Big Data and Specific Analysis Methods for Insurance Fraud Detection

Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA
University of Economic Studies, Bucharest, Romania
ramona.bologa@ie.ase.ro, razvanbologa@ase.ro, alexandra.florea@ie.ase.ro

Analytics is the future of big data because only transforming data into information gives them value and can turn data in business in competitive advantage. Large data volumes, their variety and the increasing speed their growth, stretch the boundaries of traditional data warehouses and ETL tools. This paper investigates the benefits of Big Data technology and main methods of analysis that can be applied to the particular case of fraud detection in public health insurance system in Romania.

Keywords: *Big Data, Social Networks, Data Mining, Fraud Detection*

1 Introduction

Health budgets are a common target of fraudulent practices. Due to the complicated nature of medical processes, frauds have always found a favorable environment in the health insurance system.

Since fraud is on, the increase holistic fraud prevention is required. According to the well know market research organization Gartner [9]: “Security and fraud risk exposure is increasing as organizations are threatened at multiple points of vulnerability. Companies are re-evaluating how they tackle security since a fragmented approach is consistently leaving organizations at greater risk of attack. A more holistic approach to security ensures all layers of protection function together”.

Electronic health cards with smart chips have been implemented in order to fight fraud in health insurance. The use of eHealth cards has been generating a huge amount of data that needs to be processed. The conventional database technologies are not suitable for performing this type of analysis due to their inner limitations. The new big data technologies are yet to be understood and the benefits they provide in fighting fraud need to be investigated.

By performing big data analysis, common repetitive errors that are "hidden" inside huge repositories of data can be identified and corrected. Such errors would go undetected in the absence of big data technologies because the human brain is not capable to correlate the huge quantities of data available in the medical sector.

In order to prevent insurance fraud, big data analytics should use the following technologies: business rules, anomaly detection, text mining, database searches and social network analysis. These technologies will be approached during the following sections.

2. Big data technology - advantages and challenges

The subject of big data is of major interest to the scientific community as the size of the databases has been growing beyond the limits of current technologies. As indicated in the bibliography section, there is consistent research of big data technologies with applications into the health sector.

There are currently many research initiatives for developing big data technologies. Some of these initiatives are funded by private companies such as IBM, ORACLE, SAS and Microsoft. Other initiatives are funded by public bodies and/or the open source community. A notable technology open source is Hadoop which is often integrated with commercial technologies.

However, the applications of such technologies are still to be developed as they are highly dependent on each sector of activity and on each geographical area.

Although massive data amounts were produced during the last two years, the term "big data" was present in the research literature starting with 1970s, but it has seen an explosion of publications since 2008 [3].

Wikipedia defines big data as "a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications" [12].

The industry standard definition of Big Data projects it along four dimensions: volume, velocity, variety and veracity [17].

Volume dimension refers to the huge data volume produced or manipulated by a company that must be further manipulated in order to get useful information.

Velocity dimension refers to speed of data processing, as some activities need real-time responses. Distributed and parallel processing algorithms become very important for this reason. During the fraud detection process it is very important to analyze day-by-day big data flows, millions of detailed record that must be scrutinized to get a behavior pattern identification.

Variety dimension refers to various types of data that are manipulated by a company, both structured and unstructured (text, audio, video, click streams, log files, sensor data etc.).

Veracity dimension refers to the information level of trust granted by the business decision factors.

More than that, when working with big data, the meaning of each event can be interpreted only in relationship with preceding events. So, we heave streams of data that must be analyzed all together, like a sequence, so the traditional analytic methods work poorly on these cases. Traditional analytic tools approach data at entity level, as each entity provides useful information. The shift to detailed stream data changes the needs and requires for complex ETL tools.

First used by Internet giants like Yahoo, Ebay or Facebook, Apache Hadoop is the most popular big data platform. Hadoop is an open source platform for processing big data that uses distributed processing across clusters of servers. It has become the "de facto" standard for storing, processing and analyzing huge amounts of data.

