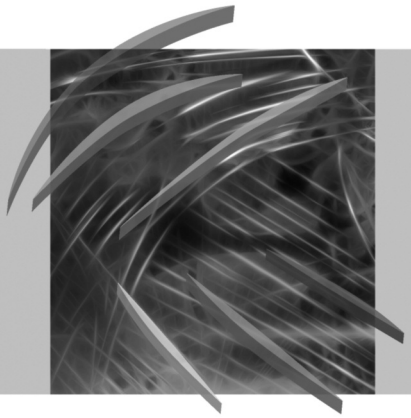
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

206

Advanced Information Technologies for Management – AITM 2011

Intelligent Technologies and Applications



edited by
Jerzy Korczak, Helena Dudycz, Mirosław Dyczkowski

Publishing House of Wrocław University of Economics
Wrocław 2011

Reviewers: Frederic Andres, Witold Chmielarz, Jacek Cypryjański, Beata Czarnacka-Chrobot, Bernard F. Kubiak, Halina Kwaśnicka, Antoni Ligęza, Anna Ławrynowicz, Mikołaj Morzy, Stanisław Stanek, Ewa Ziemba

Copy-editing: Agnieszka Flasińska

Layout: Barbara Łopusiewicz

Proof-reading: Marcin Orszulak

Typesetting: Adam Dębski

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com

Information on submitting and reviewing papers is available on the Publishing House's website
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-182-9

The original version: printed
Printing: Printing House TOTEM

Contents

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 206
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS

Advanced Information Technologies for Management – AITM 2011 ISSN 1899-3192
Intelligent Technologies and Applications

Krzysztof Michalak, Jerzy Korczak*
Wrocław University of Economics, Wrocław, Poland

EVOLUTIONARY GRAPH MINING IN SUSPICIOUS TRANSACTION DETECTION

Abstract: Money laundering may involve complex organizational schemes designed to obfuscate the real purpose of money transfers. In this paper, we present a graph mining method that allows detection of transaction subgraphs containing suspicious transactions. Suspicious subgraph model is parameterized using fuzzy numbers which represent parameters of transactions and some structural features of the transaction subgraphs itself. The method presented in this paper uses fuzzy matching of graph structures which allows detecting money-laundering schemes which differ to some extent from those annotated by an expert.

Keywords: graph mining, evolutionary algorithms, money laundering.

1. Introduction

Individual transactions involved in money laundering are usually obfuscated, so that they look as regular, legal transactions. Therefore, the best chance of detecting criminal activities is given by analyzing relations between transactions and entities which send and receive them. Money laundering process is usually divided into three stages: placement, layering and integration [Truman, Reuter 2004] which involve various schemes of money transferring between bank accounts. Examples of such schemes, which involve transferring money via a number of intermediaries, are shown in Figure 1. The first scheme is created in order to avoid exceeding a transfer amount which automatically triggers a suspicious transaction alarm (e.g. $10,000). A much larger amount is transferred from the sender to the receiver but it is split to transfers that are below the alarm threshold. The second scheme makes the connection between the sender and the receiver harder to discover than in the case of a direct transfer.

In Figure 2 a small subgraph of transactions (in which vertices are shaded in gray) is shown. Such a subgraph may indicate an attempt to obfuscate transferring money from the account #31639 to the account #24075. It is, however, possible that similar subgraph is formed by completely legal activities. Also, accounts used for illegal transaction structuring may also be involved in a number of legal transactions.

* e-mails: {krzysztof.michalak,jerzy.korczak}@ue.wroc.pl.

