Traffic characterization and classification for web applications

Student: Chiu, Yen-Chun

Advisor: Prof. Lin, Po-Ching

National Chung Cheng University

Ming-Hsiung, Chiayi 621, Taiwan

shuaichiou@gmail.com

Abstract

Software defined network (SDN) is a next-generation concept that allows network administrator to manage network flows with ease. It is programmable, centrally managed, and being able to adapt to the change of topology. Meanwhile, these characteristics also leads to new security problems, which some have already been discussed and proven. And people also give out great protection mechanisms and suggestions. However, there are still more possibilities of attacks in different scenario left to be discovered. And as SDN evolves, it will definitely cause some more new issues. In this paper, we analsis several types of topology-related attacks, and create a method that is able to detect topology poisoning attacks by using the standard functionality of OpenFlow switch. We evaluate the effectiveness of our method under different conditions, and discuss the possibility of future works.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowaday, as the popularity of public clouds like Google cloud, Microsoft Azure, Amazon EC2 increase, we can see how cloud computing offer a new way of deploying applications and service. While people try to move everything onto the cloud, the growth of volume and complexity of data center network (DCN) are getting out of hand, and server virtualization is also becomming more and more common. Networking organization are under increasing pressure to be more efficient, agile, and maintainable than is possible with the traditional approach to networking.

In traditional network, most of the network functionality heavily rely on hardware because they are implemented in network devices. They are under the control of manufacturers. As a result, the evolution of those functionality will be limited. Furthurmore, implementing a network-wide policy requires configuring at the device-level. And similarly, tasks such as provisioning, change management, and de-provisioning are also very time-consuming, error-prone and require a lot of manwork. It will surely be a formidable job for network administrators to manage a large scale network with traditional network. Scalbility is also a problem, link oversubscription is no longer reliable enought to keep up with the growing demand on data centers [24].

Software defined network (SDN) is a dynamic, manageable, cost-effective, and adaptable network network structure. It gains great popularity among enterprises as well as academia recent years. One key concept of SDN is separating the control plane from data plane, it is centrally contrallable by software application. Comparing to lagacy network, it is more flexible, maintainable and agile. Also, components of legacy network such as a regular switch is totally compatible within a SDN network structure.

Nevertheless, new technology often comes with new security problems. In addition to switches and hosts, which of course are potential targets for hackers, SDN uses controllers to realize centralized control. If the applications in controllers are compromised, the whole network might fall into attacker's hand [6]. We will talk more about these issues in later chapters.

Beside those hardware components, attackers might want to leverage some crucial software components of SDN, like OpenFlow protocol, applications in controller, and topology services. The topology-related services provides a manner to adjust to the modification of network topology automatically. Nevertheless, some parts of it is still not very mature and has proven to be insecure. It may lead to Man In The Middle attack (MITM) or Deny of Service (DoS) in a similar way to how it could happen in regular network [1]. In this thesis, we will be focusing on those topology-related issues.

The motivation of our work is that, although some work has been done to deal with all those security problems in SDN, some aspect are missing. For example, most of them do not consider the situation that switches are compromised. And to the best of our knowledge, there has not been an effective way to detect if there is any switch being compromised in the network.

With the program-configurable and managible traits of SDN, one is able to implement network function like IDS, firewall, WAF etc. We believe it is possible to implement similar defensive solution with the aids of those SDN properties. *The main goal of this thesis* is to promotes existing method and try to propose a new detection method using the standard functionality of OpenFlow switch. With this work, we hope we are able to set an example to inspire others and draw more attention from the community to concern about the security issues in SDN, rise the security awareness of SDN users, and ultimately resulting in a more mature SDN environment.

During the research, we study the specification of components as well as the potential threats in SDN. After discussing about those attacks, we try to cover the situations that were left by the works of the others. Then some threat models and the frame of our innovative method is proposed hypothetically. In the experiments, we found that XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX Eventually, we figure out a way to detect whether is an switch acts like the way it should be. Finally, we will assess how effective our method is.

The *main contributions of this paper* are as follow:

1. Analsis several types of attacks that influence the visibility of network.
2. Discusses about previously-created counter measurements.
3. Create a inovative detection measure that is efficient and simple to implement.

4. Evaluate our method under different conditions.

The following chapters in this thesis will be: Chapter 2 gives detail background knowledge of the used technology, discuss about possible threats and countermeasure. ——————————— Chapter 3 tells about our threat model and the theory of out own detection methd. Chapter 4 contains the experimental details including setup, considerations, simulations of attacks and evaluation methods. In Chapter 5, the proposed method will be evaluated under different conditions. Finally, the conclusion and future expectation of this work will be in Chapter 6.

Chapter 2

Background and related work

In this chapter, we will be looking at the structure of SDN,

2.1 Software-defined network

Software defined network (SDN) is a dynamic, manageable and cost-effective next-generation network structure. It is designed for dealing with large scale, complex, dynamic data center network existing today. There are a lot of new concepts behind SDN. First of all, it is software programmable. One can configure, manage, and optimize network resources very quickly via dynamic, automated SDN programs. For instance, companies are able to develop new applications that can further improve manageability. Secendly, it is very flexible. Abstracting control from forwarding lets administrators dynamically adjust network-wide traffic flow to meet changing needs. Furthurmore, a complete scope of the network is maintained by the central unit - controller. It is also in charge of the flows inside the network. Finally, just like all other opensource softwares, it is open standards-based and vendor-neutral. SDN simplifies network design and operation because instructions are provided by SDN controllers instead of multiple, vendor-specific devices and protocols. To sum up, SDN provides greater visibility into network, reduces manual intervention, improves maintainability, and requires less hardware budget. Figure 2.1 simply shows the architectural components of SDN, which are applications, network devices, Northbound and Southbound interface. In a SDN architecture, network devices are all*switches* instead of *switches and routers* in legacy network.

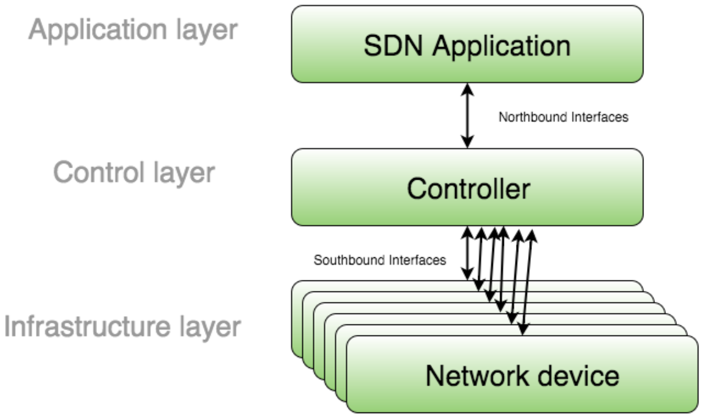


Figure 2.1: SDN architectural structure.

2.1.1 SDN Application

SDN applications are programs that communicate their requirements and desired behavior to the SDN controller via a northbound interface (NBI). An SDN Application consists of one SDN Application Logic and one or more NBI Drivers. They may consume an abstracted view of the network for their internal decision making purposes. Routing, Intrusion Detection System (IDS), firewall, network balancer, network monitor and even a modular composable service like FRESCO are all good examples of what one can build with SDN applications [18]. These funtionalities can be done at different layers of protocol stack.

2.1.2 SDN Controller

The SDN controller is a centralized entity in charge of the rules of network flows. It is responsible for managing the forwarding rules in every SDN switches and provides the abstract view of the network, incuding network topology and the state of network resources, SDN Applications need. The interface for the communication between controller layer and infrastructure layer is called Southbound Interface (SBI). Some examples are OpFlex, SoftRouter, Puppet and OpenFlow [6]. OpenFlow is the first and most popular Southbound protocol for the time being. An SDN Controller consists of NBI Agents, the SDN Control Logic, and the SBI. NOX, POX, Floodlight, Ryu and Maestro are some open source controllers that are widely used [2,3].

2.2 OpenFlow switch

OpenFlow is the first and most popular standard sounthbound interface. It enables network controllers to determine the path of packets by adding, modifying and removing matching rules in forwarding tables of the switches. It is able to realize more advanced network funtionalities than routing protocols we use in legacy network. It allows switches from different vendors as long as the devices support OpenFlow. An *OpenFlow Logical Switch* is a switch that support OpenFlow protocol. An OpenFlow switch consists of ports, flow tables and group table, Figure 2.2 is an overview of the whole OpenFlow switch structure. When a packet comes in from the *ingress port*, it will go through flow tables, group table, and will be processed in corresponding to the matching flow entry. Aside from physical switches, there are also software implementations of virtual switch, such as *Open vSwtich*. Typically, OpenFlow switches separate OpenFlow traffic and non-OpenFlow traffic with OpenFlow instances, they do not interfere with each other. [25]

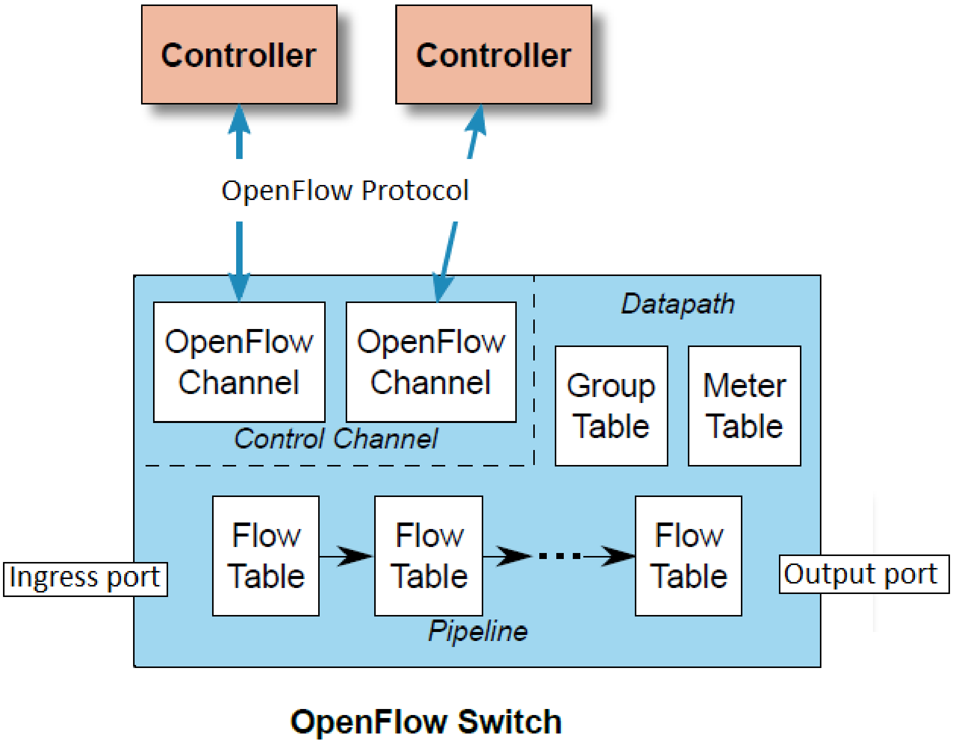


Figure 2.2: OpenFlow switch structure overview

[20]

