EVALUATING FLOW FEATURES FOR NETWORK

APPLICATION CLASSIFICATION

CARLOS ALCANTARA

Master's Program in Computer Engineering

APPROVED:

Michael P. McGarry, Ph.D., Chair

Patricia Nava, Ph.D.

Eric Smith, Ph.D.

Stephen Crites, Jr., Ph.D.
Dean of the Graduate School

EVALUATING FLOW FEATURES FOR NETWORK APPLICATION CLASSIFICATION

by

CARLOS ALCANTARA, B.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Electrical and Computer Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

May 2020

Acknowledgements

I wish to acknowledge my thesis advisor Dr. Michael P. McGarry of the Electrical and Computer Engineering department of the University of Texas at El Paso. I am thankful for his mentorship and guidance not only during the completion of this work, but throughout my academic career. The time and effort dedicated to assist me has never gone unnoticed and is deeply appreciated. Thank you for always being available and willing to help. I am a better student, researcher and professional because of him.

I would also like to thank my colleague Christopher A. Mendoza for his help and insight on the work conducted in the lab. He is a great researcher and I am glad I had the opportunity to have worked with him.

I am deeply thankful for my parents and sister. None of my success would be possible if it wasn't for their love and support. Thank you for all that you constantly do to enable me to pursue my goals.

And finally, I am extremely grateful for my girlfriend. She has always encouraged and assured me that I would succeed. Words cannot express how appreciative I am of her unconditional support and encouragement through this and all my endeavors. I am truly fortunate to have such a wonderful woman by my side.

Abstract

Communication networks provide the foundational services on which our modern economy depends. These services include data storage and transfer, video and voice telephony, gaming, multimedia streaming, remote invocation, and the world wide web. Communication networks are large-scale distributed systems composed of heterogeneous equipment. As a result of scale and heterogeneity, communication networks are cumbersome to manage (e.g., configure, assess performance, detect faults) by human operators. With the emergence of easily accessible network data and machine learning algorithms, there is a great opportunity to move network management towards increasing automation. Network management automation will allow for a reduced likelihood of human error in network configuration, improved productivity from network managers as redundant tasks are automated, simplified scalability, and greater insight into network operation. Network application classification, the process of identifying the network application associated with trains of packets called flows, is a critical task in the automation of network management. This association of network applications with network traffic is critical for improving network management as it will allow setting application-specific policies to optimize network operation, enhancing security measures by blocking certain applications with improved firewall configurations, and developing a more reliable quality of service by prioritizing time-sensitive applications.

This work studies the classification performance of a basket of network flow features. We utilized three categories of flow features: inherent, derived, and engineered. In our first experimental analysis, we set out to uncover the inherent and derived feature's ability to classify network flows. We developed an expert system to generate application labels to serve as training data, which is used to train our models on two inherent and one derived feature. Flows are analyzed by implementing three supervised machine learning techniques for classification: k-nearest neighbors, decision trees and random forest. These experiments varied the number of applications and type of flows, all or only large, in a traffic data capture

from UKY's university network. For our subsequent experimental analysis, we engineered three flow features based on host behavior presented by the authors of BLINC and examined their influence on traffic classification performance when combined with the features from the previous experiments. A new UKY data set is captured using deep packet inspection to obtain training labels and the same three machine learning techniques are employed. In these subsequent experiments, we varied the set of features used for classification by always including the three inherent and derived features and one combination of adding the three engineered features. Our initial experiments reveal that the inherent and derived features can adequately classify a subset of applications while focus on large flows slightly reduces performance. Our subsequent experimental analysis concludes that the use of engineered features provides a statistically significant improvement on classification performance for decision tree and random forest, while KNN is most effective with only the original three inherent and derived features.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Communication networks are the pillar holding up our modern civilization. The creation of communication networks has transformed the world by allowing the development of a global society while redefining many aspects of our lives in the process. First, communication networks have reshaped our idea of community. Traditional communities bound by geographic location are no longer our only avenue for socialization. The capacity to interact with anyone across the world has developed virtual communities, where people with similar hobbies and interest can come together to find a sense of inclusion and belonging. Our professional environment has also been impacted by permitting communication and collaboration across the globe. This has allowed companies to operate through multiple branches located in every corner of the world while still focusing on a cohesive product. As for employees, it has given many the option to work from home. Finally, communication networks have enabled e-commerce, where we can buy and sell goods and services from anywhere in the world. Whether its something as simple as a toothbrush or as significant as a car or home, we are able to make these purchases completely online. Anything that we can ever want or need is just a click away.

