UNIVERSITI PUTRA MALAYSIA

EGO-CENTRIC APPROACH FOR PREDICTING FRAUDULENT COLLABORATION IN TELECOMMUNICATION

ROSMAWATI BINTI AB RAUB
FSKTM 2010 1

EGO-CENTRIC APPROACH FOR PREDICTING FRAUDULENT COLLABORATION IN TELECOMMUNICATION

ROSMAWATI BINTI AB RAUB

MASTER OF SCIENCE

UNIVERSITI PUTRA MALAYSIA

2010

EGO-CENTRIC APPROACH FOR PREDICTING FRAUDULENT COLLABORATION IN TELECOMMUNICATION

By

ROSMAWATI BINTI AB RAUB

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

January 2010

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

EGO-CENTRIC APPROACH FOR PREDICTING FRAUDULENT COLLABORATION IN TELECOMMUNICATION

By

ROSMAWATI BINTI AB RAUB

January 2010

Chair : Associate Professor Hj. Ramlan bin Mahmod

Faculty : Computer Science and Information Technology

Recently, there has been a surge of interest in social networks ever since the tragic event of September 11, 2001 attacks on The World Trade Center in the United States. E-mail traffic, disease transmission, criminal activity and communication network can all be modeled as social networks. Ego-centric is an approach used in social network analysis. In the social network parlance, the focused person is referred to as “ego” and his or her affiliate, friend or relative is known as “alters”. An ego-centered network positions an individual at the center of a social network team for the person to traverse his or her relationships with other team members. Through social network analysis, enforcement officers can recognize how information flows through social ties, how people acquire information and resources and how cleavages and coalitions operate. In this thesis, based on social network theories and link analysis; a data mining technology, a social network analysis model is developed to facilitate in detecting fraudulent collaboration, after which an evaluation is then made on the performance of the developed model. This study aims to explore the

usage of embedding social network analysis functions into fraudulent collaboration investigation in call details records. Two types of social network data collection approaches are discussed; (i) social network with centrality measures values and (ii) social network without centrality measures values, where the first approach is based on the previous research while the second is based on the current research experimented. Performance of the models produced by both approaches are measured based on a standard measurement. Performance is tested using statistical models which include Bayesian Network, Naïve Bayesian and Binary Logistic Regression Model is performed. These statistical models are used in order to prove and determine which model is the 'best' that can produce a better prediction of fraudulent collaboration. The outcome of this research is thought to be of help to any enforcement agency or relevant authority in its future operations or measures to detect fraudulent activity in social networks.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENDEKATAN "EGO-CENTRIC" UNTUK MERAMALKAN PENJENAYAH YANG BERSUBAHAT DALAM PENIPUAN TELEKOMUNIKASI

Oleh

ROSMAWATI BINTI AB RAUB

Januari 2010

Pengerusi : Professor Madya Hj. Ramlan bin Mahmod, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Sejak berlakunya tragedi serangan ke atas World Trade Center di Amerika Syarikat pada 11 September 2001 yang lalu, minat terhadap rangkaian sosial telah bermula dan semakin diminati sehingga kini. Kesesakan yang berlaku pada "e-mail", penyakit berjangkit yang menular, aktiviti penipuan yang berleluasa dan rangkaian komunikasi, kesemuanya boleh dimodelkan menjadi suatu rangkaian sosial. Pendekatan "Ego-centric" adalah satu pendekatan yang digunakan dalam analisis rangkaian sosial**.** Di dalam penggunaan perkataan pada analisa rangkaian sosial, orang yang diberi tumpuan dinamakan "ego" manakala jiran-jirannya dikenali sebagai "alters". Rangkaian "ego-centered" memposisikan seseorang itu ditengah-tengah rangkaian sosial supaya hubungannya dgn jiran-jirannya boleh diketahui. Melalui analisa rangkaian sosial, pegawai penyiasat boleh mengetahui aliran maklumat yang keluar dan masuk melalui hubungan sosial sesama manusia, bentuk hubungan atau kerjasama yang terjalin antara manusia, bagaimana seseorang itu mendapat maklumat dan dari mana sumber maklumat tersebut diperolehi. Kertas penyelidikan ini berdasarkan teori analisa rangkaian sosial dan analisa hubungan

