Applying Data Mining in Investigating Money Laundering Crimes

Zhongfei (Mark) Zhang
Jingzhou Hua Ruofei Zhang
SUNY Binghamton
Binghamton, NY 13902-6000
(607) 777 2935
zhongfei@cs.binghamton.edu

John J. Salerno
Air Force Research Laboratory
AFRL/IFEA
Rome, NY 13441-4114
(315) 330 3667
John.Salerno@rl.af.mil

Philip S. Yu
IBM Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
(914) 784 7141
psyu@watson.ibm.com

ABSTRACT

In this paper, we study the problem of applying data mining to facilitate the investigation of money laundering crimes (MLCs). We have identified a new paradigm of problems --- that of automatic community generation based on uni-party data, the data in which there is no direct or explicit link information available. Consequently, we have proposed a new methodology for Link Discovery based on Correlation Analysis (LDCA). We have used MLC group model generation as an exemplary application of this problem paradigm, and have focused on this application to develop a specific method of automatic MLC group model generation using Hierarchical Composition Based (HCB) correlation in the LDCA methodology. A prototype called CORAL is implemented in the MLC group model generation application, and preliminary testing and evaluations based on a real MLC case data are reported. The contributions of this work are (1) identification of the uni-party data community generation problem paradigm, (2) proposal of a new methodology LDCA to solve for problems in this paradigm, (3) formulation of the MLC group model generation problem as an example of this paradigm, (4) application of the LDCA methodology to develop a specific solution to the MLC group model generation problem by introducing HCB correlation measure using fuzzy logic, and (5) development, evaluation, and testing of the CORAL prototype in a real MLC case data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining.*

General Terms

Algorithms, Measurement, Documentation, Design, Experimentation, Legal Aspects.

Keywords

Money Laundering Crimes (MLCs), MLC Group Models, Uni-Party Data, Bi-Party Data, Community Generation, Link Discovery based on Correlation Analysis (LDCA), CORAL, Histogram Based Clustering, Hierarchical Composition Based (HCB) Correlation, Timeline Analysis.

1. INTRODUCTION

Money laundering is considered as a major crime in criminology [12,31], and is identified as one of the top group crimes in today's society [5,12]. With the development of global economy, the increasing applications of the Internet, and the advancement of the e-business and especially the e-banking, it is predicted that money laundering crimes (MLCs) will become more prevalent, more difficult to investigate, and more detrimental to the healthy development of the economy and the stabilization of the financial systems [12,31]. With the threat of global terrorism, preventing money laundering becomes ever more important to stop the financing of terrorist activities.

The investigation of MLCs involves reading and analyzing thousands of textual documents in order to generate (or hypothesize) crime group models. With the generated models, the investigators are able to provide evidence to support the prosecution against the defendants, to identify other individuals who might also be involved in the crimes to bring them to justice, and to predict and prevent crimes with similar patterns from occurring [12]. At present, this model generation process is performed completely manually, and thus, is extremely expensive and very tedious and labor-intensive, typically costing at least several man-month effort to the government, hence making the investigation and prosecution a prolonged process. Consequently, it is highly desirable to automate or semi-automate this process of identifying suspected crime groups in MLC, so that the governmental manpower can be significantly saved, and the crime investigation and prosecution time may be significantly reduced, ultimately leading to effective stopping or suppression of the MLCs.

This Paper describes part of an on-going and government-sponsored research project involving multiple universities, industrial labs, and government agencies to develop a methodology and related tools of semi-automatic MLC group model generation and analysis. This research project consists of

an automatic component of model generation, and a manual component for user validation of the generated models based on the domain expertise, resulting in a semi-automatic approach. While user validation is an important and indispensable component to model analysis for effective crime investigation [4,12,8], this paper focuses on the automatic component of MLC group model generation, as this is where data mining techniques are applied.

The contributions of this work are (1) identification of the uni-party data community generation problem paradigm, where no explicit links between data items are available, (2) proposal of a new methodology on **L**ink **D**iscovery based on **C**orrelation **A**nalysis (LDCA) to solve for problems in this paradigm, (3) formulation of the MLC group model generation problem as an example of this paradigm, (4) application of the LDCA methodology to develop a specific solution to the MLC group model generation problem by introducing a hierarchical composition based (HCB) correlation measure using fuzzy logic, where the correlations among two persons are inferred from correlations among the timings of their activities, and (5) development of a prototype system called CORAL, and evaluation and testing of CORAL in a real MLC case data.

The paper is organized as follows. After this introduction section, we briefly review the related work, present the challenges this work faces, and identify a new problem paradigm. We then make the problem statement and discuss the data set given in this problem, based on which we propose a general methodology to address the community generation issue in the new problem paradigm. This is followed by a presentation of a specific method of solving for the MLC group model generation problem, which is an example of the identified new problem paradigm. Finally, experimental testing and evaluations are reported for a prototype system of the proposed method for MLC group model generation before the paper is concluded.

2. RELATED LITERATURE AND CHALLENGES

Data mining has been recently extended from the traditional structured data mining [6] to the unstructured data mining [15,21], and in particular, to the time-series data mining [3,16,26], text mining [9,15,24,21], and Web mining [2,36,20,27,11]. This work is about automatic generation of a community of data items in a particular application. Specifically, the project focuses on mining tagged free text data to generate MLC group model, a community generation problem in law enforcement application. Community generation, though there is extensive research in the literature in recent years, in general is still considered as one of open and challenging problems in data mining research [7,22,34].

Of the reported community generation efforts in the literature, all the work focuses on automatic generation of a community based on a binary relationship given between data items. Examples of these efforts include mining on Web community or topic related documents [2,36,20,27,11], collaborative filtering [30,33], and social network analysis [32]. All of these community generation efforts assume that the input data are *bi-party data*, i.e., there is explicit link information given between data items (e.g., Web links, user-item mappings, or scoring of items assigned by users).

