Social Network Analysis via Matrix Decompositions: al Qaeda

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

August 2004

Abstract

Social network analysis investigates the structure of human groups using pairwise links among their members. We show how matrix decompositions can be used to extend the standard repertoire of social network and link analysis tools to allow, for example, the inclusion of other information about individuals, and higher-order information about the relationships among them. We show how these extensions can be applied by analyzing the structure of al Qaeda and its related terrorist organizations. Much of the information about, for example, relative importance of al Qaeda members can be extracted from simple relational information.

1 Introduction

Social network analysis explores the structure of groups in human society by modelling individuals, places, and objects as nodes of a graph, and adding links between nodes to represent relations among them. For example, important subgroups can be identified as cliques in the graph; individuals in particular positions of control can be identified by their centrality (using a number of measures); and substructures of particular interest (for example, communication chains) can be discovered [9, 12, 19].

Social network analysis has been applied to both terrorist and criminal networks. For example, Baker and Faulkner [2] relate location in a criminal network to length of eventual sentence; Sageman uses SNA to validate his division of al Qaeda members into four classes: leadership, core Arab, South-East Asian, and Maghreb [16].

1.1 The challenge

Transnational terrorism groups present a new problem for the countries against whom their actions are directed, usually characterized as *asymmetric* or *4th generation warfare* [17]. Unlike the case of military conflicts between nations, such terrorist groups have a membership that is hard to define, few visible fixed targets, the ability to operate across borders relatively freely, and independent sources of funding, removing indirect ways to pressure them via sponsors. The resources that must be expended by both sides differ by several orders of magnitude: around 25 men and expenditures estimated to be only ~\$500,000 were sufficient for al Qaeda to attack the World Trade Center. In contrast, U.S. spending in response is at least \$100 billion [3, 5].

Counterterrorism efforts face corresponding difficulties in attempting to detect and preempt attacks. A metaphor suggested for al Qaeda is that it is a venture capitalist for terror [6]; proposals for attacks are brought to the leadership and those that are approved receive support in the form of training and financing. This means that attackers may have only minimal contact with the main part of the organization until quite close to the time an attack is mounted. This suggests that *every* contact with known al Qaeda members, no matter how fleeting, needs to be treated as significant. The transnational nature of al Qaeda also makes it clear that a profile of a "typical" al Qaeda member does not exist – and there is some evidence that the group is trying to recruit members who appear even less like a hypothetical Salafist terrorist.

Al Qaeda is only the most visible of a number of movements whose grievance extends beyond a single geographical region, whose aim is not simple visibility for their cause, and who have discovered that nation-states are vulnerable to asymmetric warfare attacks. Counterterrorism technologies will, sadly, be of use even when al Qaeda has been defeated.

1.2 Link analysis technology

The techniques of social network analysis have some limitations as tools to explore the graphs that model social groups. First, it is not straightforward to extract 'higher-order' information, that is information that is associated not with a connected pair of objects but with a larger set. Second, it is not easy to introduce and use information that is not naturally associated with edges of the graph, for example demographic information. Third, social network analysis typically depends on the precise connection structure of the graph, so that small changes in the graph may produce large changes in its properties. This is a problem because information about terrorist groups necessarily misses some data, and it is also likely that some data is wrong.

In this chapter, we show to how to use the machinery of *matrix decompositions* to extract more information from a graph that models a social group. We use three such decompositions:

1. Singular value decomposition (SVD). Although SVD is commonly used for dimension reduction, we use it both as a graph partitioning tool (an approach known as *spectral graph partitioning*) and as a way to detect the most anomalous, and hence most interesting, nodes in a graph. SVD transforms data based on correlation, and so can extract structure that is incomplete; it does not require prespecification of the structures of interest.

2. Semidiscrete decomposition (SDD). SDD partitions data into subsets with similar attribute values, in the process creating an unsupervised hierarchical classification tree. Hence it is a clustering tool that works in a different way to both SVD and metric-based clusterers such as k-means that are, in any case, unreliable in high dimension.

3. Independent component analysis (ICA). ICA partitions data into the least Gaussian components possible. In a graph context, this amounts to selecting components that are the most like cliques as possible.

These techniques largely avoid the weaknesses of conventional link analysis: they include higher-order correlation information, they can use extra information associated with both edges and nodes, and they are robust in the presence of missing values (because these are often implied indirectly by other values) and wrong values (because correlation rather than equality is the basic comparator).

We illustrate the application of these methods on a dataset containing information about 366 members of al Qaeda (current as of the beginning of 2004). The dataset contains typical relationship information, such as members who are related, who are friends, or who have encountered one another since joining the organization. However, it also contains demographic information (age, countries of origin and joining the group, education and marital status, etc.) and we are able to include this information in our analysis.

2 Matrix decompositions

We begin with a dataset containing information about n objects (people in our context), with m attributes about each one. Some of these attributes might be categorical (they come from a fixed set of choices such as marital status), numeric (such as age), or representative of connections among the objects. In this last case, we will use an adjacency representation: the friendships among the 366 people will be represented by 366 different attributes, with a 0 value at position ij indicating that persons i and j are not friends and a 1 value indicating that they are. Of course, this region of the dataset will typically be sparse (i.e. mostly 0s).

Such a dataset is naturally viewed as a matrix, A, with n rows and m columns. A *matrix decomposition* expresses the matrix A as a product of other matrices in a way that reveals A's structure. Hence a typical matrix decomposition can be expressed as a matrix equation:

$$A = C\,S\,F$$

where C is $n \times m$, S is an $m \times m$ diagonal matrix (off-diagonal entries are all 0), and F is $m \times m$. Typically, the sizes of the matrices on the right hand side are restricted to force the decomposition to represent the data more compactly, so that the decomposition, truncated to k, becomes:

$$A \approx C_{n\times k}\,S_{k\times k}\,F_{k\times m}$$

Matrix decompositions are related to Expectation-Maximisation, where each different decomposition imposes extra conditions on the way in which the partitioning is done.

There are several different ways to interpret a matrix decomposition, and each sheds different light on the underlying data. In the *factor* interpretation, the rows of F are interpreted as underlying or latent factors and the entries of C as ways to mix these factors to produce the observed data. The diagonal entries of S are *weights*, whose magnitude gives the relative importance of each factor. This view is commonplace and productive in the social sciences as *factor analysis*. Sometimes the factors can be regarded as axes in some space, in which case the entries of C are coordinates of points in this space.

The (outer) product of the ith column of C, the ith entry on the diagonal of S, and the i row of F is a matrix of the same shape as A, and in fact A can be expressed as the sum of all of these matrices. This allows a *layer* interpretation of the decomposition. A is obtained by sandwiching all of the outer-product matrices together, and so each of them can be regarded as making some contribution to all of the values of A. Once again, the magnitude of the diagonal element provides information about how important each layer is to the total dataset.

2.1 Singular Value Decomposition

SVD transforms data in a way that converts correlation to proximity [4, 18]. In the decomposition:

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

the matrices U and V are orthogonal (the superscript dash indicates transposition), and the diagonal entries of S, called the singular values, are non-increasing.

Because V is orthogonal, a geometric interpretation is natural. The rows of U can be understood as the coordinates of points corresponding to the objects. The axes of the transformed space are such that the greatest variation in the original data lies along the direction of the first axis (the first row of V), the greatest remaining variation along the second axis, and so on. Hence truncating at some k gives a representation in a lower-dimensional space that captures the correlative structure as accurately as possible.

It is conventional to scale the data so that the relative magnitudes of each attribute are the same, and also to subtract the mean from each column of attributes. If this latter is not done, the first singular vector represents the average magnitude of the data and is typically of less interest. However, when the data represents, for example, the adjacency matrix of a graph, it may not be sensible to normalize the entries.

The SVD is completely symmetric with respect to rows and columns of the original matrix, so that all of the analysis that can be done for objects can trivially be repeated for the attributes as well.

SVD can be used in a number of ways to analyze a dataset:

- Dimensionality reduction. This is the most common use of SVD in data analysis since it provides a way to reduce high-dimensional data (i.e. with many attributes) to lower dimension, losing as little information as possible in the process. When the original data contains noise, this dimensionality reduction can be regarded as denoising as well.

  One of the benefits of dimensionality reduction is that choosing $k = 2$ or 3 allows the rows of U to be plotted. This often makes it possible to understand at least the most significant structure of a dataset by visual inspection.

- Clustering. In a transformed and truncated space, the relationships among the points have been clarified and consequently clustering might be expected to work more effectively. There are, broadly, two approaches, although each contains many competing variants. The first is to use metric-based clustering, for example k-means, in the new space. The second is to use the properties of the SVD directly in an approach called spectral clustering [10]. For example, those points which lie in the cone around the first axis (those whose dot product with the axis is less than 1/2) are placed in one cluster; those with the same property with respect to the second axis in the second cluster, and so on. This produces k clusters, of which the last one is the 'everything else' cluster. In some settings, it is obviously correct to include in each cluster the points that lie within the cone corresponding to the negative direction of each axis as well. These points are negatively correlated with the others with which they are being lumped, but they are correlated nevertheless. In low dimensions, no formal clustering algorithm is required because the clusters can usually be seen.

