# Wireless User Identification with Traffic Signatures

Jiayang Liu Xiaozhu Lin

## ABSTRACT

Recently the privacy issue of wireless network receives more and more attention, especially for user identification and location tracking. In this project, we propose to study user identification by analyzing the packet size and timing from aggregated traffic sniffed in a wireless LAN. Although there are many works on user identification from the semantic information leaked in packets or from physical layer signatures, this work is the first effort to discover the relation between the user identity and the traffic characteristics deriving from only packet size and timing. We conduct a preliminary spectrum analysis on a small set of wireless traffic trace. The results show the possibility to extract distinct features of different users' traffic using very simple algorithm. We plan to use machine learning algorithms to 1) identify a single user's traffic and 2) further classify aggregated traffic containing multiple users' traffic. We hope our work will answer the question that whether users can be identified from their traffic characteristics and to what extent they can be identified.

## 1. INTRODUCTION

As wireless networks grow prosperously, ubiquitous computing has become an integral part of more and more people's daily life. 802.11 is one of the most popular wireless technologies for anytime anywhere network connection. At the same time, however, the open nature of 802.11 networks has raised lots of concerns on wireless users' privacy. In 802.11 networks, a third party can easily capture the packets of other wireless users in the radio range. Obviously, the wireless users' privacy is in danger if the packets are in clear text. A third party can infer what applications they are using and what websites they are visiting by monitoring the IP addresses and port numbers. Existing encryption methods can hide the data payload of 802.11 MAC packets so that the communication contents won't be exposed. But even when the data payload is encrypted, one can still identify a user from implicit traffic fingerprints [3]. J. Pang later proposed an identifier-free link layer protocol [5] to further encrypt those fingerprints, including MAC address and other MAC header fields. Then the only information can be monitored by a third party is the packet size and arrival time if the most strict encryption method is used. But does that mean we are safe from software sniffing now?

The answer is not obvious. Although numerous work has been done for identifying applications from traffic characteristics [8][9], it is still open whether one can identify a user from only packet size and arrival time. A wireless user may have unique patterns in packet size and timing for mainly three reasons. Firstly, each user may have a unique set of frequently used applications. Since different applications have different traffic characteristics, a unique combination of different applications can generate distinct traffic pattern. For example, a user who often watches online video must have a very different pattern from one whose dominant application is browsing website.

Secondly, even using the same application, different people tend to have different usage habits. Take web browsing, the most popular network application, as an example. One user may like to open a page at a time and wait for the loading to finish; another user may usually open multiple pages at once and send many requests before the previous one get responded. And someone may mostly visit news websites in US with text, images and videos; someone may frequently visit forum websites oversea with mainly text and images; someone may spend a lot of time on watching YouTube. Another example is that many people use Outlook to manage Email account. They may have different intervals in checking new messages, which will generate periodical request packets with different intervals.

Thirdly, the software and hardware configurations of the computer can also cause variations in traffic characteristics. The operating system may have different response time to process a new packet; the network interface driver may have different setting in the maximal packet size; the TCP protocol on the computer may be implemented with different congestion control methods. Theses variations will all contribute to the unique pattern in packet size and timing.

In the proposed work, we plan to analyze the traffic characteristics of different users, design a classification algorithm to identify a user form only packet size and timing, and evaluate it on a large data set. Our classification algorithm will be based on machine learning. We assume we can get a target user's traffic to train the algorithm by monitoring the medium when only this user is in the network or when the MAC address is not encrypted. Then our first step is to identify this user given a single unknown user's traffic; the second step is to tell whether the user's traffic is inside an aggregated traffic of multiple users, i.e. to track the user's location by deciding if he is inside the wireless network.

Our work is the first to reveal whether we can and how to identify a user from only packet size and arrival time. The answer will help us to understand the privacy protection in today's wireless networks. If we can find a simple algorithm to reliably identify a user, we basically point out a fundamental problem for wireless networks, since packet size and arrival time cannot be eliminated form explosion by any encryption method. To protect against such privacy violation, one can only distort the traffic pattern by inserting dummy bytes into a packet to change its size or hold outgoing packet for extra time to change its timing. Either way will result in considerable communication overhead. Therefore, once demonstrated, the problem will not be easily solved. We hope our work will bring more attention to this problem.

## 2. RELATED WORK

In recent years, researchers are interested in the identification of users within wireless context, even when wireless packets are

**Table 1: Feature Size of Four Users**

| User | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Feature Size | 1099 | 218 | 154 | 309 |

partially or fully encrypted. For instance, 802.11 null data frames can be used to identify wireless devices, since different vendors apply 802.11 null data frames in different applications [1]. [2] observes that low layers of wireless networking protocol stack will leak some information in plain text about end hosts. [3] proposes that in wireless LAN, privacy threats could come from service discovery procedure, like unencrypted announcements by infrastructures, or probe activities from end systems. It also argues, by tracking features from such sources over a certain period, specific users can be "fingerprinted". The authors present a system architecture called *Tryst* to minimize sensitive information exposure in 802.11 service discovery. [4] proposes a technique called *stack virtualization* to provide random identifiers for every layer of protocol stack. In [5], the authors propose a novel wireless link layer protocol to eliminate the uniqueness of MAC addresses, as well as to obfuscate every bit in 802.11 packets. Such kinds of methods, if feasible, will fail any fingerprinting techniques based on semantics in plain text. However, even with the most strict encryption method or protocol, the packet size and timing can still be obtained by a third party as long as they talk over a common PHY layer. Then the ability to identify a user through packet size and timing becomes a fundamental issue to understand privacy protection in wireless network. That is right the goal of this work.

