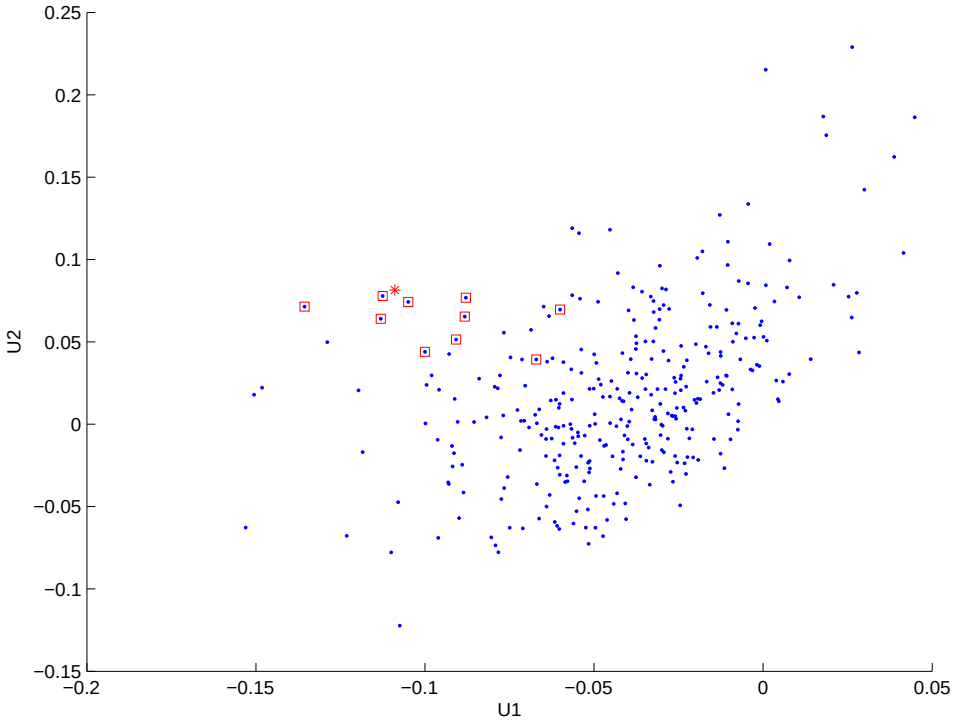

Figure 13: Same plot showing only the 337 individuals still correlated with the target.

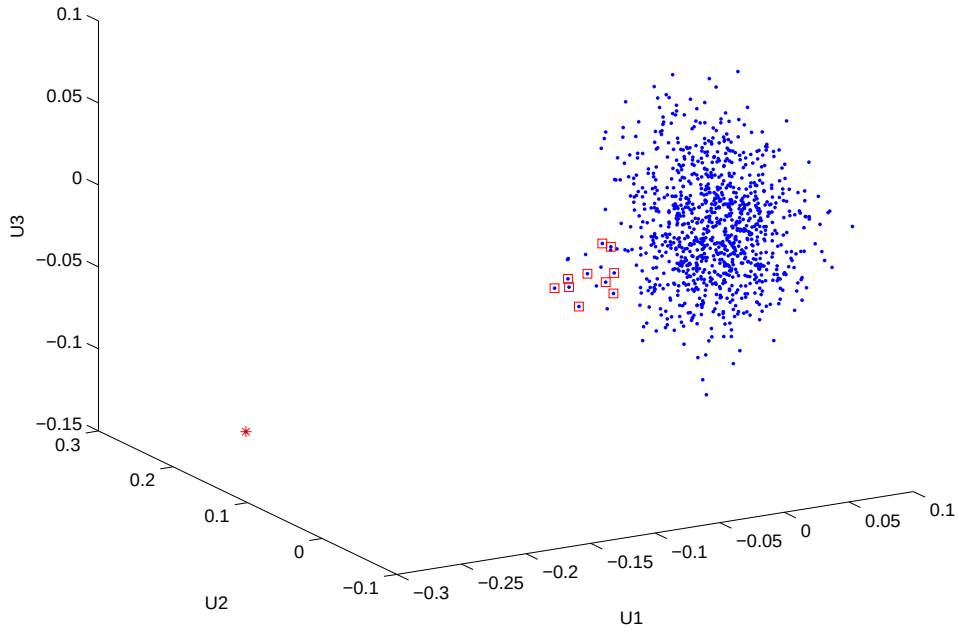

Figure 14: 10 normal clusters with variance 1 with centers drawn from a normal distribution with mean 0 and variance 1; terrorist cluster normally distributed around a randomly chosen individual with variance 1; weight of 4 on the target.