  When the matrix represents the adjacency matrix of a graph, the clustering produced by SVD is often similar to the clique structure of the graph.

- Ranking objects by their interestingness. We have already explained that each row of U can be identified with a point in a k-dimensional space. Suppose that an arrow is drawn from the origin of the space to each of these points. Then the angles between these vectors reveal the correlation among the points. Two points that are strongly positively correlated will have vectors that are close together. Their dot products, which correspond to the cosine of the angle between the vectors, will be large and positive. Two points that are strongly negatively correlated will point in almost opposite directions, and will have a dot product that is large and negative. Two points that are uncorrelated should have a dot product that is close to zero, and it here that a problem arises. One way in which such a dot product can arise is that the two vectors are almost at right angles to each other. However, typically the number of available dimensions (k) is much smaller than the number of uncorrelated points (which could be n). There is another way in which the dot product can be close to zero and that is that the point itself is close to the origin. Hence points that are uncorrelated with most of the other points will tend to be placed near the origin. For similar reasons, a point that is correlated with almost all of the other points will also tend to be placed near the origin.

Hence in the transformed space, points that are located far from the origin correspond to objects that are interesting in the sense that their correlations with the other objects is unusual. Conversely, points that are close to the origin correspond to objects that are less interesting, either because they are randomly correlated with other objects, or correlated similarly with all of them. Ranking the objects in order of the distance of their points from the origin allows the most interesting objects to be selected.

Because SVD is symmetric with respect to objects and attributes, exactly the same idea can be used to discover the relative interestingness of the attributes.

2.2 Semidiscrete Decomposition

SDD [11, 15] decomposes a matrix A as:

$$A = X\, D\, Y$$

where the entries of X and Y are from $\{-1, 0, +1\}$ and D is a diagonal matrix with non-increasing entries (a variant of the original SDD as described in [13]).

The natural interpretation of SDD is the layered one based on the outer product matrices. The product of the ith column of X and the ith row of Y is a matrix which contains rectilinearly aligned patterns of -1s and $+1$s against a background of 0s. The non-zero values can be regarded as a stencil of locations within A where a set of values of similar magnitude (d_i) can be found. The locations where there is a $+1$ correspond to positive values of this magnitude and those where there is a -1 correspond to negative values of this magnitude.

Hence, whereas SVD analyzes the data in a geometric space, SDD analyzes the data within the matrix itself, decomposing it into sets of hills and valleys, such that the sum of all of the sets recreates the original data.

Furthermore, the values in the X matrix provide an unsupervised hierarchical classification of the objects. At the top level, those objects whose entry in the first column of X are $+1$ are in one branch, while those whose entries are -1 are in an opposite branch. Those objects whose entries are 0 are in yet a third branch, so that the classification tree is ternary. The tree is hierarchical because the clusters with the largest value of d_i appear first.

Although SDD was originally developed as a storage-efficient analogue of SVD, there is no necessary link between the classifications each produces. When the data naturally clusters into many small, well-separated clusters, SDD and SVD tend to agree. It also often happens that the

Short name	Year joined the jihad
Full name	Age joined the jihad
Date of birth	Place joined the jihad
Place of birth	Country joined the jihad
Youth national status	Acquaintance links
Family socioeconomic status	Friend links
Religious background	Nuclear family links
Educational achievement	Relative links
Type of education	Religious leader
Occupation	Ties not in sample
Marital status	Role in organization
Children	Operation(s) involved
Social background	Fate
	Links after joining

Figure 1: Dataset attributes.

top-level classification from SDD is aligned with the first axis of SVD, so that the $+1$ points are at one extremity and the -1 points at the other – but this does not necessarily happen.

2.3 Independent Component Analysis

ICA [1, 7, 8, 14] decomposes the data matrix, A, into components that are as statistically independent as possible (in contrast to SVD which decomposes the data into components that are *linearly independent*). We use the FastICA algorithm for convenience.

The ICA of a matrix A is:

$$A = W\ H$$

(note that there is no 'weight' matrix in this case, and hence no natural ordering on the components). H represents the statistically independent factors and W the way in which these factors must be mixed to recreate A.

3 al Qaeda Dataset

We will illustrate the power of these matrix decomposition techniques by using them to investigate the structures and relationships within al Qaeda, to the extent that they are publicly known. We use a dataset collected by Marc Sageman from a wide variety of public sources. The dataset contains information about 366 members of al Qaeda as of the beginning of 2004. The available attributes are shown in the table in Figure 1.

Many of these attributes are demographic in nature, but several describe the links among al Qaeda members under various categories. Of course, there are many missing values because not all information is publicly available.

Name	BL centrality	Name	BL centrality
bin Laden	298	Jarrah	234
Zawahiri	240	Shehhi	235
Banshiri	226	Mihdhar	220
M Atef	254	Hada	227
Sheikh Omar	222	Harithi	227
Islambuli	230	Ayiri	232
Zubaydah	260	Aktas	222
Makkawi	242	Sungkar	229
Hawsawi	227	Hambali	253
Taha	230	Faruq	233
KSM	250	Ramda	251
Zarqawi	221	Melouk	220
Qatada	221	Doha	225
Hage	221	Trabelsi	244
Khadr	222	Moussaoui	235
Ghayth	224	Bahaiah	229
Khallad	241	Khabab	228
Shah	232	Khalifah	227
Atta	246	Tabarak	222
Shibh	260		

Figure 2: Al Qaeda members with high Bavelas-Leavitt centrality.

We also use a subset of dataset, a link or adjacency matrix containing all of the links among members, whether as family, relatives, friends, or members of the group. The graph of these relationships has 366 nodes (of course) and 2171 edges. The maximum degree of the graph is 44 (but of course this number is probably higher in practise), and the mean degree is 6.44. (This value for the mean degree is interesting because it matches the rule of thumb that members of a group much have connections to about 6 others if they are to remain in the group.) The diameter of the graph is 11.

Figure 2 gives the Bavelas-Leavitt centrality values that exceed 220. For each node, this measure is the ratio of the sum of all of the shortest paths to and from that node to the sum of all of the shortest paths in the entire dataset. Accordingly, it measures how close the node is to the center of the graph of links in some notional space.

Many of the members with high scores are the leadership of al Qaeda as expected. However, there are several surprises: Hada, Harithi, Ayiri, Aktas, Faruq, Ramda, Melouk, Trabelsi and Bahaiah. Examination of the data suggests that these members get such high centrality scores because they have links to Osama bin Laden and several others of the top leadership. In the absence of other knowledge, this complicates the use of a centrality measure as an analysis device because it does not distinguish well between the important leadership and those with little importance but who are directly connected to the leadership. (Of course, this is further complicated by the fact that such people may be hangers-on, but may also be *eminences grises*.)

The University of Arizona group have analyzed this dataset and used multidimensional scaling to produce a picture of the group's connectivity (Jie Xu, personal communication, 2004). This shows that the dataset is naturally clustered into 13 almost-cliques, with about 60 members not allocated to a single clique.

A graph of the links within al Qaeda is maintained by Intelcenter and can be viewed on their web site (www.intelcenter.com/linkanalysis.html). While the graph is compendious, it is hard to extract actionable information from it.

4 Analysis using matrix decompositions

4.1 Using the links between individuals

In this section we consider only the results of enhanced link analysis, that is we consider the graph of relationships among al Qaeda members. The base dataset is a 366×366 adjacency matrix for the graph that includes: acquaintances, family, friends, relations, and contacts after joining.

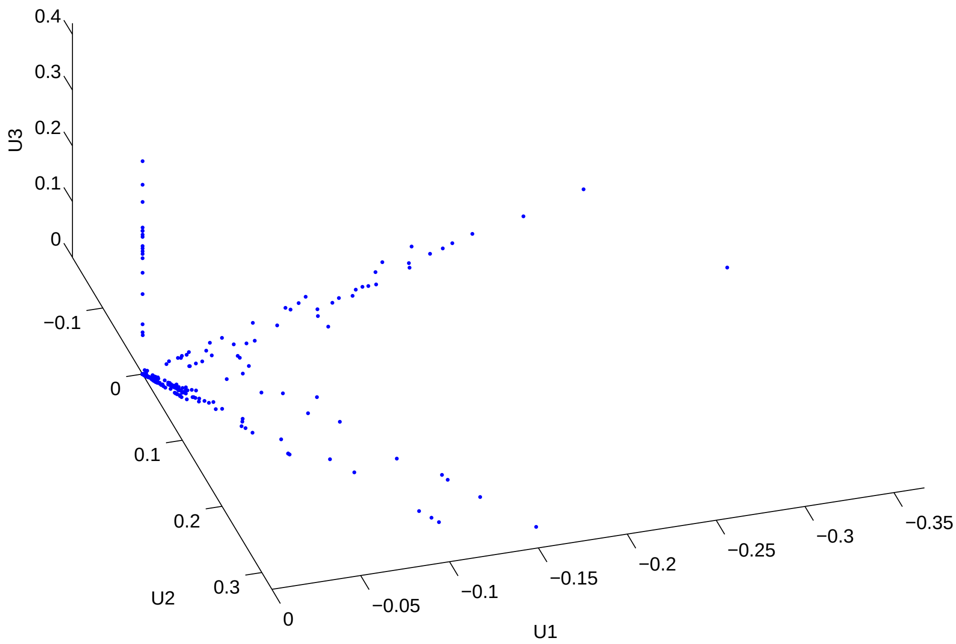

Figure 3: SVD plot of al Qaeda members using only relationship attributes.

