P2P network classification

A both port and payload agnostic approach

P. J. Molijn

Student: 851437445
Date: 9/9/2014

Open Universiteit
www.ou.nl

P2P NETWORK CLASSIFICATION

A BOTH PORT AND PAYLOAD AGNOSTIC APPROACH

by

P. J. Molijn

in partial fulfillment of the requirements for the degree of

Master of Science
in Software Engineering

at the Open University, faculty of Management, Science and Technology
Master Software Engineering
to be defended publicly on Tuesday September 9, 2014 at 15:00 PM.

Student number:	851437445	
Course code:	T75317	
Thesis committee:	Prof. dr. M. C. J. D. van Eekelen (chairman),	Open University
	Dr. ir. H. P. E. Vranken (supervisor),	Open University

An electronic version of this thesis is available at http://dspace.ou.nl/.

This thesis is dedicated in loving memory of my late mother, who suddenly passed away on Jan. 8th 2014. Her guidance and encouragement have enabled me to fulfill my potential.

ELEONORE EUNICE MOLIJN
(1951-2014)

ACKNOWLEDGEMENTS

Writing this thesis would not have been possible without the support of several people whom I need to gratefully thank.

I wish to thank my supervisors prof. dr. Marko van Eekelen and dr. ir. Harald Vranken for providing me with the opportunity to write this thesis under their guidance. Words cannot describe my gratitude towards Marko and Harald. Despite their busy schedules, they reserved some moments of their valuable time, they provided me with excellent feedback and guided me into the world of academia. It has been a real pleasure working with both of you and I am looking forward to our next project together. *To PhD or not to PhD?* That is the question!

I would also like to thank dr. Anda Counotte-Portman for her valuable advices and assuring that there were no delays in study progress due to the Open University's part. I am thankful to dr. Bastiaan Heeren for his availability and willingness to answer many of my questions during my study.

I would like to thank my parents. They introduced me to the wonderful world of computing by purchasing my first computer, the commodore 128. My father, Hedwig, explained complex math problems in such a way that it was a real joy to find the solution myself when I was a child. My late mother, Eleonore, who loved reading books, passed this characteristic on to me. I really enjoy reading and am thankful for that. Thanks, Mom and Dad, for believing in me and providing me with the opportunities to pursue higher education.

When my younger and only brother, Delano, achieved his BSc degree, he motivated me to pursue my Masters. Thank you for just being my brother and one of the first people of implicit or explicit motivation on my journey.

Finally, without the love and support from my wife Gracia, my son Qylan and daughter Kaylin, I would never have succeeded. Gracia, my utmost respect for keeping up with all the late nights / early mornings during my study, always there providing me with the necessary drinks and/or food to keep going, not to mention my grumpy moods you gracefully dealt with when there were some setbacks. Qylan, thanks for the time we spend together with the Martial Arts Training. Just an example of the good moments we experienced to take the mindset from research onto a different topic. Kaylin, the funny faces you can put on, combined with the most hilarious story telling. Thanks for these example moments as well. I thank you both, love you very much, am proud of you and I am proud to be your father. Although all three of you had to cope with my absence often, you seldom complained. Hopefully now, as the thesis is completed, we can spend more family time together.

P. J. Molijn
Lelystad, September 2014

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

The popularity of Peer-to-Peer (P2P) applications, and consequently the P2P traffic on the internet, has increased in the last years. This increase in traffic usage of P2P applications is besides benign P2P applications also due to malicious P2P software such as P2P botnets. To cope with the increasing threats imposed by malicious P2P botnets, botnets should be combated actively. A first step is to detect which internet traffic originates from P2P botnets. In this research, a start has been made by looking at whether internet traffic can be classified as either P2P traffic or non-P2P traffic, yet regardless of whether it concerns benign or malicious traffic.

Classification of P2P traffic is challenging since traditional techniques, that mainly analyze port numbers or payload data, are becoming ineffective against applications that use random ports or encryption. This research proposes, based on literature study, Machine Learning (ML) as a method for P2P traffic classification, using the algorithms J48, REPTree and AdaBoost for analysis of statistical flow features, which are both port and payload agnostic.

The classifier is trained with a data set consisting of network traffic derived from four P2P applications, two P2P botnets and non-P2P traffic. Classifier metrics were obtained by utilizing test data sets, in such a way that each individual set is disjunct with all the other sets(including training set). The results of this quantitative empirical research show that the proposed method can achieve high accuracy, outperforming comparable existing approaches for classification of P2P traffic.

