Fraud Detection on Large Scale Social Networks

Yaya Sylla, Pierre Morizet-Mahoudeaux, Stephen Brobst

Fraud detection on large scale social networks

Yaya Sylla (1), (2), Pierre Morizet-Mahoudeaux (1)
(1) UMR-CNRS 7253 Heudiasyc
University of Technology of Compiègne
Compiègne, France
Email: yaya.sylla@utc.fr, pierre.morizet@utc.fr

Stephen Brobst (2)
(2) Teradata Corporation
Massy, France & Dayton, USA
Email: yaya.sylla@teradata.com
stephen.brobst@teradata.com

***Abstract*—The incredible growth of the internet use for all kinds of businesses has generated at the same time an increase of fraudulent activities, which calls for developing new methods and tools for detecting fraud and other crimes against banks and customers. Fraud detection needs to analyze and link data, which are gathered from heterogeneous data repositories, and to address problem solving algorithms optimization and parallelization, new knowledge representation paradigms, association mechanisms for linking data, and graph analysis for clustering and partitioning. We present in this paper the motivation of our study and the first steps of the work. We will focus on the emergence of new coding models based on MapReduce and SQL extensions, and on graphs paths issues.**

***Keywords*-Large scale graphs analysis; graph partition and clustering; parallel processing; fraud detection**

I. 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, non-certified 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.

II. 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 subnetwork.

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 petabytes of social network data, we will use the SQL/MapReduce framework that is a practical approach to self-describing, polymorphic, and parallelizable user defined functions. SQL MapReduce (SQL/MR) features enhance large data sets through parallelized execution and makes 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 programing language such as Java, C or C++. SQL/MR allows the use of standard library data structures and open-source 3rd party libraries.

III. Research plan

In this work, we will focus on the emergence of new coding modes based on MapReduce and SQL extensions, and on graphs paths issues in the case of large scale networks architectures. We began studying alternative approaches such as the k-NN, and the conditions for finding convergence criteria in the case of recursive algorithms.

A certain number of approaches and algorithms have allowed developing efficient indexation and information search functions. The HDFS file system linked to Hadoop allows to distribute the data storage and to run efficient data analyses thanks to the Map/Reduce model, which allows distributing one operation on several nodes in order to get parallelism for the execution. Model coding based on SQL combined with SQL Map-Reduce allows algorithms to run on parallel platforms linking several massive data sets *i.e.* structured data stored in a relational database combined with unstructured data stored on Hadoop systems. SQLMR also allows to implement iterative functions using SQL language and Java programming.

The main steps of the research will be:

- Select social network dataset or link several social networks data together (Facebook, Twitter, LinkedIn, Google) by defining a large scale social network for analysis
- Use an algorithm for determining community from the social network such as in [1]. We will develop extensions to this work and show the impact on the results based on the specific use case. We will compare these different approaches and re-combine the best into a new algorithm derived on k-partite graph clustering/partitioning algorithms that can be applied to such graphs. We will use these algorithms iteratively on parallel infrastructures like Hadoop according to convergence criteria.
- Once a given community will be identified (with a known structure or not), we will apply different fraud detection algorithms on clusters/partitions to the identified community matrix. We will focus in the first step of this study on Identity Theft Fraud detection.

IV. Conclusion

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] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ser. WebKDD/SNA-KDD '07. New York, NY, USA: ACM, 2007, pp. 16–25. [Online]. Available: http://doi.acm.org/10.1145/1348549.1348552

[2] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Bipartite graph partitioning and data clustering," in *Proceedings of the tenth international conference on Information and knowledge management*, ser. CIKM '01. New York, NY, USA: ACM, 2001, pp. 25–32. [Online]. Available: http://doi.acm.org/10.1145/502585.502591

[3] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for large-scale graph processing," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 135–146. [Online]. Available: http://doi.acm.org/10.1145/1807167.1807184

[4] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," 09 2010. [Online]. Available: http://arxiv.org/abs/1009.6119

[5] W. Eberle and L. Holder, "Anomaly detection in data represented as graphs," *Intell. Data Anal.*, vol. 11, no. 6, pp. 663–689, Dec. 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1368018.1368024

[6] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, "Netprobe: a fast and scalable system for fraud detection in online auction networks," in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 201–210. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242600