International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

USAGE Classification of internet traffic in mobile messaging apps

GULLA MANASA **EMMADI GOUTHAM**

PG Scholar, Department of CSE, Vaagdevi College of Engineering, Autonomous, Bollikunta, Warangal Telangana, Mail id:gmanasa39@gmail.com

Assistant Professor Department of CSE,Vaagdevi College of Engineering, Autonomous Bollikunta, Warangal Telangana, Mail id: goutham.emmadi@gmail.com

ABSTRACT:

The rapid adoption of mobile messaging Apps has enabled us to collect massive amount of encrypted Internet traffic of mobile messaging. The classification of this traffic into different types of in-App service usages can help for intelligent network management, such as managing network bandwidth budget and providing quality of services. Traditional approaches for classification of Internet traffic rely on packet inspection, such as parsing HTTP headers. However, messaging Apps are increasingly using secure protocols, such as HTTPS and SSL, to transmit data. This imposes significant challenges on the performances of service usage classification by packet inspection. To this end, in this paper, we investigate how to exploit encrypted Internet traffic for classifying in-App usages. Specifically, we develop a system, named CUMMA, for classifying service usages of mobile messaging Apps by jointly modeling user behavioral patterns, network traffic characteristics, and temporal dependencies. Along this line, we first segment Internet traffic from trafficflows into sessions with a number of dialogs in a hierarchical way. Also, we extract the discriminative features of traffic data from two perspectives: (i) packet length and (ii) time delay. Next, we teach a service usage predictor to classify these segmented dialogs into single-type usages or outliers. In addition, we design a clustering Hidden Markov Model (HMM) based method to detect mixed dialogs from outliers and decompose mixed dialogs into sub-dialogs of single-type usage. Indeed, CUMMA enables mobile analysts to identify service usages and analyze enduser in-App behaviors even for encrypted Internet traffic.

1.INTRODUCTION:

Recent years have witnessed the increased popularity of mobile messaging Apps, such as WeChat and WhatsApp. Indeed, messaging Apps have become the hubs for most activities of mobile users. For example, messaging Apps help people text each another, share photos, chat, and engage in commercial activities such as paying bills, booking tickets and shopping. Mobile companies monetize their services in messaging Apps. Therefore, service usage analytics in messaging Apps becomes critical for business, because it can help understand in-App behaviors of end users, and thus enables a variety of applications. For instance, it provides in-depth insights into end users and App performances, enhances user experiences, and increases engagement, conversions

and monetization. There are four modules in our system including traffic segmentation, traffic feature extraction, service usage prediction, and outlier detection and handling. Specifically, we first built a data collection platform to collect the traffic-flows of in-App usages and the corresponding usage types reported by mobile users. We then hierarchically segment these traffic from traffic-flows to sessions to dialogs where each is assumed to be of individual usage or mixed usages. Also, we extracted the packet length related features and the time delay related features from traffic-flows to prepare the training data. In addition, we learned service usage classifiers to classify these segmented dialogs. Moreover, we detected the anomalous dialogs with mixed usages and segmented these mixed dialogs into multiple sub-dialogs of single type usage. Finally, the experimental results on real world WeChat and WhatsApp traffic data demonstrate the performances of the proposed method. With this system, we showed that the valuable applications for in-App usage analytics can be enabled to score quality of experiences, profile user behaviors and enhance customer care. Traditional methods for traffic classification rely on packet inspection by analyzing the TCP or UDP port numbers of an IP packet or reconstructing protocol signatures in its payload For example, an IP packet usually has five tuples of protocol types, source address, source port, destination address and destination port. People estimate the usage types of traffic by assuming that messaging Apps consistently transmit data using the same port numbers which are visible in the TCP and UDP headers. However, there are emerging challenges for inspecting IP packet content. For example, messaging Apps are increasingly using unpredictable port numbers. Also, customers may encrypt the content of packets. In addition, governments have imposed privacy regulations which limit the ability of third parties to lawfully inspect packet contents. Moreover, many mobile apps use the Secure Sockets Layer (SSL) and its successor Transport Layer Security (TLS) as a building block for encrypted communications.

