Abstract

We discuss the use of selected passive wireless device fingerprinting and prior location inference techniques from a forensic perspective. With specific focus on 802.11 signals, we examine how fingerprints based on the Preferred Network List (PNL) enumerations can be utilized to obtain the locational dimension. This dimension may act as a bridge between the digital fingerprint and the physical world.

Using a data set of 147,944 network names contextualized to a particular geographical area, we discuss the associated opportunities, challenges and limitations.

1. Introduction

Ongoing growth of the number of smart devices such as smartphones, tablets and wearables indicates their widespread adoption and suggests the ever-increasing likelihood of people carrying at least one smart device with them at all times [5, 7]. Such devices are commonly equipped with a number of features to support connectivity over communication protocols, such as Global System for Mobile Communications (GSM), Bluetooth and 802.11 wireless commonly known as “Wi-Fi”. The latter protocol has been suggested to carry signals of possible evidentiary value [24, 34, 35]. Passive wireless signal capture and subsequent analysis can be used to infer possible prior locations of a client device using wireless access point (AP) geolocation based on disclosed network names.

The ubiquitous presence of Wi-Fi coupled with the unrestricted nature of network naming practices mean that for any given network name there could be a number of potential matching locations. While these matches can represent a distinct set of possibly unique locations from a single area, they can also refer to locations from largely dispersed geographical zones spreading across multiple countries. In this paper, we look into the practical aspects of past location inference using wireless AP geolocation and examine the associated opportunities and challenges of this method from a digital forensics perspective.

The rest of the paper is organized as follows. Section 2 presents the related work and introduces the digital forensic process model used to illustrate the identified challenges and opportunities, which are described in Sections 3 and 4. A discussion of the key findings is presented in Section 5. The paper is concluded and future research directions are outlined in Section 6.

2. Related work

The active access point (AP) discovery mechanism specified by the 802.11 protocol can lead to disclosure of sensitive information that can subsequently be exploited to carry out various inference attacks [28, 29]. The discovery mechanism relies on storing the list of previously associated networks known as the Preferred Network List (PNL) by the device and periodically sending out signals possibly containing network names from the PNL [6, 9].

These signals, also known as probe request frames, carry various attributes of potential forensic significance including the Media Access Control (MAC) address of the probing client, the Service Set Identifier (SSID) of the preferred network and received signal strength that indicates the client proximity to the receiver [17]. Probe request interception can be facilitated by passive wireless traffic monitoring and does not require specialized tools – rather, a combination of low-cost hardware and open-source software components can provide a basic sensor for less than US$60 [26]. The introduction of even cheaper and smaller multi-protocol boards such as the Intel Edison is also likely to reduce barriers to entry in the context of large-scale wireless signal monitoring [19].

The passive manner in which probe requests can be collected has enabled a number of studies into their associated applications.
2.1. Passive fingerprinting

In [40], the authors provide a broad overview of wireless device fingerprinting methods and associated challenges. In addition to passive techniques, a number of active approaches are described. However, we are not exploring active methods further due to the inherent need for client device interaction, which is deemed undesirable in a forensic context. In some jurisdictions it is also an illegal undertaking without proper warrants, court orders or appropriate isolation.

As one example of a passive approach, looking at the externally observable characteristics of signal frames, wireless device driver fingerprinting techniques can be utilized to distinguish between various drivers and chipsets due to small variations in probing timing and patterns [12, 14]. It is unclear whether these techniques are consistently useful in the context of devices that are based on identical drivers and hardware.

Another approach based on the contents of the PNL can be used to derive a digital fingerprint that may be unique for devices that probe for a diverse set of networks [10, 27]. The MAC address of the probing device can be collected regardless of whether the frame contains a non-blank SSID. While a technically minded individual can spoof their MAC address, certain detection techniques can be employed to mitigate this practice [12].

In a bid to reduce device information leakage during service discovery, Apple incorporated MAC address randomization into the service discovery mechanism implemented by iOS 8 and above. However, the feature does not appear to be operating on a full-time basis and requires possibly uncommon device settings to be selected prior to activation [2, 33]. Measures for defeating randomization in practice have already been proposed [15].

2.2. Inference and analytics

PNL-based fingerprints can be used to track visitors at mass events and identify potential social relationships between the owners of different devices based on the commonalities in respective fingerprints [3, 4, 10]. A timeline-driven analysis of device manifestation can reveal the daily habits of its owner [39]. Commercial entities offer visitor and location analytics solutions that also employ timeline-based presence analysis but are not necessarily concerned with the contents of the PNL [8, 13].