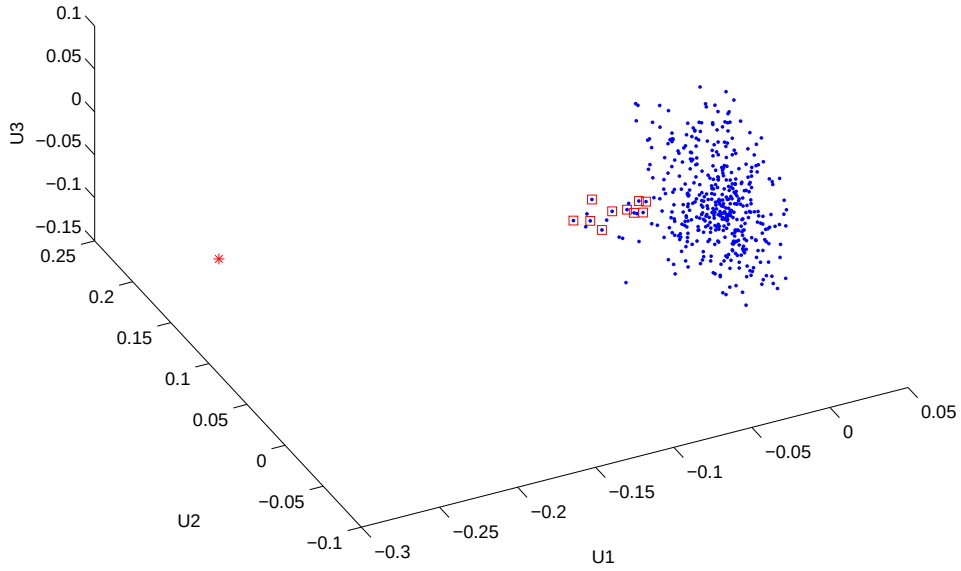

Figure 15: Same plot showing only the individuals correlated with the target.

normally distributed with variance 1 are generated around randomly chosen points in the second level. One of these 'third' level clusters is chosen as the terrorist cluster and its center as the target.

Figure 16 shows that both projection and proximity find the terrorist cluster with reasonable accuracy. Figure 17 shows the plot of the points correlated with the target.

Experiment 7. In the dataset, the local environment of each of the second level cluster centers is the same and we can choose any of them as possible terrorist clusters. On the other hand, the local environment of all of the other points is quite different. Figure 18 shows the sizes of the sets of points correlated with a particular point, when that point is a second-level cluster center (a possible target) and when it is one of the other points.

Those points that are targets have neighborhoods that start out smaller and shrink more rapidly than the neighborhoods of points that are not targets. The difference between the two types of points is marked, even by the third round.

Experiment 8. In our experiments so far, the number of terrorists has been about 1% of the total number of individuals. This fraction is too large to be realistic, even if a substantial prescreening process is applied before this kind of data mining is used.

Figure 19 shows the three-dimensional plot of a dataset with 10000 rows, normally distributed around the origin with variance 1, with a 10-terrorist cluster normally distributed with variance 1 generated around one of the ordinary individuals. The terrorist cluster is now much more diffuse. However, note that the extremal point along the projection on the vector from the origin to the target is a terrorist, and several others project on this vector further from the origin than the target.

