Demographics Inference Through Wi-Fi Network Traffic Analysis

Huaxin Li
Shanghai Jiao Tong University
Shanghai, China
Email: lihuaxin003@sjtu.edu.cn

Zheyu Xu
Shanghai Jiao Tong University
Shanghai, China
Email: lanesra@sjtu.edu.cn

Haojin Zhu
Shanghai Jiao Tong University
Shanghai, China
Email: zhu-hj@sjtu.edu.cn

Di Ma
University of Michigan-Dearborn
Dearborn, USA
Email: dmadma@umich.edu

Shuai Li
University of Minnesota
Twin Cities, USA
Email: shuai@cs.umn.edu

Kai Xing
University of Science and
Technology of China, P.R. China
Email: kxing@ustc.edu.cn

***Abstract*—Although privacy leaking through *content* analysis of Wi-Fi traffic has received an increased attention, privacy inference through *meta-data* (e.g. IP, Host) analysis of Wi-Fi traffic represents a potentially more serious threat to user privacy. Firstly, it represents a more efficient and scalable approach to infer users' sensitive information without checking the content of Wi-Fi traffic. Secondly, meta-data based demographics inference can work on both unencrypted and encrypted traffic (e.g., HTTPS traffic). In this study, we present a novel approach to infer user demographic information by exploiting the meta-data of Wi-Fi traffic. We develop a proof-of-concept prototype, Demographic Information Predictor (DIP) system, and evaluate its performance on a real-world dataset, which includes the Wi-Fi access of 28,158 users in 5 months. DIP extracts four kinds of features from real-world Wi-Fi traffic and proposes a novel machine learning based inference technique to predict user demographics. Our analytical results show that, for unencrypted traffic, DIP can predict gender and education level of users with an accuracy of 78% and 74% respectively. It is surprising to show that, even for HTTPS traffic, user demographics can still be predicted at a precision of 67% and 72% respectively, which well demonstrates the practicality of the proposed privacy inference scheme.**

I. Introduction

The wide deployment of public wireless access points and the prevalence of portable mobile devices allow people to have ubiquitous wireless access to the Internet. According to research by iPass, it is estimated that by 2018 there will be over 340 million public Wi-Fi hotspots globally. It is also expected that the number of Wi-Fi-enabled devices will grow to more than 7 billion by 2017 [1]. Compared with 3G/4G services, Wi-Fi access is one user-preferred connectivity option when using popular applications, such as Skype, Netflix or Facebook due to its superiority of cost and connectivity.

While public Wi-Fi provides convenience and free access, it may potentially pose a serious threat to the privacy of mobile users by leaving their computer and other electronic devices open to hacking. Even though there exist a series of security solutions which provide link-to-link security (e.g., WPA2-AES) and end-to-end encryption (e.g., HTTPS), mobile users are still facing a big security challenge due to the lack of security protection, inappropriate implementation of security protocols, and untrusted/fake hotspot service providers. Existing research reveals potential privacy leakage in public hotspots by examining user end activities such as web browsing, search engine querying and smartphone apps' usage [2]. Most of existing studies are based on the assumption of unencrypted traffic or a full knowledge of user behaviors and they cannot work in the case of incomplete information [2]–[4].

In this study, we raise the following question: *can an attacker infer the sensitive information (e.g., gender, age, education) of targeted users by observing the meta-data of Wi-Fi traffic (e.g. IP, Host) ?* The answer to this question is not straightforward. Firstly, mobile users usually stay at hotspots for short durations and thus public Wi-Fi traffic represents a partial view of its full traffic. This problem is more challenging in the case that a certain percent of websites utilizing HTTPS protocol to encrypt the browsing traffic, which prevents any external observer from accessing the traffic contents. According to a recent report in 2015, HTTPS traffic reaches 46% for browser traffic (increased 7% in 12 months) and 61% for app traffic (increased 9%) [5]. Due to these reasons, by eavesdropping and analyzing the content of Wi-Fi traffic, less than 10% mobile users have their gender information leaked out [2].

In this study, we answer the question above by studying how to infer user demographic information from the meta-data of Wi-Fi network traffic. The proposed approach is motivated based on the observation that even for the encrypted traffic, it is still possible for the eavesdropper to obtain the meta-data of Wi-Fi traffic, which leaves a new attack interface for the insider attackers (e.g., fake/untrusted service providers) and external attackers (e.g., external hackers who break the password). Our insight is that users sharing the similar attributes will have similar network characteristics. To achieve this, we extract four kinds of attributes which can create distinct signatures for different demographics. Then, we propose a novel Random Forest based demographic information inference scheme and develop a proof-of-concept prototype system named Demographics Information Predictor (DIP). Our

978-1-4673-9953-1/16/$31.00 ©2016 IEEE

study is based on a large real-world dataset which involves 98 Wi-Fi access spots and 28,158 users. The contributions of this work can be summarized as follows:

- Based on the large-scale real world dataset, we demonstrate that the network traffic originated from users with different demographics has distinct signatures. We extract four kinds of features which can create distinct signatures for different demographics.
- We propose a Demographics Information Predictor (DIP) system that can learn users' demographics by just passively monitoring users' traffic flows. DIP employs a novel Random Forest based prediction technique to predict users' demographics. We evaluate DIP using our real-world Wi-Fi traffic dataset and show that DIP can predict gender and education level with an accuracy of 78% and 74%, respectively.
- We also measure to what extent the demographics will be leaked through encrypted network traffic, such as HTTPS traffic. We consider the lower bound of information leakage, i.e., assuming all HTTP traffic are encrypted as HTTPS traffic. Surprisingly, users' demographics can still be predicted at precision of 67% and 72%.

To the best of our knowledge, this is the first work that uses the large scale real-world data set to measure the privacy violations in the meta-data of the Wi-Fi network traffic. This study aims to call for the attention of the society and shedding light on the measures of protecting Wi-Fi traffic.