2 SURVEY

2.1 Traffic Analysis of Encrypted Messaging Services: Apple iMessage and Beyond "This survey refer that's the course of the past decade instant messaging services Have gone from a niche application used on desktop Computers to the most prevalent form of communication in the world, due in large part to the growth of Internet enabled phones and tablets. Messaging services, like Apple Message and Whats App, handle tens of billions of messages each day from an international user base of over one billion people. Given the volume of messages traversing these services and ongoing concerns over widespread eavesdropping of Internet communications, it is not surprising that privacy has been an important topic for both the users and service providers. To protect user privacy, these messaging services offer transport layer encryption technologies to protect messages in transit, and some services, like iMessage and Telegram, offer end-to-end encryption to ensure that not even the providers themselves can eavesdrop on the messages. As previous experience with Voice-over-IP and encrypted HTTP tunnels have shown us, however, the use of state-of-the-art encryption technologies is no guarantee of privacy for the underlying message content. In this paper, we analyze the network traffic of popular encrypted messaging services to (1) Understand the breadth and depth of their information leakage, (2) Determine if

attacks are generalizable across services, and (3) Calculate the potential costs of protecting against this leakage. Specifically, we focus our analysis on the Apple iMessage service and show that it is possible to reveal information about the device operating system, finegrained user actions, the language of the messages, and even the approximate message length with accuracy exceeding 96%, as shown in the summary provided in Table 1. In addition, we demonstrate that these attacks are applicable to many other popular messaging services, such as WhatsApp, Viber, and Telegram, because they target deterministic relationships between user actions and the resultant encrypted packets that exist regardless of the underlying encryption methods or network protocols used. Our analysis of countermeasures shows that the attacks can be completely mitigated by adding random Padding to the messages, but at a cost of over 300% overhead, which translates to at least a terabyte of extra data per day for the service providers. Overall, these attacks could impact over a billion users across the globe and the high level of accuracy that we demonstrate in our experiments means that they represent realistic threats to privacy, particularly given recent revelations about widespread metadata collection by government agencies."

2.2 Real-time Classification for Encrypted Traffic

"Survey of Classifying network flows by their application type is the backbone of many crucial network monitoring and controlling tasks. Basic network management functions such as billing, quality of service, network equipment optimization, security and trend analyzers, are all based on the ability to accurately classify network traffic into the right corresponding application. Historically, one of the most common forms of traffic classification has been the port-based classification, which makes use of the port numbers employed by the application on the transport layer. However, many modern applications use dynamic ports negotiation making port-based classification ineffective with accuracy ranges between 30% and 70%. The next step in the evolution of classification techniques was Deep Packet Inspection (DPI) or payload-based classification. DPI requires the inspection ofthe packets' payload. The classifier extracts the application payload from the TCP/UDP packet and searches for a signature that can identify the flow type. Signatures usually include a sequence of bytes/strings and offsets that are unique to the application and characterize it. DPI is widely used by today's traffic classifier vendors. It is very accurate but suffers from a number of drawbacks. Recently, we have witnessed a dramatic growth in the variety of network applications. Some of these applications are transmitted in an encrypted manner, posing a great challenge to the DPI paradigm. Such applications may choose that encryption both for security and to avoid detection. Common P2P applications such as BitTorrent and eMule have recently added encryption capabilities (primarily to avoid detection). As a significant share of the total bandwidth is occupied by P2P applications and since current DPI based classifiers must see the packet's payload, encryption may become a real threat for ISP's in the near future. The inability of port-based and payload-based analysis to deal with the wide range of new applications and techniques used in order to avoid detection has motivated the study of other classification techniques. Two examples include behavior based classification and

classification based on a combination of learning theory and statistics."