In this research, we have identified a new paradigm of problems in which the data are given as *uni-party data*, i.e., there is no explicit binary relationship between the data items, but the goal is to generate communities based on a yet-to-be-determined binary relationship between the data items. In the MLC documents collected by the law enforcement agency, most of them contain only uni-party data, e.g., the monetary activities of a single person. An example of uni-party activity can be that Fred Brown took $950 cash from his bank account on Feb. 2, 1994 or John Smith purchased a used Honda with $1100 cash on Feb. 4, 1994; there is no explicit relationship between Fred Brown and John Smith reflected in the documents. Moreover, even if for some documents there might be explicit binary relationship available for the financial transactions (money sender and recipient relationship), due to the current technology limitation of Information Extraction (IE), the IE tool is not able to capture verbs from the text robustly, resulting in the explicit binary relationship in the text becoming unavailable. On the other hand, the generation of the MLC group models is essentially building up the communities of a group of persons based on certain relationships between the persons inferred from the documents. Hence, this is a typical uni-party data community generation problem. Another example of this type of problems is to generate communities of countries based on the smuggling activities of massive destruction weaponry between them from the news data. Here the smuggling relationships may not be given from IE or may not even be explicitly reported in the news, but a solution to the problem is to "infer" these relationships through the data to generate the communities of these relationships among a group of countries.

Another application of the problem paradigm is that the data per se are intrinsically uni-party data. Examples include the generation of network intrusion models from intrusion data records in all the nodes of a network; generation of a traffic accident correlation model from traffic record data monitored at all the locations in a traffic network. Note that problems in this scenario actually is a generalized problem of finding association based on "inferring" the implicit binary relationships among a group of the data items.

Regarding the application areas, while data mining techniques have been applied to many areas in research and commercial sectors [14], there is little work reported in the applications in the law enforcement community [1], and to our knowledge, no research has been done in the application of MLC investigation.

3. PROBLEM STATEMENT

The goal of automatic model generation problem in MLC investigation is to generate a community of data items, the MLC group model. Here the data items are those individuals involved and committed to a specific MLC being investigated. In law enforcement practice, an MLC group model is often referred to a group of people linked together by certain "attributes". These "attributes" typically are identified by the investigators based on their experiences and expertise, and consequently are subjective, and may differ in different MLC cases by different investigators.

Since in the literature, no one has addressed this problem before, we propose to use a certain correlation as the "attributes" for link discovery in order to build up the community for model

generation. The correlation defined will be problem-specific, and in this MLC group model generation problem, we have developed a new approach to defining and determining the correlation, which is one of the contributions in this work. Based on the correlation defined, we build a tool to formally construct an MLC group model as a graphic representation with the following information: (1) all the crime members of this group, (2) the correlation relationships between different group members, (3) the link or communication structure among the group members which may be used to infer the different roles every members play in the group (e.g., who is in charge of the group or who are the core members of the group), (4) all the financial transaction history of each member in the group, and (5) the personal information of each group member.

The input data to the MLC model generation problem are typically free text documents, and sometimes also contain tables, or other more structured data. The types of the data may vary from different sources, such as bank statements, financial transaction records, personal communication letters (including emails), loan/mortgage documents, as well as other related reports. Ideally, if the semantics of these documents were understood completely, the link discovery based on correlation analysis would become easier. However, the current status of natural language understanding is far from being able to robustly obtain the full semantics of the documents [25]; instead, what we are able to robustly obtain is through IE [13] to identify those key entities that are relatively easy to identify and extract in the text. The key entities typically include the four W's: what, who, when, and where.

In this project, we have a data set consisting of 7,668 free text, physical documents regarding a real MLC case provided by National Institute of Justice (NIJ). The documents are first converted to digital format under OCR, and then key entities are tagged in the documents using a BBN developed IE tool [10]. The tagged documents are represented as XML files. The tagged key entities include person names, organization names, financial transaction times and dates, location addresses, as well as transaction money amounts; no link information is tagged, making the whole problem a typical uni-party data community generation problem. Figure 1 shows part of a typical tagged document used as input in this project, and Figure 2 shows the goal of MLC group model generation using a hypothetical example. Note that the correlation between people in Figure 2 is not illustrated, and typically a model may be a general graph as opposed to a hierarchical tree.

4. GENERAL METHODOLOGY

Given the problem statement, the solution to the general MLC group model generation problem consists of two stages: text processing (including OCR conversion and IE tagging), and community generation. The output of text processing is the tagged text, i.e., uni-party data served as the input to the community generation. Text processing is not the focus of this project. In this section, we propose a general methodology, called Link Discovery based on Correlation Analysis (LDCA), as a solution to the general uni-party data community generation problem. LDCA uses a correlation measure to determine the "similarity" of patterns between two data items to infer the strength of their linkage, where the correlation measure may be defined in fuzzy logic to accommodate the typical impreciseness of the "similarity" of patterns.

Figure 3 shows the components of LDCA as well as the data flow of these components. In principle, LDCA consists of three basic steps. For each problem in the uni-party data community generation paradigm, assume that the data item set is *U*. *Link Hypothesis* hypothesizes a subset *S* of *U,* such that for any pair of the items in *S* there exists a mathematical function (or a procedural algorithm) *C* that applies to this pair of items to generate a correlation value in the range of [0, 1], i.e., this step defines the correlation relationship between any pair of items in *S:*

$$\forall p, q \in S \subseteq U, C : S \times S \to [0,1] \qquad (1)$$

Link Generation then concerns with applying the function *C* to every pair of the items in *S* to actually generate the correlation values. This results in a complete graph *G(S,E)* where *E* is the edge set of this complete graph with computed correlation values. Finally, *Link Identification* defines another function *P* that maps the complete graph *G* to one of its subgraph $M \subseteq G$ as a generated community. In the next section, we present a specific method of LDCA in the application of MLC group model generation.

