Graph-Network Models and Methods Used to Detect Financial Crimes with IAFEC Graphs IT Tool

Zbigniew Tarapata[1,*], *Rafal* Kasprzyk[1], *Kamil* Banach[1]

[1] Institute of Computer and Information Systems, Faculty of Cybernetics, Military University of Technology, Warsaw, Poland

Abstract. The article outlines graph-network models used to detect financial crimes. The general graph-network model of transaction participants was defined. The article also provides some examples of the graph-network models used to detect financial crimes. The authors also proposed an original method for detecting financial crimes based on measurements of the network characteristics and graph similarity, used in the R&D project: "Advanced information technologies supporting the (mainly financial) data analysis processes in the area of financial crimes" sponsored by National Centre of Research and Development (NCBiR), Poland. The method is based on the use of measures of the selected graph characteristics and similarity of graph elements. Some examples of the use of the proposed measures and methods on the basis of the IAFEC Graphs IT tool, which was created as part of the project, were described. Tools options, such as filtering graphs, determining clusters, determining centrality measures, or links between network elements are presented. The IAFEC Graphs architecture was presented both from the software point of view (system-to-layer division) and hardware requirements. In addition, an example of the use of the described functions of the tool for examining links between entities contained in public registers in Poland is presented.

1 Introduction

The models and methods for detecting financial crimes are often based on mathematical models connected with graphs and networks [1], [2]. This is due to the fact that by using the graph, it is possible to describe the structure of any real object, i.e. the set of interdependent elements. The real objects may be in the form of links between persons, companies, chattels, real properties, transactions as well as financial transactions. The majority of methods described in the literature consist in grouping all available cases, analysed using a combination of the following techniques:

- clusterization and pattern recognition [3],
- link analysis [4], [5], [6],
- association rules [5],
- connected graphs [4], [5],
- mining frequent patterns [7], [8].

Only some of the approaches use the temporary analysis of patterns, e.g. dynamic Bayesian networks [9] or probabilistic approach [5]. According to the majority of methods, it is necessary for the user and/or expert to define some of the parameters, i.a. threshold values. However, it is not a general rule. The described methods create and analyse the structures (of data) including (single) transactions, accounts, institutions and groups of institutions, whereas the two-last mentioned include more premises (of data about the participants, transactions and/or attributes thereof) concerning the analysed crimes. The methods of analysis of individual and group cases (of participants in transactions) are usually executed separately if such separation is introduced.

In the paper original method for detecting financial crimes based on measurements of the network characteristics and functions of some IT tool (IAFEC Graphs) implementing the method were presented.

2 General graph-network model of transaction participants

A graph may constitute a model for the network of transaction participants. Generally speaking, two basic types of graphs [2]: directed graph and undirected graph are distinguished.

Directed graph G_s:

$$G_s = \langle V, A \rangle \quad A \subset V \times V \tag{1}$$

Undirected graph G:

$$G = \langle V, E \rangle \quad E \subset \{\{x, y\} \subset V\} \tag{2}$$

Mixed graph G_m:

$$G_m = \langle V, E \cup A \rangle \tag{3}$$

where:
V – *set of nodes, vertices*,
E – *set of edges*,

* Corresponding author: zbigniew.tarapata@wat.edu.pl

A – set of arcs (sometimes referred to as directed edges).

Let $|V| = N$, $|E| = M$ ($|A| = M$).

In compliance with the above-mentioned definitions, the graph edge shall be defined through a two-element subset of the set of vertices (order of elements in the subset insignificant – set property), whereas the graph arc - through the ordered pair of vertices (order of elements in pair significant – property of the ordered pair).

Directed network (weighted directed graph):

$$S_s = \left\langle G_s, \{\xi_i\}_{i=\overline{1,I}}, \{\Psi_j\}_{j=\overline{1,J_s}} \right\rangle \tag{4}$$

Undirected network:

$$S = \left\langle G, \{\xi_i\}_{i=\overline{1,I}}, \{\Phi_j\}_{j=\overline{1,J}} \right\rangle \tag{5}$$

Mixed network:

$$S_m = \left\langle G_m, \{\xi_i\}_{i=\overline{1,I}}, \{\Phi_j\}_{j=\overline{1,J}} \cup \{\Psi_j\}_{j=\overline{1,J_s}} \right\rangle \tag{6}$$

where:

$$\xi_i : V \to X_i \quad , \quad i = \overline{1,I}$$
$$\Psi_j : A \to Y_j \quad , \quad j = \overline{1,J_s}$$
$$\Phi_j : E \to Y_j \quad , \quad j = \overline{1,J}$$

most often: $X_i = R$, $Y_j = R$, where R – the set of real numbers.