2.2.1 OpenFlow Channel and Control Channel

The *OpenFlow channel* is the interface for the controller to config switch and receive packet from switch, and the *Control Channel* on switches may support one or more OpenFlow channel, allowing more than one controller to co-manage the switch. The connection may be initiated by either switch or controller. The OpenFlow channel is usually encrypted by TLS, but may be configured to run over plain TCP. The network used for OpenFlow channel can be either *in-band* or *out-of-band*. When using in-band network, it uses the network managed by the OpenFlow switch and is only logically separated from data plane, and the switch must set up the proper set of flow enties for the connection. If it uses out-of-band network, OpenFlow switches are connected to controller using separate dedicated network, and switches should make sure that traffic of OpenFlow channel doesn't run through the OpenFlow processing pipeline. [20]

2.2.2 OpenFlow port

OpenFlow switches are logically connected to each other by OpenFlow ports. Similarly to ports in traditional network, OpenFlow ports are the network interfaces for packets transmission. A switch sends and receives packets via ingress ports and output ports. There are *physical ports*, ports that correspond to a hardware interface of switch, and *logical port*, higher abstraction ports that may map to various physical ports. When ports are added or removed, the content of flow tables remain unchanged . Packets forwarded to deleted or non-existent ports are just dropped, and if a same port number is reused after deletion, any remaining flow entries referencing to that perticular number may be re-targeted to the new port, which possibly lead to undesirable result. Therefore, controller should clean up the reference of a port if a port is deleted. [20]

2.2.3 OpenFlow table and entry

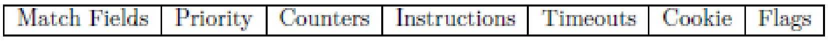
When packets arrive at a switch, it will go through the prcoessing pipeline. Figure 2.4 is a complete picture of packet processing flow inside the switch. In the pipeline, there are one or more *flow tables*, and each packet is associated with an empty *action set* at the begining. Processing always starts with the first table in ingress processing. Each OpenFLow table contains multiple flow entries. Figure 2.3 is the columns of a flow entry. A flow entry is identified by its match fields and priority. Packets will be matched with *match field*, which consists of ingress port, packet header and other optional pipeline fields. When a flow entry is matched, it will modify the action set according to its instruction. Some example of the actions are output to a specific port, drop and change TTL in the packet.

There is an entry that wildcards all field, resulting in matching all value, with a priority of 0. When a packet does not matched by all other entries in all the flow tables, it is processed according to this entry. Normally it will be encapsulated and sent the the controller, the controller decides how it should be processed, and a new flow entry will be added according to the result.

If a flow entry does not specify next flow table, all actions in the action set are executed and the whole process inside the switch is complete. Egress processing happens after the output port is determined, its structure and mechanism is almost identical to ingress processing. It is an optional feature.

It if it is not enabled, the packet is then processed by output port and in most case forwarded out of the switch. [20]

Another type of OpenFlow table is called *group table*. A flow entry may point to a group, providing additional method of forwarding. The columns of group entry is shown in Figure 2.5. Each group entry is identified by a 32-bit unsigned integer as identifier. Group types specify how the many buckets should be executed in the group. Actions in a action bucket is applied as an actions set.



Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

Figure 2.3: Columns of a flow entry

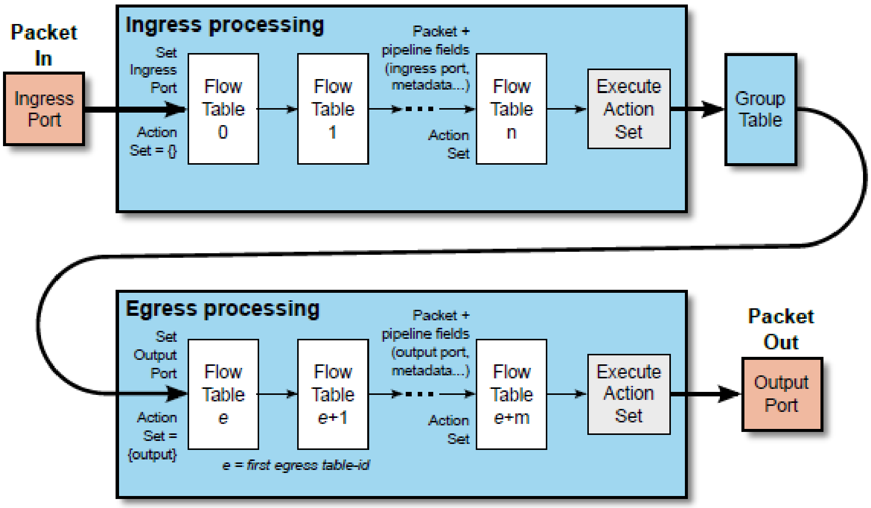


Figure 2.4: Packet processing pipeline in switch

Group Identifier	Group Type	Counters	Action Buckets

Figure 2.5: Columns of a group entry

2.3 Topology discovery services

Topology managing in SDN is quite different from it is in regular network. To realize centralized control and high programmability, controllers need to maintain the visibility of the whole network. Therefore, topology discovery services play a important role in SDN. These services are try to achieve auto adjustment when the topology alternate, which reduce the need of man effort significantly when it comes to network managing. The topology system include three parts: (1) switch discovery, (2) host tracking, (3) internal link. [7] We will describe how these parts work individually in the following subsection.

2.3.1 Switch discovery service

Switch discovery is rather intuitive. When a switch try to initiate a connection with controller, the OpenFlow channel will be established just like we mentioned in 2.2.1, the information of the switch will be sent to the controller and stored for future usage.

2.3.2 Host tracking service

n a controller, Host Profiles are maintained to keep track of the location of a host. When a Packet-Miss happen like we stated in the second paragraph of 2.2.3, a Packet-In message, along with the missed packet's information, will be sent to the controller. The controller will lookup the Host Profiles it maintains. If the Host Profile of the host cannot be found, controller assume a new host join the network and add this information. But if there is a conflict between the Host Profile and the Packet-In message, the controller treat this as a host migration and update the location information of the Host Profile. Figure 2.6 is a illustration of this service.

2.3.3 Link discovery service

When we refer to Link discovery, we mean the procedure of discovering the link between switches. Since there has not been a standard for the link discovery in OpenFlow controller, we will be using the term *OpenFLow Discovery Protocol* (OFDP) when mentioning it. Currently, although there are be some minor difference in detail, all the main stream controllers support OFDP.

The OFDP leverages the Link Layer Discovery Protocol (LLDP) with subtle modifications to perform topology discovery in an OpenFlow network. LLDP is originally implemented by Ethernet switches to exchange its identity and capabilities with a physically adjacent layer 2 peer, its EtherType field is 0x88cc. LLDP packets are sent regularly via each port of a switch [22]. The information learned from received LLDP packets is stored by all the switches and the packets will not be forwarded after being sent accross a single hop. Figure 2.7 shows the structure of LLDP Ethernet frame structure. Each LLDP Data Unit contains a sequence of type-length-value (TLV) like Figure 2.8. As OFDP use a normal destination multicast MAC, OFDP advertisements would be forwarded by backbone switches. OFDP will not interfere with LLDP messages that the backbone networks may have implemented as it uses a different MAC address. [21]

However, OFDP operates quite differently from LLDP. Just like we mentioned in the begining of this section 2.3, the topology information is kept by the controller instead of OpenFlow switch, and an OpenFlow switch will do nothing more than forward the LLDP packet. The simplified process is shown in Figure OFDP. All switches have a pre-installed rule in their flow table, sending any LLDP packet received from any port, except the controller port, back to controller via Packet-In. Initially, controller creates an individual LLDP packet for each port on each switch via Packet-Out message, each LLDP packets has their own Chassis ID. In the example figure, after receiving LLDP packet from controller, S1 send it out on Port 1 and received by S2 on Port 3. As a result to the pre-installed forwarding rule, switch S2 forwards the received LLDP packet to the controller via a Packet-In message. This Packet-In message contains meta data such as the ID of the switch and the ingress port via which the packet was received. Thus, the controller can now infer that there exists a link between Port 1 of S1 and Port 3 of S2, and this information will be added to controller's topology database. After running this process through all the switches, controller is able to obtain all link between switches in the network. The entire discovery process is performed periodically with a typical default interval size of 5 seconds. [9]

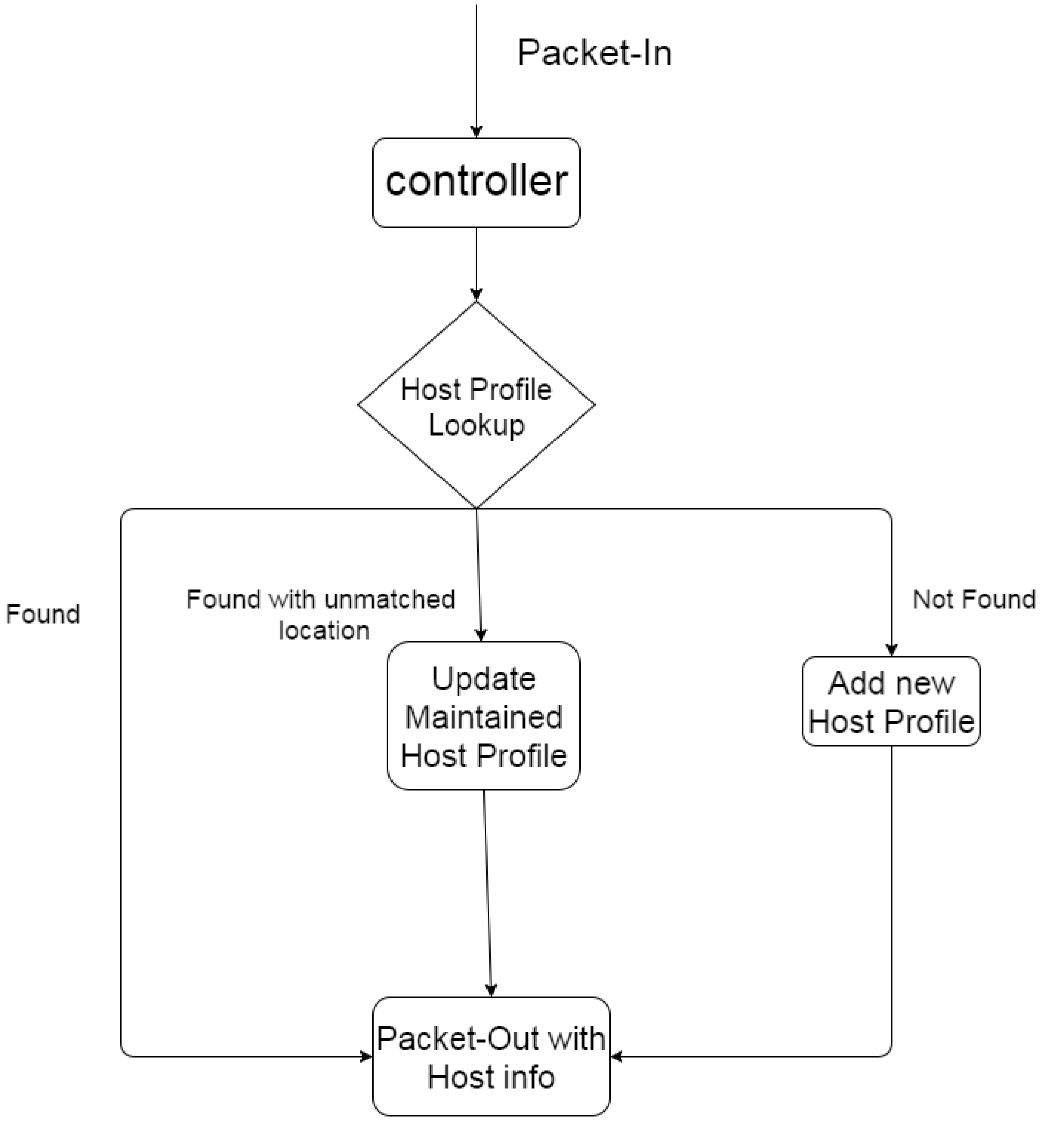
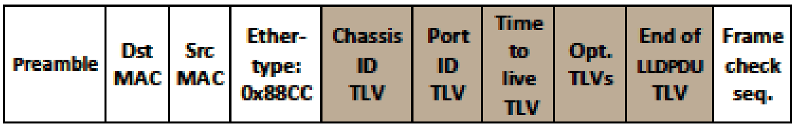
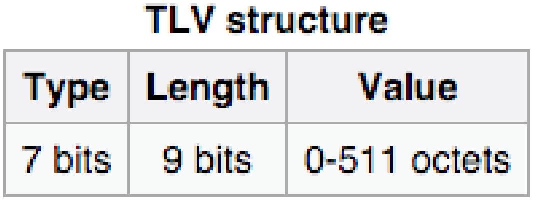

Packet-In
controller
Host Profile Lookup
Found
Found with unmatched location
Not Found
Update Maintained Host Profile
Add new Host Profile
Packet-Out with Host info

Figure 2.6: Host tracking service



Preamble	Dst MAC	Src MAC	Ether-type: 0x88CC	Chassis ID TLV	Port ID TLV	Time to live TLV	Opt. TLVs	End of LLDPDU TLV	Frame check seq.

Figure 2.7: LLDP packet frame structure [22]



TLV structure

Type	Length	Value
7 bits	9 bits	0-511 octets

Figure 2.8: TLV structure [22]

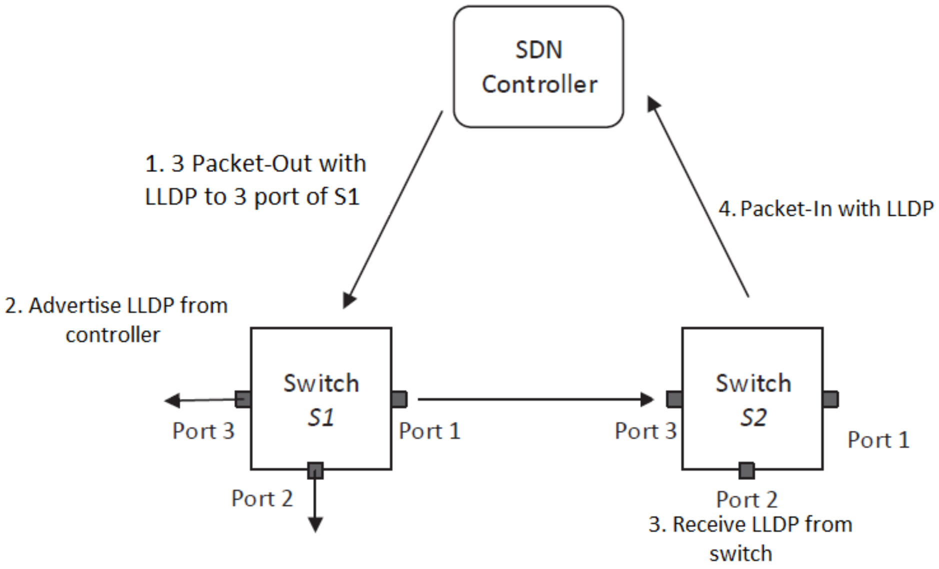

SDN Controller
1. 3 Packet-Out with LLDP to 3 port of S1
4. Packet-In with LLDP
2. Advertise LLDP from controller
Switch S1
Port 3
Port 1
Port 2
Switch S2
Port 3
Port 1
Port 2
3. Receive LLDP from switch

Figure 2.9: A illustration of OFDP procedure

2.4 SDN security

As SDN brings fascinating development to network technology, its security has become a top concern. As a consequence of introducing a new structure, new protocols and new concepts, there will be more vectors that we should be concerned about in SDN. For instances, applications inside the controller might have flaws, control channel should be secured carefully, take what a malicious user inside the network can achieve into consideration, and a lot more. And don't forget, beside those new elements, attacks may also happen in non-SDN-specific ways. Figure **??** is a whole picture of where the known attacks might possibly take place inside SDN, and Table **??** contains further discription of these threats. In this chapter, we will talk about SDN security-related issues. We will focus on some of them and discuss about the attack scenarios, possible consequences and countermeasures. With all the wonderful features of SDN, we belive it is possible to come up with good solutions to deal with the security problems.

Table 2.1: Pros and cons of each classification.

Classification	Features	Pros and Cons	Used in studies
Port-based	well-known port numbers	simply and fast but can't be used in the tunnelling traffic	[34]
Payload-based	string-based matching, regular expression matching and IP protocol	accurate classification capability and easy development but slow and can't be used in encrypted traffic	[36] , l7-filter
flow-based	packet size, packet inter-arrival time, flow duration, packet numbers,etc.	can detect the class of yet unknown applications but cannot deal with multiple transport transport protocols simultaneously	[39], [40], [41], [46], [47], [41–45]

2.4.1 General security

There are quite a few hardware devices, such as controllers, switches and hosts in SDN network. One should never overlook the devices security, regardless of whether it is inside a SDN network or not. For example, all these devices might possibly contain some software vulnerabilities inside their system. Some vulnerabilities found in recent years in the systems are Hearthbleeding, Poodle and Shellshock [26–28]. In addition to the system, other software parts of SDN, such as applications and Northbound APIs might also contain some software vulnerabilities. Moreover, one might be able to gain unauthorized access to the network physically or virtually, by using trivial methods like bruteforcing the administrative interface or a physical break in.

To deal with these kind of threat, all the best practices for hardening servers are applicable. Autonomic trust management technique could be used to harden these components in the network [17]. Also, Diego et al. propose replication, dynamic device association and self-healing concepts in their work to reinforce the security [8]. As to the administrative interface, organizations would definitely want to implemen and activate Role-Based Access Control (RBAC) policies and event logging [14]. With all these methods, the damage will be reduced if controller compromise should happen.

2.4.2 Controller and channel related attack

SDN is a centralized controllable network structure. One can easily imagine that how a malicious controller can bring great hazard. If the controller is compromised, it will give the access to control of the flow in the network under that controller to attacker. By manipulating the flows, one may cause Deny Of Service (DOS) between the desired connections or Man In The Middle (MITM) with spoofed Southbound api message to redirect flow to host they have access. Besides that, all the sensitive information including information of devices, network topology and all the croptographic keys will fall into attacker's hand, resulting in damage expansion. It is also hard to detect such a threat, since with what one can do with a controller, it is quite possible to avoid many intrusion detection methods. Therefore, it is often known as the most servere threat in SDN.

The possibilities of controller being compromised including malicious applications that send unwanted command through northbound api, vulnerabilities on controller or administrative stations etc. One may also forge fake control message the to the controller, especially if in-band connection method is used. A host also have chance to perform a DOS to the controller by deliberately crafting malicous packets for controller to prcoess slowly. Nevertheless, this kind of attack depends on the design and implementation of the controller heavily [5].

As a countermeasurement to malicious application, the security policy enforcement kernel, FortNox, introduced by Phillip et al. implements a rule detection engine and role-based authentication to mediate all OpenFlow rule insertion requests [19]. As to securing the channel, using out-of-band network for control traffic could reduce the chance of undesired control channel manipulation. Also, one should always use cryptocrophic methods such as SSL/TLS to secure the channels. However, it is not enough, vulnerabilities of SSL/TLS have been proposed and proven [16]. Kreutz et al. propose the concept of dynamic trust model and replication to maintain a trustful relationship among the connections between controller and design a secure and dependable control platform to enforce security [8].

