International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 191-202
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

SOCIAL NETWORK ANALYTICS (SNA) FRAUD

Mrs.N.Mathimagal M.C.A.,(M.Phil.)[1]

[1] N.Mathimagal, [2] Dr.S.Thiruniraisenthil [3] Dr.A. Muthukumaravel
[1]M.Phil-CS Research Scholar, Department of MCA, BIHER, Chennai, Tamil Nadu, India
[2] Associate professor, Department of MCA, BIHER, Chennai, Tamil Nadu, India
[3] Dean-Faculty of Arts & Science, & HOD-Department of MCA, BIHER, Chennai, Tamil Nadu, India

In this first portion, we set the site for what's ahead by launching fraud analysis using explanatory, predicting and social chain method. We start off by explain and identify fraud and discuss various types of fraud. Next, fraud detection (FD) and prevention is explained as a means to address and limit the amount and over-all impact of fraud. Big data and analysis provide mighty tools that may increase an organization's FD system (FDS). We explain in detail how and why these tools complement traditional expert based FD approaches. Future, the fraud analytics process model is established; prove the maximum level overviews of the step that are follow the growing and executing a data driven FDS. The portions conclude by the analyzing the characteristics and skills of a better fraud data researcher, follow by the scientific predicting on the topic.The combination of all features (i.e., intrinsic and network features) is fed to the machine learning algorithms. This is the Gotcha! Model. As the creation of network features drastically increases the Number of features to learn from, ensemble methods like Random Forest are used to train the models.Gotcha! Can easily be mapped to other fraud detection applications, like credit card fraud detection.

Keywords: *SNA, FD, Gotcha, big data.*

1. INTRODUCTION

The incredible growth of the internet use for all sort of applications such as data production and storage, business transactions, professional, cultural and personal information management, etc. are

pushing back the frontiers of traditional computer and digital data management. This overwhelming activity allows all kinds of players to propose new services and offers.

Unfortunately, some did not hesitate to take advantage of this space to be engaged in fraudulent activities, such as Identity Theft Fraud. The objective of this study is to work on a new way to address large scale social network fraud detection by combining real-time processing and batch processing in data warehouse and Hadoop Distributed File System (HDFS).

Fraud is often characterized by irregular concentration of activities on subsets of nodes in subnetworks of the internet, particularly on online social networks (OSN).

This calls for linking data, which were not likely to be linked, because they do not belong to the same networks. Linking social networks data, spread upon different heterogeneous data repositories, calls for addressing several challenging problems such as algorithms optimization and parallelization, new knowledge representation paradigms for heterogeneous, redundant, noncertified or false information, association mechanisms, graph analysis for clustering and partitioning.

To address this multi-dimensional problem, we will adopt the following approach: 1) identify community subnetworks by using community detection algorithms running in a parallel environment, 2) represent data and knowledge stored in these networks in a common knowledge scheme, 3) apply iterative algorithms for clustering and partitioning.

The paper is organized as follows. We present in the second section, some of the main characteristics of OSN data, specially in the case of fraudulent activity. Then, we describe some recent works in different areas such as community detection in social networks, the analysis of large graphs, the clustering and partitioning of bi-partite graph and fraud detection.

Then we introduce the basis of our approach. In the third section, we present how we intend to develop our study, and how we are going to test the proposed solutions through experiments. In the last part, we will give some preliminary conclusions.

2. SOCIAL MEDIA, SOCIAL NETWORK AND BIG DATA

The volume of data recorded and exchanged on networks requires developing new management approaches

for data storage, update, search, visualization and analysis. In addition, these data are not stored in a unique digital format, but are heterogeneous, structured or not, and multimedia. In that project, we will focus more precisely on these networks formed by potentially linked data, due to the fact that they share the same fraudulent activity.

The objective is to be able to give traits to theses nodes and links, to show how they are grouping, forming interest communities or even emerging structures. The links are built based on certain information exchanges between individuals, organisms or entities.

There are communication links representing the messages exchanged between people, membership links representing structures (companies, social or professional groups, services, product categories, etc.) and association links between entities. A first distinction can be done at this level between static links representative of structures and dynamic links representative of actions.