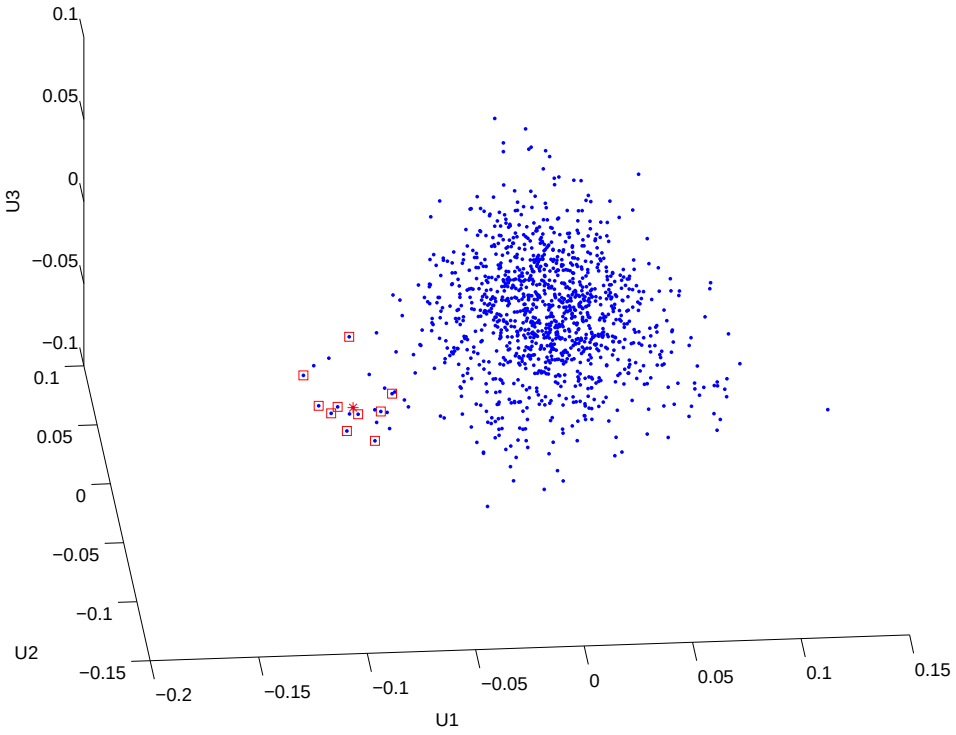

Figure 16: Three levels of clusters: first level of 100 cluster centers normally distributed around 0 with variance 1; second level of 10 cluster centers normally distributed with variance 1 around these; then 20 clusters of size 10 distributed around these. One second-level cluster center designated as the target.

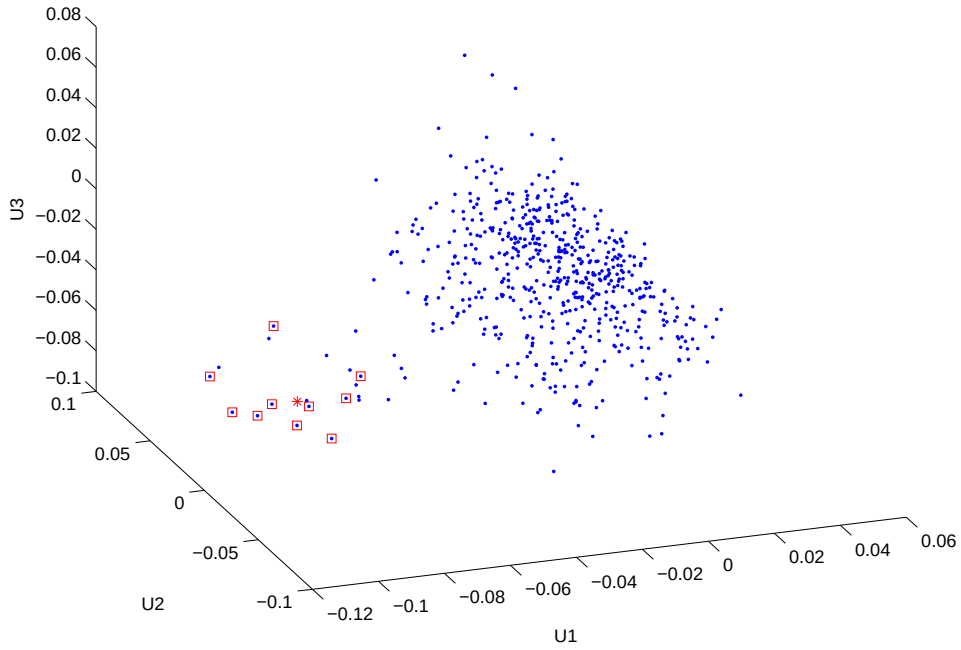

Figure 17: Same plot showing only those individuals correlated with the target.

After rnd	Size of sets correlated with a point that is a target			that is not a target			
1	145	419	199	831	370	586	416
2	20	27	47	513	90	194	150
3			20	461	48	86	78
4				400	42	56	65

Figure 18: Sizes of correlated sets after elimination of uncorrelated individuals. Initial size of all sets is 1200.

