Characterizing Privacy Leakage of Public WiFi Networks for Users on Travel

Ningning Cheng[1], Xinlei (Oscar) Wang[1], Wei Cheng[1], Prasant Mohapatra[1], Aruna Seneviratne[2]
[1]Department of Computer Science, University of California, Davis, CA, US
[2]National ICT of Australia, ATP, Sydney, Australia
Email: [1]{nincheng, xlwang}@cs.ucdavis.edu
[1]weicheng@ucdavis.edu [1]prasant@cs.ucdavis.edu [2]Aruna.Seneviratne@nicta.com.au

***Abstract*—Deployment of public wireless access points (also known as public *hotspots*) and the prevalence of portable computing devices has made it more convenient for people on travel to access the Internet. On the other hand, it also generates large privacy concerns due to the open environment. However, most users are neglecting the privacy threats because currently there is no way for them to know to what extent their privacy is revealed. In this paper, we examine the privacy leakage in public *hotspots* from activities such as domain name querying, web browsing, search engine querying and online advertising. We discover that, from these activities multiple categories of user privacy can be leaked, such as identity privacy, location privacy, financial privacy, social privacy and personal privacy. We have collected real data from 20 airport datasets in four countries and discover that the privacy leakage can be up to 68%, which means two thirds of users on travel leak their private information while accessing the Internet at airports. Our results indicate that users are not fully aware of the privacy leakage they can encounter in the wireless environment, especially in public WiFi networks. This fact can urge network service providers and website designers to improve their service by developing better privacy preserving mechanisms.**

I. Introduction

The increasing deployment of public wireless access points (also known as *hotspots*) and the prevalence of portable computing devices such as tablets and smartphones have made it more convenient for people to access information on the Internet. Nowadays, businessmen, frequent travelers and people on vacation can easily access to a network from any public wireless access points. As of June 2012, there are 776,871 WiFi hotspots in 144 countries, according to JiWire's hotspot registry [1]. While public WiFi provides convenience and free access, it greatly compromises users' privacy, because with open access medium, anyone within the communication range might eavesdrop and interpret the traffic.

Currently, several protection techniques are available on 802.11 based wireless networks, such as securing the access by WEP, WPA or WPA2 encryption, using virtual private network (VPN) as a tunnel to the Internet and adopting captive portals for access controls [2]. However, none of these mechanisms have been particularly popular in public hotspot networks. Unlike enterprise networks, small business networks or home networks where network owners/users have the motivation to protect their information assets by encryption and are able to configure their networks, a public hotspot needs to be more open to provide any on-the-go users an easy connection to the Internet. To keep this openness nature, hotspot providers usually leave the networks open without applying any security strategies. Hence, users of public hotspots are responsible for protecting their own privacy [3]. Another security protection technique available is VPN, which can encrypt or hide data from being exposed in plain text. There are commercial VPN services (e.g., $OpenVPN$, $PRIVATEWiFi^{TM}$) that ensure packet encryption. However, such protection service providers generally charge a monthly fee, making it less accessible to users. Such problem also exists when using captive portal hosts, where users need to make extra payments for a secure network. Ordinary users may find it difficult to set up the security channel by themselves, hence end up skipping privacy preserving protections. In addition, users' private information can be utilized to generate revenue for online industries such as online advertisement. This incentive encourages Internet service providers collecting private information from users' online activities which further puts the users' privacy at stake.

Users can try more secured access networks such as 3G/4G networks to get more protections. However these networks are generally expensive to use. Also, users' cognition of privacy leakage is limited. For example, when mobile users have a choice of networking options (public WiFi and cellular networks), most of the time users tend to use public WiFi networks instead of 3G/4G cellular networks because WiFi is generally faster and less expensive. Users tend to neglect the privacy threats they face because currently there is no way for them to know to what extent their privacy is revealed in free public WiFi networks. Moreover, with current technology, operating systems for mobile devices make decisions on behalf of all applications in a one-fits-all solution. For example, iOS or Android system will choose WiFi networks over cellular networks when both of them exist, and there is no interface for users to configure each application's network access mode.

In this paper, we examine the potential privacy leakage in public hotspots from the user end activities such as web browsing, search engine querying and smartphone apps usage. We collect and analyze packets of users from fifteen different airports at different time and generate more than twenty datasets. We identify important network parameters that can be used to profile user's private information from open network traffic and characterize user privacy leakage based on their significance for different privacy categories. A reasoning engine is proposed to trigger different privacy protection mechanisms in order to tailor corresponding privacy requirements.

Understanding the privacy leakage of the hotspot networks has both technical impact and social impact. In terms of technical impact, it can incentivize better privacy protection mechanisms. For example, browser developers can offer a customized privacy protection interface to satisfy different privacy requirements; website developers can encrypt sensitive information such as health records and financial information; operating system developers can give users a personalized privacy protection scheme based on users' different privacy concerns. In terms of social impact, a comprehensive understanding of the privacy leakage can help people be aware of the privacy leakage problem that come about with the explosive growth of mobile technology in our daily lives, and hence reduce potential threats such as identity theft.

The rest of the paper is organized as follows, Section II gives our definition and categorization of privacy. Section III describes the potential privacy leakage. Section IV presents our reasoning engine model. Section V evaluates the privacy leakage in real traveling data. Section VI discusses the related work and Section VII concludes the paper.

II. Definition of user privacy

As Spiekermann and Cranor's definition in information technology content [4], a "privacy-friendly" system should be protected from three distinct layers: user sphere, recipient sphere and joint sphere. User sphere generates the data, recipient sphere receives the data and joint sphere hosts the data and provides services. In the public hotspots, user sphere refers to the communication devices of the travelers/customers who access the WiFi networks while on-the-go. Recipient sphere refers to the servers or databases that receive users' requests for certain information such as a webpage. Joint sphere refers to any third parties that involve in the communication process, such as the network providers that provide the WiFi service in the airports (or coffee shops) or the content delivery servers that store the website information.

