Restructuring Transactional Data for Link Analysis in the FinCEN AI System*

Henry G. Goldberg and Raphael W.H. Wong**

U.S. Department of the Treasury, Financial Crimes Enforcement Network (FinCEN),
2070 Chain Bridge Road, Vienna VA 22182
**Current address: National Association of Securities Dealers (NASD) Regulation, Inc.,
9513 Key West Avenue, Rockville MD 20850
goldberh@nasd.com, wongr@fincen.treas.gov

Abstract

Due to the nature and costs of data collection, many real-world databases consist of large numbers of independent transactions. Finding evidence of structured groups of entities reflected in this data is a task aptly suited to Link Analysis. However, the databases usually must be restructured to allow effective search and analysis of the linkage structures hidden in the original transactions. The FinCEN AI System (FAIS) [Senator 1995] is an example of such an application. We briefly discuss the process of database restructuring and show how it is used to support the discovery and analysis of evidence of money laundering in a database of cash transactions.

Introduction

Transactional Databases

In our modern world, much of human activity is initially recorded in terms of individual transactions. Both the government and commercial sectors produce tremendous quantities of transactional data, often with little additional analysis being done at the time of collection. Interpretation, search, organization, even error correction and re-formatting may be left for the analyst at some later time. This may be due to the cost of doing such analysis on all the data, or it may be that the particular questions to be answered were not known when the data collection system was designed. Legal and regulatory prohibitions may also stand in the way of collecting certain data about every actor, indiscriminately.

In any case, we are often faced with a large number of relatively uninteresting transactions, with little data in each that might be used to select out the relevant ones for further analysis. In situations where we are looking for fraud in transactional data, the fraudulent activities are likely to be camouflaged to look like normal activities. The actors involved may spend much of their time in normal, uninteresting, activities and only occasionally in fraudulent ones. [Goldberg 1997] It is clear to many analysts who work with this data that the inter-relationships among transactions hold the key to select the proper subset and to understand the implicit activities hidden in the data -- activities which are incompletely reflected in the recorded transactions. The analysis of these relationships, called Link Analysis, is a vital technique used by law enforcement and intelligence analysts the world over. [Andrews 1990]

Finding Hidden Structure

Statistical and other data mining methods (which often are formulated to model and characterize populations of similar instances) can analyze sets of transactions to show consistent or novel relationships among data within a single transaction or among subsets of transactions which are clearly identified by inherent identifiers. However, finding structural relationships among transactions, which reveal the structure of the enterprises or organizations involved, requires that the database be restructured to make these relationships more explicit. [Goldberg 1995] Once restructured, the database may be queried using a number of well known approaches for on-line analytical processing (OLAP). The costs of restructuring, borne once, can then benefit a number of subsequent analyses.

We will discuss three of the most common restructuring processes, which are employed by FAIS in restructuring the Bank Secrecy Act (BSA) data for analysis, **disambiguation**, **consolidation**, and **aggregation**. We will then describe some of the specific challenges of doing this with the BSA data, and finally, we will demonstrate how the restructured database supports two levels of link analysis for the detection of money laundering and other financial crimes.

---

* The authors of this paper are (or were) employees of the Financial Crimes Enforcement Network of the U.S. Department of the Treasury, but this paper in no way represents an official policy statement of the U.S. Treasury Department or the U.S. Government. The views expressed herein are solely those of the authors. This paper implies no general endorsement of any of the particular products mentioned in the text.

Transforming Transactions into Links and Nodes

Nodes: Finding Relevant Entities

An essential, initial activity in preparing data for link analysis is to identify relevant entities and actors in the transactional data. Applying domain specific knowledge is very important for this task. For example, in a database of medical transactions, every unique individual must be considered a separate entity. However, in a database of commercial transactions, families or businesses may represent the correct level of granularity for analysis. Representing the real-world entities which underlie the transactions, and which are reflected in various identifiers in the data, is the role of **consolidation**. Correcting for errors (and intentional misrepresentations) in this data is the process of **disambiguation**.

These two activities are closely related and can be quite resource consuming. For example, in [Stolfo 1997] a process called "merge-purge" is employed to disambiguate identifiers on credit card transactions. No single identifier is relied upon exclusively, but a distance measure involving a number of identifiers is used through a cost-effective means of cross-comparing limited sub-sets of transactions [Hernandez 1995]. The resulting sets of records can then be consolidated into a profile of activity that can be analyzed for the likelihood of fraud.

In cases where relatively error-free identifiers are available in the data, it is still necessary to consolidate multiple transactions in order to evaluate the activities of the identified entities. Once the sets of records are consolidated, straightforward **aggregation** functions such as sums and averages, as well as more complex functions (e.g. determination of the most likely spelling of a name given a number of near misspellings), may be easily applied to produce a representative profile of the consolidated records.

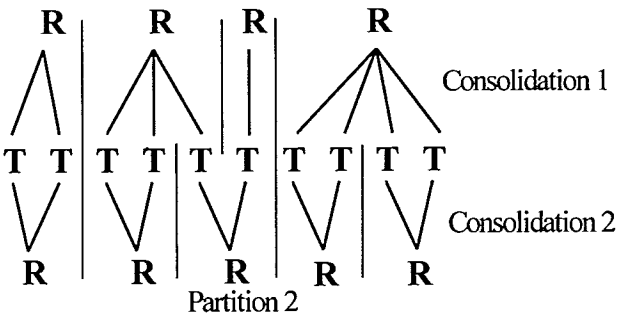

Figure 1: Multiple Consolidations

We should note that the activities of consolidation and aggregation are essential in many domains which deal with multiple occurrences of similar activities. For example, in a task domain of learning abstract concepts from robot sensor data, [Rosenstein 1998] refers to clustering instances into categories, and abstracting from categories to form prototypes. The critical point about data for link analysis, is that the consolidation and aggregation must not replace the individual transactions, since they provide the relationship information which will eventually connect the consolidated entities. In Figure 1 (from [Goldberg 1995]) consolidation on transactions, T (defined to have two actors), has identified a set of entities, R, which may be linked together by the transactions in which they "participate". Domain specific knowledge will indicate whether the sets are disjoint (e.g. depositors and accounts), intersecting (e.g. doctors and patients)., or identical (e.g. traders in a stock market).[1]

Links: Finding Meaningful Relationships

In order to support effective link analysis, linkages between entities should represent meaningful relationships as defined by the domain. Thus, linking an individual with an account into which he makes large cash deposits is meaningful for the task of finding money laundering, while linking him to the other depositors at that bank is probably not. This is domain specific -- apparent only from our understanding of how banks (and money laundering) operate.