2.4.3 Topology poisoning attack

So far, the topology poisoning attacks have been widely-disucssed. [] An attacker-controlled host can also exploiting host tracking service to XXXXXXXXX, causing the controller falsll

In [10]

TopoGuard

- Port Manager - Self-maintained_port property db - Any, Switch, Host mode, change according to received LLDP - If receive LLDP from HOST or First_Hop from switchḊetected! - Host prober - Host migration - Precondition: controller receive SHUT_DOWN_FLAG, maintain this to detect - Post condition: Host become unreachable from previous location(send ICMP to check) - Topology Update Checker - Mostly take use of Port Manager and Host prober - Check post condition according to host list maintained by Port manager - Roll back and raise alert if attack detected - Two-switch tunnel - Using a relay node to full controller about the existence of connection between to compromised switch - IMO, any communication will do(ex: web, ftp), possible? - extended two-switch tunnel - single-switch tunnel

Chapter 3

Method of malicious switch detection

A compromised switch can also drop the packet it receives and cause DOS. The detection of this kind of attack is pretty straight-forwawrd,

we will focus on the MITM types of attack

Compromised switches not only have the same capabilities as the malicious hosts, but they are also capable of performing more dynamic and severe attacks. First, a compromised switch can be used for traffic eavesdropping. Both data and control traffic passing through the compromised switch can be replicated and sent to the attacker for further processing. Furthermore, the attacker can interfere with the control traffic passing through the compromised switches to perform man-in-mhe-middle attacks [8]. By doing so, the attacker can act as the controller to some target switches. The attacker can also spoof control messages to the controller on behalf of the target

Below is our attack scenario assumption:

- All switches support the OpenFlow - protocol and are controlled by one OpenFlow controller - All switch follows the controller - Cloud service provider is trusted - potential flaw is unintentional z - Adversary either gain unfair resource, dos, info stealing - The switch has access to internet - Attacker can add,remove, modify entries - Attacker cant change the way of switch processing - Attacker has no physical access - The network uses OpenFlow as southbound protocol - The control channel is properly protected with TLS protocol, meaning that it provides confidentiality for the control traffic as well as mutual authentication between the controller and switches. - Non-compromised switch operate normally

Attacker sure can drop LLDP, resulting in DOS. However, its very noticeable.

Packet pair

- The dispersion in normal vs attack scenario, regardless of congestion

Manet protection

- Protection against packet dropping =¿ Keep trust schema according to neighbors

Possible solutions

- Auth the host

- Implement public-private key structure - Bring space/computation overhead, not very practical - Verify legitimacy of migration (Adopted)

- Controller receive Port_Down signal before host migration - Host become unreachable from previous location - Add light overhead to pack-in - Auth for LLDP packets - Add TLV authenticator to packet - Calculated by DPID and Port number - Fail to protect against LLDP relay

The network-wise visibility is one of a key point we can use to solve some problem, comparing to the ordinary network.

Most of the existing works assume that all of the openvswitches are trustable Scenario of switch compromised

- 0days - login password guessed - Not the main purpose of this paper. We will be discussing the hazards, detection and prevention

The capabilities of the attacker are largely determined by two factors: how TLS is used to protect control-plane communication, and whether the network uses in-band or out-of-band control channels. Additionally, some attacks are only possible if the attacker can perform control-channel hijacking instead of only being able to modify the flow-table of the compromised switch.

Packet forwarded to non-existent ports are just dropped

-==

Precise traffic classification relies on a clear understanding of web traffic. However, as application protocols and development techniques of web applications keep evolving (particularly, the HTTP/2 protocol [58] and new development techniques such as Ajax and node.js in recent years), the traffic characterization of web applications in prior studies may not reflect the state-of-the-art. Therefore, it is necessary to re-examine the traffic characteristics from typical web applications, and find out effective features for precise classification. Typical features in traffic characterization usually include application payloads, packet size, connection length and duration, packet inter-arrival time, and so on. We argued in [33] that the features may be unstable due to the variations of Internet traffic. In this work, we focus primarily on the application-level features that are not subject to the conditions in underlying networks. We study the traffic characteristics in popular web applications, including office applications (Google document), maps (Google maps), file sharing (Google drive and dropbox), video streaming (Youtube, Dailymotion, Tudou), and online games (Tetris battle and Dungeon Rampage on Facebook), in terms of the features on various browsers. We then extract the features from the main connection described in next section, so we do not need to compute the average and the standard deviation of the request/response lengths like [33].

3.1 Data collection

A web application usually involves highly user interactions between the browser and the web server. Developers often use techniques such as Ajax to improve user experiences, e.g., by actively pushing web content to the browser before the user requests it. Moreover, new application protocols, particularly SPDY (`www.chromium.org/spdy`) primarily developed at Google and HTTP/2 based on SPDY, support features to reduce the latency of loading web pages for efficient web browsing. Major browsers such as IE, Firefox and Chrome have supported SPDY and HTTP/2, and most of them have enabled the support by default at the time of this thesis writing. The web traffic from the Google services covered in this work is all sent over SPDY. However, the traffic from the other web applications (see Table 3.1) also use SPDY, except that from Facebook, Dailymotion and Tudou. Thus, the traffic studied in this work reflect the latest status in the development of web applications.

Since a browser may keep prior web content in a local cache to speed up web browsing, we use the guest mode of a browser (in which the web content in prior browsing will not be preserved) when interacting with a web application to ensure a complete set of packets during the interaction can be collected. We browse only a specific web application at a time to ensure the web traffic is all from that application, and use Wireshark (`www.wireshark.org`) to collect the packets.

In this work, we consider several typical scenarios of using web applications, and capture the web traffic from them. This work covers totally five types of 9 web applications, as listed in Table 3.1. We implore users to operate each web application on either Chrome or Firefox in the scenarios described in this table. A user is requested to interact with the applications for a period from one to two minutes as usual. We do not use Internet Explorer for several reasons. First, the browser is not supported on many operating systems, e.g., Mac OS and Android [59]. Second, the usage of IE is reported to be dropped to only 7.1% [60]. Third, Windows 10 no longer supports IE.

3.1.1 Extract statistical signature

A web application server may offer two or more services, so identifying the application based solely on the server's IP address is unreliable. For example, Google offers all its services on the same back-end server infrastructure (e.g., Google document and Google map can be provided by the same IP address 74.125.23.102 in our observation.), so users can reuse existing TCP connections to Google servers to access the other services [61]. Thus, classification with statistical features to distinguish the traffic from various web applications is necessary.

Table 3.1: The scenario of each web application.

no.	type	application	scenario
1	Document	Googledoc	arbitrarily typing and editing
2	Map	Googlemap	typing a location name, arbitrarily browsing the map and zooming-in and zooming-out
3-5	Video	Youtube/ Tudou/ Dailymotion	typing a video name and arbitrarily moving to a specific time position during watching
6-7	File transfer	Google drive/ Dropbox	up/downloading
8-9	Gaming	Facebook : Tetris Battle/ Dungeon Rampage	arbitrary operation

There are several requirements for the training packet traces to acquire precise application-level statistical features:

1. The collected packet traces must be generated only from the targeted web application. We set the filter of Wireshark to capture the packets from or to port 80 and 443, and ask the user runs only a web application on the browser for every time of packet collection.

2. The packets in the beginning of connections should be preserved for TCP state tracking, which is necessary for packet reassembly to recover the application-level features.

3. The collected traffic should be sufficient and diverse because it may affect the accuracy and reliability. The collected traffic must contain the packet traces from the user interactions from the targeted web application. Moreover, we performed various interactions on a same web application for every collection to ensure the diversity of packet traces. For example, we changed the frequency of typing extremely every time when we collected the traffic from Google document.

We set the target IP address and ports and follow the scenarios listed in Table 3.1 to run the web applications when collecting the training packet traces. The quantity of collected traffic is important to reflect practical user behavior precisely. We take advantage of this to observe whether diverse user behaviors will impact on the classification or the characterization. Not only the user interactions, but also the environment will influence the results. For example, some browsers support the SPDY protocol, but Mozilla Firefox did not until 2011. The difference will influence the patterns of some flows when the packets are carried over different protocols.

After collecting a new set of packet traces for a web application in the training set, we find the main connection by using the developer tool affiliated to the browser (e.g., Firebug on Firefox), which allows developers to view information about the transmitted messages. Hence we confirm which connection is the most representative of the user interactions by referring to the parameters in the HTTP messages summarized in Table 3.2. Nevertheless, many simultaneous connections may load web information in Table 3.3 to provide smooth user experiences, so we choose the longest one as the main connection. The only exception is the main connection for Tetris Battle, which transfers the elements related to the game platform instead of the game interactions. Thus, our mechanism takes advantage of this property to decide the main connection when a set of packet traces are analyzed. We can effectively segregate the noise, such as embedded advertising, with the method. This work uses the `libnids` library (`libnids.sourceforge.net`), which offers IP defragmentation, TCP stream reassembly and TCP port scan detection, to reconstruct the request-response streams of all the connections. The tool needs to track the states of TCP connections, so the beginning time of packet capturing is crucial to ensure important state transitions, e.g., 3-way handshaking, are not missing. We extract the source/destination IP addresses and source/destination ports to recognize each flow.

3.1.2 Feature definition

We assume that f_m is the main connection when a user interacts with a certain web application, and extract all n messages to be analyzed (after processed by `libnids`). Let s_i be the i-th message size in the request connection, where $i = 1 \ldots n$. We do not adopt the features in the bi-direction because including the features from the responses will introduce more ambiguity between some web applications. For example, the features generated from file download applications are sometimes similar to those from games and maps, and will decrease accuracy by 6% according to our preliminary experiment (not shown in Chapter 4). This work counts the occurrence frequency of each message size, and takes the top five frequent sizes to be represented as a vector v_i for every main connection.

Table 3.4 presents an example, in which we arbitrarily typed in a Google document on Chrome for around two minutes. Thus, the feature to characterize this interaction is (38, 34, 220, 224, 116). However, the number of

Table 3.2: The judgment of the main connection.

	Referred parameter	Judgment
Document	bundles	contain some arrays to store the characters we typed
Map	content-type	(1)image/png : comparing whether the preview picture and the map showed on the website are the same or not (2)application/vnd.google.octet-stream-compressible;charset=x-user-defined : this message continuously appear as we arbitrarily send actions to the map
Video	content-type	For example, audio/mp4, video/webm and video/f4v.
File transfer	content-disposition content-type	(1)checking the filename in the content-disposition match with the file we transfer (2)comparing the content-type match with the file we transfer
Game	content-type	the content-type is application/x-shockwave-flash game applications always use .swf to transfer files

different message sizes may be less than five to form a five-tuple vector. We take two ways to solve this problem. Table 3.5 is generated by watching a video clip on Dailymotion with Chrome. It describes that there is a vacancy within the vector. We can fill up the missing tuple with the mode (i.e, the most common value) of the other message sizes [62], and the vector becomes (636, 665, 658, 589, 636). Table 3.6 is generated by downloading a file from Dropbox on Firefox. The occurrence time of all message sizes are the same in this table. We choose the first one to fill up the vacancies, and the vector is (753, 202, 162, 753, 753). The message sizes mentioned in these three tables are sorted by the occurrence times in the decreasing order. The features are stable and the testing results are showed in Section 4.4.