Usually in hotspot networks, users are considered to be responsible for their own privacy protection by not accessing sensitive information in this open communication environment. However, this might not be enough. Our work focuses on detecting privacy leakage in the first sphere and tries to figure out reasons why leakage still happens. In addition to the user sphere, our work can also be related to the second and third sphere by showing how more secured design/management of the second and third sphere can help to reduce the privacy leakage in the first sphere.

In order to quantify privacy leakage, we define *privacy unit* as follows:

DEFINITION 1: **Privacy unit:** a piece of information that includes user privacy. It is the smallest unit used to measure an incidence of privacy leakage.

For example, "JD@gmail.com" is an email address with one privacy unit. But a more specified email address such as "Jane_Doe_19880203@ucdavis.edu" includes five pieces of privacy information: the user's name, email address, gender, date of birth and organization, and hence has five privacy units.

The discovery of privacy unit is highly dependent on what kind of information is considered to be private. In this paper, we categorize privacy into two types: user privacy and infrastructure privacy.

User privacy relates to user information such as identity, name, address, job, and interests. We further categories them into five sub-categories: identity privacy, location privacy, financial privacy, social privacy and personal privacy.

Identity privacy refers to a person's name, SSN, driver license number and other information that can identify who the person is.

Location privacy includes a user's location traces, such as where he is, where he has been, and what place he frequently goes to.

Financial privacy is a person's financial habits or condition, such as his online transactions, the merchandize he recently browsed, his stocks and other financial related information.

Social privacy includes a user's social information such as relationship and intimacy with his friends, family members, colleagues, or club members.

Personal privacy is the kind of information that can reflect a user's personal traits. For example, his hometown, medical conditions, marriage status, habits and hobbies, sexual orientation, political views, personality and other personal information.

Infrastructure privacy includes device identification, access points, service plan, operation system and other information related to the communication infrastructure.

User privacy as well as infrastructure privacy could be released through user actions. Based on the definition of privacy unit and categorization of privacy, we examine the privacy exposure by collecting the in-the-air traffic in public WiFi networks.

III. Privacy leakage in WiFi networks

A. Privacy leakage detection

Due to the open air nature of WiFi networks, it is easy to eavesdrop users within their communication range as long as they are in the same WiFi channel. The packets being sniffed can include network parameters such as MAC address, IP address, MAC layer flags, IP layer flags, protocol names, protocol fields and the content information in the payload if not encrypted. After doing a deep packet inspection, it is possible to study how many user privacy units can be leaked from various network parameters. In order to do this, we conduct different kinds of network activities such as turning on the WiFi interface, accessing the Internet and surfing different websites. Then we examine each parameter on different network layers and list those that can be used to infer users' information.

B. Sources of privacy leakage

Although the privacy leakage is detected at the user end, the source of the private information can come from not only the user sphere, but also the recipient sphere and the joint sphere. To be more specific, privacy leakage can be traced back to three types of sources: users' devices, website content and profiled advertisements. Here, we give three examples that reveal users' privacy from different sources:

Scenario one: name resolution in multicast. In this scenario, consider a traveler Ginger, who brings an iPhone named "Ginger's iphone" with her. When it sends out a domain name query by the Multicast DNS (mDNS) protocol with the device name "Ginger's iphone.local" in the mDNS query (Figure 1), in this case, it is easy to infer that the user names her device by her own name. Combined with her IP address, which is also included in the same frame, we can link any communication activities with the person named "Ginger". Other than the

```
Domain Name System (query)
Flags: 0x0000 (Standard query)
Queries
  GINGERs-iPhone.local: type ANY, class IN, "QU" question
```

Fig. 1. MDNS leaks hostname of user

```
Link-local Multicast Name Resolution (query)
Flags: 0x0000 (Standard query)
Queries
   Chirag-PC: type ANY, class IN
```

Fig. 2. LLMNS leaks host name of user

mDNS protocol, the Link-local Multicast Name Resolution (LLMNR) periodically sends out query messages like “Chriag-PC” (Figure 2), letting anyone in the communication range know the username of her device.

Scenario two: content in the HTTP conversation. A traveler viewing an Australian news website such as “$http$: $//www.smh.com.au/$” may indicate that he/she comes from Australia, and hence reveals his/her previous location or his/her home country. Website content can reveal more than just location or nationality information. For example, Amazon gives information about users’ shopping interests. Youtube shows people’s music interests and financial sites such as Bank Of America or Chase indicate that user has a bank account at that bank. All these personal information can be generated from the communication traffic as long as that user surfs the website and clicks the link.

Scenario three: profiled advertisement. Other than web content providers, there are third party servers such as advertisement aggregaters that use users’ previous browsing history to profile users and send relevant ads back to them. In this case, we can infer users’ personal information by the ads sent from the third party servers. For example, a user receives ads from “$http$: $//www.pgatour.com/$” implies that he/she is interested in golf. Advertisements of a golf club in a specific city may even reveal the user’s location. It is reasonable to assume that a user who receives a golf club ad from a club located in San Francisco (SF), may also live in SF as well. A shaver ad means that the user is most likely to be male. Cosmetics ads on the other hand are more likely to be profiled for a female user.

C. Leakage from users

From the user’s device, privacy is usually leaked from network protocols that include user privacy such as their device names (which are possible the users’ own names), email addresses, or infrastructure privacy such as network names (SSIDs), MAC addresses and network providers.

At the 802.11 MAC layer, one of the parameters that can reveal a user’s private information is SSID. Whenever a WiFi interface is activated, by default, the system will broadcast a list of its previous accessed networks’ SSIDs. In this case, it is possible to leak the user’s company, frequently visited places and other information if their names are included in the SSIDs. SSIDs can also show the devices’ service providers. For example, devices probing for a network named “attwifi” is known as AT&T’s hotspot network name and “t-mobile” is known as T-mobile’s hotspot name. The MAC address named “Recipient address” in Acknowledge (ACK) frames shows the device has just sent some data to the AP.

At the 802.11 application layer, different protocols have different parameters. Devices that use multicast DNS protocol need to announce themselves by their host names when attached to the hotspot network. As scenario one shows, the user’s name will be revealed in these protocols if a user names her device by her own name. For example, iPhones/iPads send out standard mDNS queries containing its device name. Microsoft devices use LLMNR protocol which also contains devices’ names in the query.

Another type of protocol containing privacy units is email related protocol, such as POP3, SMTP and HTTP email protocols, other than email address, it may also reveal a user’s name, age, work affiliation when the user put these pieces of information in his/her account name.

D. Leakage from websites

Website information is generally derived from DNS packets and HTTP packets. For example, the URL of a website can be generated by combining the host, directory and file name of the HTTP header. The domain name and IP address in a DNS packet reflect the location and country of the website. Last but not least, information in the website content can generate implication of identity privacy, location privacy, financial privacy, social privacy and personal privacy.

User queries in popular search engines like Google or Bing reflect private information, too. Especially when users search for sensitive key words such as those related to their medical conditions.

In order to test how many user privacy units can be leaked (even though users are only surfing a regular website without inputting sensitive information), we collect popular websites’ traffic from different categories and investigate their privacy units on different privacy categories. In order to characterize the personal privacy leakage in these scenarios, we examine over 50 popular websites with their sublinks and list top 5 widely adopted third party advertisers for packet analysis. We target our study on popular websites, such as Google, Yahoo, Amazon and other top five websites from different areas such as health, politics and shopping websites regarding their traffic flows statistics given by Alexa [5]. We investigate the leakage of different aspects of users’ private information. A detailed illustration of popular websites’ leakage condition is shown in Table I, where “full content” means the whole website can be exposed by concatenating the “host”, “directory” and “filename” in the HTTP header fields.

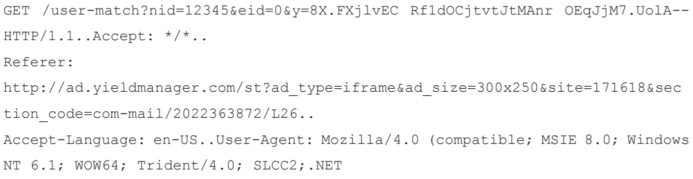
E. Leakage from third party advertisers and aggregators

Third party advertisers conduct Online Behavioral Advertising (OBA) to deliver advertisements tailored for users. Figure 3 gives an example of the advertisements sent by HTTP protocol. By looking at the content of the packets sent from the advertisers, it is possible to infer users’ private information based on profiled advertisements.

In this part, we focus on detecting profiled advertisements in the top websites. Therefore, we can discover the most frequently accessed third-party advertisers that profile user online activities. The method is as follows. each time we open

TABLE I
USER PRIVACY LEAKAGE IN POPULAR WEBSITES

Type	Website	Leaked Info	Type	Website	Leaked Info
Overall	Google	query string	Health	National Institutes of Health	website content
	Facebook	profile photoes		WebMD	full content
	Youtube	search query, vedio content		PubMed	full content
	Yahoo	full content		MayoClinic	website name
	Wikipedia	full content, query string		Mercola	full content
Politics	Slate Magazine	full content	Shopping	eBay	item viewed
	NewsMax	full content		Netflix	sign up page
	Infowars	full content		Wal-Mart Online	full content
	Salon	full content		Groupon	location
	Daily Kos	login page		Amazon	recently viewed
Sport	ESPN	full content	News	CNN	full content
	Yahoo Sports	full content		New York Times	full content
	ESPN Cricinfo	full content		Google News	website name
	NBA	full content		The Weather Channel	full content
	MLB	full content		Reddit	website name
Travel	Booking.com	query string, location	Religion	Bible Gateway	full content
	TripAdvisor	query string, location		Church of Jesus Christ of Latter-day Saints	full content
	XE	website name		Astrodienst	full content
	Expedia	website name		Online Parallel Bible	query string
	Universal Currency Converter	website name		FamilySearch	nothing(https)
Sex orientation	The Advocate	full content	Financial	PayPal	nothing(https)
	Towleroad	full content		Wells Fargo	nothing(https)
	Queerty	full content		American Express	nothing(https)
	AfterEllen	full content		Bank of America	nothing(https)
	Gay.com	full content		Chase	nothing(https)



```
GET /user-match?nid=12345&eid=0&y=8X.FXjlvEC Rf1dOCjtvtJtMAnr OEqJjM7.UolA--
HTTP/1.1..Accept: */*..
Referer:
http://ad.yieldmanager.com/st?ad_type=iframe&ad_size=300x250&site=171618&sec
tion_code=com-mail/2022363872/L26..
Accept-Language: en-US..User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows
NT 6.1; WOW64; Trident/4.0; SLCC2;.NET
```

Fig. 3. Link from third party advertisement

a popular webpage, if the advertisement on it is user-tailored, we click it to generate a traffic to the advertising server. An example of privacy leakage from the top five advertising servers are listed in Table II, where servers are ordered by their popularity (based on the total number of clicks from the advertiser).

TABLE II
EXAMPLES OF PRIVACY LEAKAGE IN AD CONTENTS

	Privacy information				
Advertisers	Identity	Location	Finance	Social	Personal
Doubleclick	Profiled	Disneyland	Bank	Veterans	News
2mdn	×	Spa	Bank	×	Car
Atdmt	×	Hospital	Bank	×	HP toner
Google -syndication	×	×	Bank	×	Politics
Google -analytics	×	Groupon	Bank	×	Insurance

F. Infrastructure privacy leakage vs. user privacy leakage

After inspection of the packets, we discover there are different implications between infrastructure privacy leakage and user privacy leakage. Infrastructure privacy such as MAC addresses, device IDs, network SSIDs, operation systems and other information is mostly used to link two traffic flows, or two devices to the same user. For example, we can target a specific user's conversation sessions by filtering out its MAC address, or link two devices to the same user with a unique SSID that only belongs to this user.