Hadoop is a Java based framework and uses simple parallel programming models using clusters of inexpensive servers that locally store and process huge volumes of data. The result is a fundamental decrease of data storage cost. The analysts are free to write code in almost any contemporary language using the streaming APIs available in Hadoop.

The platform offers a high level of scalability as processing requirements are distributed on thousands of machines and its software is designed to detect and solve failures at application level. This way, its clusters are very resilient.

Core Hadoop has two main systems:

- ***Hadoop Distributed File System (HDFS)***: self-healing high-bandwidth clustered storage.
- ***MapReduce***: distributed fault-tolerant resource management and scheduling coupled with a scalable data programming abstraction.

In the beginning (2008), Hadoop had significantly less capabilities then relational databases and had limited supporting tools. But now, it has more robust SQL capabilities and access to all SQL-based applications. Cloudera was the first that introduced commercial support for Hadoop in 2008, followed by MapR and Hortonworks. IBM and EMC have each its own Hadoop distribution. Microsoft and Teradata offer complementary software Hortonworks' platform. Oracle resells and supports Cloudera, while HP, SAP work with multiple Hadoop software providers [16].

Classic business intelligence tools use

relational databases for storage and query execution. In order to use the traditional analysis methods and techniques, there have been many efforts to develop SQL-like languages for big data access, query and manipulation: BigSQL, HiveQL, CassandraQL, JAQL, Sparql, Shark etc, each of them associated with a specific big data platform.

Many users concluded that no type of big data is optimal for all their requirements. Today there are many implementations of hybrid big data architecture, which combine two or more technologies in specialized roles (see Fig.1). For example, combining Hadoop for unstructured data staging with in-memory business intelligence tools for query acceleration, with stream computing for continuous data provision and with massively parallel processing RDBMS for data warehousing and data management [18].

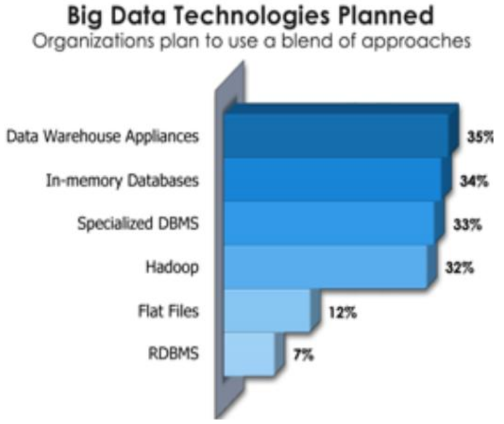


Fig. 1. Hybrid technology approach [21]

Big Data Connectors are used to combine RDBMS with Hadoop for deeper analysis, using data mining or statistical analysis.

More than that, in order to enable more flexible topologies of Hadoop and non-Hadoop solutions, a standard query virtualization layer can be developed to support a transparent SQL access to any platform [18].

A scan of the list of projects reported as using big data techniques by PoweredBy Hadoop [20], suggests that the big data approach is best suited for business problems that meet one or more of the following criteria [5]:

- Data-restricted throttling
- Computation-restricted throttling
- Large data volumes
- Significant data variety
- Benefits from data parallelization.

The use of big data for fraud prevention has a huge potential as the data generated by the current transactional systems is enormous and current database technologies are unable to process it. All the five criteria listed before are respected, which make fraud prevention a perfect suited application for big data analytics.

The next section will address traditional methods of analysis for the detection of fraud and how these can be exploited in the context of the big

3. Fraud in the health insurance system in Romania

The main issue in fraud detection is the fact that the collections of data are impossible to be processed by a human brain. For instance, a controller could observe that in a short period of time all the inhabitants located on a single street did a set of expensive laboratory tests, say for hormonal disorders, which a medical lab is charging to the health insurance system. This is clearly a fraud as it highly unlikely that a very limited number of persons located on a single street go, in a short period of time, to take such rare and expensive tests.

It is easy to suspect that the doctor who signed the test results probably did not act on his own and that such frauds most probably happened before. The controller might want to analyze not only the activity of the doctor that signed the test results, but also the activity of the persons from his/her social network (colleagues, managers, previous managers and others).