Table 3.4: A feature of editing by Google document.

	s_1	s_2	s_3	s_4	s_5	s_6	...	s_n
Message size (bytes)	38	34	220	224	116	199	...	m_n
Count	177	177	37	27	13	12	...	c_n

The vector is (38, 34, 220, 224, 116).

Table 3.5: A feature of downloading by Dailymotion.

	s_1	s_2	s_3	s_4	s_5
Message size (bytes)	636	665	658	589	
Count	4	3	2	1	

The vector is (636, 665, 658, 589, 636).

Table 3.6: A feature of downloading by Dropbox.

	s_1	s_2	s_3	s_4	s_5
Message size (bytes)	753	202	162		
Count	1	1	1		

The vector is (753, 202, 172, 753, 753).

3.2 System Workflow

Figure 3.1 illustrates the workflow of the classification process. A user may use either Mozilla Firefox or Google Chrome to run a web application as mentioned above. Simultaneously, the capture filter for the web application traffic is set to port 80 and 443 on the PC. The developer tool and Wireshark is employed to find the main connection that reflects user interactions most. Next, `libnids` will reassemble the captured packets to extract application-level messages. In the last step, these features extracted from the messages are fed into Weka (`http://www.cs.waikato.ac.nz/ml/weka`) for traffic classification with four machine learning methods: NBTree, Random Forest, J48graft and Naive Bayes.

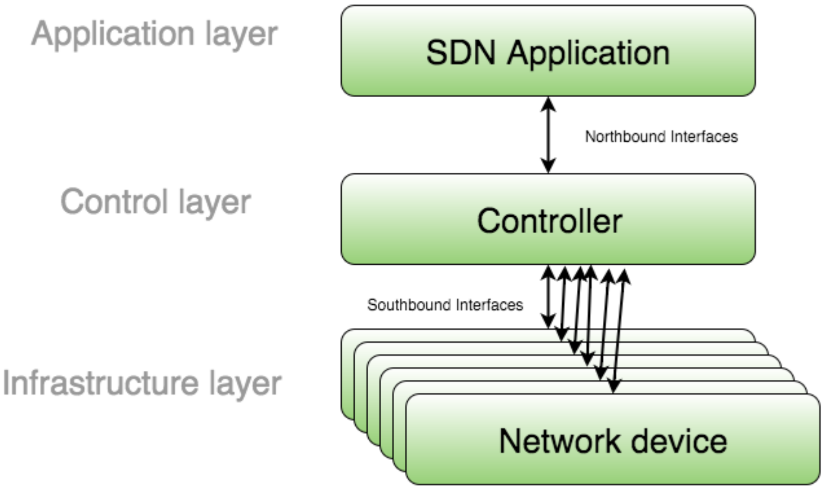


Figure 3.1: System workflow of classification.

Table 3.3: The contents of other connections

Label	Web application	Content-Type
Document	Google doc	application/json, text/html, text/javascript, font/woff, image/png, image/gif, etc.
Map	Google map	application/json, text/javascript, image/gif, image/png, etc.
Game	Dungeon Rampage	text/html, application/x-javascript, image/png, application/x-shockwave-flash, etc.
	Tetris Battle	image/gif, image/png, application/xml, application/x-shockwave-flash, audio/mpeg, etc.
File Transfer	Google Drive	text/javascript, application/json, text/css, text/html, image/gif, etc.
	Dropbox	application/x-javascript, image/gif, image/png, application/octet-stream, text/css, etc.
Video Stream	Youtube	audio/mp4, video/mp4, text/javascript, text/css, application/x-shockwave-flash, image/gif, etc.
	Dailymotion	application/x-shockwave-flash, text/css, text/xml, image/jpeg, application/x-javascript, etc.
	Tudou	application/x-shockwave-flash, image/jpeg, application/octet-stream, text/xml, image/gif, etc.

Chapter 4

Traffic Characterization and Classification Evaluation

In this chapter, we will discuss how to collect the traffic of web applications in detail and present not only the traffic characterization of various web applications, but also the accuracy of their classification.

4.1 Scenarios of Packet Collection

The network traffic is generated from user interactions on individual applications. We choose the following web applications for analysis: (1) Google docs for the office application, (2) Google map for the map application, (3) Youtube, Tudou, Dailymotion for the video streaming applications, (4) Google drive and Dropbox for the file transfer applications, and (5) Tetris Battle and Dungeon Rampage on Facebook for the game applications. In this work, we repeatedly performed the following actions ten times on either

Chrome (version 41.0.2272.89) or Firefox (version 36.0.1) for a period from one to two minutes each time.

The interactions in the office application involve arbitrarily typing characters and changing the colors. The interactions in the map application involve typing an arbitrary location name to be searched in the map, zooming-in and zooming-out the map, and browsing the map. The interactions in the game application involve (1) arbitrarily moving a role and fighting with the other roles in Dungron Rampage and (2) moving and rotating the blocks in Tetris Battle. We also consider the download functions in the video streaming and the file sharing applications. In the former application, we typed an arbitrary video name and moved to a specific time position when watching a video or just silently watched a video until the end. In the latter application, we just downloaded an arbitrary file and waited silently until the completion of the file up/downloading.

4.2 Traffic Characterization of Web Applications

A typical web session may involve several parallel connections to speed up fetching the requested content, and may also contain irrelevant connections such as those for advertisements. Thus, we manually inspected the content of requests and responses with the developer tools affiliated with the browsers when collecting the related packet traces, and focused only on the connections over which the user interactions are carried. To determine the connections associated with the user interactions, we watched the ongoing connection activities in the developer tool when interacting with a web application, and selected the longest one of these active connections as the main connection. Table 4.1 presents the average number of connections and the standard deviation in each web application. The values demonstrate that Chrome creates more connections than Firefox on average to transfer the data for the same web applications.

Table 4.1: The average number of connections and the standard deviation (SD) for each web application.

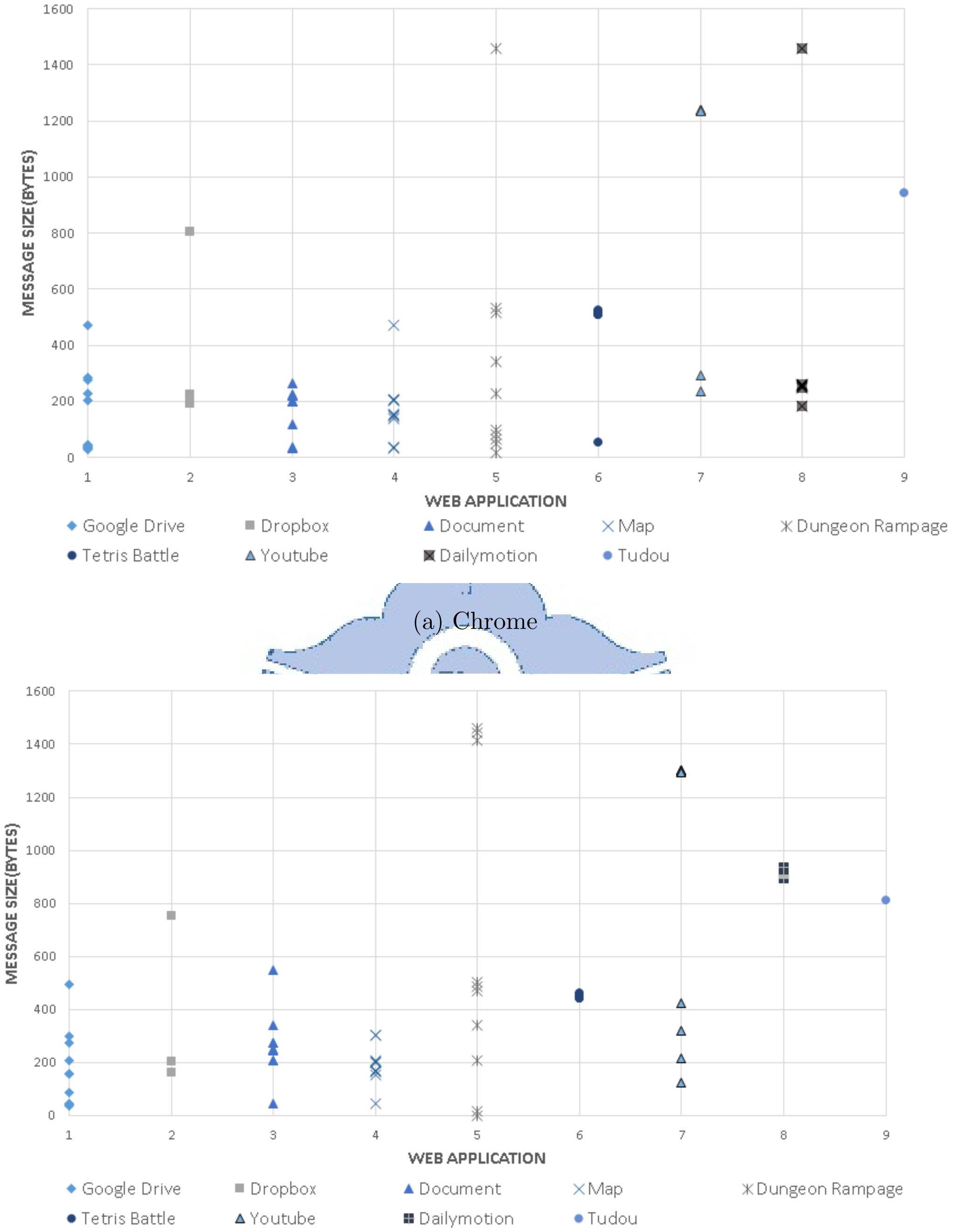
		Chrome		Firefox	
Label	Web application	Avg.	SD	Avg.	SD
Document	Google doc	37.30	4.69	33.70	2.71
Map	Google map	27.60	2.50	19.40	1.65
Game	Dungeon Rampage	125.40	10.01	108.30	7.45
	Tetris Battle	218.30	14.77	174.70	8.65
File Transfer	Google Drive	37.50	3.17	31.70	0.95
	Dropbox	68.00	11.26	61.30	3.30
Video Stream	Youtube	52.3	6.02	23.70	2.63
	Dailymotion	380	118.08	130.50	8.67
	Tudou	150.80	16.60	196.30	25.19

We analyzed the traffic from the web applications with the developer tools and Wireshark. The traffic characteristics are summarized as follows. In the office application, the characters from user typing are sent in short messages in the same connection from both Chrome and Firefox, and the length of the short massage size for each browser is either 34 bytes or 46 bytes. When we enter the website of the map application, it will display the map of the user's location according to the source IP address. Furthermore, we search a new location, the application will transfer all the data associated with the region to the browser at a time, so no further packets are transferred when we slightly zoom-in and zoom-out the map. When we play Tetris Battle, we find that around 17.28% of the packets in our each collection contain the TCP PSH flags to "push" the packets from the receive buffer to the server application because this game is highly interactive. In the file sharing applications, the client keeps transferring short messages during downloading a file on Google drive. In contrast, there are few messages transferred from the client during file downloading on Dropbox. In video applications except Youtube, the connection that downloads the video clip will be reset when the user changes the time position.