2.3 ITCM: Real Time Internet Traffic Classifier Monitor "The Internet traffic is changing continuously and this contribute to difficult the characterization of network behavior and structure. Massive games and cloud and grid services increase every day their percentage participation in total network traffic. Traffic monitoring Systems generally make use of flow information. Examples are Net Flow or IETF IPFIX, which defines a standard to exporting flow information by routers and switches. Such systems are widely used in network service providers and corporations to gain knowledge about critical business Applications, analyze communication patterns prevalent in traffic, collect data for account, or detect anomalous traffic patterns. A vital issue for corporations and ISPs (Internet Service Providers) is to identify traffic application types which are transmitted on their networks. Pattern recognition and machine learning models have given significant attention to semi supervised learning. In network traffic areas, encryption and processing restrictions, protocol obfuscation and use of ephemeral ports make the task of construct classification models difficult. The large amount of Internet traffic flowing through networks makes the use of approaches that combine labeled and unlabeled data to construct accurate classifiers suitable. There are a large number of papers in the traffic monitoring and traffic classification area. Most papers usually focus on either traffic flow reassembly or traffic classification and identification, But not on their combination. This paper describes the architecture of a real time Internet traffic classifier monitor for use in corporate networks. It also evaluates different machine learning methods for network traffic classification. The classifier monitor is based on concept of bidirectional flow. This means that the fundamental object to be classified in a determined pattern is the traffic flow, either complete or as sub flow. A flow is defined by one or more packets between a host pair with the same quintuple: source and destination IP address, source and destination ports and protocol type (ICMP, TCP, UDP)."

2.4 On different ways to classify Internet traffic: a short review of selected publications "Internet traffic classification or identification is the act of matching IP Packets to the application that generated them. Traffic classification is important for managing computer networks: for example, it is used for traffic shaping, policy routing, and packet filtering. From business point of view, it provides valuable marketing information via customer profiling, whereas scientific and government agencies employ it to identify global Internet trends. Given just a single IP packet it is difficult to classify it there is no application name in the protocol headers. In the past, the service port number was used for discriminating the traffic class, but this became ineffective in the early 2000s due to peer-to-peer (P2P) traffic. Another popular and de facto standard classification method is Deep Packet Inspection (DPI): pattern matching on full packet contents. Despite being accurate, it is computationally expensive and brings privacy concerns. Moreover, traffic encryption makes DPI increasingly irrelevant. Instead, novel classifiers investigate groups of packets in order to find distinguishing features of entire application protocols. Usually, a flow of packets is statistically summarized (e.g. by average packet size and inter packet arrival time) and the resultant feature vector is classified using Machine Learning (ML) (e.g. Neural Network or Support Vector Machine). Such methods

are largely resistant to misuse of the port number and to encryption: the overall behavior of a particular protocol or host is examined instead of seeking for a strict match in a single packet. Latest methods tackle the problem of classification from many perspectives: counting packets, analyzing the DNS context, adopting multi classification, and more. Our "Multilevel Traffic Classification" develops an algorithm that combines different methods to increase classification completeness and accuracy. The aim of this work is to discuss diversity in classification methods. This survey also share our findings on the quality of traffic classification papers. For the review, we selected publications that: (a) Present differentiated methods, (b) Were published recently (2009-2012), and (c) Are interesting in our opinion. Comparing with existing surveys namely and our paper focuses on different time span. We review newer works that were not mentioned in these studies: they represent novel developments in traffic classification."

3. PROPOSED

In this paper, we aim at developing data mining solutions for classifying encrypted Internet traffic data generated by messaging Apps into different service usage types. If properly analyzed, the patterns could be a source of rich intelligence for classifying service usages. Furthermore, from the security and privacy perspective, the underlying issue we leverage is that current privacy protection technology conceal the content of a packet, while they do not prevent the detection of networks packets patterns that instead may reveal some sensitive information about the user's preference and behaviorAlong this line, in this paper, we propose a method to classify in-App service usages using encrypted Internet traffic data by jointly modeling behavioral structure, flow characteristics and temporal dependencies. Specifically, we first segment Internet traffic from traffic-flow to sessions to dialogs by combining hierarchical clustering as well as thresholding heuristics. Besides, we extract the discriminative features of these segmented dialogs from two perspectives: (1) packet length and (2) time delay. In addition, we learn a service usage predictor by feeding the extracted features and the reported usage types into the chosen classifiers. Moreover, for those outlier dialogs with mixed usages, we exploit a clustering method to further segment these dialogs into sub-dialogs. Furthermore, we develop a system, named CUMMA, for classifying service usages in mobile messaging Apps using the proposed method. It has been incorporated into the SmartCare service of a company for the purpose of enhancing end-user experiences.