For infrastructure privacy leakage, it is easy to detect specified privacy leakage. As long as the network parameter exists, the correspondent privacy leakage will be detected. For user privacy leakage, it will be more complicated. The type of privacy leakage will depend on the content of the website, users' queries in search engines and the advertisements they received.

In order to study the inference of user privacy from network parameters, we generate a coherence/dependance table (Table III) between network parameters and privacy categories. This table includes the network parameters we examined in traffic examiners (Wireshark and NetWitness) and the user privacy we detected from the 50 popular websites in Table I. It illustrates the relationship between privacy leakage and the network parameters.

In the next step, a reasoning engine is proposed as a privacy leakage detection mechanism when certain network parameters exist in a traffic. The core of the engine is the privacy inference rules that are designed based on dependencies of Table III. As an output of the reasoning engine, it triggers privacy protection actions to protect user privacy.

IV. REASONING ENGINE

We propose a rule-based reasoning engine which uses network parameters as inputs, deducts the privacy units leaked using existing parameters and triggers privacy-preserving actions.

The reasoning engine is shown in Figure 4. It has following main components: a *knowledge database*, *leakage detection rules* and an *alert processor*.

A. Overview

The procedures in the reasoning engine is shown in Figure 4. It takes network traffics as the input, and checks the network

TABLE III
POTENTIAL PRIVACY LEAKAGE IN DIFFERENT NETWORK PARAMETERS

Privacy units		Network parameters								
		MAC	IP	mDNS host name	Website	SSID	Domain name	Applications	Email	Ads content
Identity privacy	Name	×	×	direct	c.d.[1]	indirect		c.d.	direct	c.d.
	SSN	×	×	×	c.d.	×	×	c.d.	×	c.d.
	Driver licence	×	×	×	c.d.	×	×	c.d.	×	c.d.
Location privacy	Previous	×	×	×	indirect	indirect	×	c.d.	×	c.d.
	Most often	×	×	×	c.d.	indirect	×	c.d.	×	c.d.
	Most recent	×	×	×	c.d.	indirect	indirect	c.d.	×	c.d.
Financial privacy	Previous transaction	×	×	×	direct	×	indirect	c.d.	×	c.d.
	Interested merchandize	×	×	×	direct	×	×	c.d.	×	c.d.
	Stock	×	×	×	direct	×	×	c.d.	×	c.d.
	Bank info	×	×	×	indirect	×	×	c.d.	×	c.d.
Social privacy	Family member	×	×	×	indirect	×	×	c.d.	×	c.d.
	Friends	×	×	×	indirect	×	×	c.d.	×	c.d.
	Other social group	×	×	×	direct	×	×	c.d.	×	c.d.
	Intimacy of relationship	×	×	×	c.d.	×	×	c.d.	×	c.d.
Personal privacy	Hometown	×	×	×	indirect	×	indirect	c.d.	×	c.d.
	Marriage status	×	×	×	c.d.	×	×	c.d.	×	c.d.
	Medical condition	×	×	×	indirect	×	×	c.d.	×	c.d.
	Hobbies and habits	×	×	×	direct	×	×	c.d.	×	c.d.
	Organization	×	×	×	direct	×	×	c.d.	direct	c.d.
	Religion	×	×	×	direct	×	×	c.d.	×	c.d.
	Political view	×	×	×	indirect	×	×	c.d.	×	c.d.
	Sexual orientation	×	×	×	direct	×	×	c.d.	×	c.d.
	Personality	×	×	×	c.d.	×	×	c.d.	×	c.d.

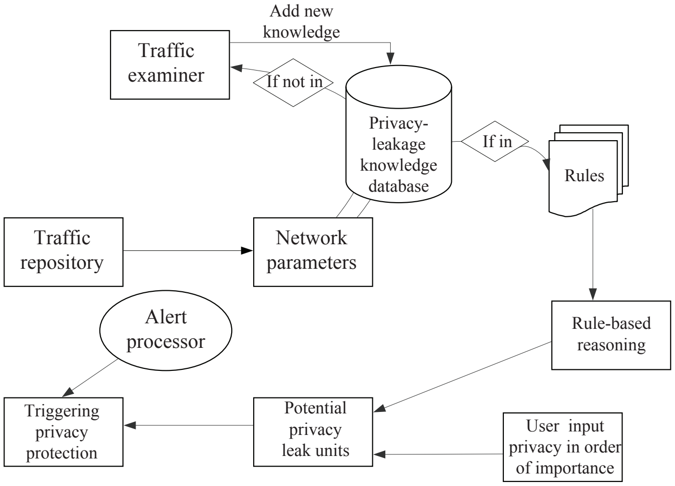

Add new knowledge
Traffic examiner
If not in
Privacy-leakage knowledge database
If in
Rules
Traffic repository
Network parameters
Alert processor
Rule-based reasoning
Triggering privacy protection
Potential privacy leak units
User input privacy in order of importance

Fig. 4. Procedures in the reasoning engine

parameters in the traffic. If a parameter has been analyzed and collected in the privacy-leakage knowledge database, it will be processed with privacy inference rules and conflict resolving rules (if needed). The privacy units leakage is deducted based on the rules. By combining privacy units being leaked and a user input partial order, the seriousness of the privacy leakage is measured. When the leakage is over a certain threshold, the alert processor will notify the user by alarms and trigger a privacy protection mechanism of the system.