5. LDCA IN MLC GROUP MODEL GENERATION

In this section we take the MLC group model generation as an example of the new uni-party data community generation problem paradigm, and describe a specific method to solve for this problem by applying the general LDCA methodology to the MLC investigation context. Below we follow the general steps of LDCA to present this method.

5.1 Link Hypothesis

The Link Hypothesis in the MLC group model generation problem states as follows:

- The data set *U* is the set of all extracted individuals from the collection of the given documents.
- For each individual, there is a corresponding financial transaction history vector (may be null) along timeline.
- The correlation between two individuals is defined through a correlation function between the two corresponding financial transaction history vectors.
- If two individuals are in the same MLC group, they should exhibit similar financial transaction patterns, and thus, should have a higher correlation value.
- Any two individuals may have a correlation value (including 0), i.e., *S* = *U*.

Since we only have access to the isolated, tagged entities in the documents, we must make an assumption to reasonably "guess" the associated relationships between the extracted time/date stamps and the money amount of a specific transaction with the extracted individual. Therefore, when we parse the collection of documents to extract the financial transaction history vectors for every individuals, we follow the following proposed *one way nearest neighbor* principle:

- For every person name encountered, the first immediate time instance is the first time instance for a series of financial activities; the second immediate time instance is the second time instance for another series of financial activities, etc.
- For every time instance encountered, all the following financial activities are considered as the series of financial activities between this time instance and the next time instance.
- Financial activities are identified in terms of money amount; money amount is neutral in terms of deposit or withdrawal.
- Each person's time sequence of financial activities is updated if new financial activities of this person are encountered in other places of the same document or in other documents.
- The financial activities of each time instance of a person is updated if new financial activities of this time instance of the same person are encountered in other places of the same document or in other documents.

```
The issuer of the Letter of Commitment is a Major US and
International Securities Firm. The firm has confirmed and verified,
with our escrow company and Mr. <ENAMEX TYPE="PERSON">Richard
Alan</ENAMEX>, attorney at law, that they are the issuer of the
Letter of Commitment to the <ENAMEX TYPE="ORGANIZATION">George Creek
Company, Inc.</ENAMEX>, in the amount of <NUMEX TYPE="MONEY">Three
Million Five Hundred Thousand dollars</NUMEX>. The commitment was
issued on <TIMEX TYPE="DATE">October 23, 1992</TIMEX>.

The firm also confirmed on <TIMEX TYPE="DATE">October 23,
1992</TIMEX>, to underwrite for <ENAMEX TYPE="ORGANIZATION">George
Creek Company, Inc.</ENAMEX>, the sum of <NUMEX TYPE="MONEY">Three
Million Five Hundred Thousand dollars</NUMEX>. The firm verified
that one of their very substantial clients requested, issuance
of the Letter of Commitment and underwriting, to <ENAMEX
TYPE="ORGANIZATION">George Creek Company, Inc.</ENAMEX>,
under terms and conditions of an established business relationship,
their client has with <ENAMEX TYPE="ORGANIZATION">George Creek
Company, Inc.</ENAMEX>
```

Figure 1. A part of an input XML document. Note that there is no direct, explicit link information given in the tagged data.

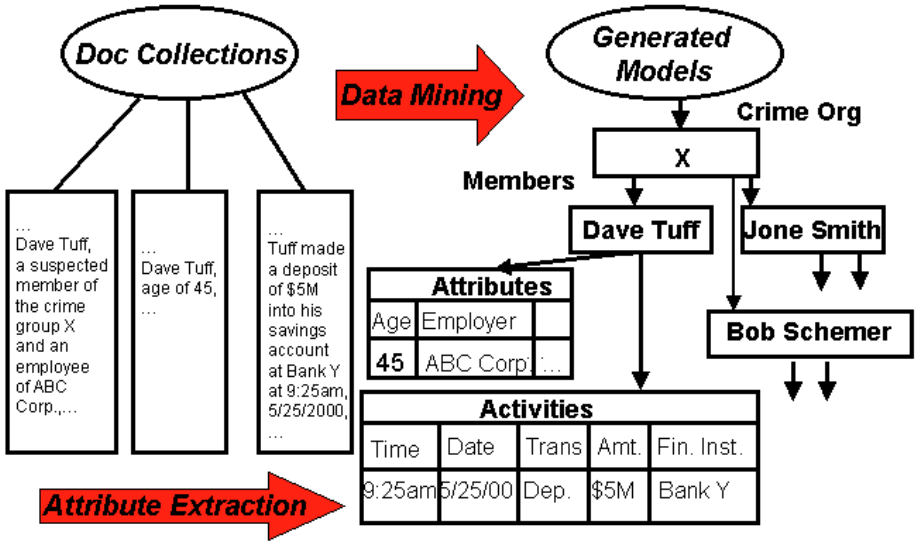

Figure 2. A hypothetical example of the automatic MLC group model generation.

Based on this parsing principle, we define and generate an event-driven, three-dimensional, nested data structure for the whole data set *U*: whenever a new individual's name is encountered, a new PERSON event is created; whenever a new time instance is encountered, a new TIME event is created under a PERSON event; whenever a new financial transaction is encountered, a new TRANSACTION event is created linked to both corresponding TIME and PERSON events. All the events

are represented as vectors. Figure 4 illustrates the data structure. Based on this data structure, after parsing the whole collection of the documents, we map the whole data structure into a timeline map illustrated in Figure 5, where each timeline represents the financial transaction history vector of each individual. The time axis of the timelines is "discretized" into time instances. Each node in the timelines is called a *monetary vector* that records the part of the financial transaction history of the corresponding person between the current time instance and the next time instance.