Locational analysis is possible on the basis of two sources. First, data collection sensors can be equipped with the Global Navigation Satellite System (GNSS) receivers to capture an approximation of client device location. A dense sensor deployment can provide the basis for accurate location inference based on Wi-Fi positioning algorithms [41]. Second, a database of physical AP locations such as WiGLE.net can be interrogated to determine geographic coordinates of wireless APs based on the network name [3, 4, 6, 9, 10, 16, 24].

We focus on the second technique. Previous works featuring WiGLE.net in similar contexts do not necessarily examine the general characteristics of the utilized geolocation process. This research aims to fill this gap by looking at the custom data set of 147,944 SSIDs representing the WiGLE.net data available for the metropolitan area in Perth, Western Australia.

2.3. WiGLE.net and SSID geolocation

Various war-driving data repositories have emerged as the result of the sharing culture pertinent to the war-driving communities [22]. At the time of writing, WiGLE.net is the largest community-driven repository in operation since 2001 that contains location information on over 200 million wireless APs. The data collection is based on a crowd-sourced model with an Android app available to the community members and the public. Support for direct data file uploads in a variety of different file formats such as Kismet and NetStumbler is also provided [37]. The available end-user search can be used to locate SSIDs taking into account specific granularity requirements, such as bounding the search to a particular geographical area.

Figure 1 presents the extract of a SSID lookup result set showing the raw attribute display format. Historically, the query would involve impersonating a client browser request to mimic end-user search to download the raw results page and extract the required
information from specific markup elements [25]. With the introduction of an undocumented JavaScript Object Notation (JSON) endpoint in late 2014, alternative and more efficient techniques for accessing the database have become available.

### 2.4. Digital Forensics Context

There are a number of key method requirements around collection of wireless signals for forensic purposes based on the principles of minimal interaction, accuracy, repeatability and legal compliance [34]. For PNL fingerprint enumeration, these principles mean that device signal capture must be conducted using passive wireless monitoring that only collects the required frame header attributes (such as MAC address, SSID and RSSI) disregarding the transmitted payload information. As described in [34, 36], this approach could be suitable for warrantless interception in the Australian context and possibly in other jurisdictions.

Passive wireless monitoring does not guarantee that probes from all devices in the vicinity of the sensor will be observed as various wireless network interface card (NIC) configurations will yield different capture and coverage rates [15]. Therefore, it is not guaranteed that the collected PNL fingerprints will be complete for any given device because it may not be practical or possible to capture all of the emitted probes.

While the requirements and limitations associated with 802.11 probes collection and PNL fingerprint enumeration are relatively well understood, the forensic aspect of SSID geolocation using 802.11 traces is ambiguous. To identify a structured way of establishing its place in this context we look at the existing digital forensic investigation frameworks.

The authors of [1] provide a survey of digital forensic models described between 1995 and 2014 that includes the Integrated Digital Forensic Process Model (IDFPM) as the most recently introduced concept. IDFPM is presented as being an “all-encompassing standardized” multi-model derivative that facilitates a uniform approach to digital investigations [21]. IDFPM consists of 5 phases each comprising a defined set of steps that are to be followed in a prescribed manner. As shown in Figure 2, SSID geolocation based on 802.11 device traces could be mapped to two of the five phases of this model - preparation and investigation, because we anticipate that appropriate readiness must be achieved before SSID geolocation can be integrated into the latter phase. Furthermore, we deliberately separate trace capture from geolocation-based analysis to reflect the focus of this paper. Subsequently, we assume that device traces have been captured using a forensically sound method in accordance with the previously outlined requirements and that we have a pool of trusted PNL fingerprints available to us for subsequent examination.

### 3. Preparation

The preparation phase is labeled as the “single most critical process” and mainly focuses on forensic readiness from the policy, procedural, infrastructure and operational perspectives [21]. We only discuss the infrastructure readiness aspect, as we believe it will have critical influence on the practical feasibility of the proposed technique. Other perspectives will also need to be given due consideration when establishing readiness from a holistic perspective.

#### 3.1. Infrastructure readiness challenges

To enable SSID geolocation, a reliable, complete and current source of AP locations is required. Community-driven databases such as WiGLE.net may be prone to potential quality and coverage issues.
The collection techniques used by the suppliers of the data will undoubtedly affect its quality. For example, antennas with different gains can be used to capture AP beacons from a number of distances and this can affect the precision of the associated reported geographic coordinates. Optimization algorithms could be employed to account for this scenario, but in areas of low community activity optimization may not be possible due to limited volume of reported observations.