It is worth distinguishing between the terms *graph* and *network*. The network is defined as the quantified (weighted) graph, i.e. the graph, on whose vertices and/or edges/arcs, some functions, whose interpretation depends on the type of the object modelled by the network (e.g. in the transaction network: number of transactions between two participants in the transaction modelled by the edge/arc, value of such transaction, etc.), were described. The terms "graph" and "network" are often used interchangeably, even though it is incorrect from a formal point of view. In the subsequent part of the article, where such distinction may give rise to confusion, the terms "graph" and "network" shall be used interchangeably, however, bearing in mind the fact that the graph describes only the structure of a real object (system), whereas the network, apart from the structure, also describes quantitative characteristics of such object (system).

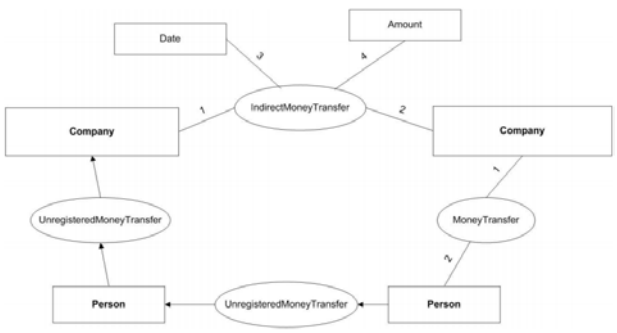

Fig. 1. Graph pattern of a financial crime, source: [10]

Fig. 1 presents graph pattern of a financial crime. Persons, companies, transfers, amount represent the graph vertices (directed edges), whereas arcs represent the links between such vertices. Some functions described on the vertices or edges can represent quantitative or qualitative characteristics of the structure.

3 Original method for detecting financial crimes based on measurements of the network characteristics

In this chapter, the original method for detecting financial crimes based on measurements of the network characteristics and graph similarity is presented. Some of the selected elements of the method were implemented in the *IAFEC Graphs* tool described in [10].

The fundamental question arises as to the method for measuring the network characteristics. The measurement method should allow to find such network characteristics, which are important from the point of view of the analysis of the system modelled by the network. The following centrality measures of vertices were proposed [1], [11] including their interpretation in terms of detecting financial crimes. The centrality measures are used to measure characteristic network vertices and to determine which vertex of the network is the most important from the point of view of a given measure.

Normalized degree dc_i of vertex i:

$$dc_i = \frac{k_i}{N-1} \tag{7}$$

where: k_i means a degree (total number of edges "adjacent" to the vertex and arcs coming "in" and "out" of the vertex) of vertex i on graph G.

The vertex, which has the highest degree (thus, the greatest proximity), obtains the highest value of such measure.

Radius rc_i of vertex i:

$$rc_i = \frac{1}{\max\limits_{j \in V} d_{ij}} \tag{8}$$

where: d_{ij} – length of the shortest route (alternating sequence of vertices and edges/arcs adjacent to each other, starting on a given initial vertex and ending on a given final vertex) on graph G between vertices i and j (length of the route on graph between vertices i and j = number of edges and arcs on the route from i to j).

The vertex that is as close as possible to all the furthest network vertices (the shortest distance separating the vertex from the furthest vertex) gets the highest rating.

Closeness cc_i of vertex i:

$$cc_i = \frac{N-1}{\sum\limits_{j \in V} d_{ij}} \tag{9}$$

According to this measure, the vertex is the more central, the closer it gets to all other network vertices. As a result, the measure allows to determine which of the two random vertices requires fewer steps to "communicate" with any other network vertex.