A clear example of just how truly vital communication networks have become to our society can be seen in the face of this COVID-19 pandemic. Communication networks have had an immediate impact as they have enabled the health industry to share information in real time to facilitate tracing the spread of the virus. A domino effect has also been felt on our day-to-day interactions with communication networks. First, the use of social media and news networks to disseminate information regarding the preventive measures, such as the stay-at-home orders and social distancing guidelines. These measures have

pushed society to find new ways to keep in touch with family and friends. Video and audio communication applications have become very useful tools for us to do so. Despite the pandemic, communication networks provide the infrastructure to allow many to continue to work from home and continue their education remotely. Tools like Blackboard Collaborate Ultra and Zoom, that allow for virtual classes and work meetings to take place, are the new norm for at least the near future. Finally, we have never been so reliant on e-commerce. Whether purchasing essentials, like groceries and take-out, or recreational items, like games or books to keep us busy in these difficult times, it is clear that communication networks have never been so necessary as they are today. Communication networks have been so ingrained in our daily routines, even before this pandemic, that it is nearly impossible to imagine a world without these services.

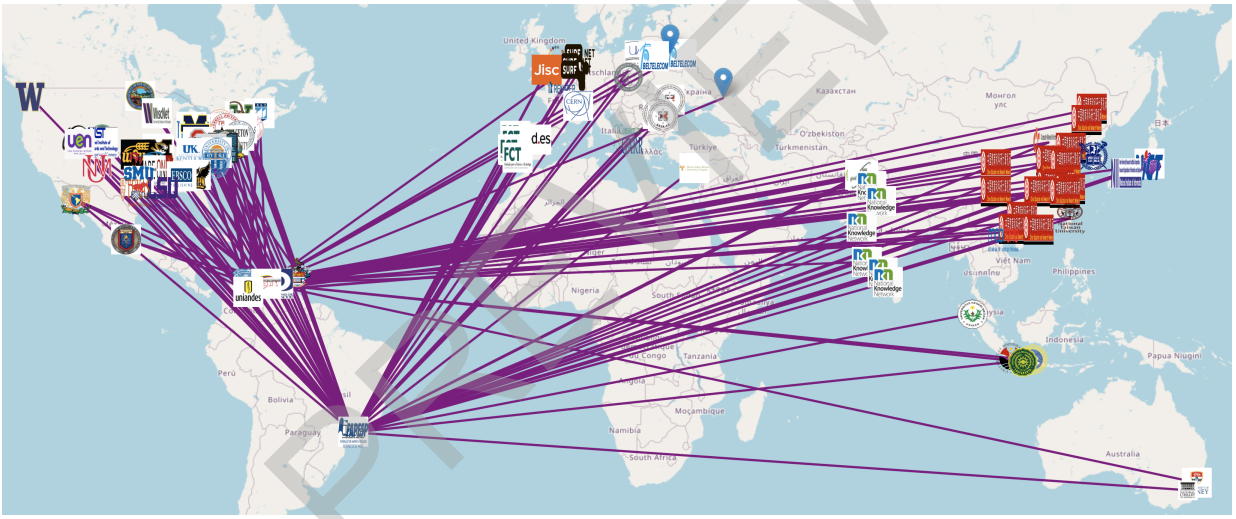

Figure 1.1: Map of autonomous systems communicating over the Internet [1].