B. Knowledge database

The knowledge database keeps the inference relationships about the protocol or website content and privacy information being exposed. In the privacy leakage knowledge database, knowledge is generated in a way similar to the tables shown in Table I and Table III. If the relation between a network parameter and a privacy unit is "direct", it means the network parameter contains the privacy unit. The knowledge is expressed as:

- <mDNS.host name: name→name>
- <email: name→name>
- <email:organization→organization>
- <website:religion→religion>

If the relation is "indirect", it means although the network parameter does not include the privacy unit explicitly, it can infer the privacy unit implicitly. For example, indirect knowledge can be:

- <website: language→home country>
- <website: profile photo→gender>
- <search: medicine→medical condition>

A "c.d." relation is short for "condition-dependant" relation, it means the privacy unit is exposed only under certain condition. For example:

- <website has weather report: location→previous location>
- <website is a shopping site: item listed→previous viewed item>
- <ads is GoogleAds: content→hobby>
- <ads is GoogleAds: address→location>

All the relation between network parameters and privacy units is stored in the privacy leakage knowledge database, accessed by the reasoning rules to deduct potential privacy leakage.

C. Rules

The privacy detection rules generate users' privacy information by a list of deduction rules and the facts in the knowledge database.

The following inference rules give an example of a simple inference based on knowledge database:

Privacy inference rules:

[1] Here "c.d." is short for "condition-dependant".

- if x(MAC), y(email:name) then x(name: y)
- if x(MAC), z(email.organization) then x(works at: z)

It can be concluded that x, named as y, is working at z, simply from the facts in the knowledge database.

However, the rules can conflict with each other given more network parameters. For example,

- if x(MAC), Amazon(web:host), cosmetics(web: items viewed) then x(gender: female)
- if x(MAC), facebook(web:host), (profile photo: male) then x(gender: male)

In this case, we need to resolve the conflicts.

Conflict resolving rules: Given a set of privacy inference rules which conflict with each other, we follow next three steps to make a final decision:

- Step 1: Record all conflict rules.
- Step 2: Apply a *majority rule* to the conflict rules concerning the same subject. In the previous example, if there is another rule resulting in deducting x as male, then combining three rules, x is decided to be male.
- Step 3: When the *majority rule* cannot resolve the conflicts, give higher priority to the rules that generate less conflicts. For example, if inference rules based on "items viewed" has generate 3 more conflicts for other subjects/users, and inference rules based on profile photos only has one conflict. Then x is decided to be male according to the less conflicting rule.

After performing reasoning rules on the network parameters, we are able to determine what type of privacy information is leaked. However, the users might also want to know how serious their privacy leakage is. Since the seriousness of privacy leakage is largely dependent on the user's attitude towards privacy, in our reasoning engine, we accept user inputs to decide the seriousness of privacy leakage, based on an order of the privacy units given by the users concerning their importance.

D. User input

User input is a partial order set of privacy units $< P, \prec >$ from the least sensitive to the most sensitive information. The seriousness of the privacy leakage is based on both the de facto leakage and the partial order given by the users.

DEFINITION 2: **Seriousness of privacy leakage, S:**

Let

$< P|p_i \epsilon P >$ be the partial order privacy set given by the user,

$< L|l_i \epsilon L >$ be the leakage set that deducted from the rules ($l_i = 1$ if p_i is leaked, $l_i = 0$ if p_i is not leaked),

$< W|w_i \epsilon W >$ be the percentage weight assigned to each privacy unit in P,

Seriousness of privacy leakage $S(P, L) = \sum w_i \cdot p_i \cdot l_i$

E. Alert processor

Alert processor is used to trigger privacy a protection mechanism to prevent the sensitive privacy from being leaked when the *seriousness of privacy leakage* exceeds a threshold. The privacy protection mechanism can be alerting alarms, switching interfaces, or sending popup messages to users.

TABLE IV
NUMBER OF DEVICES IN THE AIRPORT

Airport & date	**Time**	**# of devices**
IAD0509	12.24-12.39	160
IAD0511	12.40-12.55	284
IAD0511	13.04-13.55	273
SMF0515	18.45-18.58	47
IAD0517	12.15-12.45	113
IAD0517	13.04-13.19	223
PHL0525	17.00-17.12	354
SMF0526	16.07-16.35	29
SFO0526	20.51-22.08	32
SYD0531	17.45-18.33	392
LAX0609	13.38-13.53	143
FRA0610	2.09-2.29	66
DEL0622	3.48-4.16	12
DEL0622	8.43-8.57	15
EWR0623	2.04-2.17	8
LAS0630	21.06-21.48	155
SFO0707	21.17-21.32	34

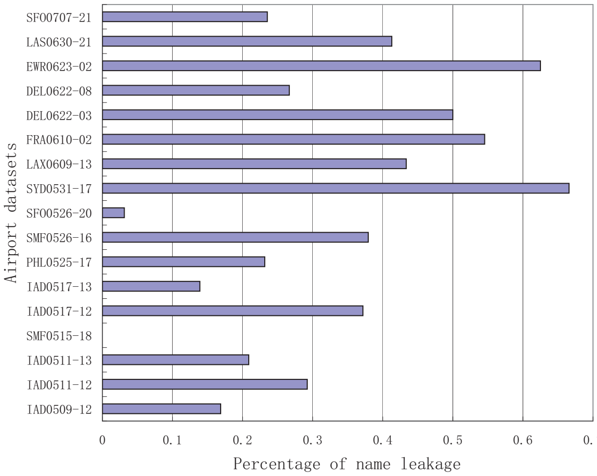


Fig. 5. Percentage of names leaked in multicast DNS protocol

V. STATISTICS FROM REAL WORLD EVALUATION

To evaluate the privacy leakage in real-world traveling scenarios, we collect over 20 airport traffic datasets from 15 airport in four countries starting from May 9th to July 7th. Packets are collected by the traffic monitor software "Wireshark" on a Windows PC with an AirPcap Nx wireless adapter. It logs all the traffic within the communication range to a dataset named by the airport name, date, time and channel number. The time of the data collection varies from 15 minutes to 60 minutes. The datasets include more than 1 million packets and over 150,000 traffic sessions. To examine the privacy leakage in each airport entry, we count the private units leaked in different privacy categories. Next, we will discuss the statistical results based on different privacy leakage sources.

