ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.8, August 2008, pp.1980−1994
DOI: 10.3724/SP.J.1001.2008.01980

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于社会网络可视化分析的数据挖掘*

杨育彬[1,2+], 李 宁[1], 张 瑶[3]

[1](南京大学 计算机软件新技术国家重点实验室, 江苏 南京 210093)

[2](School of ITEE, UNSW@ADFA, Canberra, Australia)

[3](南京大学 金陵学院, 江苏 南京 210093)

Networked Data Mining Based on Social Network Visualizations

YANG Yu-Bin[1,2+], LI Ning[1], ZHANG Yao[3]

[1](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

[2](School of ITEE, UNSW@ADFA, Canberra, Australia)

[3](Jinling Institute, Nanjing University, Nanjing 210093, China)

[+]Corresponding author: E-mail: yangyubin@nju.edu.cn

Yang YB, Li N, Zhang Y. Networked data mining based on social network visualizations. *Journal of Software*, 2008,19(8):1980−1994. http://www.jos.org.cn/1000-9825/19/1980.htm

Abstract: Studies in social network theory focus on characterizing complex social relationships by firstly mapping and visualizing them into a graph, and then subsequently identifying the corresponding graph properties. This paper provides an integrated approach, which combines social network analysis and data mining theory with the necessary geographical attributes to analyze 1417 instances of terrorism that occurred world wide during the period 1980−2002. The study reveals interesting patterns on the evolution of these terrorist organizations over two decades. The proposed method can be easily generalized to be applied to other types of large-scale networked datasets, such as micro-array data, and genomic networked data, etc.

Key words: social network analysis; data mining; network dynamics; network evolution

摘 要: 把社会等复杂系统看作网络的思想由来已久.利用社会网络分析的方法,能够对各种社会关系进行精确的量化表征和分析,从而揭示其结构,对一系列当代社会的现象进行更加深入而具体的解释.结合社会网络可视化分析和数据挖掘的理论与方法,引入相关的地理信息,对包含 1980−2002 年间世界范围内 1417 例恐怖袭击事件的数据库进行数据分析,以这些恐怖袭击事件各要素节点之间关系作为基本分析单位,对恐怖组织之间的活动模式和发展特点等内在规律进行挖掘与解释,得出有意义的结果.提出的方法可以有效地推广应用于蛋白质结构分析、生物基因分析以及各类社会问题的分析过程.

关键词: 社会网络分析; 数据挖掘; 网络动态模式; 网络发展模式

中图法分类号: TP311 **文献标识码**: A

* Supported in part by the Key Program of the National Natural Science Foundation of China under Grant Nos.60723003, 60505008 (国家自然科学基金); in part by the Natural Science Foundation of Jiangsu Province of China under Grant Nos.BK2007520, BK2006116 (江苏省自然科学基金); in part by the Australian Research Council (ARC) Centre for Complex Systems under Grant No.CEO0348249 (澳大利亚复杂系统研究中心项目)

Received 2008-01-17; Accepted 2008-04-18

1 Introduction

Data mining[1] is the process of discovering novel patterns in databases. Traditional data mining techniques assume that the attributes are independent and that the values that each of these attributes can take are also independent. This assumption makes theoretical analysis of data mining techniques feasible, but unfortunately it is an unrealistic assumption in many real life situations. Relaxing the inaccurate assumption that the attributes are independent and the values that each attribute can take are also independent would naturally mean that there are virtual links connecting these attributes and their values. As such, one can visualize the relationship between these attributes and their values as a network. This is a much richer representation that can be used to substantiate data analysis with some of the rigorous measures used in social network theory. This approach we call networked data mining.

In networked data mining, a number of variations exist in the literature. One is known as link analysis[2−4]. Link analysis research uses search and probabilistic approaches to find structural characteristic in the network such as hubs, or identifying potential relationships for relational data mining. Link analysis alone is insufficient as it looks at one side of the coin and ignores complex nonlinear relationships that may exist between the attributes. Another approach depends purely on visualization, such as NetMap[5]. Unfortunately, these tools that depend on visualization alone - despite being useful to provide some insight - they are insufficient and rely on the user to carry out many tedious and time consuming tasks, many of which could be automated.

In addition to the previous discussion, most of the work on link analysis or network visualization ignores the temporal dimension. Uncovering a relationship among or within attributes (connecting the dots) is an important step, but in many domains it is more important to understand how this relationship evolved over time. Hence, understanding network dynamics and evolution is needed to complete the picture. Once we understand the dynamics and evolution of these relationships, we can search for ways to disconnect the dots when and if needed.

This paper presents a case study in the area of terrorist networks combining social network analysis and data mining theory with the necessary geographical attributes. The structure of the rest of this paper is as follows. In Section 2, a short preliminary literature review is presented to point the reader to several key papers in the literature. Section 3 presents an overview of the system framework. A case study is then introduced in Section 4 and this is followed by a description of the network measures used for this study, and by a comprehensive analysis of the case study with those network measures. Conclusions are then drawn in Section 7.

2 Review

2.1 Social network theory

Studies in social network theory focus on characterizing social relationships by first mapping them into a graph and subsequently identifying the corresponding graph properties. Entities (known as actors or agents) in a social relationship (such as friendship, drug dealers, or terrorists) are mapped to nodes in a graph, while the links (arcs) between these nodes either reflects the existence and type, or alternatively the absence of a relationship. Links can be directed (implying a relationship from one side to another, but not vice versa) or undirected (implying a relationship in both directions). For example, if A and B are two mobile phones talking to each other, then A can talk to B and B can also talk to A; i.e. the link is undirected. In some relationships, such as the mother-daughter relationship, if A is the mother of B, then B cannot be the mother of A. In this latter case, the graph must be directed. Most studies in social network theory assume an undirected graph. Nodes and/or links can also be weighted, although most studies in social network theory focus on unweighted graphs, because the emphasis of these studies is on the properties of

the topology of the graph itself.

Networks were traditionally studied as collections of nodes with a uniform random probability that any two nodes were connected. This type of networks has been known as random graphs. Albert and Barabasi[6] provide a comprehensive overview of networks from both this traditional as well as a more recent, complex system, perspective. They examine a number of key network properties, including connectivity, clustering (the probability that two nodes linked to some other node, are in turn directly linked), and degree and path length distribution, to name but a few. In particular, they emphasize that real networks often differ from random networks in terms of having larger clustering coefficients and also a more skewed degree distribution (especially in the form of scale-free networks, such as the internet[7]).

Network measures have become sufficiently established, that there are now a range of software packages that process networks to obtain these and other properties. However, while it is possible to define statistical properties of networks, there are related problems of both how to process and understand very large networks. Wu, *et al.*[8] take advantage of the characteristic of scale-free networks that some nodes are more connected, and are hence, more 'important' to simplify networks by only considering the more 'important' nodes.

A branch of social network theory, known as network sampling, identifies appropriate approaches to collecting information about a social network. There are generally two network sampling approaches, either a sample of all possible candidates, or a snowball sample[9]. This latter approach has been adapted in different schemes, but the basic idea is the same. To sample a network, one starts with a subset, then each member in the subset is asked for recommendations for extra nodes, this process of adding more recommendations (nodes) is repeated until a reasonable sample is collected. The key characteristic in this approach is that individuals are chosen for their membership; not for their characteristics. Global samples are viewed as generating more reliable representations of a network, while snowball sampling may be required in situations where the network under consideration is only a small proportion of the overall population or where there may be a significant cost associated with determining the characteristics and linkages of a node.

Unfortunately, snowball sampling is insufficient for what is known as "hidden populations". For example, when examining the population of drug dealers, it is not easy to start with a subset, and it is not possible to ask a drug dealer to talk about other drug dealers. Researchers in social networks have developed many techniques to handle this problem, such as: ego centered networks, targeted sampling, and respondent driven sampling[10].

In certain applications, such as terrorism, the previously enumerated approaches for dealing with hidden populations fail. First, one cannot ask a terrorist, to ask other terrorists, to tell him how many terrorists they themselves know. Second, communicating with a terrorist is usually not possible until he/she is captured, etc. Lastly, but not least, a key question in security is, when is a suitable time to capture the network? These difficulties hinder previous methods. In fact, in most of the recent studies on terrorist network analysis, the terrorist network was usually viewed a statical social network created at a certain time point, without considering its growth and evolution along time axis[11,12]. By doing so, the network is analyzed only once so that some important knowledge, such as "how did the size and structure of a network evolve", is not able to be fully discovered and then clearly visualized. To address this issue, a novel terrorist network analysis method based on dynamic network visualization is proposed in this paper, in order to connect the dots over time (i.e. to capture the temporal effect in social network visualization and analysis).

2.2 Link analysis

The field of link analysis has two main directions. One of these is to use an understanding of the link structure of a network to identify desired characteristics of the network. The other is to form a, usually probabilistic, model

of the network in order to make predictions about characteristics of the network.

A first order analysis of links within a network is to analyze which nodes are highly connected. The relevance of those nodes is, however, dependent on the problem domain. Staab[2] considers the domain of marketing, where he theorizes that if strongly, socially connected individuals could be identified, then those individuals would be worthwhile targets for advertising. Interestingly, Kao, *et al*.[3] use the opposite approach when attempting to improve web search technology. In their work, they use the observation that many web pages contain sub-sections that primarily exist for navigational purposes, and that do not usually contain other forms of information. These sections are identified by their high level of connectivity to other pages. The work in this paper shows improved search results may be obtained by ignoring the content of those highly connected sections.

In the domain of link prediction, a probability model is often generated that attempts to reflect the linkages within the network. Such a model may then be used to predict either the presence of links in a similar network, or of other parts of the same network. For example, Tasker, *et al*.[4] consider the network formed by the nodes of faculty, staff, research scientist, staff, research group, research project, course and organization within a school of a university. This work attempts to determine if particular relationships (such as Advisor(faculty, student)) exists between nodes, based on the shared links between other nodes. To do this, it developed a model which reflected the probabilities that such relationships occurred between links, based on data from two schools. This model was then used to discover the same relationships based on the network of another school.

Similar approaches have been used on citation references extracted from CiteSeer[13]. Hill[14] forms a network between authors, publications and cited works. This network is then used to identify the authors of an unseen publication based solely on its citations. In this case, probabilities are not used, but a variant is considered where only links that occur more often are considered. Popescul and Ungar[15] form a similar network, and then try to predict probable citations for a paper, based on its authors, where the paper was published and other citations of the paper.

2.3 Visualization

Two aspects of visualization methods in data mining may be considered: how data may be visually presented, and how such presentations may be manipulated or explored.

Regarding presentation, simple techniques regarding one and two dimensional data have been long established in the forms of graphs and maps. Keim[16] gives a comprehensive taxonomy of visual techniques in terms of the data type to be visualized, the visualization technique used and the interaction and distortion technique used. The key complication of visualizing multidimensional data is the question of how it may be mapped to a two-dimensional display. Some approaches, such as Self-Organizing Maps[17], attempt to map the multi-dimensional data into a two-dimensional state while preserving certain properties, such as attempting to position data points so that their two-dimensional Euclidean distance matches their distance given by some other measure. Other approaches attempt to represent different dimensions by symbolic representations, i.e. differing symbols or colors or sizes. Chen, *et al*.[18] developed a representation of the characteristics of the pages of an internet site, where colors were used to represent how long a page was viewed for, the thickness of links represented how often they were used, and size represented how many times a page was viewed. Multiple dimensions may also be treated by either only viewing certain dimensions, or by collapsing multiple dimensions into a single one.

Due to the sheer size of many data mining problems, to be able to effectively explore the data it must be possible to manipulate the representation chosen. In visualizations where only certain dimensions are shown, or where dimensions are collapsed in some way, it is common to allow users to alter the choices that are made for these

elements. For example, Pampalk[19] extends Self-Organizing Maps to develop a series of Visualizations that are generated by differing the weights of the elements that contribute to the distance measure used by each visualization, while Chen[18] allows a user to specify which measures or collection of measures that they wish to see on a particular visualization.

Another way of dealing with scale, is allowing the user to filter the data examined. Goldberg and Wong[20] give an example of gradually refining the data under consideration to identify abnormal elements of a network. The refining process may take several approaches, it may be that data fitting a user's query may be used, or data may be automatically thresholded by either magnitude or ranking (i.e. being the largest, or first).

3 The System Framework

We have developed an interactive social network visualization and analysis tool to work with multi-field data from multiple heterogenous sources. It is domain independent; being simply configurable and scalable for any dataset. The data model is based on a network of associations - each field in a dataset represents a layer in the model and each possible value for an ordinal field is a node in that layer. Traditional data records - a set of fields and their particular values - then exist as a set of connected nodes. Multiple records then form a graph that may be connected or disconnected - depending on commonality of field values between different records. For example, consider a dataset concerning terrorist attack records. Each record includes four fields: attack type, terrorist organization who initiated the attack, country and year. If one record concerned a terrorist organization A performing a bombing in 1996 in country C_1, then a four node connected graph would be established, each node representing one of the four attributes. If a second record concerned terrorist organization A with a hijacking in country C_2 in 1992, then a 7-node connected graph would be built - node A serving as a common node (potential linkage) between the two records. This data model facilitates the extraction of linkages and associations between data that could likely be missed in a traditional record-based approach.