Transactions in databases generally refer to entities (actors) by a fixed set of identifiers. They may be viewed as vectors of identifiers of entities and transaction details, as in:

T=[i1,i2,...,j1,j2,...,t1,t2,...]
where i1, i2,... are attributes of entity i;
j1, j2,... are attributes of entity j;
and t1, t2,... are attributes of the transaction.

disambiguation may then be seen as a clustering operation in the space(s) of attributes, I, J, etc. I may be identical to J if the transactions relate entities from the same population, e.g. in telephone calls between two phone numbers, of they may be separate, e.g. deposits by a person into a bank account. Such a transaction is seen as providing primary evidence for a link, L[i,j]. It is the realization of these links in the supporting database which allows rapid search and display of the linkage structures inherent in the transactional data.

The transaction above also provides evidence for additional links, since the association of attribute values is a relationship which may be of secondary interest to the analyst as well. For example, after finding a set of bank deposits which relate two individuals to some accounts owned by a business, further analysis might find that they occasionally use the same social security number. This may be a case of the same person using an alias, or of two people using one id. Knowledge of the transaction collection methods will help to determine the *a priori* likelihood of each. Finer grained link analysis will show

[1] Consolidation may also be viewed as a kind of link analysis, with an identity match providing a link between transactions. The consolidated set of transactions is the transitive closure over these links.

these associations, and may be of interest in evaluating the network. We will see that these two levels of granularity of linkage are used by FAIS to support first search and retrieval and then detailed analysis of networks.

The FinCEN AI System: Cash Transactions to Money Laundering Enterprises

The FinCEN AI System was designed to exploit the Bank Secrecy Act data to support the detection and analysis of operations engaged in money laundering and other financial crimes. We will provide a brief description of the problem, the data, and the database. We will then discuss how the data is restructured for link analysis and how link analysis is performed. Finally, we will mention some challenges which are seen as areas for future improvements in the system.

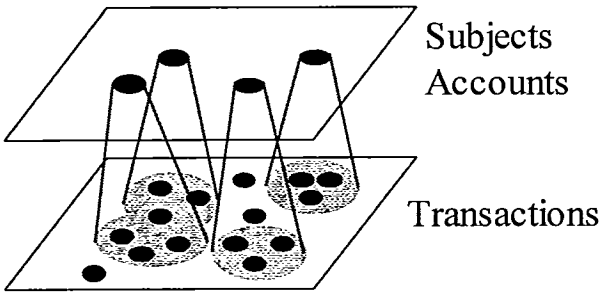

Figure 2: Clusters of Transactions

Problem Description

Money Laundering is a complex process of placing the profit, usually cash, from illicit activity into the legitimate financial system, with the intent of obscuring the source, ownership, or use of the funds. Money laundering typically involves a multitude of transactions, perhaps by distinct individuals, into multiple accounts with different owners at different banks and other financial institutions.

Generally, money laundering schemes usually involve two phases: **placement** and **layering**. Money is placed in various bank accounts, "spent" at casinos and other businesses, and used to "purchase" phantom goods. Financial transactions are often **structured** to avoid the reporting rules. Multiple stages of money transfer are employed to disguise the true origin and ultimate ownership (layering). Money is transferred among accounts and in and out of the country by a variety of instruments, including checks, wire transfers, and smuggled currency. The ownership of these accounts is hidden through a variety of means. Names and locations of businesses involved are continually changed. Prior to strict enforcement of the reporting laws, bank officials often cooperated wittingly or unwittingly in these schemes. There is still the danger that legitimate financial (and other) businesses will be suborned by the criminal organizations behind these activities.

The BSA Data. To combat money laundering, the BSA requires reporting of cash transactions in excess of $10,000. In addition, related transactions which in aggregate exceed the limit and any other transactions which appear suspicious to the bank officials are to be reported as well.

This record keeping preserves a financial trail for investigators to follow and allows the Government to systematically scrutinize large cash transactions. FAIS accumulates about 12 million transactions per year. The majority are CRTs (Currency Transaction Reports) filed by banks, credit unions, etc. A small number of related reports are filed by gambling casinos. Another, smaller number of CMIR reports (Cash and Monetary Instruments) are accumulated. These are filed by anyone entering or leaving the USA with currency or other negotiable instruments valued at $10,000 or more. The data reported on the forms are subject to errors, uncertainties, and inconsistencies that affect both identification and transaction information.

These three types of reports are designed to draw a protective curtain around the legitimate financial industry, protecting it from most sources of undocumented large amounts of cash. Since most criminal activities depend upon cash at the street level, it was believed that this data would suffice. While other channels are available for moving cash (and new ones are coming in various forms of electronic currency and the Internet), these data instruments have served to capture some "footprints" of many money laundering operations.