The rest of paper is organized as follows. Section II discusses related research works. Section III introduces real-world traffic leakage, research motivations, a traffic privacy model, and our real-world dataset. Section IV presents our DIP system and describes its functions. Section V discusses feature selection and prediction model in detail. In section VI, we evaluate the privacy leakage in different scenarios. Section VII discusses limitations and mitigation suggestions and Section VIII concludes this paper.

II. Related Works

This paper is to understand the level of user privacy leakage through meta-data analysis of Wi-Fi traffic. The presented work is related to the following areas of research.

Network Traffic Privacy. Privacy leakage in network traffic is receiving increasing attention. Cheng et al. [2] captured Wi-Fi network traffic at 20 airport hotspots in four different countries. Their analysis reveals that two thirds of travelers leak privacy sensitive data by DNS queries, web browsing, or querying search engine. Das et al. [3] present PCAL (Privacy-Aware Contextual Localizer) which can learn users' contextual locations (such as residence and cafe) just by passively monitoring user's network traffic. Xia et al. [4] focus on association between the browsing traces and OSN's ID of a user. They present a framework to correlate the user identity extracted from the social network traffic to its online behavior. Konings et al. [6] collected the mDNS announcements in a semi-public Wi-Fi network at a university. Their study shows that, of 2,957 unique device names, 59% contained both real names of users, with 17.6% containing first and last name of the user. Yan et al. [7] propose a novel privacy-preserving scheme against traffic analysis in network coding. Even though network traffic is encrypted, privacy violation is still possible [8], [9]. Different from previous works, our work takes a new approach to infer user demographic information by exploiting the meta-data of Wi-Fi traffic.

Demographics Inference. Inference on demographic information has been discussed using various signatures. Hu et al. [10] extract content-based features and category-based features from webpage click-through logs to infer users' gender and age. Seneviratne et al. [11] employ Naive Bayes model and Support Vector Machine to reveal users' gender from their installed apps. Schwartz et al. [12] apply differential language analysis *Facebook* on status update messages to predict user demographics. Bi et al [13] show how user demographic traits such as age and gender, and even political and religious views can be efficiently and accurately inferred based on their search query histories using a model trained from *Facebook* likes. Chaabane et al. [14] infer OSN users' undisclosed (private) attributes (e.g. gender, relationship, age and country) by using public attributes (e.g. hobby) of other users who share similar interests. Different from previous works, our work selects Wi-Fi traffic meta-data which can be sniffed passively as features to infer demographics leakage.

III. Motivations and Problem Formulations

A. Traffic Leakage in Real World

Previous research demonstrates the insecurity of public Wi-Fi, which may potentially leak user privacy information from Wi-Fi traffic. In the following, we summarize various cases which lead to different Wi-Fi traffic leakage.

1) Public Open Wi-Fi or Rogue Hotspots: Although there are existing Wi-Fi security solutions such as 802.11i proposed in 2004, open public Wi-Fi networks without any protection are still popular due to the free and simple wireless connections. In fact, a typical selling-point of many restaurant chains nowadays is that they offer free Wi-Fi connections to customers. In an open public Wi-Fi environment, wireless connections are vulnerable to Man-in-the-middle (MITM) Attack, which allows the attacker to tap into wireless channels and obtain the Wi-Fi traffic.

Unauthorized 'rogue' hotspots allowing back-door access to the network, and honeypot APs used in attacks that lure end users to connect to unsecured external networks, represent two other kinds of threats to Wi-Fi traffic. Rogue Wi-Fi Containment is not an easy job in practice due to the great difficulty of accurate Rogue Wi-Fi detection. There has been significant interest in the industry on improving the security against rogue hotspot, e.g., using certificates to authenticate the Wi-Fi [15]. However, it is far from being widely deployed in practice.

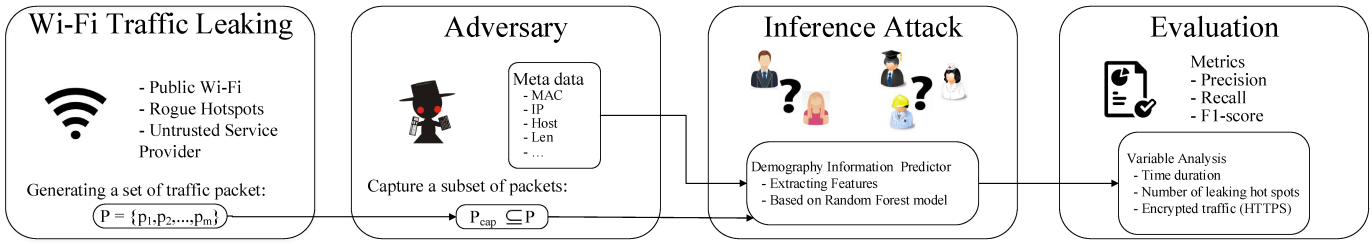


Fig. 1: Overview of framework

2) Security Enabled Wi-Fi without Proper Implementation: IEEE 802.11i provides important security features for Wi-Fi. However, without appropriate implementations, security vulnerabilities can still be exploited by the attackers. For example, using a pre-shared key (PSK) can be strong, but using a single passphrase limits security to its weakest link, the human factor. Further, protocol attacks ranging from key discovery to multi-layer Evil Twin impersonation are periodically being discovered [16], [17]. In the case of being hacked by the adversary, the Wi-Fi traffic will also be exposed to the attackers.

3) Untrusted Service provider: For Wi-Fi service provider, an important business model is advertisement. According to a report of Cisco, from mobile advertisement, Wi-Fi service provider is achieving a 24 Cost per mille (CPM) in a mall in Canada and another is commanding 40 CPM for a mall in Singapore. Targeted advertisement is expected to be an important way for improve the CPM while target ad is based on user location and demographic information. Therefore, service providers have the incentive to collect user traffic and infer corresponding sensitive information.

B. HTTPS traffic

Utilizing SSL to encrypt traffic data is regarded as an important approach to enhance the Wi-Fi security. With the popularity of HTTPS protocol, more and more websites employ HTTPS protocol to secure communication between server and client. HTTPS is the result of layering the HTTP on top of the SSL or TLS protocol, thus adding the security capabilities of SSL/TLS to standard HTTP communications. The main goal of HTTPS is to provide authentication of the visited website and to protect the privacy and integrity of exchanged data.