Our system follows a server-client model. Clients are heterogenous data-sources and transmit their information to the centralized server. Each client's data may be partial or incomplete and the first task for the server is the alignment, reconstruction and amalgamation of data from multiple clients (single clients with complete data are, of course, also supported). The server-client element of our system also incorporates a temporal aspect. As data is received by the central server it is incorporated and any visualized graph updated based on the newly arrived data - allowing visualization and analysis of the data at any stage - not simply on a "complete" dataset. This model matches the way that many real-world datasets are gathered and, ideally, analyzed - for instance a supermarket chain collecting and analyzing sales from all its branches, or a government agency collating intelligence from a number of heterogenous sources.

The server not only collates all data centrally, it also acts as an interactive visualization, and data plus network analysis tool. Various 2D graph representations are provided as views of the data, and a number of path and network analysis tools can be invoked through the application's interface. It is this network linkage based lens on the data, with the associated visualization and analysis tools that lies at the core of our system and arguably provides new insights into complex and/or rich datasets.

Both client and server are interactive, graphical applications. For clients, users select which fields from their database they wish to transmit to the server. Similarly on the server side the user may select any subset and ordering for the data fields it is receiving to visualize and analyze. For instance the database used for this paper has fields including City, Country, Organization, Attack Type, Target Type and Casualties, but most of the analysis and visualizations presented will concern only three layers (fields) - Country, Organization, and Attack Type. Interactive

exploration of data is thus facilitated by supporting the selection of subsets of fields and their particular order - links are drawn between nodes from adjacent layers (that derive from an original data record).

The server's primary interface presents a 2D networked view of the data. Several automatic layouts are provided on the basis of spatially separating and grouping the nodes on the basis of layer (data field). Together with the "random" layout there is the arrangement of layers vertically, horizontally, and as a group of concentric circles. Figure 1 shows three layers - Country, Organization, and Attack Type arranged as concentric circles. Further, users are able to "deform" any graph by dragging individual nodes to new positions. Together with this, users may select a subset of a graph by choosing a particular node and being shown only those nodes that share a connection with it. Figure 2 shows one of the powerful features of our system - its graphical link analysis. Link analysis is a tool for

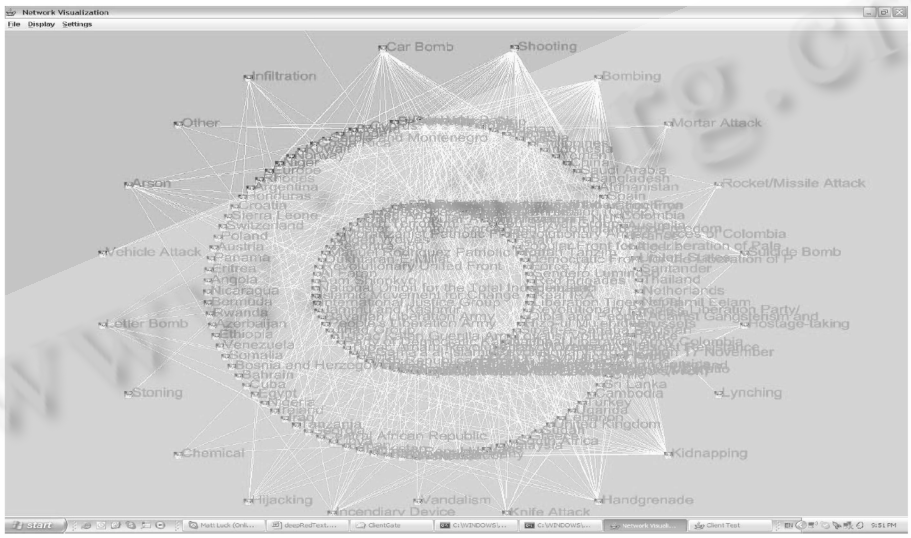

Fig.1 The server's interface and network representation. The network shows the three layers of Country, Organization, and Attack Type arranged as concentric circles

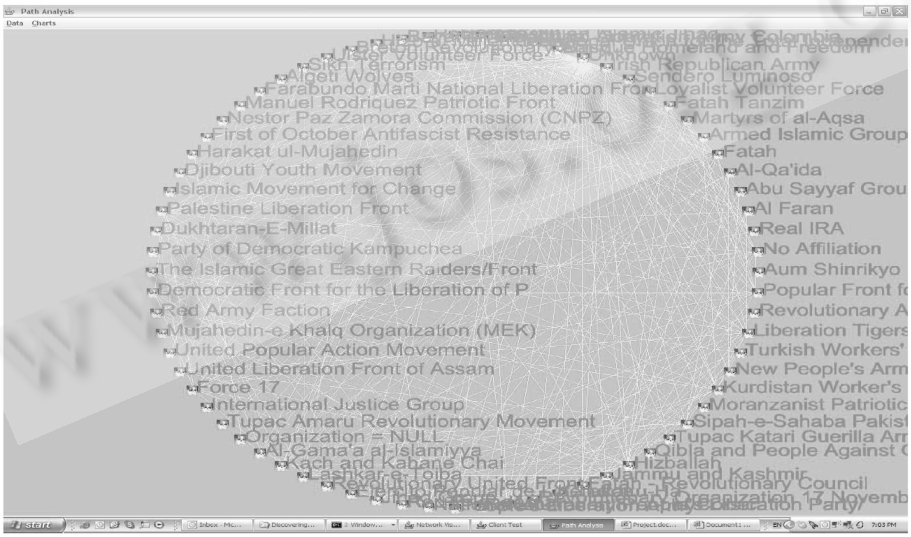

Fig.2 Linkage visualization, generated from a query to show links between organizations that have carried out attacks in the same countries

finding links between nodes at the same level, through connections with a common node from another layer. Figure 2 shows links between terrorist organizations on the basis of the Location - if the graph shows an edge between two organizations it means that have carried out an attack in the same location as one another. In using the link analysis feature the user selects the layer to be intra-linked, direction to look for links in, and "distance" (number of edges) within which links are considered to exist.

A link in Fig.2 is a multi-node link. Consider a record consisting of the two tuples $\langle a1, b1, c1\rangle$ and $\langle a2, b1, c2\rangle$. The fact that $b1$ is connected to $c1$ through the first tuple and $a2$ is connected to $b1$ through the second tuple, it does not necessarily mean that $a2$ is connected to $c1$. Thus a link in this graph is defined by multiple nodes (3 in this example). This makes any tracing through the network meaningful and efficient.