Consolidation and Aggregation in the FAIS Database. The FAIS database consists of individual transaction records of the three form types mentioned above. Subject (people and businesses) and bank account records are created during the process of disambiguating input identifiers and consolidating transactions with those of prior subjects and accounts. This operation results in consolidated sets of transactions (called **clusters**), illustrated in Figure 2, with associated collections of

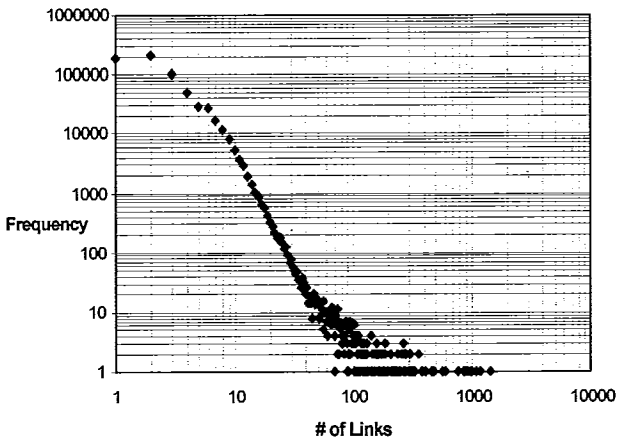

Figure 3: Cluster "Fan-out"

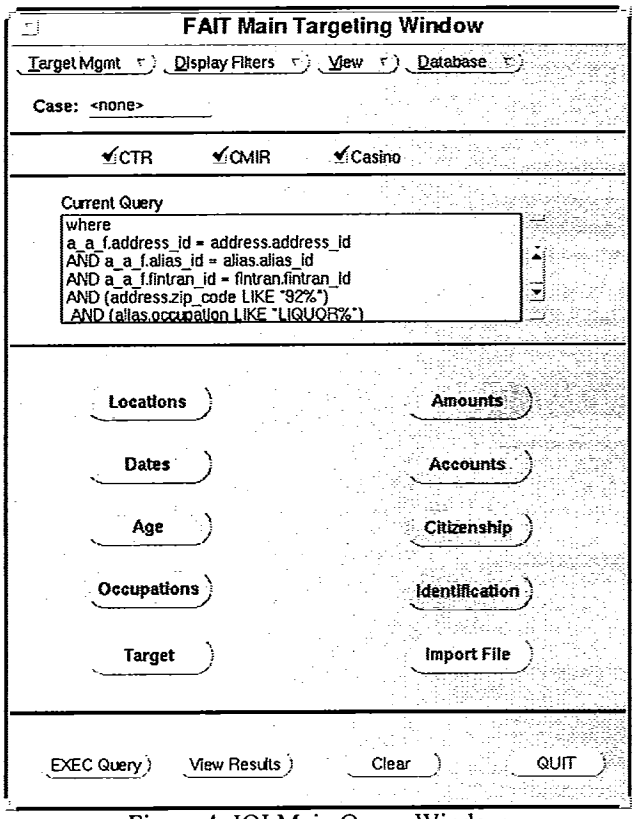

Figure 4: IQI Main Query Window

identifiers, aggregated money and transaction data, and the linkage of subjects and accounts to individual transactions (and thence to one another). The disambiguation process is tuned to be very conservative, since it is easier to aggregate further, if needed, than to split clusters that mistakenly combined separate entities. This consolidation process transforms the database from a purely transactional one to a form which can support direct link analysis via (efficient) database operations.

Figure 3 shows the distribution of cluster "fan-out" - the number of clusters immediately linked by one transaction to each cluster. This distribution is an estimate of the size of the networks which are available for analysis, although conservative disambiguation tends to make these smaller than their true size, and as we will show, we rely upon human intervention to gather additional networks together. It may be seen that the distribution is highly skewed, with a few clusters (primarily businesses engaged in largely cash sales, such as supermarkets) linking to large numbers of others (usually employees and bank accounts of many separate retail branches). It is tempting to try to eliminate this data from the input, however, due to regulatory as well as practical reasons, this has not been done.

The Role of Link Analysis in FAIS

Link analysis plays a critical role in FAIS, both in network acquisition and analysis. In this section, we discuss the role of link analysis in identifying relationships, exposing structure of organizations, and characterizing the roles of key actors. We describe two levels of link formation which enable these analyses to be performed over a huge database. Finally, we illustrate the process of finding and analyzing a network in FAIS through a series of illustrations.

Identifying Relationships. Money laundering is based on relationships and the methods for hiding them from the knowledge of law enforcement. Disentangling the results of layering requires that we detect and represent the relationships hidden in the transactional data. These relationships include both the channels for movement of money, and the ownership and operational relationships among entities in the organization. Identifying the money transfer relationships allows users of FAIS to infer ownership and operational relationships (e.g. common depositor, joint ownership, multiple businesses owned by a single entity, etc.)

Key Actors and Roles. Certain entities in the network may play essential roles in its operation, yet be hidden from immediate access in the transactions. Roles such as courier, pass-through account, or shell corporation are more readily seen in the structural map of the organization than in the raw data. Furthermore, poor quality disambiguation (due to the difficulties of dirty data, intentional misidentification, and missing transactions) can be corrected by the analyst when common relationships are seen to imply common identity.

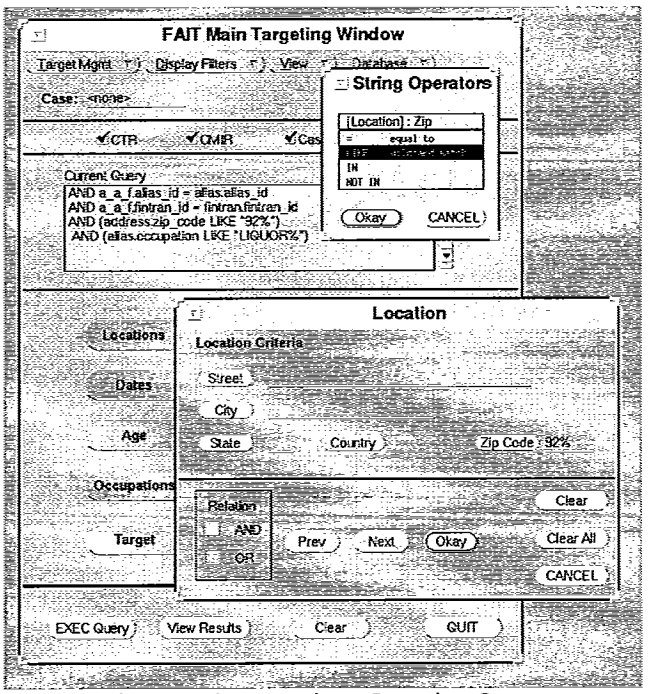

Figure 5: Constructing a Location Query

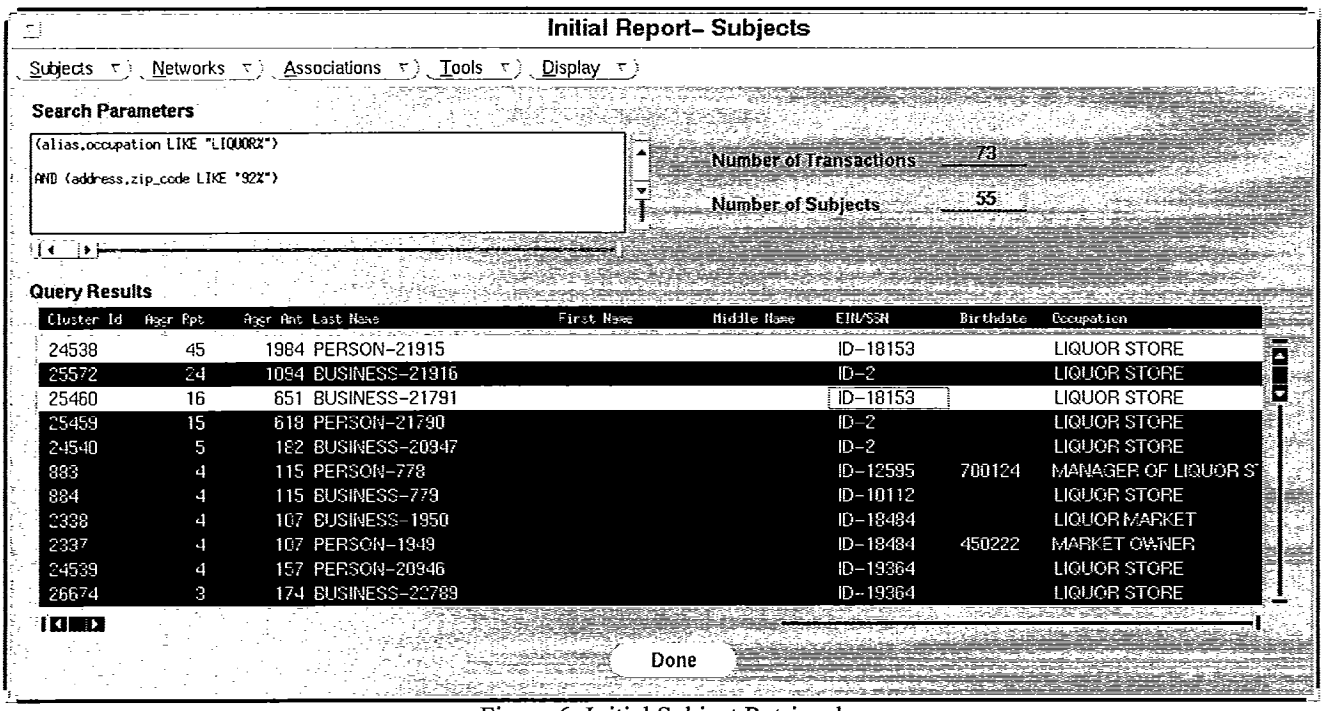

Figure 6: Initial Subject Retrieval

Structure of Enterprises. The overall structure of an organization can explain a great deal about its operation. recognition of this structure and the roles of the individual entities. (While most of the illustrations in this paper are taken from a "scrubbed", demo database, Figure 13 is taken from a real money laundering case. It is included to give an accurate feel for the link analyses which FAIS supports.)

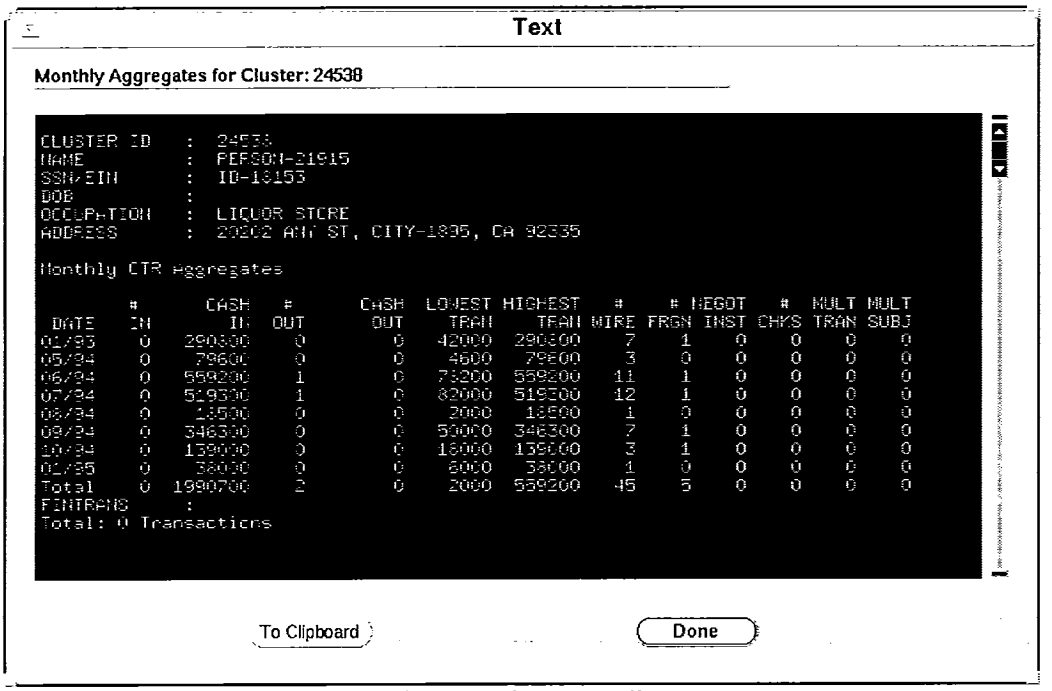

Figure 7: Cluster Details

In the final figure we see a large link diagram taken from a case involving many "legitimate" businesses funneling laundered money through a single account. The relationship of the two sets of businesses and individuals on either side of this account becomes clear through the