The data sets and some source codes used in the thesis will be made available to the research community to enable validation and extension of the work.

Keywords: P2P traffic, port agnostic, payload agnostic, classification, Machine learning

SAMENVATTING

De populariteit van Peer-to-Peer (P2P) toepassingen, en daarmee ook het P2P verkeer op het internet, is in de laatste jaren sterk toegenomen. Deze toename is naast het gebruik van goedaardige P2P toepassingen ook te wijten aan kwaadaardige P2P toepassingen zoals P2P botnets. Om de toenemende bedreigingen van P2P botnets te pareren, is actieve bestrijding ervan noodzakelijk. Een eerste stap daarin is om te detecteren welk internetverkeer deel uitmaakt van P2P botnets. In dit onderzoek is daarmee een start gemaakt door te kijken of internetverkeer geclassificeerd kan worden als P2P verkeer en niet-P2P verkeer, nog ongeacht of dat goed- of kwaadaardig verkeer betreft.

Classificatie van P2P verkeer is uitdagend aangezien traditionele technieken, die hoofdzakelijk poortnummers of payload-informatie analyseren, ineffectief zijn tegen toepassingen die willekeurige poorten of encryptie gebruiken. In het onderzoek is, op basis van literatuuronderzoek, Machine Learning (ML) gebruikt als methode voor classificatie van P2P verkeer, waarbij de algoritmen J48, REPTree en AdaBoost gebruikt zijn voor analyse van statistische flow features die zowel poort- als payload agnostisch zijn.

Het classificatie mechanisme leert P2P gedrag van een data set die bestaat uit zowel goedaardig P2P-verkeer, kwaadaardig P2P-botnet verkeer en niet-P2P verkeer. De nauwkeurigheid van de classifier op de daadwerkelijke test data bepaalt hoe effectief er onderscheid kan worden gemaakt tussen P2P en niet-P2P verkeer. De performance metrieken van de classifier zijn allen gebaseerd op het gebruik van test data sets, waarbij elke individuele set disjunct is met de overige sets(inclusief de training set). Uit de resultaten van dit kwantitatief empirisch onderzoek is gebleken dat hiermee een hoge nauwkeurigheid kan worden bereikt, die vergelijkbare bestaande benaderingen voor classificatie van P2P verkeer overtreft.

De datasets en enkele broncodes die tijdens het onderzoek werden gebruikt zullen publiekelijk ter beschikking worden gesteld om bijvoorbeeld validatie of uitbreiding van dit werk mogelijk te maken.

Trefwoord: P2P traffic, port agnostic, payload agnostic, classification, Machine learning

1

INTRODUCTION

1.1. BACKGROUND

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at elit ac mi elementum fermentum. Proin cursus mi non lectus interdum lacinia. Maecenas hendrerit congue augue. Aliquam porttitor quam dolor, et dapibus tortor pellentesque eget. Ut blandit eget nunc at consectetur. Suspendisse tempor pretium elit, vitae porta justo gravida in. Cras lobortis lorem sapien, et porta libero egestas sed. In ut blandit justo. Etiam in diam quis sapien porta mattis. Phasellus ac arcu auctor, tempus ligula viverra, maximus sapien. Duis quis faucibus mauris. Suspendisse at est et ex ultrices bibendum. Vivamus maximus dapibus dignissim. Nulla purus nisl, iaculis vitae fringilla in, commodo nec dolor.

Donec magna purus, vestibulum sed tellus id, cursus imperdiet odio. Quisque sodales, massa sit amet mollis ornare, nibh ipsum posuere dolor, eget venenatis ex erat sit amet ligula. Nullam et libero et ipsum ultricies bibendum. Sed vel eros dignissim, ornare purus eu, posuere tortor. Aenean sed elementum velit, non ornare augue. Maecenas porttitor, urna sed lacinia auctor, dui eros hendrerit lorem, fringilla auctor risus justo a nisl. Duis gravida vitae velit dapibus maximus. Nam ultrices, erat nec mattis euismod, neque quam pulvinar mi, sed tristique tortor augue nec lectus. Maecenas bibendum finibus ante ac venenatis. Aenean eu velit eu augue semper facilisis feugiat in velit. Maecenas maximus luctus tortor vel vulputate. Nullam fringilla commodo lobortis.