In the field of social network analysis many approaches are based on networks decomposition into subnetworks, such as in the case of community detection in social networks [1].

An agglomerative technique allows identifying all maximal cliques representing relationships. The kernels of eligible communities are formed by iteratively adding the left vertices to their closest kernels to obtain a fractional community that represent the fractional sub network. Bipartite graph partitioning and data clustering are particularly promising approaches for graph analysis [2].

The problem is formulated as a bi-partite graph to cluster/partition nodes by minimizing an edge density function using Singular Value Decomposition. A framework composed of model and MR functions that include several graph analysis functions can be used for large graph processing [3]. Different types of fraud measurement and detection techniques have already been proposed, some of which are using community construction based on indirect links between individuals [4]–[6].

For working on these massively distributed peta bytes of social network data, we will use the SQL/Map Reduce framework that is a practical approach to self-describing, polymorphic, and parallelizable user defined functions. SQL Map Reduce (SQL/MR) features enhance large data sets through parallelized execution and make it possible to test the

algorithm with massive volumes of data about users, devices, and activities.

Thus, the exploration and investigation of data to identify relationships indicative of likely fraud becomes easier with custom MR functions using programming language such as Java, C or C++. SQL/MR allows the use of standard library data structures and open-source 3rd party libraries.

3. SOCIAL NETWORK ANALYSIS FOR FRAUD DETECTION

In the last decade, the use of social media websites in everybody's daily life is booming. People can continue their conversations on online social network sites like Facebook, Twitter, LinkedIn, Google+, Instagram, and so on and share their experiences with their acquaintances, friends, family, and others. It only takes one click to update your whereabouts to the rest of the world.

Plenty of options exist to broadcast your current activities: by picture, video, geo-location, links, or just plain text. You are on the top of the world—and everybody's watching. And this is where it becomes interesting. Users of online social network sites explicitly reveal their relationships with other people. As a consequence, social network sites are a (almost) perfect mapping of the relationships that exist in the real world. We know who you are, what your hobbies and interests are, to whom you are married, how many children you have, your buddies with whom you run every week, your friends at the wine club, etc.

This whole interconnected network of people knowing each other, somehow, is an extremely interesting source of information and knowledge. Marketing managers no longer have to guess who might influence whom to create the appropriate campaign. It is all there—and that is exactly the problem. Social network sites acknowledge the richness of the data sources they have, and are not willing to share them as such and free of cost. Moreover, those data are often privatized and regulated, and well-hidden from commercial use.

On the other hand, social network sites offer many good built-in facilities to managers and other interested parties to launch and manage their marketing campaigns by exploiting the social network, without publishing the exact network representation. However, companies often forget that they can reconstruct (a part of) the social network using in-house data. Telecommunication providers, for example, have a massive transactional data base where they record call behaviour of their customers.

Under the assumption that good friends call each other more often, we can recreate the network and indicate the tie strength between people based on the frequency and/or duration of calls. Internet infrastructure providers might map the relationships between people using their customers' IP-addresses. IP-addresses that frequently communicate are represented by a stronger relationship. In the end, the IP-network will envisage the relational structure between people from another point of view, but to a certain extent as observed in reality.

Many more examples can be found in the banking, retail, and online gaming industry. Also, the fraud detection domain might benefit from the analysis of social networks. In this chapter, we underline the social character of fraud. This means that we assume that the probability of some one committing fraud depends on the people (s) he is connected to.

These are the so-called guilt-by-associations (Koutra et al. 2011). If we know that five friends of Bob are fraudsters, what would we say about Bob? Is he also likely to be a fraudster? If these friends are Bob's only friends, is it more likely that Bob will be influenced to commit fraud? What if Bob has 200 other friends, will the influence of these five fraudsters be the same?

In this paper, we will briefly introduce the reader to networks and their applications in a fraud detection setting. One of the main questions answered in this chapter is how unstructured network information can be translated into useful and meaningful characteristics of a subject. We will analyze and extract features from the direct neighbourhood (i.e., the direct associates of a certain person or subject) as well as the network as a whole (i.e., collective inference). Those network-based features can serve as an enrichment of traditional data analysis techniques.

4. FRAUD PREVENTION

Since fraud is thus arduous to prove in courts, most organizations and people attempt to forestall fraud from happening by blanket measures. This includes limiting the quantity of harm the fraudster will impact on the organization moreover as early detection of fraud patterns. for instance, mastercard corporations will cut the mastercard limit across the board in anticipation of many negative fraud cases. Advertisers will forestall advertising campaigns with low variety of qualifying events. And anti-terrorism agencies will forestall folks with bottles of pure water from boarding the planes.

These actions area unit typically in distinction with the corporate efforts to draw in additional customers and end in general discontentment. To the rescue area unit new technologies like Hadoop, Influence Diagrams and theorem Networks that area unit computationally pricy (these area unit NP-hard in applied science terminology) however area unit additional correct and prophetic .

5. WHY HADOOP?

Apache Hadoop may be a distributed system for process giant amounts of knowledge. in an exceedingly recent Hadoop Summit 2010 Yahoo, Facebook, and different corporations declared that they presently method many TBs of knowledge per day and also the volumes area unit growing at exponential rates. Hadoop is very important for finding the fraud detection drawback because:

- Sampling doesn't work for rare events since the possibility of missing a fraud in fact case ends up in important deterioration of model quality.
- Hadoop will solve abundant tougher issues by leverage multiple cores across thousands of machines and search through abundant larger drawback domains.
- Hadoop is combined with different tools to manage moderate to low response latency needs.

Let's bear these reasons one by one. Sampling may be a common technique for modeling rare events. one amongst the issues with sampling is that we tend to cannot afford to throw away rare positive cases. Even in an exceedingly stratified or sampling theme one needs to retain all positive cases since the model accuracy heavily depends on them (one will typically discard some negative cases though). Given the on top of, the system still needs to bear the total dataset to sieve through the positive and negative cases.

Hadoop is understood for its gnawing power. Nothing will compare with the output power of thousands of machines every of that has multiple cores. As was according recently at the Hadoop Summit 2010, the most important installations of Hadoop have two,000 to 4,000 computers with eight to twelve cores every, amounting to up to forty eight,000 active threads yearning for a pattern at constant time. this enables either (a) searching through larger periods of your time to include events across a bigger timeframe or (b) taking additional sources of data under consideration. it's quite common among social network

corporations to comb through twitter blogs in search of relevant knowledge.

Finally, one amongst the fraud interference issues is latency. The agencies wish to react to an occurrence as shortly as attainable, typically inside many minutes of the event. Yahoo recently according that it will alter its activity model in an exceedingly response to a user click event inside 5-7 minutes across many hundred of innumerable customers and billions of events per day. Cloudera has developed a tool, Flume, which will load billions of events into HDFS inside many seconds and analyze them victimization MapReduce.

Often fraud detection is like "finding a needle in an exceedingly haystack". One needs to bear mountains of relevant and on the face of it unsuitable info, build dependency models, appraise the impact and thwart the fraudster actions. Hadoop helps with finding patterns by process mountains of data on thousands of cores in an exceedingly comparatively short quantity of your time.

6. GOTCHA

We introduce GOTCHA!, a new, generic, scalable, and integrated approach on however (social) network analytics will improve the performance of ancient fraud detection tools during a social insurance context. We identify 5 challenges that concur with fraud; that is, fraud is associate degree uncommon, well-considered, time-evolving, carefully organized, and unnoticeably hid crime that appears in many alternative varieties and forms. Whereascurrent analysis fails to integrate of these dimensionsinto one encompassing approach, GOTCHA! is that the 1st to address every of those challenges along in onehigh-performance, time-dependent detection technique.

In short, GOTCHA! contributes to the fraud detectiondomain by proposing a completely unique approach on the way to spread fraud through a (i) time-weighted networkand options extracted from a (ii) bipartite graph.We have a tendency to exploit dynamic network-based options area unit hidden (dashed line).derived from the direct neighborhood and develop a new propagation rule that infers associate degree initialexposure score for every node victimisation the complete network.