4 Case Study

Terrorism is probably one of the most serious challenge we face in this century. The data we are using were generated by querying the online publicly available repository of terrorists "attacks" at the Institute on Counter Terrorism of USA. The data was processed and transformed into an Access database table with the following fields: date of attack, organization, attack type, target type, location, number of people killed, number of people injured, number of terrorists if known, and text explaining each incident. The complete data set summarizes 1417 terrorist operations since 1980. Astonishingly, the data reveals that 46% of the incidents were cases for which it is still unknown which terrorist organization was responsible for the attack. The original data set (we call OriginalDB in the rest of this paper) was further split into two subsets: one that includes only those records with known organizations (we call KnownDB) and the other consists of the records with unknown organizations (we call UknownDB). As the original data was generated in Israel, it contained many records on terrorist activities that occurred in Israel. We thought it would be interesting to see if there is a bias in the dataset because of possibly skewed data collection. As such, we divided the KnownDB into two subsets: one for attacks carried out in Israel (we call KnownIDB) and the other for the rest of the world (we call KnownNoIDB). Table 1 provides summary statistics of these datasets.

Table 1 A summary statistics of the different datasets

Measure	OriginalDB	KnownDB	UnknownDB	KnownIDB	KnownNoIDB
Number of instances	1417	767	650	351	416
Number of organizations	70	69	N/A	12	61
Number of attack types	22	20	22	10	17
Number of targets	44	44	44	28	37
Number of locations	96	59	83	3	56

Given a particular data set, a number of different networks may be formed. We have used the notation X-Y to denote a network where the nodes of the network are the unique labels in X, and those nodes are linked if they share some common attribute Y. In particular, the networks generated are:

- Org-Loc Each Organization is a node, and two nodes are linked if they have both carried out an attack in at least one common location.
- Loc-Org The previous networks inverse, i.e. each Location is a node, and two nodes are linked if there exists one organization that has carried out an attack in both locations.
- Org-Attack Two organizations are linked if they have both used at least one common type of attack.
- Org-Target Two organizations are linked if they have both attacked the same type of target.

- Loc-Attack Two locations are linked if the same type of attack has occurred in both of them.

5 Network Measurements

The following graphics display a particular dataset when being analyzed by a particular measure. These measures form time series, as at each point in time, the network measured is that network generated by only those attacks up to that particular time point. Thus, all the networks weakly monotonically increase in size as time progresses. It is also worth noticing that, in order to capture the temporal evolution of the network structure, all the measures adopted in this paper are calculated on the network level, rather than on the node level. Therefore, node-level measures such as Centrality, Betweeness, Closeness, etc., which are widely used in other social network research to identify each node's importance, is not adopted in our system since they are not suitable for our analysis task. We emphasize much more on analyzing the whole network's evolution along time axis, which leads us to choose the following network-level measures to perform our analysis.

5.1 Number of nodes

The number of nodes, n, denotes the total number of nodes in a network.

5.2 Number of edges

The number of edges, e, denotes the total number of edges (or links) in a network. In this report, the network is represented as an undirected graph. Note that two nodes will only ever be connected by one link, and that a node is never connected to itself.

5.3 Average degree

The degree, $\mathrm{d}(u)$, of a node, u, is the number of edges it has. The average degree of a network is calculated by:

$$deg = \frac{\sum_{\forall u} \mathrm{d}(u)}{n} \tag{1}$$

5.4 Degree distribution

The degree distribution of a network is defined as a histogram of the occurrence of the node degree values, $\mathrm{d}(u)$. It can be used to analyze the structural difference between different networks.

5.5 Cluster coefficient

The cluster coefficient is a measure of how well linked are the neighbors of each node in a network. It measures how close the neighborhood of each node comes, on average, to being a complete graph.

$$coef = \frac{1}{n}\sum_{i=1}^{n}\frac{C_i}{N_i(N_i-1)} \tag{2}$$

where C_i represents the number of direct links between all the immediately neighboring nodes of a node i, and N_i is the number of neighbors of that node.

5.6 Number of clusters

Trivially, the number of disconnected clusters in a network.

5.7 Giant cluster size

The giant cluster of a network is defined as the disconnected cluster which has the maximum number of nodes in it. This measure is the number of nodes contained in that giant cluster.

6 Analysis of Measures

This section first considers several network measures introduced above, on each of the networks, and on each of the data sets. Figure 3 and Figure 4 comprises five separate plots - all being time-series of the same network measure. The five plots correspond to the five different networks constructed from the data: Organizations connected by common location; Locations connected by common organization; Organizations connected by common attack type; Organizations connected by target type; and Locations connected by attack type. On each plot three curves can be found: Clean - all records from the database for which there is an associated organization (previously referred to as KnownDB); I(srael) - a subset of the Clean dataset that is geographically Constrained to the locations Israel, West Bank, and Gaza Strip (KnownIDB); and NoI - the "Clean" dataset with all the I(srael) Data records removed (KnownNonIDB).Figure 3 plots the average clustering coefficient for the five networks, as a function of time. The clustering coefficient measures the degree of (local) interconnectedness between nodes that share a common neighbor. All plots show a sharp increase in coefficient value after 1988, but which reach a relatively stable value (i.e., don't systematically increase or decrease with time). In most cases the clustering coefficient is quite high - compared to other natural networks. In the case of Organizations linked by Location and vice-versa the coefficient value is around 0.5. Foreshadowing a later figure (number of clusters) we interpret this result to show that Organizations tend to have geographic bounds on their activities and that those same geographic bounds channel the targets of organizations operating in the same region into the same subset of countries - in a network sense there are a number of separate clusters, each with a relatively high clustering coefficient. When networks are linked by attack type or target type, the clustering coefficient is particularly high - approximately 0.8. Once again this appears to substantiate the interpretation that modus operandi - attack method, and target choice - are global in nature. That is; the degree of interconnectedness within these networks is extremely high and we know (from the following analysis) that there are very few disconnected clusters in these networks.