With HTTPS, the content of packets, including the headers, request URL, query parameters, and cookies (which often contain identity information about the user), are successfully masked via encryption, which is shown in the red box of Fig 2. However, HTTPS cannot hide IP addresses, port numbers and some statistics, such as Seq and Len, because they are a part of the underlying TCP/IP protocols. In practice, this means that attackers can still acquire the IP address and port number of the Wi-Fi access point or the web server that one is communicating with, as well as the duration of session and amount of data transferred of the communication, as shown in the green box of Fig 2.

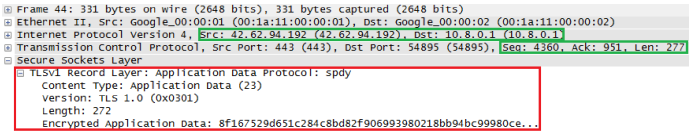


Fig. 2: A demo of HTTPS traffic packet

C. Research Motivations

Previous studies have investigated the privacy leakage problem by having a detailed analysis on Wi-Fi traffic contents [2]. Different from previous works, we propose a demographic inference system which can predict user demographic information through *meta-data analysis* of Wi-Fi traffic. The proposed system is expected to have the following desirable features.

- *More Scalable*: To address the severity of privacy leakage, the proposed system should cope with a large amount of network traffic and predict demographics of a large group of people.
- *Larger Target Coverage*: In previous works which study privacy leaking based on network traffic contents, only a small percent (less than 10%) of users are found that their demographic privacy is leaked [2]. The proposed system exploits meta-data of Wi-Fi traffic to predict user demographic information and is expected to work well in the case of lack of complete information.
- *HTTPS Traffic Tolerance*: HTTPS traffic represents a great challenge for content based traffic analysis since the encrypted Wi-Fi data are immune to the traffic analysis except the meta-data. Therefore, the proposed system is expected to exploit the available meta-data, which cannot be protected by HTTPS protocol, to infer user demographic information.

D. Traffic Privacy Model

In our problem, we consider a set of users $\mathcal{U}$ who generate a series of traffic packet $\mathcal{P}$ within a time duration $\mathcal{T}$. $\mathcal{P}$ consists of a sequence of traffic packets $\mathcal{P} = \{p_1, p_2, ..., p_m\}$, where a traffic packet $p_i \in \mathcal{P}$ contains meta-data fields $\mathcal{F}_i = \{f_1, f_2, ..., f_n\}$ in different layer's protocols, such as "MAC address", "ip address", "Host", "User-agent", "Seq", "Len". We model the traffic privacy profile of a specific user u as a function extracting the meaningful fields $\mathcal{F}$ from traffic packets $\mathcal{P}$, i.e. $\alpha_u : \mathcal{P} \rightarrow \mathcal{F}$.

An attacker, under different attack scenarios, may capture a subset of whole traffic packets, $\mathcal{P}_{cap} \subseteq \mathcal{P}$, from one or more sources of network traffic $\mathcal{L} = \{l_1, l_2, ..., l_q\}$. And fields $\mathcal{F}_{cap} \subseteq \mathcal{F}$ extracted from $\mathcal{P}_{cap}$ will be exposed to the attacker and leak privacy information directly or indirectly, from the perspective of an attacker. Under different conditions, $\mathcal{F}_{cap}$ contains different kinds of contents and different amount of information. For example, if an attacker is interested in MAC address, IP address, host and User-agent in traffic packet, the attacker can observe all of them in a HTTP packet, i.e. $\mathcal{F}_{cap}$ = {MAC, IP, host, User-agent}, while he can't observe host and User-agent in a HTTPS packet because the application level was encrypted in HTTPS protocol. Nevertheless, the attacker can still observe the MAC address and ip address in a HTTPS packet, i.e. $\mathcal{F}_{cap}$={MAC,IP}.

Using $\mathcal{F}_{cap}$, the goal of the attacker is to infer demographics, which is considered privacy leakage issue of the mobile users in this work. Formally, it is a function β translating $f_i \in \mathcal{F}$ into information which can be used to infer demographics information $\mathcal{DI}$: $\beta(f_i) \rightarrow \mathcal{DI}$.

E. A Real-world Traffic Dataset

We obtain network traffic from 98 Wi-Fi hotspots deployed on a university campus. The real-world Wi-Fi traffic dataset contains 28,158 users' network traffic with a duration of 5 months (2014.09-2015.01). It contains more than 12.7 million Wi-Fi connection *sessions*. A *session* here is defined as a continuous time duration in which a user connects to Wi-Fi before the timeout. If timeout for more that 5 minutes, the next connection is considered as a new session.

To preserve privacy of users, we sanitize the traffic data first. We anonymize users' id and remove personal identity related information. After sanitization, each session contains meta-data including connection start time, duration time, ip address, server host and some statistics such as packets size, HTTP flow number and so on. Meanwhile, the dataset also provides anonymized user attributes such as gender and education level, which provides the ground truth to evaluate the performance of the DIP system.

IV. Demographic Information Predictor (DIP)

In this section, we present Demographic Information Predictor (DIP) system, which can extract information from traffic and predict users' demographics based on the meta-data of Wi-Fi traffic. DIP aims to automatically extract information from traffic and generate profile signatures to predict users' profiles. If untrutsted service providers or external adversaries are able to monitor traffic passively, they can exploit DIP to predict user's profile. Our insight is based on the fact that it is highly possible that users having similar demographics have similar network usages [10]. Network access behaviors and mobility characteristics will also share the similar demographic features, which is supported by the previous work on web browsing analysis. The system architecture is shown in Fig. 3.