A communication network is defined as an interconnection of computing devices with the purpose of sharing information. The largest and most obvious example of a communication network is the Internet, where network-enabled devices such as computers, smart phones and tablets are able to transmit images, audio, video and files over the network. The Internet is comprised of a series of interconnected networks. The networks that constitute the Internet are called autonomous systems (AS). Each AS in the Internet is assigned a

unique number for identification. For example, the University of Texas at El Paso (UTEP) is AS# 16461. Figure 1.1 is a visualization of the Internet, where each icon on the map represents an AS and the connecting purple lines correspond to the information transferred amongst these AS. Although this image represents a relatively small subset of Internet traffic, it provides a clear perspective of the global scale of communication networks and how information is shared across the world. Now, let us consider a topology of an AS such as the sample network of an educational institution as illustrated in Figure 1.2, where the variety of networking devices that are constantly interacting is exemplified. Core, layer 3, and layer 2 switches and routers connecting end users to file, web, and mail servers are just a few of the devices coordinating within an AS to provide communication services. This makes evident that a communications network can be a very complex system making network management a complicated task. The fact remains that none of the wonderful services mentioned earlier would be possible as they are today without communication networks. It is therefore in our best interest to make every effort so that these communication networks, that are so important to us, are operating as efficiently as possible. In order to do that, we must understand how to control, or manage, network operation.

Network management is best described using the International Telecommunications Union (ITU) M.3400 FCAPS model which delineates the responsibilities of network management as:

- **F**ault detection and correction
- **C**onfiguration and operation
- **A**ccounting and billing
- **P**erformance assessment and optimization
- **S**ecurity assurance and protection

For example, a network manager will account for the usage of its network in order to fine-tune the network configuration to provide optimal operation. Similarly, a new security

