Wi-Fi Traffic Analysis using Advanced Demographic Information Predictor System

ANNA DENNY
CSE, ASIET, Ernakulum, India

Abstract: 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. In this study, we present a novel approach to infer user-demographic information by exploiting the meta-data of Wi-Fi traffic. 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. Results show that, for unencrypted traffic, DIP can predict gender and education level of users with an accuracy of 78% and 74% respectively.

I. Introduction

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. In this study, we present a novel approach to infer user-demographic information by exploiting the meta-data of Wi-Fi traffic. 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. Results show that, for unencrypted traffic, DIP can predict gender and education level of users with an accuracy of 78% and 74% respectively.

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. Zeng [1] presented an approach to predict user’s gender and age from their web browsing behaviors. Web page view information is treated as a hidden variable to propagate demographic information between different users. Learning from the web page clicks through data, web pages are associated with user’s age and gender through a discriminative model. The user’s age and gender are predicted from the demographic information of the associated web pages through a Bayesian framework. Users with similar demographic information visit similar web pages. Only content based and category based features are used. This method is not much efficient to identify and refines the profiles which contain fake information. [2] Kosinski presented an approach which is based on search query histories. Train predictive models based on the publically available dataset containing users Facebook likes. Then, match likes with search queries using Open Directory Project categories. Finally, we apply, the model trained on Facebook likes to large-scale query logs of a search engine. Its main difficulty is to maintain query logs of search users. Das [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. Pathak [4] introduced a Traffic monitoring, and it is classified into two: statistical and application-based. It consists of set of attributes from user’s network traffic and predicts user’s location with high accuracy. Instead of querying, PACL learn user’s location and hence reduces energy consumption. Its disadvantage is payload of packets is expensive. Srivatsa [5] focus on a set of location traces can be deanonymized by using “contact graphs”. It identify meetings between anonymized users in a set of traces can be structurally correlated with a social network graph. Also, it exploited structural similarities between two sources of user correlations and deduce mapping between nodes in the contact graph and deanonymized it. Its demerit is high computational cost.

Different from previous works, our work selects Wi-Fi traffic meta-data which can be sniffed passively as feature to infer demographics leakage. The proposed system is demographic inference system, which can predict user demographic information through meta-data analysis of Wi-Fi traffic. The features of proposed system are:

1) More Scalable: For addressing 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.
2) Larger Target Coverage: 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.

3) HTTPS Traffic Tolerance: The proposed system is expected to exploit the available meta-data, which cannot be protected by HTTPS protocol, to infer user demographic information.

III. Advanced demographic Information Predictor (ADIP)

In this, we present Demographic Information Predictor (ADIP) system, which can extract information from traffic and predict users’ demographics based on the meta-data of Wi-Fi traffic.

Fig.1 DIP system architecture

ADIP aims to automatically extract information from traffic and generate profile signatures to predict users’ profiles. If fake service providers or external adversaries are able to monitor traffic passively, they can exploit ADIP 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. Network access behavior and mobility characteristics will also share the similar demographic features, which are supported by the previous work on web browsing analysis. The system architecture is shown in Fig. 1.

A. Traffic Process Engine

ADIP data is collected and information which is used to identify users’ profiles. Given a series of traffic as input, ADIP parses traffic packets and extracts targeted fields. ADIP uses MAC addresses to identify devices and aggregates flows from the same devices or the same IP addresses. Then ADIP 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, ADIP 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 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.

B. Profile Signature Generator

Profile Signature Generator is used to extract features for predicting users’ profiles. With information generated from Traffic Process Engine, the features are classified 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.
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 user’s demographics. The prediction model of Profile Predictor is based on Random Forest model [3], 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.

ADIP Features And Prediction Model

Here we can see the core components of DIP in details, which mainly consist of feature selection and prediction model.

D. 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 host’s 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 \( A = \{ \text{male}, \text{female} \} \) or education where \( \theta \) 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. 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 female users. 5 applications with the lowest entropy for the users with education level of bachelor, masters and doctors respectively. It can be observed that bachelor students are more interested in Electronic Product, while Job is most popular among masters and Marriage among doctors.

E. Category-based Features

Applications in our dataset are classified 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 the gender attribute. It shows that Social Networks, Job and Finance are most popular in bachelors, masters and doctors respectively. The category-based features can also be used to distinguish different groups of users.

F. 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 those IP addresses of Wi-Fi access points are highly correlated to locations of users. 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. Thus location based features are supposed to show a strong correlation with demographics.

G. Statistical Features

Statistical features reveal distinct information that can distinguish users with different demographic information. Male users have higher time duration than female users do. Male users will stay a longer time for each network access. Master students have the highest average time duration, and bachelor students’ time duration is slightly longer than doctor students’ time duration. 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 master students’ HTTP number per flow. Male users have higher HTTP size per flow than female users. The bachelor students have the minimal HTTP size per flow, and master 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) [3], a machine learning model. 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. Then, many features we selected are dependent on each other with non-linear relationships.
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.

IV. Conclusion

In this paper, we 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. Some of the results show that the best accuracy of predicting gender and education level. Even in encrypted traffic, i.e. HTTPS, users’ demographics can be predicted.

References

Journal Papers: