SSIDs in the Wild: Extracting Semantic Information from WiFi SSIDs

Suranga Seneviratne, Fangzhou Jiang, Mathieu Cunche, Aruna Seneviratne

SSIDs in the Wild: Extracting Semantic Information from WiFi SSIDs

Suranga Seneviratne*†, Fangzhou Jiang*†, Mathieu Cunche‡§, Aruna Seneviratne*†
* School of EET University of New South Wales , † NICTA, Australia
† Email: first name.last name@nicta.com.au
‡ University of Lyon, France, § Inria, France
§ Email: mathieu.cunche@inria.fr

***Abstract*—WiFi networks are becoming increasingly ubiquitous. In addition to providing network connectivity, WiFi finds applications in areas such as indoor and outdoor localisation, home automation, and physical analytics. In this paper, we explore the semantics of one key attribute of a WiFi network, SSID name. Using a dataset of approximately 120,000 WiFi access points and their corresponding geo-locations, we use a set of similarity metrics to relate SSID names to known business venues such as cafes, theatres, and shopping centres. Such correlations can be exploited by an adversary who has access to smartphone users preferred networks lists to build an accurate profile of the user and thus can be a potential privacy risk to the users.**

I. Introduction

WiFi networks have gained a rapid popularity in recent years and it is predicted that there will be one WiFi hotspot for every 20 users by 2018 [1]. In addition to its presence in enterprise environments and public places such as shopping malls and transit stations, WiFi networks are extensively used in residential settings as the means of sharing broadband connections among multiple devices [2]. The pervasive use of WiFi networks has enabled the provision of various other services. For example, WiFi fingerprinting is being used as a means of localisation and is now commonly available in smart devices [3], [4]. Customer analytics based on data collected from WiFi hotspots is another such example[1,2].

One attribute of a WiFi network is the SSID name that is generally assigned by the owner of the network. In public and enterprise settings, meaningful names are assigned for the SSIDs so that users can easily find the relevant network through their devices. Most of the current smart devices store information and configuration of the networks user has connected to in the past, to expedite re-connection to that network when the user comes within the coverage area next time. It has been reported that some advertisement networks and analytics libraries collect this information by directly accessing the preferred network list or indirectly by collecting information on the currently connected WiFi network [5], [6].

This paper explores the potential privacy implication of smartphone users, of such data collection. Specifically, we investigate how the semantic information available in WiFi SSIDs can be exploited to correlate them to business venues that potentially can be used to profile users. By analysing data from approximately 120,000 WiFi networks, we first provide insights on how network owners tend to name their WiFi networks. Then we introduce and evaluate similarity metrics that can be used to correlate business entities with observed SSIDs in a confined geographic region. Our results show that at higher similarity levels a precision as high as 97% can be achieved.

The remainder of the paper is organised as below. In Section II we present the related work. Section III describes the datasets used and Section IV provides a basic characterisation of the dataset and the filtering methods we used. Section V presents similarity metrics that can be used to correlate business entities to WiFi SSIDs and an evaluation of their performance. Section VI discusses the privacy implications of our findings and potential future research directions.

II. Related Work

Multiple work have studied privacy leakages related to WiFi connectivity in Smartphones [7], [8], [9], [5]. Greenstein et al. [7] presented the privacy issues of 802.11 service discovery, that results in the device MAC address as well as the identifiers (SSID) of previously connected networks being leaked on the wireless channels. For example, authors showed that those SSIDs can be resolved to a set of locations, thus revealing the previously visited locations. Cunche et al. [8] showed that SSIDs leaked by wireless devices can be used to infer social links between the owners. Cheng et al. [9] extended this social-link inference framework by including physical proximity and spatio-temporal behaviour. Achara et al. [5] investigated the abuse of the Wi-Fi related permission of the Android system and showed that it can be exploited by apps to obtain sensitive information such as users' location.