Fusce urna erat, molestie et felis sit amet, sollicitudin pharetra lectus. Nam congue, mauris vel consequat feugiat, nisi diam pretium turpis, scelerisque suscipit enim mi a felis. Integer sem mauris, mollis tincidunt lobortis eleifend, pellentesque sit amet massa. Integer condimentum mi sed metus rhoncus, eget venenatis ex gravida. Nam maximus sagittis dolor, at ultrices sapien sagittis at. Nunc dapibus euismod purus, sollicitudin commodo mauris consectetur vitae. Nunc tincidunt mi nec bibendum pulvinar. Fusce hendrerit tortor eget lacus gravida ultrices eu at turpis. Nulla tincidunt dui ac luctus accumsan. Duis porttitor neque sed nulla auctor condimentum. Etiam mattis blandit massa, suscipit venenatis ex aliquet eu. Fusce non elit congue, vehicula magna a, faucibus magna.

Duis scelerisque ex ac lectus mattis laoreet. Sed mattis tempus leo nec tincidunt. Donec eu ante et purus porttitor pharetra. Ut pretium condimentum ligula, vel lacinia ante fermentum nec. Cras id varius ipsum, vel volutpat dolor. Morbi eget magna aliquet, dapibus elit sed, scelerisque turpis. Duis in mi vel nisl efficitur condimentum. Praesent elementum, ipsum ut semper imperdiet, dolor leo volutpat augue, vitae rhoncus urna diam ut magna.

Praesent fermentum augue nisi, a fringilla turpis efficitur eget. Nunc erat nulla, lacinia laoreet felis id, rutrum ultricies eros. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In egestas, lacus ut egestas scelerisque, quam urna rhoncus est, sed mollis erat tortor id felis. Nunc tortor enim, faucibus a orci nec, ullamcorper consequat nunc. Nulla quam quam, sagittis quis sollicitudin quis, convallis sit amet est.

Nam mattis, odio sagittis varius mollis, enim velit dictum lorem, a placerat nunc erat non turpis. Nullam non urna at tellus faucibus vehicula in sed orci. Cras id enim vitae lorem aliquam tincidunt eu eget mauris. Phasellus sed felis nulla. Donec laoreet nec sapien ut facilisis. Ut risus orci, dignissim vel tincidunt at, porta ac ante. Vivamus eleifend arcu nisi, in consectetur elit tempus eget. Maecenas dictum quam ut tincidunt laoreet. Sed lorem augue, ultrices in ultrices eget, fringilla nec odio. Phasellus lacinia, orci quis tincidunt venenatis, enim est maximus velit, a vulputate orci sapien a eros. Donec lacinia malesuada vestibulum. Aenean non fermentum tortor. Suspendisse congue at metus nec aliquet. Pellentesque posuere, urna ut bibendum condimentum, tortor lectus dapibus tortor, eget iaculis dolor diam at quam. Suspendisse potenti. Morbi libero ipsum, dignissim ac urna sed, fringilla bibendum sem.

The primary security risk is brought upon us from vulnerabilities in software which is then utilized by malicious software. Malicious software is also known as malware. McGraw and Morrisett [MM00] define malicious code as *"any code added, changed, or removed from a software system in order to intentionally cause harm or subvert the intended function of the system."*

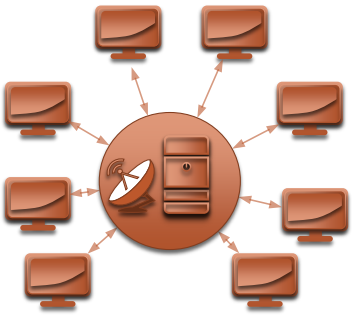
1.2. RESEARCH CONTRIBUTION

Suspendisse potenti. Vestibulum accumsan elementum magna, non viverra ligula. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Curabitur tempus aliquet bibendum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi sed libero vel magna dictum commodo. Integer dictum lacus lectus, quis dapibus sem iaculis eu. In hac habitasse platea dictumst. Suspendisse leo nisl, sodales at ligula nec, laoreet aliquam lorem. Duis in diam ex. Aenean ac augue volutpat, faucibus est quis, suscipit ante. In in magna urna. Proin scelerisque magna non ligula luctus fermentum. Morbi lectus massa, aliquam nec laoreet vitae, vulputate a mi. Cras sodales, ex quis viverra fermentum, ipsum nunc vulputate odio, quis tincidunt nisi orci suscipit nisl. Duis dignissim vehicula metus at volutpat.