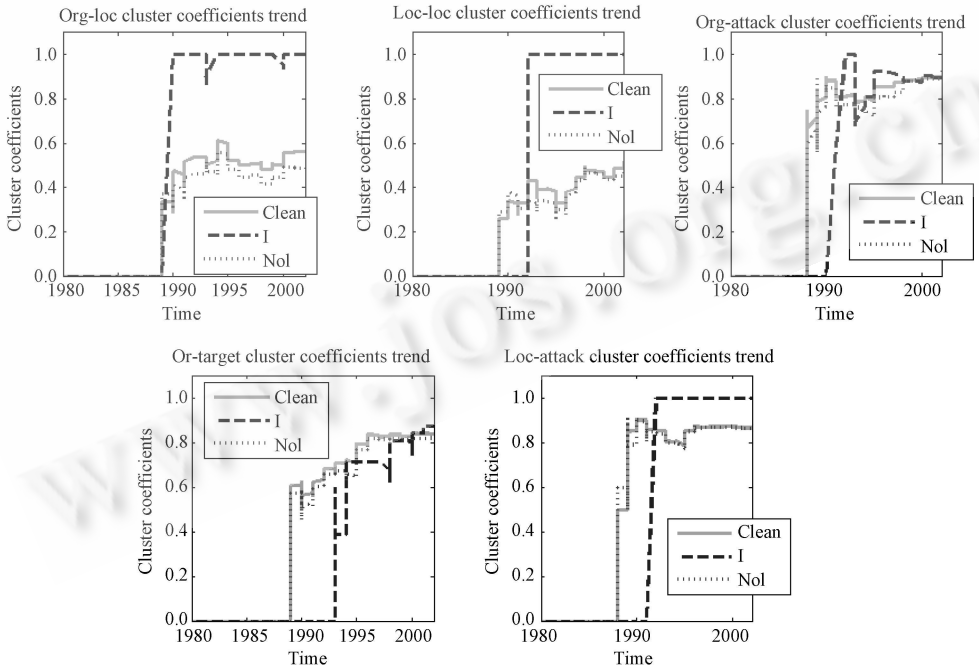

Fig.3 Clustering Coefficients as a function of time for the five networks; and three partitions of the data

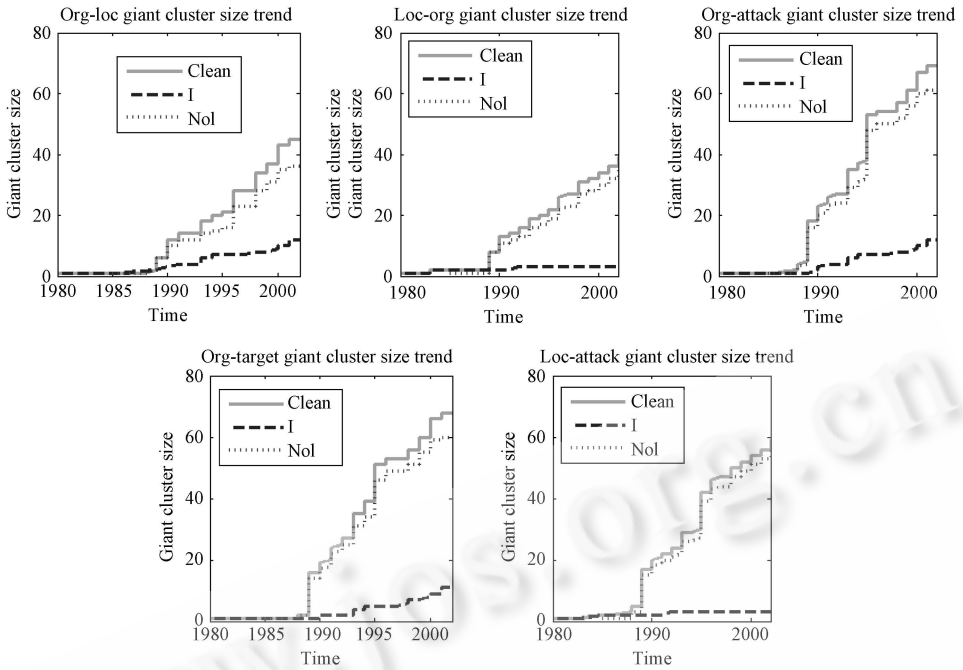

Fig.4 Giant Cluster Size as a function of time for the five networks; and three partitions of the data

Figure 4 plots the size (in terms of number of nodes) of the largest cluster (set of nodes that are connected) as a function of time. Contrasted with the total number of nodes, it shows what portion of the network is connected in a single "central" graph. From the previous figure we know that networks linked on the basis of attack type or target type are primarily a single cluster. Therefore we see the giant cluster size for such networks grow rapidly and automatically as each new organization or location is added to the dataset. The Organization by Location and Location by Organization giant cluster size also increase over time, but not as rapidly. Again, this is attributed to the fact that some organizations are single location specific; therefore they cannot join the giant cluster on the basis of organizations sharing a location or locations sharing an organization.

Figures 5 to 7 are 3D surfaces that show the degree distribution for all five networks - and the three variants of each - through time. That is, at each time step they plot the count of nodes with a certain degree of connectivity. This can be far more revealing than a single average-degree figure as we have a sense of the distribution of connectivity values; helping to address such questions as whether a network follows a small-world, random, or other such distribution. The six plots in each figure represent 2 perspectives on three different datasets: Clean - all records from the database for which there is an associated organization (previously referred to as KnownDB); I(srael) - a subset of the Clean dataset that is geographically Constrained to the locations Israel, West Bank, and Gaza Strip (KnownIDB); and NoI - the "Clean" dataset with all the I(srael) Data records removed (KnownNonIDB). In each plot, time in years is drawn along one axis, degree distribution along the the other; with the vertical axis being the count of the number of nodes that possess that degree distribution.

Figure 5 shows the degree distribution for the network in which nodes are terrorist organizations and connecting edges reflect a common location (country) of attack. It can be seen that the preponderance of nodes possess very low degree; although there is a shift to (average) higher degree as time progresses. This bears out the earlier analysis that there exist a number of unique organization-location pairings - a terrorist organization that acts in only one place, and that place possessing only a single terrorist organization - and hence a number of small, disconnected clusters,

plus the larger giant cluster. It also shows that there are a very few organizations that operate in locations in which many other organizations have operated. To some extent, this network exhibits scale-free characteristics.

Figure 6 shows the degree distribution for the network in which nodes are organizations and connecting edges reflect common attack types. It is evident that by the late 1990s nearly all organizations have a high degree of connectivity; reflecting a commonality of attack methods between the different organizations. Prior to that time an almost linear progression in the increase in degree can be seen for all organizations.

