Title
Contextual Localization Through Network Traffic Analysis

Permalink
https://escholarship.org/uc/item/53m934w1

ISBN
9781479933600

Authors
Das, Aveek K
Pathak, Parth H
Mohapatra, Prasant
et al.

Publication Date
2014-04-01

DOI
10.1109/infocom.2014.6848021

Peer reviewed

Contextual Localization Through Network Traffic Analysis

Aveek K. Das*, Parth H. Pathak*, Chen-Nee Chuah†, Prasant Mohapatra*
*Computer Science Department, †Electrical and Computer Engineering Department,
University of California, Davis, CA, USA.
Email: {akdas, phpathak, chuah, pmohapatra}@ucdavis.edu

***Abstract*—The rise of location-based services has enabled many opportunities for content service providers to optimize the content delivery based on user's location. Since sharing precise location remains a major privacy concern among the users, many location-based services rely on *contextual location* (e.g. residence, cafe etc.) as opposed to acquiring user's exact physical location. In this paper, we present PACL (Privacy-Aware Contextual Localizer), 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 100 gigabytes of total data. Our results show that PACL (built using decision tree) can predict user's contextual location with the accuracy of around 87%.**

I. Introduction

In recent years, tremendous growth has been observed in location-based services. At large, current location-based services can be classified into two categories. The first category of services require *precise* location of the users, for example, smartphone navigation system where exact latitude and longitude information is essential. The second type of services only need contextual information about location. For example, knowing that a user is at a cafeteria or a shopping mall is sufficient (and necessary) to provide services specific to that location category. Determination of *contextual location* information is also extremely important for content providers and Content Distribution Networks (CDNs) to optimize the content delivery and provide recommendations based on user's location type. Third party services, also, can provide targeted advertisements related to the contextual location of the user. In this paper, we present first-of-its-kind 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.

Note that content providers can use existing services such as FourSquare to map user's precise location to contextual information but this requires users to share their physical location. Due to increasing concerns about location privacy, more and more users are unwilling to provide their location information, especially for contextual location-based services. This led to the *Do Not Track Me Online Act of 2011* [1] which gives users an option to disable tracking of its location by content providers or websites. As an example of privacy preferences, 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). Once the contextual location has been identified, this information can be shared with content providers using recently proposed ISP-CDN collaboration model [2], [3].

First, we show that network traffic originating from different types of locations (such as cafe, university campus, residence etc.) have built-in distinct signatures. 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 originated at different *types* of locations have distinct signature embedded in them. To establish this, we have collected nearly a 100 gigabytes 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 built using decision tree that can predict user's contextual location efficiently and accurately. We evaluate PACL using our network traces, and show that PACL can predict contextual location with an overall accuracy of 87%.

This paper is structured as follows. We start out with discussion of related research works in Section II. In Section III, we introduce the PACL system and describe its functioning in details. Section IV includes details about the dataset used for analysis. The features which differentiate each contextual location are discussed in Section V. In Section VI, we present the prediction model and the prediction results observed using our proposed model, followed by conclusions in Section VII.

II. 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.

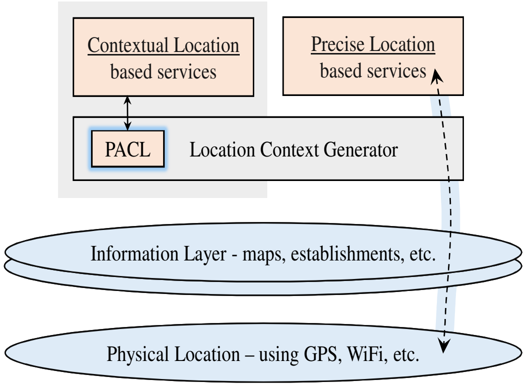

Contextual Location based services
Precise Location based services
PACL
Location Context Generator
Information Layer - maps, establishments, etc.
Physical Location – using GPS, WiFi, etc.

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

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 [4], [5]. 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 [6] are used to improve the accuracy. Location privacy preserving techniques have attracted a lot research starting from initial studies such as [7]. Methods such as cloaking [8] and obfuscation [9] 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 ([10] 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.

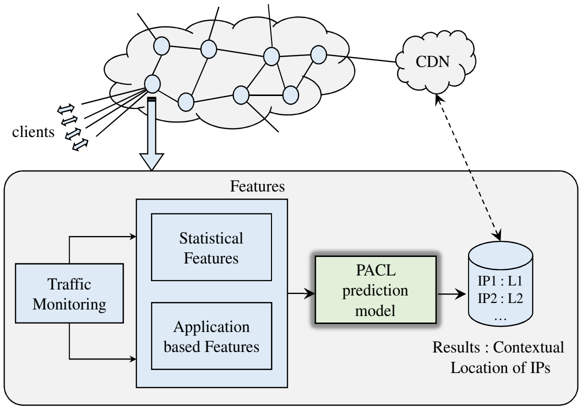


Fig. 2: Architecture of the PACL system: Network traffic is monitored for a number of features, which when used in the PACL model gives contextual location prediction of an IP.

Another research work relevant to ours is [11] in which Trestian et al. provide a detailed study on applications accessed by users at different locations and show that they tend to be different at work and home, irrespective of the time of the day. Our model not only profiles the usage of applications and services by users at different locations but also combines them with other statistical features to predict their contextual location.