A tool that allows such analysis for the health insurance system does not exist on the current market. During discussions with global leading companies, the Romanian Health Insurance Agency discovered that such tools exist for the banking industry but nobody adapted them to the health sector.

National Health Insurance System in Romania has been continuously restructured during the last 20 years and the following paragraphs summarize the main legislation relating to these changes, according to the site of the National House of Health Insurance [19].

The Law Social Health Insurance - Law no. 145/1997 was adopted in June 1997. This followed the type Bismarck insurance model with compulsory health insurance based on the principle of solidarity and operating under a decentralized system. It came into force on January 1, 1999. In consequence, from 1 January 1999, Insurance Houses have functioned as autonomous public institutions, led by representatives of the insured and employers through the boards, as well as the National House of Health Insurance (19).

O.U.G.no.150/20.11.2002 on “Organization and functioning of health insurance" repealed law no. 145/1997. This allowed conceptual and structural changes of the health insurance system as a unified system of financing care and promoting health. According to this normative act, social health insurance system in Romania has three major components:

- Insured person;
- Health care providers (doctors, hospitals, pharmacies);
- Health insurance houses (tertiary payer).

Law no. 95/2006 has produced a new health system reform in Romania, imposing more flexibility and dynamism, giving clear responsibilities to ensure both logistics for the coordinated functioning of the health insurance system (by collecting and efficient use of funds), and appropriate means for representing, informing and supporting the interests of the insured.

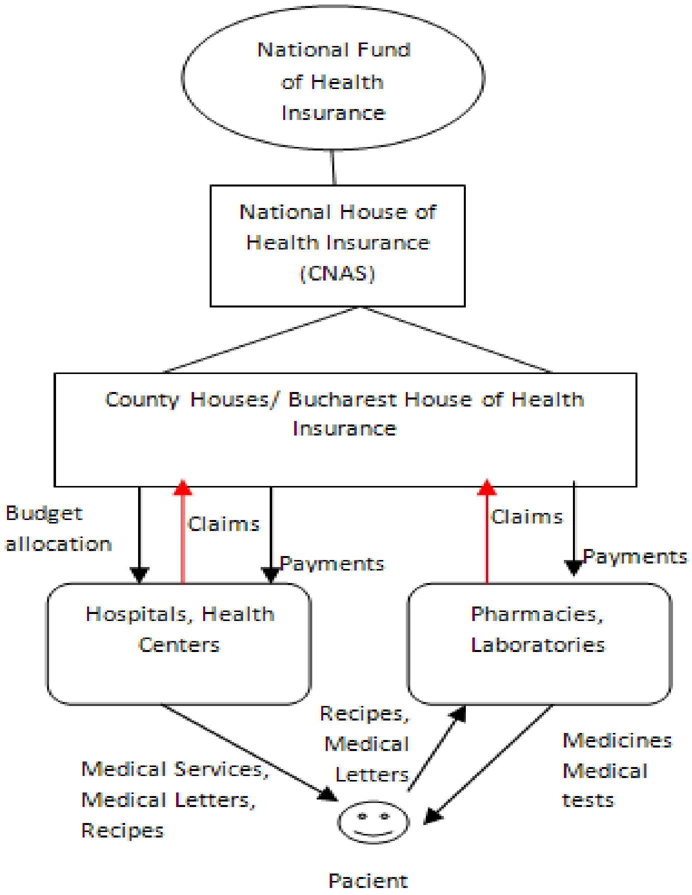

National Fund of Health Insurance
National House of Health Insurance (CNAS)
County Houses/ Bucharest House of Health Insurance
Budget allocation
Claims
Payments
Claims
Payments
Hospitals, Health Centers
Pharmacies, Laboratories
Recipes, Medical Letters
Medical Services, Medical Letters, Recipes
Medicines Medical tests
Pacient

Fig. 2. Main entities involved Health Insurance System

Basic principles for the organization and functioning of the National House of Health Insurance are:

- Free choice of health insurance house;
- Solidarity and subsidiarity in the collection and use of funds;
- Free choice of family doctor, physician and health unit;
- Mandatory participation to health insurance contribution payment for the formation of National Fund of Health Insurance;
- Participation of insured persons, employers and government to the National Fund of Health Insurance management;
- Provide a package of basic health services, equitably and without discrimination of any insured;
- Transparency health insurance system.