4.3 Feature analysis

4.3.1 Message size distribution

Figure 4.1 presents the top 10 frequent message sizes from the two browsers in each scenario. The messages are reassembled packet content in the main connections by `libnids`, and their sizes are ordered by the occurrence frequency. This figure demonstrates that the message sizes in different web applications vary significantly, and the occurrence frequencies of each message size also vary with the applications. The significant variations in different applications imply that the feature of message sizes should be effective for classification. Moreover, the features in the same web applications on different browsers are similar, meaning that the classification is expected to label the applications correctly even though they are running on different browsers.



(a) Chrome

(b) Firefox

Figure 4.1: The message size distribution of each web application from two browsers.

Performing the actions on similar web applications will result in different features. For example, the message size distribution of video streaming applications (Youtube, Dailymotion, Tudou) differ obviously according to Figure 4.1. However, we do not need to classify individual applications in practice because the management policies (e.g., bandwidth management or QoS) for similar applications are likely to be the same. Thus, we argue that it suffices to classify similar applications into the same category.

Figure 4.2 shows the categorization procedure. We divide the web applications into five categories: file sharing, office, map, game and video streaming. A category consists of one or multiple similar applications. If an application is correctly classified, it is also classified into the correct category. Even though the features are occasionally ambiguous between the applications in the same category (e.g., Dailymotion and Tudou), the applications can be still classified into the category of video streaming applications.

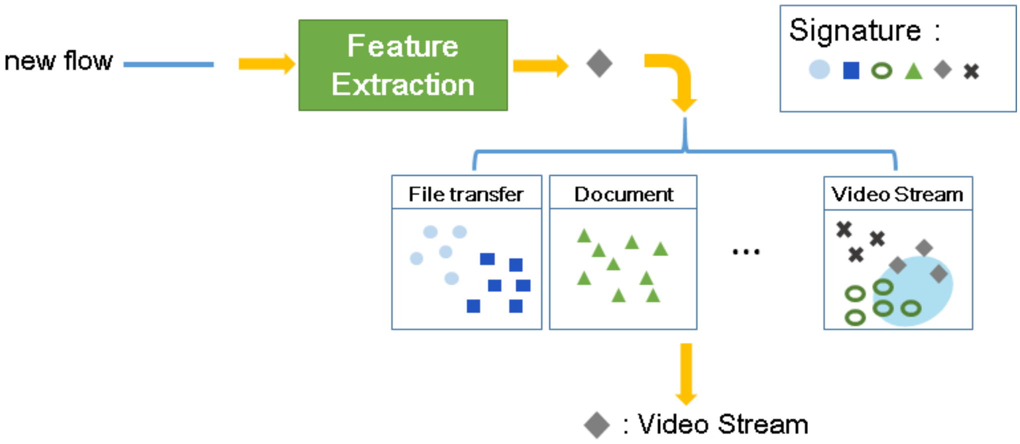


Figure 4.2: Categorization of web applications.

4.4 Classification results

After we collect sufficient data from the browsers, we employ Weka to evaluate whether the feature is effective or not. Weka supports a collection of machine learning algorithms for data mining, among which we choose NBtree, Random Forest, J48graft and Naive Bayes to compare the classification accuracy by using different algorithms. We use 10-fold cross-validation to ensure the reliability. When testing the performance, we divide the web applications into two groups, i.e., interaction applications (document, map, game) and file transfer applications (video streaming, file transfer), since we want to compare the result with the previous method, which tests the accuracy of interaction applications and file transfer applications separately. In addition, the possibility of early classification is also evaluated and it will be described in detail on Section 4.4.4.

We choose the previous study to compare with our mechanism. Lin et al. presented the method based on fine-grained classification and used the average and the standard deviation of the request/response lengths as the features for classification. The difference between our targets is that we only want to classify the kinds of web applications; however, they want to distinguish what action a user does. In the work, they subdivide the flows into search keystrokes, editing, action and download. To dislodge the unnecessary message, they artificially input the first request time or the first request bytes. Nevertheless, that method is more suitable for offline analysis. The reason we choose this study as our comparison is that the way of features extraction ,which is independent of the network condition under the application layer, is similar to our mechanism.

4.4.1 Classification with Similar Interactions in Each Run

Table 4.2 shows the classification accuracy of each web application. The accuracy of all classifiers reach 95% or above with our method, but the values still are slightly lower than previous work. There are some reasons may affect the result. For instance, our targets are different and we do not divide search keystroke messages from the flows we extract, since search packets and interaction packets may be transferred in the same connection. The true-positive rate in Google Map does not achieve 1 is because a few instances of Google Map are misidentified as those of Google document.

Table 4.2: True/false positive rates of classifying the interactive connections.

	Correctly classified (%)		Document TP/FP		Map TP/FP		Game TP/FP	
	Prior	New	Prior	New	Prior	New	Prior	New
NBtree	99.80	97.50	0.993/0.001	1/0.033	0.999/0.005	0.90/0	1/0	1/0
RandomForest	99.89	98.75	1/0.001	1/0.017	0.999/0	0.95/0	1/0	1/0
J48graft	99.28	97.50	0.993/0.004	1/0.033	0.996/0.019	0.90/0	0.95/0	1/0
NaiveBayes	99.49	95.00	0.987/0	1/0.067	0.996/0.009	0.80/0	1/0	1/0

Prior: Previous work / New : Our method

We also chose three additional web applications to disturb classification:

Google sheets (type: document), Bing map (type: map) and Dungeon Blitz (type: game). We followed the scenarios described in Table 3.1, and operated each web applications ten times on either Chrome or Firefox. We gathered the features generated from these additional applications and the original training set to verify the accuracy after the disturbance. The result is showed in Table 4.3 and the accuracy after 10-fold cross-validation can be still up to 97.14% for NBtree. In this work, we add additional applications only to the group of interaction applications to verify the accuracy because the variation of operations for file transfer applications are usually low.

Table 4.3: True/false positive rates of classifying the interactive connections with additional web applications.

	Correctly classified (%)		Document TP/FP		Map TP/FP		Game TP/FP	
	Orig.	Add.	Orig.	Add.	Orig.	Add.	Orig.	Add.
NBtree	97.50	97.14	1/ 0.033	0.975/ 0.03	0.90/0	0.925/ 0.01	1/0	1/0
RandomRorest	98.75	95.71	1/ 0.017	0.95/ 0.04	0.95/0	0.90/ 0.02	1/0	1/0
J48graft	97.50	92.14	1/ 0.033	0.90/ 0.06	0.90/0	0.85/ 0.04	1/0	0.983/ 0.013
NaiveBayes	95.00	95.71	1/ 0.067	0.975/ 0.05	0.80/0	0.875/ 0.01	1/0	1/0

Orig.: Original training set / Add.: With additional web applications

We also differentiate between downloading of video stream and general file transfer. We typed a video name and arbitrarily switched to multiple time positions when watching a video or just silently watched a video until the end on Youtube, Dailymotion and Tudou on the two browsers, but we up/downloaded files by Google drive and Dropbox. In summary, we watched 60 videos and transferred 40 files. After extracting the features, we used 10-fold cross-validation to evaluate the classification. It is presented in Table 4.4 that the classification accuracy can be up to 97% for all the classifiers, even better than previous work.

Table 4.4: True/false positive rates of classifying downloading connections.

	Correctly classified (%)		Video Streaming TP/FP		Files Transfer TP/FP	
	Prior	New	Prior	New	Prior	New
NBtree	92.72	97.00	0.907/ 0	0.967/ 0.025	1/ 0.093	0.975/ 0.033
RandomForest	92.33	97.00	0.902/ 0	0.967/ 0.025	1/ 0.098	0.975/ 0.033
J48graft	91.57	97.00	0.893/ 0	0.967/ 0.025	1/ 0.107	0.975/ 0.033
NaiveBayes	92.72	97.00	0.907/ 0	0.967/ 0.025	1/ 0.093	0.975/ 0.033

Prior : Previous work / New : Our method

4.4.2 Classification with Diverse Interactions in Each Run

In this subsection, instead of separating the web applications into interaction function and download function, we classify the traffic into five categories: document, map, game, video stream and file transfer. In the prior study of [33], the authors evaluated classification separately because the features for interaction function and download function are different in that study, so we also separate the evaluation as well in the last subsection to compare with the prior study fairly. However, we use the same feature, i.e., top-5 most frequent message sizes from the requests, for all the web applications, so we classify all the categories for evaluation in this subsection. We also operated each web application ten times on either Chrome or Firefox, but deliberately performed an extremely different action in the same scenario. For example, we may just continuously type or intermittently edit a document for each time we collected in document function. The classification accuracy after 10-fold cross-validation also can be up to 93.89% for Random Forest, meaning that the feature is stable and can be used for the classification with high accuracy.

Table 4.5: True/false positive rates of classifying all connections.

	Correctly classified (%)	Document TP/FP	Map TP/FP	Game TP/FP	Video Streaming TP/FP	Files Transfer TP/FP
NBtree	92.22	1/ 0.05	0.80/ 0.006	0.95/ 0.007	0.917/ 0	0.925/ 0.029
RandomForest	93.89	0.90/ 0.013	0.90/ 0.019	0.975/ 0.021	0.933/ 0.017	0.95/ 0.007
J48graft	92.22	0.95/ 0.031	0.80/ 0.013	0.975/ 0.021	0.917/ 0.008	0.925/ 0.021
NaiveBayes	92.22	0.95/ 0.025	0.95/ 0.013	0.95/ 0.029	0.917/ 0.017	0.875/ 0.014

4.4.3 Classification with Interactions from Multiple Users

To ensure the accuracy does not too depend on specific users, we invited eight users to generate the traffic for the classification. In this subsection, we compare only interactive functions because the their actions are more diverse than download functions. The users were asked to perform similar actions to those from a single user and repeated three times for each web applications on Chrome and Firefox. The total number of packet traces is 192: 48 for Google document, 48 for Google map, 48 for Dungeon Rampage and 48 for Tetris Battle.