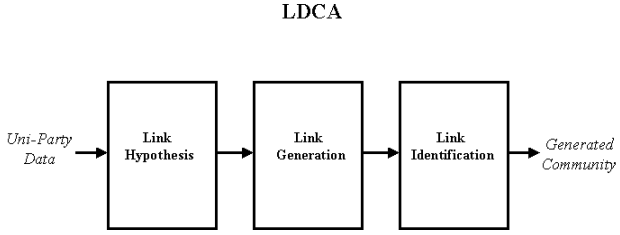

Figure 3: LDCA components and the data flow.

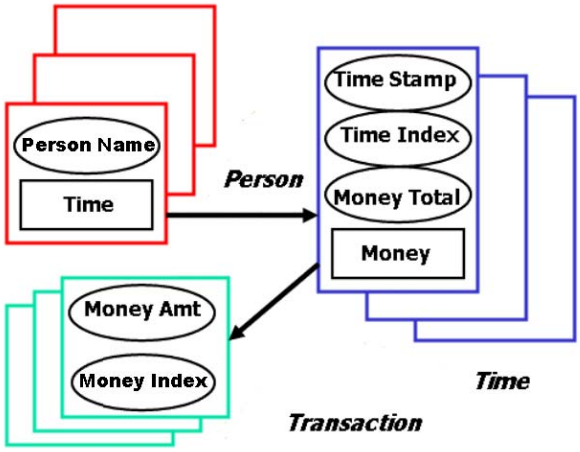

Figure 4: Event-driven, three-dimensional, nested data structure.

While the above "one way nearest neighbor" parsing principle may not be necessarily true in all the circumstances, we propose this principle based on the following two reasons. (1) this is the best we can do with the absence of the actual association information in the data, and (2) the experimental evaluations show that the generated models based on this principle are reasonably accurate.

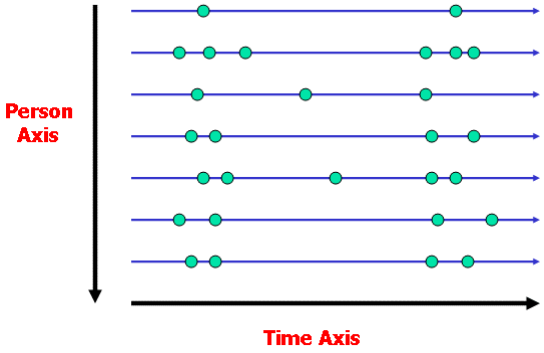

Figure 5: The timeline map from the three-dimensional, nested data structure; here each node is a monetary vector.

5.2 Clustering Transaction Activities

Given the generated timeline map, based on the Link Hypothesis, in order to accurately determine the financial transaction correlation between two individuals, ideally we wish to be able to determine which monetary vectors are "useful", i.e., they are truly related to the money laundering case being investigated, and which are just noise (e.g., a "normal" financial transaction of an individual such as a "normal" purchasing activity, or a false association between one's monetary activity and someone else due to the one way nearest neighbor parsing principle). However, since we do not have the true semantics of the documents, this information is not available to us. Fortunately, during the data collection process (i.e., the law enforcement investigators manually attempt to collect all the documents that might be related to the case) the investigators typically have the intention to collect all the documents that are related to those either suspiciously or routinely related to the case; thus, it is expected that for those individuals who might be involved in the crimes, the majorities of their monetary vectors should be well clustered into several "zones" in the timeline axis where the actual MLCs are committed. We call this assumption as the *focus* assumption. Based on the focus assumption, we only need to pay attention to the "clusters" of the monetary vectors in the timeline map, and can ignore those monetary vectors that are scattered over other places of the timeline map. This allows us to maximally "filter" out the noise when determining the correlation between two individuals.

Assume that there are n individuals extracted in total. This clustering problem is then a standard clustering problem in an $n+2$ dimensional Euclidean space (n PERSON dimensions, 1 TIME dimension, and 1 TRANSACTION dimension). This problem may be solved through applying the standard K-means algorithm [14,23]. However, taking advantage of the fact that all the n individuals share the same timeline, we can further simplify this general $n+2$ dimensional clustering problem as follows.

When we discretize the whole timeline into different time instances, each monetary vector is viewed as a node in this one-dimensional timeline space. We first simplify the problem by collapsing all the monetary vectors into scalar variables w.r.t.

the accumulated transaction frequency for each monetary vector. This provides a timeline histogram of financial activities for each individual. We then project all these timeline histograms of individuals into a single timeline axis to form a composite histogram. Consequently, the clustering problem is reduced to a segmentation problem in the (composite) histogram [18]. Figure 6 illustrates how the composite timeline histogram is generated from all the monetary vectors along the timeline. Since the projection and the histogram segmentation may be performed in linear time in the timeline space [18], this clustering algorithm significantly improves the complexity and avoids the iterative search the *K*-means algorithm typically requires. The resulted number of "hills" (i.e., segments or time zones) in the composite histogram becomes the *K* clusters.

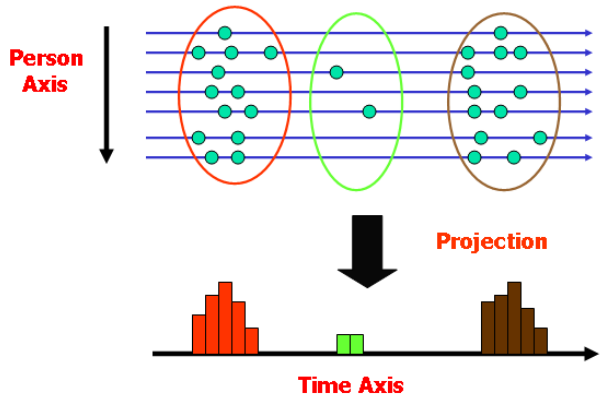

Figure 6: Clustering transaction activities based on histogram segmentation.