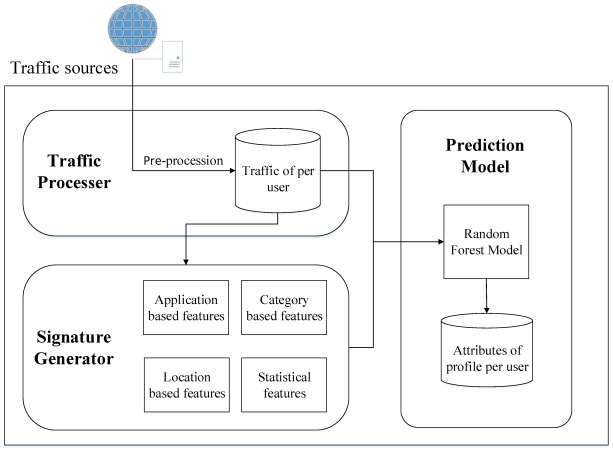


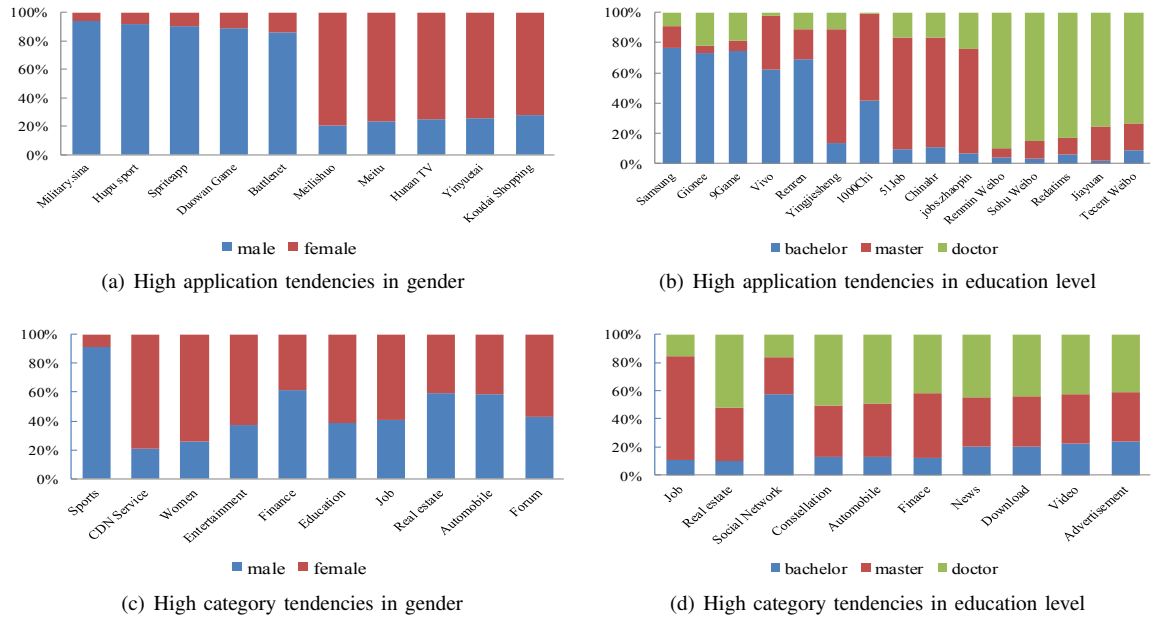
Fig. 3: DIP system architecture

A. Traffic Process Engine

DIP collects information which is used to identify users' profiles. Given a series of traffic as input, DIP parses traffic packets and extracts targeted fields. DIP uses MAC addresses to identify devices and aggregates flows from the same devices or the same ip addresses. Then DIP filters out packets with targeted meta-data such as Host, User-agent and URL in HTTP protocol and preserves the data sequence of the users. For further analysis, DIP also handles some procession including aggregating domains addresses from the same service providers. For example, "a.domain.com" and "b.domain.com" are two addresses from the same application's different servers, we aggregate them according to the text similarity. So DIP can be deployed either in an ISP or a Wi-Fi hotspot to perform the traffic processing in a real time manner.

B. Profile Signature Generator

Profile Signature Generator is used to extract features for predicting users' profiles. With information generated from Traffic Process Engine, we classify features into four categories: application based features, category based features, location based features and statistical features. Application based features are extracted from Host field of HTTP protocol. It usually describes which websites users visited or which applications of smartphone users used. Application based features reflect application usage of users and we further classify applications into different categories, such as communications, news, shopping, etc. Location based features describe where a user gets access to a network. Location based features can be extracted from IP address because IP address of the access point reflects coarse location of a user and is highly correlated to contextual location [3]. Location based features can be viewed as mobility of users. Statistical features describe statistic of traffic flows or traffic packets, such as number of HTTP requests per session, size of a HTTP packets, duration of a session.



(a) High application tendencies in gender

(b) High application tendencies in education level

(c) High category tendencies in gender

(d) High category tendencies in education level

Fig. 4: Application based features and category based features

C. Profile Predictor

Once Profile Signature Generator generates different kinds of features, Profile Predictor uses features to predict demographics of users. The Profile Predictor employs supervised machine learning techniques to learn a machine learning model and predict users demographics. The prediction model of Profile Predictor is based on Random Forest model [18], which runs efficiently on large data bases and can handle thousands of input features without feature selection. In this work, we assume part of users' information is available in public and can be used to train the model. The generated model can be saved for future use in other scenarios.

V. DIP Features and Prediction Model

In this section, we present the core components of DIP in details, which mainly consists of feature selection and prediction model.

A. Features selection

1) Application-based Features: Application-based features are represented as hosts in HTTP protocol in our problem. Application-based features indicate which websites users visited and which apps users ran or services users enjoyed. Since there are a large number of hosts in our dataset and some of the hosts are from the same service providers or organizations, we aggregate them according to hosts similarity, and only select applications used by at least 10% of the users.