Table 4.6 shows the accuracy for each algorithm. We also used 10-fold cross-validation for this classification. The accuracy of our method for Naive Bayes is even higher than the results in Section 4.4, and can be up to 97.92%

with random forest. Compared the result with the training set shown in last subsection, the accuracy is degraded just slightly, meaning that the feature is stable.

Table 4.6: True/false positive rates of classifying the interactive connections for multiple users.

	Correctly classified (%)		Document TP/FP		Map TP/FP		Game TP/FP	
	Prior	New	Prior	New	Prior	New	Prior	New
NBtree	96.63	96.35	0.939/ 0.025	0.938/ 0.028	0.971/ 0.044	0.938/ 0.021	0.993/ 0	0.99/0
RandomForest	98.28	97.92	0.954/ 0.009	1/ 0.007	0.99/ 0.032	0.979/ 0.014	1/0	0.969/ 0.01
J48graft	97.65	93.75	0.939/ 0.011	0.917/ 0.035	0.986/ 0.044	0.917/ 0.035	0.993/ 0.001	0.958/ 0.021
NaiveBayes	95.74	95.83	0.87/ 0.017	0.938/ 0.028	0.98/ 0.091	0.938/ 0.028	1/ 0.001	0.979/ 0

Pre: Previous work / New : Our method

4.4.4 Early Classification

The classification method in this work can be applied to manage various web applications, even though the web traffic is encrypted. The specifics of classification may vary in practice, depending on the purpose of management. If the purpose is accounting the usage of web applications in a network, it would suffice to classify the web applications offline based on the features from *the entire packet traces* that have been observed. However, if the purpose is access control, classification after extracting the features from the entire packet traces will be too late to block the traffic in time. Thus, we also evaluate the effectiveness of early classification with the features from the partial messages in the beginning of the connections. In early classification, a connection can be blocked as soon as the web application is recognized based on the features from the partial messages.

The packet traces for evaluating early classification are same as those described in the last subsection. We extracted the first k messages from each main connection ($k = 15, 20, 25, 30, 50, 100$) and used the size distribution of only these messages for the evaluation. Like the features from the entire packet traces, we also extracted the top five frequent message sizes in terms of occurrence frequency.

The previous analysis show that the true-positive rate of almost all kinds of functions are at least up to 0.90; however, the true-positive rate for the map application is decreased to 0.80 because its features are not so stable as others. We choose Dungeon Rampage on Chrome as example for stable one shown as Figure 4.3 and Google map on Chrome as example for unstable one shown as Figure 4.4. The line chart is created by entire messages within a main connection and the scatter chart is created by the first k messages in each figure.

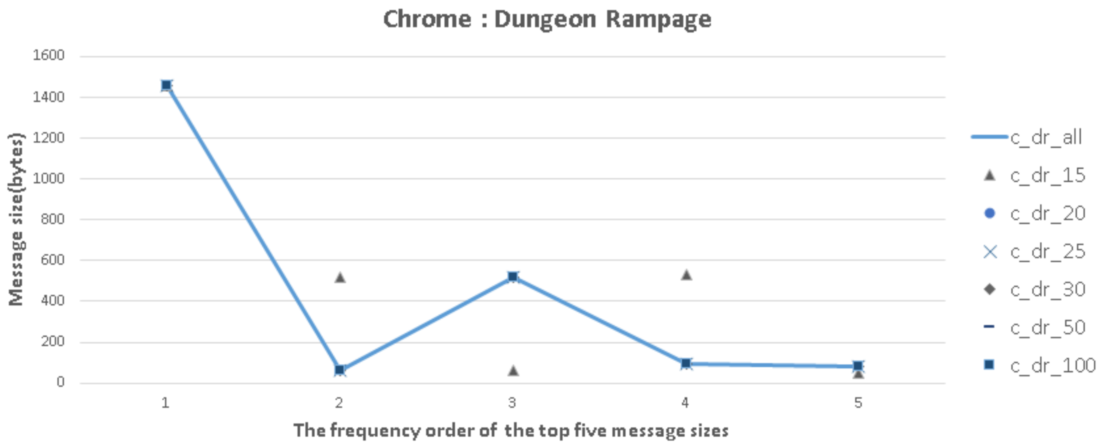


Figure 4.3: The message size distribution for Dungeon Rampage.

Figure 4.3 depicts that the scatter charts are similar with the trend of line chart when we extract more than the first 20 messages. Figure 4.4 depicts that the scatter charts are almost similar with the trend of line chart when we extract more than the first 30 messages. So we finally extracted the first 30 messages from each web applications as our feature for early classification. The classification accuracy after 10-fold cross-validation can be at least 86.67% and even can be up to 93.89% for Random Forest.

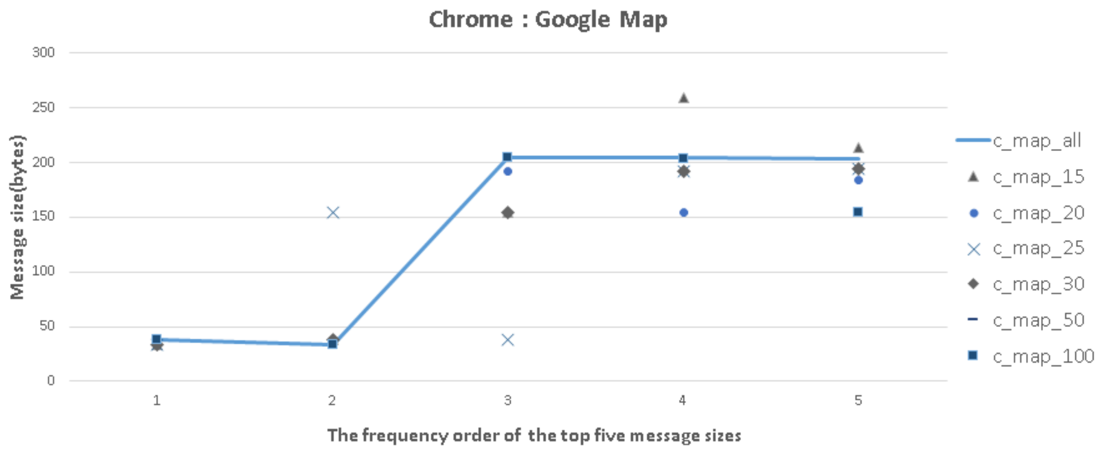


Figure 4.4: The message size distribution for Google map.

Table 4.7: Early classification true/false positive rate in all connections.

	Correctly classified (%)	Document TP/FP	Map TP/FP	Game TP/FP	Video Stream TP/FP	File Transfer TP/FP
NBtree	86.67	0.70/ 0.038	0.95/ 0.019	0.925/ 0.007	0.933/ 0.025	0.75/ 0.079
RandomForest	93.89	0.85/ 0.013	1/0	1/ 0.014	0.933/ 0.017	0.90/ 0.036
J48graft	88.33	0.85/ 0.019	1/ 0.019	0.95/ 0.029	0.85/ 0.042	0.825/ 0.043
NaveBayes	87.22	0.85/ 0.063	0.95 /0.013	0.925/ 0.021	0.95/ 0.017	0.675/ 0.043

4.5 Practice and Limitation

Even though the classification accuracy is high for extracted packet traces from real user interactions with web applications, there are still some issues that should be addressed to deploy this classification in practice.

First, although one IP address may be associated with more than one web application, as we demonstrated earlier in this work, we can still record the mapping between the IP addresses and their associated applications from earlier classification in a list to speed up classification. Since the mapping may be one-to-many, if a remote IP address can be found in the list and it is mapped to only one application, we can leverage the result of earlier classification to label the traffic to that application directly; otherwise, we can follow the procedure described in Chapter 3 to classify the web application. The result from earlier classification can at least help reduce the scope of possible labels.

Second, the traffic for each packet trace is just generated from a specific web application in this work, but in reality, we will need to analyze multiple applications at the same time and extracting the main connection is a problem. To solve this issue, we will observe the traffic density and quantity of each IP address during a period time and if the values both reach the thresholds, we will extract it as the main connection for classification. After classification, if the device that employs the design determines to block the main connection, the connections having the same pair of source/destination IP addresses as the main connection and happening around it will be blocked as well. Furthermore, users may use a web application not in the training set. Thus, after traffic classification, we will have to compute the distance between the feature vectors of the analyzed traffic and the application(s) that the analyzed traffic is supposed to be. If the distance is too long, than the analyzed traffic will be labeled as unknown.

Third, if the main connection is identified after the web application has been executed for a period of time, it may be too late to block an application for access control since the function is likely to have already completed its execution when the flows are collected and analyzed. The early classification described in Section 4.4.4 can address this issue. Moreover, the accuracy may be decreased if the execution time of an web application function is too short to extract meaningful feature.

Chapter 5

Conclusion and future work

We can only try to anticipate what the attackers may try to target with SDNs. The deployments are new, the protocols are new, the controller software is new, and the history of past SDN attacks is unknown. Based on the SDN architecture, we can predict where an attacker may be likely to strike. If we put ourselves in the attackers shoes, we might be able to spot a weakness to exploit. Then we can harden that weakness ahead of time.

Before an organization embarks on an SDN deployment project, they should consider how they will secure the system during the early design stage. Dont leave security until the final clean-up phase. If an organization waits until it is working, then hardening the northbound and southbound control messages may cause service-affecting problems. Like most things, setting it up right from the start will save organizations many problems down the road.

Bibliography

[1] Adnan Nadeem and Michael P. Howarth, “A Survey of MANET Intrusion Detection & Prevention Approaches for Network Layer Attacks,”, In Proc. of IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 4, FOURTH QUARTER 2013.

[2] Natasha Gude, Teemu Koponen, Justin Pettit, Ben Pfaff, Martn Casado, Nick McKeown and Scott Shenker “NOX: Towards an Operating System for Networks”, ACM SIGCOMM Computer Communication Review 38.3 (2008): 105-110.

[3] EugeneNg, ZhengCai AlanL Cox TS., “Maestro: Balancing fairness, latency and throughput in the openflow control plane”, Tech. rep., Rice University, 2011.