There are many online third-party software tools which claim to predict the geographical location of an IP address [12]. However these services only provide city-level information of the IP address but neither the exact location or the contextual location is available. Some of these tools provide geographical coordinates, but those mostly refer to the coordinates of the ISP the IP address is registered to.

III. Privacy Aware Contextual Localizer (PACL) System

In this work, we design Privacy Aware Contextual Localizer (PACL) system, which can determine the *category* of user's location. PACL is built on a simple fundamental idea that user's network activity is highly dependent on user's contextual location. If one is able to identify the attributes of network traffic that are sufficiently different across different contextual location, ISP or any third party entity capable of passively monitoring traffic, can use the same set of attributes to determine user's location context. This location context can then be shared with content service providers who can optimize the content deliver accordingly. The foremost advantage of the PACL system is that users are not required to share their

TABLE I: Dataset Used For Location Signature Analysis

Location Type	Traces	No. of IPs	Total IPs.	Total flows	Packet Count (Million)	Duration (Hours:Minutes)	Size of Network Trace
Residential	Apartment-1	91	315	16695	16.47	7:40	7.2 GB
	Apartment-2	78		20505	31.15	10:40	14.9 GB
	Apartment-3	72		14396	17.45	3:22	7.9 GB
	Apartment-4	52		6465	14.82	2:44	6.8 GB
	Apartment-5	22		12469	8.38	3:16	3.1 GB
University Campus	Department hall	114	529	14887	27.34	5:12	5.9 GB
	Library-1	313		20153	83.62	7:55	21.9 GB
	Library-2	102		26861	65.29	8:19	19.2 GB
Cafeteria/Restaurant	Starbucks-1	234	450	39532	12.89	8:03	5.6 GB
	Starbucks-2	216		44720	12.73	8:48	4.9 GB
Airport/Travel	Washington-1	88	458	10682	2.01	0:18	682 MB
	Sydney-1	80		8586	4.05	1:24	1.4 GB
	Orlando	63		2280	1.35	0:20	499 MB
	Washington-2	55		3201	1.00	0:13	209 MB
	Denver	53		7264	2.02	0:21	515 MB
	Washington-3	40		1338	1.37	0:20	340 MB
	Los Angeles	39		2691	1.01	0:15	411 MB
	Sydney-2	23		872	0.84	0:25	190 MB
	San Francisco	17		2024	1.17	0:15	624 MB

precise location with anyone, and at the same time, they can be served using the content that is optimized based on their location context. The components of the PACL system are shown in Fig. 2.

Traffic Monitoring: PACL can be deployed within traffic monitoring systems of an ISP or an AS (Autonomous System). Flows originating from user IPs can be monitored for a fixed amount of time after which PACL determines its contextual location. Note that PACL is similar to traditional Internet traffic classification methods as it performs better when complete bi-directional network traffic of end-user IPs can be monitored. Since this is the first attempt towards determining type of location purely using network traffic, we restrict our study to the case where PACL is deployed on traffic monitors with complete bi-directional network flows.

In our measured dataset, we collect network traffic over the edge at WiFi hotspots deployed at different types of locations (details in Sec. IV). We build and verify PACL using the traces of over a 100 gigabytes collected at different location over the period of 20 days.

Identifying Location Signature: In the PACL, we first identify specific attributes of IPs which are likely to be correlated to IP's location. In the training phase, we use the available ground-truth of location to find the correlation between the attributes with the location. The attributes (or features) we use can be classified in two categories - statistical features and application-based features. Examples of statistical features include number of flows originated by an IP, packet length distribution of all packets of an IP etc. On the other hand, in the application-based features, we classify user's network flows in different categories of applications (such as emails, games, social-networks etc.). To understand what kind of content users are interested in (independent of which application they use to access it) when at a specific location, we also classify flows into different interest categories. We show that both statistical and application-based features can generate a distinct signature for different locations.

Applying Location Signatures to Determine Location Context: Once the location signature has been identified, PACL prediction model predicts the contextual location of a user based on location signature mentioned above and the observed statistical and application-based features associated with the particular user (or IP address). As shown in Fig. 2, the results are stored in a repository, which can be accessed by the content providers to optimize content delivery and provide location-specific services. However, even after prediction of contextual location of an IP address, PACL continues to predict contextual location as dynamic reallocation of IPs might change IP's location category. The prediction model is built using a decision tree with reduced error pruning. It is observed that the combination of both the statistical features and application based features give better prediction of location context than using each set individually. Application of this model on our dataset of over 1700 users yeilds a prediction accuracy of over 87%.

Before describing PACL in details, we discuss the application scope and limitations of PACL. First and foremost, PACL can not be used for location-based services where user's precise location is essential. In other words, it can not be used for applications where precise location is more important than preservation of privacy. Second, PACL is capable of predicting most common "location types" but its current form can not characterize traffic from short-term gatherings (such as a sports event).

IV. Network Traffic Collection and Datasets

One major challenge we faced in developing the PACL system is to acquire network traffic traces which precisely originate at specific locations. If network traces from ISP or AS are used, they might not always have the ground-truth location for different IPs. To address this challenge, we capture the network traffic at the edge at different WiFi hotspots deployed at different locations. The details of the datasets are presented in Table I.