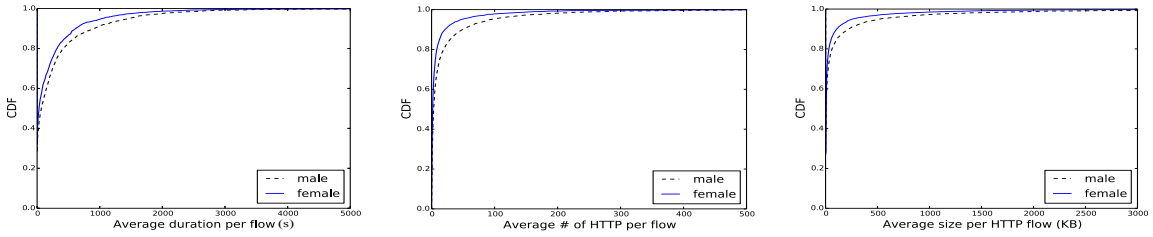
Certain applications show strong tendency towards attributes of demographics, such as gender and education. To characterize the tendency, we calculate the *entropy* of each application with respect to attributes. Let A be a kind of demographic attributes, e.g. gender $A = \{\text{male, female}\}$ or education $A = \{\text{bachelor, master, doctor}\}$. Then entropy ε can be presented as:

$$\varepsilon(A) = -\sum_{a \in A} \theta(a) log_2 \theta(a) \quad (1)$$

where θ is the user distribution of an attribute.

Entropy measures the uncertainty of each attribute. Entropy has the maximum value when the probability of each tendency follows a uniform probability and has the minimum value when the probability of one tendency is dominant. So lower entropy of an application indicates it is distinguishable with respect to an attribute. Since the number of users of each attribute is imbalance, we *under sample* users of each kind of the attributes to keep balance. Fig. 4(a) shows the 5 applications with lowest entropy for male users and female users respectively. In general, *Game* and *Sport* are more popular among male users while *Fashion* and *Shopping* are more popular among male users. Fig. 4(b) shows the 5 applications with the lowest entropy for the users with education level of bachelor, masters and doctors respectively. It can be observed that bachelors are more interested in *Electronic Product*, while *Job* is most popular among masters and *Marriage* among doctors.

2) Category-based Features: We classify applications in our dataset into 39 categories. To evaluate tendency of each category, we calculate *entropy* again. Fig 4(c) shows the 10 categories with lowest entropy for the gender attribute. It shows that *Sports*, *Finance* and *Real Estate* are more popular in male users and *Women*, *Entertainment* are more popular in female users. Similarly, Fig. 4(d) shows the 10 categories with lowest entropy for education level attribute. It shows that *Social Networks*, *Job* and *Finance* are most popular in bachelors, masters and doctors, respectively. The results show



(a) Time duration per flow of Gender (b) HTTP number per flow of Gender (c) HTTP size per flow of Gender

(d) Time duration per flow of Education (e) HTTP number per flow of Education (f) HTTP size per flow of Education

Fig. 5: Cumulative Distribution Functions of statistical features

that the category-based features can also be used to distinguish different groups of users.

3) Location based Features: Different Wi-Fi access points have different IP addresses. We can extract IP addresses from IP layers of traffic packets so we can know where users access Wi-Fi access points. Location is a strong indicator of users' demographics. On one hand, previous work has pointed out that IP addresses of Wi-Fi access points are highly correlated to locations of users [3]. On the other hand, location based features can be viewed as mobility of users, and the mobility is highly correlated to users' profile attributes such as hobbies, habits and relationship [19]–[21]. Thus location based features are supposed to show a strong correlation with demographics.

4) Statistical Features: As we discussed above, users' demographic information shows correlation with network usages, and similar Internet applications exhibit similar characteristics in the aggregated traffic. Implicitly, statistical features reveal distinct information that can distinguish users with different demographic information. For examples, Fig. 5(a) and Fig. 5(d) show different time duration per flow of different groups of users. Fig. 5(a) shows that male users have higher time duration than female users do. It reflects that male users will stay a longer time for each network access. Similarly, Fig. 5(d) shows that master students have the highest average time duration, and bachelor students' time duration is slightly longer than doctor students' time duration. Fig. 5(b) and Fig. 5(e) show average HTTP number per flow of different attributes. As we can see, male users have higher HTTP number per flow than female users. Master students have the maximal average HTTP number per flow, and bachelor students' HTTP number per flow is slightly larger than doctor students' HTTP number per flow. Similarly, Fig. 5(c) shows that male users have higher HTTP size per flow than female users. Fig. 5(f) shows that bachelor students have the minimal HTTP size per flow, and master students' HTTP size per flow is slightly larger doctor students' HTTP size per flow.

B. The proposed Prediction Model

To effectively predict the attributes of demographic information in our problem, we propose a novel prediction approach, which is based on Random Forest (RF) [18], a machine learning model. There are many reasons of choosing random forest model in our study. Firstly, RF model randomly chooses items in training set so it is effective to avoid over-fitting. It chooses part of features for each tree, so it can cope with high feature dimension and does not need feature reduction. Secondly, many features we selected are dependent on each other with non-linear relationship. Decision trees in Random Forest are employed to address this issue and Random Forest can detect feature interactions. Thirdly, Random Forest runs fast and efficiently on large data bases. Significant improvements in classification accuracy come from generating an ensemble of trees and letting them vote for the most popular class.

1) Algorithm: Based on the traffic privacy model which is presented in Section III, our prediction model takes a set of users $\mathcal{U}$, whose demographic information is known by the adversary, as prior knowledge. Then the model aims to predict demographics of some other users $\mathbf{u}$. So it can be formulated as a classifier Ψ which predicts demographic class label $j \in \mathcal{J} = \{1, ..., J\}$ at the input $\mathbf{u}$ over independent replicates of the learning set $\mathcal{U}$, which is denoted as:

$$\Psi(\mathbf{u}, \mathcal{U}) = j \tag{2}$$

Our Random Forest based classifier consists of a collection of decision trees $\{h_k(\mathbf{u}, \mathcal{M}), k = 1, ...\}$, where the $\mathcal{M}$ are features extracted from packets fields $\mathcal{F}_{cap}$, which are introduced in previous sections. Then each tree casts a unit vote for the most popular class at input $\mathbf{u}$.