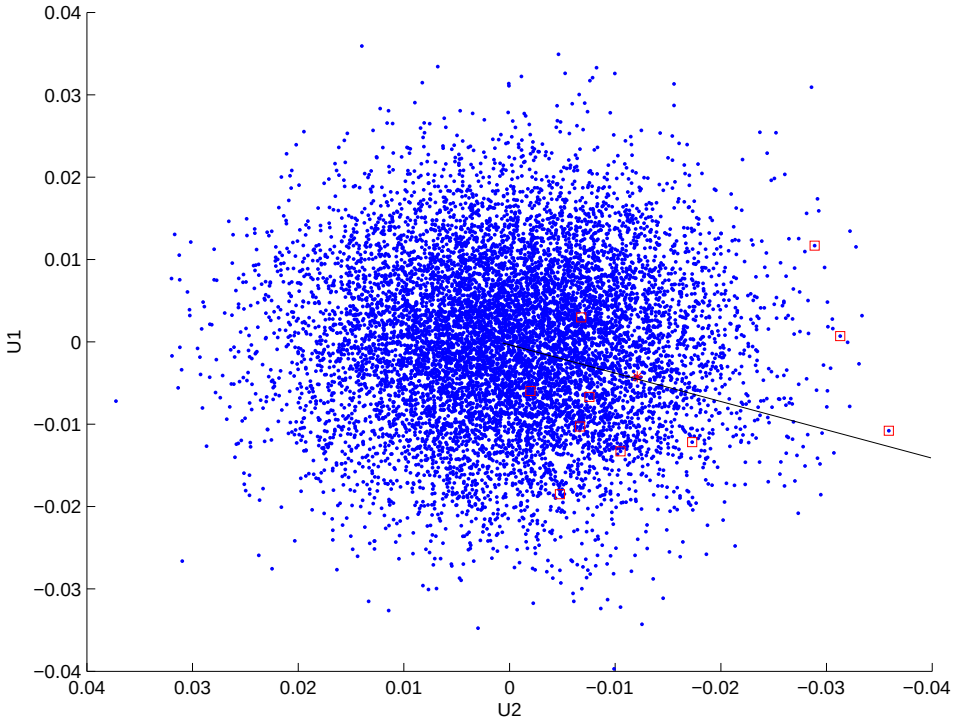

Figure 19: Plot of a 10-terrorist group and a dataset of 10000 ordinary individuals.

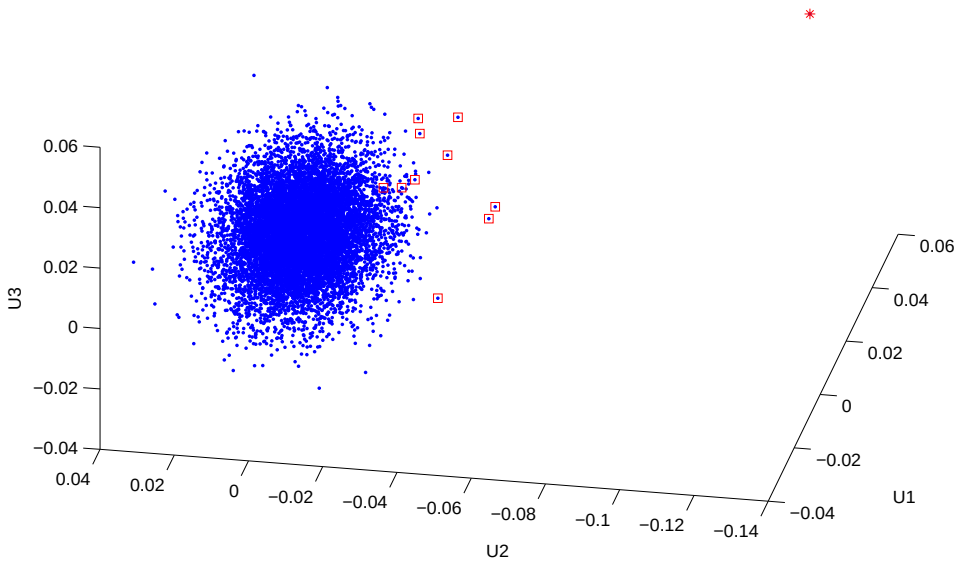

Figure 20: Same dataset with the target weight scaled by 4.

As Figure 20 shows, adding a weight of 4 to the target clearly selects the majority of the terrorist group, so even in large datasets SVD is effective.

Experiment 9. Fact 1 suggests that sparseness in datasets will not cause difficulties for SVD. This illustrates one of the strong properties of SVD – it is capable of detecting correlation even between individuals who have no (non-zero values of) attributes in common, via higher-order correlations.

Figure 21 shows a plot of a dataset similar to that of Experiment 1, but with 80% of the values set to zero. Although many of the terrorist cluster are not close to the target, several members still are. A dataset like this represents a situation where underlying attributes exist but are missing for some reason.

Another kind of sparse dataset is one in which there are no meaningful values for the zero entries. Figure 22 shows the plot of such a dataset. Here all 1010 rows are generated using a normal distribution with mean zero and variance 1, and a random row among the first 1000 is selected as the target. The rows of the terrorist cluster are then correlated with the target in the following way: if a target attribute has a non-zero value then, with 70% probability, the terrorist row is changed to a value drawn from a normal distribution whose mean is the value of the target attribute and whose variance is 1; otherwise the value is left unchanged. The correlation of the terrorist cluster with the target is plainly visible. Figure 23 shows the same plot with the points uncorrelated with the target removed.

