SCALABLE FINANCIAL ANOMALY DETECTION IMPLEMENTATION IN A DISTRIBUTED GRAPH DATABASE SERVER

Kasun Prabhath Dharmadasa

199318L

Thesis submitted in fulfilment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2023

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 14/7/2023

S. R. R. B. K. K. P. Dharmadasa

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor: Date:

Signature of the supervisor: Date: 14/7/2023

Dr. Miyuru Dayarathna

Abstract

Financial transactions have become a prominent part of the economy in the world. Over 1 billion transactions are being processed daily around the world and a considerable portion of those transactions accounts for various fraudulent activities that take place around the world. Terrorist Financing, Money Laundering are popular examples that generate fraudulent transactions. Financial institutions are obligated to have the capability to detect such transactions and perform necessary measures to mitigate and report the parties involved with such fraudulent transactions. Several implementations exist that map the financial transactions in the form linked/graph data and detect any anomalies using the structural features of the graph such as PageRank, Degree Distribution, etc. However, these implementations require the transactions to be executed in a single graph database and this limits the capability to scale horizontally when the number of transactions increases. This research proposes an extension to a C/C++ based distributed graph database server called JasmineGraph that is capable of handling large amounts of graph data and performing anomaly detection algorithms in a distributed manner. We generate Degree Distribution and PageRank scores for the graph network in a distributed manner and use these graph structural features to train a machine learning model for anomaly detection. Our distributed anomaly detection approach has been able to predict anomalous transactions with an F1-score up to 0.98 and was able to reduce the execution time by 79.5% in comparison to the non distributed approach when detecting anomalies. For large datasets (PaySim-2M), the non distributed approach failed to process due to lack of memory but was successful after using the distributed approach making it more efficient to use our distributed anomaly detection for large financial transaction networks. As future work, we plan to expand our anomaly detection approach on streaming graphs for real time anomaly detection.

Keywords: Graph Databases, Fraud Detection, Anti Money Laundering, Scalability, System Performance

Acknowledgement

I am deeply grateful to my advisors, Prof. Sanath Jayasena and Dr Miyuru Dayarathna, for their immense support on my master's program. Their knowledge and guidance has been a great help to me and have paved the way for the success of this thesis.

I am deeply thankful to my family and my friends for their love and support during this process. Without their encouragement and motivation, I would not have been able to complete this journey.

Finally, I would like to extend my sincere gratitude to all of the participants in my study. Their willingness to share their experiences and insights has been invaluable to my research and has helped to make this thesis a success. Thank you for your time and contribution.

Table of Contents

Declaration i
Abstract ii
Acknowledgement iii
List of Figures vi
List of Tables vii
List of Abbreviations viii
List of Appendices ix

1. INTRODUCTION 1
    1.1. Research Problem 3
    1.2. Motivation 4
    1.3. Aim and Objectives 5
    1.4. Contributions 5
2. RELATED WORK 7
    2.1. Financial Anomaly Detection 7
    2.2. Distributed Graph Database Servers 9
    2.3. Graph Structural Features 10
    2.4. Graph based Machine Learning 11
    2.5. Summary 12
3. USE OF GRAPH STRUCTURAL FEATURES FOR ANOMALY DETECTION IN DISTRIBUTED GRAPHS 13
    3.1. JasmineGraph Database Server 13
    3.2. Degree Distribution 14
    3.3. PageRank 15
4. IMPLEMENTATION 17
    4.1. Distributed Degree Distribution 17
    4.2. Distributed PageRank Approximation 20
    4.3. Binary classification ML Model Architecture 21
5. EVALUATION 23
    5.1. Experimental Setup 23
        5.1.1. Computing Infrastructure Setup 23
        5.1.2. Experiment Datasets 24
    5.2. Degree Distribution and PageRank 25
    5.4. Scalability Evaluation 30
    5.5. Results Discussion 31
6. CONCLUSIONS 33

6.1. Summary 33
6.2. Limitations 34
6.3. Future Work 34
REFERENCES 35
Appendix I - Out Degree Distribution Algorithm 39
Appendix II - In Degree Distribution Algorithm 40
Appendix III - PageRank Algorithm 42

List of Figures

Figure 3.1: Overview of JasmineGraph 14

Figure 4.1: Central Store implementation in JasmineGraph 17

Figure 4.2: Propagation of the PageRank score of node A within the graph 20

Figure 5.1: Experiment setup for 4 worker nodes in JasmineGraph 24

List of Tables

Page

Table 1.1: Comparison between Rule based anomaly detection and ML based anomaly detection 2

Table 5.1: Dataset properties 25

Table 5.2: Experiment results for AMLSim-100K dataset 26

Table 5.3: Experiment results for PaySim-500K dataset 26

Table 5.4: Experiment results for AMLSim-1M dataset 26

Table 5.5: Experiment results for PaySim-2M dataset 27

Table 5.6: Non-distributed vs distributed approach using JasmineGraph server for AMLSim-100K 28

Table 5.7: Non-distributed vs distributed approach using JasmineGraph server for PaySim-500K 28

Table 5.8: Non-distributed vs distributed approach using JasmineGraph server for AMLSim-1M 29

Table 5.9: Non-distributed vs distributed approach using JasmineGraph server for PaySim-2M 29

List of Abbreviations

AML	Anti Money Laundering
CPU	Central Processing Unit
GB	Giga Byte
GHz	Giga Hertz
MB	Mega Byte
ML	Machine Learning
MRI	Magnetic Resonance Imaging
RAM	Random Access Memory
VM	Virtual Machine

List of Appendices

Appendix	Description	Page
Appendix I	Out Degree Distribution Algorithm	39
Appendix II	In Degree Distribution Algorithm	40
Appendix III	PageRank Algorithm	42