A. User name leakage

According to our dataset, the most conspicuous user privacy being leaked at user end is a person's name, because users like to name their devices with their own names. Over the 2000 unique devices we detected, more than 600 device names contain their owner's names.

In order to find the total device number, we record each device by its MAC address, then the devices that send beacon frames are filtered out because they are obviously access points. This way, the remaining addresses belong to travelers' devices.

We discover that 3 of the 20 airports' dataset do not contain any user devices. The number of devices of the remaining 17 airports are shown in Table IV.

Figure 5 shows the percentage of name leakage in each airport dataset. According to this result, we can find out the user name leakage can be up to 68% in some airport. Because people do not know their mobile devices' names are frequently broadcasted by network protocols such as mDNS and LLMNR queries, they are potentially leaking this private information to anyone within their communication range.

B. Website content leakage

Of all the 625 user names we detected in the datasets, 587 of them are leaked from the mDNS and LLMNR messages. Remaining 38 are detected from the website contents. Since most of the websites store their resource in files, the website content can be reached by combining the host URL, directory and file name in the HTTP protocols, revealing all the information in the website the user is browsing.

To evaluate the privacy leakage in website contents, we first study the users' preference of websites when they are on travel. Statistical result of the website access is shown in Figure 6. It shows the number of users accessing the website in the total airport datasets. The statistical result of privacy information leaked by these websites is shown in Figure 7, where the leaked privacy units are the privacy leakage discovered in all the airport datasets. We can see that the popularity of websites such as Google and Facebook for travelers are mostly consistent with their overall popularity, and the business websites (such as stock and investment websites) and news websites (such as CNN, NPR and NYTimes) are more important for travelers than ordinary users.

Detailed distribution of users accessing different websites on each airport dataset is shown in Figure 8, where the number of users accessing a website is normalized by the total number of users in the airport. In the figure, the search query content is from the online search engine website such as Google, Bing and AOL. Hobby is inferred from the website content. Location is from the file directories of the website resources. For example, NYTimes and Craigslist organize content in directory named by cities. An image folders named by a city in Flicker implies that user has been to the city before and stored the photos in this folder. Other private information such as merchandize, health information, social groups of interest can also be inferred from website contents. Personal photos stored in social networks' photo servers, such as "$profile.ak.fbcdn.net$" reveals the user profile photos on Facebook. Pictures of the merchandize that the user previously browsed in Amazon is also an indication of shopping interest.

C. Advertisement and application leakage

Figure 9 shows the number of users accessing the advertisement servers or the application servers in each dataset.

Most of the advertisement and smartphone apps use HTTP protocol for packet transmission. For third-party advertisers, the referrer's link and the content of the ad may contain personal information which has been tailored to match the user's profile. Similar to website examiner, ad contents can also be examined by its host name, directory and file name. Also some advertisers such as Doubleclick obfuscates the referrer's URL by stuffing it with characters "$2\%F$", it is possible to figure out the original URL by substitute these character with subfolder mark "\". Table II gives an example of different type of privacy unit inferred by the ad's content.