The exposure score measures the extent to that a nodeis influenced by dishonest nodes. We have a tendency to integrate eachintrinsic and network-based option into one scalable algorithm. We have a tendency to argue that fraud may be a time-dependentphenomenon, and as a consequence, GOTCHA! Isdesigned specified a subject's characteristics and

fraudprobability will modification the over time.

7. RESULT AND ANALYSIS

By using the above logic, we studied the execution of fraud detection by theatres ticket booking. Here depending upon the GOTCHA method the unwanted processing charges can be eliminated. So that the user can be save the amount form online booking charges & other things. The results obtained are below

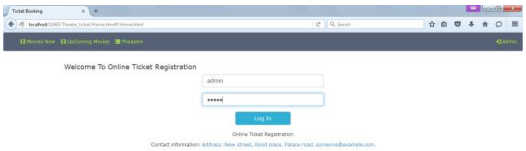

Fig 1: Admin Login

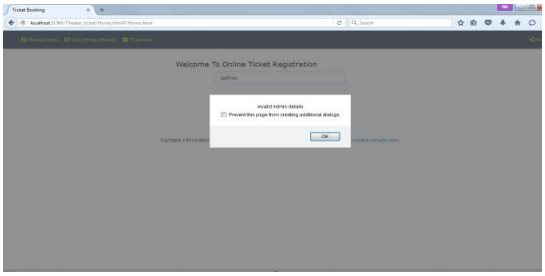

Fig 2: login ok

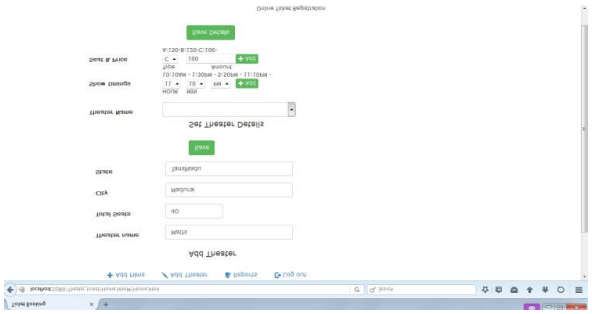

Fig 3: Admin movie entry details

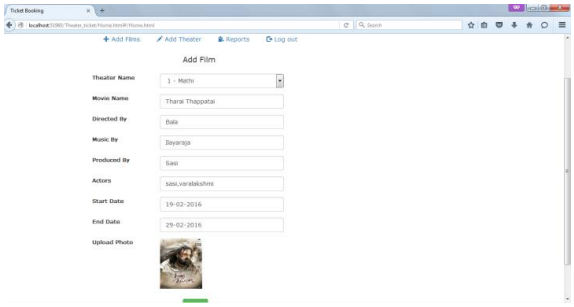

Fig 4: Admin movie entry final page

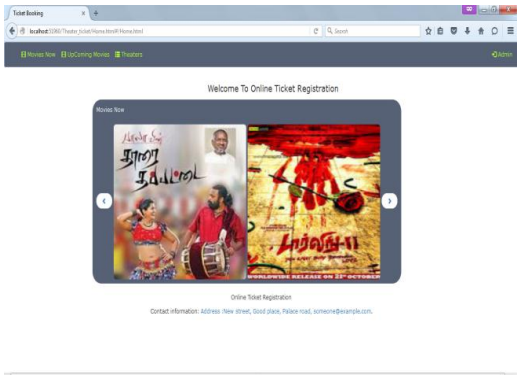

Fig 5: online movie display page

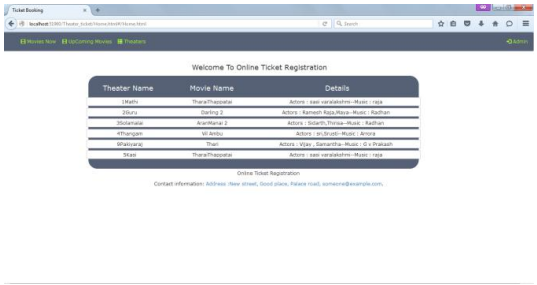

Fig 6: User need to select their option of theatre it will redirect to ticket booking page