Besides software-based sniffing, some researchers indicate that even though networking layer data is obscured, similar goal can be achieved via physical layer characteristics caused by hardware variations. [6] fingerprints wireless device through radiometric signatures coming from hardware imperfections that are unique for every piece of wireless interfaces. Therefore, by identifying their wireless interfaces, users can be tracked with a high probability. [7] proposes to discriminate Access Points according to their clock imperfects. However, monitoring physical characteristics is much more different than software-based sniffing and requires special equipments. Therefore, software-based sniffing is a more severe problem in wireless network.

# 3. INITIAL ANALYSIS

We analyzed a small dataset of wireless network traces collected at Rice. The result shows it is highly possible to identify a user from traffic pattern. It also provides us valuable information on how to design the classification algorithm.

## 3.1 Network Traces

Since at this moment we are trying to find some unique patterns in wireless traffic for a single user, the traces we use were collected from the laptops of 4 users, other than monitoring the aggregated traffic of a wireless network. The traces record the traffic of each user at various locations, i.e. office, home, for one week. We parsed the traces and only kept the packet size and arrival time information.

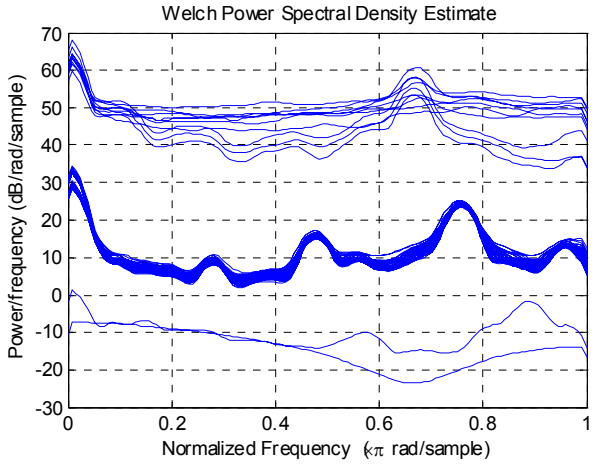


**Figure 1: Frequency distribution of packets in feature size for user 1**

## 3.2 MatLab Analysis

We observe that many applications generate packets with fixed sizes, such as http request packets. The distribution of packets of same size in time scale reflects the usage of certain application. We also find that there are many periodic traffic flows inside every user's traces, such as Outlook periodic synchronization. Therefore, our initial analysis is based on the periodic patterns of packets with the same size, realized by frequency analysis.

Our analysis for each user is in five steps: 1) find the set of frequently occurred packet sizes, noted as S; 2) in each trace file of the user, extract the arrival time of the packets with the same size for each value in S; 3) analyze the arrival time in frequency domain and get a vector of the frequency distribution per trace file; 4) calculate the distance between the frequency vectors of every two trace files; 5) find out the packet size that generates the minimum vector distance on average, i.e. the packets having the most similar distribution in frequency domain.

Therefore, we obtain the packet size with the most stable patterns across all trace files for each user. We call it the feature size. We find every user has a different feature size as shown in **Error! Reference source not found.**. For example, the frequency distribution of User 1's feature size is shown in Figure 1. Each curve represents a trace file. We can see that most trace files exhibit very close frequency distribution.

# 4. REFERENCES


[1] Gu, W., et al., On Security Vulnerabilities of Null Data Frames in IEEE 802.11 Based WLANs, in *International Conference on Distributed Computing Systems*, 2008.

[2] Aura, T., et al., Chattering Laptops.

[3] Pang, J., et al., 802.11 user fingerprinting, in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking,* 2007.

[4] Lindqvist, J. and J. Tapio, Protecting Privacy with Protocol Stack Virtualization, in *WPES*, 2008.

[5] Greenstein, B., et al., Improving wireless privacy with an identifier-free link layer protocol, in *Proceeding of the 6th*

*international conference on Mobile systems, applications, and services*, 2008.

[6] Brik, V., et al., Wireless device identification with radiometric signatures, in *Proceedings of the 14th ACM international conference on Mobile computing and networking*, 2008.

[7] Jana, S. and S.K. Kasera, On fast and accurate detection of unauthorized wireless access points using clock skews, in *Proceedings of the 14th ACM international conference on Mobile computing and networking*, 2008.

[8] Haffner, Pl, et al., ACAS: automated construction of application signatures, in *Proceeding of the 2005 ACM SIGCOMM workshop on Mining network data*, 2005.

[9] Moore, A. and Zuev, D., Internet traffic classification using bayesian analysis techniques, ACM SIGMETRICS Performance Evaluation Review, v.33 n.1, June 200