Computer Networks 118 (2017) 24–36

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Privacy-aware contextual localization using network traffic analysis☆

Aveek K. Das[a,*], Parth H. Pathak[b], Chen-Nee Chuah[c], Prasant Mohapatra[a]

[a] Computer Science Department, University of California, Davis, CA, USA
[b] Computer Science Department, George Mason University, Fairfax, VA, USA
[c] Electrical and Computer Engineering Department, University of California, Davis, CA, USA

ARTICLE INFO

Article history:
Received 20 August 2015
Revised 20 December 2016
Accepted 16 February 2017
Available online 24 February 2017

Keywords:
Localization
Contextual location
Wireless traffic monitoring
Internet measurement

ABSTRACT

The rise of location-based services has enabled many opportunities for content service providers to optimize the content delivery to user's wireless devices based on her location. Since the sharing precise location remains a major privacy concern among the users, certain location-based services rely on *contextual location* (e.g. residence, work, etc.) as opposed to acquiring user's exact physical location. In this paper, we present PACL (Privacy-Aware Contextual Localizer) model, which can learn user's contextual location just by passively monitoring user's network traffic. PACL can discern a set of vital attributes (statistical and application-based) from user's network traffic, and predict user's contextual location with a very high accuracy. We design and evaluate PACL using real-world network traces of over 1700 users with over 100GB of total data. Our results show that PACL, when built using the Bayesian Network machine learning algorithm, can predict user's contextual location with the accuracy of around 89%.

1. Introduction

A tremendous growth has been observed in location-based services, in the last few years. On a broad scale, current location-based services can be classified into two categories. Users navigate to specific locations, search for restaurants and businesses near a certain location, check-in on social networks, etc. using these location-based services. The first category requires *precise* user location to provide its services. One example for such a service is smartphone navigation system where exact latitude and longitude information is essential. The other type of services only need contextual information about the users' location. For example, knowing that a user is at an airport or a shopping mall is sufficient (and necessary) to provide certain services specific to that location category. *Contextual location* information is also important for content providers and Content Distribution Networks (CDNs) which can use this knowledge to optimize the content delivery and provide useful recommendations based on user's location type. Third party services, also, can provide targeted advertisements related to the contextual location of the user. Most users believe that contextual location based services are based on precise user location, which they are not comfortable to share, in most occasions, to receive contextual location based services. If these services can be provided to users without compromising their privacy (about precise location), we believe users would be benefited by such services. In this paper, we present a privacy-preserving system that can determine user's location category (or contextual location) just by passively monitoring and learning from aggregate network traffic from different categories of location.

Existing services such as FourSquare [1] can be used by content providers to map a user's precise location to her contextual location category but this requires the user to share their precise physical location. Increasing concerns about location privacy, have prompted more and more users to be unwilling about provide their location information, especially for contextual location-based services. This insecurity among users have led to the *Do Not Track Me Online Act of 2011* [2] which provides users with an option to disable tracking of its location by content providers or websites. As an example of privacy preferences, we can say that users are willing to share their GPS location for Google Maps Navigation but when services such as YouTube ask for user's location, users often deny the request even though content delivery could have been optimized by YouTube if the location was available.

In this paper, we propose a network traffic analysis technique whereby an ISP or any third-party entity capable of passively monitoring network traffic can determine user's contextual location (without knowing user's exact physical location). The ISP can

☆ An earlier version of this work was submitted and accepted for publication at IEEE INFOCOM 2014. The work was titled "Contextual Localization through Network Traffic Analysis".

* Corresponding author.
E-mail addresses: akdas@ucdavis.edu (A.K. Das), phpathak@gmu.edu (P.H. Pathak), chuah@ucdavis.edu (C.-N. Chuah), pmohapatra@ucdavis.edu (P. Mohapatra).

http://dx.doi.org/10.1016/j.comnet.2017.02.011
1389-1286/

use the traffic analysis technique to determine the users' location category. Once the contextual location has been identified, CDNs can probe to obtain this information from the ISPs using the proposed ISP-CDN collaboration model [3,4]. This information can then be utilized by the CDNs to provide contextual location based services to users, like targeted advertisements. For example, at work, a person would prefer to get an advertisement of a word-processing software on sale rather than get an advertisement for a movie ticket. Thus, one of the major applications of the proposed technique is to provide location context based advertisements to users.

Our method can also work without an ISP accessing the contents of the packets (such as website being accessed or payload). Protocol identification and relevant statistical features are sufficient for location categorization. As we see later in the paper, statistical features of flow, packets and protocols in the user created network data can be used to achieve an accuracy of location prediction which is as high as 83%. This is accomplished without looking at the content of the packets. This kind of inspection is often carried out by the ISP for traffic engineering and security purposes. Hence, we believe that ISPs can assist in location categorization using our technique while adhering to the privacy acts. After determining the location category, the ISPs can also fine-tune their security policies, as public locations (like cafeteria/restaurants) needs different policies as compared to private locations (like apartments). For example, certain ports and flows in a public location context can be blocked to provide more security to users from attackers.