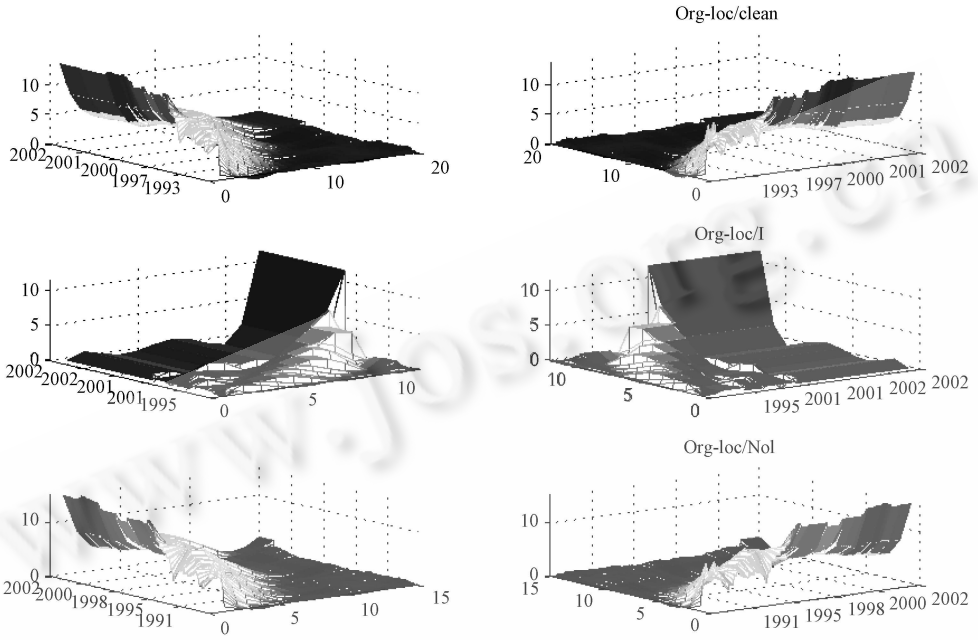

Fig.5 Degree Distribution of org-loc network as a function of time

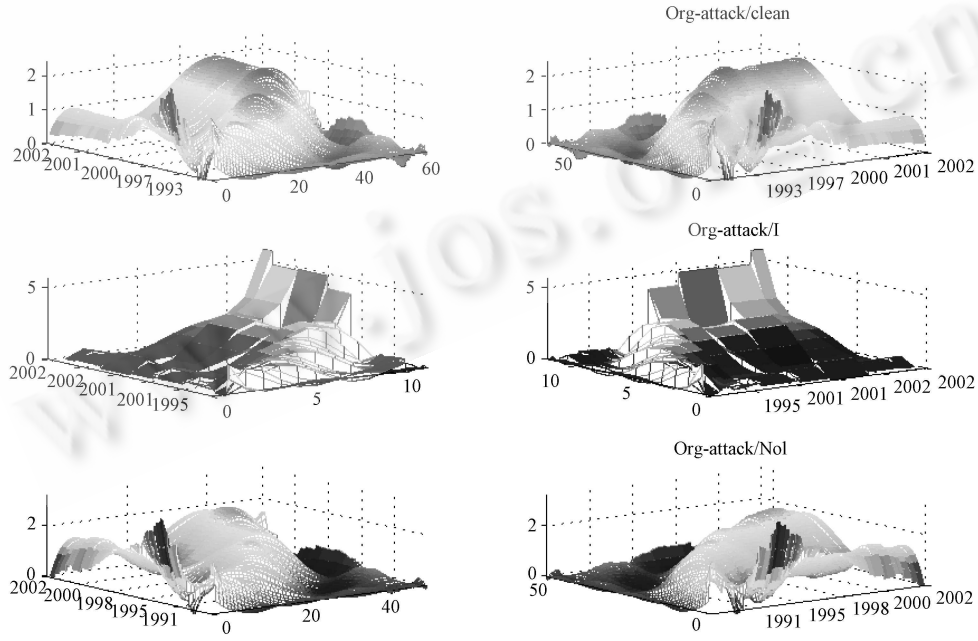

Fig.6 Degree Distribution of org-attack network as a function of time

Figure 7 shows the degree distribution for the network in which nodes are terrorist organizations and connecting edges reflect occurrences of the same target types for attacks. As for attack types, it can be seen that most organizations have a high degree reflecting a commonality in choice of target types between different organizations. Once again, stability is reached by the late 1990s with a rapid linear increase prior to that time. In order to better understand the spatio-temporal nature of the data, a number of movies were generated. Each movie shows a projection of the earth's surface; with terrorist attack activity being plotted based on the country attacked (as also done for the previous analysis). Successive frames in each movie represent the advance of time; typically one month; but in one case; each attack. In order to provide a consistent location for the country under attack, the CIA World Factbook[21] published latitude and longitude figures for countries were employed.

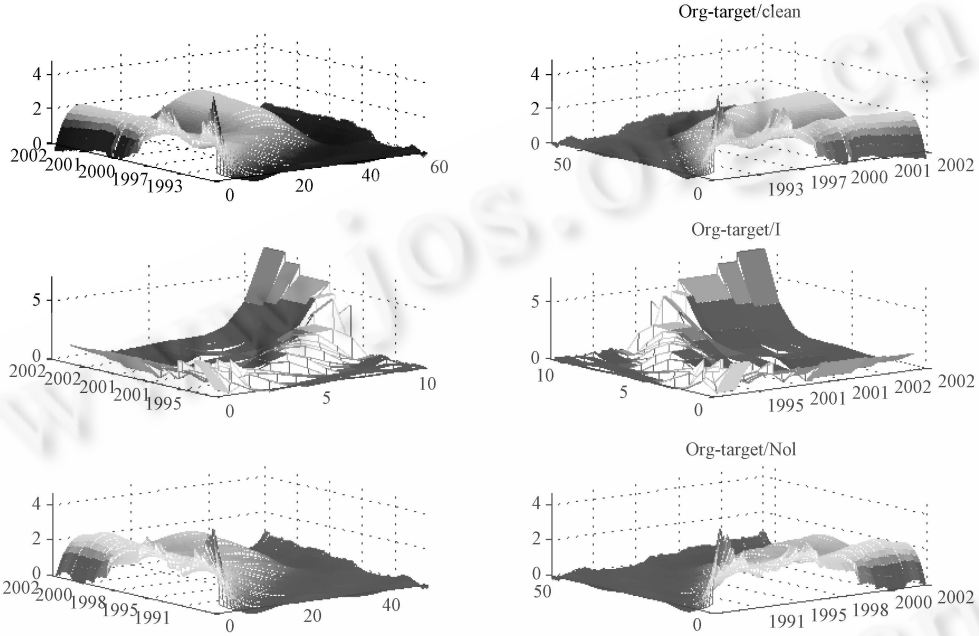

Fig.7 Degree Distribution of org-target network as a function of time

The movie temporalNetwork, a frame from which can be in figure 8, shows a different type of network - one in which locations (countries) are the nodes but edges represent temporal adjacency. That is, two countries are connected if a terrorist act in the first country is "immediately succeeded" (i.e., the next terrorist act to occur globally) by an act in the second country. As there are over 1400 acts of terrorism recorded in the database, this leads to a rather long movie. While it is hard to extract patterns from the animation, the very complexity of the graph again reveals the global and emergent property of terrorism - there appears to be no central coordinating authority that sequences the attacks either temporally or spatially. Further, this technique, of creating a graph based on temporal adjacency, appears to have considerable scope for other datasets in which there is a temporal dimension.