Given a D dimensional feature vector $\mathcal{M}_k = \{m_1, ..., m_D\}$, a decision tree h_k is a collection of nodes n_i organized in

a hierarchical tree structure. Node can be a split node or a terminal leaf node. Assuming a binary decision tree, for each split node n_i, the splitting function $f(\mathcal{M}_k, \pi_i, \phi_i)$ can be represented as:

$$f(\mathcal{M}_k, \pi_i, \phi_i) = \begin{cases} 1 & \text{if } \mathcal{M}_{\pi_i} > \phi_i \\ 0 & \text{if } \mathcal{M}_{\pi_i} < \phi_i \end{cases} \tag{3}$$

where $\pi_i \in \{1, ..., D\}$ is feature index and ϕ_i is the threshold to divide two classes. For demographic class labels $\mathcal{J}$, prediction result $j \in \mathcal{J}$ of the decision tree h_k can be formulated as decision:

$$d_k(\mathbf{u}, j) = 1 \tag{4}$$

while

$$d_k(\mathbf{u}, \mathcal{J}/\{j\}) = 0 \tag{5}$$

Given an input set $\mathbf{u}$ with a size of N and feature vector $\mathcal{M}$, the detailed algorithm is as follows:

- The classifier Ψ first randomly samples N items $\mathbf{u}_k$ with replacement from the $\mathbf{u}$. This sample will be the training set for growing a tree $h_k(\mathbf{u}_k, \mathcal{M})$.
- Each decision tree h_k grows with a number of features $\mathcal{M}_k$ ($\ll \mathcal{M}$) specified at each node n_i. The $\mathcal{M}_k$ features are selected at random out of the $\mathcal{M}$. The best splitting on these $\mathcal{M}_k$ is used to split the node and form splitting function $f(\mathcal{M}_k, \pi_i, \phi_i)$ (which is introduced above). The criterion of Gini impurity $I(\cdot)$ is taken as reference when splitting the node. Feature index π_i^* and threshold ϕ_i^* can be chosen as:

$$\pi_i^*, \phi_i^* = \underset{\pi,\phi}{\operatorname{argmax}} I(\mathcal{M}_k, \pi, \phi) \tag{6}$$

- Each tree grows to the largest possible extent and does not prune. For all trees $\{h(\mathbf{u}_k, \mathcal{M}_k), k = 1, ...\}$, the classifier performs majority voting for classes $\mathcal{J}$ to obtain final result c_{fin}:

$$c_{fin} = \underset{k=1}{\operatorname{argmax}} \sum d_k(\mathcal{X}_k, j) \tag{7}$$

VI. Experiments and Evaluations

In this section, we present our real-world experiments and analysis results.

A. Experiments

To analyze privacy leakage in different scenarios, we consider different time periods $\mathcal{T}$, different leakage source $\mathcal{L}$ and packet fields $\mathcal{F}$ as introduced in Section III. For each group of experiments, we randomly selected 50% of users data as training set and 50% as testing set, and we represent each user as a feature vector using the features mentioned in Section V-A. To measure an average performance, each group of experiments is repeated for 5 times. *Precision, Recall* and *F1-score*, which are well known and broadly used metrics in classification problem, are used to evaluate the results. For each class, *Precision* is the number of true positive results divided by the number of all positive results; *Recall* is the number of true positive results divided by the number of positive results that should have been returned; *F1-score* is defined as $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. A dummy classifier that predicts by randomly guessing was used as the baseline. Thus for gender prediction, the baseline of accuracy is 50% and for education level prediction, the baseline of accuracy is about 33%.

B. Evaluations

1) Short time v.s. Long time (Time variation): In practice, network traffic leakage may last for different time durations. An attacker can choose to sniff Wi-Fi traffic for a long time while he can also sniff Wi-Fi for a short time. Similarly, a victim can connect to a compromised Wi-Fi for a long time or a while. So we consider different time duration of sniffing the traffic corresponding to different traffic leakage scenarios. To simulate different time duration, we randomly select different percentage of traffic of each user from 10% to 100%. The results of predicting gender and education are shown in Table. I respectively.

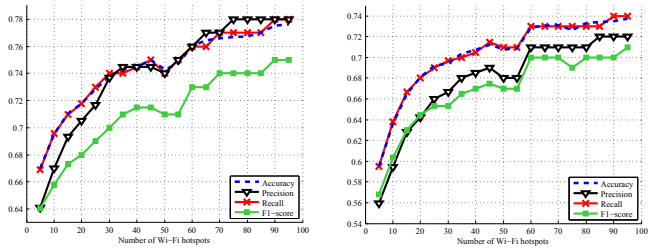
TABLE I: Inference with different percentage of traffic

	Gender				Education			
%	Acc.	Pre.	Rec	F1	Acc.	Pre.	Rec.	F1
20%	0.75	0.75	0.75	0.71	0.74	0.72	0.74	0.71
40%	0.76	0.76	0.76	0.73	0.74	0.72	0.74	0.71
60%	0.77	0.77	0.77	0.74	0.73	0.71	0.73	0.70
80%	0.77	0.78	0.77	0.75	0.74	0.72	0.74	0.71
100%	0.78	0.78	0.78	0.76	0.74	0.73	0.73	0.71

The results show that even only a small part of Wi-Fi traffic reveals users' demographics with a high accuracy. For gender attributes, accuracy, precision, recall and f1-score with 10% of traffic are 0.73, 0.73, 0.73 and 0.69 respectively, and for education level attributes, accuracy, precision, recall and f1-score with 10% of traffic are 0.74, 0.71, 0.74 and 0.70 respectively. Meanwhile, different attributes show different tendencies when the percentage of traffic increases. For gender attributes, all metrics increase while for education level attributes, metrics are not obviously affected by the amount of traffic. It reflects that the small parts of traffic is enough to reach a considerable accuracy of education prediction. Traffic leakage in a short time also poses a serious threat to users privacy.