A. WiFi Packet Captures

The data is collected by passively sniffing WiFi packets from the air near the WiFi hotspot. We chose four different categories of locations - residential, university campus, cafeteria/restaurants and airport/travel (see Table I). For each category, we collected traces at multiple different locations of that category to extract/learn the category-specific characteristics.

The traces were collected using TP-Link WN722N WiFi USB adapters [13] connected to a laptop running Linux. The WiFi adapters run in monitor mode of ath9k driver [14] and Wireshark is used to capture the packets. We connect three different adapters to each laptop in order to simultaneously capture on 3 different channels (channels 1, 6 and 11 of 2.4 GHz IEEE 802.11 b/g/n). The traces account for a total of over 100 gigabytes of data captured over 20 different days. The airport traces were captured in 2012 as described in [15].

The dataset and the subsequent analysis is based on classification of contextual location into four classes. However, the PACL model can be extended to incorporate other location categories, provided the model is trained beforehand based on the features from those locations. The analysis done here is based on wireless network traces, but the analysis is applicable for wired network traffic. We use WiFi traces as they can be collected easily in public settings, and in any case, most of the devices that are used at these locations are wireless devices.

B. Data Sanitization

Before processing the data as input to the PACL learning model, we sanitize the network traces. The process of the sanitization phase is divided into two steps. First, the collected dataset is anonymized to remove any personal identity related information. The second step involves removing all the packets from the network traces which will not be forwarded to the ISP. In this step, all the MAC layer frames (such as WiFi beacons etc.) as well as MAC layer headers are removed from all IP packets as these information is not forwarded beyond WLAN.

V. Finding Location Signature

We propose a traffic analysis system, which can passively monitor network traffic and extract the **statistical features** and **application and service based features**, on a per-IP basis, to be used for learning and prediction.

A. Statistical Features

For each IP address in the trace, we calculated the statistical features listed below. They are divided into 4 subsets as shown below. Type I and II attributes hold single numerical values, while the attributes of Type III and IV are distributions, which are represented using <min, max, average, median, standard deviation, skewness, kurtosis>. Note that, a flow is identified using a 5-tuple <source IP, source port, destination IP, destination port, protocol>.

Type I - Coarse-grain statistics:

1) Total number of flows
2) Average number of concurrent sessions
3) Percentage ON time - ratio of number of 10 second blocks when IP was active (had at least one flow) to the total time of the trace
4) Number of activity periods (one activity period = a period of time when the IP was continually active, i.e. had at least one flow active)
5) Number of bytes transferred
6) Number of packets transferred
7) Average throughput

Type II - Protocol level statistics:

8) Number of HTTP flows
9) Number of HTTPS flows
10) Number of TCP (non-HTTP/HTTPS) flows
11) Number of UDP flows

Type III - Flow level statistics:

12) Flow length
13) Flow throughput
14) Bytes transferred per flow
15) Packets transferred per flow

Type IV - Packet level statistics:

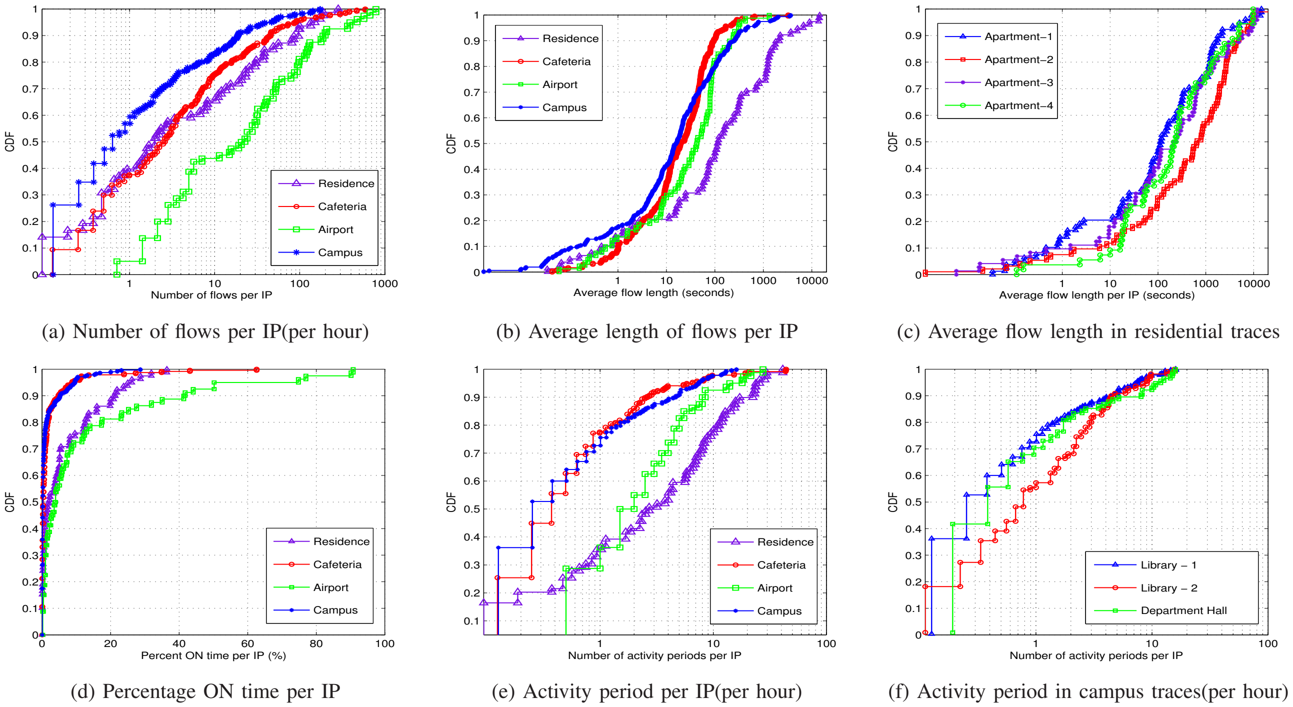
16) Packet inter-arrival time
17) Packet size

The total number of statistical features are 53 (1 feature each for Type I and II and 7 features for each distribution for the statistics of Type III and IV).

During the entire time of the trace, the DHCP lease to a particular device does not expire and thus for all calculations, we assume one IP address is assigned to one device (we also verify this by checking the MAC addresses corresponding to each IP address). For the calculation of activity period, percentage ON time and concurrent flows per IP address, the entire trace duration was divided into bins of 10 second intervals each and the analysis was done based on the whether an IP address created any flow during each of these time bins. The statistical attributes which are directly dependent on the total time of the trace (e.g., total flows per IP, total number of HTTP flows, etc.) were normalized on a per hour basis, to eliminate any biases due to difference in the duration of different traces.

Analysis of Statistical Features: The statistical attributes reveal distinct information that can serve as location signature and in turn, used to predict contextual location. Some of these characteristics are shown in Fig. 3. As we can see, airport trace has the highest number of flows per IP per hour as compared to the other locations, where as Campus has the lowest, as seen in Fig. 3a. Airport and cafeteria traces have mostly smartphone based network traffic and thus each device generates a large number of flows (due to background applications and ads). On the other hand, campus traces have a large number of IP addresses with very low flow count - as there are users who pass by the WiFi hotspot and their devices, which are connected to the campus network, by default, may generate traffic for that transient period of time.

Figs. 3b and 3e show the length of flows and the number of activity periods per IP are the largest in case of residence



(a) Number of flows per IP(per hour)

(b) Average length of flows per IP

(c) Average flow length in residential traces

(d) Percentage ON time per IP

(e) Activity period per IP(per hour)

(f) Activity period in campus traces(per hour)

Fig. 3: Statistical attributes: Figures (a), (b), (d) and (e) represent variation of attributes across four different location classes, while Figures (c) and (f) represent the variation of a particular attribute across the different traces of the same location class.

as compared to others. This is expected, as in residential buildings users tend to keep their devices on for longer duration, even though the usage can be in on-off manner and not continuously. From Fig. 3b we can observe that more than 50% of the IP addresses in the residential traces have flow lengths greater than top 10% IP flow-lengths in cafeteria trace. This is because most users tend to stay for a very short time in cafeterias. This proportion of users is smaller in campus as many users prefer to sit at once place. However there are several IP addresses with very small flow-lengths in campus trace, generated due to users who happen to pass by, as mentioned above.

Activity Period: One of the most distinct attributes among different location categories is activity period, as we will later see in Section VII. We calculate activity period count as the number of times an IP was continuously generating at least one flow in each of the 10 second time intervals, the whole trace was divided into. Fig. 3e indicates the higher number of activity periods in apartments, but questions may arise as to why such a trend is observed in airports too. This is because the activity period is normalized on a per-hour basis and the activity periods actually calculated are for approximately 15-30 minute traces. Hence we see higher number of activity periods in airport trace. Around 90% of IP addresses at campus and cafeteria have activity period count less than five. This is mainly as a result of passer-by user devices in campus traces and users in cafeteria traces who connect to the network for a few specific purposes.

Percentage ON Time: The percentage ON time of each IP address represents the total time an IP was active, as a percentage of the entire time of the trace. As seen in Fig. 3d, apartment and airport traces have the highest ON time percentage of all the four locations as most user devices are usually on for almost the entire time of the trace (note that airport traces are very short in duration). ON time percentages in cafeteria is smaller than those in campus, but there are some devices with very high percentage ON time in the cafeteria dataset. This is most likely to be due to the employees of the establishment who were present at that location during the entire data collection time.

Variation across datasets for the same location category: Figs. 3c and 3f show the variation of two specific attributes across more than one trace of a particular location. These two figures help us to show that the variation of a particular attribute across multiple traces at the same category of location behaves similarly, inspite of the fact that the trace was collected in a different date and at a different location (but same contextual location). Similar trend across different traces at same location category is seen for almost all of the above mentioned features, which help us to assign a specific signature for each type of location.

