Poster: The Privacy Leaky Bucket - Your Mobile Devices

Jie Chu
The Graduate Center, CUNY
365 Fifth Avenue,
New York, NY, USA 10016
jchu1@gradcenter.cuny.edu

Ping Ji
John Jay College, CUNY
524 West 59th. Street
New York, NY, 10019
pji@jjay.cuny.edu

ABSTRACT

Booming wireless technology makes personal mobile device a norm (smartphone, tablet, etc) and the wireless communication becomes fundamental in daily life. The goal of this paper is to assess how much personal information is leaked (or can be extrapolated) from a systematically collected open-in-air message exchanges. We monitor the wireless traffic and we apply data mining technique to answer the question of how much personal information can be observed and inferred through open wireless medium.

Keywords

Privacy; Mobile Network; Wireless monitoring

1. INTRODUCTION

Booming wireless technology makes personal mobile device a norm (smartphone, tablet, etc). Nowadays, people spend 60% of their online time on mobile devices[1]. The bandwidth of cellular network had been increasing, but the cost of building denser cellular network makes it hard to keep up the same pace with the rapid growth of mobile data traffic, and thus ubiquitous Wi-Fi is expected to offload most of mobile data traffic.[2] As the wireless communication becoming fundamental in daily life, the characteristics of wireless communication, i.e. shared media, is a double-sided coin. It makes the wireless communication so convenient but also leads to the issue that communication channel can be easily monitored. The content of communication is usually protected by encryption on different layers. 802.11 provides encryption on link layer and many websites utilize transportation or application layer encryption to protect the traffic. There are many researches on the security of encryption and protocol design. However, even without breach the encryption, due to the constraints of communication protocol, there are still many information publicly available. For example, in order to make the communication possible, the two parties have to disclose their identities, MAC address in 802.11, as well as the timestamp. Anyone who obtain

an off-shelf wireless network card could observe such public information.

The goal of this paper is to assess how much personal information is leaked (or can be extrapolated) from a systematically collected open-in-air message exchanges. In particular, we want to answer the following questions: 1. *How much association relationship, which could reflect share of owner, between devices can be identified?* 2. *Can we predict a device's behavior based on its history?* 3. *Given a piece of new data, can we identify the device sharing the same pattern? What if the device ID can be spoofed?*

Our methodology is to set up wireless monitor point in a crowded metropolitan area. We monitor the open-in-air mobile traffic keeping only the public information including MAC address, timestamp, location, SSID etc. Then we apply time series model for correlation across devices and prediction on each individual devices. Our initial results show interesting findings from the dataset.

2. DATA COLLECTION

In order to demonstrate the characteristics of wireless network among different locations, we setup laptop as monitor nodes, of which the wireless card are set to run in monitor mode to record 802.11 traffic, at several locations. In our experiment, we use Macbook Pro running Mac OS X 10.6, with wireshark and kismet installed to passively capture the 802.11 raw traffic. We put up a monitor node to monitor a single channel which is used by the closest wireless access point(AP) for three months on a school campus. This monitor node remains static during the entire session. Although there are many channels need to be observed, the single channel that the nearby wireless AP operates on should cover a great number of devices with continuous traffic which would suit our need for studying the characteristic of client's traffic. In addition, many 802.11 clients would actively send probe request on every channel to look for a 802.11 hotspot to join in, thus would potentially increase the number of clients that can be observed by our monitor nodes.

We processed raw 802.11 packets such that for each packet we only keep meta data and some key information on link layer, such as source and destination MAC address, timestamp, received signal strength, channel, packet length, SSID (if applicable), etc. We did not attempt to extract any information beyond layer 2 under the assumption that adversary can not decrypt packet transmitted via secure 802.11 connection. And thus we don't keep any of the user data.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MobiHoc'16 July 04-08, 2016, Paderborn, Germany

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4184-4/16/07.

