Faculty of Engineering and Information Technology
University of Technology Sydney

Data Mining of Classification for Sybil User Detection

Thesis submitted in partial fulfilment of
the requirements for the degree of
Master of Analytics by Research

By

Anand Arun Chinchore

June 2016

Dedicated

To my late father....

Mr Arun Yashwant Chinchore

CERTIFICATE OF AUTHORSHIP / ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as a part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Acknowledgements

I would like to express my special appreciation and thanks to my supervisor Dr Guandong Xu. You have shown tremendous patience with me as one of my mentors. I would like to thank you for encouraging my research and for allowing me to grow as a researcher in analytics. Your advice on both my research and my career have been priceless.

I would also like to extend my appreciation to my co-supervisor Dr Frank Jiang for providing me with continuous support throughout my Master of Analytics by research and study. Without your professional guidance, persistent help, patience, motivation and immense knowledge, this thesis would not have been possible. Words are not enough.

I thank the management team members of the Advanced Analytics Institute, Mr Colin Wise and Mr Austin Andrade, for their encouragement and guidance towards the completion of my research journey. All of you have been there to support me, when required, over the last two and half years.

I would also like to thank the many other faculties, staff, lecturers, professors and doctors of the University of Technology Sydney, who guided and helped me along the path of my research within UTS.

I thank my workplace manager, and my colleague, Mrs Roula Christodoulides who helped and supported me to balance my time in work and study.

A special thanks to my mother Smt Vasudha Arun Chinchore and my family. Words cannot express how grateful I am to you for all of the sacrifices that you've made on my behalf. Your prayers for me have sustained me thus far. I would also like to thank all of my friends who have supported me in writing and given me incentive to strive towards my goal.

Anand Arun Chinchore

June 2016 @ UTS

Table of contents

List of figures

List of Tables

List of publications

Papers published

Chinchore, A, Jiang, F, Xu, G 2015, 'Intelligent Sybil attack detection on abnormal connectivity behaviour in mobile social networks', in the proceedings of the Springer International Publishing 10th International Conference, KMO (2015), Maribor, Slovenia, 24-28 August, vol. 224, pp. 602-617.

Abstract

Data analytics and Big Data application research, along with new structures in complex data, are reveling the secrets of their own complexity and patterns with valuable and critical, but challenging, issues through newly designed tools, techniques and models in data science technology. A common example concerns the interconnectivity of social network users on mobiles, involving content and information sharing through mobile social networks.

There have been a large number of studies on mobile networks. Many focus on a variety of secured applications that attempt to exploit social connections, impersonate users or attack social groups. Such applications are often created with the intention of collecting confidential information, laundering money, blackmail or to perform other criminal activities.

Existing methods for identifying such activity, such as distributed systems, social graph-based sybil detection, behaviour classification, and local ranking systems that estimate the trust level between users, rely on the dependencies between random nodes of connection on mobile social networks. These models aim to detect suspicious connections and have the advantage of learning the relationships between nodes and data. However, their detection patterns tend to impose the behavioural patterns typically associated with community-based and external networks.

In data mining, the graph-based and classification models used for pattern collection can accurately predict patterns in data in targeted categories. Decision trees, commonly used for classification, are trees in which each branch represents a choice between a number of alternatives, and each leaf represents a classification, or decision. For example, a decision tree may help an institution decide whether a node in a dataset is suspicious, or considered to be sybil, if a

decision tree can be induced from a set of data about its instances and the - classifications of those instances. It could also provide the flexibility to demonstrate data distribution. Thus, researchers have tried to combine different techniques and methods into network-based models to detect various patterns generated by sybil nodes within a network.

The purpose of this thesis is to abridge existing classification and regression techniques to identify sybil nodes, and the correlation of those nodes with time, to address these research limitations.

Classification and regression techniques predict behaviour based on continuous or categorical responses. For example if the predicted response is continuous, then it is called a regression tree. If the response is categorical, it is called a classification tree. At each node of the tree, the value of one the connected input nodes is checked and a binary answer − yes or no − determines whether one continues to the left or right sub-branch. When a leaf is reached, a prediction follows from a series of entropy calculations and graphing techniques.

This thesis introduces a novel classification model for sybil detection in mobile social behaviour that identifies dependencies using connection duration and other attributes. Roger Quinlan's C4.5 algorithm, its resulting decision tree and a random forest simplify the step-by-step identification process, while maintaining its merits. Partial correlations between nodes are simplified using Rattle programming, and the dataset is divided into majority nodes to assist processing.

This research also includes a behavioural survey of the nodes and an extended analysis using a classification system for sybil detection, with a particular focus on sybil attacks in mobile social network environments. Each sybil node is tracked and identified based on the frequency and duration of its connections with other nodes.

An outline of how the classified model identifies behaviour is also included, along with an explanation of the flow of the decision tree and the C4.5 algorithm process, which press-gangs identified sybil nodes based on the results of entropy calculations and information gain. The calculated entropy for each node connection across the all datasets informs the information gain. The

maxGain calculations for individual node bring the final stage of draw decision tree and helped to predict the sybil nodes, compare and justify the sybil attackers .

These processes and new models applied to sybil detection provide insight into the behaviour of connections, through deep analytics and entropy gain. The evidence gleaned from this research brings significant knowledge to data analytics and data science in the identification of threats on mobile social networks.