Experiment 10. We now show that similar effects hold for distributions other than the normal distribution. The Poisson distribution with mean 1 generates many values close to 1, with

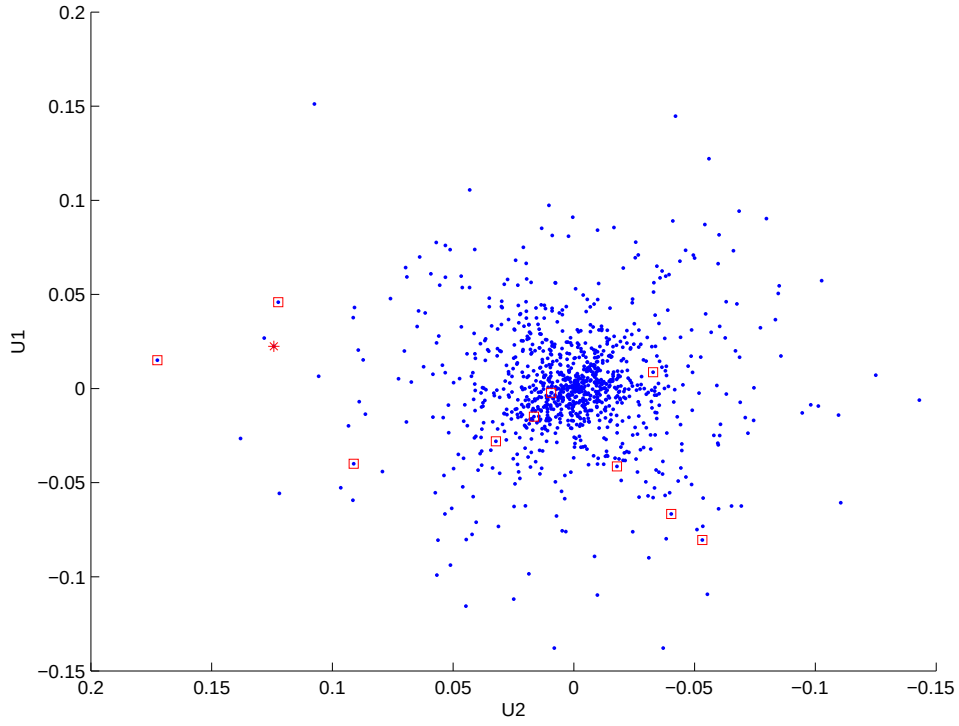

Figure 21: A sparse dataset generated by setting 80% of the values in a dense dataset from Experiment 1 to zero.

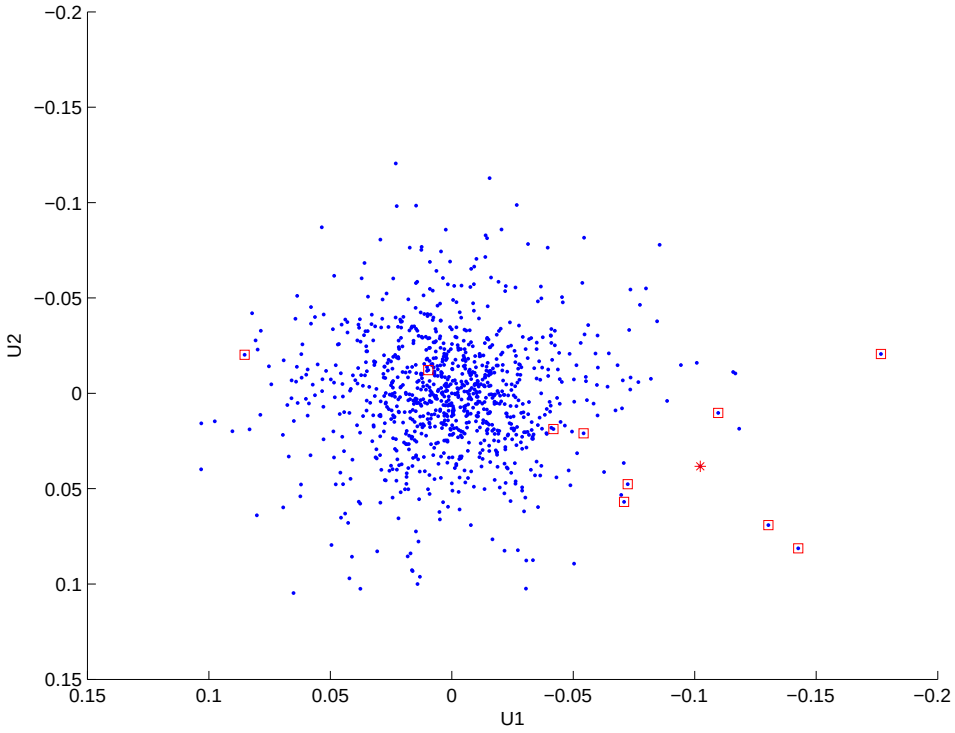

Figure 22: A sparse dataset generated directly.