Two Levels of Link Analysis

Database Restructuring for Node Evaluation and Rapid Search. As we have described above, the database of BSA transactions is restructured to consolidate transactions and link subjects and accounts. This restructuring makes explicit in the database the subject-account relationships inherent in the individual transactions (as well as relationships between multiple subjects such as owner-depositor, or multiple accounts). The system can make use of these relationships to provide rapid search and retrieval to the analyst. Additionally, consolidation links multiple identifiers (from many

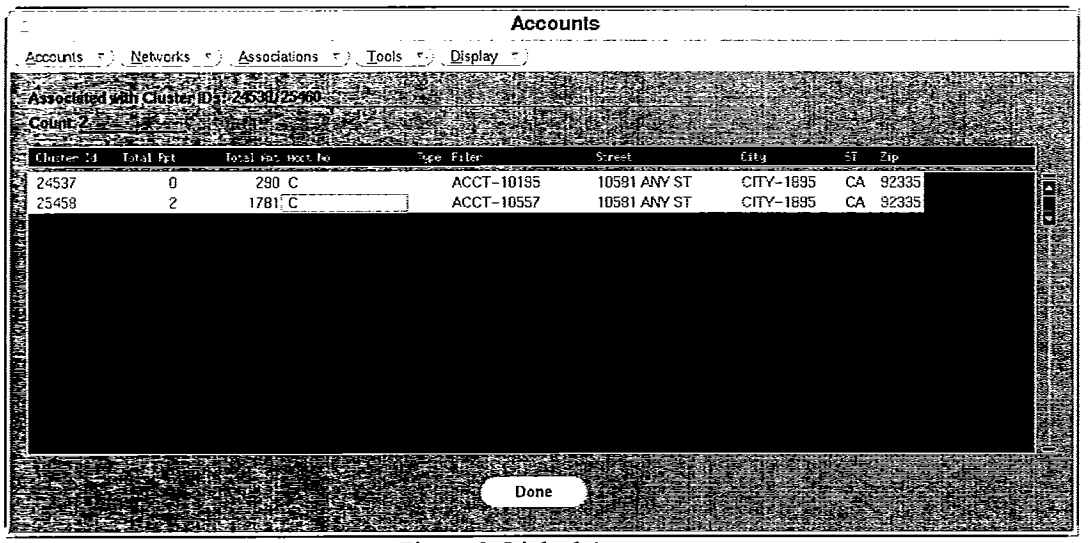

Figure 8: Linked Accounts

transactions) each subject or account, increasing the ability of the analyst to retrieve relevant data based on a structure query.

In addition to human evaluation of the clusters produced by this process, automated evaluation using rule-based inference is also implemented. This aspect of the system is discussed in greater detail in [Senator 95].

laundering is known to involve certain types of businesses. He can focus on this location, and specify certain types of businesses.[2]

The initial retrieval is directed at a set of subjects (or accounts) fitting the analyst's specifications. This set (Figure 6) may contain unrelated clusters. Also, important ones may be missing. However, the user can inspect the entities retrieved in detail (Figure 7), since the consolidation process associates aggregated data from all transactions in the cluster with the entity. After isolating those clusters which seem to be of interest, the analyst can use the IQI to widen the search by retrieving other clusters linked to the specified set by one or more transactions. Figure 8 shows all accounts linked to the highlighted subjects selected in Figure 6.

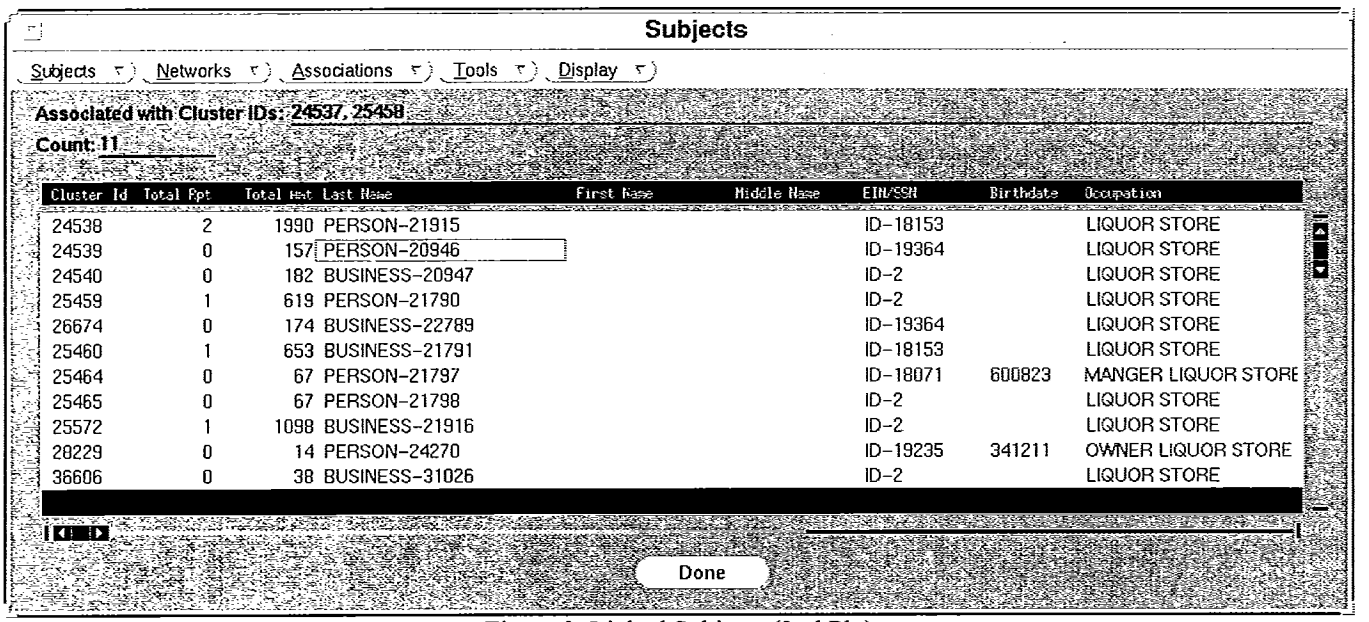

Figure 9: Linked Subjects (2nd Ply)