Figure 3 shows a 3-dimensional (truncated) view of the relationships among al Qaeda members extracted from their links. The most obvious fact is that there is a clear division into three (perhaps four) clusters. This radial pattern is typical: those points at the extremities represent individuals with the most interesting connections to the rest of the group. Many members are either connected in limited ways, or little is known about them. All such members resemble each other, and so tend to be located close to the origin.

The structure is made clearer by adding name labels (we follow Sageman's usage) and removing points (and so individuals) that are located close to the origin. Figure 4 shows those points that are more than 1.5 times the median distance from the origin, while Figure 5 removes even more points. It now becomes possible to identify the visible structure.

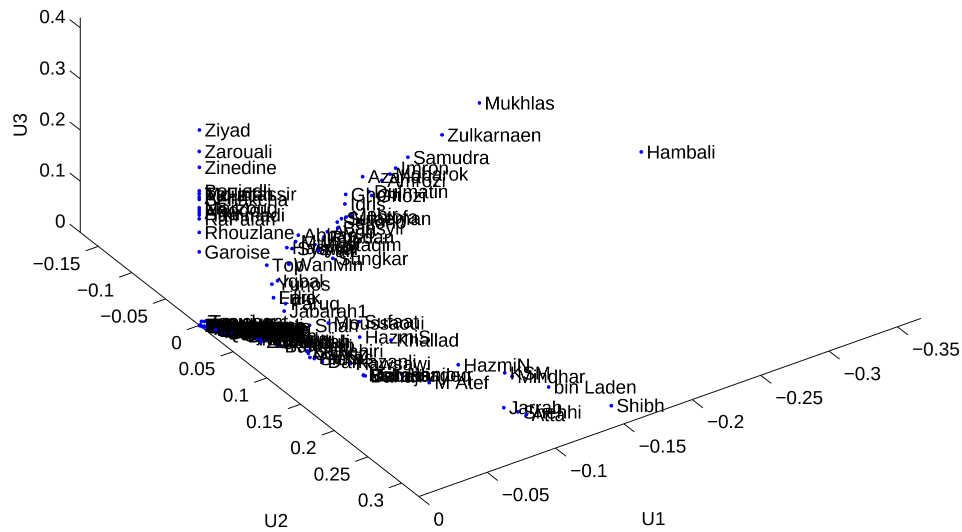

Figure 4: SVD plot of 143 interesting members (greater than 1.5 times the median distance from the origin) labelled with short identifiers.

There are three clusters in these figures: a group of Algerians arranged vertically in the figures; a group of South East Asian members stretching to the right; and a group of leaders and some core Arabs towards the front. It is clear from these figures that Hambali plays a pivotal connecting role between the SE Asian group and the leadership group; but further, the separation into two parallel lines of the leadership group is entirely due to whether or not they have a link to Hambali. The fact that Hambali is well-connected is obvious from the raw data – but it is not so obvious how integral these connections are to holding al Qaeda together. The strong presence of the Algerian cluster is slightly surprising; while these members have been active over a long period, they are not obviously the most important members of al Qaeda's European operations in the raw data.

Each of the clusters arranges the more important members farthest from the origin as expected. Notice that bin Laden is not the most extremal member of the leadership cluster – this appears to be partly due to good tradecraft (he is not directly involved in operations) and to relative inactivity over the past few years. Note that Figure 5 selects the highest profile al Qaeda members well.

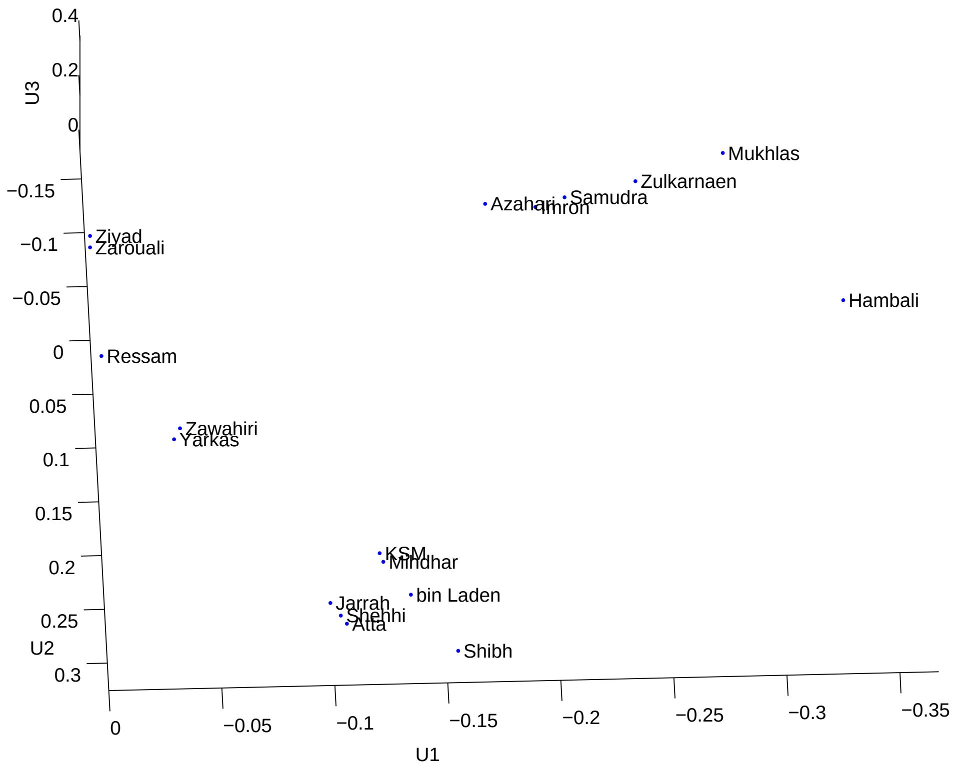

Figure 5: SVD plot of 18 interesting members (greater than 4 times the median distance from the origin) labelled with short identifiers.

Figure 6 shows the 143 interesting members, but using dimensions 4 to 6 of the SVD (in other words, relationships in less important dimensions). Here again there are 3 clusters, although they do not bear the same close relationship to attack teams (although the group to the right are in fact the September 2001 attackers and their support group). It is clear, once again, that the most important members of the group are placed far from the origin.

Figure 7 overlays the SVD plot in dimensions 1–3 with information about which cultural group each member comes from. This figure shows the strong, separated, groups from the Middle East and from South East Asia. Although the vertical group are different from everyone else, most of the Maghreb/Algerian/French members resemble core Arabs.

Figures 8 and 9 are the same plots, but with the color and shape labelling derived from the SDD classification of the points. The top-level division in Figure 8 is coded by color: on one side, the important leaders, core Arabs, and SE Asians (red), in the middle the majority of the members (green), and on the other side, the Algerians (blue). The subsequent two levels are indicated by the symbol shape as shown in the following table:

+1	+1	dot	0	+1	+	−1	+1	diamond
+1	0	circle	0	0	star	−1	0	triangle down
+1	−1	cross	0	−1	square	−1	−1	triangle up

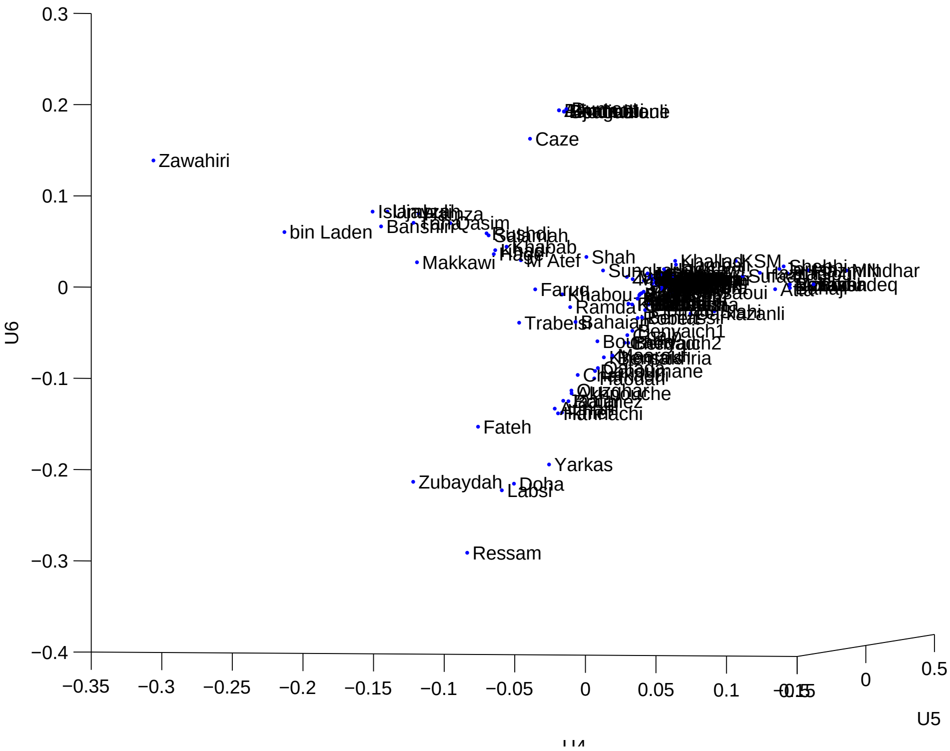

Figure 6: SVD plot in dimensions 4–6 of 143 interesting members (greater than 1.5 times the median distance from the origin) labelled with short identifiers.