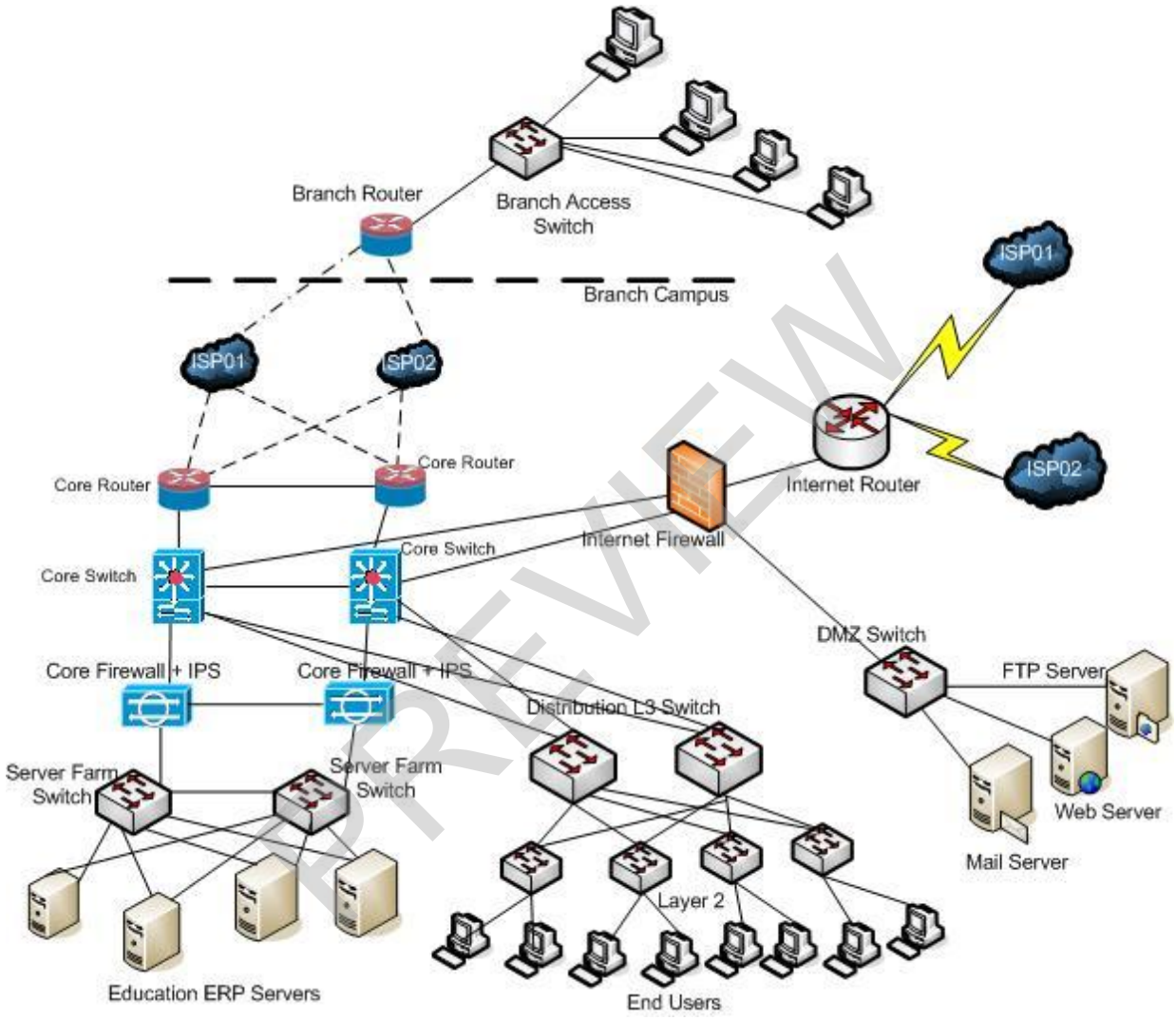

Figure 1.2: Network topology of a generic university AS [2].

policy is implemented by the network manager who then performs an assessment of network performance once this new policy is set. These responsibilities that define network management come with many challenges. There are four main challenges associated with network management. The first challenge is the scale of communication networks. This is not only in the sense that the Internet contains over 96,000 AS [3], but as Figure 1.3 illustrates, the size of a single AS can become overwhelming. The next challenge in communication networks is the heterogeneity of devices that constitute a network. Not only are there many types of devices that serve distinct functions within a network, but there are also different manufacturers with various models of these devices. Although these manufacturers design their equipment such that they are able to interoperate, most have proprietary software which indicates that the configuration of these devices is not uniform. This means that the network manager, when needing to update the Dell and D-Link switches along with the Linksys and Cisco routers to be able to access the new HP server, will need to write multiple versions of the same configuration file to update all these network devices. Now imagine having to check each of those distinct networking devices to determine which need the latest software update or a new security patch because of a new found vulnerability. Network management can quickly become a cumbersome task. The third challenge is the induction of human error. Whenever a human is interacting with any computing system there is a possibility of generating an error. Anyone who has ever written any computer code can attest to the devastating effects that a simple typo, such as a missing semicolon or closing bracket, can have on the functionality of a system. And finally, the last challenge is the frustration of a fault in network operation. The frustration can manifest in the end user detecting a misconfiguration and becoming upset with the network manager because the network that is so essential to them is not working as they expect. This can create further frustration for network managers as they are unable to detect and correct faults before the end users are affected. All of these challenges need to be handled by the network managers, who for the most part currently need to do all of this via a command line interface or SNMP. As is evident, there is a lot of work being asked of network managers.

In an effort to make network management a less burdensome task, there is a growing trend towards network management automation.