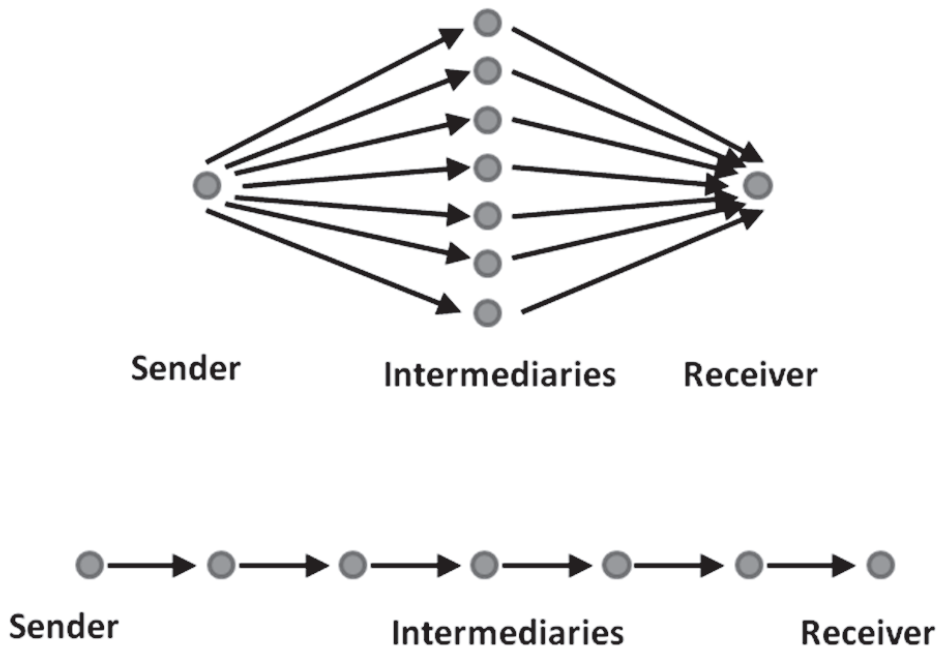


Figure 1. Examples of subgraphs which may indicate money-laundering activities

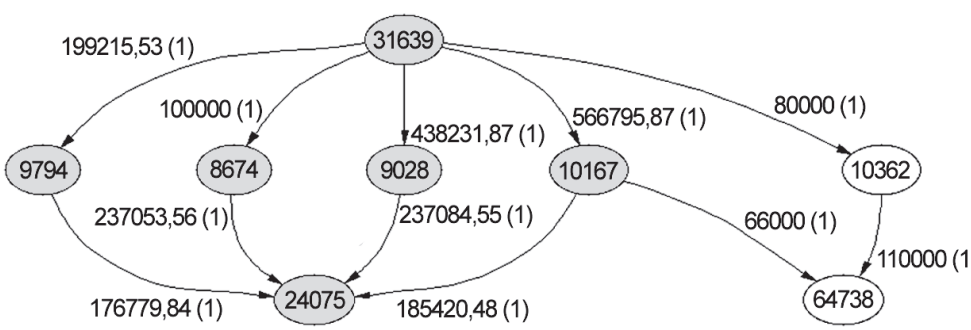


Figure 2. Suspicious transaction graph (with vertices shaded in grey) connected with other, possibly legal activities. Vertex labels are account identifiers, edge labels contain number of transactions between the two account and the total amount transferred

2. Evolutionary graph mining

Graph mining [Cook, Holder 2007] seems to be a promising approach for money laundering detection, because it makes it possible to detect complex dependencies between transactions and to take into account properties and relations of entities in-

volved in sending and receiving the transfers. Following the paradigm of machine learning, we would like to be able to detect illegal transactions in unseen data using a suspicious subgraph model built using training data provided by human expert who annotates transactions. Annotated transactions can also be obtained by performing data mining in data warehouses [Korczak, Marchelski, Oleszkiewicz 2008; Korczak, Oleszkiewicz 2009] and can subsequently be used for training a suspicious subgraph model used in this paper. Because it is hard to predict what transaction schemes may be invented by criminals who try to perform illegal acts, we propose a method in which a suspicious subgraph model is built from smaller building blocks in an evolutionary manner. General model for a transaction subgraph detected by the presented method is shown in Figure 3.