Here the similarities are between the Algerian group and the SE Asians (indicated by diamonds). In all of the clusters, a few marginal members can be discerned, indicated by pluses. Although the SDD classification generally agrees with that of SVD, the benefit of the extra information is that it provides substructure: a better view of boundaries and more detail within clusters.

Independent component analysis is naturally interpreted in a layered way: each component describes some aspect of the dataset. In this case, ICA works as an effective clique detector. It finds small groups of individuals who are much more closely linked than usual.

Recall that we compute

$$A = W\ H$$

From this, we can compute the set of outer product matrices formed by multiplying the ith column of W with the ith row of H (giving a matrix with the same shape as A).

Figure 10 shows one example of such an outer product matrix. Because the rows of the original dataset tend to have been organized in rough groups, the clique of connected individuals discovered by ICA happens to be located almost contiguously in the dataset, but there is no necessary connection.

We extract the individuals associated with each outer product by applying a threshold function to each such matrix (in this case, 0.2 of the maximum value) to produce a 0-1 matrix. We then list

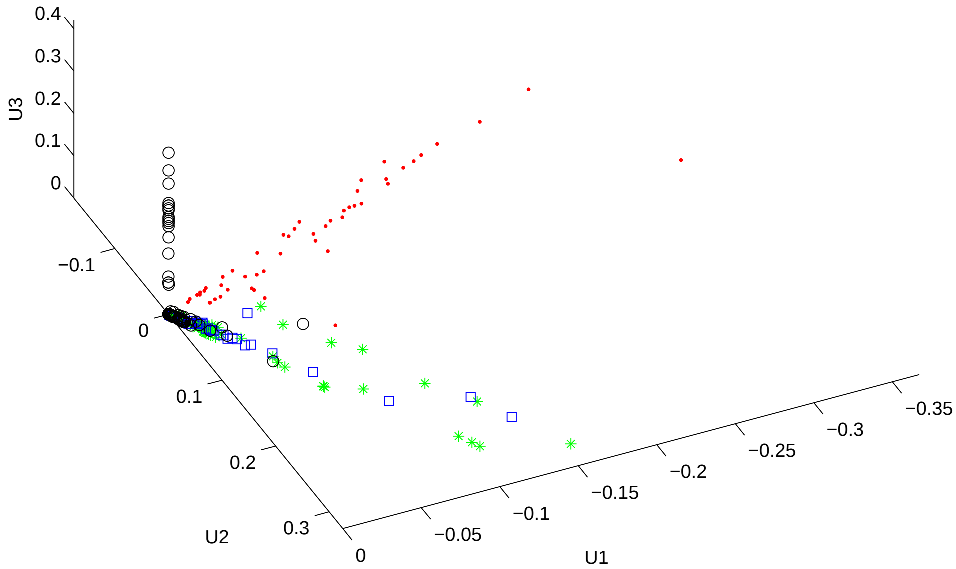

Figure 7: SVD plot showing cultural group membership (squares – leadership, stars – core Arabs, circles – Maghreb, dots – S.E. Asian).

those objects (individuals) with a 1 anywhere in their row.

Recall that ICA does not impose an importance ordering on components (at least not directly). Some clusters that arise from the link dataset are:

- Christophe Caze, Lionel Dumont, Rachid Souimdi, Saad el Aihar, Amar Djouina, Mouloud Bougelane, Hassan Zemiri, Hocine Bendaoui, Seddick Benbahlouli, Laifa Khabou, Fateh Kamal (Groupe Roubaix, France 1994).

- Rachid Ramda, Ali Touchent, Boulem Bensaid, Safe Bourada, Smain Ait Ali Belkacem, Mohamed Drici, Ali ben Fatoum, David Vallat, Khaled Kelkal, Karim Koussa, Adelkader Maameri, Abdelkader Bouhadjar, Nasserdine Slimani, Farid Melouk, Ahmed Zaoui (France 1995).

- Osama bin Laden, Mohammed Atef, Mustafa Ahmed al-Hawsawi, Khalid Sheikh Mohammed, Waleed Tawfiq bin Attash, Mohamadou Ould Slahi, Mamoun Darkazanli, Mohammad bin Nasser Belfas, Mounir al-Motassadeq, Abdal Ghani Mzoudi, Said Bahaji, Mohammed Atta, Ramzi bin al-Shibh, Ziad Jarrah, Marwan el-Shehhi, Zakarya Essabar, Hani Hanjour, Nawaf al-Hazmi, Khalid al-Mihdar, Saleem al-Hazmi, Fayez Ahmad el-Shehri, Ahmed al-Nami, Christian Ganczarski, Encep Nurjaman (Hambali) (largely organizers and participants of World Trade Center attack, 2001)

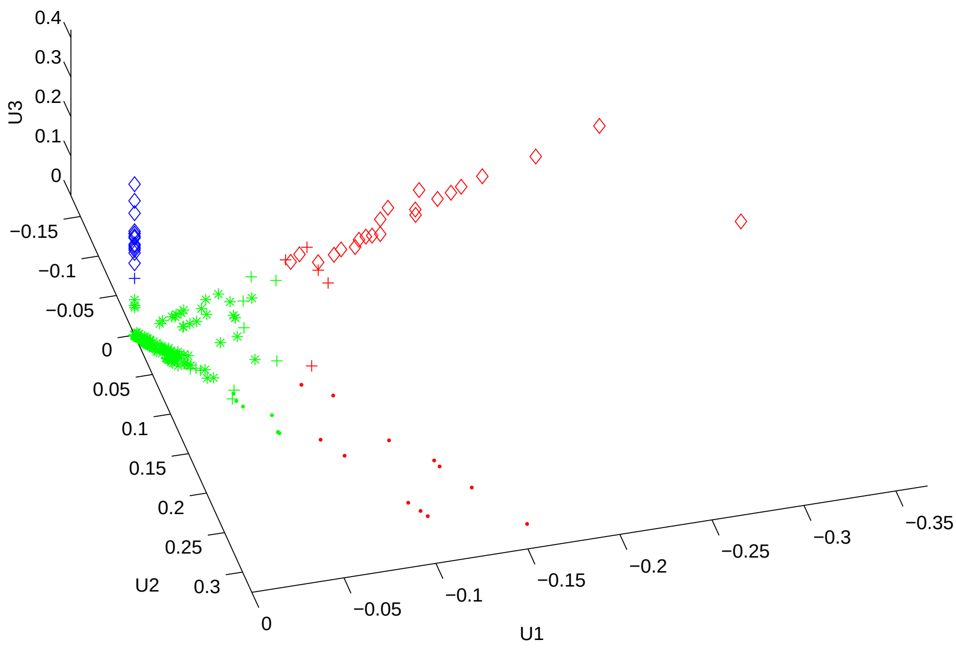

Figure 8: SVD plot of members with SDD color and shape labelling, showing extra boundary information.

- Mohamed Zinedine, Abdelilah Ziyad, Abdelkrim Afkir, Hamel Marzoug, Abdeslam Garoise, Radouane Hammadi, Stephane Ait Iddir, Mohamed Azil, Abdelaziz Rhouzlane, Abderrahmane Boujedli, Kamel Benakcha, Rachid Falah, Tarek Falahm El Moustapha ben Haddou, Farid Zarouali, Abderrazak Mountassir (same group identified by SVD).

- Abu Bakar Baasyir, Abdullah Sungkar, Encep Nurjaman (Hambali), Ali Ghufron bin Nurhasyim (Mukhlas), Yassin Syawal, Rahman al-Ghozi, Abdul Aziz (Samudra), Enjang Bastaman (Jabir), Amrozi bin Nurhasyim, Ali Imron bin Nurhasyim, Hutomo Pamungkus (Mobarok), Faiz bin Abu Bakar Bafana, Hasyim bin Abbas, Mohammed Nasir bin Abbas (Sulaeman), Abdul Rahim Ayub, Azahari bin Husin, Aris Sumarsomo (Zulkarnaen), Suranto Abdul Ghoni, Noordin Mohammad Top, Jhoni Hendrawan (Idris), Pranata Yudha (Mustofa), Wan Min bin Wan Mat, Umar Dul Matin, Abbas Edy Setiono, Thoriqudin (Rusdan), Mustaquim, Muhajir (JI members, SE Asian attacks).