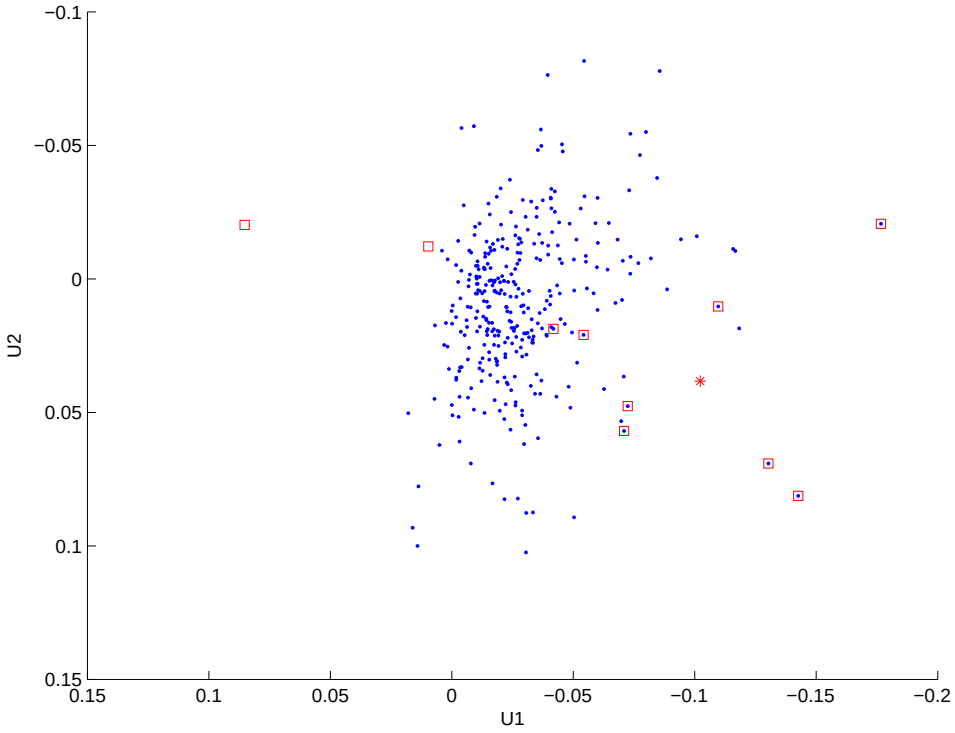

Figure 23: The same plot without the points uncorrelated with the target.

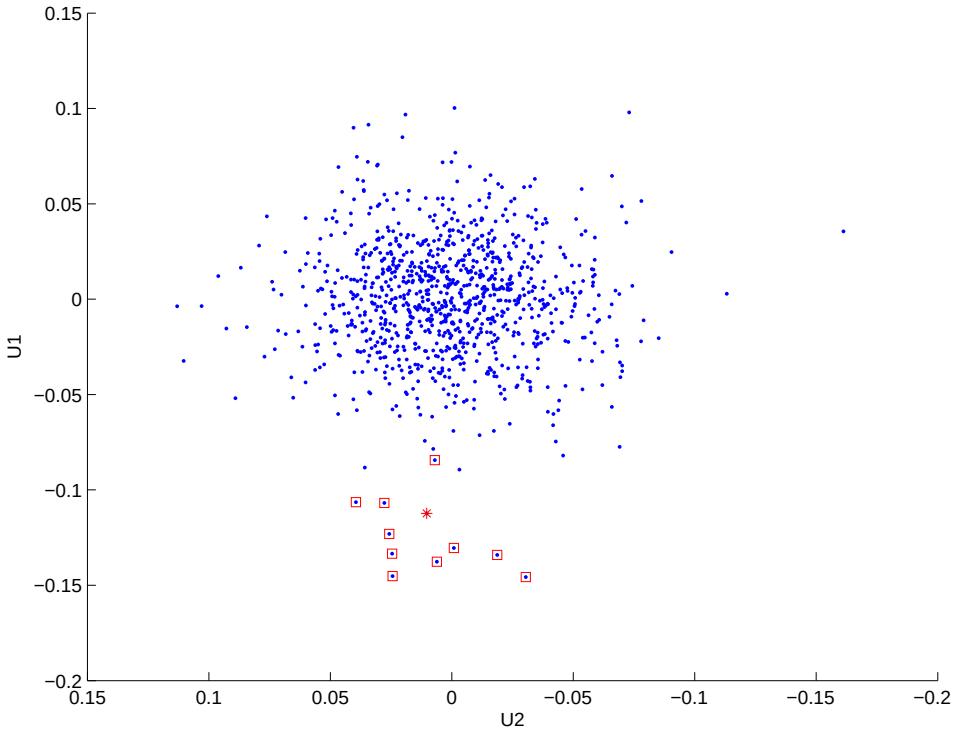

Figure 24: Ordinary individuals generated from a Poisson distribution, terrorist cluster normally distributed around a randomly chosen target.

the frequency decreasing rapidly with magnitude. We build a dataset of a 1000 rows from this distribution, subtracting λ to make the values approximately zero mean.