From a financial perspective, the Budget of the National Health Insurance Fund is approved by the Annual State Budget Law and takes into consideration:

- Current requirement of medical activity;
- Amounts representing arrears recorded in hospitals and open circuit pharmacies;
- Financial resources in each county.

In Figure 2, flows which are recording most cases of fraud are marked with red color. The main types of fraud that can be identified in the Romanian insurance health system are mostly similar to other European health systems [5]:

- Unusual high number of invoices for a particular insured person in a short time (3-4 days);
- Use of false identities for claiming false hospitalization, false prescriptions or other false health care services;
- Claiming medical invoices having dates outside the insurance period;
- Excessive number of medical claims in a certain period;
- An excessive number of manual invoices requests whose values are usually lower than the limit of inspection;
- Claims having payable amounts higher than the billed amounts that insurance house will pay.

4. Analysis methods for detecting fraud in health insurance

Insurance fraud can be defined as “knowingly making a fictitious claim, inflating a claim or adding extra items to a claim, or being in any way dishonest with the intention of gaining more than legitimate entitlement” [2].

The current health issuance fraud is about 5% of the health budgets. In Romania alone this amount represents around 250-300 million Euros/year [15]. Most of this fraud would be detectable by clever data analytics. The area of fraud prevention has been traditionally correlated with data mining and text mining. Even before the “big data” phenomena started in 2008, text mining and data mining were used as instruments of fraud detection. However, the limited technological capabilities of the pre-big data technologies made it very difficult for researchers to run fraud detection algorithms on large amounts of data.

Frauds in Health Insurance system can be specific to each country, usually based on gaps or weaknesses of legislation. Models are constantly changing fraud, malicious individuals seeking ever new ways to circumvent the law. Consequently, methods for identifying and preventing fraud must always be adjusted and ready to rediscover the fraudulent actions.

In general we can identify two types of fraud [13]:

1. ***Opportunistic fraud,*** when a person takes advantage of the deliberate padding or inflating of a legitimate insurance claim. This type of fraud is very common, but the incident is related to a reduced amount.
2. ***Professional fraud***, usually done by organized groups of people who may have multiple, false identities. They know very well how to organize the

system and often work together with people within the system. The incidence of these events is lower, but the amount related to an incident is much higher.

The data used for analysis are taken from the database of the National House for Health Insurance and contains all necessary information on the partners involved in events claim payments for medical services.

Specific attributes are used to detect frauds that are usually the same. Thus, in the field of health insurance it can be taken into account: patient demographics (age, gender), details of the medical services provided to the patient, and details of the claim. [11]

The complex nature of the data used in fraud detection has been well described by [1]:

- Volume of both fraud and legal classes will fluctuate independently of each other; therefore class distributions (proportion of illegitimate examples to legitimate examples) will change over time.
- Multiple styles of fraud can happen at around the same time. Each style can have a regular, occasional, seasonal, or once-off temporal characteristic;
- Legal characteristics/behavior can change over time.
- Within the near future after uncovering the current modus operandi of professional fraudsters, these same fraudsters will continually supply new or modified styles of fraud until the detection systems start generating false negatives again.

Techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence.[7] Examples of ***statistical data analysis techniques*** are:

a. Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
b. Calculation of various statistical parameters such as averages, quantiles, performance metrics, probability distributions, and so on. For example, the averages may include average length of call, average number of calls per month and average delays in bill payment.
c. Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
d. Computing user profiles.
e. Time-series analysis of time-dependent data.
f. Clustering and classification to find patterns and associations among groups of data.
g. Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.

Fraud management is a knowledge-intensive activity. The main ***AI techniques*** used for fraud management include [AI]:

a. Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
b. Expert systems to encode expertise for detecting fraud in the form of rules.
c. Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either automatically (unsupervised) or to match given inputs.
d. Machine learning techniques to automatically identify characteristics of fraud.
e. Neural networks that can learn suspicious patterns from samples and used later to detect them.

