Journal of Network and Computer Applications 116 (2018) 9–23

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Review

A survey on network data collection

Donghao Zhou[a], Zheng Yan[a,b,*], Yulong Fu[a], Zhen Yao[a]

[a] State Key Laboratory of ISN, School of Cyber Engineering, Xidian University, Xi'an, China
[b] Department of Communications and Networking, Aalto University, Espoo, Finland

ARTICLE INFO

Index Terms:
Intrusion detection
Attack detection
Network data collection
Network management
Network security
Packet capture

ABSTRACT

Networks have dramatically changed our daily life and infiltrated all aspects of human society. At the same time when we enjoy the convenience and benefits brought by the networks, we also suffer from a great amount of intelligent attacks and malicious intrusions. As a fundamental procedure of network security measurement, network data collection executes real time network monitoring, supports network performance evaluation, assists network billing, and helps traffic testing and filtering. Thus, it plays a crucial and essential role for dealing with network intrusion detection and unwanted traffic control. But an adaptive and effective data collection mechanism that can be pervasively applied into heterogeneous networks is still lacked. The literature we have hunted rarely comments and compares the performance of existing data collection mechanisms. In this paper, we conduct a survey on existing data collection methods, mechanisms and architectures. According to a number of proposed assessment criteria, we evaluate the performance of existing data collection mechanisms and summarize their characteristics. Furthermore, we figure out some open issues based our investigation and forecast future research directions.

1. Introduction

Networks have dramatically changed our daily life and infiltrated all aspects of human society. Every day, we use a wide variety of network services and applications, which produces a huge amount of network data. Although most of the data generated in the networks are meaningless to us, a part of them contain useful and sensitive information that should be well collected, protected and managed.

On the other hand, at the same time when we enjoy the convenience and benefits brought by the networks, we also suffer from a great amount of intelligent attacks and malicious intrusions. Various security threats are cause by different types of attacks, e.g., Denies of Service (DoS), Distributed Denies of Service (DDoS), viruses, wormhole, and password guessing or stealing attacks. For detecting these attacks, some network data should be collected in order to figure out network vulnerabilities. According to the vulnerabilities, network administrators can take corresponding actions, recover network functions, predict future network threats and enhance network security and robustness.

Network data collection can greatly help network attack detection and assist network administration. Through real-time monitoring, testing, configuring, controlling and evaluating based on network data, network administrators are able to obtain network system performance, evaluate Quality of Service (QoS) and find out network fault points. For Internet Service Providers (ISPs), network data plays the basis of traffic accounting. Statistical volume information of traffic impacts the policy of ISPs.

With the acceleration of the 5G technologies and the promotion of the Internet of things (IoT) services, large scale and high-speed networks become the focus of current research and development. In order to collect and analyze network data effectively, researchers and operators proposed a lot of systems and applications. Though numerous surveys of network traffic analysis were published, there are few investigations on network data collection (Liu et al., 2018; He et al., 2018). This survey focuses on network data collection to make up for this missing study.

For specific network scenarios and specific collection puposes, the requirements of network data are different. Thus, in the process of network data collection, it is not necessary to collect all available data from the networks (Lin et al., 2018). Since data collectors are required to collect useful information, useless and meaningless information should be dropped. Redundant information should be fused. Then the reserved data can be aggregated to generate useful features, which serve as the basis for attack identification, intrusion detection, and furthermore network security measurement.

Regarding network data collection for security measurement, some important types of data should be normally collected. Network packet is still the most common data format in current network environments. Thus, network packets are usually considered as the main objectives

* Corresponding author. State Key Laboratory of ISN, School of Cyber Engineering, Xidian University, Xi'an, China.
E-mail addresses: dhzhou@stu.xidian.edu.cn (D. Zhou), zyan@xidian.edu.cn, zheng.yan@aalto.fi (Z. Yan), ylfu@xidian.edu.cn (Y. Fu), 695154820@qq.com (Z. Yao).

https://doi.org/10.1016/j.jnca.2018.05.004
Received 18 December 2017; Received in revised form 9 April 2018; Accepted 7 May 2018

that should be investigated in the field of network data collection. But, existing approaches of packet data collection usually suffer from packet loss, especially when coming across overwhelming traffic (Morariu and Stiller, 2008). And for high-speed lines, existing approaches are usually becoming useless because of substandard capability. Flow is a group of packets with same features. Commonly, the five-tuple features, including source and destination Internet Protocol (IP) addresses, source and destination ports, protocol types, are the features of the packets. A flow-based data collection mechanism, as an alternative of a packet-based mechanism, screens the flow rather than all packets (Lee et al., 2014; Kundu et al., 2009). The flow-based data collection mechanism reduces the tasks in packet analysis and performs much better than the packet-based mechanism in gigabit networks. However, it lacks fraction fidelity because of packet and flow filtering. Log file is a data storage form widely distributed in network devices. As one of the ways in data inspect, log analysis utilizes abundant log resources, such as system logs, device logs and Web logs, to extract and parse valuable information. However, it is troublesome to implement log analysis because log files are always with huge data quantity, low information density and disordered formats (Oliner et al., 2011). In the literature, there are various kinds of other data formats and collection mechanisms (Morariu and Stiller, 2008; Oliner et al., 2011; Lee et al., 2014; Kundu et al., 2009), with different pros and cons. Due to the importance of network data collection, it is essential to review the state-of-the art in order to summarize its current advance and figure out open issues for future investigation.