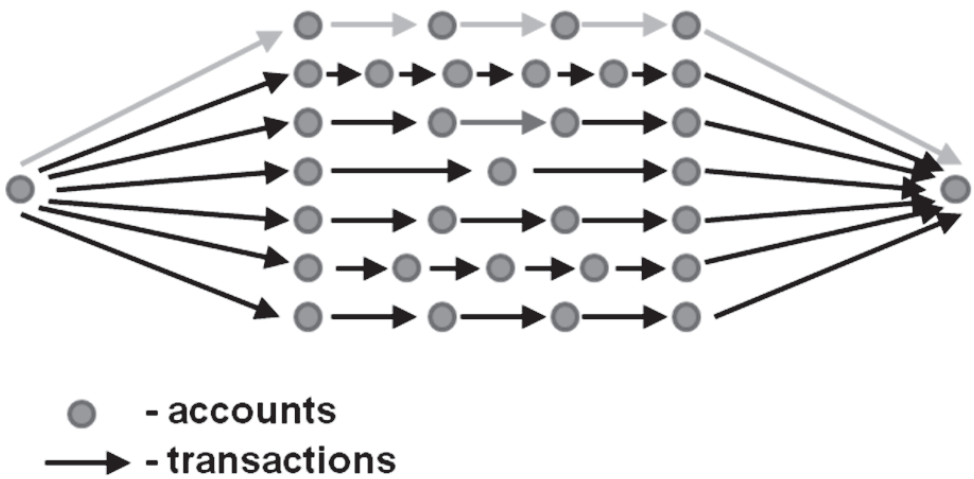


Figure 3. General structure of a subgraph detected by the presented method

Evolutionary approach involves a population of subgraph models (specimens) which are evaluated according to their ability to detect suspicious transactions. Existing specimens are mutated by introducing small random changes to parameters of specimens. Information is distributed in the population during a crossover phase in which parts of subgraph model are interchanged between specimens. Selection procedure which promotes specimens with higher values of fitness function (i.e. those that perform better in suspicious transaction detection) ensures that in consecutive generations performance of specimens in the population improves. Further discussion of genetic algorithms and their applications can be found in [Goldberg 1989; Goldberg, Sastry 2011]. Evolutionary approach makes it possible to train the model with respect to transaction parameters as well as graph structure.

The model is built from three blocks: individual transaction pattern (TR), transaction chain pattern (SER) and parallel paths pattern (PAR). This model is para-

meterized using polygonal fuzzy numbers [Buckley, Eslami 2002; Fetz et al. 1999] (denoted using the hat (^) symbol). We use simplified polygonal fuzzy numbers represented by only 6 real numbers: $\hat{x} = \langle x_1, x_2, x_3, x_4, m_2, m_3 \rangle$ for which a membership function $\mu_{\hat{x}}$ is calculated as follows:

$$\mu_{\hat{x}}(x) = \begin{cases} 0 & \text{for } x \le x_1 , \\ m_2 \cdot \frac{x - x_1}{x_2 - x_1} & \text{for } x \in (x_1, x_2] , \\ m_2 + (m_3 - m_2) \cdot \frac{x - x_2}{x_3 - x_2} & \text{for } x \in (x_2, x_3] , \\ m_3 \cdot \frac{x_4 - x}{x_4 - x_3} & \text{for } x \in (x_3, x_4) , \\ 0 & \text{for } x > x_4 . \end{cases}$$

The $TR = \langle \hat{a}, r(\cdot), s(\cdot) \rangle$ block contains a polygonal fuzzy number $\hat{a}$ representing transaction amount and two functions $s(\cdot)$ and $r(\cdot)$ that assign weights to classes to which transfer senders and receivers belong. These classes represent types of entities such as "company", "person" or "tax office". Functions $s(\cdot)$ and $r(\cdot)$ have discrete domains and are represented by arrays of real values, one value (weight) per class. The $SER = \langle \hat{m}, \hat{\delta} \rangle$ block describes chains in which transactions are connected in series, using a polygonal fuzzy number $\hat{m}$ to represent the number of transactions in a chain and a polygonal fuzzy number $\hat{\delta}$ (which is intended to measure how much money is transferred transparently through a chain of intermediaries) to represent the ratio of the amount transferred in the last transaction to the amount transferred in the first transaction. The $PAR = \langle \hat{n}, \hat{\Delta}, \tau \rangle$ block describes transaction chains (described by the *SER* pattern) connected in parallel with one account originating all the chains and one account receiving the money at the opposite end. A polygonal fuzzy number $\hat{n}$ represents the number of chains connected in parallel, a polygonal fuzzy number $\hat{\Delta}$ (which is intended to measure how much money is transferred transparently through a network of intermediaries) represents the ratio of the sum of amounts received by the receiving account to the sum of amounts sent from the sending account and an acceptance threshold τ is used for deciding which transaction subgraphs match the pattern. The entire pattern *PAT* used to match suspicious subgraphs consists of three elements which describe the subgraph at each of three levels of hierarchy $PAT = \langle TR, SER, PAR \rangle$. Alternatively it can be denoted as $PAT = \langle \hat{a}, r(\cdot), s(\cdot), \hat{m}, \hat{\delta}, \hat{n}, \hat{\Delta}, \tau \rangle$.