Figure 3 shows the analysis methods depending on the types of fraud and the types of frauders [SAS1].

Data mining techniques can be used for fraud detection for large sets of data from health insurance system. These techniques detect behavior patterns in large data sets, so based on several cases considered fraudulent can calculate the probability that each record be fraudulent.

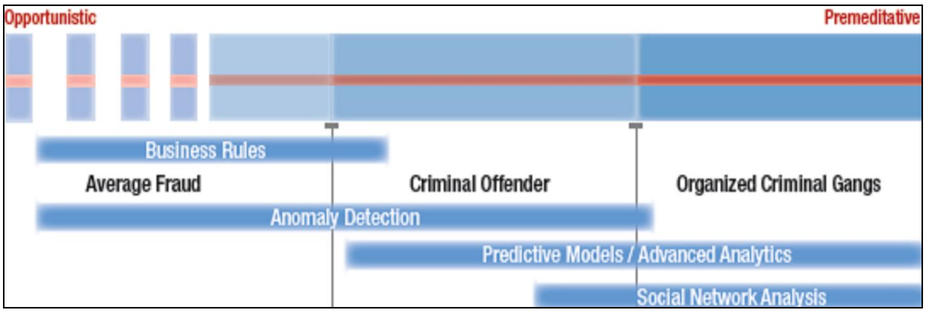


Fig. 3. Techniques for fraud detection ([SAS1])

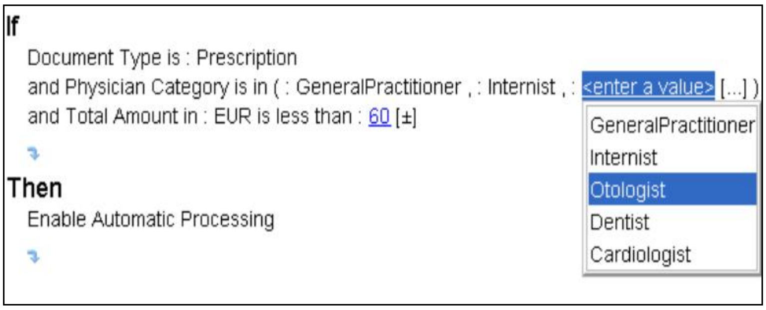
This data must be available, relevant, adequate and clean. There are two main criticisms of data-mining fraud detection tools: the dearth of publicly available data for analysis and the lack of published well-known methods and techniques that are specifically efficient for this field [8].

One of the most commonly used techniques for detecting fraud is **anomaly detection.**

Anomaly detection algorithms are very simple to set and functions automatically. Some key performance indicators are for an event chosen and then thresholds are set. If a threshold is exceeded, then the event is signaled for further investigation. The effectiveness of this method is influenced by the choice of indicators to be monitored, of the analysis period, and of the threshold value settings.

Business Rules

If fraud patterns are known, one can resort to checking every transaction by applying business rules. Based on an aggregate score or exceeding a set threshold, a transaction can and marked as suspicious, and then carefully investigated. Figure 4 presents an example of a business rule used for validation in the claim processing application:



If

Document Type is : Prescription

and Physician Category is in (: GeneralPractitioner , : Internist , : <enter a value> [...])

and Total Amount in : EUR is less than : 60 [±]

Then

Enable Automatic Processing

Fig. 4. Example of business rule [4]

This technique is very simple to apply, once the system was originally set. Its weaknesses are two: setting initial parameters can lead to many false alarms that require further investigation, and the system is flexible to adapt to new methods to defraud the system, new business rules. You can add new business rules only if they meet the new method of fraud.

Database searching

For records detected as suspicious further investigation. One approach is the use of the database searching services, which can give investigators a large amount of information from multiple sources. Was the suspicious person involved in illegal activities? Had he attempted fraud in other areas in the past? Information can be obtained by searching the data to other companies that can help solve the case.

Predictive modeling

Predictive modeling is very successful in detecting fraud. By applying data mining tools, fraud propensity scores can be calculated. Then, using predictive models, they can automatically tell the probability that data is fraudulent and it must be subjected to detailed analysis.