- Osama bin Laden, Zain al-Abidin Mohammed Hussein (Zubaydah), Omar ibn Mahmoud Omar Othman (Qatada), Mohamed Heidar Zammar, Mamoun Darkazanli, Amar Makhlulif (Doha), Mohamed Bensakhria, Essid Sami ben Khemais, Tarek Maaroufi, Imad Eddin Barakat Yarkas, Anwar Adnan Mohammad Salah, Mohammed Galeb Zouaydi, Tayssir Alluni, Oussama Dara, Mohammed Bahaiah, Jose Luis Galan Gonzalez, Abdelaziz Benyaich, Salahed-

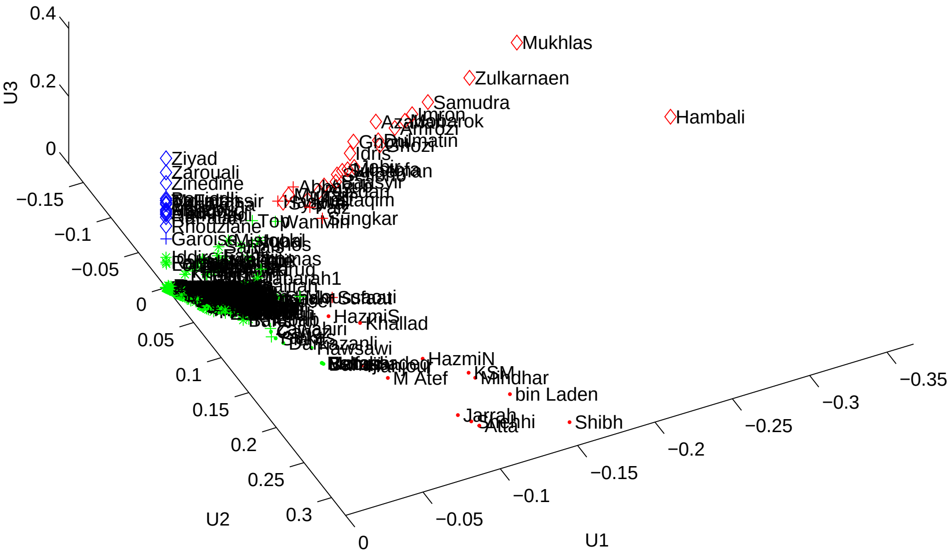

Figure 9: SVD plot of members with SDD color and shape labelling and short identifiers.

dine Benyaich, Said Chedadi, Driss Chebli, Najib Chaid Mohamed, Mohamed Fizazi (North African and European attacks).

- Zain al-Abidin Mohammed Hussein (Zubaydah), Safe Bourada, Laifa Khabou, Fateh Kamel, Abdellah Ouzgar, Zoheir Choulah, Said Atmani, Abderraouf Hannachi, Ahmed Ressam, Mustapha Labsi, Mourad Ikhlef, Adel Boumezbeur, Samit Ait Mohamed, Abdel Majit Dahoumane, Mokhtar Haouari, Amar Makhlulif (Doha), Yacine Akhnouche, Omar Chaabani (Jaafar), Rabah Kadri, Slimane Khalfaoui, Hassan Zemiri, Adil Charkaoui (Los Angeles millennium attack).

Other groups include: those involved with early attacks in Egypt and the early leadership of al Qaeda, and those involved in the Casablanca attack in 2003.

The interesting things about these groups are:

- Although they are based purely on link data, they correspond well to patterns of terrorist attacks. This shows that al Qaeda's functional structure (who plans, leads and carries out an attack) is heavily derived from existing familial and relationship connections among its members. (Although some group link structure is present in the raw data, it is by no means enough to determine attack groups.)

- Several people appear in multiple groups, thus revealing their role as the glue that binds disparate groups together. Notice that many groups with close geographical and relationship

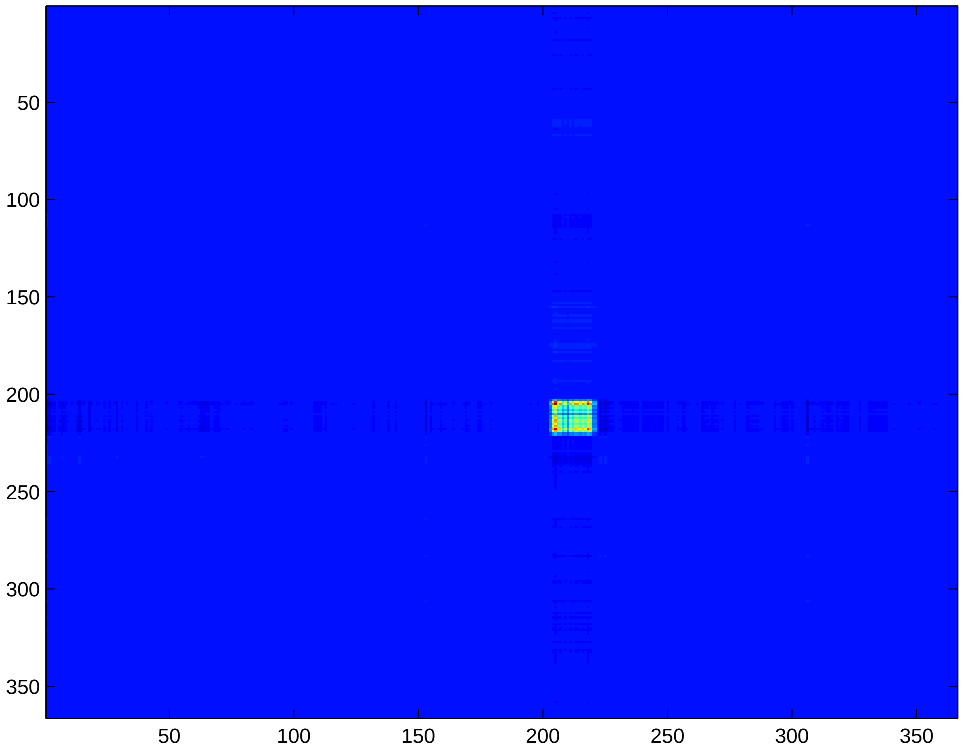

Figure 10: An example of an outer product matrix from the ICA of the relationship matrix. The presence of a small, connected group is easily visible.

ties still include one or two members of the al Qaeda leadership, showing how long-distance relationships maintain overall group cohesiveness.

- There are some individuals who ought, on the face of it, to appear as members of a group but do not. This may happen simply because not enough is known about them so they appear to be relatively unconnected generally. However, it is also possible that such people are deliberately trying to lie low, so it may be useful to apply extra scrutiny to them.

The choice of threshold affects the tightness of the boundary of each cluster – increasing the threshold reduces the membership of each cluster and removes some apparently anomalous individuals.

A sense of the overall ICA can be obtained by examining an image of the W and H matrices. Each column of the W matrix corresponds to one component. The presence of high values in this column indicates objects that are associated with this component (recall that the organization of the data tends to place similar people in adjacent rows already, which is why the clusters are so obvious). For example, column 1 reveals the cluster at rows ∼230–250, and column 3 reveals the cluster at rows ∼60–80 but with some weaker connections to the leadership (early rows).

Similarly, each row of the H matrix corresponds to a component, and indicates which attributes play a role in the selection of that component. The matrices are shown in Figures 11 and 12 respectively.

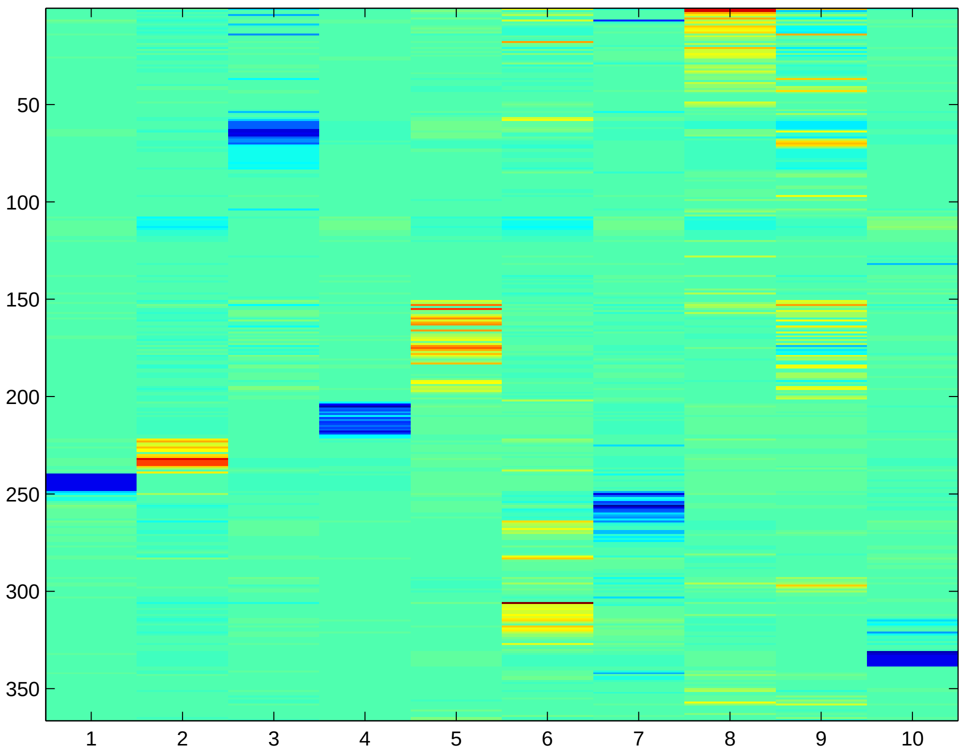

Figure 11: Image of the W matrix. Each row corresponds to one component, and the distinctive coloring to the members of that component.