Using the IQI for Network Acquisition. FAIS supplies the user with an Interactive Query Interface enabling rapid search and retrieval of relevant data. Figure 4 shows the main IQI window. Queries based on a number of different types of identifier data are constructed (removing the need for analysts to learn complex SQL). For example, Figure 5 shows the construction of a location query. The analyst may have an interest in a particular location where money

This process may continue as long as necessary.

[2] The latter is rather unreliable, since the "occupation or business" field in the CTR is supplied by the transactor rather than the financial institution. However, these data problems may be compensated for by the consolidation and linkage searches implemented in FAIS.

(Transitive closure is usually reached in a few iterations.) At any ply, the user may search for subjects or accounts depending on their understanding of the structure of the network and data.[3] Figure 9 shows the set of subjects linked to the accounts of Figure 8. Since further searches yield no new clusters, the process is ended here and the second phase of link analysis is entered.

Detailed Link Analysis of Isolated Networks. A set of clusters may be selected and extracted for analysis in Alta Analytics' NetMap tool [Davidson 1993] using one of a number of prepared link analysis models. Since the database links all transactions to the clusters, and since clusters are accumulated in each ply of the IQI search process, the full set of transaction data may be specified in this way. We go back to the raw transaction data to acquire the full set of relationships -- both the transactional associations, e.g. person X deposited into account Y, and the consolidation associations, e.g. name A occurred with address B -- in a uniform model amenable to link analysis.

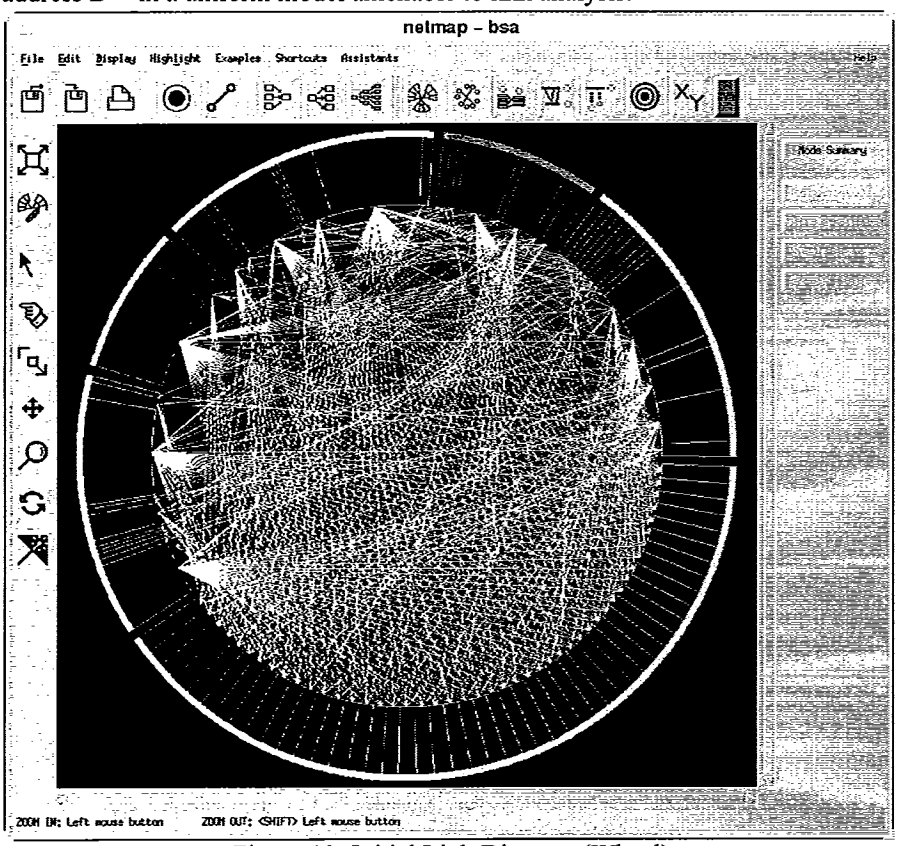

Figure 10: Initial Link Diagram (Wheel)

Figure 10 shows the initial NetMap display (wagon wheel format) of the data underlying the selected subjects from Figure 9. This display may be manipulated in a number of ways. For example, we remove the transactions, since we are more interested in the entities involved, and convert the display to a node and edge form familiar to many law enforcement analysts (Figure 11). This display is shown in a partial state of being manipulated to bring out the structural aspects of the organization behind these transactions. All underlying data which has been used to form linkages, and any additional data we feel may be useful is available for the analyst to drill down, and can be seen in closeup zoom (Figure 12).

An example of a finalized display is shown in Figure 13. It is drawn from a successful case involving a large money laundering operation. The role of the circled account is very clear from the structure of the network, as well as the relative roles of the businesses and individuals in the two separate sub-nets. Common features of the subsets of clusters (e.g. ethnic origin of the names of individuals and foreign destinations of money transfers) lead to further understanding of the hidden relationships of the organization thus uncovered.

