

Detecting patterns of fraudulent behavior in forensic accounting*

Boris Kovalerchuk¹, Evgenii Vityaev²

¹Dept. of Computer Science, Central Washington University
Ellensburg, WA 98926, USA

borisk@cwu.edu

²Institute of Mathematics, Russian Academy of Sciences,
Novosibirsk, 630090, Russia

vityaev@math.nsc.ru

Abstract. Often evidence from a single case does not reveal any suspicious patterns to aid investigations in forensic accounting and other forensic fields. In contrast, correlation of sets of evidence from several cases with suitable background knowledge may reveal suspicious patterns. Link Discovery (LD) has recently emerged as a promising new area for such tasks. Currently LD mostly relies on deterministic graphical techniques. Other relevant techniques are Bayesian probabilistic and causal networks. These techniques need further development to handle rare events. This paper combines first-order logic (FOL) and probabilistic semantic inference (PSI) to address this challenge. Previous research has shown this approach is computationally efficient and complete for statistically significant patterns. This paper shows that a modified method can be successful for discovering rare patterns. The method is illustrated with an example of discovery of suspicious patterns.

1. Introduction

Forensic accounting is a field that deals with possible illegal and fraudulent financial transactions [3]. One current focus in this field is the analysis of funding mechanisms for terrorism where *clean money* (e.g., *charity money*) and *laundered money* are both used [1] for a variety of activities including acquisition and production of weapons and their precursors. In contrast, traditional illegal businesses and drug trafficking *make dirty money appear clean* [1].

There are many indicators of possible suspicious (abnormal) transactions in traditional illegal business. These include (1) the use of several related and/or unrelated accounts before money is moved offshore, (2) a lack of account holder concern with commissions and fees [2], (3) correspondent banking transactions to offshore shell banks [2], (4) transferor insolvency after the transfer or insolvency at the time of

* Kovalerchuk, B., Vityaev, E., [Detecting patterns of fraudulent behavior in forensic accounting](#). In Proc. of the Seventh International Conference “Knowledge-based Intelligent Information and Engineering on Systems”, Oxford, UK, Sept, 2003, part 1, pp. 502-509

transfer, (5) wire transfers to new places [4], (6) transactions without identifiable business purposes, and (7) transfers for less than reasonably equivalent value [5].

Some of these indicators can be easily implemented as simple flags in software. However, indicators such as wire transfers to new places produce a large number of 'false positive' suspicious transactions. Thus, the goal is to develop more sophisticated mechanisms based on interrelations among many indicators. To meet these challenges link analysis software for forensic accountants, attorneys and fraud examiners such as NetMap, Analyst's Notebook and others [4-7] have been and are being developed.

Here we concentrate on fraudulent activities that are closely related to terrorism such as transactions without identifiable business purposes. The problem is that often an individual transaction does not reveal that it has no identifiable business purpose or that it was done for no reasonably equivalent value. Thus, we develop a technique that searches for suspicious patterns in the form of more complex combinations of transactions and other evidence using background knowledge.

The specific tasks in automated forensic accounting related to transaction monitoring systems are the identification of suspicious and unusual electronic transactions and the reduction in the number of 'false positive' suspicious transactions by using inexpensive, simple rule-based systems, customer profiling, statistical techniques, neural networks, fuzzy logic and genetic algorithms [1]. This paper combines the advantages of first-order logic (FOL) and probabilistic semantic inference (PSI) [8] for these tasks. We discover the following transaction patterns from ordinary or distributed databases that are related to terrorism and other illegal activities:

- a normal pattern (NP) – a Manufacturer Buys a Precursor & Sells the Result of manufacturing (MBPSR);
- a suspicious (abnormal) pattern (SP) – a Manufacturer Buys a Precursor & Sells the same Precursor (MBPSP);
- a suspicious pattern (SP) – a Trading Co. Buys a Precursor and Sells the same Precursor Cheaper (TBPSPC);
- a normal pattern (NP) -- a Conglomerate Buys a Precursor & Sells the Result of manufacturing (CBPSR).