5.3 Link Generation Using HCB Correlation

After the clustering, each individual's financial transaction history vector may be represented as a timeline histogram partitioned into *K* clusters or time zones, which may in turn be represented as *K* histogram functions of time *t*: $\langle f_i(t) \rangle$, where $f_i(t)$ is the financial transaction histogram of this individual in cluster *i*. In order to capture the unique characteristics of the MLC data, we introduce a *hierarchical composition based* (HCB) approach to estimating the correlation between two individuals $\langle x, y \rangle$. This *HCB correlation* is defined as a combined global correlation of all the local correlations between the two individuals, whereas the local correlation is defined as the correlation between two clusters of the timeline histograms of the two individuals. Figure 7 illustrates the process of determining the global correlation from local correlations between two individuals *x* and *y*. The reason why the HCB correlation is defined here as this "two level" function is due to the unique nature of the problem --- individuals in the same MLC group may exhibit similar financial transaction patterns in different time "zones" (which is captured by the local correlation), but the difference in the timeline of their financial activities should not be too large (which is captured by the global correlation). While the local correlation is defined following a standard approach in Pattern Recognition literature to determine a fuzzified "similarity" between two functions [37], the global correlation here is defined based on the unique nature of this problem to further constrain the overall "similarity" between the financial transaction patterns along the timeline of two individuals' activities.

To define a reasonable correlation function, it is noted that the concept of similar financial transaction patterns is always fuzzy (e.g., if two individuals belong to the same crime group and are involved in the same MLC case, it is unlikely that they would conduct transactions related to the crime simultaneously at the exact time, nor is it likely that they would conduct transactions related to the crime at times that are a year apart; it would be likely that they conduct the transactions at two different times close to each other). Consequently, we apply fuzzy logic [35] in both definitions of the local and global correlations to accommodate the actual "discrepancy" of the occurrences in the extracted financial transaction activities between different individuals at different times. As we shall see later in section 6, applying the simple cosine function based correlation [29] on the timeline histogram of each individual will not be able to capture the link relationships and thus perform poorly.

5.3.1 Local Correlation

Let $fx_i(t)$ and $fy_j(t)$ be the financial transaction histogram functions of individual *x* and *y* in cluster *i* and *j*, respectively. Following the standard practice to define a fuzzified correlation between two functions [37], we use the Gaussian function as the fuzzy resemblance function *within* cluster *i* between time instance *a* and *b:*

$$G_i(a,b) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-\frac{(a-b)^2}{2\sigma_i^2}}. \quad (2)$$

$$\sigma_i = \frac{2}{W_i(W_i-1)} \sum_{a=1}^{W_i} \sum_{b=a+1}^{W_i} |a-b|. \quad (3)$$

where σ_i is defined accordingly based on the specific context in this problem, and W_i is the width of the cluster *i*.

The purpose of using Gaussian function is that it gives a natural decay over the time axis to represent the fuzzy resemblance between two functions [37]. Consequently, two transactions of two individuals occurred at closer times results in more resemblance than those occurred at farther away times.

It can be shown [37] that after applying the fuzzy logic using the Gaussian function as the resemblance function, the resulting fuzzified histogram is the original one convolved with the fuzzy resemblance function.

$$gx_i(t) = \sum_{t'=1}^{W_i} fx_i(t') G_i(t,t'). \quad (4)$$

Thus, the local correlation between $fx_i(t)$ and $fy_i(t)$ is defined as the maximum convolution value

$$g(x_i, y_j) = \max_{t=0}^{W_i} \sum_{t'=-W_j}^{W_j} gx_i(t') gy_j(t-t'). \quad (5)$$

We note that the local correlation is shift invariant. It is only used to measure the shape similarity of two histograms $fx_i(t)$ and $fy_i(t)$, where the implication on the time distance between the histograms is captured in the global correlation to be described next.

5.3.2 Global Correlation

Assuming the timeline axis is clustered into K time zones, based on the definition of the local correlation, for each individual x, at every cluster i, there is a set of K local correlations with individual y $\{g(x_i, y_j), j = 1, \ldots, K\}$. We give the fuzzy weights to each of the elements of the set based on another Gaussian function to accommodate the rationale that strong correlations should occur between financial transactions of the same crime group closer in time than those farther away in time. Thus, we have the following series:

$$\{g(x_i, y_j)\, S(i, j), j = 1, \ldots, K\} \quad (6)$$

where

$$S(i,j) = e^{\frac{(c_i - c_j)^2}{2\sigma_i^2}}. \quad (7)$$

and c_i and c_j are the centers of cluster i and cluster j along the timeline.

The correlation between individual x in cluster i and the whole financial transaction histogram of individual y is then defined based on the *winner-take-all* principle:

$$C(x_i, y) = \max_{j=1}^{K} \{g(x_i, y_j) S(i, j)\}. \quad (8)$$

Finally, defining the vectors

$$Cy(x) =< C(x_i, y), i = 1,..., K > . \quad (9)$$

$$Cx(y) =< C(y_i, x), i = 1,..., K > . \quad (10)$$

the global correlation between x and y is defined as the dot product between the two vectors:

$$C(x, y) = Cy(x) \bullet Cx(y) = \sum_{i=1}^{K} C(x_i, y) C(y_i, x). \quad (11)$$

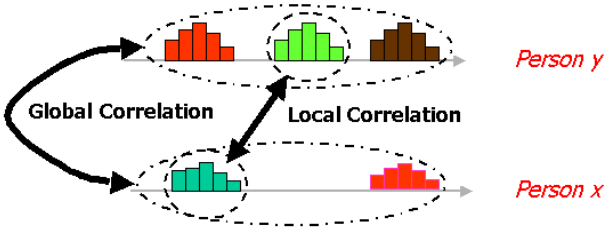

Figure 7. Illustration of the HCB approach to determining the correlation between two individuals *x* and *y*.

