On the Difficulty of NOT being Unique: Fingerprinting Users from Wi-Fi Data in Mobile Devices

Mariana Cunha
CRACS/INESC TEC, CISUC, and Department of Computer Science, Faculty of Sciences, University of Porto
Porto, Portugal
mccunha@dei.uc.pt

Ricardo Mendes
CISUC and Department of Informatics Engineering, University of Coimbra
Coimbra, Portugal
rscmendes@dei.uc.pt

Yves-Alexandre de Montjoye
Imperial College London, Exhibition Road, South Kensington
London, United Kingdom
demontjoye@imperial.ac.uk

João P. Vilela
CRACS/INESC TEC, CISUC, and Department of Computer Science, Faculty of Sciences, University of Porto
Porto, Portugal
jvilela@fc.up.pt

Abstract

The pervasiveness of mobile devices has fostered a multitude of services and applications, but also raised serious privacy concerns. In order to avoid users' tracking and/or users' fingerprinting, smartphones have been tightening the access to unique identifiers. Nevertheless, smartphone applications can still collect diverse data from available sensors and smartphone resources. Using real-world data from a field study we performed, this paper demonstrates the possibility of fingerprinting users from Wi-Fi data in mobile devices and the consequent privacy impact. From the performed analysis, we concluded that a single snapshot of a set of scanned Wi-Fi BSSIDs (MAC addresses) per user is enough to uniquely identify about 99% of the users. In addition, the most frequent Wi-Fi BSSID is sufficient to re-identify more than 90% of the users, a percentage that goes up to 97% of the users with the top-2 scanned BSSIDs. The Wi-Fi SSID (network name) also leads to a re-identification risk of about 83% and 97% with 1 and 2 of the strongest Wi-Fi Access Points (APs), respectively.

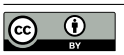
CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Ubiquitous and mobile computing**;

Keywords

Privacy, Mobile Devices, Fingerprinting, Re-identification Risk, Wi-Fi

ACM Reference Format:
Mariana Cunha, Ricardo Mendes, Yves-Alexandre de Montjoye, and João P. Vilela. 2025. On the Difficulty of NOT being Unique: Fingerprinting Users from Wi-Fi Data in Mobile Devices. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), March 31-April 4, 2025, Catania, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3672608.3707966



SAC '25, March 31-April 4, 2025, Catania, Italy

ACM ISBN 979-8-4007-0629-5/25/03
https://doi.org/10.1145/3672608.3707966

1 Introduction

The prevalence of smartphones in the current digital society has brought a rich opportunity to collect large amounts of heterogeneous data in a multitude of contexts. While beneficial to both users and services, such data might contain sensitive information and raise serious privacy concerns [11]. Within the smartphone's context, several works have demonstrated possible manners of leaking data through applications [6, 38], logs [28], and misuse of permissions [2, 4, 21, 44] (e.g. location tracking without permissions). In addition, recent news [9, 23] report examples of privacy-breaches, where applications were being used to collect side information, such as Wi-Fi data, without user consent.

In an attempt to enhance user's control and rights over their personal data, regulations on information privacy have been created, namely the General Data Protection Regulation (GDPR) in European Union [18]. In practice, smartphones give users some control through permission managers, where users can allow/deny permissions and consequent access to smartphone data/resources. Nonetheless, one of the main challenges when regulating and protecting user's privacy is the access to unique identifiers (IDs). In this regard, Android defines best practices for developers related to the selection and use of unique IDs [14], such as choosing user-resettable IDs, avoiding hardware IDs, or respecting the purpose of the advertising ID. The latest versions of Android [5] restrict the access to hardware identifiers (e.g. IMEI and serial number) to applications that are device or profile owner applications, have special carrier permissions, or have the READ_PRIVILEGED_PHONE_-STATE privileged permission [14].

Despite the developments to empower users to regain control over their data [27, 30], the current permission model of smartphones still has limitations and fails to account for data correlation and contextual dependency. For instance, while the permission to access location data is considered as runtime/dangerous and, hence, requires a permission prompt, obtaining location through side information, such as Wi-Fi [2, 4, 49] or Bluetooth, is still possible without explicit permission [44]. This raises serious privacy concerns that go beyond physical safety, since human mobility traces are highly unique and might reveal the user's identity, habits, social relationships or even health conditions [7, 12, 36]. Taking this into

consideration, this paper demonstrates the possibility of using data available to smartphone applications (e.g. Wi-Fi Access Points (APs) BSSIDs and SSIDs) as a fingerprint of users, even when other types of identifiers are blocked from access.