There are a number of existing surveys about network data collection mechanisms researched and deployed in network architectures. For instance, Sperotto et al. made a survey about IP flow data collection mechanisms (Sperotto et al., 2010). Xu et al. (2016) compared and analyzed collection mechanisms regarding Deep Packet Inspection (DPI). Other surveys focus on investigating data collection mechanisms (Davis and Clark, 2011; Moindze and Konate, 2014; Callado et al., 2009; Lin et al., 2018) with limited research scopes, so that researchers and practioners are hard to find a mechanism to satisfy their working purposes. The literature still lacks a thorough survey on network data collection that summarizes previous results by evaluating their performance with uniform criteria in order to instruct future research.

In this survey, we made a comprehensive review on network data collection mechanisms. We first introduce the types of data carriers for data collection. Then, we propose a number of criteria for evaluating the performance of different data collection mechanisms. By employing the proposed criteria, we review the current state-of-art in order to summarize current advance, find open issues and direct future research. Specifically, the contributions of our paper can be summarized as below:

1 We propose a series of criteria to evaluate current network data collection mechanisms.
2 We thoroughly review existing network data collection mechanisms and analyze their advantages and disadvantages by employing the proposed criteria as an evaluation measure.
3 We figure out a number of open issues and indicate several promising directions to instruct future research.

The rest of the paper is organized as follows. Section 2 introduces relevant knowledge of network data formats and data collection. Section 3 proposes and justifies the evaluation criteria of data collection. Section 4 provides a classification of different data collection mechanisms, followed by a review on the current state-of-art by employing the criteria as an evaluation measure in Section 5. Then we figure out open research issues and propose future research directions in Section 6. Finally, a conclusion is drawn in the last section.

2. Overview of data carriers and data collection

This section briefly introduces the carriers of network data. They are significant for data collection mechanisms. Packets, flows, logs are widely used in mainstream data collection mechanisms. Besides, some network components, such as the controllers of Software Defined Network (SDN) implement and assist data collection. The data collection mechanisms monitor data flow locally and record available information for network quality measurement, traffic estimation and attack prevention. In what follows, we will briefly introduce three basic types of network data collection methods.

2.1. Packet based data collection

Packet is a very significant data carrier in the networks based on the TCP/IP protocol. In a packet exchanging network, effective information is divided and encoded into packets. A source node sends packets that include source and destination addresses to a destination node. When the destination acquires the packets, decoding and aggregation are executed to get expected data.

The packet has various formats according to the types of networking protocols. The packet commonly consists of two parts: packet header and its payload. The header plays the role to guide the packet to transmit in a network and mark the source information of the packet. In many data collection methods, the header becomes important to identify and filter packets. For example, some header-based methods (Davis and Clark, 2011; Kim and Reddy, 2008) classify the packets into multiple flows according to their IP addresses, ports and protocols contained in the header. The payload contains the data exchanged between communicating parties, though some of them could be encrypted.

Packet capturing is a traditional method for information acquisition in network management. It is also the most commonly used scheme (Qadeer et al., 2010; Ficara et al., 2008; Morariu and Stiller, 2008; An and Liu, 2016; Antichi et al., 2012) to accomplish the goal of network data collection. A libpcap function library provides the capability to collect all the contents of a packet flow. Nevertheless, with the wide network access from mobile devices and the popularity of cloud services, high-speed and large-scale network systems become common. The volume of network will overwhelm packet capturers. Though some improved mechanisms were presented, there is still a dilemma with regards to packet capturing. It is therefore essential to consider alternatives to packet capturing mechanisms. Some existing researches (Zhao et al., 2007; Zhang et al., 2009; Kamiyama and Mori, 2006; Ji et al., 2009) abandon the idea to capture all packets, but to adopt packet sampling mechanisms. Simple sampling and stratified sampling are two instances of them. The simple sampling mechanism randomly extracts packets from all the traffic, while stratified sampling classifies the packets and drops some packets according to groups. Moreover, some other sampling mechanisms were also proposed to suit for real-time traffic data collection.

2.2. Flow based data collection

Network flow collection is another important way for network data collection. Flow is a set of packets with the same characteristics passing through a specific observation point over a period of time. Network flow monitoring can occur at every location of the network. But network core devices are the most effective nodes to be monitored and controlled since these devices can obtain important data about cyber threats and attacks. Therefore, flow collection at network core devices is the most prevalent data collection mechanism currently. Flow collection also exists in network edge nodes and hosts. Contrary to core devices, hosts only monitor the flows pass through the hosts and collect the flow records accordingly. In edge nodes and gateways, network flows of inside switches and hosts are monitored. By applying the monitoring and collection mechanisms, inbould and outbound flow