Betweenness bc_i of vertex i:

$$bc_i = \frac{\sum_{l \in V} \sum_{k \neq l \in V} \frac{p_{l,i,k}}{p_{l,k}}}{(N-2)(N-1)} \quad (10)$$

where: $p_{l,i,k}$ – number of routes in graph G between vertices l and k crossing i.

The measure is defined as a fraction of the number of the shortest routes joining two random network vertices, which include a given vertex. Usually, the value of such measure is normalized by including a maximum potential number of the shortest routes in the graph, which is a complete graph. A complete graph is the graph, where every vertex is connected with all other vertices. In graph, with N vertices, it is possible to have $N(N\text{-}1)$ of such connections

Clusterization gc_i of vertex i:

$$gc_i = \frac{2E_i}{k_i(k_i-1)}, \; k_i > 1 \quad (11)$$

where: E_i – number of edges between neighbours of vertex i.

The measure describes a "probability" that the first (i.e. the closest) neighbours[a] of vertex i are also their own first neighbours. It turns out that the measure has an interesting sociological interpretation. When social networks are being shaped, people with similar views, interests, etc. are clustered (grouped) together. What is more, third parties may also affect our relationships with other people. Therefore, sociologists analysing the interactions in social networks started to examine the relationships between third parties as well. The relationships inside such a triangle are much more complex and may lead to some very interesting conclusions. The clusterization process reflects exactly this kind of a situation, when a ratio of the number of triangles[b] in the network to the number of all triangles that may potentially occur is taken into account. A good example of such triangle is the subgraph constructed on vertices numbers: 9, 10, 11 in Fig. 2.

The graph in Fig. 2 describes a certain network of 11 transaction participants (persons are vertices, whereas edges represent direct transactions between such persons, e.g. wire transfers). The following central vertices (i.e. the vertices, in case of which the measure value is the highest) shall be obtained for such network and for the particular centrality measures:

- $dc_i - i = 3$; meaning that person number 3 was involved in the largest number of transactions in the aforesaid network (i.e. the person made and/or received the largest number of wire transfers, equal to 4 (from/to persons: 1, 2, 4, 5));
- $rc_i - i \in \{6, 7\}$; meaning that persons numbers 6 and 7 need the lowest number of indirect transactions (through persons, with whom they directly execute transactions, and persons, with whom such persons directly execute transactions, etc.) to execute the transaction with the furthest person in the network;
- $cc_i - i = 6$; meaning that person number 6 needs the lowest number of transactions in this network so that every person in the network could be involved in the transaction (directly or indirectly) with such person;
- $bc_i - i = 8$; meaning that person number 8 shall be most often asked for mediation in the transactions between persons from the entire network, i.e. to execute the transaction with some persons in the network, person number 8 must act as a mediator (e.g. each person 1,...,6 must ask person 8 about the transaction with person number 9, thanks to which they shall be able to execute transactions with persons numbers 10 and 11);
- $gc_i - i = 9$; meaning that person number 9 knows the highest number of such persons (exactly 2 persons), with whom person number 9 may be involved in direct transactions and who also execute such direct transactions between themselves (10 and 11).

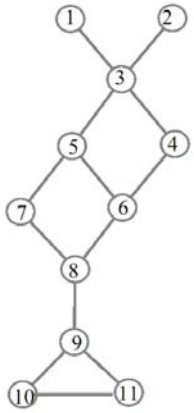

Fig. 2. Graph example to analyse the centrality measures of vertices, source: [1]

The previously described characteristics concerning the graph (network) vertices. Two basic characteristics of the network: average distance and (average) clusterization of the network shall be defined.

Average distance L (average length of the shortest routes) in the network:

$$L = \frac{\sum_{i \neq j \in V} d_{ij}}{N(N-1)} \quad (12)$$

Clusterization (average) C of the network:

$$C = \frac{1}{N} \sum_{i \in V} gc_i \quad (13)$$

[a] In plain language, the first neighbors of a given vertex are such vertices that are directly connected with the edge (arc). Further (second, third, etc.) neighbors are the vertices, with which a given vertex is linked indirectly through its neighbors (closer and further).