Figure 24 shows the results when the terrorist cluster is generated using a normal distribution with variance 1 around a randomly chosen row. Figure 25 shows the results when the terrorist cluster is also generated by the same Poisson distribution around the target (i.e the mean of the terrorist distribution is roughly the target). Figure 26 shows the same plot with only the correlated points shown.

Some settings have data that is binary in nature; each person did, or did not do some action, or does or does not have some particular attribute. The Poisson distributions above are quite close to such datasets because we used a mean of 1 and the results are similar.

7 Related Work

Techniques used for detecting outlying objects or outlying processes, for example Independent Component Analysis (ICA) [8], and 1-Class Classification [15, 16] seem less likely to provide good solutions for terrorism detection, although they may be effective when the *values* are completely beyond the control of individuals.

Singular value decomposition has been known since 1873, and used extensively in computing since Golub discovered an effective algorithm to compute it [7]. There is a vast literature on its use for dimensionality reduction; and it has been used for information retrieval where proximity

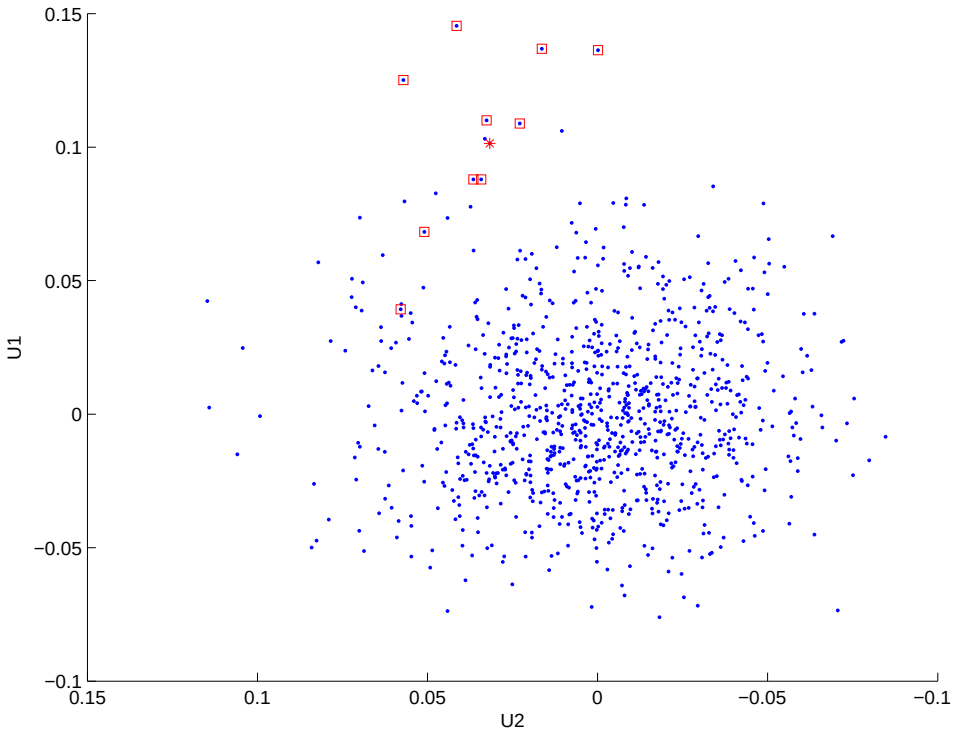

Figure 25: Ordinary individuals generated from a Poisson distribution, terrorist cluster Poisson distributed around a randomly chosen target.