The movies varMovement and varMovementAve, a frame from one of which can be found in figure 9, show the movement in terrorist activity on a month by month basis. Hence, each frame in the movies correspond to one month. Attacks that took place within the time window are aggregated - the mean location (in latitude and longitude) is drawn as a red asterisk, representing a theoretical center to the activity, while the standard deviation in each of latitude and longitude is used to drawn a blue box centered about that mean location. Hence, the animations represent both the "average" or "center" of terrorist activity; and also its geographic spread. The movies differ, in that each frame in varMovement only plots terrorist attacks for the month in question; while varMovementAve keeps

a moving window of one year (of attacks) centered on the month in question. The later approach tends to smooth out "noise" in the data; better showing long term trends in the data, such as shifts in location or extensions to the area of activity. The movies clearly show the centrality of the Middle East in terrorist attacks, though periodic variability in scope is also visible. The movies CHMovement and CHMovementAve, a frame from one of which can be found in Fig.10, also show terrorist activity on a month by month basis. In this case a Convex Hull - a minimal polygon that encloses all locations - is drawn atop the map. Arguably, this provides a better feel for the geographic extent of attacks. As for the previous movies, these differ in that CHMovement considers only 1 month of data, while CHMovementAve provides a smoothing of the data by using a moving 1 year window centered on the month in question. Again, the global nature of terrorist attacks is clearly evident.

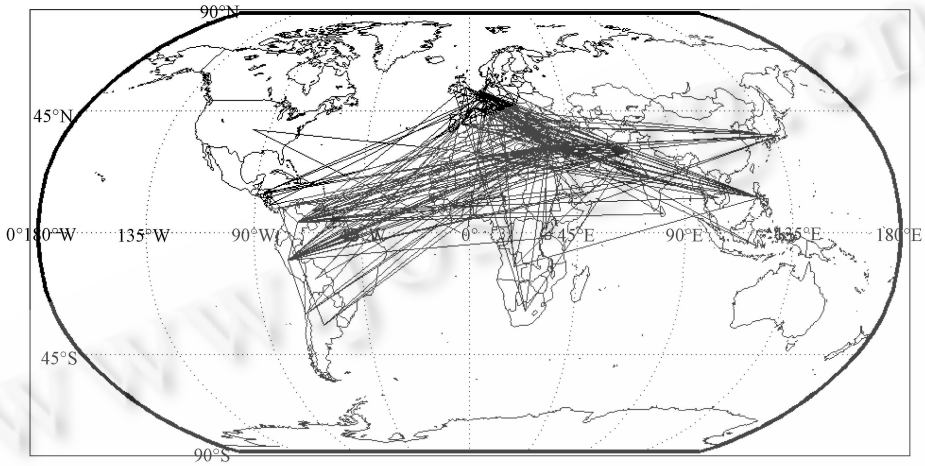

Fig.8 A frame from the temporal network movie, corresponding to the state of the network after the 500'th act of terrorism. The network is formed by connecting locations based on a temporal ordering of attacks - the location of the next (in time) attack is connected to the previous attack's location

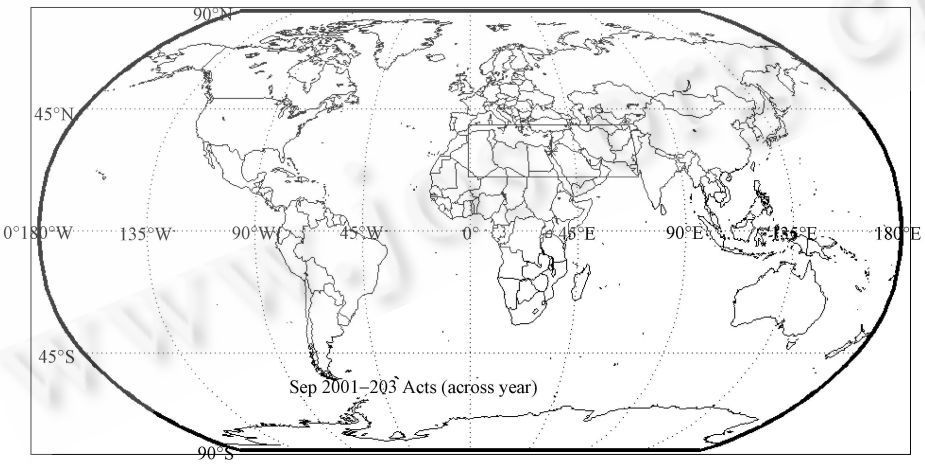

Fig.9 A frame from the varMovieAve movie. The frame shows the mean location (an asterisk) and variability (the surrounding box - 1 standard deviation) of attacks in a 1 year period centered on the month of September 2001. Each frame of the movie corresponds to an advance of one month

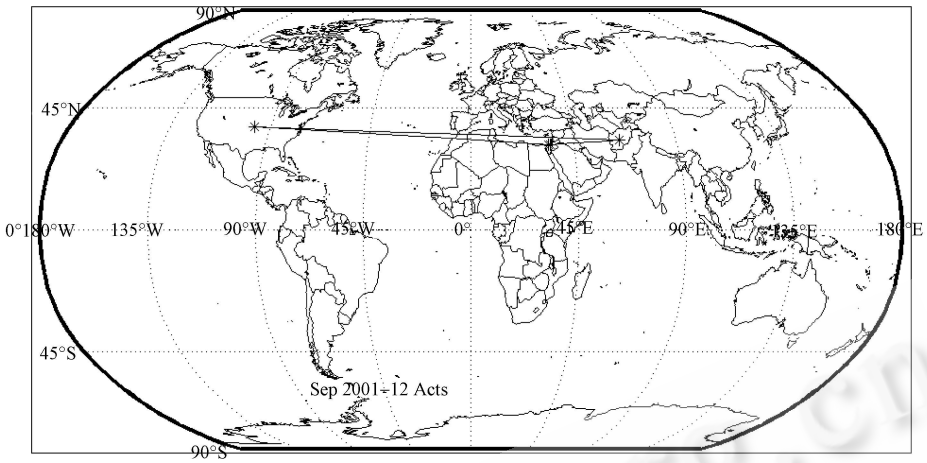

Fig.10 A frame from the CHMovement movie. The frame shows the convex hull - a polygon that encloses the region that terrorist attacks occurred in, in the month of September 2001. Each frame of the movie corresponds to one month