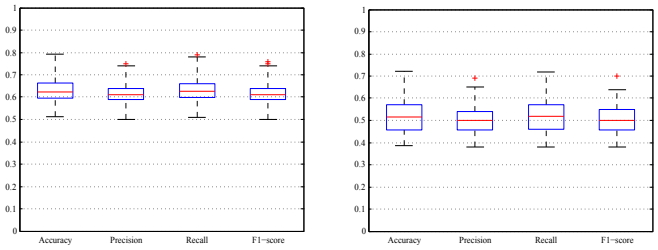
2) One location v.s. All locations: In the real-world attacks, an attacker can sniff one or more Wi-Fi hotspots and a victim may also connect to one or more Wi-Fi hotspots. More sniffed Wi-Fi hotspts means more information leaked. We classify network traffic according to the sources of Wi-Fi hotspots and perform demographic inference by traffic from different Wi-Fi hotspots. As shown in Fig. 6(a) and Fig. 6(b), metrics increase when the number of sniffed Wi-Fi hotspots increases.

To evaluate a lower bound on privacy breach, we consider the scenario that an attacker only gets access to one Wi-Fi hotspot to sniff the traffic. The results are illustrated in Fig. 7(a) and Fig. 7(b). They show the minimum, first quartile (bottom edge of the box), median, second quartile (top edge of the box) and maximum of metrics values using network traffic



(a) Gender (b) Education

Fig. 6: Inference with traffic from different number of Wi-Fi



(a) Gender (b) Education

Fig. 7: Inference with traffic from single Wi-Fi

from a single Wi-Fi hotspot to predict. For gender prediction, the median precision exceeds 60% and the maximum precision exceeds 75%. For education level predication, the median precision exceeds 50% and the maximum precision is close to 70%. The results show that the attacker has a high chance of breaching user privacy even if only one Wi-Fi hotspot is sniffed.

TABLE II: Results of prediction in HTTPS traffic

Demographics	Features	Precision	Recall	F1-score
Gender	location-based	0.67	0.64	0.65
	statistics	0.62	0.69	0.63
Education	location-based	0.72	0.74	0.72
	statistics	0.49	0.54	0.48

3) Encrypted Traffic: To further analyze the extent of privacy leakage through network traffic, we consider the scenario that traffic is encrypted with HTTPS. As discussed in Section III, not only plain text data but also some semantic fields are encrypted, which prevents the direct privacy leakage due to content analysis on network traffic. Whereas, there is still information that can be extracted from a HTTPS-enabled network traffic packet, such as MAC address, IP address and some statistics. So even in HTTPS network traffic, it is possible to infer a user's demographics by observing the metadata of Wi-Fi traffic. We use IP address to generate location features and summarize statistics to get statistical features, as introduced in Section V. We assume that all HTTP packets are encrypted as HTTPS, which is a lower bound of privacy leakage. The results of demographics prediction are shown in Table II.

The results show that even in encrypted traffic, demographics can still be effectively predicted. Location-based features reach high precisions, i.e. 0.67 for gender attributes and 0.72 for education attributes. It is obvious that location-based features perform better than statistical features. It is reasonable because mobility directly relates to person's demographics while statistics of traffic packets are implicit reflection of network activities and not so distinguishable for demographics. But if we only consider predicting results of statistics features, the precision still achieve 0.62 for gender attributes and 0.49 of education attributes, which are higher than baseline with only 12% and 16% accuracy. It shows that relying on encryption cannot address all of the problems and the attacker can still infer the users' demographics by observing the encrypted data.

C. Comparison of Prediction Models

In the last part of this section, we compare the results of different prediction models in comparison to the Random Forest. We tried a set of basic classification algorithms (Decision Tree, Perception, Support Vector Machine, Naive Bayes, K-Nearest Neighbors) to compare the performance. We implemented these algorithms using *scikit-learn* package [22] of Python. The results in Table. III show only Support Vector Machine (SVM) performs better that Random Forest. But the time consumption of Support Vector Machine is 12.79 times longer than the time consumption of Random Forest. So it is a trade-off between the performance and cost.

TABLE III: Inference with different prediction models

Model	Attribute	Precision	Recall	F1-score
Decision Tree	Gender	0.70	0.70	0.70
	Education	0.68	0.64	0.65
Perception	Gender	0.77	0.73	0.74
	Education	0.71	0.65	0.67
SVM	Gender	0.79	0.78	0.79
	Education	0.76	0.72	0.73
Naive Bayes	Gender	0.76	0.76	0.76
	Education	0.61	0.63	0.6
K-Nearest Neighbors	Gender	0.69	0.69	0.69
	Education	0.63	0.66	0.63
Random Forest	Gender	0.77	0.78	0.77
	Education	0.72	0.74	0.71

VII. Discussion

In this section, we discuss limitation in our work and mitigation for the demographic inference.

A. Limitations

For the sake of our inability to access the generic Wi-Fi network traffic, our dataset only includes the Wi-Fi traffic on a university campus. While we confess that this limitation may result in the bias of our dataset, we argue that it does not invalidate our approach – privacy inference through metadata analysis. Our study is based on the observation that the users sharing similar demographics usually have similar classification features, which has been validated by previous works under different contexts such as web browsing, smartphone apps, and mobile social networks [10], [11]. Therefore, our proposed approach can be applied to other datasets (e.g., public Wi-Fi traffic dataset) although the considered features may have some differences. Our study confirms that the threat

of leaking sensitive user information through the meta-data of Wi-Fi traffic is realistic. As one of our future works, we will consider a more resourceful adversary which can collect public Wi-Fi traffic in order to understand the level of privacy leakage through public Wi-Fi data.

B. Mitigation

In the following, we discuss several privacy enhancing techniques that could be used to defend against our traffic demographics inference attack.