Transaction subgraphs are evaluated using *TR*, *SER* and *PAR* patterns in the following manner. A transfer *T* of amount *a* for which the sender belongs to a class c_s and the receiver to a class c_r is assigned a weight $w_{TR}(T)$ which is calculated, based on $TR = \langle \hat{a}, r(\cdot), s(\cdot) \rangle$, as: $w_{TR}(T) = \mu_{\hat{a}}(a) \cdot s(c_s) \cdot r(c_r)$. Weight of a transaction chain *L* is calculated using $SER = (\hat{m}, \hat{\delta})$ based on its length *m* and the ratio

of amount transferred in the last transaction to the amount transferred in the first transaction δ as:

$$w_{SER}(L) = \frac{\sum_{T \in L} w_{TR}(T)}{m} \cdot \mu_{\hat{m}}(m) \cdot \mu_{\hat{\delta}}(\delta),$$

where T denotes transactions which belong to the transaction chain L. Weight of a subgraph P containing n parallel paths is calculated based on the ratio of the sum of amounts received by the receiving account to the sum of amounts sent from the sending account Δ. The weight $w_{PAR}(P)$ is calculated using $PAR = \langle \hat{n}, \hat{\Delta}, \tau \rangle$ as:

$$w_{PAR}(P) = \frac{\sum_{L \in P} w_{SER}(L)}{n} \cdot \mu_{\hat{n}}(n) \cdot \mu_{\hat{\Delta}}(\Delta),$$

where L denotes transaction chains which belong to the subgraph P.

Fuzzy numbers $\hat{a}, \hat{m}, \hat{\delta}, \hat{n}$ and $\hat{\Delta}$ are represented by 6 real numbers each. Functions $r(\cdot)$ and $s(\cdot)$ are represented by sets of weights assigned to each entity class. $31 + 2k$ real numbers, where k is the number of entity classes, are thus adequate to represent the entire *PAR* pattern. In the genetic algorithm mutation of each of the $31 + 2k$ real numbers in each of the specimens is performed with equal probability P_{mut}. Parameters x_i of fuzzy numbers are mutated by adding a value drawn with a uniform probability from the range . The fuzzy number containing the changed number is then corrected so that $\left[-\frac{R_x}{2}, \frac{R_x}{2}\right]$ the fuzzy number components x_1, x_2, x_3 and x_4 are in correct order and the condition $L_x \le x_1$ and $x_4 \le U_x$ is satisfied. Parameters R_x, L_x and U_x are defined for each fuzzy parameter of the model (i.e. $\hat{a}, \hat{m}, \hat{\delta}, \hat{n}$ and $\hat{\Delta}$) separately. Numbers that represent fuzzy number parameters m_2 and m_3, weights assigned to entity classes and the value of acceptance threshold τ are mutated by adding a random value drawn with uniform probability from the range [–0.005, 0.005]. The value obtained by this addition is then clipped to the range [0, 1]. Selection is performed using a standard roulette-wheel selection procedure [Zhong et al. 2005] and a standard single-point crossover operator [Hasancebi, Erbatur 2000] is used. Probability of performing a crossover on any two specimens is equal to the parameter P_{cross}. For a specimen S which represents a pattern $PAT(S) = < TR(S), SER(S), PAR(S) >$ the evaluation function is calculated in the following way. First, a set P is constructed from all those subgraphs G that match the pattern PAT(S) for which $w_{PAR(S)}(G) > \tau$. Transactions satisfying this condition are deemed suspicious. Transactions in each subgraph G are assigned scores according to the status assigned by a human expert (for example, "illegal" transactions 1.0, "legal" transactions 0.0 and not classified transactions 0.1). Denote a total number of transactions in subgraph G as $t_n(G)$ and a sum of weights of transactions as $t_w(G)$. Then, the evaluation function of specimen S is calculated using values of $t_n(G)$ and $t_w(G)$ obtained for all $G \in P$ as:

$$F(S) = \frac{\sum_{G \in P} w_{PAR(S)}(G) \cdot t_w(G)}{\sum_{G \in P} t_{n(G)}}.$$

Higher values of the evaluation function indicate specimens that are expected to perform better in identifying suspicious transactions in previously unseen data.

3. Experiments

Real life datasets are hard to obtain due to confidentiality of banking data. Thus, in the experiments performed so far we used artificially-generated data which represent transactions from a period of one year from a simulated "mini-economy" in which three classes of economic entities are defined: companies, individual persons and offices (tax offices and social security offices). Companies are characterized by a distribution of the number of employees $N(m_E, \sigma_E)$, a distribution of salary amount $N(m_S, \sigma_S)$, a distribution of the number of goods sold per year $N(m_G, \sigma_G)$, and a distribution of prices of goods $N(m_P, \sigma_P)$. First, a predefined number of companies n_c is generated and for each of them a number of employees n_E is drawn from the Gaussian distribution $N(m_E, \sigma_E)$. Employees are added to the model and for each of the 12 months in a year a transaction representing a salary with the amount a_s drawn from the Gaussian distribution $N(m_S, \sigma_S)$ is generated. For each company the number of goods sold during the year n_G and price of each good are drawn from Gaussian distributions $N(m_G, \sigma_G)$ and $N(m_P, \sigma_P)$. Buyers are selected at random from all employees of all companies. A number n_T of tax offices and the number n_F of social security offices are added to the model. Offices are characterized by a distribution of tax rate $N(m_T, \sigma_T)$ and a distribution of social security fee rate $N(m_F, \sigma_F)$. To each company one tax office and one social security office are assigned at random. For each company a tax rate α_T is drawn from the Gaussian distribution $N(m_T, \sigma_T)$ and a social security fee rate α_F which is drawn from the Gaussian distribution $N(m_F, \sigma_F)$. This represents the variation in tax deduction due to costs etc. Tax amount a_T is calculated based on the sum of payments C_p received by the company as $a_T = C_p \cdot \alpha_T/100$. Social security fee a_F is calculated in a similar fashion based on the sum of salaries paid by the company in each month C_s. To the set of transactions described above n_{ML} money laundering schemes are added which consist of a sender, a receiver and a number n_B of intermediaries who relay money from the sender to the receiver. Generation of money laundering schemes is characterized by a distribution of the amount sent from the sender to one intermediary $N(m_Q, \sigma_Q)$, a distribution of the number of intermediaries $N(m_B, \sigma_B)$ and a distribution of the fraction of the amount received by the intermediary that is forwarded to the receiver: $N(m_\Delta, \sigma_\Delta)$. Tax and social security fee transactions are annotated as "legal", transactions belonging to the generated money

laundering schemes as "illegal" and all the remaining transactions (salaries and payments for goods) as "unknown".

Table 1. Parameters controlling generation of companies for SMALL datasets

Parameter	Company class		
	large	medium	small
n_C	2	4	25
$m_E(\pm\sigma_E)$	5 000 (±1 000)	500 (±100)	50 (±20)
$m_S(\pm\sigma_S)$	6 000 (±1 500)	5 000 (±1 200)	4 000 (±1 000)
$m_G(\pm\sigma_G)$	100 000 (±30 000)	1 000 (±300)	100 (±30)
$m_P(\pm\sigma_P)$	50 (±10)	500 (±100)	500 (±100)