7 Conclusion

In this paper, we presented a short overview of the value added when analyzing network evolution for a terrorist data set. We also presented an overview of the system which we have developed for network visualization and analysis. In general, networked data mining has been shown to be a promising and potentially active area for research. The paper presented interesting patterns gleaned from the data. In particular, it has shown that choosing a highly interconnected network, such as X-attack, or X-target, can limit the effectiveness of certain network measures. In contrast, the less connected networks loc-org and org-loc exhibited more interpretive behavior.

In future research, a further investigation on the relationships among the data attributes used to construct the social networks, including “location”, “attack type”, and “organization”, will be conducted in order to disclose some correlations that may help in improving our network visualization and analysis process. Moreover, the other interesting research topic is to consider how the visualization and analysis method proposed in this paper can be applied to other types of large-scale networked datasets, such as microarray data, and genomic networked data, etc. Those datasets provide much more types and much larger volumes of data than terrorism dataset we used, and are more possible to discover more interesting patterns by doing social network visualization and analysis.

Acknowledgments This work is funded in part by the Key Program of the National Natural Science Foundation of P.R. China under Grant No.60723003, in part by the National Natural Science Foundation of P.R.China under Grant No.60505008, in part by the Natural Science Foundation of Jiangsu Province under Grant No.BK2007520, in part by the Natural Science Foundation of Jiangsu Province under Grant BK2006116, and in part by the Australian Research Council (ARC) Centre for Complex Systems under Grant No.CEO0348249. The author would like to thank Prof. Hussein A. Abbass, Dr. Michael Barlow, and Dr. Daryl Essam in UNSW@ADFA for their valuable suggestions and contributions to this manuscript.

References:

[1] Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: Towards a unifying framework. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. Advances in Knowledge Discovery and Data Mining. AAI/MIT Press, 1996.

1–36.

[2] Staab S. Social networks applied. IEEE Intelligent Systems, 1986,124:317–28.

[3] Kao H, Lin S, Ho J, Chen M. Entropy-Based link analysis for mining Web informative structures. In: Proc. of the 11th ACM CIKM 2002. New York: ACM Press, 2002. 574–581.

[4] Taskeru B, Wong M, Abbeel P, Koller D. Label and link prediction in relational data. In: Gottlob G, Walsh T, eds. IJCAI Workshop on Learning Statistical Models from Relational Data. Morgan Kaufmann Publishers, 2003.

[5] Barlow M, Galloway J, Abbass HA. Mining evolution through visualization. In: Beyond Fitness: Visualising Evolution, a Workshop at the 8th Int'l Conf. on the Simulation and Synthesis of Living Systems (ALife 8). MIT Press, 2002. 103–111.

[6] Albert R, Barabasi A. Statistical mechanics of complex networks. Reviews of Modern Physics, 2002,74:47–97.

[7] Albert HJR, Barabasi A. Diameter of the world wide web. Nature, 1999,401:130–131.

[8] Wu A, Garland M, Han J. Mining scale-free networks using geodesic clustering. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 719–724.

[9] Software SS. Sampling methods. http://www.statpac.com/surveys/sampling.htm, accessed 2006-08-12.

[10] Watters J, Biernacki P. Targeted sampling: Options for the study of hidden populations. Social Problems, 1989,36(4):416–430.

[11] Hussain A. Terrorist networks analysis through argument driven hypotheses model. In: ARES 2007: Proc. of the 2nd Int'l Conf. on Availability, Reliability and Security. Washington: IEEE Computer Society Press, 2007. 480–492.

[12] Yang CC, Liu N, Sageman M. Analyzing the terrorist social networks with visualization tools. In: Mehrotra S, *et al*., eds. Proc. of Intelligence and Security Informatics. Berlin: Springer-Verlag, 2006. 331–342.

[13] Lawrence S, Giles CL, Bollacker K. Digital libraries and autonomous citation indexing. IEEE Computer, 1999,32(6):67–71.

[14] Hill S. Social network relational vectors for anonymous identity matching. In: Gottlob G, Walsh T, eds. Workshop on Learning Statistical Models from Relational Data, IJCAI. Morgan Kaufmann Publishers, 2003.

[15] Popescul A, Ungar L. Statistical relational learning for link prediction. In: Gottlob G, Walsh T, eds., Workshop on Learning Statistical Models from Relational Data, IJCAI. Morgan Kaufmann Publishers, 2003.

[16] Keim D. Information visualization and visual data mining. IEEE Trans. on Visualization and Computer Graphics, 2002,7(1):100–107.

[17] Honkela T. Self-Organizing maps in natural language proc. [Ph.D. Thesis]. Espoo: Helsinki University of Technology, 1997.

[18] Chen J, Sun L, Zaiane OR, Goebel R. Visualizing and discovering web navigational patterns. In: Sihem AY, Luis G, eds., Proc. of the 7th ACM SIGMOD Int'l Workshop on the Web and Databases (WebDB 2004). New York: ACM Press, 2004. 13–18.

[19] Pampalk E, Goebl W, Widmer G. Visualizing changes in the structure of data for exploratory feature selection. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the SIGKDD2003. New York: ACM Press, 2003. 157–166.

[20] Goldberg H, Wong R. Restructuring transactional data for link analysis in the FinCEN AI system. In: Jensen D, Goldberg H, eds., 1998 Fall Symp. on Artificial Intelligence and Link Analysis. Menlo Park, California: AAAI Press, 1998. 38–46.

[21] Agency CI. The World Factbook. http://www.cia.gov/cia/publications/factbook, accessed 2006-02-12.

YANG Yu-Bin is an associate professor at the Department of Computer Science and Technology, Nanjing University, Nanjing, China and a CCF member. His research areas are networked data mining, social network analysis and content-based multimedia retrieval .

LI Ning is an associate professor at the Department of Computer Science and Technology, Nanjing University, Nanjing, China and a CCF member. Her research areas are data mining and machine learning.

ZHANG Yao a lecturer at Jinling Institute, Nanjing University, Nanjing, China. Her research areas are digital library and information retrieval.