4.2 Using demographic and relational information

We now add demographic information including: year of birth (dob), country of birth (birthplace), cultural group (clump), national status (natstatus), socioeconomic status (fses), religious background (religbgnd), type of school attended (school), education level attained (educ), type of education (edtype), occupation (occup), marital status (married), number of children (kids), possession of a criminal background (crimbgnd), year of joining al Qaeda (yrjoin), age at joining al Qaeda (agejoin), place at which member joined (placejoin), country in which member joined (countryjoin), fate, and year left the group (yrleft) usually by death. The number of demographic attributes and their amount of variation they show produces plots with much less clustering.

Figure 13 shows the basic clustering among al Qaeda members based on SVD. It is clear that the group is fairly homogeneous, except for a distinct cluster towards the bottom of the figure. As we shall see, this cluster represents a subgroup of members who have a stronger religious background and religious education than the majority.

Analysis of the attributes, shown in Figure 14, shows that dimension 1 captures the variation in educational attainment, dimension 2 captures variations in locations such as where members were

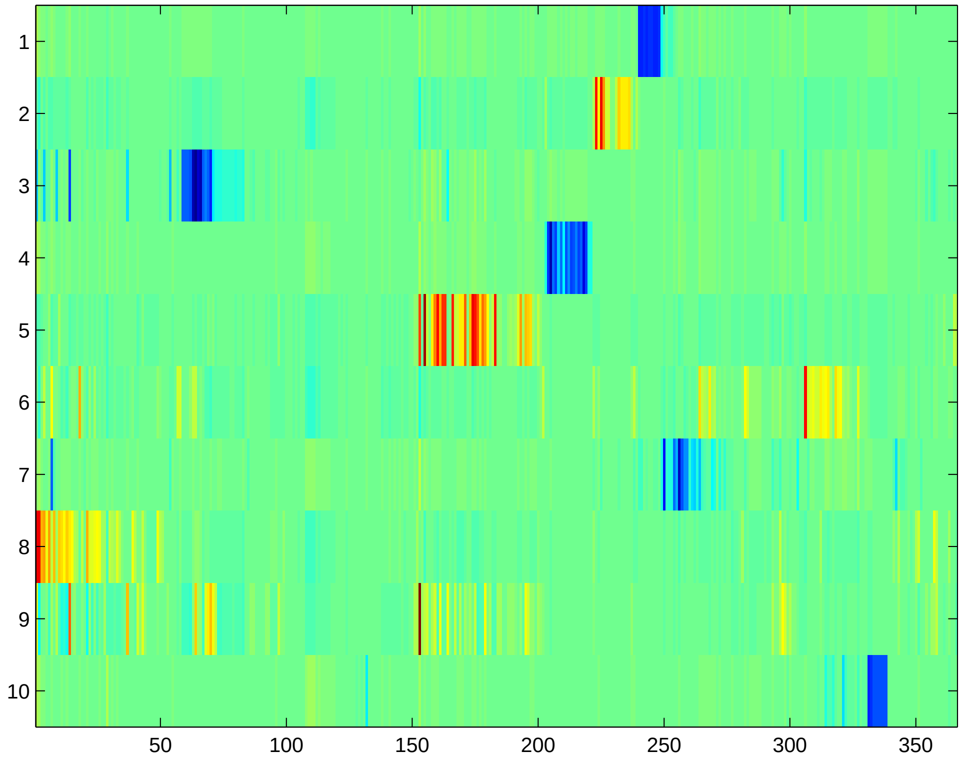

Figure 12: Image of the H matrix. Note that it is the transpose of the W matrix because the data matrix is (almost) symmetric.

born and joined the organization, and dimension 3 captures differences in religious background and schooling.

These relationships among the dimensions in the attribute space must be consistent with the relationships among members in the object space. Figures 15 and 16 show the most unusual members, projected in different dimensions. We can see that, for example, al-Zawahiri and bin Laden are well-educated while Omar Khadr and Abdul Karim Khadr are not (separation along axis U1). The second dimension captures differences in country of birth and country in which the member joined the jihad. Since the countries are coded alphabetically, this reveals no absolute information about the structure of al Qaeda, although it may reveal some relative information. For example, Abdallah ibn Mohammad al-Rashoud was born and joined the jihad in Saudi Arabia, while Chellali Benchellali was born in Algeria and joined in France and Wadih el-Hage was born in Lebanon and joined in Afghanistan.

Figure 16 shows the relationship between education and religious background. Now the vertical dimension represents degree of religious background, with Hage a non-Muslim with a secular education, and Mukhlas and Zulkarnaen from a religious background and pupils of a Madrassa. Note the small cluster in the lower left-hand corner of members who are both religious and highly educated: Sheikh Omar Abdel Rahman, who has doctorate, and Abu Bakar Baasyir and Abdullah Sungkar, who both have Masters degrees.

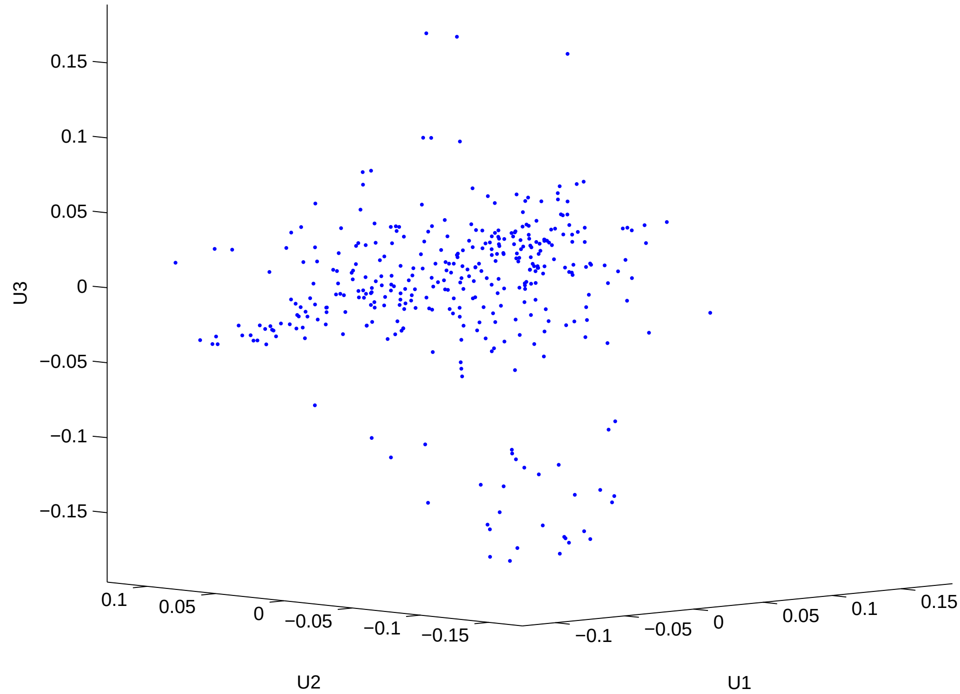

Figure 13: SVD plot of al Qaeda members using both demographic and relationship attributes.

The size of the singular values (the diagonal of S) give some indication of the relative importance of the factors corresponding to each dimension. These values are 13.5, 11.9 and 9.6 indicating that education is about 40% more important as an explanation for variation among al Qaeda members than religious background (13.5/9.6).

Figure 17 shows the effect of cultural background. It is clear that the leadership (squares) are slightly different from the remainder of the members, but this is not surprising because of the group's history. Its leaders come from similar backgrounds and are of a similar age. There are very little difference between the characteristics of the other groups, although the Maghreb members (circles) show some systematic differences from the core Arab (stars) members.

Figure 18 shows an SVD overlaid with information obtained from SDD. In this case, we have used a more powerful combination of the two called the JSS methodology: SVD is applied to the data matrix, A, the decomposition is truncated at some k and the component matrices multiplied to give a modified version of A. SDD is then applied to the correlation matrix obtained from the modified version of A. This correlation matrix captures higher-order correlation information and tends to provide a clearer picture of complex data than using SDD directly on A.