[b] A triangle in the graph (network) is the part (subgraph) composed of three vertices, where each vertex is connected with the every other vertex out of the two remaining vertices through the edge (arc). In other words, it is the so-called clique in the graph (a subgraph, where each vertex from the clique is connected with every other vertex through the edge), size 3.

4 Description of IAFEC Graphs Tool

The tool was created as an online application available via the web browser (i.e. web application). It allows to use any environment, where modern web browser is available. Additionally, more than one analysts may use the tool simultaneously.

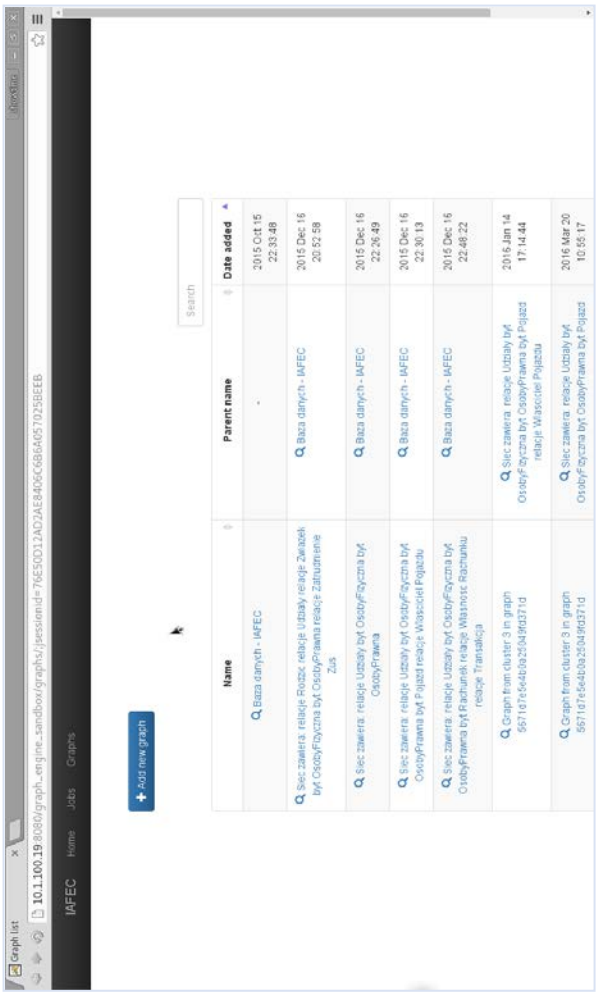

Fig. 3. Window with the list of the networks available in IAFEC Graphs, source: [10]

IAFEC Graphs allow to perform a number of operations enabling the analyst to analyse the network. One of such operations, which is usually the first step of the analysis, is the refiltering of the network in such a manner so that it includes vertices and edges/arcs of the types in terms of the analysed case. Therefore, it is possible to obtain the network, which includes, for example, the information only on people, legal entities and shares (such network is shown in Fig. 3, it is the third network in the table counting from the top).

Fig. 4 shows the structure of links between entities from public registers as presented in *IAFEC Graphs* [10]. The graph vertices represent appropriate entries in public registers in Poland, such as: Personal Identification Number (PESEL), National Business Registry Number (REGON), Tax Identification Number (NIP), Vehicle Identification Number (VIN), Number of Land Register, Number of the National Court Register (KRS), etc. The edges or arcs (directed edges) of the graph represent the type of links between different vertices: property, relationship, parent-child, employment, etc.