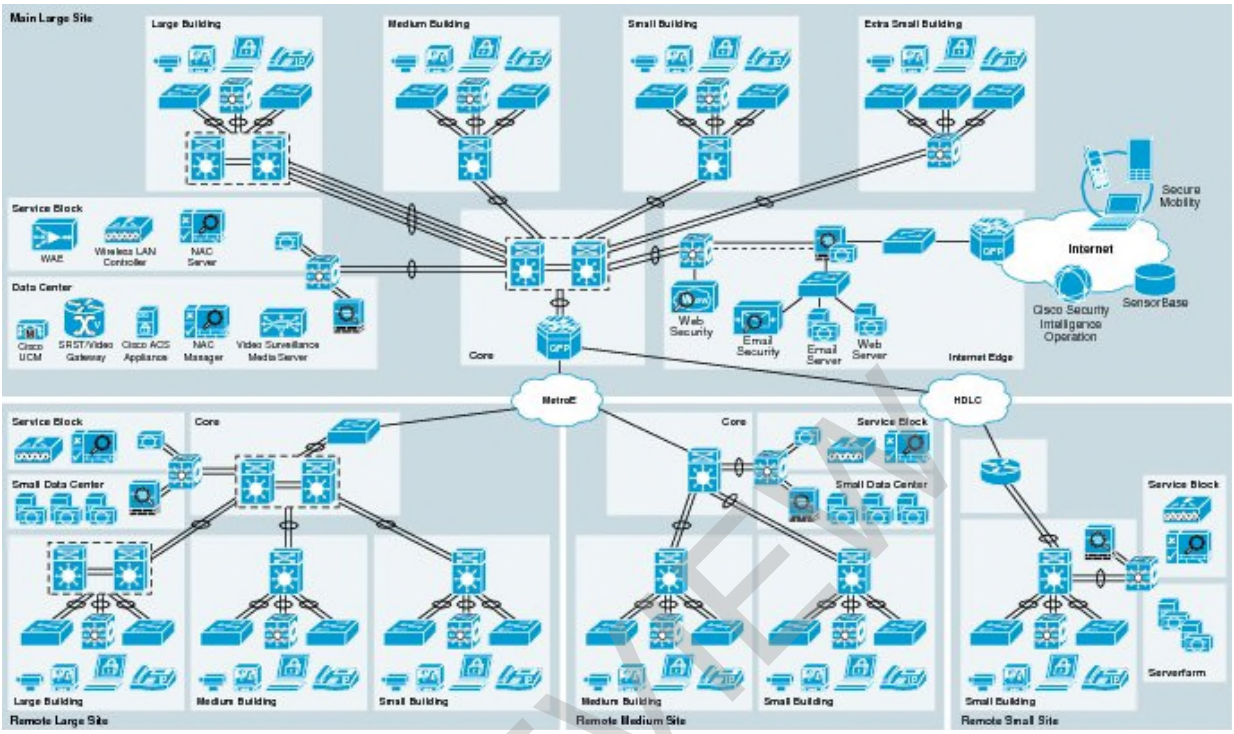

Figure 1.3: A network topology can become complex which makes them difficult to manage [4].

1.1 Automating Network Management

Network management automation is the process in which software is utilized to configure, provision, manage and test communication networks with minimal human assistance. The objective is not that of eliminating network managers from network management, but rather to limit the redundant and time-consuming tasks currently assigned to network managers by enabling these tasks to be performed automatically.

The introduction of software defined networking has opened the door for network management automation. Software defined networking (SDN) allows the separation of the data

and control planes in networking devices. While each networking device will control its data plane, SDN allows multiple network devices to be dynamically configured with software by a central controller. This means that to manage the network, changes are only made to the network controller which forwards these changes to all the networking devices within the network. Figure 1.4 provides a visual representation of SDN in comparison with traditional network management. This gives network managers simplified access to their network devices from a single location as opposed to connecting to each individual network device which becomes difficult to handle with the issues of scale and heterogeneity. Having a network that can autonomously detect and correct connection problems, self-optimize, or recognize and block security concerns such as cyberattacks would prove invaluable, and SDN gets us one step closer to achieving that goal. However, in order to make network management automation a reality, a deeper understanding of network utilization is necessary.