It is clear from the figure that the extra information agrees with the clustering given by SVD. Note that the group of well-educated, religious members is captured as a subgroup. It is also noticeable that the well-educated cluster displays more variability than the matching cluster of less educated members. There are substantial overlaps between the well-educated cluster and

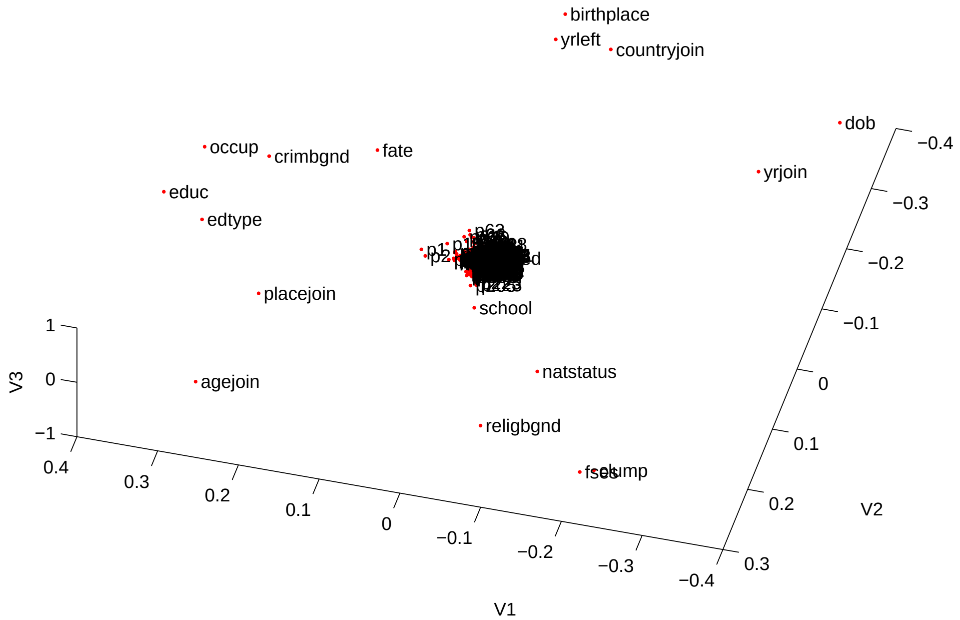

Figure 14: SVD plot of dataset attributes, showing the relationships among them. The large cluster in the center are the attributes associated with explicit relationships.

the group's leadership, providing further evidence that the stereotype of terrorists as ignorant, brainwashed, or psychotic does not apply to al Qaeda.

Independent Component Analysis is not useful on the demographic data because it tends to select small groups who resemble each other on the basis of a few demographic attributes – which is both misleading, and obvious from the raw data.

5 Discussion

5.1 Methodology

We can see from these results that the major benefit of SVD is its ability to select and order objects (in this case al Qaeda members) from most to least interesting. This is partly because al Qaeda is a fairly homogeneous organization, so that there are few significant demographic clusters within it. Even the clustering visible in the relationship data is important only for the more unusual/important members – most of the rank and file are quite similar. SDD allows more detailed and discriminative analysis, as it is able to provide boundaries between subgroups more precisely.

The major benefit of ICA is its ability to find and select closely coupled groups of individuals. Unlike a traditional clique-discovery algorithm, ICA allows an individual to participate in several

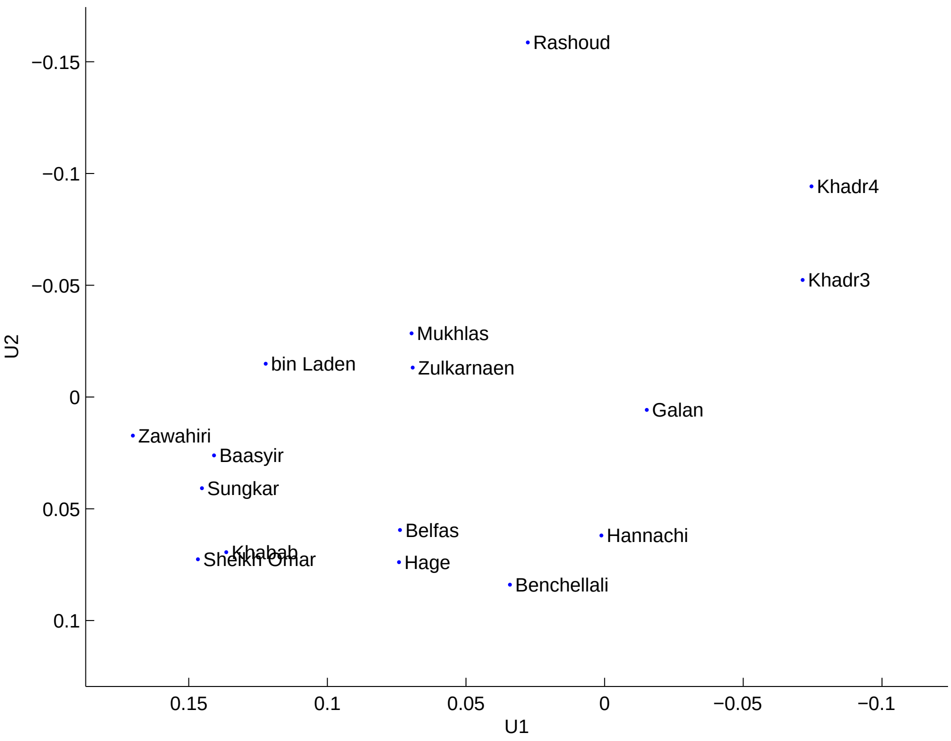

Figure 15: SVD plot of 16 interesting members (greater than 1.5 times the median distance from the origin) in dimensions 1 and 2. Dimension 1 represents variation in education; dimension 2 variation in place of origin.

groupings which is both more flexible and more realistic.

A number of parameter choices were made in these algorithms: the number of dimensions at which to truncate the SVD for visualization, and for preparation for SDD, the scaling of the array entry magnitudes for SDD, the boundaries for considering objects interesting, the number of components used for ICA, and the thresholds used for using ICA components to select groups of members. Sensible values for all of these were chosen, but other structures might conceivably be revealed by other parameter choices. At present, no principled ways to choose these parameters are known.

A major advantage of matrix decompositions over typical social network and link analysis tools is complexity. The matrix decompositions used here have complexities that are typically cubic in n, the number of people being considered. However, when the data is sparse, as relationship data usually is, this can be reduced to linear, which remains feasible even for much, much larger datasets. In contrast, measures such as centrality have complexities that are at least cubic in n and often worse (because they often consider all paths in a graph). Furthermore, the software tools used for link analysis often assume quite small networks and so are not optimized for datasets even of this size, let along larger ones.

Link analysis also has a number of other drawbacks. First, such analyses are at the mercy

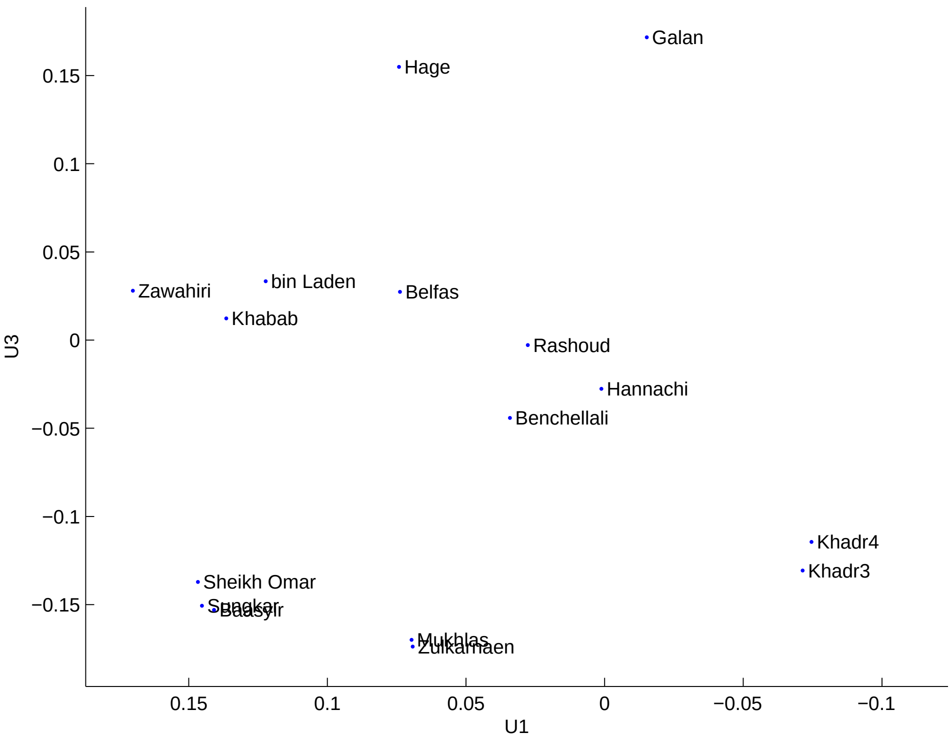

Figure 16: SVD plot of 16 interesting members (greater than 1.5 times the median distance from the origin) in dimensions 1 and 3. Dimension 1 represents variation in education; dimension 3 variation in religious background.

of their graph-drawing algorithms, which may create a misleading impression of the importance of an individual through an accident of placement. Second, the individuals to which attention is drawn are those with many connections. This is useful, but is easily extracted from the raw data, and fails to show either centrality measures or higher-order connections. Third, the graphs quickly become large, so that only small pieces can be seen at a time, which makes it hard to extract global information or see large-scale patterns. Hence, although visualization via link analysis plays to human strengths in seeing patterns, the size and complexity of the graphs involves tends to make this difficult to achieve in practise.

5.2 al Qaeda

It is clear from this analysis that al Qaeda is better regarded as a loose confederation of groups with related aims than as a hierarchically-controlled, functionally-organized single group. Repeatedly, the structure that emerges from considering relationships among members matches the structure related to groups that have carried out attacks. In al Qaeda, it is who you know, not what you know that determines your role in the organization. This observation has also been made by Sageman [16], Gunaratna [6], and others.