5.4 Link Identification for MLC Group Model Generation

After applying the correlation function to every pairs of individuals in the data set U, we obtain a complete graph $G(U, E)$, where U is the set of all the individuals extracted from the given collection of the documents, and E is the set of all the correlation values between individuals such that for any correlation $C(x, y)$, there is a corresponding edge in G with the weight $C(x, y)$ between the two nodes x and y in U.

For the problem of MLC group model generation, we define the function P in Link Identification as a graph segmentation based on a minimum correlation threshold T. The specific value of T may be obtained based on a law enforcement investigator's expertise, where our CORAL prototype allows the investigator to play with different thresholds so as to be able to validate different models generated based on his/her expertise.

Given T, from the literature [28] there are efficient algorithms such as the breadth-first search with complexity $O(|E|)$ available to conduct this segmentation. Note that there may be multiple subgraphs M generated, indicating that there may possibly be multiple MLC groups identified in the given document collection. It is also possible that the original graph $G(V, E)$ may not necessarily be connected (the complete graph G may have edges with correlation values 0, resulting in virtually an incomplete graph).

6. EXPERIMENTAL RESULTS

We have implemented the specific method of LDCA in the MLC group model generation problem into a prototype system, called CORAL. In this section we first discuss the scenario of a real MLC case used in the experiments with the data given by NIJ. CORAL is tested and evaluated based on this data set. Since the data are not considered as public domain data, we have replaced all the real names of the involved individuals and organizations with faked names in this paper for the purpose of the discussion and references.

6.1 The Case Scenario

The documents used in this project were collected concerning the practices of a group of businesses, their clients and associates involved in an alleged money laundering case. The documents were obtained from an investigation of a fraudulent scheme to offer and sell unregistered prime bank securities throughout the United States. The U.S. Securities and Exchange Commission, the Securities Division of the Utopia Corporation Commission, the U.S. Customs Service and the Utopia Attorney General's Office jointly investigated the case.

It was alleged that Bob Schemer and his company, Acme Finance, Ltd., along with a group of other individuals and organizations developed a fraudulent trading scheme. Religious and charitable groups and individuals investing retirement funds were targeted. Approximately $45 million dollars was raised from more than three hundred investors. To encourage investors, Schemer et al misrepresented the use and safety of investors' funds. Investors were told that their funds would be transferred to a foreign bank, secured by a bank guarantee and used as collateral to trade financial instruments with the top fifty European banks. The investors were also told that this trading activity would provide annual returns of 24% to 60%. This was not the case. Schemer et al did not send any of the funds to Europe for use in a trading program, and the funds were not

secured by any type of guarantee. Instead, Schemer et al misappropriated the investment funds for unauthorized and personal uses. Schemer also used the funds to make Ponzi payments, which is an investment scheme in which returns are paid to earlier investors entirely out of the money paid into the scheme by new investors.

6.2 The Test and Evaluation Results

There were 7,668 documents in total as the whole collection that were provided by NIJ. Due to the gross OCR errors and the IE tagger errors (which was partially caused by the OCR errors), we had to manually clean up all the documents before they could be used as the CORAL input. We manually cleaned 332 documents and used this collection as the testing bed for CORAL. We ran CORAL on the 332 documents. There were 252 person names extracted with 2,104 monetary vectors in total. The distribution of the monetary vectors along timeline was not even, with the majority obtained from those involved in the MLC case, which verified that the focus assumption was correct.

CORAL ran the collection of the 332 documents and took about 20 minutes to complete the model generation in a platform of P-III/800, 512 MB, Windows 2000. Compared with the typical effort required in a manual model generation, this demonstrates the savings automatic model generation can offer. Figure 8 shows a pair of financial transaction history timeline histograms of two individuals included in a generated model as part of the same MLC group based on the "similarity" of their timeline histograms, and Figure 9 shows the intermediate interface of the timeline map of Bob Schemer. The user can click any node in the timeline map and CORAL pops up a separate window to show the summary of the monetary vector this individual has conducted at the corresponding time instance. Figure 10 shows the models generated by CORAL with thresholds 0.35 and 0.18.

At this time we do not have access to the court documents in terms of the complete list of the individuals convicted in this case as well as the roles every convicted individual played in the MLC group. However, from what is reported in the news, we know that with sufficiently high correlation thresholds (such as 0.18), the individuals identified by CORAL are the major crime group members convicted. This shows that the LDCA based method for MLC group model generation does have the capability to identify the correct MLC group members as well as to link them together to generate the groups. To further show the strength of the HCB correlation method we have developed, we have also implemented CORAL using the simple cosine function based correlation [29]. With a threshold selected to identify a similar number of MLC group members, the members generated based on this "direct" correlation do not match the major crime group members convicted. This is due to the fact that the "direct" correlation does not address the unique characteristics of the MLC data --- the transaction patterns exhibit zoning effects in the timeline, and the transaction activities exhibit "discrepancy" in time. On the other hand, taking the example of the model generated at the threshold 0.18 by CORAL using HCB correlation, we have 7 individuals that were identified as the MLC group members from the original 252 individuals extracted in the 332 documents. This shows that the elimination rate is 245/252 = 97%! Based on this MLC case evaluation, we are confident that the LDCA general methodology offers great promise for automatic community generation based on uni-party data sets.

7. CONCLUSIONS

We have identified a new paradigm of problems in this paper, which is the community generation from mining uni-party data. Unlike the traditional community generation problems such as Web mining, collaborative filtering, and social network analysis, in which the data sets are given as bi-party data, here we do not have direct and explicit access to the link information between data items. We have proposed a general methodology to solve for the problems in this paradigm, called Link Discovery based on Correlation Analysis (LDCA). As an example of these problems, we formulate and address the money laundering crime (MLC) group model generation problem. Since the conventional cosine function based correlation is not able to capture the link relationship, we develop a Hierarchical Composition Based (HCB) correlation analysis along timeline to generate the MLC group model. We have implemented this method to develop the CORAL prototype system, and tested and evaluated CORAL using a data set of a real MLC provided by NIJ. The preliminary testing and evaluations have demonstrated the promise of using LDCA in automatically generating MLC group models based on HCB correlation, as well as validated the LDCA methodology.