[4] Anurag Khandelwal, Navendu Jain and Seny Kamara “Attacking Data Center Networks from the Inside”,

[5] Markku Antikainen, Tuomas Aura, and Mikko Srel, “Spook in Your Network: Attacking an SDN with a Compromised OpenFlow Switch”, In proc. of Springer International Publishing Switzerland 2014.

[6] Lakshman, T. V., Nandagopal, T., Ramjee, R., Sabnani, K. and Woo, T., "The softrouter architecture", In Proc. of ACM SIGCOMM Workshop on Hot Topics in Networking. Vol. 2004. 2004.

[7] TIEN THANH BUI, "Analysis of Topology Poisoning Attacks in Software-Defined Networking", Degree project in security and mobile computing, Second Level Stockholm, Sweden 2015.

[8] Diego Kreutz, Fernando Ramos, and Paulo Verissimo, "Towards secure and dependable software-defined networks", In Proc. of the second ACM SIGCOMM workshop on Hot topics in software defined networking. ACM, 2013.

[9] Farzaneh Pakzad, Marius Portmann, Wee Lum Tan and Jadwiga Indulska, "Efficient Topology Discovery in Software Defined Networks", In Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on, pp. 1-8. IEEE, 2014.

[10] Po-Wen Chi, Chien-Ting Kuo, Jing-Wei Guo and Chin-Laung Lei, "How to Detect a Compromised SDN Switch", In Network Softwarization (NetSoft), 2015 1st IEEE Conference on (pp. 1-6). IEEE.

[11] Sungmin Hong, Lei Xu, Haopei Wang and Guofei Gu, "Poisoning Network Visibility in Software-Defined Networks: New Attacks and Countermeasures", In NDSS. 2015.

[12] Alharbi, Talal, Marius Portmann, and Farzaneh Pakzad, "The (In) Security of Topology Discovery in Software Defined Networks.", In Proc. of Local Computer Networks (LCN), 2015 IEEE 40th Conference on. IEEE, 2015.

[13] Kreutz, Diego, Fernando Ramos, and Paulo Verissimo, “Towards secure and dependable software-defined networks”, In Proc. of the second ACM SIGCOMM workshop on Hot topics in software defined networking. ACM, 2013.

[14] Ferraiolo, David F., and D. Richard Kuhn., “Role-based access controls”, arXiv preprint arXiv:0903.2171 (2009).

[15] Hu, Ningning, and Peter Steenkiste, “Estimating available bandwidth using packet pair probing”, No. CMU-CS-02-166. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2002.

[16] Ralph Holz, Thomas Riedmaier, Nils Kammenhuber, Georg Carle, “X. 509 Forensics: Detecting and Localising the SSL/TLS Men-in-the-middle” , In Computer securityesorics 2012 (pp. 217-234). Springer Berlin Heidelberg.

[17] Yan, Zheng, and Christian Prehofer, “Autonomic trust management for a component-based software system”, Dependable and Secure Computing, IEEE Transactions on 8.6 (2011)

[18] Shin, S., Porras, P. A., Yegneswaran, V., Fong, M. W., Gu, G., and Tyson, M., “FRESCO: Modular Composable Security Services for Software-Defined Networks”, in NDSS 2013.

[19] Phillip Porras, Seungwon Shin, Vinod Yegneswaran, Martin Fong, Mabry Tyson and Guofei Gu, “A Security Enforcement Kernel for OpenFLow Networks”, In Proc. of the first workshop on Hot topics in software defined networks. ACM, 2012.

[20] https://www.opennetworking.org/images/stories/downloads/sdn-resources/onfspecifications/openflow/openflow-switch-v1.5.1.pdf

[21] http://groups.geni.net/geni/wiki/OpenFlowDiscoveryProtocol

[22] https://wiki.wireshark.org/LinkLayerDiscoveryProtocol

[23] http://www.networkworld.com/article/2840273/sdn/sdn-security-attack-vectors-and-sdn-hardening.html

[24] http://www.wipro.com/documents/insights/the-thinking-network.pdf

[25] http://h10032.www1.hp.com/ctg/Manual/c04495114

[26] https://web.nvd.nist.gov/view/vuln/detail?vulnId=CVE-2014-6271

[27] https://web.nvd.nist.gov/view/vuln/detail?vulnId=CVE-2014-3566

[28] https://web.nvd.nist.gov/view/vuln/detail?vulnId=CVE-2014-0160

[29] http://usenix.org/legacy/publications/library/proceedings/usits01/full_papers/lai/lai_html/node2.html ————-

[30] T. T. Nguyen and G. Armitage, “A Survey Of Techniques for Internet Traffic Classification Using Machine Learning,” In Proc. of IEEE Comm. Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Oct.-Dec.2008.

[31] J. S. Park, S. H. Yoon and M. S. Kim, “Performance Improvement of Payload Signature-based Traffic Classification System Using Application

Traffic Temporal Locality," In Proc. of the 15th Network Operations and Management Symposium (APNOMS), Sept. 2013.

[32] M. Roughan, S. Sen, O. Spatscheck and N. Duffield, "Class-of-servicemapping for QoS: a statistical signature-based approach to IP traffic classification," In Proc. of ACM SIGCOMM Conference on Internet Measurement, Oct. 2004.

[33] P. C. Lin, S. Y. Chen and C. H. Lin, "Towards Fine-grained Traffic Classification for Web Applications," In Proc. of Australasian Telecommunication Networks and Applications Conference (ATNAC), Nov. 2014.

[34] P. Schneider, "TCP/IP traffic Classification Based on port numbers," Diploma Thesis, Division of Applied Science, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA, 1-6

[35] Y. Xue, D. Wang and L. Zhang, "Traffic Classification: Issues and Challenges," In Proc. of the International Conference on Computing, Networking and Communications (ICNC), Jan. 2013.

[36] http://www.ntop.org/products/deep-packet-inspection/ndpi/

[37] http://l7-filter.clearos.com/

[38] B. Hullar, S. Laki and A. Gyorgy, "Efficient Methods for Early Protocol Identification," IEEE Journal on Selected Areas in Communications, vol. 32, issue 10, Oct. 2014.

[39] L. Bernaille, R. Teixeira and K. Salamatian, "Early application identification," In Proc. of the Conference on Future Networking Technologies, 2006.

[40] H. M. An, M. S. Kim and J. H. Ham, "Application traffic classification using statistic signature," In Proc. of the 15th Asia-Pacific Network Operations and Management Symposium (APNOMS), Sept. 2013.

[41] B. Schmidt, D. Kountanis and A. Al-Fuqaha, "Artificial Immune System Inspired Algorithm for Flow-Based Internet Traffic Classification," In Proc. of the IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), Dec. 2014.

[42] S. Tapaswi and A. S. Gupta, "Flow-Based P2P Network Traffic Classification Using Machine Learning," In Proc. of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Oct. 2013.

[43] F .Dehghani, N .Movahhedinia, M. R. Khayyambashi and S. Kianian, "Real-Time Traffic Classification Based on Statistical and Payload Content Features," In Proc. of the 2nd International Workshop on Intelligent Systems and Applications (ISA), May 2010.

[44] S. Huang, K. Chen, C. Liu, A. Liang and H. Guan, "A statistical-feature-based approach to internet traffic classification using Machine Learning," In Proc. of the International Conference on Ultra Modern Telecommunications and Workshops (ICUMT), Oct. 2009.

[45] M. Kohara, T. Hori, K. Sakurai, H. Lee and J. C. Ryou, "Flow Traffic Classification with Support Vector Machine by Using Payload Length," In Proc. of the 2nd International Conference on Computer Science and its Applications (CSA), Dec. 2009.

[46] M. Roughan, S. Sen, O. Spatscheck and N. Duffield, "Class-of-Service Mapping for Qos: A Statistical Signature-based Approach to IP Traffic

Classification," In Proc. of the Internet Measurement Conference (IMC), Oct. 2004.

[47] J. Zhang, C. Chen, Y. Xiang and W. Zhou, "Classification of Correlated Internet Traffic Flows," In Proc. of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), June 2012.

[48] S. Valenti and D. Rossi, "Fine-grained behavioral classification in the core: the issue of flow sampling," In Proc. of the 7th International Wireless Communications and Mobile Computing Conference (IWCMC), July 2011.

[49] B. Park, J. W. Hong and Y. J. Won, "Toward Fine-Grained Traffic Classification," IEEE Communication Magazine, vol. 49, issue. 7, pp. 104-111, July 2011.

[50] B. Augustin and A. Mellouk, "On Traffic Patterns of HTTP Applications," In Proc. of the IEEE Global Communications Conference (GLOBECOM), Dec. 2011.

[51] R. Archibald, Y. Liu, C. Corbett and D. Ghosal, "Disambiguating HTTP: Classifying web Applications," In Proc. of the 7th International Wireless Communications and Mobile Computing Conference (IWCMC), July 2011.

[52] P. Casas and P. Fiadino, "Mini-IPC: A minimalist approach for HTTP traffic classification using IP addresses," In Proc. of the 9th International Wireless Communications and Mobile Computing Conference (IWCMC), July 2013.

[53] http://libnids.sourceforge.net.

[54] M. Lee, R. R. Kompella and S. Singh, "Ajaxtracker: active measurement system for high-fidelity characterization of Ajax applications," In Proc. of the 2010 USENIX conference on Web application development (WebApps '10), 2010.

[55] S. Veres and D. Ionescu, "Measurement-based traffic characterization for Web 2.0 applications," In Proc. of the IEEE Instrumentation and Measurement Technology Conference (I2MTC), May 2009.

[56] L. Shuai, G. Xie and J. Yang, "Characterization of HTTP behavior on access networks in Web 2.0," In Proc. of the International Conference on Telecommunications (ICT), June 2008.

[57] S. Lin, Z. Gao and K. Xu, "Web 2.0 traffic measurement: analysis on online map applications," In Proc. of the 18th International Workshop on Network and Operating Systems Support for digital audio and video, ser. NOSSDAV 09, 2009.

[58] M. Belshe, BitGo, R. Peon and M. Thomson, Ed., "Hypertext Transfer Protocol Version 2 (HTTP/2)," RFC 7540, May 2015.

[59] http://internet-browser-review.toptenreviews.com/.

[60] http://www.w3schools.com/browsers/browsers_stats.asp.

[61] F. Schneider, S. Agarwal, T. Alpcan and A. Feldmann, " The New Web: Characterizing AJAX Traffic," In Proc. of the Passive and Active Measurement Conference, 2008.

[62] "The Data Mining and Knowledge Discovery Handbook," In Oded Maimon and Lior Rokach Tel-Aviv (eds.), University Israel, pp. 40-41.