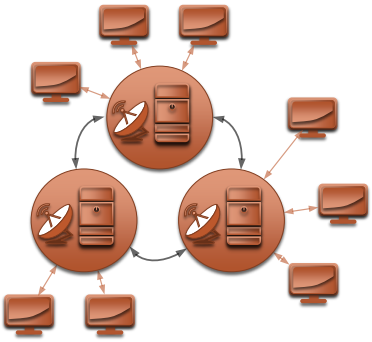
1.3. DELIVERABLES

The deliverables of the research project are the following:

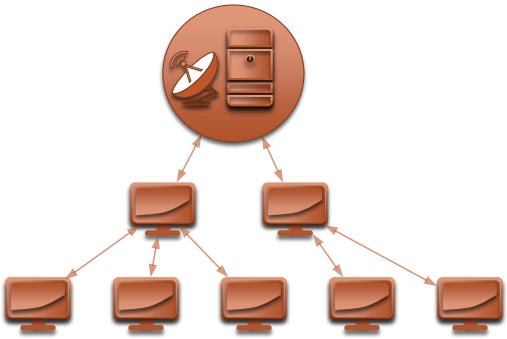
- Tool(s) for flow feature extraction.
- Algorithm for P2P Traffic classification.
- Definition of relevant features for P2P traffic classification.
- A traffic classification approach not relying on port nor payload combined with flow analysis.



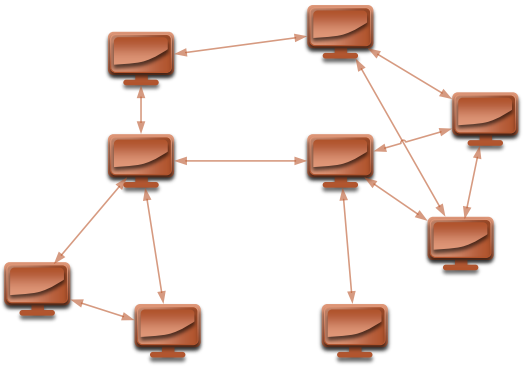
(a) Centralized botnet topology.



(b) Multi-Server botnet topology.



(c) Hierarchical botnet topology.



(d) P2P botnet topology.

Figure 1.1: Botnet communication topology.

2

PEER-TO-PEER SYSTEMS

This chapter provides background information regarding P2P systems. The most important classification of P2P systems is their degree of centralization and their network structure. A brief description of each of the P2P classifications along with their advantages and disadvantages are described.

2.1. INTRODUCTION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean sodales a ex sit amet fringilla. Pellentesque eu libero eget leo tempor imperdiet. In tempor lorem pharetra neque mollis, quis efficitur odio porta. Donec ultrices sodales tellus, ac ornare quam pulvinar vitae. Proin in cursus nisi. Vestibulum et lectus quis mi consectetur interdum sit amet id dui. Pellentesque non magna vitae tellus malesuada egestas. Nunc sed lectus vel quam sollicitudin semper vel vehicula lectus. Sed quis massa eget dui lacinia pretium. Vestibulum aliquet felis nec mauris mattis varius sed in massa. Etiam sodales at arcu eget maximus. Donec in rutrum lectus. Morbi malesuada eu velit vitae aliquam. Curabitur enim nulla, porttitor et luctus at, tristique quis odio.

Suspendisse potenti. Vestibulum accumsan elementum magna, non viverra ligula. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Curabitur tempus aliquet bibendum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi sed libero vel magna dictum commodo. Integer dictum lacus lectus, quis dapibus sem iaculis eu. In hac habitasse platea dictumst. Suspendisse leo nisl, sodales at ligula nec, laoreet aliquam lorem. Duis in diam ex. Aenean ac augue volutpat, faucibus est quis, suscipit ante. In in magna urna. Proin scelerisque magna non ligula luctus fermentum. Morbi lectus massa, aliquam nec laoreet vitae, vulputate a mi. Cras sodales, ex quis viverra fermentum, ipsum nunc vulputate odio, quis tincidunt nisi orci suscipit nisl. Duis dignissim vehicula metus at volutpat.