B. Application based Categorization

To detect the interest of users in various kinds of applications at different locations, we use a keyword based search on the content of the captured packets, a method similar to the one used in [11]. Packets include the HTTP objects like GET, POST and URLs as well as DNS queries and answers. For the keyword based search, we created a keyword list, currently

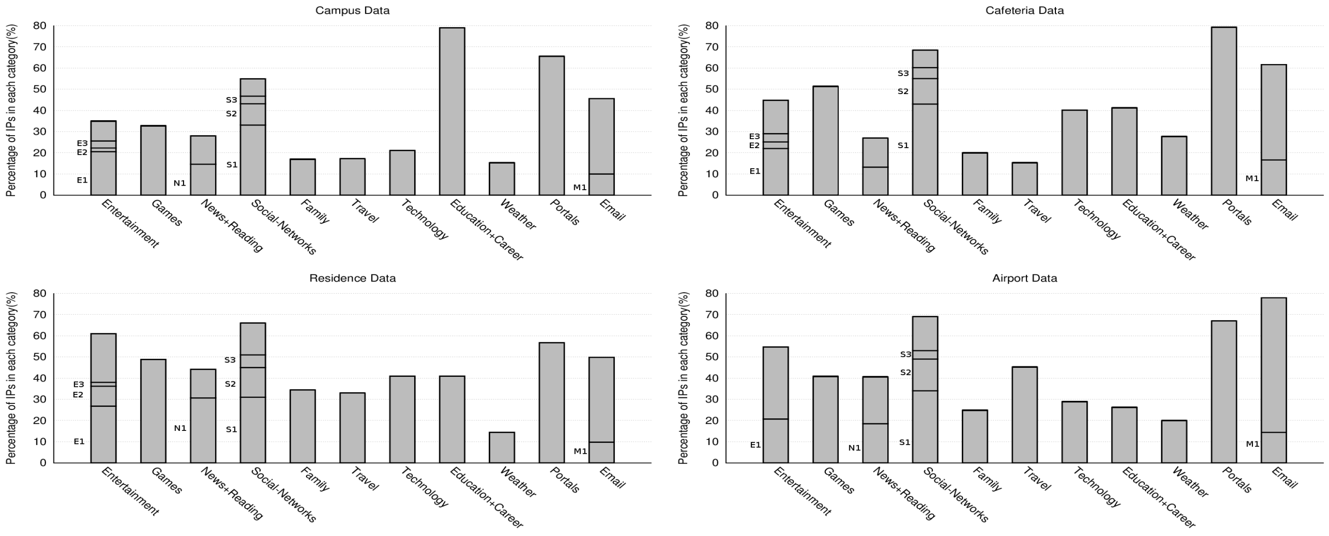


Fig. 4: Representation of interest categorization (E1: Youtube, E2: Netflix, E3: Pandora, N1: CNN, S1: Facebook, S2: Twitter, S3: Instagram, M1: Gmail)

around 50 keywords for each category - generated using the common words of the *Keyword Tool* from Google Adwords [16] collected over one week, for each of the categories. Based on this search, we used the percentage of packets for a particular IP that had a keyword-match in any category as the score of the IP for that category. Apart from the 21 categories, we also did the above analysis on 12 commonly used services and used the scores as attributes. The 33 attributes in this category, combined with 53 statistical features, result in 86 attributes, in total.

TABLE II: Application Categories and Services

Categories	Entertainment, Games, News-Reading, Finance, Weather, Social network, Sports, Education-Career, Email, Portals, Family, File-sharing, Technology, Food-Culture, Travel, Health, Fashion, Politics, Shopping, Automobiles, Science
Services	Youtube, Netflix, Pandora, Amazon, Craigslist, CNN, Twitter, Facebook, Instagram, ESPN, Gmail, Dropbox

The keyword search on the trace showed that in general, around 60-70% of the IP addresses could be profiled on the basis of interest category. A particular IP address is considered to be interested in a specific application category if there is at least one packet that gives a keyword-match for that category. However, we observed that when a particular IP address was profiled to be belonging to a certain application category there were substantially large count of packets for which there was a keyword match in the same category. Table II shows the list of categories and services used for as the features in this category and Table III shows a few keywords of some of the categories. Fig. 4 represents the percentages of IP addresses that were profiled to be interested in one specific category.

TABLE III: Categories and Keywords

Interest Category	Keywords
Entertainment	youtube, netflix, itunes, mp3, video, music
Games	zynga, xbox, games, puzzles, trivia, aws
News and Reading	nytimes, bbc, cnn, blogspot, news, magazine
Sports	espn, mlb, soccer, olympics, fifa, ncaa, nba
Social Networks	facebook, twitter, friends, social, plus.google
Travel	maps, expedia, airlines, tripadvisor, yelp
Technology	endgadget, cnet, bestbuy, techcrunch, gizmo
Education and Career	.edu, stackoverflow, github, courseera, school
Shopping	craigslist, amazon, ebay, target.com, groupon
Email	gmail, pop3, imap, smtp, hotmail, yahoomail

Interpretation of Application based Categorization: The residential traces have the highest interest percentage in entertainment. Apart from that, food, family, shopping, politics, fashion and automobiles have higher percentage with lower interest in mails and portals as compared to the other locations. Mail and portals are not accessed by users at their own homes as compared to outside, like at work or when on the go. Also access to file-sharing websites are mostly seen in apartment traces. Traces collected in a campus WiFi hotspot have a very high percentage of IPs interested in education related websites, portals and emails, as can be expected. Music, video and games are accessed much less in a campus environment as compared to the others. Results in Fig. 4 verify this claim.

Cafeteria and airport traces have very high number of IPs with interest in social-networks, portals and email. Outdoor locations are expected to have high percentage of users checking weather, as is observed in cafeteria and airport traces. There is a high number of IP addresses accessing travel related websites in the airport, as compared to other traces, which is an expected trend. Users interested in entertainment are much higher in apartment and cafeteria. Gaming websites or applications are found to be very high in the cafeteria trace (due to smart-phone games) and in apartments (due to dedicated gaming services, such as, xbox).