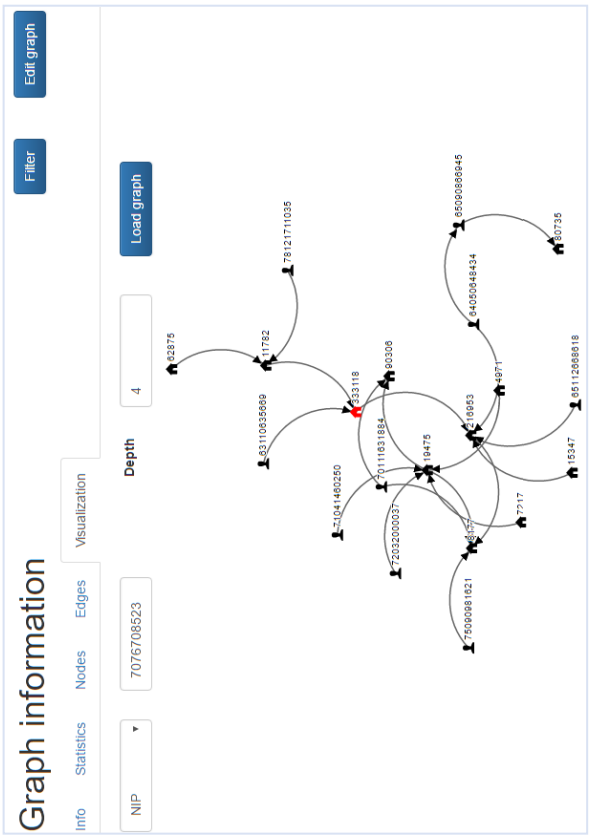

Fig. 4. Part of the graph representing the structure of links between entities from public registers as presented in IAFEC Graphs, source: [10]

It is possible to set clusters (i.e. subnetworks, which are not interconnected in a given network) and then use the network for the purpose of the analysis in the same way as in the filtered network. It allows to focus on the largest interconnected groups of entities during the analysis.

When using the tool, the analyst has a number of graph-network algorithms for their disposal, which are based on the original method for detecting financial crimes as described in previous section. Therefore, the analyst may set the shortest route between two random entities, average length of routes in a given network as well as various measures for the network characteristics in the form of centrality measures, thanks to which it is possible to determine which vertex is the most important in the network (from the point of view of a given measure).

The tool also provides for a possibility of calculating the following centrality measures of vertices:

- normalized degree,
- closeness,
- radius,
- betweenness,
- clusterization.

The IAFEC Graphs IT tool was divided into six layers, responsible for implementing particular fragments of the functionalities (see Fig.5). Each of them

works independently, using only such methods and services that are made available by those other layers. It uses MongoDB as a distributed database and Neo4j graph platform.

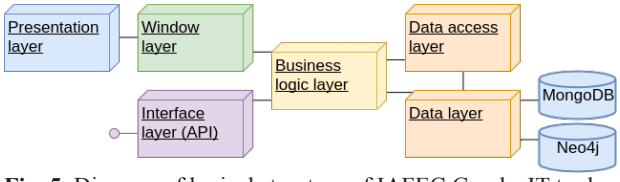

Fig. 5. Diagram of logical structure of IAFEC Graphs IT tool

1. Presentation layer: the layer is responsible for presenting and making available the user windows. The user's browser constitutes the area of operations of particular components included in this layer.
2. Interface layer (API): the layer makes it possible for the system client (presentation layer, which the user uses via the browser) to use a set of services allowing to modify individual business objects of the system, such as graphs, vertices, branches, clusters or tasks performed or completed by the system.
3. Window layer: the layer is responsible for preparing user windows and handling the incoming requests.
4. Business logic layer: the layer is responsible for the operations changing the status of the business objects. It is used by the interface layer and provides additional abstraction layer between the window layer and data access layer. Therefore, it is possible to easily "exchange" the data access layer so that it may use another database "engine".
5. Data access layer: the layer allows to save, read, modify and delete business objects from the database. To that end, it provides appropriate methods for performing such operations.
6. Data layer: the layer includes the data model structure processed by the application – i.a. the structure of the graph vertex and branches.

5 Example of experiments

5.1 Finding connections between two entities

Analysts common practice is finding connections between two entities (i.e. between person and legal entity). It can be done by calculating shortest route between them. This task can be done using IAFEC Graphs. Analyst can run a job to find a route between two vertices in graph by specifying their identification number (i.e. ID or VAT number). After calculation result is presented on graph.

Fig. 5 shows the shortest route between person $s = 69121797880$ and person $t = 45082422315$ in the structure of links between entities included in public registers, as disclosed in *IAFEC Graphs* [10] (see: description in subchapter 3). The analysis of the shortest route between two vertices allows to determine:

- whether the two given persons (vertices) are in any way linked (indirectly; the direct link is shown in the graph as an edge or arc between the vertices). If the route between them exists, they are connected: =the shortest route, the strongest link. One of the methods for hiding links between the persons and companies (through bank accounts, companies, real estates, etc.) is an attempted extension of such a route (i.e. "mules", large number of fictional or actual intermediaries, etc.) to make it more difficult to find any links;
- who or what (persons, companies, real estates, etc.) and how (through persons, companies, real estates, etc.) connects such two vertices; all vertices on the route from *s* to *t* are linked indirectly or directly.