This paper shows the fingerprinting and re-identification risk of accessing Wi-Fi data through an installed smartphone application, that is, from the user's perspective. In spite of the efforts to avoid the use of other and more explicit unique identifiers and the well-known risks that advent from accessing such data, we conclude on the possibility of using Wi-Fi data as unique identifiers to fingerprint users. Until Android 9 (circa 2018), access to Wi-Fi data was possible with install-time permissions, which means without the user perception and without the possibility of revoking such access. This problem is transversal to other devices that still allow access to either location or Wi-Fi data (e.g. laptops). Despite the enhancements in more recent Android permissions, the fact that full network access is still asked in 99% of the smartphone applications, view Wi-Fi connections in 72% of the apps, and that one third of the apps request location permissions [16], emphasizes the importance of exploring the user's fingerprinting through location and Wi-Fi data.

Towards this goal, we relied on real-world data collected from a field study conducted with participants that carried our smartphones for at least one week [29, 30]. Relying on the contextual data (e.g. location and scanned Wi-Fi devices) collected through a smartphone application, we perform the fingerprinting and re-identification analysis. In this paper, we demonstrate that fingerprinting users with Wi-Fi data is possible and leads to a re-identification of about 99% of the users for a single snapshot of scanned Wi-Fi BSSIDs (i.e. an adversary with a single set of scanned Wi-Fi BSSIDs). When considering an adversary with information about the Wi-Fi AP with the strongest signal, nearly 83% and 97% of the users were re-identified with 1 and 2 highest signal strength SSIDs/network names, and 94% and 99% for the 1 and 2 highest signal strength BSSIDs. In line with these results, the re-identification risk assessment showed that 81 out of 82 users single scanned at least one Wi-Fi BSSID (MAC address), with the most frequent AP being enough to re-identify more than 90% of the users and the top-2 scanned BSSIDs sufficient to re-identify over 97% of the users. Similarly, over 95% of the users can be re-identified with the top-4 of Wi-Fi SSIDs, Wi-Fi locations or GPS locations (which is in line with previous work [12]). This highlights the privacy risks of accessing Wi-Fi information and is a call for action in raising user's privacy awareness and in the development of privacy-preserving mechanisms that take into account the data correlations among heterogeneous sources.

The remainder of this paper is structured as follows. Section 2 provides an overview of background concepts and related work, whereas Section 3 details the dataset. Section 4 presents the performed analysis on fingerprinting and evaluates the privacy risk through the re-identification metric. Section 5 discusses the privacy implications of Wi-Fi fingerprinting and corresponding privacy-preserving strategies, and Section 6 draws the main conclusions. Throughout the paper, smartphone application might simply be referred to as app.

2 Background and Related Work

The growing and indispensable use of smartphones has allowed the access to a variety of personal and sensitive data. This has fostered personalized services and novel applications that take into account the user's profile and preferences. With the claimed purpose of providing a better service suited to the user, smartphone applications collect Personally Identifiable Information (PII), such as unique identifiers (IDs), from mobile devices to distinguish devices/users, but also track them [33, 47].

The Android Operating System (OS) offers a number of IDs with different characteristics in terms of scope (i.e. which systems can access the ID), resettability and persistence (i.e. the lifespan of the ID and how it can be reset), uniqueness (i.e. the likelihood of collisions), and integrity protection and non-repudiability (i.e. a difficult-to-spoof ID) [14]. To ensure that the provided IDs are properly handled, Android defines best practices for developers related to the selection and use of unique IDs [14], such as choosing user-resettable IDs, avoiding hardware IDs, or respecting the purpose of the advertising ID. The latest versions of Android [5] restrict the access to hardware identifiers (e.g. IMEI and serial number) to applications that are device/profile owner apps (i.e. apps with administrative control over the device or a specific profile), have special carrier permissions (i.e. permissions limited to apps affiliated with mobile carriers), or have a special privileged permission (READ_PRIVILEGED_PHONE_STATE) only available to system apps that are part of the firmware or installed by the device manufacturer.

In spite of the introduced constraints and the official guidelines for working with Android IDs [14], previous research demonstrated that unique IDs and, specifically, persistent IDs are being accessed and often used for tracking users [31, 37, 38]. The authors of [31] identified 51 unique vulnerabilities that evidence the pervasive mishandling of user-unresettable identifiers (UUIs) in the latest Android phones. In this paper, we further demonstrate that the current countermeasures are insufficient and neglect the risk of identifying users through other and less explicit IDs. In particular, we conclude on the possibility of using Wi-Fi data as unique identifiers to fingerprint users.