VI. PACL Prediction Model and Results

In this section we describe a model, created on the basis of the aforementioned features to efficiently predict users contextual location.

TABLE IV: Comparison of Prediction using Different Feature Subsets

Set of features	No. of Features	Correctly Classified Instances (%)	Size of Tree (Average)	TP Rate	FP Rate	ROC Area	Attributes with highest information gain
Coarse-Grain	7	1335 (76.2)	132	0.762	0.085	0.919	Activity period, Percentage ON time, Flow count, Concurrent flows
Protocol Based	4	1461 (83.4)	144	0.834	0.060	0.952	HTTP flow count, UDP flow count
Flow Level	26	1095 (62.5)	116	0.625	0.138	0.846	Flow length:max, Bytes per flow:mean, Bytes per flow:std. devn., Throughput per flow:mean, Flow length:min
Packet Level	14	1277 (72.9)	135	0.729	0.099	0.906	Packet size:min, Packet size:median, Packet inter-arrival time:max, Packet inter-arrival time: median
Application Based	19	952 (54.3)	107	0.543	0.173	0.774	Education and Career Emails, Portals, Games
Entire Set	**70**	**1527 (87.16)**	**101**	**0.872**	**0.046**	**0.978**	Activity Period, Flow length:max, Education and Career UDP flow count, Concurrent Flows

A. *Feature Selection*

Before creating the model for prediction, we need to identify the specific features that contribute towards differentiating between location categories. For this purpose, Chi-squared statistic evaluation [17] is applied to the 86 attributes and a score is assigned to each one of the features, which symbolizes the relation between the attribute and the class.

Chi-Squared Statistic: This statistic is used to evaluate the "distance" between the distribution of each class for an attribute. Initially, the values of an attribute are divided into separate intervals. Based on this division, the frequency of instances in each interval and class is calculated. Then the *Chi2* value is calcuated based on Equation 1 (with n=2) for each pair of sorted adjacent intervals to ascertain if the relative frequencies of the classes are similar enough to justify their merging. If the *Chi2* distance is smaller than a certain threshold for the pair, the intervals are merged. Merging continues till all adjacent pairs have a *Chi2* value greater than the threshold (20 in our case).

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \qquad (1)$$

- ○ A_{ij} = frequency of i^{th} interval and j^{th} class.
- ○ E_{ij} = expected frequency of A_{ij} = $\frac{R_i * C_j}{N}$
- ○ R_i = number of values in i^{th} interval = $\sum_{i=1}^{n} A_{ij}$
- ○ C_j = number of values in j^{th} class = $\sum_{j=1}^{k} A_{ij}$
- ○ k = number of classes
- ○ n = number of intervals
- ○ N = total number of values

At the end of this step, if an attribute has been merged into one interval then the attribute is considered irrelevant in representing the original data and hence has a *Chi2* value of 0. Otherwise, the score is calculated as per Equation 1. Fig. 5 represents the normalized Chi-squared statistic score of the statistical attributes based on a) coarse-grain features b) protocol-based features c) packet-based features and d) flow-based features. On the basis of the results, we remove 16 attributes from our data-set which end up with a score of zero and build our model for prediction based on the remaining 70 features.

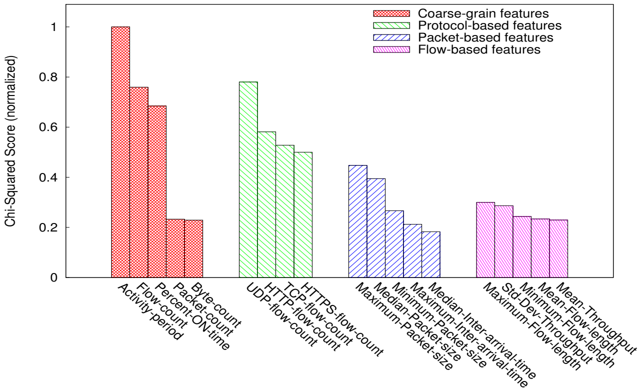


Fig. 5: Chi-Square statistic score for the highest-correlated features for each subset of statistical attributes.

B. *Model : Random Subspace with Decision Tree*

Predicting the location category from the statistical and application based features is non-trivial as many of the statistical features are dependent on each other and their interrelationship is non-linear. To address this issue, we use a machine learning approach to create the model involving these individual features. For this purpose we use Random Subspace algorithm. The algorithm implements a decision tree with reduced error pruning but also utilizes meta-learning on it. Due to non-linear nature of the attributes the most prevalent algorithm used is decision trees. Decision tree models employ simple if-then-else statements which predict classes efficiently and are also human readable. Another very important advantage is that they do not require the features to be independent among themselves.

Decision tree with reduced error pruning: The algorithm implements a C4.5 decision tree using the information gain ratio of different features. The information gain of an attribute is the expected reduction in entropy because of knowing the value of the attribute [18]. Attributes with higher information gain are likely to be more distinct among the classes, hence they are chosen first while building the decision tree from root to the leaves. The next step is the pruning of the tree. Reduced

TABLE V: PACL Prediction Results : TP and FP rate is calculated for one class against all other classes.