- VPN or Tor: VPN or Tor: A potential strategy to mitigate demographics inference attack in Wi-Fi traffic is to prevent the attackers from obtaining the meta-data of network traffic. For example, a user can exploit virtual private network (VPN) or anonymity network Tor to prevent the attackers from tracking the routing information or collecting the traffic characteristics. However, such solutions may incur significant network overhead and suffer from reduced network performance.
- Randomized MAC: Another strategy of thwarting the demographics inference is anonymizing the users via MAC Randomization, which prevents the attackers from linking a specific user with his Wi-Fi traffic. MAC randomization has been introduced on iOS 8 operating system. With MAC randomization, a user can change his MAC address whenever accessing a new Wi-Fi. We perform an experiment by using this strategy. Our experiment results show that it can reduce the accuracy of the proposed demographic inference attack by 15% for gender and 22% for education.
- Dummy Traffic: Dummy traffic is a technique to defend against statistical analysis attack [23]. Therefore, adding dummy packets with random statistics helps hiding user traffic characteristics under the proposed demographic analysis. Our experiments show that, by simply adding 10% of dummy traffic, the inference accuracy is decreased by 9% for gender and 13% for education. Due to the high speed and low expense of Wi-Fi network accesses, adding dummy traffic packets only causes a limited impact on network performance.

In practice, the users can adopt multiple defending strategies to enhance their privacy under the demographic analysis attack. It is noted that there is always a tradeoff between privacy and utility. A practical security defense strategy should strike a balance among multiple factors, including users' convenience, privacy requirement, and network performance.

VIII. Conclusion

In this paper, we use Wi-Fi traffic from 28,158 users in 5 months to analyze demographics leakage and propose the Demographic Information Predictor (DIP) system. DIP extracts four kinds of features from real-world Wi-Fi traffic and applies machine learning technique to predict users' demographics. We consider different scenarios with different time durations, traffic sources and whether data are encrypted or not. The results show that the best accuracy of predicting gender and education level achieve 78% and 74% respectively. Even in encrypted traffic, i.e. HTTPS, users' demographics can be predicted at precision of 67% and 72%. The privacy leakage through Wi-Fi network traffic should become a more serious concern.

ACKNOWLEDGMENT

This work is supported by National High-Tech R&D (863) Program (no. SS2015AA011309), NSFC (no. 61272444, U1401253, U1405251, 61411146001, 61332004).

References

[1] Pablo Valerio, "WiFi Offloading To Skyrocket". http://www.networkcomputing.com/wireless-infrastructure/wifi-offloading-to-skyrocket-/d/d-id/1321007

[2] N. Cheng, X. Wang, W. Cheng, P. Mohapatra, and A. Seneviratne, "Characterizing privacy leakage of public WiFi networks for users on travel," In *Proc. of INFOCOM'13*, IEEE, 2013.

[3] A. K. Das, P. H. Pathak, N. C. Chuah and P. Mohapatra, "Contextual localization through network traffic analysis," In *Proc. of INFOCOM'14*, IEEE, 2014.

[4] N. Xia, H. H. Song, Y. Liao, M. Iliofotou, A. Nucci, Z. L. Zhang and A. Kuzmanovic "Mosaic: Quantifying privacy leakage in mobile networks," *ACM SIGCOMM Computer Communication Review*, 2013.

[5] Wandera, "http://www.realwire.com/releases/Wandera-reveals-Q2-2015-mobile-security-and-data-usage-figures"

[6] B. Konings, C. Bachmaier, F. Schaub, and M. Weber, "Device names in the wild: Investigating privacy risks of zero configuration networking," In *Proc. of MDM'13*, IEEE, 2013.

[7] Y. Fan, Y. Jiang, H. Zhu, and X. Shen, "An efficient privacy-preserving scheme against traffic analysis attacks in network coding," In *Proc. of INFOCOM'09*, IEEE, 2009.

[8] M. U. Ilyas, M. Z. Shafiq, A. X. Liu, and H. Radha, "Who are you talking to? Breaching privacy in encrypted IM networks," In *Proc. of ICNP'13*, IEEE, 2013.

[9] M. Korczynski, and A. Duda, "Markov chain fingerprinting to classify encrypted traffic," In *Proc. of INFOCOM'14*, IEEE, 2014.

[10] J. Hu, H. J. Zeng, H. Li, C. Niu and Z. Chen, "Demographic prediction based on user's browsing behavior," In *Proc. of WWW'07*, ACM, 2007.

[11] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, "Your installed apps reveal your gender and more!". In *Mobile Computer Communication Review*, 2015.

[12] Schwartz, H. Andrew, et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one* 8.9 (2013): e73791.

[13] B. Bi, M. Shokouhi, M. Kosinski, T. Graepel, "Inferring the demographics of search users: Social data meets search queries," In *Proc. of WWW'13*, ACM, 2013.

[14] A. Chaabane, G. Acs, M. A. Kaafar, "You Are What You Like! Information Leakage Through Users' Interests," In *NDSS'12*, 2012.

[15] WiFi Alliance. Wi-Fi CERTIFIED Passpoint. https://www.wi-fi.org/hotspot-20-technical-specification-v100, June 2012.

[16] A. Bittau, M. Handley and J. Lackey, "The Final Nail in WEP's Coffin," In *Proc. of SAP'06*, IEEE, 2006.

[17] E. Tews and M. Beck, "Practical attacks against WEP and WPA," In *Proc. of WiSec'09*, ACM, 2009.

[18] L. Breiman, "Random forests," *Machine learning* 45.1 (2001): 5-32.

[19] M. Srivatsa and M. Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *CCS'12*, ACM, 2012.

[20] S. Li, H. Zhu, Z. Gao, X. Guan, K. Xing and S. Shen. "Location Privacy Preservation in Collaborative Spectrum Sensing," In *Proc. of INFOCOM'12*, IEEE, 2012.

[21] M. Li, H. Zhu, Z. Gao, S. Chen, K. Ren, L. Yu, and S. Hu. "All Your Location are Belong to Us: Breaking Mobile Social Networks for Automated User Location Tracking," In *MobiHoc'14*, ACM, 2014.

[22] scikit-learn: Machine learning in python. http://scikit-learn.org/stable/.

[23] W. M. Shbair, A. R. Bashandy, and S. I. Shaheen. "A New Security Mechanism to Perform Traffic Anonymity with Dummy Traffic Synthesis," In *CSE'09,* IEEE, 2009.