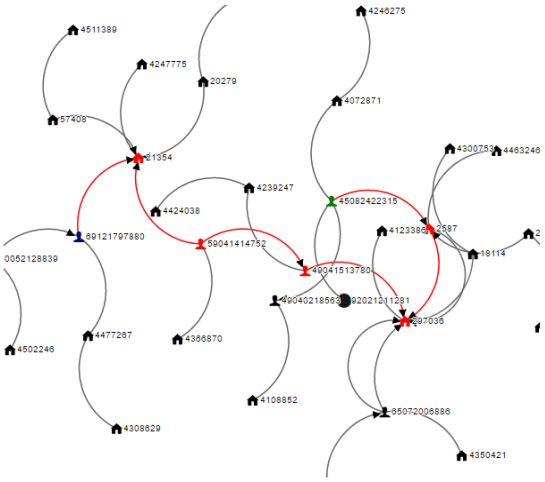

Fig. 6. The shortest route between person s = 69121797880 and person t = 45082422315 in the structure of links between entities included in public registers, as disclosed in IAFEC Graphs, in the subgraph, where - for nodes existing on the route from s to t - the nodes at a distance of maximum 2 therefrom were shown

The analyst would like to set the most important vertex for selected network using the centrality measures. Therefore, the analyst shall launch the betweenness algorithm for the new graph. The result of such algorithm in the form of a table is in the Fig. 7. The figure shows that the most important vertex, from the point of view of such measure, is vertex number NIP 7076708523. The tool also allows to view the graph, starting from a given vertex, so as to show the nodes, which are at a certain distance from the vertex. The figure includes such visualization – the initial vertex is the one described above and all vertices at a distance of 4 therefrom are also shown.

5.2 Finding most important entity

Another common task for analyst is finding most important entity in some network. The analyst, after selecting network, can run job that will calculate centrality measure for it. In mentioned example analyst used *betweenness* centrality.

Fig. 7 shows the values of the *betweenness* measure for the vertices of the graph, whose subgraph is shown in Fig. 6. Vertex 333118 (identical to 7076708523) has the

highest measure value, which means that such vertex most often mediates between the subgraph vertices. Such vertex shall most often (in comparison with other vertices) appear on the shortest route between any pair of the subgraph vertices. On the other hand, the value of vertices nos. 65090866945 and 71041460250 is equal to zero, which means that such vertices do not appear on any shortest route connecting two random subgraph vertices.

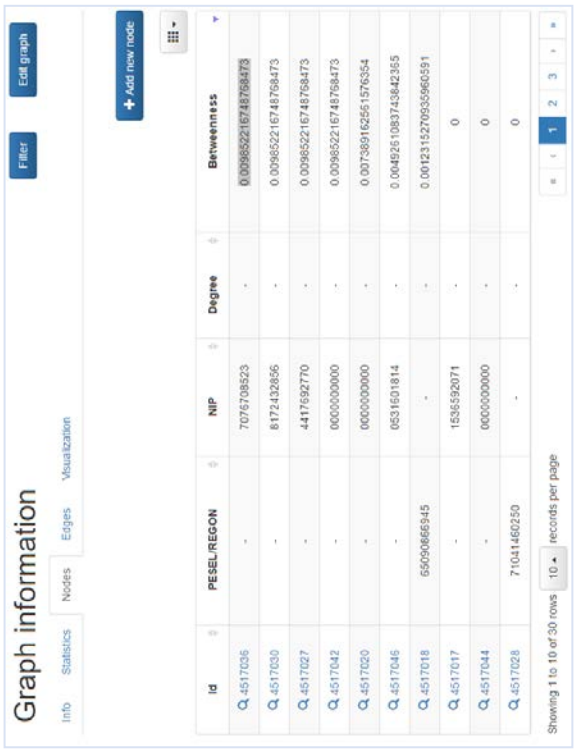

Fig. 7. Values of the betweenness measure for the vertices of the graph, whose subgraph is shown in Fig. 6. The vertices are sorted in descending order with respect to the measure value, thus, some of them are not visible on the first page of the table, source: [10].