Location Class	TP Rate	FP Rate	Precision	ROC Area
Airport	0.961	0.029	0.922	0.996
Cafeteria	0.860	0.051	0.852	0.979
Campus	0.883	0.070	0.846	0.976
Residence	0.740	0.025	0.866	0.953
Combined Results	**0.872**	**0.046**	**0.872**	**0.978**

error pruning starts at the leaves and each node is replaced by the most popular class. If the accuracy of the prediction of the class is not altered then the change is kept and steps are repeated. Using the decision tree with pruning enables our model to run faster as the tree size reduces.

Meta-learning: The metalearning classifier consists of multiple trees constructed systematically by pseudo-randomly selecting subsets of the feature vector, that is, trees are constructed using random feature subsets. Then the decision of each tree on the data used for prediction is combined together by averaging the conditional probability of each class at the leaves [19].

C. PACL Prediction Accuracy

For prediction of location category, the representative features are extracted from an IP address. These features are then used as an input in the aforementioned model and a location category is predicted. To check the prediction accuracy of our model we divide the entire data set into *n*-folds and use *n*-1 folds for training and use the remaining one fold as test data to predict the location class. We repeat this step for the remaining *n*-1 sets of data. Here, we consider *n* = 10.

We measure the efficiency of prediction of the location classes on the basis of the following characteristics:

1) **True Positive Rate:** The fraction of instances correctly classified as class A, among all instances actually belonging to class A = $\frac{|TP|}{|TP|+|FN|}$, where TP = number of true positives and FN = number of false negatives.
2) **False Positive Rate:** The fraction of instances which were wrongly classified as class A, among all instances not belonging to class A = $\frac{|FP|}{|FP|+|TN|}$, where FP = number of false positives and TN = number of true negatives.
3) **Area under ROC Curve:** The Receiver Operating Characteristics curve (ROC) plots the variation of false positive rate vs. true positive rate for all the instances of the test data and for each class. The ideal ROC curve approaches the top left corner for 1 true positive rate and 0 false positive rate. The area under the ROC curve ($\in$ [0,1]) gives an estimate of the effectiveness of the prediction model. A perfect model has a ROC area of one.
4) **Precision:** The fraction of instances which actually belong to class A, among all classified as class A = $\frac{|TP|}{|TP|+|FP|}$.

The results of our model is presented in Table V along with the confusion matrix for prediction as shown in Table VI. Overall, our model predicts 1527 out of the 1752 instances correctly giving a prediction rate of 87.16%. The ROC curve for the 4 location categories are shown in Fig. 6b. The figure as well as Table V shows that the prediction is most effective for airport traces where as residence traces show least effectiveness. The exact ROC values are in Table V. Cafeteria and campus dataset show similar prediction efficiency.

TABLE VI: Confusion Matrix - Each element is represented as (x,y) where x is row number representing the number of IPs actually belonging to that class, and y is column number representing the number of IPs predicted in the corresponding class.

Classified Class	Airport	Cafeteria	Campus	Residence
Airport	**440**	6	6	6
Cafeteria	8	**387**	42	13
Campus	12	33	**467**	17
Residence	17	28	37	**233**

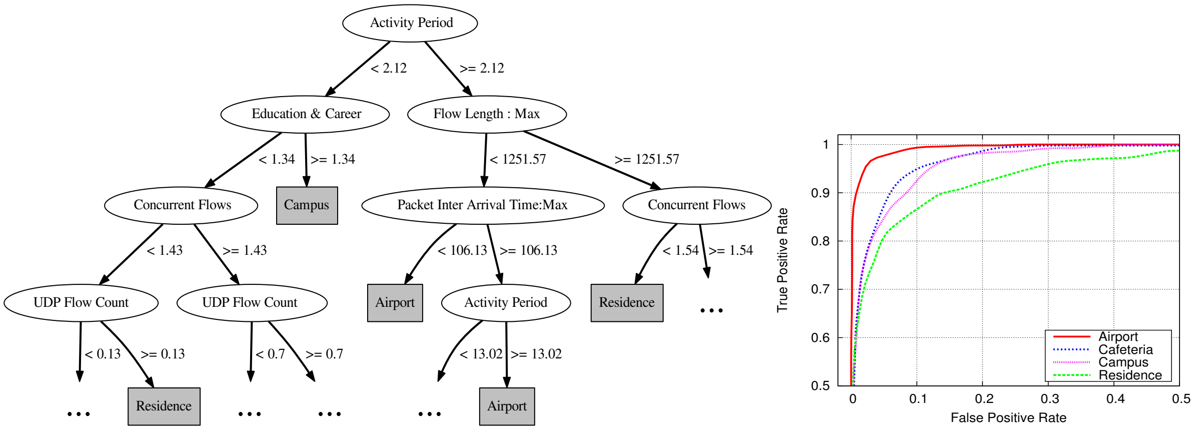
In Fig. 6a, we plot a pruned version of our decision tree model (built using all the features). The model shows that the attribute “activity period” has the highest information gain. Fig. 3e shows that the variation of activity period across different location classes is very distinct and hence activity period is most effective in distinguishing the location categories. Fig. 5 shows that this attribute has the highest Chi-squared statistic score. The nodes near the root of the tree includes attributes that belong to all the different subset of features, which shows that the combination of the features are required for efficient prediction.