Historical observations can be kept in the database indefinitely and, therefore, may increase the number of potential matches for a given network, even though some of them may refer to no longer operational APs. While filtering based on age can be used to discard the possibly inexistent locations, it may also result in the removal of genuine elements.

### 3.1.2. Coverage

Coverage and timeliness of observations pertaining to new deployments also depend on community activity levels [22]. As shown in Figure 3, a large portion of the reported locations is associated with major roads with poor coverage of residential areas located in between. A subset of a community-driven database could be used to establish the basis for a specialized service in a rapid fashion [11]. Commercial providers subsequently rely on consumer devices to automatically maintain the accuracy and currency of Wi-Fi positioning services [32].

### 3.2. Specialized source opportunity

To facilitate a reliable and up-to-date source of AP location information, an agency could establish and maintain a dedicated specialized database using sensors deployed via an existing public transport system as well as municipal and other vehicles that are likely to travel away from major roads. This approach would expedite improved coverage and timely discovery of newly added APs. Needless to say, the collection and preservation of beacons and geographic coordinates would need to be performed in pre-defined, consistent and a forensically sound manner. The discovery of on-premise corporate networks, however, would still present a challenge because of the potential need to enter restricted access premises to capture the associated beacons.

### 4. Investigation

The investigation phase is the epicenter of IDFPM.
The phase involves following a defined process to develop findings for subsequent presentation. For illustrative purposes, we assume that we have identified the pool of devices of possible interest represented by their MAC addresses and respective PNL fingerprints and we are interested in being able to infer prior device locations based on SSID enumerations. We now examine the possible uses of SSID geolocation in this context.

4.1. Opportunities

4.1.1. Context-bound geolocation

We assume that PNL-based fingerprints collected in a particular area are likely to represent devices that belong to local individuals. In other words, associated device owners are possibly recurring visitors to this area and may work or reside in relatively close proximity, since average gyration radius of human mobility is thought to be less than 100 km [18]. Therefore, we suggest that context-bound geolocation, or limiting SSID location search to a particular area of interest, can be used to obtain a more representative location set in a localized context.

Figure 4 shows the cumulative distribution function and density based on SSID location count for both the entire data set and networks that have up to 10 associated locations. The majority (88%) of SSIDs in the analyzed data set appear to be uniquely identifiable (one known associated location) with vendor default network names like “DLINK” and “NETGEAR” representing the least locatable items with the most number of locations.

While the latter finding is not surprising, we recognize that this context-bound dataset presents a feasible source of prior device location inference. To challenge this view, we also describe the potential limitations associated with non-contextualized geolocation in Section 4.2.1.

4.1.2. Specialized network naming conventions

A number of Australian Internet Service Providers (ISPs) often pre-configure the SSID of supplied consumer routers and wireless adapters using a string that contains pseudo-random alphanumeric sequences following a set prefix. Table 1 lists some of these prefixes together with the corresponding SSID count and overall data set percentages in the context of SSIDs with one associated location.

In total, such networks represent at least 16% of the analyzed data set and this selection is not exhaustive. At the same time, a large portion of consumers may not have the ability to modify the pre-configured network name due to lack of required technical knowledge and the original value is likely to remain static for prolonged periods of time if not for the entire lifetime of the AP.

Table 1. Selected prefixes used in pre-configured SSIDs that have one associated location in the dataset

<table>
<thead>
<tr>
<th>Prefix (case-insensitive)</th>
<th>Number of SSIDs with 1 location</th>
<th>Data set percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigPond</td>
<td>16,224</td>
<td>11%</td>
</tr>
<tr>
<td>Telstra</td>
<td>3,112</td>
<td>2%</td>
</tr>
<tr>
<td>Optus_</td>
<td>2,835</td>
<td>2%</td>
</tr>
<tr>
<td>pocketwifi</td>
<td>1,918</td>
<td>1%</td>
</tr>
</tbody>
</table>

Figure 5. Box plot describing the Shannon entropy (A), SMAZ compression rate (B) and zlib compression rate (C) for unique and non-unique network name strings in the context of a random sample representing 5% of the entire data set for each category. Horizontal lines depict thresholds used in the algorithm.
4.1.3. Name-based locatability inference

In some cases, it may be possible to gauge the potential location count of a given SSID based on selected characteristics of the value string. To illustrate this feature, we calculated the Shannon entropy value [31] and compression rates for two algorithms - SMAZ and zlib, as measures of the associated randomness for randomly selected sets of uniquely (single location) and non-uniquely (more than one location) locatable SSIDs representing 5% of each category. The SMAZ algorithm is used because it offers effective compression of short strings and SSIDs are limited to 32 characters [30]. The zlib algorithm is used due to its widespread adoption.