Summary

The article presents graph-network models and methods used to detect financial crimes. The general graph-network model of transaction participants was defined. The authors proposed an original method for detecting financial crimes based on measurements of the network characteristics and graph similarity. The method is based on the use of measures of the selected graph characteristics and similar graph elements. Some examples of the use of the proposed measures and methods on the basis of the IAFEC Graphs IT tool [10] were described. The presented method can be extended using weighted graphs similarity approach [12], [13], [14] for graph-based financial crime pattern recognition.

References

1. C. Bartosiak, R. Kasprzyk, Z. Tarapata, Application of Graphs and Networks Similarity Measures for Analyzing Complex Networks, Biuletyn Instytutu Systemów Informatycznych, vol. **7**, 1-7 (2011)
2. R. Diestel, *Graph Theory*, Springer-Verlag, Berlin Heidelberg, (2005)
3. R. Dreżeski, W. Filipkowski, System Supporting Money Laundering Detection, Digital Investigation, Vol. **9(1)**, 8-21, (2012)
4. Cz. Jedrzejek, J. Bak, M. Falkowski, Graph Mining for Detection of a Large Class of Financial Crimes, *Proceedings of the 17th International Conference on Conceptual Structures*, Moscow, Russia, 26–31 July, (2009)
5. G. Krishnapriy, M. Prabakaran, Identifying Money Laundering Groups in MultiMode Network Using Data Mining, International Journal of Enhanced Research in Science Technology & Engineering, Vol. **3(5)**, 291–295 (2014)
6. Zhongfei Zhang, Philip S.Yu, Applying Data Mining in Investigating Money Laundering Crimes, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 24 – 27, (2003)
7. G. Krishnapriya, M. Prabakaran, A Probabilistic Model Using Graph Based Sequential Pattern Mining Algorithm for Money Laundering Identification, International Journal of Innovative Research and Development, Vol. **2(7)**, 497–502 (2013)
8. X. Luo, Suspicious transaction detection for anti-money laundering, International Journal of Security and Its Applications, Vol. **8(2)**, 157–166 (2014)
9. S. Raza, S. Haider, Suspicious activity reporting using dynamic Bayesian networks, Procedia Computer Science, 3, 987–991 (2011)
10. R. Kasprzyk, Z. Tarapata, K. Banach, M. Parada, D. Bocian, Narzędzie informatyczne IAFEC Graphs do wykrywania zależności między elementami zbiorów rejestrów publicznych, w: M. Kiedrowicz (red.), *Zaawansowane modele i metody wykorzystywane w zwalczaniu przestępstw finansowych*, WAT, Warszawa, (2018)
11. R. Kasprzyk, *Complex Systems Evolution Models and Methods to Investigate Their Characteristics for Computer Identification of Potential Crises* (Polish title: *Modele ewolucji systemów złożonych i metody badania ich charakterystyk dla potrzeb komputerowej identyfikacji potencjalnych sytuacji kryzysowych*), PhD thesis, Military University of Technology, Warsaw, Poland, (2012)
12. Z. Tarapata, R. Kasprzyk, Graph-based optimization method for information diffusion and attack durability in networks, Lecture Notes in Artificial Intelligence, Vol. **6086**, 698-709, (2010)
13. Z. Tarapata, Multicriteria weighted graphs similarity and its application for decision situation pattern matching problem, *Proceedings of the 13th IEEE/IFAC International Conference on Methods and Models in Automation and Robotics* (MMAR'2007), 27–30 August, Szczecin, Poland, pp. 1149–1155, (2007)
14. M. Chmielewski, M. Paciorkowska, M. Kiedrowicz, A semantic similarity evaluation method and a tool utilised in security applications based on ontology structure and lexicon analysis, *Fourth International Conference on Mathematics and Computers in Sciences and in Industry*, pp. 224-233, DOI 10.1109/MCSI.2017.46 (2017)