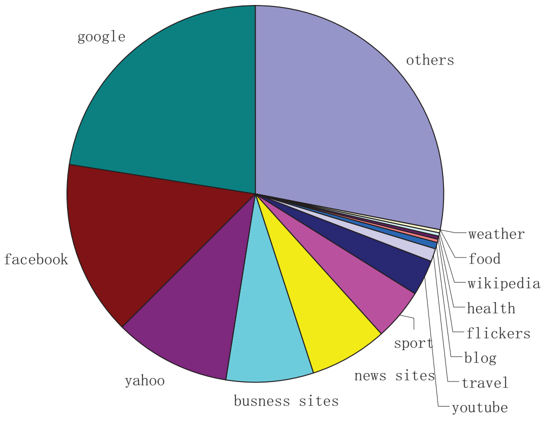


Fig. 6. Websites being accessed in the airport datasets

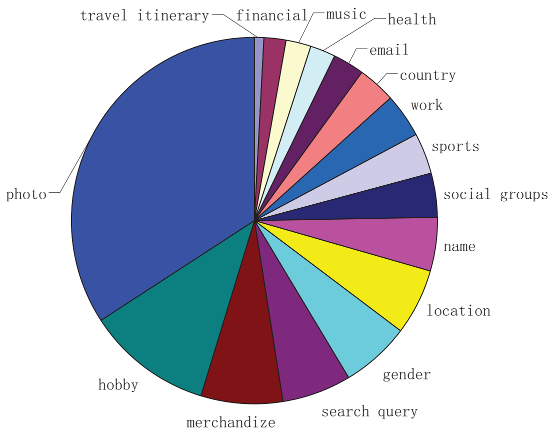


Fig. 7. Privacy units leaked by website access in the airport datasets

The apps installed on smartphones can be implied by

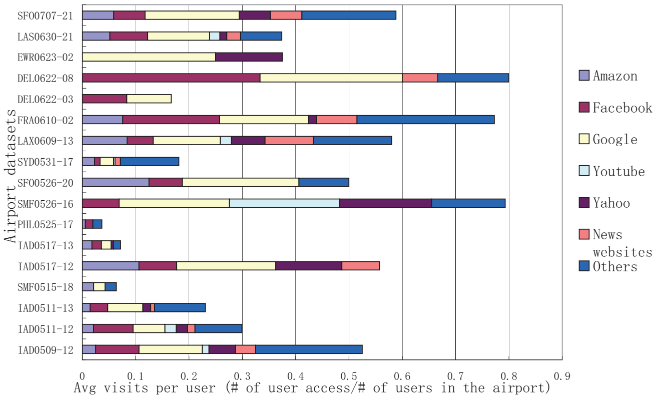


Fig. 8. Websites accessed in each airport dataset

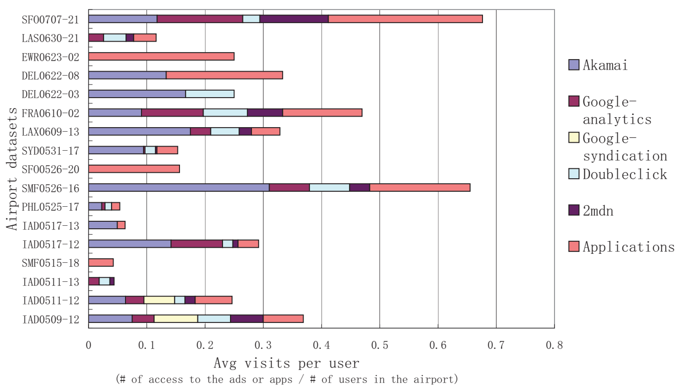
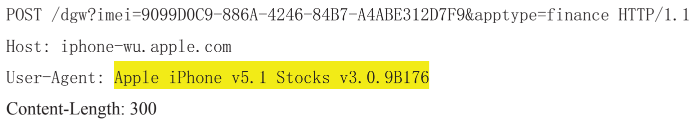


Fig. 9. Advertisement accessed and apps detected in the airport in each dataset



```
POST /dgw?imei=9099D0C9-886A-4246-84B7-A4ABE312D7F9&apptype=finance HTTP/1.1
Host: iphone-wu.apple.com
User-Agent: Apple iPhone v5.1 Stocks v3.0.9B176
Content-Length: 300
```

Fig. 10. iPhone revealing its stock app when it is in use

the domain name queries. For example, a device querying for "*api.twitter.com*" or "*api.facebook.com*" implies the smartphone is installed with Twitter or Facebook apps. A smartphone may also include its apps name by itself when communicating with the application servers. Figure 10 gives an example where an iPhone reveals the version of its Stock app when posting a message to the application server.

D. Overall privacy leakage

The overall statistical results of users privacy leakage distribution in each dataset are shown in Figure 11 and Figure 12, where x-axis is the number of privacy unit in the dataset (normalized by the total number of the user), and y-axis is the date and airport we collected the dataset. Each bar represents the type of privacy information can be discovered in the dataset. The privacy information can be generated based on different sources of network parameters as shown in Figure 13.

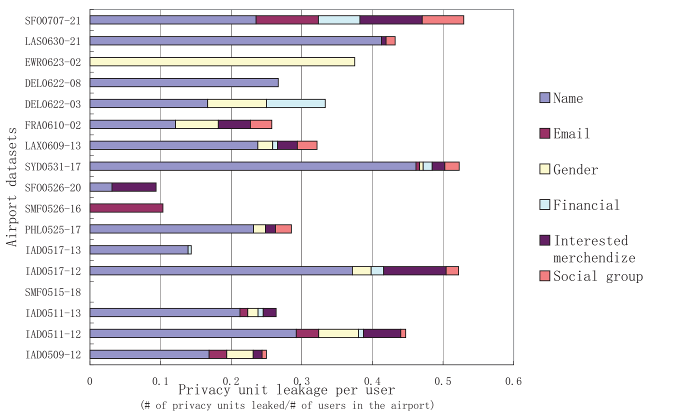


Fig. 11. Privacy units leaked in the airport (regarding name, gender, financial, interested merchandize and social group)

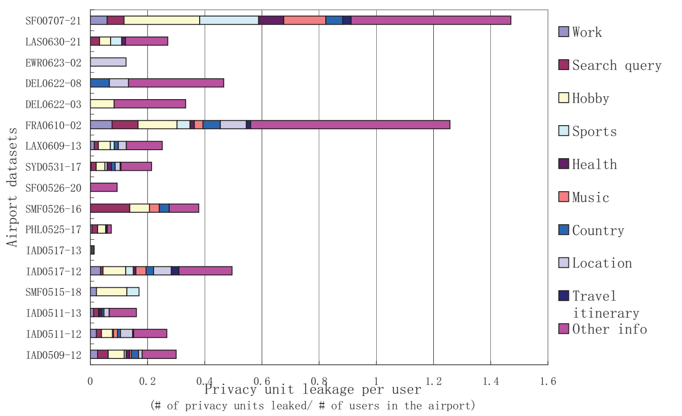


Fig. 12. Privacy units leaked in the airport (regarding work, search query, hobby, sports, health, music, home country, location, travel itinerary and other information)

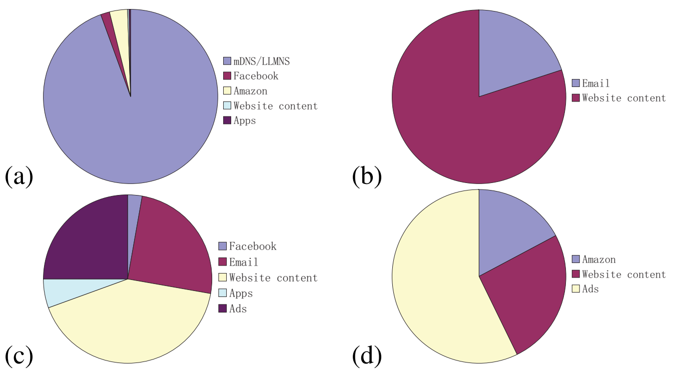


Fig. 13. Privacy units based on different network parameters(a).Name from mDNS, facebook, email and website content (b).Home country from email and web content (c).Location from facebook, email, web content, apps and ads (d).Interested merchandize from web content and ads

From the figures above, we can see that in public hotpots, privacy is not only leaked by users themself, but also by protocol designers (multicasting device name in the mDNS service), website hosts (coupling user privacies even when not necessary), advertisement brokers (profile based on users' browsing history) and application service providers (leak personal information).

E. User profiling with leaked privacy data

The personal information we get can also be used to profile user. These information may include user's device name, gender, age, location and other personal information. When combining these information with social network platforms, such as Facebook, LikedIn and Twitter, it is possible to identify some users and their account on the social network. Next, we give an example how we use this information to identify a user. In order to protect user anonymous, we substitute user's private information with symbol characters.