Fig 7: Ticket booking page

6. CONCLUSION

In this paper, we tend to improve the performance of ancientclassification techniques for Social Securityfraud detection by together with domain-driven networkinformation mistreatment GOTCHA!, a brand new fraud detectionapproach. We tend to begin by distinguishing the challenges thatconcur with fraud and style GOTCHA! Specifiedit addresses every of those challenges to find future fraud. In this paper we have presented our motivations to study large scale social networks for characterizing communities. Our study will address the problems of linking information spread over several heterogeneous networks, algorithms parallelization and optimization for network analysis, and graph partitioning and clustering for structure extraction. We expect that this work will provide an answer to fraud detection.

REFERENCES

[1] Armstrong, J. S. (2001). Selecting Forecasting Methods. In J.S. Armstrong, ed. Principles of Forecasting: A Handbook for Researchers and Practitioners. New York: Springer Science + Business Media, pp. 365–386.

[2] Baesens, B. (2014). Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications. Hoboken, NJ: John Wiley & Sons.

[3] Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. Statistical Science, 17 (3): 235–249.

[4] Caron, F., VandenBroucke, S., Vanthienen, J., &Baesens, B. (2013). Advanced Rule-Based Process Analytics: Applications for Risk Response Decisions and Management Control Activities. Expert Systems with Applications, Submitted.

[5] Chakraborty, G., Murali, P., & Satish, G. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. Cary, NC: SAS Institute.

[6] Dr Ananthi Sheshasaayee, R. Megala, “A Conceptual Framework For Resource Utilization In Cloud Using Map Reduce Scheduler”, International Journal of Innovations in Scientific and Engineering Research (IJISER), Vol.4, No.6, pp.188-190, 2017.

[7] Cressey, D. R. (1953). Other People's Money; A Study of the Social Psychology of Embezzlement. New York: Free Press.

Duffield, G., &Grabosky, P. (2001). The Psychology of Fraud. In Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology (199).

[8] Elder IV,, J., & Thomas, H. (2012). Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. New York: Academic Press.

Fawcett, T., & Provost, F. (1997). Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1–3 (3): 291–316.

[9] Grabosky, P., & Duffield, G. (2001). Red Flags of Fraud. Trends and Issues in Crime and Criminal Justice, Australian Institute of Criminology (200).

[10] Han, J., &Kamber, M. (2011). Data Mining: Concepts and Techniques, Third Edition: Morgan Kaufmann.

[11] Hand, D. (2007, September). Statistical Techniques for Fraud Detection, Prevention, and Evaluation. Paper presented at the NATO ASI: Mining Massive Data sets for Security, London, England.

[12] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining.

Cambridge, MA: Bradford.

[13] Jamain, A. (2001). Benford's Law. London: Imperial College.

[14] Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive Modeling with Big Data: Is Bigger Really Better? Big Data 1 (4): 215–226.

[15] Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed.). Hoboken, NJ: Wiley.

[16] Maydanchik, A. (2007). Data Quality Assessment. Bradley Beach, NC: Technics Publications.

[17] Navarette, E. (2006). Practical Calculation of Expected and Unexpected Losses in Operational Risk by Simulation Methods (Banca &Finanzas: Documen-tos de Trabajo, 1 (1): pp. 1–12).

[18] Petropoulos, F., Makridakis, S., Assimakopoulos, V., &Nikolopoulos, K. (2014). "Horses for courses" in demand forecasting. European Journal of Operational Research, 237 (1): 152–163.

[19] Schneider, F. (2002). Size and Measurement of the Informal Economy in 110 Countries around the World. In Workshop of Australian National Tax Centre, ANU, Canberra, Australia.

[20] Tan, P.-N. N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining.

[21] Boston: Addison Wesley.

[22] Van Gestel, T., &Baesens, B. (2009). Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and

Regulatory Capital. Oxford: Oxford University Press.

[23] Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., &Baesens, B. (2015). Gotcha! Network-based Fraud Detection for Social Security Fraud. Management Science, Submitted.

[24] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., &Baesens, B. (2012). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. European Journal of Operational Research 218: 211–229.