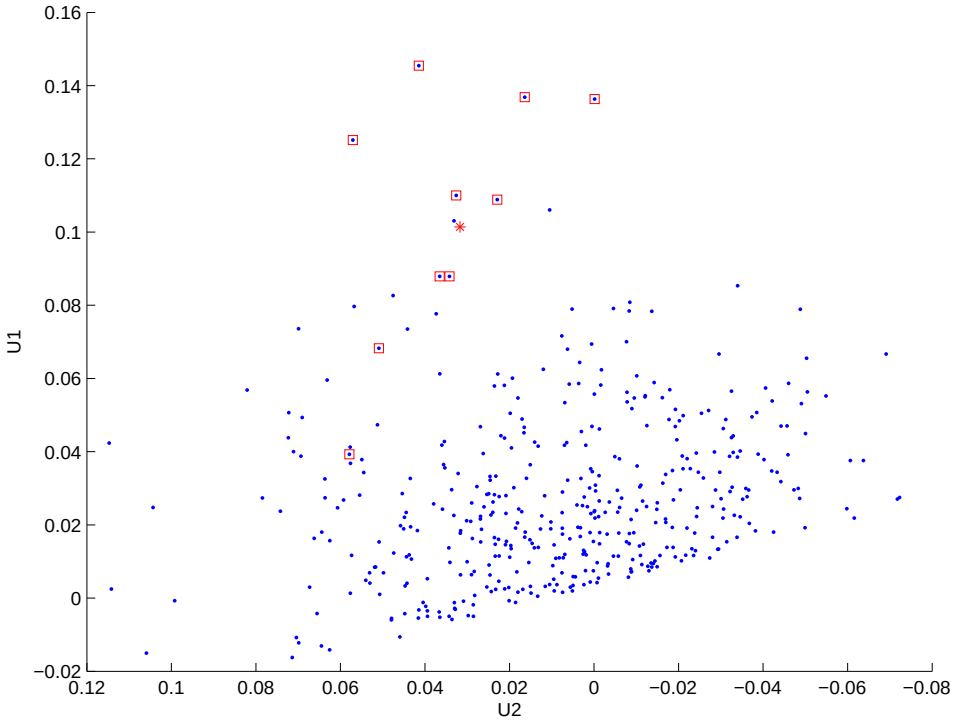

Figure 26: The same plot without the points uncorrelated with the target.

(in the sense of cosine similarity) serves as a proxy for correlation [3].

Social Network Analysis [6] study the interactions between individuals and derives global properties from the structure of the resulting graphs. There are two serious drawbacks to the use of SNA techniques for terrorism detection:

- Networks are built by adding pairwise links between two individuals, and this will not scale well since the number of potential links is quadratic in the number of individuals. In practice, SNA seems to have been used when a particular individual threat has been identified as a tool to discover his or her collaborators. In other words, SNA has a bootstrap problem (but may be useful once the kind of prescreening we suggest here has been applied).

- Links are made between individuals as the result of some interaction between them, rather than because of some correlation between them. In other words, SNA may discover two collaborators who *meet* at a target site, but will not discover them simply because they both *visit* the target site.

Link or traffic analysis has similar drawbacks: it can be useful once at least one member of a terrorist group has been identified; but it has the same limitation of only detecting direct relationships between two individuals, rather than their correlated actions. Traffic analysis has been used to detect unusually strong patterns of interaction, but only on the basis of a handful of attributes.

The paper [13] describes experiments using Inductive Logic Programming on relational datasets recording nuclear smuggling and contract killing. This work could presumably be generalized to counterterrorism.

8 Conclusion

We have shown that SVD is able to detect small correlated clusters, representing terrorists, against a variety of backgrounds representing degrees of innocent correlation. Qualitatively, in every case there exists a mechanism that identifies at least one (usually more) of the terrorist cluster based on proximity to the target, either directly in a low-dimensional space or by projection along a vector derived from the target. The number of false positives induced by these procedures is not trivial, but it is arguably reasonable. Our results do not suggest an optimal strategy for applying SVD for terrorist detection – rather they suggest a number of effective techniques. More experience will be required to determine how best to combine these techniques.

Many questions remain: are the generated datasets used for these experiments reasonable analogues of real-world datasets, can the *ad hoc* detection procedures used here be codified and automated, and does the performance remain acceptable as datasets become larger, perhaps much larger?

References

[1] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 2001.

[2] Y. Azar, A. Fiat, A.R. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *ACM Symposium on Theory of Computing*, pages 619–626, 2001.

[3] M.W. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[4] S. Chakrabarti and A. Strauss. Carnival booth: An algorithm for defeating the computer-assisted passenger screening system. Course Paper, MIT 6.806: Law and Ethics on the Electronic Frontier, `http://www.swiss.ai.mit.edu/6805/student-papers/spring02-papers/caps.htm`, 2002.

[5] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

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

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

[8] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[9] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. In *Proceedings of the 41st Foundations of Computer Science (FOCS '00)*, page 367, 2000.

[10] A. Kontostathis and W.M. Pottenger. Detecting patterns in the LSI term-term matrix. Technical Report LU-CSE-02-010, Department of Computer Science and Engineering, Lehigh University, 2002.

[11] A. Kontostathis and W.M. Pottenger. Improving retrieval performance with positive and negative equivalence classes of terms. Technical Report LU-CSE-02-009, Department of Computer Science and Engineering, Lehigh University, 2002.

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

[13] R.J. Mooney, P. Melville, L.R. Tang, J. Shavlik, I de Castro Dutra, D. Page, and V.S. Costa. Relational data mining with Inductive Logic Programming for link discovery. In *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*, November 2002.

[14] M. Newman and D. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263:341–346, 1999.

[15] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

[16] D.M.J. Tax. *One Class Classification*. PhD thesis, Technical University Delft, 2000.