Ut mollis eget sapien ut feugiat. Vivamus vehicula purus sed nisi congue tempus. Morbi ac est ac elit hendrerit pulvinar nec vitae quam. Ut viverra cursus urna, ac ultrices metus maximus non. Cras eros erat, blandit ut sapien sit amet, posuere scelerisque justo. Sed tempor luctus nunc, sit amet blandit metus convallis ac. Nulla sed leo vehicula, placerat ex non, malesuada tellus. Maecenas rhoncus tellus ac luctus venenatis. Aliquam mi libero, commodo nec nisi vitae, posuere tempus lectus. Quisque eu felis nec erat iaculis

scelerisque quis quis est. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos.

Curabitur eget mauris dapibus, mollis enim rhoncus, volutpat lorem. Vivamus ac ex eget magna ultricies dictum. Donec mollis vitae velit non egestas. Aliquam tristique lectus nunc, et commodo quam lobortis ut. Suspendisse potenti. Cras vel ipsum id mauris tincidunt commodo in sit amet lectus. Nullam eleifend velit nulla, rhoncus porta ex cursus sed.

Proin lorem augue, vestibulum id velit non, vulputate vestibulum elit. Integer at justo eu felis iaculis congue. Morbi congue, mauris convallis blandit sollicitudin, justo diam dignissim ante, id fermentum nisi ex in justo. Morbi non sapien in odio aliquam dictum consequat nec turpis. Vivamus auctor id ante in mollis. Sed eleifend at magna eget varius. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla tempor finibus elit, aliquet commodo urna fringilla non. Aenean ac tortor vitae ligula tempor cursus.

BIBLIOGRAPHY

ACADEMIC ARTICLES

[MM00] Gary McGraw and Greg Morrisett. “Attacking malicious code”. In: *IEEE software* 5 (2000), pp. 33–41.

ACRONYMS

ANN Artificial Neural Network.
AOC Area Under Curve.

C&C Command and Control.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.
DDoS Distributed Denial of Service.
DHT Distributed Hash Table.
DNS Domain Name System.
DPI Deep Packet Inspection.
DT Decision Tree.

EM Expectation-Maximization.

FN False Negative.
FP False Positive.
FPR False Positive Rate.
FTP File Transfer Protocol.

GMM Gaussian Mixture Model.
GTVS Ground Truth Verification System.

HTML Hypertext Markup Language.
HTTP Hypertext Transfer Protocol.
HTTPS Hypertext Transfer Protocol Secure.

IANA Internet Assigned Numbers Authority.

KNN *K* Nearest Neighbor.

LC Linear Classifier.
LDA Linear Discriminant Analysis.

MCC Matthews Correlation Coefficient.
ML Machine Learning.
MLA Machine Learning Algorithm.

NB Naïve Bayes.
NN Nearest Neighbor.

QDA Quadratic Discriminant Analysis.
QoS Quality of Service.
QT Quality Threshold.

REPTree Reduced Error Pruned Tree.
ROC Receiver Operating Characteristic.

SVM Support Vector Machine.

TDG Traffic Dispersion Graph.
TN True Negative.
TNR True Negative Rate.
TP True Positive.
TPR True Positive Rate.

VoIP Voice over IP.

GLOSSARY

bot A bot is a compromised computer with malicious software installed.

botherder *See* botmaster.

botmaster User who controls a botnet.

botnet A botnet is a network of bots and are controlled by a botmaster.

centroid A centroid is a data point (imaginary or real) at the center of a cluster.

malware malicious software.

P2P A Peer-to-Peer (P2P) is a type of decentralized and distributed network architecture in which individual nodes in the network (called "peers") act as both suppliers and consumers of resources, in contrast to the centralized client–server model where client nodes request access to resources provided by central servers..

servent A servent is a host within a computer network acting as both a **SERV**er and a cli**ENT**.

Supervised learning Supervised learning algorithms are trained on labelled examples, i.e., input where the desired output is known. The supervised learning algorithm attempts to generalise a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs..

swarm A swarm is a collection of peers that are interested in distributing the same content.

Unsupervised learning Unsupervised learning algorithms operate on unlabelled examples, i.e., input where the desired output is unknown. Here the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalise a mapping from inputs to outputs..