Challenges

Among the challenges for improving FAIS, the following are especially relevant to link analysis and the role it plays in the system. While FAIS has been a great success in helping analysts in investigating leads or developing new leads and in analyzing complex criminal organizations, we have identified the following as high payoff or critical areas for additional effort.

Network Analysis. It is apparent that the final stage of link analysis in FAIS is still quite labor-intensive. Algorithmic support for graphical re-organization of networks would reduce analyst time significantly, allowing more time for investigative analysis. Other supporting link analysis "methods" are required as well including for example, search for similar sub-nets and automatic identification of "key" nodes. While the "man in the loop" mode of operation that FAIS implements has been one source of its success, analysts need more powerful tools to deal with this data in a timely manner.

[3] If the subject has a number of Casino or CMIR reports, which do not link to accounts, the analyst would probably search for more subjects at the next ply.

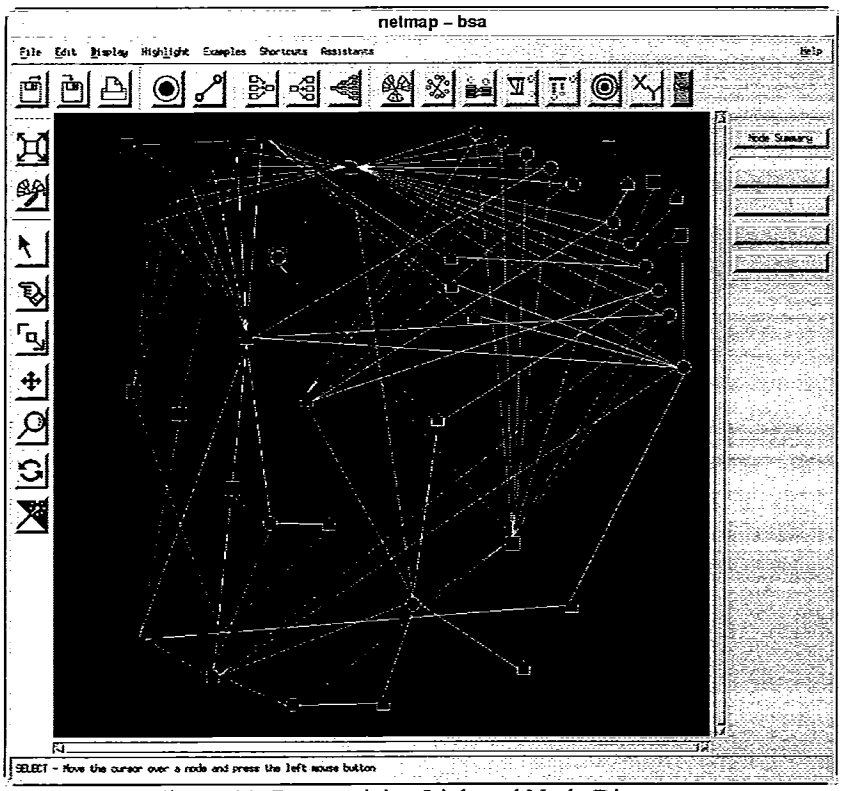

Figure 11: Reorganizing Link and Node Diagram

Processing Time. One challenge for FAIS is to reduce the search time necessary to build the clustered subject and account nodes. We have shown how, once built, these database structures enable the analyst to conduct their research on the BSA data in a more meaningful, abstracted space of entities and relationships. The cost of maintaining this transformed database as the data grows over time is so high that a number of simplifying assumptions have been made. The name and address matches have been decomposed into indexed database queries to be computable. Search techniques from CBR may be useful here. Richer, more knowledge-based[4] disambiguation, can be implemented in a practical system when better database search techniques allow more processing time to be allocated to this activity.

Merging Other Data with the BSA. Another challenge of interest to FinCEN is the need to cross reference the information in the BSA data with other sources of high value information. For example, the Suspicious Activity Report (SAR) record, a new data instrument now being collected since April 1, 1996, offers a potential source of linkages, associations, and relationships from information residing in different data sources. We see as critical the ability to identify and cross reference common entities, and merge the linkage data to allow analyses over the expanded networks. While the example mentioned above is an established data source, and a special application may be developed to support this merger, many sources of additional data are case-specific, e.g. seized records, the results of court ordered wiretaps, etc. An easy, uniform method for support these mergers is required. Such an approach might be based on an overall semantic description of the domain, in which various data sources can be modeled.

Temporal Link Analysis. The final challenge is our interest in finding temporal linkages in the data. Time is a key feature that offers an opportunity to uncover linkages that might be missed by more typical data analysis approaches. A simple example of telephone call records can make this point clear. Assume that we have the subpoenaed records of calls involving two telephone numbers, A and B. They may contain records of two other numbers, X and Y, which never contact one another. However, we might find from temporal analysis that there exists an inferred link by showing that a call from A to X is temporally proximate to a call from B to Y in a high percentage of instances. We may infer that the two are correlated and hypothesize a linkage between X and Y,

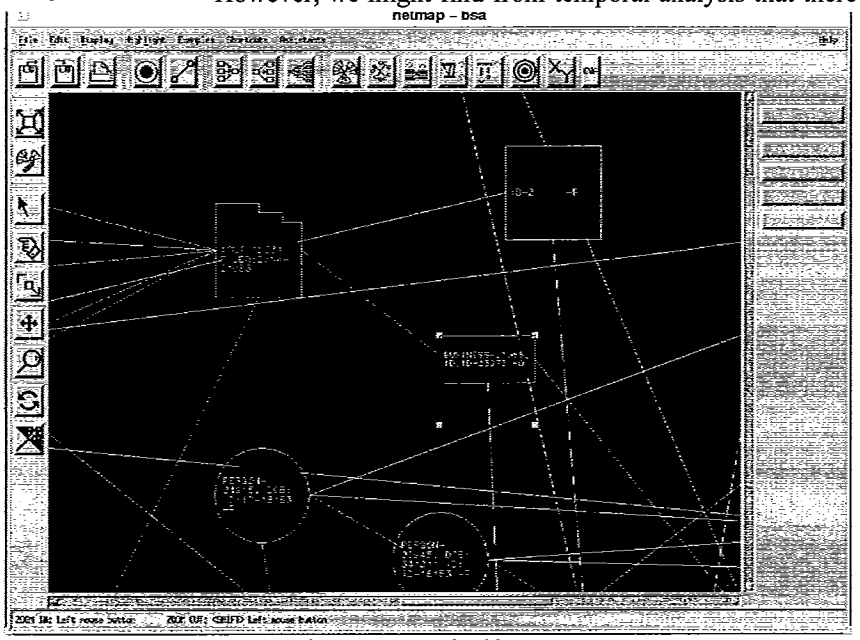

Figure 12: Node Closeup

[4] For example, a better understanding of the nature of ethnic name conventions and variations would improve subject disambiguation.