8. ACKNOWLEDGMENTS

This work is supported in part by Air Force Research Laboratory through contract F30602-02-M-V020.

9. ADDITIONAL AUTHORS

Jingzhou Hua and Ruofei Zhang, Computer Science Department, SUNY Binghamton, Binghamton, NY 13902, and Maureen Regan and Debra Cutler, Dolphin Technology, Inc., Rome, NY 13441.

10. REFERENCES

[1] Adderley, R. and Musgrove, P.B., Data mining case study: modeling the behavior of offenders --- who commit serious sexual assaults*, Proc. ACM KDD*, 2001.

[2] Aggarwal, C.C., Al-Garawi, F., and Yu, P.S., Intelligent crawling on the World Wide Web with arbitrary predicates, *Proc. ACM WWW,* 2001.

[3] Agrawal, R. Lin, K.-I., Sawhney, H.S., and Shim, K., Fast similarity search in the presence of noise, scaling, and translation in time-series database, *Proc. VLDB*, 1995.

[4] Ankerst, M., Human involvement and interactivity of the next generation's data mining tools, *Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[5] Baldwin, F.N., Money laundering and wire transfers: when the new regulations take effect will they help?, *Dickinson Journal of International Law*, 14(Spring), 413-454, 1996.

[6] Chen, M.-S., Han, J., and Yu, P.S., Data mining: an overview from a database perspective, *IEEE Trans. Knowledge and Data Engineering*, 8(6), 866-883, 1996.

[7] Domingos, P. and Hulten, D., Catching up with the data: research issues in mining data streams, *Proc. Workshop on*

Research Issues in Data Mining and Knowledge Discovery, 2001.

[8] Famili, A., Shen, W-M., Weber, R., and Simoudis, E., Data preprocessing and intelligent data analysis, *Intelligent Data Analysis*, 1, 3-23, 1997.

[9] Feldman, R. and Hirsh, H., Finding associations in collections of text, In R.S. Michalski, I. Bratko, and M. Kubat, editors, *Machine Learning and Data Mining: Methods and Applications*, 223-240, John Wiley & Sons, 1998.

[10] Fox, H., Schwartz, R., Stone, R., Weischedel, A., and Gadz, W., Learning to extract and classify names from text, *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA, Oct., 1998.

[11] Gibson, D., Kleinberg, J., and Raghavan, P., Inferring Web communities from link topoploy, Proc. HyperText98, 1998.

[12] Graycar, A. and Grabosky, P., (Eds.) *Money Laundering in the 21st Century: Risks and Countermeasures*, Australian Institute of Criminology, 1996.

[13] Grishman, R., Tipster architecture design document version 2.3, *Technical Report,* DARPA, 1997.

[14] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2000.

[15] Hearst, M.A., Untangling text data mining, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[16] Hosking, J.R.M., Pednault, E.P.D., and Sudan, M., A statistical perspective on data mining, *Future Generation Computer Systems*, 13, 117-134, 1997.

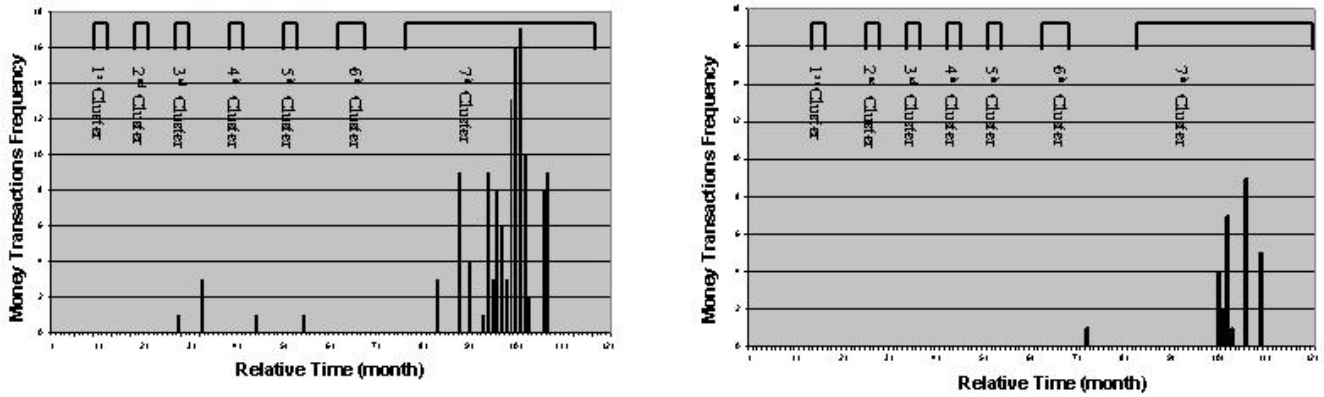

Figure 8: The financial transaction history histograms for Bob Schemer (left) and Fred Brown (right) with the clustered time zones overlaid; there were 7 clusters identified.

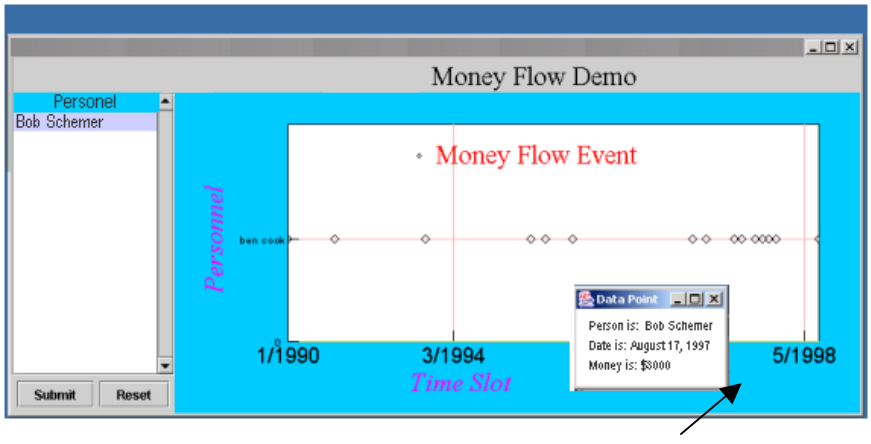

Figure 9: A CORAL view of the interface for a segment of the timeline display for Bob Schemer and his financial activities. Clicking on a dot will bring up a second window showing a summary of the financial activities at the corresponding time instance. The timeline shows Bob Schemer's activities from January 1990 through May 1998.