4 CONCLUSION

In this survey, we will conclude a system for classifying service usages using encrypted Internet traffic in mobile messaging Apps by jointly modeling behavior structure, network traffic characteristics, and temporal dependencies. There are four modules in our system including traffic segmentation, traffic feature extraction, service usage prediction, and outlier detection and handling. Specifically, we first built a data collection platform to collect the traffic-flows of in App usages and the corresponding usage types reported by mobile users. We then hierarchically segment these traffic from traffic-flows to sessions to dialogs where each is assumed to be of individual usage or mixed usages. Also, we extracted the packet length related features and the time delay related features from traffic-flows to prepare the

training data. In addition, we learned service usage classifiers to classify these segmented dialogs. Moreover, we detected the anomalous dialogs with mixed usages and segmented these mixed dialogs into multiple subdialogs of single type usage. Finally, the experimental results on real world WeChat and WhatsApp traffic data demonstrate the performances of the proposed method. With this survey, we showed that the valuable applications for in-App usage analytics can be enabled to score quality of experiences, profile user behaviors and enhance customer care.

REFERENCES

1] M. Kumar, J. Meena, R. Singh and M. Vardhan, "Data outsourcing: A threat to confidentiality, integrity, and availability," Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on, Noida, 2015, pp. 1496-1501.

2] B. K. Samanthula, Y. Elmehdwi and W. Jiang, "kNearest Neighbor Classification over Semantically Secure Encrypted Relational Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1261-1273, May 1 2015. 3] S. Thurner, M. Grün, S. Schmitt and H. Baier, "Improving the Detection of Encrypted Data on Storage Devices," IT Security Incident Management & IT Forensics (IMF), 2015 Ninth International Conference on, Magdeburg, 2015, pp. 26-39.

4] Y. Fu; H. Xiong; X. Lu; J. Yang; C. Chen, "Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps," in IEEE Transactions on Mobile computing, vol.PP, no.99, pp.1-1

5] M. D. Singh, P. R. Krishna and A. Saxena, "A privacy preserving Jaccard similarity function for mining encrypted data," TENCON 2009 - 2009 IEEE Region 10 Conference, Singapore, 2009, pp. 1-4.

6] F. Liu, W. K. Ng and W. Zhang, "Encrypted SVM for Outsourced Data Mining," Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on, New York City, NY, 2015, pp. 1085-1092.

7] Y. Rahulamathavan, R. C. W. Phan, S. Veluru, K. Cumanan and M. Rajarajan, "Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud," in IEEE Transactions on Dependable and Secure Computing, vol. 11, no. 5, pp. 467-479, Sept.-Oct. 2014.

8] J. Xu, W. Zhang, C. Yang, J. Xu and N. Yu, "TwoStep-Ranking Secure Multi-Keyword Search over Encrypted Cloud Data," Cloud and Service Computing (CSC), 2012 International Conference on, Shanghai, 2012, pp. 124-130.

9] H. Hu, J. Xu, C. Ren and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," Data Engineering (ICDE), 2011 IEEE 27th International Conference on, Hannover, 2011, pp. 601-612.

10] Y. Huang, J. Katz and D. Evans, "Quid-Pro-Quotocols: Strengthening Semi-honest Protocols with Dual Execution," Security and Privacy (SP), 2012 IEEE Symposium on, San Francisco, CA, 2012, pp. 272-284.

AUTHOR'S PROFILE:

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

GULLA MANASA PG Scholar, Department of CSE, Vaagdevi College of Engineering, Autonomous, Bollikunta, Warangal Telangana, Mail id:gmanasa39@gmail.com

EMMADI GOUTHAM,

Assistant Professor Department of CSE,Vaagdevi College of Engineering, Autonomous Bollikunta, Warangal Telangana, Mail id: goutham.emmadi@gmail.com