given our understood linkage of A and B. Another example might involve deposit to an account. Periodic deposits which fit the business cycle of the "legitimate" cover may be discounted, simplifying the analysis, while

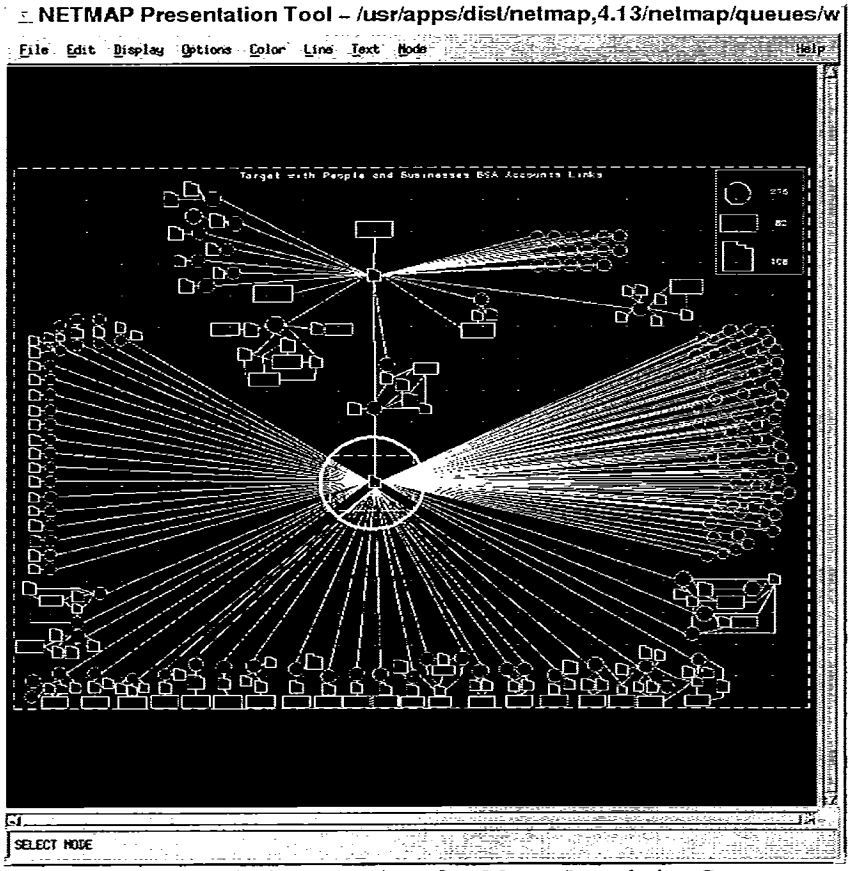

Figure 13: Final Presentation of an Money Laundering Case

deposits at odds with this cycle may indicate the flow of laundered funds (e.g. cash deposits by a ski resort in the summer).[5] These kinds of temporal analyses offer the opportunity for greatly improved detection of criminal activity, as well as the opportunity to filter out legitimate activities at an earlier stage in the process.

Summary

We have discussed the fundamental role of database restructuring and link analysis in detecting and analyzing complex activities hidden in transactional data. We described the processes of entity disambiguation and transaction consolidation and the utility of creating explicit linkages in the supporting database. We demonstrated how these processes are combined in FAIS to support search, retrieval, and analysis of complex networks of transactions and entities which are the database "footprint" of criminal organizations engaged in money laundering and other financial crimes. Finally we introduced some areas where algorithmic support would greatly improve the efficiency and effectiveness of the system -- more powerful network analysis tools, better and faster consolidation, merger of other data into the model, and temporal analysis of links.

[5] Or they may indicate some other unusual activity. One story of such an analysis involves the observation of a great deal of cash being transferred from banks in a particular city over one weekend. Observed and investigated, the reason turned out to be the Super Bowl.

References

Andrews, P. P. and Peterson, M. B. eds. 1990. *Criminal Intelligence Analysis.* Loomis, CA: Palmer Enterprises.

Davidson, C. 1993. What Your Database Hides Away. *New Scientist* 1855:28-31 (9 January).

Goldberg, Henry G. and Senator, Ted E., "Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation," *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Montreal, August, 1995.

Goldberg, Henry G. and Senator, Ted E., "Break Detection Systems," *Proc. of Workshop on AI and Fraud Detection,* (at 14th National Conferemnce on Artificial Intelligence, AAAI-97), Providence, July 1997.

Hernandez M A and Stolfo S J (1995), A Generalization of Band Joins and The Merge/Purge Problem, Department of Computer Science, Columbia University, New York, NY.

Rosenstein, Michael T. and Cohen, Paul R., "Concepts from Time Series", *Proc. 15th National Conf. on AI*, July 1998.

Senator, T.E., Goldberg, H.G., et. al., "The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions," in *Proc. 7th Annual Conf. IAAI,* Aug. 1995.

Stolfo, Salvatore J., Fan, David W., et. al., "Credit Card Fraud Detection Using Meta Learning: Issues and Initial Results," Department of COmputer Science, Columbia University, New York, 1997.