Using basic descriptive analysis, we observed slight differences in the summaries of each characteristic between the values corresponding to the two groups, as presented in Figure 6. We attempted to quantify the possibility of deducing whether a particular SSID is likely to have a single associated location using selection thresholds based on basic descriptive characteristics.

In testing of the algorithm described in Figure 6, we observed that using this simple logic we were able to correctly identify the likelihood of potential location uniqueness for 81% of SSIDs from the remaining 95% of the data set. This approach or its more sophisticated derivative could be utilized as part of a greater SSID location model. The model could support feature selection without the need to query a source of locational information. This selection can be useful in situations where access restrictions or limits are being imposed or the overhead associated with executing an immediate search against an AP location database is not desirable.

4.1.4. Location prediction

PNL fingerprints providing a set of distinct locations can be used as input into location prediction algorithms, since even a sparse data set can be utilized as a viable basis for this purpose [20]. Additionally, location types can be inferred using residential and business lookup services or via SSID string analysis [9]. The possible day times associated with inferred device locations can potentially be estimated based on location types – such as late morning and afternoon for commercial or business locations and evening and early morning for residential locations. This can be useful when an entity is interested in a possible physical sighting of an individual carrying the device and needs to limit the potential pool of locations to monitor.

4.2. Challenges and limitations

While SSID geolocation presents a number of prior device location inference opportunities, it is not without challenges and limitations. We describe some of these aspects in this section.

4.2.1. Bounds selection

Although context-bound SSID geolocation may be effective when using fingerprints belonging to devices with a relatively short gyration radius, it may only provide a limited view for those that belong to individuals who travel in a broader context. Therefore, to obtain a bigger picture, any location contextualization filters could then be removed.

To illustrate the associated impact we extract SSIDs with a single known location from our dataset and perform non-contextualized location lookups using

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of locations (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>7,608 (41%)</td>
</tr>
<tr>
<td>USA</td>
<td>6,427 (35%)</td>
</tr>
<tr>
<td>Canada</td>
<td>880 (5%)</td>
</tr>
<tr>
<td>UK</td>
<td>632 (3%)</td>
</tr>
<tr>
<td>Germany</td>
<td>558 (3%)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>257 (1%)</td>
</tr>
<tr>
<td>Russia</td>
<td>160 (&lt;1%)</td>
</tr>
<tr>
<td>New Zealand</td>
<td>132 (&lt;1%)</td>
</tr>
<tr>
<td>China</td>
<td>127 (&lt;1%)</td>
</tr>
<tr>
<td>Sweden</td>
<td>108 (&lt;1%)</td>
</tr>
</tbody>
</table>
WiGLE.net to assess their locatability outside of the original context. We were able to perform lookups for 6,023 SSIDs, which represents 4% of the 131,653 SSIDs with one known location. While 3,971 (66%) of the SSIDs were still found to have a single associated location, queries for 2,052 (34%) items returned more than one match – representing a total of 57,418 different locations.

Subsequently, we reverse-geocoded the coordinates of 18,428 (32%) non-contextualized locations and observed a wide geographic spread across more than 90 different countries. The corresponding selection of the top 10 counties is shown in Table 2.

Non-bound geolocation has the potential to produce high volume of matches possibly with a significant amount of noise. Selecting appropriate bounds represents a potential challenge, because different use cases will carry different requirements. Depending on the objectives and context, one may choose to focus on specific areas of interest, whether locally or on an interstate or multinational level.

To emphasize the associated challenges, we analyzed the results of non-bound lookups further. First, we produced the subset of the associated per-SSID location counts for locations within the original bounding box denoting the Perth metropolitan area. Second, we produced the same subset for locations that were within the bounding box denoting Australia but outside Perth metropolitan area (the context of the initial data set). Finally, we produced the same subset for all locations corresponding to the rest of the world. The approximations of bounding boxes used are shown in Figure 6.

Figure 7 shows the cumulative distribution function of SSID location count for the three resulting subsets. The evident curve differences illustrate the effect that bounds selection can have upon the associated SSID location count. Specifically, we observed that a very small portion of the original SSIDs had more than one location within the original context (a change from 1 location to 2), possibly due to recent updates to the WiGLE.net database.

Subsequent bounds expansion resulted in diminishing locatability of SSIDs that were originally thought to have only one associated location. In the context of the same country, natural migration of the population could possibly result in physical relocation of the wireless APs. Furthermore, the pseudo-random SSID naming practices described in section 4.1.3 could result in suffix collisions meaning that the same network name is automatically assigned to wireless access devices purchased by consumers in different locations. The increase in the number of matches across different countries could be attributed to the...
likelihood of SSID reuse across multiple locations.

4.3. Integrating SSID geolocation

Given our firm belief in the potential of this technique, we discuss how a structured SSID geolocation process could fit into the Digital Forensic Investigation phase of the IDFPM as one of possible inputs, as shown in Figure 8.

Consider a scenario where a large-scale sensor deployment can facilitate device trace capture in a forensically sound manner. Given an incident, an investigator could be interested in identifying the devices that were in close spatial and temporal proximity to this incident. To achieve the identification, historical traces can be filtered to match the required spatiotemporal dimension.

Based on a pool of identified devices, the SSID geolocation process can be followed to generate findings in an iterative fashion. First, an individual device is selected from the pool and its PNL contents are analyzed. If PNL fingerprint lacks potentially locatable SSIDs (contains a blank SSID or SSIDs associated with known common network names such as "NETGEAR") it may not be practical to proceed with the remaining steps and another device is selected from the available pool.

A single SSID is then selected from the PNL fingerprint to perform the lookup. Context definition forms an important part of the process because it will influence the number and possible uniqueness of returned locations in a given context, as discussed in Section 4.2.1. The defined context informs the query to the trusted source of AP locations, facilitated during the readiness phase.

The obtained location information is examined to inform subsequent inferences, which can result in a number of outcomes including but not limited to the expansion of the original device pool, targeted selection of one or more additional devices from the existing pool, or context redefinition. For example, if a unique location is uncovered based on a given SSID, one may be interested in identifying other devices that were observed as probing for the same SSID and, thus, were likely to share the location with the original device. Consequently, the outcomes of SSID geolocation could be used to formulate hypotheses.

From a technology perspective, the progression through the investigation phase that includes SSID
geolocation can be designed to leverage existing visualization tools such as Maltego – this approach is already implemented by a number of open-source tools that facilitate 802.11 trace capture and analysis such as snoopy-ng and Watcher [23, 38]. Figure 9 shows a screenshot depicting a possible device and SSID information visualization method. An application like Maltego could also be used to communicate these findings and facilitate the subsequent reconstruction and review processes.

5. Discussion

From a forensic perspective, SSID geolocation presents a number of opportunities and challenges. A complete and current source of AP locations can facilitate successful context-bound lookups returning possibly unique locations, as observed on the basis of the dataset contextualized to the specific metropolitan area. Certain SSID naming practices may also support the viability of this technique, as they facilitate the deployment of SSIDs that are less prone to deliberate or accidental reuse.

Our analysis of SSID string values indicates the potential practicality of automated inference based on arbitrary thresholds applicable to the characteristics of string randomness. As an additional attribute, the MAC address of the AP (BSSID) could be used to support the linkage inferences – in other words, claim that a particular device was connected to a particular access point at least once in case the device is also available for inspection and BSSIDs can be discovered and extracted.

However, if the required level of readiness is not achieved or no reliable location source is available, the associated location inferences are unlikely to bear any forensic value. Despite this, the approach can still be used as part of intelligence activities.

Although findings based on the studied dataset may be encouraging, an analysis of alternative context-bound sets is likely to produce different results. Consider an area where the majority of access points are represented by routers with default vendor SSIDs. Therefore, the described techniques are perceived to provide a varying degree of practical success. Nevertheless, the addition of the locational dimension to the digital fingerprint brings it a step closer to the physical world and can provide some basis for linking the device to its owner.

We also recognize that this technique is likely to be ineffective against devices that belong to individuals who are able and willing to modify their MAC address or PNL composition as a possible evasion measure. Finally, we suggest SSID geolocation as a supplementary mechanism that is not intended for utilization as a sole input into the finding development process.

6. Conclusion and future directions

In this paper, we discussed the SSID geolocation technique, which has noteworthy forensic potential and implications. We also presented some of the associated opportunities, challenges and limitations. Most importantly, we observed that the analyzed area provides a viable basis for possibly unique location inference, in part due to the peculiarities of the associated SSID naming and deployment practices.

At the same time, we acknowledged that our observations are heavily dependent upon the specifics of the analyzed data set, and that our findings are likely to be applicable only to the represented geographical area. In future work, we will study additional data sets and will further examine the iterative context definition process.

7. Acknowledgements

We would like to thank Robert Hagemann and the entire team behind the Wigle.net service for providing the examined data set and permitting extended access to their database.

8. References