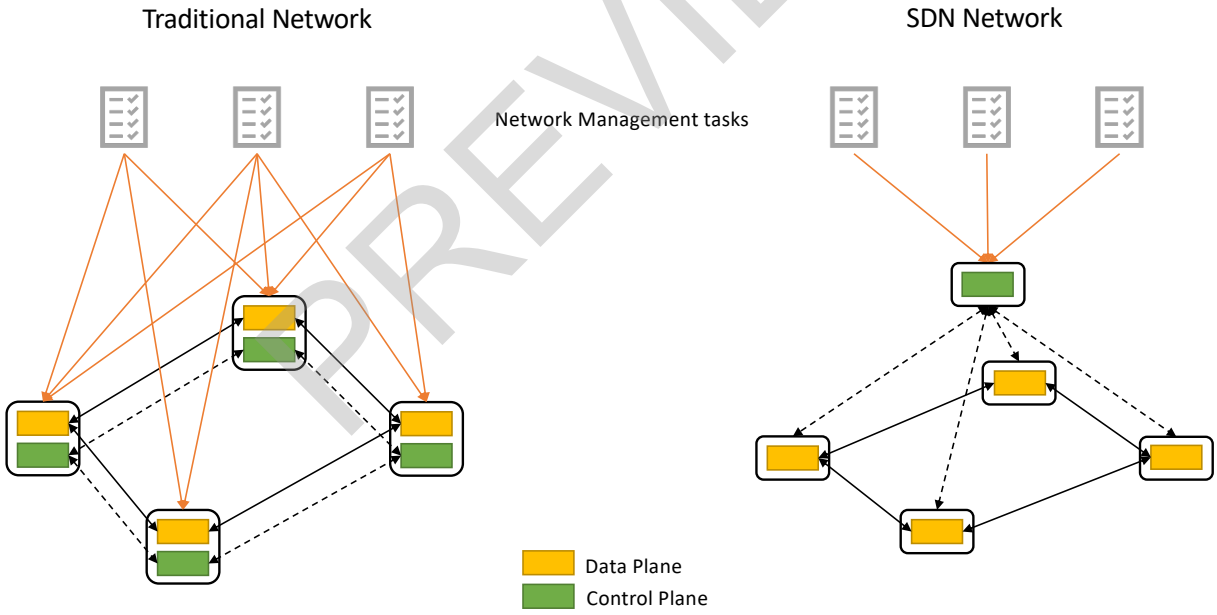

Figure 1.4: Software Defined Networking compared with traditional networking.

Let us envision a scenario where we have fine-grained knowledge of network behavior. This will allow application-specific configuration, where an application can be limited

to a percentage of network resources. For example, having an entire university network allocating all of its resources to video streaming is undesirable. By identifying traffic according to its generating application, a network manager can configure policies to limit this application to consume a third of the network resources. Understanding the performance of individual applications will also improve quality of service. By measuring latency of sensitive applications, such as live streaming, an Internet service provider can adjust the network as necessary to guarantee the service that their customers expect. Finally, recognizing how applications are communicating within a network over a period of time will allow for the generation of behavior patterns. These behavior patterns, represented by network flows, can then be used to detect faults by analyzing new network traffic with contextual anomaly detection techniques. The application that produced the network flows can serve as part of the context to detecting faults and misconfigurations before they affect the end users. The broad impact that network traffic classification by application will make on the understanding of network operation is the key to unlocking network management automation.

1.2 Network Application Classification

Network traffic consists of a series of data packets generated by a variety of applications and utilities propagating across a communications network. The process of applying a label to observed network traffic according to the program or process responsible for its creation is referred to as network application classification. The label applied is determined based on the characteristics of the network traffic (e.g. size, duration). This can be done at at the packet level, where each packet's generating application is identified, or at the flow level, where packets passing an observation point are aggregated based on similar characteristics, or features. Flows are created for the purpose of extracting a conversation communicating over the network as opposed to analyzing individual packets. Once flows are created, they are subsequently labeled according to the communicating application. The labels assigned

to the observed traffic can be coarse-grained such as peer-to-peer or bulk transfer, or fine-grained where the exact application which generated the message (e.g. BitTorrent, FTP) is identified.