To preserve accuracy, models must be constantly updated to include new types of illegal events.

Text mining

Text mining is a very useful technique as almost 80% of data generated by insurance claiming process have an unstructured form. This technique is very efficient on big data volumes. Meaningful data are extracted and then analyzed by text mining algorithms to reveal abnormal or suspicious behavior of the insured.

Social network analysis

Social network analysis is a method recently used in detecting fraud. This method involves several steps:

1. It starts from modeling the relationships between major system information components (entities) as a network;
2. Suspicious components are detected on the basis of shared characteristics and there is defined a set of indicators for tracking them;
3. Suspicious entities are detected by performing simulations;
4. The resulting reports are visualized in order to be interpreted (see Fig. 5).

SAS Company, world leader in business analytics software, included SAS Social Network Analysis palette of tools for fraud detections [14].

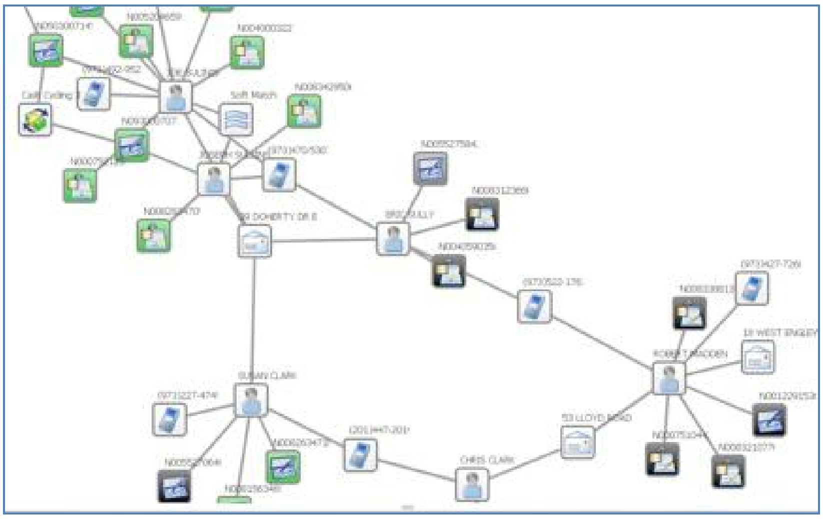


Fig. 5. Social network visualization

Social networks are linked with fraud detection because frauds are performed almost always of networks of persons rather than single individuals.

Once an individual has been identified as suspicious, the entire social network linked to him could be analyzed for searching fraud schemes. Most fraud schemes are hidden beyond the huge collections of data. If a controller would know where to look the fraud schemes would be relatively easy revealed as they are pretty simple to identify.

Conclusions

Big Data technology and distributed processing power of big data cloud bring fraud detection in insurance to another level. Not long ago, insurance fraud detection was not considered cost-effective because the cost and duration of the investigations were too high, so many companies prefer to pay claims without investigation.

Applying Big Data analysis methods can lead to rapid detection of abnormal claims, and then creates a new set of tests to automatically narrow the segment potentially fraudulent applications or to detect new patterns of fraud, previously unknown.

The article briefly presented the National Health Insurance System and the main types of fraud that are encountered.

An analysis of Big Data technology demonstrates its huge potential, but it shows that native tools for data analysis are still immature. The analysis methods applied in the field of health insurance were briefly described, each of them being effective for a particular type of fraud or a particular stage of the fraud detection process. All this leads to the conclusion that the best solution for detecting fraud in the health insurance system is, at present, a hybrid solution, both in terms of technologies and in terms of models of analysis.

References

[1] Fawcett, T., "AI Approaches to Fraud Detection and Risk Management", Papers from the 1997 AAAI Workshop, Technical Report WS-97-07. AAAI Press;

[2] Gill, K. M., Woolley, K. A., & Gill, M., "Insurance fraud: The business as a victim", in M. Gill (Ed.), Crime at work, Vol 1. (pp. 73-82), Leicester: Perpetuity Press, 1994;

[3] Halevi, G., & Moed, H., "The evolution of big data as a research and scientific topic: overview of the literature. Research Trends", Special Issue on Big Data, 30, 3-6, 2012.