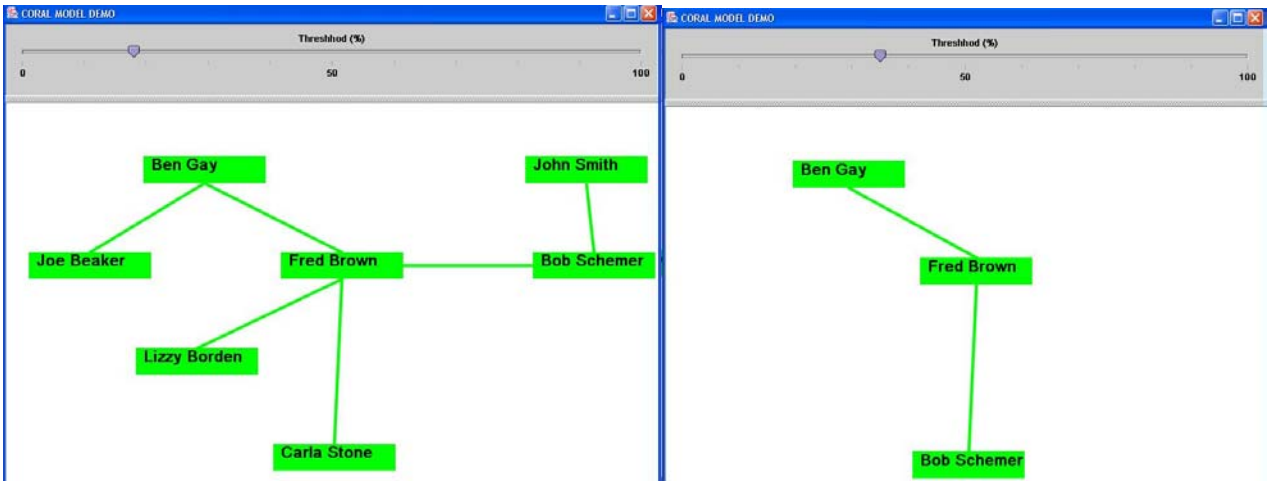

Figure 10: The models generated by CORAL with a threshold = 0.18 (on left) and 0.35 (on right).

[17] Hui, S.C. and Jha, G., Data mining for customer support, *Information and Management*, 38, 1-13, 2000.

[18] Jain, R., Kasturi, R., and Schunck, B.G., *Machine Vision,* Prentice Hall, 1995.

[19] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., and Wu, A.Y., An efficient k-means clustering algorithm: analysis and implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence,* 24(7): 881-892, 2002.

[20] J. Kleingberg, Authoritative sources in a hyper linked environment, Proc. of CAN-SIAM Symposiumon Discrete Algorithms, 1998.

[21] Mack, R. and Hehenberger, M., Text-based knowledge discovery: search and mining of life-sciences documents, *Drug Discovery Today*, 7(11), Suppl. S89-S98, 2002.

[22] Mannila, H., Local and global methods in data mining: basic techniques and open problems, *Proc. 29th International Colloquium on Automata, Languages, and Programming*, 2002.

[23] Mitchell, T.M., *Machine Learning*, McGraw-Hill, 1997.

[24] Wang, K., Zhou, S., and Liew, S.C., Building hierarchical classifiers using class proximity, *Proc. VLDB*, 1999.

[25] MUC-7, *Proc. 7th Machine Understanding Conference,* http://www.itl.nist.gov/iaui/894.02/related_projects/muc/), 1998.

[26] Rafiei, D. and Mendelzon, A., Similarity-based queries for time series data, *Proc. ACM-SIGMOD International Conf. Management of Data*, 1997.

[27] Rajagopalan, S., Kumar, R., Raghavan, P. and Tomkins, A., Trawling the Web for emerging cyber-communities. 8the WWW conference, 1999.

[28] Russell, S. and Norvig, P., *Artificial Intelligence, A Modern Approach,* Prentice Hall, 1995.

[29] Salton, G., *Automatic Text Processing,* Addison-Wesley, 1989.

[30] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., Item-based collaborative filtering recommendation algorithms, *Proc. ACM WWW,* 2001.

[31] Saxena, R., Note and comment: cyberlaundering: the next step for money launderers?, *St. Thomas Law Review*, Spring, 1998, http://web.lexis-nexis.com/universe/

[32] Scott., J. Social Network Analysis: A handbook, SAGE Publications, 1991

[33] Shardanand, U. and Maes, P., Social information filtering: algorithms for automating "world of mouth", *Proc. ACM CHI,* 1995.

[34] Smyth, P., Breaking out of the black-box: research challenges in data mining, *Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[35] Terano, T., Asai, K., and Sugeno, M., *Fuzzy Systems Theory and Its Applications,* Academic Press, 1992.

[36] Toyoda, M. and Kitsuregawa, M., Creating a Web community chart for navigating related communities, *Proc. ACM HT,* 2001.

[37] Vertan, C. and Boujemaa, N., Embedding fuzzy logic in content based image retrieval, *Proc. 19th Int'l Meeting of North America Fuzzy Information Processing Society,* 2000.