There are three approaches to network application classification:

- Port number conventions - The transport layer protocol and port numbers of each packet are identified and the corresponding application registered to that combination is assigned as the generating application.

- Packet payload - The payload, which is the section of the packet where the actual message transmitted is stored, is compared to a set of patterns or signatures in order to identify the communicating application.

- Flow features - The characteristics generated during packet aggregation as well as information about the flow itself are used to infer the application generating the traffic. It is common for additional flow features to be created by combining information obtained during flow creation or by combining the information in the flows with outside data to develop supplementary features.

Although all three approaches have their advantages and disadvantages which are described in detail in Section 2.3, the port number convention's susceptibility to port abuse and inability to classify data on dynamic ports do not make it a feasible approach moving forward. Similarly, the packet payload's privacy concern along with the inability to classify encrypted data makes this approach unfit for the future as more applications are opting to encrypt their traffic. In contrast, the flow feature's wide deployment, respect of data privacy and ability to handle encrypted traffic make it the most adequate approach. The ability of the flow feature approach to classify network traffic is based on the capacity of the features generated during and after flow creation to uncover the communicating applications. Therefore, it is crucial that there is a meticulous evaluation of which features should be created and considered for network application classification when using the flow feature approach.

1.3 Thesis Outline

The inability for network managers to have a clear picture of how their network is being utilized makes network management automation an extremely difficult task. The reality is that network traffic classification is not a new problem, so it begs the question, why now? Well, there are several reasons for optimism. Recent improvements in computing systems allow for large volumes of data to be stored and analyzed quickly. This has kindled many efforts in making tools to analyze this data, such as machine learning and neural network open-source libraries, which have been made easily accessible to all. This has empowered researchers across all disciplines to leverage these tools in their efforts to solve problems specific to their domain.

Network traffic classification is a non-trivial problem with the potential to change the way in which networks are currently managed for the betterment of its users. This is a monumental task that is very actively researched by many across the world. Like many other complex problems, it is best to apply the "divide and conquer" approach where the focus is on a section of the problem and these solutions are built on top of one another as these sub-problems are solved. With this in mind, the focus of this work is in evaluating the performance of network flow features in network traffic classification. Specifically, the aim is to understand how different combinations of flow features are able to impact the ability for machine learning techniques to correctly predict the applications generating network traffic.

The rest of this work is organized as follows: Chapter 2 presents the definition of network flows, typical algorithms used for classification, approaches for network application classification and relevant work conducted on this topic. Chapter 3 describes the plan for the efforts conducted to evaluate the combination of flow features in classification performance, with results reported in Chapter 4. Finally, Chapter 5 provides concluding remarks as well as future work in this exciting and challenging problem.

Chapter 2

Background

Network traffic classification is a complex problem with many moving parts that come together in an effort to overcome this challenging hurdle in order to improve network management operation. Therefore, it is of great value to understand the individual parts involved in this problem separately before attempting to find an appropriate solution. The process of network traffic data capture and aggregation into flows is covered in section 2.1, typical algorithms used for data classification are presented in Section 2.2, the different types of data used in network traffic classification are described in Section 2.3 and finally current efforts are explored in Section 2.4.

2.1 Network Flows

2.1.1 Definition

According to [5] a flow is defined as a "set of IP packets passing an observation point in the network during a certain time interval, such that all packets belonging to a particular flow have a set of common properties." These common properties are typically contained in packet headers. For example, source and destination IP addresses, source and destination port numbers, and information about the packet itself (e.g. transport layer protocol). The purpose of aggregating network packets into flows is to convert what appears to be an unordered set of packets communicating over a network into a set of meaningful information about these interactions without having to store all of the packets themselves. These interactions can provide an increased level of understanding of the network's behavior,