Table 2. Parameters controlling generation of companies for LARGE datasets

Parameter	Company class		
	large	medium	small
n_C	2	8	100
$m_E(\pm\sigma_E)$	5 000 (±1 000)	500 (±100)	50 (±20)
$m_S(\pm\sigma_S)$	6 000 (±1 500)	5 000 (±1 200)	4 000 (±1 000)
$m_G(\pm\sigma_G)$	1 000 000 (±300 000)	10 000 (±3 000)	1 000 (±300)
$m_P(\pm\sigma_P)$	50 (±10)	500 (±100)	500 (±100)

Table 3. The number of accounts and transactions of each type

Object type	Number of objects			
	$SMALL_A$	$SMALL_B$	$LARGE_A$	$LARGE_B$
Accounts	11 238	11 401	19 261	21 270
companies	31	31	110	110
offices	10	10	10	10
personal	11 197	11 360	19 261	21 150
Transactions	294 972	383 463	2 854 965	2 625 671
legal	744	744	2 640	2 640
unknown	289 336	377 207	2 848 435	2 619 049
illegal	4 892	5 512	3 890	3 982
annot. ratio	0.0191	0.0166	0.0023	0.0025

Using data generation method described above, we have generated four data sets: $SMALL_A$, $SMALL_B$, $LARGE_A$, $LARGE_B$ with the same parameters for offices: $n_T = 5$, $n_F = 5$, $m_T = 15$, $\sigma_T = 1.5$, $m_F = 20$, $\sigma_F = 2.0$ and the same parameters of money laundering: $m_Q = 5000$, $\sigma_Q = 1000$, $m_B = 40$, $\sigma_B = 10$, $m_\Delta = 1.0$, $\sigma_\Delta = 0.1$. In these datasets three classes of companies exist. They can be briefly characterized as large (L), medium (M) and small (S). The SMALL and LARGE data contain different numbers of companies in each class. Parameters of companies are summarized in Tables 1

and 2. The number of accounts and transactions of each type in each of the data sets is summarized in Table 3.

In the experiments a population of $N_{pop} = 20$ specimens was trained for $N_{gen} = 20$ generations of genetic algorithm on one of datasets in a LARGE/SMALL pair. The other dataset in the pair was used for testing. Crossover and mutation probabilities were $P_{cross} = 0.1$ and $P_{mut} = 0.01$ and the parameters controlling mutation of x_i components of fuzzy numbers were set as summarized in Table 4. For each pair of SMALL and LARGE datasets 10 independent iterations of tests were performed. After the training has been completed, one, the best specimen was selected from the entire population and it was used for selecting suspicious transaction subgraphs from the testing dataset.

Table 4. Parameters controlling mutation of x_i components of fuzzy numbers

Fuzzy number	Parameter		
	R_x	L_x	U_x
$\hat{a}$	200	range not limited	
$\hat{m}$	x_i not mutated, fixed at 0, 1, 2 and 3, only m_2 and m_3 are mutated		
$\hat{\delta}$	0.1	0.5	1.5
$\hat{n}$	2	3	100
$\hat{\Delta}$	0.1	0.5	1.5

To measure the quality of the detection we recorded the number of "legal", "illegal" and "unknown" transactions that were detected as suspicious. Results are summarized in Table 5. FP is a "false positive" measure, i.e. the ratio of "unknown" transactions among the detected ones. Note, that "legal" transactions were not used for calculating this measure because none of them was marked as suspicious in any of the tests.

Table 5. The number of "legal", "illegal" and "unknown" transactions that were detected as suspicious

Training dataset	Testing dataset	Number of transactions			FP
		legal	unknown	illegal	
$SMALL_A$	$SMALL_B$	0	35	219	15.98%
$SMALL_B$	$SMALL_A$	0	39	242	13.88%
$LARGE_A$	$LARGE_B$	0	66	178	27.05%
$LARGE_B$	$LARGE_A$	0	48	178	21.05%

During the experiments execution time of test iterations was recorded. Timings averaged from 10 iterations are summarized in Table 6.