D. Prediction Accuracy with Feature Subsets

We predict contextual location based on a number of features which are indicative of network usage patterns of various users. Combination of all features give a good prediction accuracy. But a question may arise as to how a certain subsets of features, calculated on the basis of a particular aspect of an IP address, contribute towards to the accuracy. Performance of the individual subsets of features using the same model and under the same experimental conditions is evaluated. The results for 4 sets of statistical features and the application based attributes mentioned in Section V and comparison with the overall results is shown in Table IV. The table also lists the attributes that have the highest information gain in each of the attribute subsets.

Extracting some of the features from the network traffic by an ISP is relatively easier and faster for some attributes compared to others. For example, coarse-grain statistics, like flow count, number of flows belonging to different protocols, packet count, activity period, etc., are easier to track, hence leading to faster prediction of the location category. It is observed from the results in Table IV that the coarse-grain and protocol based statistical features are most crucial in prediction among all the subsets. This is specifically important for real-time prediction.

In our analysis, the statistical features are calculated based on high-level statistics and header information. Payload infor-



(a) Decision tree model for prediction of contextual location based on training data of 1752 instances

(b) ROC curve for individual contextual location classes

Fig. 6: Decision tree and ROC curves for PACL prediction model

mation is used only in the categorization of application interest among users at various locations. Certain commercial tools [20] are available for extracting application based information systematically from the packet payload [21], more commonly known as Deep Packet Inspection (DPI). There are multiple issues with using DPI. First, most flows in modern day internet traffic are encrypted and hence cannot be decoded. Secondly, looking into the payload leads to privacy leakage issues from users' point of view. Thirdly, this procedure is resource and time intensive. Even though we have looked into payload for the application-based features, we have applied a keyword based search and did not look into the specific content accessed by users. An efficient tool to look into the content accessed by users might help us to distinguish between the applications better and in turn improve the result.

VII. Conclusions

In this paper, we present a model for prediction of users' contextual location by network traffic analysis. Using real world traces we train our model on the basis of of statistical and application-based features, to classify users' into four representative contextual locations. The PACL prediction model, in our test case, gives an accuracy of around 87%. There are multiple directions of future work. First, looking into the payload of packets is computationally expensive and as a result, we believe that the application based categorization has a scope for improvement. Next, the application of PACL to predict flash-mobs or events (short term gathering) is another scope of the work.

Acknowledgments

This work was supported in part by the NSF CNS-1251029 grant and the Intel Science and Technology Center for Secure Computing. The authors would also like to thank Yunze Zeng, Ningning Cheng, Chao Xu and Hao Fu for helping in data collection.

References

[1] "Do Not Track Me Online Act of 2011," http://www.gpo.gov/fdsys/pkg/BILLS-112hr654ih/pdf/BILLS-112hr654ih.pdf.

[2] I. Poese, B. Frank, G. Smaragdakis, S. Uhlig, A. Feldmann, and B. Maggs, "Enabling content-aware traffic engineering," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 5, pp. 21–28, Sep. 2012.

[3] I. Poese, B. Frank, S. Knight, N. Semmler, and G. Smaragdakis, "Padis emulator: an emulator to evaluate cdn-isp collaboration," in *ACM SIGCOMM 2012*.

[4] I. Constandache, S. Gaonkar, M. Sayler, R. Choudhury, and L. Cox, "Enloc: Energy-efficient localization for mobile phones," in *IEEE INFOCOMM*, 2009.

[5] K. Lin, A. Kansal, D. Lymberopoulos, and F. Zhao, "Energy-accuracy trade-off for continuous mobile device location," in *ACM MobiSys*, 2010.

[6] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *ACM GIS*, 2009.

[7] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *ACM MobiSys*, 2003.

[8] M. Damiani, C. Silvestri, and E. Bertino, "Fine-grained cloaking of sensitive positions in location-sharing applications," *Pervasive Computing, IEEE*, vol. 10, no. 4, pp. 64–72, 2011.

[9] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," in *Proceedings of the Third international conference on Pervasive Computing*, 2005.

[10] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices," in *ACM CoNEXT*, 2008.

[11] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and interests in a mobile 3g network," in *ACM SIGCOMM*, 2009.

[12] "Maxmind GeoIP," http://www.maxmind.com/en/home.

[13] "TP-Link TL-WN722N High Gain Wireless USB Adapter," http://www.tp-link.com/en/products/details/?model=TL-WN722N.

[14] "Ath9k - Atheros WLAN driver," http://wireless.kernel.org/en/users/Drivers/ath9k.

[15] N. Cheng, X. Wang, P. Mohapatra, and S. Aruna, "Characterizing privacy leakage of public wifi networks for users on travel," in *IEEE INFOCOM*, 2013.

[16] "Google Adwords: Online Keyword Tool," https://adwords.google.com/o/Targeting/Explorer?__c=1000000000&__u=1000000000&ideaRequestType=KEYWORD_IDEAS.

[17] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *International Conference on Tools with Artificial Intelligence*, 1995.

[18] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

[19] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832–844, 1998.

[20] "Packeteer," http://www.packeteer.com.

[21] T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, H. Kim, H. Chung, and T. Jeong, "Content-aware internet application traffic measurement and analysis," in *IEEE/IFIP NOMS*, 2004.