In the DNS queries list, we discover a query for url *www.waterheuvel.nl*. Since the website content is in Dutch and the country code is "nl", representing Netherlands, it can be inferred that the device owner might come from Netherlands. Next, we found that the query is sent from a device with MAC address "AA:BB:CC:DD:EE:FF". So in the next step, we search for "AA:BB:CC:DD:EE:FF" in the multicast DNS query list, where a respond with hostname "John Doe" is found. Now the name of the device owner is revealed. In the next step, we search the device owner's name in popular social websites such as Facebook, LinkedIn and Twitter. We discover five persons with same name in different country, of which two live in Netherlands according to "LinkedIn". In this case, we narrow down the device user to two candidates and both of their companies are revealed.

F. Strategies against privacy leakage

After previous study on the privacy leakage in public airports, we propose several possible easy strategies for reducing privacy leakage.

- At the user end, since user name is the most frequently leaked privacy unit from user devices, users should avoid using their name as device name.
- As network service providers, airports can adopt encryption based communication mechanism such as WPA. In order to give easy Internet access to travelers, they can put the password of the network on each boarding card, so travelers can access the eavesdropping-free network.
- Another strategy is that when accessing the network, users can change to a more secured network such as their 3G/4G network when opening sensitive webpages or querying sensitive keywords in the search engine.
- For website designers, they should use HTTPS protocol instead of HTTP protocol when their webpage includes privacy sensitive information or privacy sensitive advertisements.

VI. Related Work

Privacy leakage in traditional online social networks (OSN) has been widely studied such as [6]–[10]. These literature mainly focus on the privacy issues in social networks based on the user published data, such as identifying user relationship and characterizing user patterns. After the prevalence of content delivery networks (CDN) and advertisement networks, re-identification [11] in third-party aggregators becomes another privacy concern. It is possible to aggregate privacy information sent from different websites and characterize the linkable property to profile specific users on the third party servers [12], [13]. In order to overcome this problem, various privacy control mechanism for third party aggregators are studied [14]–[16].

Different from previous privacy analysis, our work does not rely on social network platforms to detect user privacy. The personal private information is tied with users by examining the communication traffics in the air, which serves as a bridge between social networks and communication networks.

Identity trial detection based on DNS is introduced in [17]. It reveals a location privacy infringe of mobile users when the user broadcasts dynamic DNS updates with her mobile IPs including geolocation information. The method needs the victim's DNS host name to perform a long-time monitoring to profile targeted users' location pattern, hence does not applied to travelers' location detection.

Privacy leakage in web searches has also been studied [18], [19]. In order to prevent such leakage, query obfuscation such as TrackMeNot [20] and network mixing mechanism is adopted such as Mix [21] and Tor [22].

VII. Conclusion

We have made an attempt to characterize the leakage of various privacy aspects in public WiFi networks, especially for users on travel at airports. We collected and analyzed packet traces from 20 different airport datasets from four different countries. Several intriguing privacy parameters are shown that can be sniffed or deduced from publicly available data in clear text. The results are quite alarming in the sense of the quantification of information that can be leaked while accessing public WiFi networks (without making much of an effort). The next step would be to develop techniques to safeguard these leakages to whatever extent possible.

References

[1] http://v4.jiwire.com/search-hotspot-locations.htm.

[2] K. Hole, E. Dyrnes, and P. Thorsheim, "Securing wi-fi networks," *Computer*, vol. 38, no. 7, pp. 28–34, 2005.

[3] B. Potter, "Wireless hotspots: petri dish of wireless security," *Communications of the ACM*, vol. 49, no. 6, pp. 50–56, 2006.

[4] S. Spiekermann and L. Cranor, "Engineering privacy," *Software Engineering, IEEE Transactions on*, vol. 35, no. 1, pp. 67–82, 2009.

[5] www.alexa.com/topsites/.

[6] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80, ACM, 2005.

[7] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th international conference on World Wide Web*, pp. 181–190, ACM, 2007.

[8] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pp. 506–515, IEEE, 2008.

[9] B. Krishnamurthy and C. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*, pp. 7–12, ACM, 2009.

[10] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," in *Proceedings of the 18th international conference on World wide web*, pp. 531–540, ACM, 2009.

[11] B. Malin, "Betrayed by my shadow: learning data identity via trail matching," *Journal of Privacy Technology*, vol. 2147483647, 2005.

[12] B. Krishnamurthy and C. Wills, "Characterizing privacy in online social networks," in *Proceedings of the first workshop on Online social networks*, pp. 37–42, ACM, 2008.

[13] B. Krishnamurthy and C. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proceedings of the 18th international conference on World wide web*, pp. 541–550, ACM, 2009.

[14] AdBlock Plus. http://v4.jiwire.com/search-hotspot-locations.htm.

[15] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For sale: your data: by: you," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, p. 13, ACM, 2011.

[16] M. Backes, A. Kate, M. Maffei, and K. Pecina, "Obliviad: Provably secure and practical online behavioral advertising," in *IEEE Symposium on Security and Privacy*, 2012.

[17] S. Guha and P. Francis, "Identity trail: Covert surveillance using dns," in *Proceedings of the 7th international conference on Privacy enhancing technologies*, pp. 153–166, Springer-Verlag, 2007.

[18] J. Castellà-Roca, A. Viejo, and J. Herrera-Joancomartí, "Preserving users privacy in web search engines," *Computer Communications*, vol. 32, no. 13, pp. 1541–1551, 2009.

[19] Y. Lindell and E. Waisbard, "Private web search with malicious adversaries," in *Privacy Enhancing Technologies*, pp. 220–235, Springer, 2010.

[20] S. Peddinti and N. Saxena, "On the privacy of web search based on query obfuscation: A case study of trackmenot," in *Privacy Enhancing Technologies*, pp. 19–37, Springer, 2010.

[21] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[22] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," tech. rep., DTIC Document, 2004.