In this work, first, we show that network traffic originating from different types of locations (such as cafe, university campus, residence etc.) have built-in distinct signatures based on the location category. Second, we propose a traffic analysis engine that can leverage information collected by existing passive traffic monitoring systems to discern the contextual location signature. The signature is composed of different attributes that may differ depending on the type of location (e.g., applications users access at different locations, flow length, packet size distributions etc.) These location signatures can be used to identify the contextual location of any IP address.

The contributions of our work are as follows:

1. First, we show that traffic originating from different *types* of locations have distinct signatures embedded in them. To establish this, we have collected nearly a 100GB of real-world network traffic traces for over 1700 users at different types of locations. We identify a number of attributes which when used together can create a distinct contextual location signature.
2. Next, we present a system (named PACL - Privacy-Aware Contextual Localizer) that can learn user's contextual location only by passively monitoring user's traffic flows. The core of PACL is a supervised machine learning engine that can predict user's contextual location efficiently and accurately. We evaluate our PACL model using our network traces, based on six machine learning algorithms. The best prediction accuracy is observed using the Bayesian Network classification algorithm which show that PACL can predict contextual location with an overall accuracy of 89%. This model not only gives overall good accuracy, the accuracy for the individual classes are also very similar and equally efficient.

This paper is structured as follows. We start out with discussion of related research works in Section 2. In Section 3, we introduce the PACL system and describe its functioning in details. Section 4 includes details about the dataset used for analysis. The features which differentiate each contextual location are discussed in Section 5. In Section 6, we present the methods used for feature selection. The prediction model and the prediction results observed using our proposed model are in Section 7, followed by conclusions in Section 8.

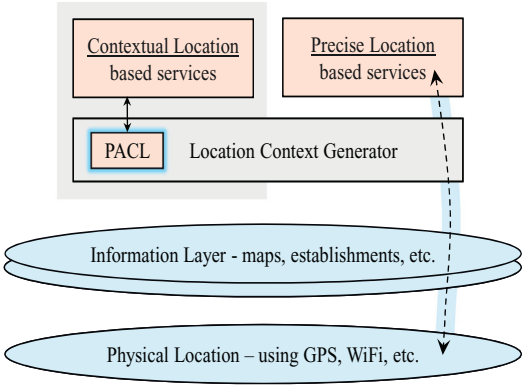

Fig. 1. PACL as compared to regular localization using precise location.

2. Background and related work

Traditional location-based services are built on top of positioning systems (e.g. GPS) and information layer (e.g. maps, database of establishments etc.). This is depicted in Fig. 1. Here, location-based services that require exact physical location typically use data from user's positioning system combined with details of information layer. This opens up many entry points for privacy invasion of users. On the other hand, certain services (such as targeted advertising, content delivery optimization etc.) do not require user's exact physical location. Also, users are less likely to provide their location for such services. Our solution, PACL, can address this challenge by eliminating the need of user's physical location in the case of contextual location-based services (see Fig. 1). Instead of querying users for precise location, PACL passively learns user's contextual location by monitoring users' network traffic.

Determining Location and Preserving Privacy: Significant amount of past research has mostly focused on two topics: (i) accurate and energy-efficient determination of user's physical location and, (ii) preserving user's privacy when sharing user's location information. In the first category of research, a variety of location determination mechanisms have been proposed like in [5,6]. The central focus of these studies is to reduce the energy consumption of determining the location while increasing the accuracy. Also, other techniques such as map matching [7] are used to improve the accuracy. Location privacy preserving techniques have attracted a lot research starting from initial studies such as [8]. Methods such as cloaking [9] and obfuscation [10] are proposed as ways to prevent privacy leakage of users using location-based services. PACL is different from these studies as it does not require actual physical location and other privacy preserving methods for protecting the physical location.

Traffic Classification: Another thread of research that is relevant to PACL is known as Internet traffic classification. The purpose of traffic classification is to monitor and analyze network traffic for determining applications and protocols being used. It is a well-established method ([11] and references therein) of profiling network traffic, anomaly detection and detecting file sharing of copyrighted content. Such traffic classification techniques and PACL share a few common characteristics. They both utilized traffic monitoring and are built using machine learning algorithms. Nevertheless, we believe that PACL takes a step forward by learning and predicting contextual location purely through network traffic analysis.

Another research work relevant to ours is [12] in which Trestian et al. provide a detailed study on applications accessed by users at