2. Example

Consider the following example. Table 1 contains transactions with the attributes seller, buyer, item sold, amount, cost and date and Table 2 describes the types of companies and items sold.

Table 1. Transactions records

Record ID	Seller	Buyer	Item sold	Amount	Cost	Date
1	Aaa	Ttt	Td	1t	\$100	03/05/99
2	Bbb	Ccc	Td	2t	\$100	04/06/98
3	Ttt	Qqq	Td	1t	\$100	05/05/99
4	Qqq	Ccc	Pd	1.5t	\$100	05/05/99
5	Ccc	Ddd	Td	2.0t	\$200	08/18/98
6	Ddd	Ccc	Pd	3.0t	\$400	09/18/98

We assemble a new Table 3 from Tables 1 and 2 to look for suspicious patterns. For instance, row 1 in Table 3 is a combination of row 1 from Table 1 and rows 1 and 4 from Table 2 that contain types of companies and items. Table 3 does not indicate suspicious patterns immediately, but we can generate pairs of records from Table 3 that can be mapped to patterns listed above using a pattern-matching algorithm A. The algorithm A analyzes pairs of records in Table 3. For simplicity, we can assume that a new table with 18 attributes is formed to represent pairs of records from Table 3. Each record in Table 3 contains nine attributes.

Table 2. Company types and item types

Record ID	Company name (seller/buyer)	Company type	Item	Item type in process PP
1	Aaa	Trading	Td	Precursor
2	Bbb	Unknown	Pd	Product
3	Ccc	Trading	Rd	Precursor
4	Ttt	Manufacturing		
5	Ddd	Manufacturing		
6	Qqq	Conglomerate		

Table 3. Combined data records

Record ID	Seller	Seller type	Buyer	Buyer type	Item sold	Item type	Amount	Price	Date
	1	2	3	4	5	6	7	8	9
1	Aaa	trading	Ttt	Manuf.	Td	Precursor	1t	\$100	03/05/99
2	Bbb	unknown	Ccc	Trading	Td	Precursor	2t	\$100	04/06/98
3	Ttt	manuf.	Qqq	Congl.	Td	Precursor	1t	\$100	05/05/99
4	Qqq	Congl.	Ccc	Trading	pd	Product	1.5t	\$100	06/23/99
5	Ccc	Trading	Ddd	Manuf.	td	Precursor	2.0t	\$200	08/18/98
6	Ddd	Manuf	Ccc	Trading	pd	Product	3.0t	\$400	09/18/98

Thus, we map pairs of records in Table 3 into patterns:
 $A(\#5,\#6)=MBPSR$, that is a pair of records #5 and #6 from Table 3 indicates a normal pattern -- a manufacturer bought a precursor and sold product ;

In contrast, two other pairs indicate suspicious patterns:
 $A(\#1,\#3)=MBPSP$, that is a manufacturer bought a precursor and sold the same precursor;
 $A(\#2,\#5)=TBSPC$, that is a trading company bought a precursor and sold the same precursor cheaper.

Now let us assume that we have a database of 10^5 transactions as in Table 1. Then Table 3 will have all pairs of them, i.e., about $5 \cdot 10^9$. Statistical computations can reveal a distribution of these pairs into patterns as shown in Table 4.

Table 4. Statistical characteristics

Pattern	Type	Frequency, %	Approximate number of cases
MBPSR	normal	55	$0.55 \cdot 5 \cdot 10^9$

MBPSP	suspicious	0.1	100
CBPSR	normal	44.7	$0.44*5*10^9$
TBSPSPC	suspicious	0.2	200

Thus, we have 300 suspicious transactions. This is 0.3% of the total number of transactions and about $6*10^{-6}$ % of the total number of pairs analyzed. It shows that finding such transactions is like finding a needle in a haystack. The *automatic generation of patterns/hypotheses descriptions* is a major challenge. This includes generating MBPSP and TBSPSPC descriptions automatically. We do not assume that we already know that MBPSP and TBSPSPC are suspicious. One can ask: “Why do we need to discover these definitions (rules) automatically?” A manual way can work if the number of types of suspicious patterns is small and an expert is available. For multistage money-laundering transactions, this is difficult to accomplish manually. It is possible that many laundering transactions were processed before money went offshore or was used for illegal purposes. Our *approach* to identify suspicious patterns is to discover *highly probable patterns* and then *negate* them. We suppose that a highly probable pattern should be *normal*. In more formal terms, the *main hypothesis (MH)* is:

If Q is a highly probable pattern (>0.9) then Q constitutes a normal pattern and not(Q) can constitute a suspicious (abnormal) pattern.