Various aspects on how users name their smart devices has been also explored [10], [11], [12], [13], [14], [15]. O'Neill et al. [10] studied Bluetooth names of 1,701 devices and showed that 42% devices tend to use default names given by the manufacturer. Kinderberg et al. [11] categorised Bluetooth names into four, *i) Identifiers:* E.g. full names, nicknames and pseudonyms, *ii) Associations:* E.g. owner's interests, such as band names, *iii) Graffiti/T-Shirt:* E.g. Text found in Graffiti, and *iv) Direct address:* a statement to a person who has discovered the device (E.g. bonjour tutti!). Konings et al. [12]

[1]https://www.skyfii.com/business.html
[2]http://retailnext.net/how-it-works/

Fig. 1: Geo-distribution of the collected SSIDs in Sydney

studied the names of user devices that are sending *mDNS* in a WiFi network and found that 59% of the device names contained real names and 17.6% contained both first name and last name. Ferreira et al. [15] investigated the effect that trust and context have on users when choosing wireless networks. For example, it was shown that users tend to connect to the networks known to them or have connected in the past, despite not completely trusting them.

To the best of our knowledge this is the first attempt of highlighting the privacy implications to smartphone users from an adversary who can extract semantic information from WiFi SSIDs.

III. Dataset

We use following two datasets in this paper.

SSID dataset contains the SSID, BSSID, and geo-location of 123,251 WiFi access points (i.e. unique BSSIDs) in Sydney, Australia. The dataset was collected by volunteers who traveled in the area with Android phones running the WiGLE[3] wardriving app. Figure 1 illustrates the geographical distribution of the collected SSIDs.

Yelp dataset contains the business names and geo-locations of 50,418 businesses in the same area obtained by crawling the popular business review website Yelp[4]. For each geo-location where an SSID was observed, we collected the businesses inside a $150m \times 150m$ square, having the SSID location as the centre.

IV. Basic Characteristics & Pre-Processing

WiFi SSID (Service Set Identifier) can contain up to 32 ASCII characters and most of the time used to assign human understandable names to the network, though there is no such specific requirement defined by the standard. This section describes the SSID naming patterns we identified and the filters we applied to remove SSIDs that do not contain semantic information relevant to our study.

[3]http://www.wigle.com
[4]http://www.yelp.com

A. Popular SSIDs

The 123,251 unique BSSIDs (access points) in the SSID dataset was corresponding to 73,899 unique SSIDs. The difference is due to a number of reasons, such as users keeping the default SSID names and large wireless networks having a significant number of access points (APs) with the same SSID.

In Figure 2 we show the 25 most popular SSIDs in our dataset and their frequencies. We found that 8,924 (7.24%) APs are not broadcasting their SSIDs (cloaked networks) and thus appear as blank SSIDs in the dataset. SSIDs such as *eduroam* (global WiFi service for users in research and higher education) and *detnsw* (Department of Education, New South Wales) are examples of large WiFi networks having a single SSID. *Netcomm Wireless, AndroidAP,* and *NETGEAR* are examples of owners continuing to use the default SSID names.

The sub-plot in Figure 2 shows the frequency of occurrence of SSIDs against the rank according to the frequency. As can be seen, while there are limited number of SSIDs occurring frequently, there is a long tail indicating significant number of SSIDs are unique. For example, 66,313 (89.73%) SSIDs were observed only once and 70,866 (95.90%) SSIDs occurred only once or twice.

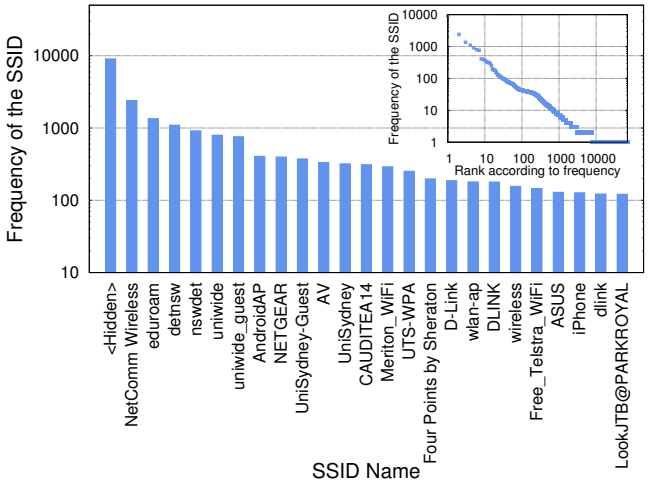

Fig. 2: SSID Popularity

B. Access Point Manufactures

For each access point we identified the device manufacturer by comparing first half of the MAC address with the *Organisationally Unique Identifier* list provided by IEEE[5] and we show the top-25 manufactures according to the percentage from total access points in Figure 3.

Netgear and *Cisco* were the top manufactures. Furthermore, it can be seen that the top-5 manufacturers' devices were used by approximately 50% of the WiFi networks and top-25 manufacturers' devices were used by approximately 88% of the networks.

[5]http://standards-oui.ieee.org/oui.txt

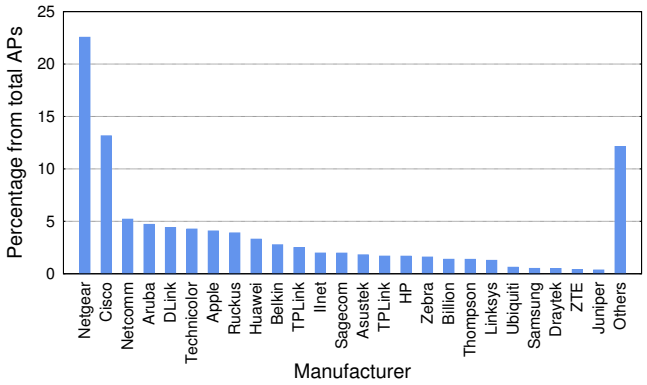

Fig. 3: Top AP manufacturers

TABLE I: Filters used on the SSID dataset

Filter	Unique BSSIDs	Unique SSIDs
Original Dataset	121,251	73,899
Cloaked Networks (Networks that do not broadcast their SSID)	8,924	1 (Blank)
Default SSID Names (SSIDs exactly matching to default SSIDs comes with the access point E.g. *Netcomm Wireless, AndroidAP, and NETGEAR*)	38,100	27,814
Generic Terms (SSIDs exactly matching to generic terms E.g. *wlan, wireless, wlan-ap)*	696	19
Non-Alphabet Characters (SSIDs containing more than 70% of non-alphabet characters)	1,068	858
SSID Length (SSIDs of length less than or equal to two characters)	1,045	394
Hybrid Filters (SSIDs with combinations of default SSID names, generic terms, and non-alphabet characters contributing to more than 70% of the total length)	1,304	890
Filtered Dataset	72,114	43,913

C. Filtering

We applied the filters described in Table I to remove data that do not provide any significant semantic information relevant to our study, i.e. SSIDs potentially related to businesses.

D. Pre-Processing & Tokenising

We pre-processed the business names and SSID names. We first changed the business names to lower case and then removed punctuation and tokenised using white spaces. However, for SSIDs, as naming patterns may not follow the standard writing English, we first tokenised by using: case changes, letter to digit transitions (or vice versa), and punctuation. Then, we changed the case to lower and removed the words which are only one character long.

V. Business Venue Identification

In this section, we present the similarity metrics we used to correlate WiFi SSIDs to business venues. For each SSID, we identified the businesses from the Yelp dataset, that lie inside $150m$ x $150m$ square[6] having the SSID location as the centre. Then, we calculated different similarity metrics between the SSID and the names of the businesses found in the neighbourhood.

[6]The size of the square was selected assuming a WiFi range of 75m.

A. Cosine Similarity (Character & Words Counts)

We calculated the *cosine similarity* between the SSID name and the business name at *character level* and *word level*. For example, in character level we represent the SSID name and the business name as a character-frequency vector, where characters were identified from both the names and calculate the cosine similarity between the SSID name (a) and the business name (b) as, $cos(a,b) = \frac{a.b}{||a||||b||}$. Similarly, word level cosine similarity was calculated by representing the names as word-frequency vectors.

B. Cosine Similarity (TF-IDF)

The two cosine similarities mentioned above can be affected by the globally and locally popular words. For example, words such as *Sydney* or *Aussi* are popular throughout the geographic region we consider. Examples for locally popular terms are suburb names that can be common among multiple business names in a smaller geographic region (i.e. *Burwood Cafe, Burwood Plaza, Burwood Library*). A match of such words must have a lower weight in deciding the final similarity value.

To this end, we first identified the top-100 words in all the corpus of business names and removed those as *stop words*. Then, for each SSID location we searched for the names of businesses in a larger square (i.e. 1.5km x 1.5km) and used that corpus to calculate *IDF (Inverse Document Frequency)* for each term. Afterwards, for each SSID in the dataset we looked for businesses within the $150m$ x $150m$ as before and represented both SSID name and business names as a *tf-idf* vectors using the previously calculated *IDF* values. Finally, we calculated the cosine similarity these two vectors as before.

C. Results

In Table II we show the SSID and business venue pairs identified in each 0.2 range of the similarity metrics. As can be seen, there are clear gaps in the number of identified pairs. For example, there is considerably smaller number pairs in the range 0.8 - 1.0 for character level cosine similarity compared to lower similarity ranges.

For the two word level cosine similarities, majority of the pairs have a zero similarity and there are only a small number of pairs in the range 0.0 - 0.2. To have a non-zero similarity value both the SSID and the business name must have a common word and a single word match usually results a similarity higher than 0.2 as the most of the times SSIDs and business names are composed of few words.

Higher similarity ranges (marked in grey in Table II) are likely to contain correct matches between SSIDs and the businesses. To identify a proper threshold, for each similarity range we selected a random sample of 100 SSID and business pairs and manually checked whether it is a correct match. Manual check was done by three researchers independently and the result is considered correct if all three researchers marked the pair as a correct match. In Table II we show

TABLE II: Number of SSID and business pairs & precision

Similarity Range	Cosine (Character)	Cosine (Word)	Cosine (Word TF-IDF)
0.0	38,963	1,424,108	1,427,468
0.0 - 0.2	140,521	68	362
0.2 - 0.4	432,819	3,867 (2%)	1,507 (4%)
0.4 - 0.6	555,065	3,913 (7%)	2,069 (15%)
0.6 - 0.8	248,006	704 (71%)	1,350 (56%)
0.8 - 1.0	13,293 (10%)	629 (97%)	1,379 (69%)

TABLE III: Examples of SSID and business matches

SSID	Business Name
CentralStationHotel	Central Station Hotel
Carmen Nicotra's Network	Carmen E. Nicotra
SydneyTAFE	Tafe NSW - Sydney Institute
.ParkRegis@FreedomInternet-9C	Park Regis City
CafeDelMarGuest2.4G	Cafe Del Mar
ManlySeniors	Manly Club For Seniors
AlphaHealth	Alpha Health Clinic
BAPTIST	Epping Baptist Church

the results as precision (i.e. how many pairs out of the 100 manually checked pairs were correct matches). According to the results, when the similarity is over 0.8, word level cosine similarity metric gives a high precision of 97%. Noticeably *tf-idf* version does not outperform the basic word level similarity. One possible reason can be the use of fixed square size than adaptively changing it according to the density of the data points. Table III shows some examples to correct matches we identified.

VI. Discussion & Concluding Remarks

A. Privacy Threats to Smartphone Users

Using data that can be collected conveniently, we showed that it is possible to match business entities to WiFi SSIDs with high precision. Under the context of smartphones, this can be exploited to identify users' fine-granular locations and most importantly semantic information of users location that may not be obtained using maps. For example, an adversary who has access to user's location can decide the user is inside a large shopping mall. By extracting the semantic information of the SSID user connects inside the shopping mall, the adversary can further reveal the exact shop the user is visiting.

In the context of smartphones, the SSIDs users connect to can be collected in multiple ways. Apps can get information about the user's current network or can collect a snapshot of networks user's have connected in the past by accessing the *preferred network list* [5] and that enables building a user profile instantly. Some WiFi enabled smartphones send out probe requests to the networks they have connected in the past [7], [8], [9] and an adversary who monitors the wireless channel can easily collect this information.

While these types of inferences can be considered as a privacy threat, it can be used to provide targeted advertisements, promotions, or coupons to users based on the semantic information obtained on their fine granular location.

B. Other Semantic Information

Apart from identifying business entities, SSIDs provides various other forms of semantic information. For example, we found approximately 1,800 SSIDs having the pattern *"'s"* which is most of the time prefixed by a person name and suffixed by a device type (*e.g. Maryanne's iPhone, Fabio's Time Capsule*). Also there were SSIDs that contain personal names without the pattern *"'s"*. It might be possible to identify those networks by querying against popular person names in the country.

The default SSID names can also be used indirectly to identify socio-economic characteristics of a locality. For example, use of expensive WiFi routers or high end broadband service providers can be a potential indication of the wealth of the people living in the neighbourhood.

C. Future Work

In this paper, we used the fixed grid sizes for the similarity measurements. Further improved results can be expected by adaptively changing the grid size of according to the density of the data points. For example, the locally popular words (e.g. suburb names) can have a smaller geographic scope in the central parts of the city and a larger scope in the outskirts. Also, it might be interesting to repeat the experiment in few other larger cities and compare the results. Finally, it is possible to collect lists of *preferred networks* from real smartphone users and quantify the amount of locations that can be related to businesses on average for a given user.

References

[1] iPass Inc., "iPass Wi-Fi growth map shows 1 public hotspot for every 20 people on earth by 2018," http://www.ipass.com/press-releases/, 2014.

[2] E. Smith, "Global broadband and WLAN (Wi-Fi) networked households forecast 2009-2018," https://www.strategyanalytics.com, 2014.

[3] Google Inc., "Location strategies," http://developer.android.com/guide/topics/location/strategies.html, 2015.

[4] Apple Inc., "Understanding location services," https://support.apple.com/en-us/HT201357, 2015.

[5] J. P. Achara, M. Cunche, V. Roca, and A. Francillon, "Short paper: Wifileaks: Underestimated privacy implications of the ACCESS_WIFI_STATE Android permission," in *ACM WiSec*, 2014.

[6] S. Seneviratne, H. Kolamunna, and A. Seneviratne, "A measurement study of tracking in paid mobile applications," in *ACM WiSec*, 2015.

[7] B. Greenstein, R. Gummadi, J. Pang, M. Y. Chen, T. Kohno, S. Seshan, and D. Wetherall, "Can Ferris Bueller Still Have His Day Off? Protecting Privacy in the Wireless Era." in *HotOS*, 2007.

[8] M. Cunche, M. A. Kaafar, and R. Boreli, "I know who you will meet this evening! Linking wireless devices using Wi-Fi probe requests," in *IEEE WoWMoM*, 2012.

[9] N. Cheng, P. Mohapatra, M. Cunche, M. A. Kaafar, R. Boreli, and V. Srikanth, "Inferring user relationship from hidden information in WLANS," in *IEEE MILCOM*, 2012.

[10] E. O'Neill, V. Kostakos, T. Kindberg, A. Penn, D. S. Fraser, T. Jones *et al.*, "Instrumenting the city: Developing methods for observing and understanding the digital cityscape," in *UbiComp*, 2006.

[11] T. Kindberg and T. Jones, ""Merolyn the Phone": A Study of Bluetooth Naming Practices," in *UbiComp*, 2007.

[12] B. Könings, C. Bachmaier, F. Schaub, and M. Weber, "Device names in the wild: Investigating privacy risks of zero configuration networking," in *IEEE Mobile Data Management*, 2013.

[13] F. Palmer and E. O'Neill, "Interpreting technology-mediated identity: perception of social intention and meaning in bluetooth names," in *CHI*. ACM, 2010.

[14] M. Persson, ""Loli: I Love It, I Live with It": Exploring the practice of nicknaming mobile phones," *Human IT*, vol. 12, no. 02, 2013.

[15] A. Ferreira, J.-L. Huynen, V. Koenig, G. Lenzini, and S. Rivas, "Socio-technical study on the effect of trust and context when choosing WiFi names," in *Security and Trust Management*. Springer, 2013.