[4] Hüsemann, S., Schäfer, M., "Building Flexible eHealth Processes using Business Rules", ECEH, volume 91 of LNI, page 25-36. GI, 2006;

[5] Loshin, D., "Business Data Suited to Big Data Analytics", October 18, 2012, http://data-informed.com/business-problems-suited-to-big-data-analytics/;

[6] Melih, K., Cuneyt, A., "A Fraud Detection Approach with Data Mining in Health Insurance", Procedia - Social and Behavioral Sciences, Volume 62, 24 October 2012, Pages 989-994, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2012.09.168

[7] Palshikar, G.K., "The Hidden Truth – Frauds and Their Control: A Critical Application for Business Intelligence", Intelligent Enterprise, vol. 5, no. 9, 28 May 2002, pp. 46–51.

[8] Phua, C., Lee, V., Smith, K., & Gayler, R., "A comprehensive survey of data mining-based fraud detection research", arXiv preprint arXiv:1009.6119, 2010;

[9] Rutrell, Y., "Analytics platform helps agencies fight cyber crime, government computer news", Jul 12, 2012, http://gcn.com/articles/2012/07/12/sas-security-intelligence-platfrom-analytics.aspx;

[10] Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M., Perspectives on Big Data and Big Data Analytics. Database Systems Journal, 3(4), 3-14, 2012;

[11] Williams, G.J., "Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries", Proceedings of PAKDD99, 1999;

[12] Big Data, http://en.wikipedia.org/wiki/Big_data ;

[13] WHITE PAPER: Combating Insurance Claims Fraud- How to Recognize and Reduce Opportunistic and Organized Claims Fraud, http://support.sas.com/resources/papers/proceedings12/105573_0212.pdf ;

[14] SAS® Social Network Analysis, http://www.sas.com/offices/europe/uk/industries/banking/fraud-detection.html ;

[15] Interviu - Ministrul alternativa al

Sanatatii: 300 de mil. de euro fraudati anual - bani pentru salariile medicilor, http://www.ziare.com/politica/opozitie/ministrul-alternativa-al-sanatatii-300-de-mil-de-euro-fraudati-anual-bani-pentru-salariile-medicilor-interviu-1260060 ;

[16] 16 Top Big Data Analytics Platforms – InformationWeek http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609 ;

[17] The IBM big data platform, http://public.dhe.ibm.com/common/ssi/ecm/en/imb14135usen/IMB14135USEN.PDF ;

[18] Big Data Debate: Will Hadoop Become Dominant Platform? http://www.informationweek.com/big-data/big-data-analytics/big-data-debate-will-hadoop-become-dominant-platform/d/d-id/1109226?

[19] Casa Nationala de Asigurari, http://www.cnas.ro/despre-noi/prezentare-generala ;

[20] Apache Hadoop http://wiki.apache.org/hadoop/PoweredBy;

[21] Ventana Research: The Challenge of Big Data Benchmark Research, 2012, http://www.ventanaresearch.com/BGD

Ana-Ramona BOLOGA (born in 1976) is associate professor at the Academy of Economic Studies from Bucharest, Economic Informatics Department. Her PhD paper was entitled “Software Agents Technology in Business Environment”. Her fields of interest are: integrated information systems, information system analysis and design methodologies, and software agents.

Razvan BOLOGA (born 1976) is associate professor at the Academy of Economic Studies in Bucharest Romania. He is part of the Computer Science department and his fields of interest include information systems, knowledge management and software ecosystems. Mr. Bologa has attended over 15 conferences presenting the results of his research.

Alexandra Maria Ioana FLOREA (born 1984) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2007 and also from the Faculty of Marketing in 2008. Since then she is a PhD candidate, studying to obtain her PhD in the field of economic informatics. At present she is assistant lecturer at the Academy of Economic Science from Bucharest, Economic Informatics Department and her fields of interest include integrated information systems, information system analysis and design methodologies and database management systems.