IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676,p-ISSN: 2320-3331,
PP 34-37
www.iosrjournals.org

Wi-Fi Traffic Analysis using Advanced Demographic Information Predictor System

ANNA DENNY
CSE, ASIET, Ernakulum, India

Abstract: *Represents a more efficient and scalable approach toinfer users' sensitive information without checking the content ofWi-Fi traffic. Secondly, meta-data based demographics inferencecan work on both unencrypted and encrypted traffic. In this study, we present a novel approach to infer userdemographic information by exploiting the meta-data of Wi-Fitraffic. DIP extracts four kinds of features from realworldWi-Fi traffic and proposes a novel machine learning basedinference technique to predict user demographics. Results show that, for unencrypted traffic, DIP can predict genderand education level of users with an accuracy of 78% and 74%respectively.*

I. Introduction

Represents a more efficient and scalable approach toinfer users' sensitive information without checking the content ofWi-Fi traffic. Secondly, meta-data based demographics inferencecan work on both unencrypted and encrypted traffic. In this study, we present a novel approach to infer userdemographic information by exploiting the meta-data of Wi-Fitraffic. DIP extracts four kinds of features from realworldWi-Fi traffic and proposes a novel machine learning basedinference technique to predict user demographics. Results show that, for unencrypted traffic, DIP can predict genderand 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 leakagethrough meta-data analysis of Wi-Fi traffic. The presentedwork 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 pageclick-through data, web pages are associated with user's age and gender through a discriminative model. Theuser'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 bypassively 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 predictsuser'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, itexploited 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 featuresto 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 trafficand predict users' demographics based on the meta-data of Wi-Fi traffic.

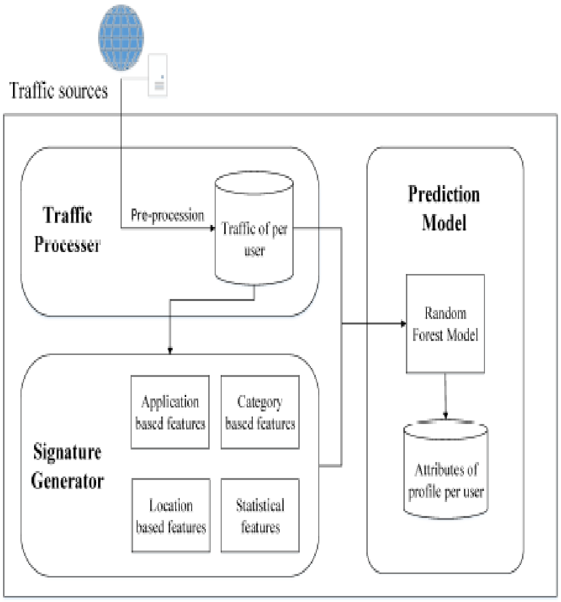

Fig.1 DIP system architecture

ADIP aims to automatically extract informationfrom traffic and generate profile signatures to predict users'profiles. If fake service providers or external adversariesare able to monitor traffic passively, they can exploit ADIP topredict user's profile. Our insight is based on the fact thatit is highly possible that users having similar demographicshave similar network usages. Network access behaviorsand mobility characteristics will also share the similar demographicfeatures, which are supported by the previous work onweb browsing analysis. The system architecture is shown inFig. 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 trafficpackets 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 withtargeted meta-data such as Host, User-agent and URL in HTTPprotocol and preserves the data sequence of the users. Forfurther analysis, ADIP also handles some procession includingaggregating domains addresses from the same serviceproviders. 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 fromTraffic 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 smart phone 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 featurescan be extracted from IP address because IP address of theaccess point reflects coarse location of a user and is highlycorrelated to contextual location [3]. Location based featurescan 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, durationof a session.

C. Profile Predictor

Once Profile Signature Generator generates different kindsof features, Profile Predictor uses features to predict demographicsof 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], whichruns efficiently on large data bases and can handle thousandsof input features without feature selection. In this work, weassume part of users' information is available in public andcan be used to train the model. The generated model can besaved 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 andprediction model.

D. Application-based Features

Application-based featuresare represented as hosts in HTTP protocol in our problem.Application-based features indicate which websites users visitedand which apps users ran or services users enjoyed. Sincethere are a large number of hosts in our dataset and some of thehosts are from the same service providers or organizations, weaggregate them according to host's similarity, and only selectapplications used by at least 10% of the users.Certain applications show strong tendency towards attributesof demographics, such as gender and education. To characterizethe tendency, we calculate the entropy of each applicationwith respect to attributes. Let A be a kind of demographicattributes, e.g. gender A = {male, female} or educationwhere θ is the user distribution of an attribute.Entropy measures the uncertainty of each attribute. Entropyhas the maximum value when the probability of each tendencyfollows a uniform probability and has the minimum valuewhen the probability of one tendency is dominant. So lowerentropy of an application indicates it is distinguishable withrespect to an attribute. Since the number of users of eachattribute is imbalance, we under sample users of each kindof the attributes to keep balance. 5 applications with lowest entropy for male users and female usersrespectively. In general, Game and Sport are more popularamong male users while Fashion and Shopping are morepopular among male users. 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, whileJob is most popular among masters and Marriage amongdoctors.

E. Category-based Features

Applications inour dataset are classified into 39 categories. To evaluate tendency of eachcategory, we calculate entropy again. Fig 4(c) shows the 10categories with lowest entropy for the gender attribute. Itshows that Sports, Finance and Real Estate are more popularin male users and Women, Entertainment are more popularin female users. Similarly, Fig. 4(d) shows the 10 categorieswith lowest entropy for education level attribute. It showsthat Social Networks, Job and Finance are most popular inbachelors, masters and doctors, respectively. The category-based features can also be used to distinguishdifferent groups of users.

F. Location based Features

Different Wi-Fi access points have different IP addresses. We can extract IP addresses fromIP layers of traffic packets so we can know where users accessWi-Fi access points. Location is a strong indicator of users'demographics. On one hand, previous work has pointed outthose IP addresses of Wi-Fi access points are highly correlatedto locations of users. On the other hand, location basedfeatures can be viewed as mobility of users, and the mobilityis highly correlated to users' profile attributes such as hobbies,habits and relationship.Thus location based featuresare 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 averagetime duration, and bachelor students' time duration is slightlylonger than doctor students' time duration. Male users have higher HTTPnumber per flow than female users. Master students have the maximal average HTTP number per flow, and bachelorstudents' HTTP number per flow is slightly larger than doctorstudents' HTTP number per flow. Male users have higher HTTP size per flow than femaleusers. The bachelor students have the minimal HTTP size per flow, and master students' HTTP size per flow B. The proposed Prediction ModelTo effectively predict the attributes of demographic informationin our problem, we propose a novel prediction approach,which is based on Random Forest (RF) [3], a machine learningmodel. RF model randomly chooses itemsin training set so it is effective to avoid over-fitting. It choosespart of features for each tree, so it can cope with high featuredimension and does not need feature reduction. Then, many features we selected are dependent on each other withnon-linear relationship. Decision trees in Random Forest areemployed to address this issue and Random Forest can detectfeature interactions. Thirdly, Random Forest runs fast andefficiently on large data bases. Significant improvements inclassification accuracy come from generating an ensemble oftrees and letting them vote for the most popular class.

IV. Conclusion

In this paper, we propose the DemographicInformation Predictor (DIP) system. DIP extractsfour kinds of features from real-world Wi-Fi traffic and appliesmachine 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:

[1] 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.

[2] 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.

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

[4] 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.

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