iaitu suatu teknologi perlombongan data, satu model analisa rangkaian sosial dibangunkan bertujuan untuk membantu dalam mencari mereka atau kumpulan yang terlibat dalam aktiviti jenayah telekomunikasi selain dari menjalani ujian menguji kecekapan model. Kajian ini juga bertujuan untuk mendapatkan keberkesanan kepenggunaan analisa rangkaian sosial dalam proses pegawai penyiasat mencari mereka yang bersekongkol atau berkerjasama dalam jenayah telekomunikasi melalui maklumat yang diperolehi dari data telefon. Dua jenis data rangkaian sosial digunakan iaitu (i) rangkaian sosial dengan nilai "centrality" dan (ii) rangkaian sosial tanpa nilai "centrality" di mana pendekatan yang pertama berasaskan penyelidikan lalu manakala pendekatan kedua adalah pendekatan yang hendak diperkenalkan dalam kajian ini. Ujian menguji kecekapan model dilakukan untuk membuktikan model yang hendak diperkenalkan ini dapat memberi ketepatan yang lebih baik dalam mencari penjenayah-penjenayah yang terlibat dalam penipuan telekomunikasi. Ujian tersebut dilakukan menggunakan model statistik iaitu "Bayesian Network", "Naïve Bayesian" dan "Binary Logistic Regression Model". Model statistik ini digunakan untuk mengetahui model manakah yang terbaik, iaitu yang dapat memberi ketepatan yang paling tinggi dalam mencari penjenayah-penjenayah yang terlibat dalam penipuan telekomunikasi. Hasil dari penyelidikan ini diharap dapat membantu agensi penyelidikan atau pihak berkuasa dalam menjejaki aktiviti penipuan telekomunikasi yang wuiud di dalam rangkaian sosial.

ACKNOWLEDGEMENTS

"In the name of Allah the Most Gracious, the Most Merciful." Praise be to Allah, the Almighty, for giving me the strength and patience to accomplish this research.

First and foremost, I wish to express my gratitude to my supervisors, Associate Professor Dr Haji Ramlan Mahmod and Puan Hajjah Zaiton Muda, for their comments, advice and guidance in the preparation of this research. Without their help, I would not have been able to complete this research.

I am also deeply grateful to all my colleagues at TM Research & Development (TM R&D) for their invaluable assistance and support. I would also like to acknowledge all the people I have worked with including my colleagues from the Fraud Management division of Telekom Malaysia who have provided me with invaluable data and materials needed for this research.

I am also greatly indebted to my company, TM R&D, which funded the research as well as my masters study program under project number R06-0652- 0.

Finally, I would like to express my appreciation to all individuals who have helped me directly or indirectly in the completion of my thesis.

Thank you.

This thesis submitted to Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Haji Ramlan Mahmod, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Hajjah Zaiton Muda
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 8 April 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ASPECT	Advanced Security for Personal Communications Technologies
BLRM	Binary Logistic Regression Model
BPH	Behavior Profile History
CBR	Case-Based Reasoning
CDR	Call Details Record
CIA	Central Intelligence Agency
CPB	Current Profile Behavior
CPT	Conditional Probability Table
CSV	Comma Separated Value
FBI	Federal Bureau of Investigation
FMD	Fraud Management Division
FMS	Fraud Management System
GRAPA	Global Revenue Assurance Professionals Association, Inc.
IDD	International Direct Dialing
ILP	Inductive Logic Programming
PIN	Personal Identification Number
PSTN	Public Switch Telephone Network
SMS	Short Messaging Service
SNA	Social Network Analysis
SPSS	Statistical Package for the Social Science
TM	Telekom Malaysia

CHAPTER 1

INTRODUCTION

1.1 Introduction

Since the tragic events of September 11, 2001 attacks on The World Trade Center in the United States, social network analysis (SNA) has increasingly been used to study terrorist network. Intelligence agencies such as the CIA and FBI are aggressively gathering information and analyzing it to investigate the so-called terrorists' activities (Hsinchun C. et. al., 2003) in finding structural properties of criminal network which helps them target critical network members for removal and surveillance. An appropriate network analysis technique is used to mine criminal networks and gain insight into these structural properties of the criminal network.

According to the report of Congress's Joint Inquiry leading from the aftermath, the attack could have been avoided if intelligence analysts emphasis more on analyzing every available pertinent information gathered by their investigators (Robert P. et. al., 2004).

Academicians have been invited for possible contributions in covering terrorist networks to enhance security and safety of the nation. A headline appeared in the Federal Computer Week: "Investigative Data Mining Part of Broad Initiative to Fight Terrorism" (June 3, 2002) which stated that FBI has selected data mining and

analytical software as a key technology to comb vast amounts of digital information to discover patterns and relationships that indicate criminal activity (Jesus Mena, 2003).

Previous works have been done in the field of finding human collaboration through their social network. For instance, shortest path algorithms; Priority First Search algorithm was proposed to identify strongest association path between entities in a criminal network (Jennifer J. X. and Hsinchun C., 2002).

One of the social network analysis technique called socio-centric approach is used in finding patterns of interaction between subgroups in a criminal network (Jennifer J. X. et. al., 2003). This approach determines the presence or absence of an association between a pair of subgroups based on a link density measure in order to predict strongest association between subgroups.

A prototype called iMiner which includes SNA specifically ego-centric technique is developed to destabilize terrorist network (Memon N. and Larsen H.L., 2006).

A research was done in finding ways to improve intelligence analysis process (Robert P. et. al., 2004). It was shown that pattern analysis is one of the key information technologies for counterterrorism that can empower analysts for better analysis.

1.2 Problem Statement

Fraudulent activity has occurred in many areas of our daily life including fraud in telecommunication industry. It is the most significant threat to telecommunication business including mobile communications and has a major impact on service levels and customer confidence. Fraud has increased drastically with the expansion of modern technology and global communication, resulting in substantial losses to the businesses. Consequentially, fraud detection has become an important issue to be explored. There are a number of passive techniques used to detect fraud in telecommunication including rule-based, profiling and neural network which detect fraud once it has already occurred as discussed in sub-Chapter 2.4.

Since the aftermath of September 11, 2001, social network analysis have become among the most important research topics for uncovering criminal networks to prevent future crimes. There are a number of works focused on collaboration prediction presented by Memon N. and Larsen H. L. (2006); Smith M. and Welser H.T. (2006) and Fisher D. (2005) but none on fraudulent collaboration prediction. A review of the literature shows that social network analysis provides an alternative approach of reducing fraud. This is performed by recognizing who the fraudster is and with whom he or she collaborates. Once the fraudster is identified, further action are undertaken by the law enforcement agencies to either *eliminate* the fraudster or place surveillance on the fraudster and his or her group members.

Through social network analysis, one can learn and examine the fraudster's communication habits, his/her connections with other members of an organized crime group and how they operate socially.

Based on the above scenario, this research will investigate:

1. The fraudulent activity that needs the involvement of at least a minimum of two persons, one of whom is the collaborator. Collaboration is defined as a form of social relationship involving cooperative activity or coordination of work between two persons/suspects without giving any thought of who initiates it first.

2. The problems in identifying the collaborator or fraudster which are:
    i) Difficulty in identifying the criteria of collaborative association between fraudsters.
    ii) Difficulty in creating a framework or process related to the above fraudulent collaboration prediction.

Hence, this thesis focuses on identifying a suitable analysis approach for the purpose of better accuracy prediction of fraudulent collaboration in telecommunication. Thus our study would be directed at ways of improving approach used in previous research which is discussed in sub-chapter 5.2. An egocentric network analysis approach is used because it examines not only a suspected person's (ego) but also his/her immediate neighbors and their associated connections. The depth of link used is limited to a maximum of two as this practice is also used by fraud analysts. Depth of link two means the nodes together with its neighbors associated with source nodes will be counted or included.

This approach focuses on the relationship between communication behaviors of a caller to a callee, in order to understand their connections to each other and to the entire network.

1.3 Research Objective

The objective of this research is to propose an ego-centric link analysis approach for predicting fraudulent collaboration in telecommunication.

1.4 Research Scope and Limitation

This thesis is based on ego-centric network approach of social network data collection and analysis of one-month duration of call detail records (CDR) which were captured on a database system located at Telekom Malaysia's exchange located in Kelana Jaya. The one-month duration or similar to 30 days duration is used as it is based on one billing cycle. Billing cycle is a periodic interval between periodic billings for services rendered normally for one month. The raw information available for every call is known as call detail record. For each phone call, data format standard is created by telecommunication industry and stored in huge network databases. These records contain detailed information about the call such as its time and duration, the caller and called number, amount bill for each call, type of calls such as international direct dialing (IDD) and its destination country.

The CDR is limited to Public Switch Telephone Network (PSTN). In short, only fixed line telephone number CDR, either incoming or outgoing calls would be captured. Any calls made from mobile number to another mobile number is not captured as these are mobile numbers comes from other service providers. It is assumed that each call that a user made is relevant to him or her.

1.5 Organization of the Thesis

The thesis is organized in accordance with the standard structure of the thesis and dissertations at Universiti Putra Malaysia. The thesis has six chapters, including the introductory chapter that covers the background information that leads to an idea of furthering in detail the concepts of social network analysis.

Chapter 2 concerns with research backgrounds, reviewing similar works that have been carried out by other researchers in social network analysis background. We will also cover a few method of telecommunication fraud detection and its transition to present fraud detection technique. This chapter is equipped with an introduction on statistical approach used which includes Bayesian Belief Network, Naïve Bayes Classification and Binary Logistic Regression Model for creating predictive fraud collaboration framework.

Chapter 3 describes the research methodology used on proposed framework. It explains all the stages in constructing the proposed framework which include modules of data preparation and measurement of performance.

Chapter 4 describes the systems design on proposed framework. It explains all the stages in constructing the proposed framework or model which include modules of network creation, behavioral structural creation, social network, network integration and data analysis.

Chapter 5 discusses on the results produced and the interpretation. Accuracy on each statistical approach is shown and comparison is made among them.

Chapter 6 provides the conclusion of the results produced with several recommendations and suggestions for future work or further research.