DOI: <http://dx.doi.org/10.1145/2942358.2942413>

Figure 1: Daily number of unique MAC address

Figure 2: Activities of two highly correlated devices

Figure 3: Results of ARIMA model

3. PRELIMINARY ANALYSIS

First we analyzed the first three weeks data to get a basic understanding on the dataset. Fig 1 shows the number of unique MAC address observed on each day of the three weeks. We can clearly observe the seasonality which matches the schedule of the school. On Monday through Thursday, when the classes are usually scheduled, the number of unique devices ranges from 5.5k to 8k, while there is a dramatic drop to less than 1k over the weekend. We can also notice that the beginning of this dataset, the number is also very small which is due to campus close on a holiday. However, the number of unique MAC address that had transmitted beacon frame on each day, which is an approximate to the number of APs, is relatively stable without the seasonality.

Since there is strong correlation between the number of devices and the school schedule in the campus dataset, we would like to further explore some patterns. By examining the overlapping devices over each day of week, we are able to identify three typical schedule patterns on campus, one will come to school campus every day, another group will come to school only on two days that they have classes, i.e. either Monday and Wednesday or Tuesday and Thursday, and the last group will occasionally come to school campus in addition to their class days. Each of these groups have a large number of devices even our observation only covers a small area, the next question is how much we can further distinguish between the devices with similar schedule.

4. MODELLING

After we preprocess the raw 802.11 packets, we further aggregate the data onto each individual devices with time window size set as l seconds, we can obtain multiple time series for all the attributes associated with a device. For example, we can count how many packets had been transmitted by the device within a time slot, the median of signal strength, etc. For further quantitative analysis, we formulate the dataset as follows: for each individual device D_i , it can be represented with multiple (possibly) dependent time series $\mathbf{X} = \{\mathbf{X}_t : t \in [0, T]\}$, where $\mathbf{X}_t = (X_{t1}, \dots, X_{tm})^T$ is a $m \times 1$ vector consists of observations of m attributes associated with this particular device in time slot t .

To explore the correlation on different devices and possibly their owners, we propose a two round approach. First we use a very large window size to group devices into sets that within each set the devices have identical activities on a coarse scale. We set l to one day and thus require devices

i and j belong to the same group if $X_i = X_j$, where we set $X_{ik} = 0$ if device i is not active on the k -th day, set $X_{ik} = 1$ if it is. Then we reduce the window size to 5 minutes, similarly we set $X_{ik} = 1$ if device i is active at slot k , otherwise as 0. Then we compute the Jaccard index defined as $f(X_i, X_j) = \frac{|X_i \wedge X_j|}{|X_i \vee X_j|} = \frac{\sum_m X_{im} \wedge X_{jm}}{\sum_m X_{im} \vee X_{jm}}$. We are able to identify the highly correlated devices as shown in Fig 2. This metric can be used to compute the self-similarity of each device, which is useful to measure the distance of two devices when we seek to match patterns in newly captured data with existing device.

We also would like to explore how much information can be extracted from historical data to predict future activities. Given the large amount of devices that we observed only have very limited history, we will focus on the most active devices with enough history which may exhibit patterns to capture. We propose to use ARMA model, which has been extensively used in modeling stationary time series process. It can be expressed as $Z_t = c + \varepsilon_t + \sum_{i=1}^p \phi_i Z_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$, where Z_t is the value of time series at time t , ε_t is the white noise error term at time t , both ϕ_i and θ_i are parameters of the model need to be estimated, p and q are the order of the ARMA model. In order to automatically select ARMA model parameter for each individual devices, we leverage Akaike information criterion (AIC) to measure the quality of ARMA model with varying parameters, and we could limit the range of p and q to a small range with analysis from sample data. AIC for ARMA model with parameter (p, q) is defined as $AIC(p, q) = \frac{2(p+q)}{T} + 2\ln(\hat{\sigma}^2)$, where $\hat{\sigma}^2$ denotes estimation of σ^2 from ARMA(p, q) which is a measure of goodness of fit. The result of fitted model for a client is shown in Fig 3.

5. CONCLUSION

In this poster draft, we describe our data collection process and modelling work on exploring the time series data extracted from 802.11 link layer. We demonstrate that with encrypted wireless traffic, it would still exhibit many patterns which could be used to reveal the identity or relationship of the mobile devices' owners.

6. REFERENCES

- [1] Major Mobile Milestones in May: Apps Now Drive Half of All Time Spent on Digital.
- [2] J. Schofield. Most mobile data will soon be offloaded to Wi-Fi networks, says Juniper Research.