Table 5 outlines an algorithm based on this hypothesis to find suspicious patterns. The algorithm is based first-order logic and probabilistic semantic inference [8].

Table 5. Algorithm steps for finding suspicious patterns based on the main hypothesis (MH)

1	<p><i>Discover patterns</i> in a database such as MBPSP in a form $MBP \Rightarrow SR$, that is, as a Horn clause $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n$ (see [8] for mathematical detail).</p> <p>1.1. Generate a set of predicates $Q = \{Q_1, Q_2, \dots, Q_m\}$ and first order logic sentences A_1, A_2, \dots, A_n based on those predicates. For instance, Q_1 and A_1 could be defined as follows: $Q_1(x) = 1 \Leftrightarrow x$ is a trading company and $A_1(a,b) = Q_1(a) \& Q_1(b)$, where a and b are companies.</p> <p>1.2. Compute a probability P that pattern $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n$ is true on a given database. This probability is computed as a conditional probability of conclusion A_n under assumption that If-part $A_1 \& A_2 \& \dots \& A_{n-1}$ is true, that is $P(A_n / A_1 \& A_2 \& \dots \& A_{n-1}) = N(A_n / A_1 \& A_2 \& \dots \& A_{n-1}) / N(A_1 \& A_2 \& \dots \& A_{n-1} \& A_n)$, where $N(A_n / A_1 \& A_2 \& \dots \& A_{n-1})$ is the number of $A_n / A_1 \& A_2 \& \dots \& A_{n-1}$ cases and $N(A_1 \& A_2 \& \dots \& A_{n-1} \& A_n)$ is the number of $A_1 \& A_2 \& \dots \& A_{n-1} \& A_n$ cases.</p> <p>1.3. Compare $P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n)$ with a threshold T, say $T=0.9$. If $P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n) > T$ then a database is “normal”. A user can select another value of threshold T, e.g., $T=0.98$. If $P(MBP \Rightarrow SR) = 0.998$, then DB is normal for 0.98 too.</p> <p>1.4. Test statistical significance of $P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n)$. We use the Fisher criterion [8] to test statistical significance.</p>
2	<p><i>Negate patterns</i>. If database is “normal” ($P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n) > T=0.9$ and $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n$ is statistically significant then negate $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n$ to produce a negated pattern $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow \neg A_n$.</p>
3	<p><i>Compute the probability of the negated pattern</i> $P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow \neg A_n) = 1 - P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n)$. In the example above, it is $1 - 0.998 = 0.002$.</p>

4	Analyze database records that satisfy $A_1 \& A_2 \& \dots \& A_{n-1} \& \neg A_n$ for possible <i>false alarm</i> . Really suspicious records satisfy the property $A_1 \& A_2 \& \dots \& A_n \& \neg A_n$, but normal records also can satisfy this property.
---	---

To minimize computations we generate randomly a *representative part* of all possible pairs of records such as shown in Table 4. Then an algorithm finds highly probable ($P > T$) Horn clauses. Next, these clauses are negated as described in Table 5.

After that, a full search of records in the database is performed to find records that satisfy the negated clauses. According to our main hypothesis (MH) this set of records will contain suspicious records and the search for “red flag” transactions will be significantly narrowed. Use of the property of *monotonicity* is another tool we use to minimize computations. The idea is based on a simple observation: If $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow B$ represents a suspicious pattern then $A_1 \& A_2 \& \dots \& A_{n-1} \& A_n \Rightarrow B$ is suspicious too. Thus, one does not need to test clause $A_1 \& A_2 \& \dots \& A_{n-1} \& A_n \Rightarrow B$ if $A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow B$ is already satisfied.