Table 6. Execution time (in seconds) averaged over 10 test iterations. These tests were performed on the same machine

Training data set	$SMALL_B$	$LARGE_A$	$LARGE_B$
accounts	11 401	19 261	21 270
transactions	383 463	2 854 965	2 625 671
Test data set	$SMALL_A$	$LARGE_B$	$LARGE_A$
accounts	11 238	21 270	19 261
transactions	294 972	2 625 671	2 854 965
Average time	1 165	2 099	2 192

4. Conclusions

In this paper we presented an evolutionary graph mining method which, contrary to data mining methods based solely on transaction features, takes into consideration dependencies between money transfers. We expect this feature to be crucial in detecting illegal activities because single transactions are often structured, so they do not raise the alarm. In the experiments no "legal" transactions were marked as suspicious and more than 2/3 of transactions marked as suspicious were actually involved in money laundering schemes. The remaining 1/3 were transactions for which the simulated expert annotation was "unknown". In real life scenario most of these transactions would actually be legal, however, this group of transactions may also include illegal ones. Computation time comparison has shown a twofold increase of computation time with the similar increase in the number of accounts. Between the same two datasets the difference in the number of transactions was about 10 times.

Further work may focus on improving the precision of the detection but also improving the completeness of the results. Improving computation speed may be important because it would allow searching for more complex subgraph patterns.

References

Buckley J.J., Eslami E. (2002), *Introduction to Fuzzy Logic and Fuzzy Sets*, Physica-Verlag, Heidelberg.

Cook D.J., Holder L.B. (2007), *Mining Graph Data*, John Wiley and Sons, Hoboken.

Fetz Th., Jager J., Koll D., Krenn G., Lessmann H., Oberguggenberger M., Stark R. (1999), Fuzzy models in geotechnical engineering and construction management, *Computer-Aided Civil and Infrastructure Engineering*, Vol. 14, No. 2, pp. 93–106.

Goldberg D. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading.

Goldberg D., Sastry K. (2011), *Genetic Algorithms: The Design of Innovation*, Springer

Hasancebi O., Erbatur F. (2000), Evaluation of crossover techniques in genetic algorithm based optimum structural design, *Computers & Structures*, Vol. 78, No. 1–3, pp. 435–448.

Korczak J., Marchelski W., Oleszkiewicz B. (2008), A new technogical approach to money laundering discovery using analytical SQL server, [in:] J. Korczak, H. Dudycz, M. Dyczkowski (Eds.), *Advanced Information Technologies for Management – AITM 2008*, Research Papers of Wrocław University of Economics No. 35, Wrocław University of Economics, Wrocław, pp. 80–104.

Korczak J., Oleszkiewicz B. (2009), Modelling of data warehouse dimensions for AML systems, [in:] J. Korczak, H. Dudycz, M. Dyczkowski (Eds.), *Advanced Information Technologies for Management – AITM 2009*, Research Papers of Wrocław University of Economics No. 85, Wrocław University of Economics, Wrocław, pp. 146–159.

Truman E.M., Reuter P. (2004), *Chasing Dirty Money: The Fight Against Anti-money Laundering*, Peterson Institute for International Economics.

Zhong J., Hu X., Zhang J., Gu M. (2005), Comparison of performance between different selection strategies on simple genetic algorithms, [in:] *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce Vol-2 (CIMCA-IAWTIC'06)*, Vol. 02, IEEE Computer Society, pp. 1115–1121.

EWOLUCYJNE DRĄŻENIE GRAFÓW W WYKRYWANIU PODEJRZANYCH TRANSAKCJI

Streszczenie: W procederze prania brudnych pieniędzy wykorzystywane są złożone schematy organizacyjne mające na celu ukrycie prawdziwego celu wykonywanych transakcji. W tej publikacji opisana została metoda drążenia grafów, która pozwala na wykrywanie podgrafów zawierających podejrzane transakcje. Model reprezentujący podejrzane podgrafy jest parametryzowany za pomocą liczb rozmytych, które reprezentują parametry transakcji oraz niektóre własności strukturalne modelowanych podgrafów. Prezentowana metoda dokonuje rozmytego dopasowania struktury grafów, co pozwala na wykrywanie także takich podgrafów, które do pewnego stopnia różnią się od tych, które zostały zaanotowane przez eksperta.

Słowa kluczowe: drążenie grafów, algorytmy ewolucyjne, pranie pieniędzy.