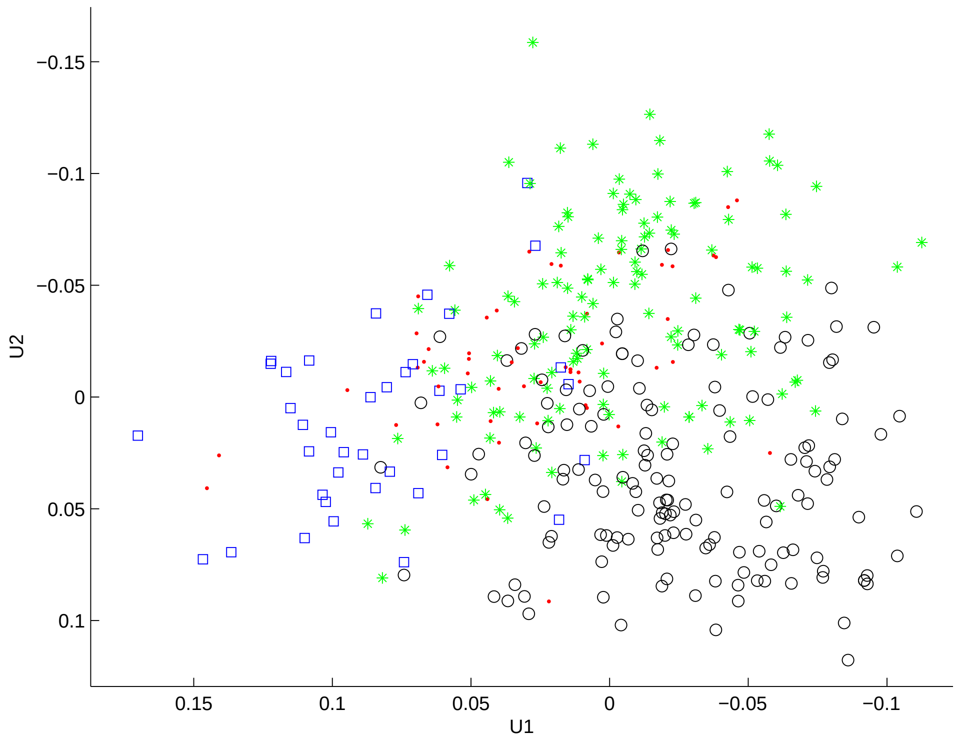

Figure 17: SVD plot based on both relationship and demographic data, showing cultural group membership (squares – leadership, stars – core Arabs, circles – Maghreb, dots – S.E. Asian).

The difference between the three cultural groupings: core Arabs, South East Asian muslims, and European/Maghreb muslims, is also strong, as is the fact that the leadership is not only made up of core Arabs, but is also much more tightly bound to this group than to the others. In fact, it is surprising that there are not more people who play the role of Hambali in connecting groups together – surprising enough that it raises the question of whether there are in fact such people, either not captured at all in this dataset, or about whom not enough is known to elicit this role.

It is also clear that al Qaeda is an egalitarian organization in the sense that there is no particular profile to its members. Although education level is the most important variable among al Qaeda members, there is no clear division across the spectrum from most to least educated. The only variable that separates the group into two subclusters is religious background and schooling; and, contrary to widespread expectation, it is the more religious cluster that is the smaller.

The use of SVD as a technique for identifying the most interesting members of a group is also quite successful at identifying either group leaders or those with an important technical role. For example, Figure 4 identifies both bin Laden and Ramzi Mohammad Abdullah bin al-Shibh (who handled money transfers for the September 2001 attacks).

Other possible analyses using matrix decompositions would be to restrict the dataset to those still alive and examine the relationships among the organization; and also to examine how the

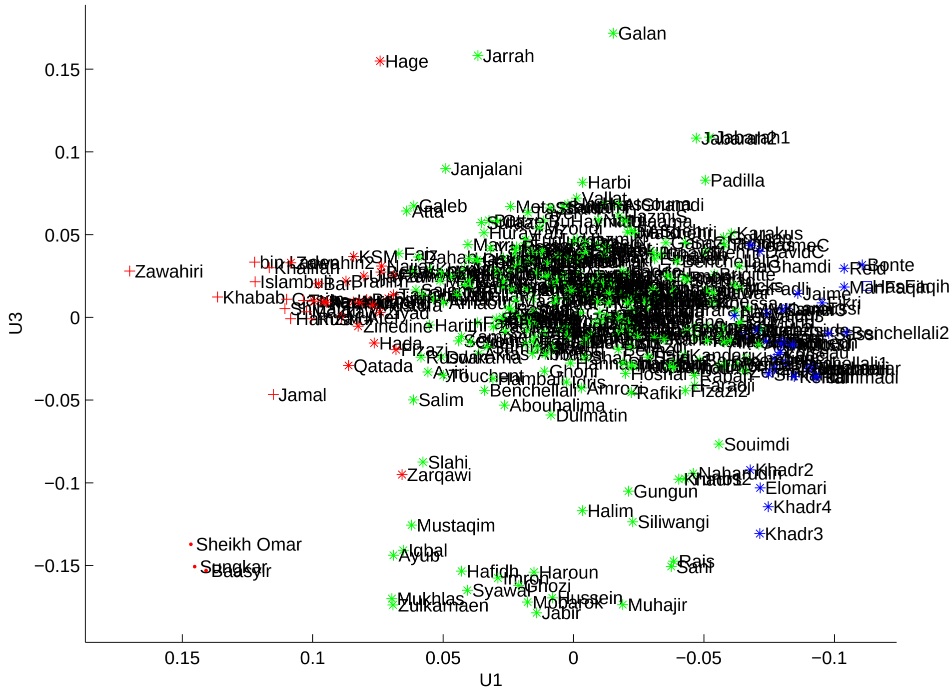

Figure 18: SVD plot with SDD color and shape labelling, using both demographic and relationship data.

relationship structures have changed over time. Of course, the addition of classified information to the data used here would also provide a more reliable picture. Together these three approaches might make prediction possible: previous attacks have often been marked by quite strong changes in connectivity and apparent demographics over a short time period by the individuals carrying out the attacks and their support cells. Such changes would certainly be visible in the matrix decomposition plots.

The dataset here does not consider connectivity as a function of communication, but clearly the addition of links based on email or telephone conversation (for example, from Echelon) would provide a better and more timely picture of connections within the organization.

6 Conclusion

Transnational terrorist groups such as al Qaeda present new challenges for counterterrorism. Like all terrorist groups, their members and actions are hard to see against the background of innocent groups and activities. Al Qaeda has demonstrated that groups formed as very loose federations can still be effective, creating a new requirement to detect sporadic or short-lived links to create a workable defense.

We have shown how matrix decomposition techniques can be used to enhance link and social network analysis. These techniques are more revealing than clique detection and centrality measures, less reliant on exact data, and much more efficient to compute.

While we discover nothing particularly new about al Qaeda, much existing knowledge is replicated from much less data. In particular, the techniques we have used are able to detect and rank the importance of members of the group solely based on their relationships. This is a powerful addition to the arsenal of counterterrorism data analysis techniques.

Acknowledgement: I am deeply grateful to Marc Sageman for making the al Qaeda dataset available.

References

[1] F.R. Bach and M.I. Jordan. Finding clusters in Independent Component Analysis. Technical Report UCB/CSD-02-1209, Computer Science Division, University of California, Berkeley, 2002.

[2] W.E. Baker and R.B. Faulkner. The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review*, 58:837–860, December 1993.

[3] J. Corbin. *Al-Qaeda: In Search of the Terror Network that Threatens the World.* Thunder's Mouth Press, 2002.

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

[5] United States Government. *Final Report of the National Commission on Terrorist Attacks Upon the United States.* 2004.

[6] R. Gunaratna. *Inside al Qaeda.* Berkley Publishing Group, 3rd edition, 2003.

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

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

[9] D. Jensen and J. Neville. Data mining in social networks. Invited presentation to the National Academy of Sciences Workshop on Dynamic Social Network Modeling and Analysis, November 2003.

[10] 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.

[11] G. Kolda and D.P. O'Leary. A semi-discrete matrix decomposition for latent semantic indexing in information retrieval. *ACM Transactions on Information Systems*, 16:322–346, 1998.

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

[13] S. McConnell and D.B. Skillicorn. Semidiscrete decomposition: A bump hunting technique. In *Australasian Data Mining Workshop*, pages 75–82, December 2002.

[14] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis, eigenvectors and stability. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 903–910, 2001.

[15] D.P. O'Leary and S. Peleg. Digital image compression by outer product expansion. *IEEE Transactions on Communications*, 31:441–444, 1983.

[16] M. Sageman. *Understanding Terror Networks*. University of Pennsylvania Press, 2004.

[17] R.H. Shultz and A. Vogt. The real intelligence failure on 9/11 and the case for a doctrine of striking first. In R.D. Howard and R.L. Sawyer, editors, *Terrorism and Counterterrorism: Understanding the New security Environment*, pages 405–428. McGraw-Hill Dushkin, 2004.

[18] G.W. Stewart. On the early history of the Singular Value Decomposition. Technical Report TR-2855, University of Maryland, Department of Computer Science, March 1992.

[19] K.M. van Meter. Terrorists/liberators: Researching and dealing with adversary social networks. *Connections*, 24(3):66–78, 2002.