3. Hypothesis Testing

One of the technical aims of this paper is to design tests and simulation experiments for this thesis. We designed two test experiments:

1. Test 1: Generate a relatively large Table 4 that includes a few suspicious records MBPSP and TBSPC. Run a data-mining algorithm (MMDR [8]) to discover as many highly probable patterns as possible. Check that patterns MBPSR and CBPSR are among them. Negate MBPSR and CBPSR to produce patterns MBPSP and TBSPC. Run patterns MBPSP and TBSPC to find all suspicious records consistent with them.
2. Test 2: Check that other highly probable patterns found are normal; check that their negations are suspicious patterns (or contain suspicious patterns).

A positive result of Test 1 will confirm our hypothesis (statement) for MBPSR and CBPSR and their negations. Test 2 will confirm our statement for a wider set of patterns. In this paper we report results of conducting Test 1. The word “*can*” is the most important in our statement/hypothesis. If the majority of not(Q) patterns are consistent with an informal and intuitive concept of suspicious pattern then this hypothesis will be valid. If only a few of the not(Q) rules (patterns) are intuitively suspicious then the hypothesis will not be of much use even if it is formally valid.

A method for Test 1 contains several steps:

- Create a Horn clause: $MBP \Rightarrow SR$.
- Compute a probability that $MBP \Rightarrow SR$ is true on a given database. Probability $P(MBP \Rightarrow SR)$ is computed as a conditional probability $P(SR/MBP) = N(SR/MBP)/N(MBP)$, where $N(SR/MBP)$ is the number of MBPSR cases and $N(MBP)$ is the number of MBP cases.
- Compare $P(MBP \Rightarrow SR)$ with 0.9. If $P(MBP \Rightarrow SR) > 0.9$ then a database is ‘normal’. For instance, $P(SR/MBP)$ can be 0.998.
- Test the statistical significance of $P(MBP \Rightarrow SR)$. We use Fisher criterion [8] to test statistical significance.

- If the database is “normal” ($P(\text{MBP} \Rightarrow \text{SR}) > T=0.9$) and if $P(\text{MBP} \Rightarrow \text{SR})$ is statistically significant then negate $\text{MBP} \Rightarrow \text{SR}$ to produce $\neg(\text{MBP} \Rightarrow \text{SR})$. Threshold T can have another value too.
- Compute probability for a negated pattern $P(\text{MBP} \Rightarrow \neg(\text{SR}))$. In the example above it is $1-0.998=0.002$.
- Analyze database records that satisfy MBP and $\neg(\text{SR})$. For instance, really suspicious MBPSP records satisfy property MBP and $\neg(\text{SR})$, but other records also can satisfy this property too. For instance, MBPBP records (a manufacturer bought a precursor twice) can be less suspicious than MBPSP.

Thus, if the probability $P(\text{SR}/\text{MBP})$ is high (0.9892) and statistically significant then a normal pattern MBPSR is discovered. Then suspicious cases are among the cases where MBP is true but the conclusion SR is not true. We collect these cases and analyze the actual content of the then-part of the clause $\text{MBP} \Rightarrow \text{SR}$. The set $\neg \text{SR}$ can contain a variety of entities. Some of them can be very legitimate cases. Therefore, this approach does not guarantee that we find only suspicious cases, but the method narrows the search to a much smaller set of records. In the example above the search is narrowed to 0.2% of the total cases.

4. Experiment

We generated two synthesized databases with attributes shown in Table 4. The first one does not have suspicious records MBPSP and TBPSPC. A second database contains few such records. Using a Machine Method for Discovery Regularities (MMDR) [8] we were able to discover MBPSR and CBPSR normal patterns in both databases.

Table 6. Database with suspicious cases

Pattern	Probability $P(A_1 \& A_2 \& \dots \& A_{n-1} \Rightarrow A_n)$	
	In database without suspicious cases	In database with suspicious cases
Normal pattern, $\text{MBP} \Rightarrow \text{SR}$	> 0.95	> 0.9
Negated pattern $\text{MBP} \Rightarrow \neg(\text{SR})$	< 0.05	< 0.1
Normal pattern $\text{CBP} \Rightarrow \text{SR}$	> 0.95	> 0.9
Negated pattern $\text{CBP} \Rightarrow \neg(\text{SR})$	< 0.05	< 0.05

The MMDR method worked without any advanced information that these patterns are in data. In the database without suspicious cases, negated patterns $\text{MBP} \Rightarrow \neg(\text{SR})$ and $\text{CBP} \Rightarrow \neg(\text{SR})$ contain cases that are not suspicious. For instance, $\text{MBP} \Rightarrow \text{BP}$, that is, a manufacturer that already bought precursors (transaction record 1) bought them again (transaction record 2). The difference in probabilities for $\text{MBP} \Rightarrow \neg(\text{SR})$ in the two databases points out actually suspicious cases. In our computational experiments, the total number of regularities found is 41. The number of triples of companies (i.e., pairs of transactions) captured by regularities is 1531 out of total 2772 tri-

ples generated in the experiment. Table 7 depicts some statistically significant regularities found. Attributes New_Buyer_type and New_Item_type belong to the second record in a pair of records (R1,R2). Individual records are depicted in table 3.

Table 7. Computational experiment: examples of discovered regularities

#	Discovered regularity	Frequency
1	IF Seller_type = Manufacturing AND Buyer_type = Manufacturing THEN New_Item_type = product	$72 / (6 + 72) = 0.923077$
2	IF Seller_type = Manufacturing AND New_Buyer_type = Manufacturing THEN New_Item_type = product	$72 / (6 + 72) = 0.923077$
3	IF Seller_type = Manufacturing AND Item_type = precursor THEN New_Item_type = product	$152 / (59 + 152) = 0.720379$
4	IF Seller_type = Manufacturing AND Price_Compare = 1 AND New_Buyer_type = Trading THEN New_Item_type = product	$47 / (2 + 47) = 0.959184$
5	IF Seller_type = Manufacturing AND Price_Compare = 1 AND Item_type = precursor THEN New_Item_type = product	$79 / (5 + 79) = 0.940476$

5. Conclusion

The method outlined in this paper advances pattern discovery methods that deal with complex (non-numeric) evidences and involve structured objects, text and data in a variety of discrete and continuous scales (nominal, order, absolute and so on). The paper shows potential application of the technique for forensic accounting. The technique combines first-order logic (FOL) and probabilistic semantic inference (PSI). The approach has been illustrated with an example of discovery of suspicious patterns in forensic accounting.

References

1. Prentice, M., Forensic Services - tracking terrorist networks,2002, Ernst & Young LLP, UK,http://www.ey.com/global/gcr.nsf/UK/Forensic_Services_tracking_terrorist_networks
2. Don Vangel and Al James Terrorist Financing: Cleaning Up a Dirty Business, the issue of Ernst & Young's financial services quarterly, Spring 2002. http://www.ey.com/GLOBAL/content.nsf/International/Issues_&_Perspectives-Library-Terrorist_Financing_Cleaning_Up_a_Dirty_Business
3. IRS forensic accounting by TPI, 2002, http://www.tpirsrelief.com/forensic_accounting.htm
4. Chabrow, E. Tracking The Terrorists, Information week, Jan. 14, 2002, http://www.tpirsrelief.com/forensic_accounting.htm
5. How Forensic Accountants Support Fraud Litigation, 2002, http://www.fraudinformation.com/forensic_accountants.htm
6. i2 Applications-Fraud Investigation Techniques, <http://www.i2.co.uk/applications/fraud.html>

7. Evett, IW., Jackson, G. Lambert, JA , McCrossan, S. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice* 2000; 40: 233–239
8. Kovalerchuk, B., Vityaev, E., *Data Mining in Finance: Advances in Relational and Hybrid Methods*, Kluwer, 2000