Due to the privacy implications of exposing unique IDs, current research has been studying the uniqueness of human behavior from several contexts [52], including mobile apps usage data [1, 26, 42, 46] and mobility patterns [8, 12]. In terms of mobile apps usage data, four apps demonstrated being sufficient to uniquely re-identify about 90% of the users [1, 42, 46]. The analysis with this data goes further and allows re-identifying whether students are depressed or non-depressed [3]. Since Android 11 (circa 2020), accessing installed applications within a smartphone app is filtered by default and requires a QUERY_ALL_PACKAGES permission to query all installed apps on a device [15].

In regard to mobility patterns, the uniqueness of location data has become a concerning challenge. Human mobility traces are highly unique, which makes it possible to infer the user's identity, habits, social relationships or even health conditions [7]. In fact, four spatio-temporal points revealed being enough to re-identify about 95% of the individuals [8, 12]. This has led smartphone OSs to protect the collection of location data by requiring a permission to access a fine or coarse location. Nevertheless, these permissions

are often requested (by one third of the apps [16]) and apps with access to such data are able to build a fingerprint with locations to uniquely identify users [25].

Motivated by the large amounts of smartphone applications that have access to Wi-Fi information [16], we depart from previous works by demonstrating the uniqueness of such information and the possibility of creating a fingerprint through the scanned Wi-Fi access points. Despite the well-known risks of inferring location from side information, such as Wi-Fi [34], and the uniqueness of mobility traces through Wi-Fi [8], our work differs from the previous ones by focusing on the user's perspective (i.e. an app installed on the user's smartphone) and how the collection of scanned Wi-Fi APs can constitute a unique ID. Our goal is to show the privacy impact of fingerprinting users through Wi-Fi information and assess the re-identification risk that advent from scanning such data.

The existing literature has, in contrast, mainly focused on device fingerprinting from the perspective of Wi-Fi access points or other entities that are able to monitor wireless networks [50]. The fingerprints are commonly used by device identification systems that rely on relevant features to identify devices. Notwithstanding the potential benefits of using device fingerprints to enhance wireless security, several works demonstrated the possibility of not only tracking users/devices, but also using the available data (e.g. Wi-Fi probe requests sent by users' devices containing the MAC address that uniquely identifies the sending device [10, 20]) to infer information about the nearby users [32, 39, 45]. In particular, MAC address randomization emerged as a response to the resulting privacy violations [48]. In distinction to these studies, this paper examines a different perspective (user's perspective) of fingerprinting from Wi-Fi data through apps installed in users' devices, emphasizing the re-identification risk that results from the scanned Wi-Fi APs, and investigating how Wi-Fi data can constitute a unique ID in mobile devices.

3 Dataset and Overview

This section starts by describing the dataset that was used to study the privacy impact of fingerprinting users through Wi-Fi data and the resulting re-identification risk, followed by the analysis of installed applications and requested permissions.

3.1 Dataset Characterization

In order to study the user's fingerprinting within the smartphone's context, we selected the COP-MODE dataset [29, 30]. This dataset was collected in a real-world field study with 93 users, where the participants carried smartphones for at least one week with their personal applications pre-installed and an application responsible for the data collection. This app prompts users at every permission check (see Figure 1) and collects their input, as well as other contextual features at the time of the prompt. In this paper, we focus on specific contextual features that are of relevance to this work, namely:

- Datetime: timestamp of the request permission prompt.
- Location: timestamp, latitude, longitude, and accuracy.
- Wi-Fi: timestamp, BSSID, SSID, and RSSI for each scanned device.
- Semantic location: the semantic location was collected from the user input, whose possibilities were: *home*, *work*, *traveling* or *other*.

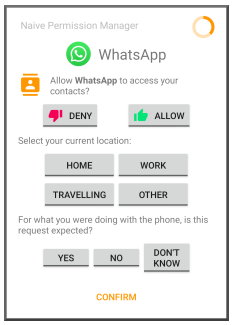


Figure 1: An example of a permission prompt issued as a result of the app *WhatsApp* checking for the contacts permission.

We should note that location data is related to the last known location reading [13] and might not correspond to the current location, since the participant might have turned location off. With respect to Wi-Fi data, the scanned Wi-Fi devices correspond to the devices in the neighborhood that were obtained from a scan attempted every 5 minutes. These considerations will be taken into account during the exploratory data analysis that follows.

The COP-MODE dataset is composed by 2180302 permission requests from 93 participants. 65261 (2.99%) of the total requests were answered by participants, while the remaining were either unhandled or answered by the 30 minutes cache, timeouts or dismissed. In this work, we will consider the user answered requests, since only these have the selected semantic location. From the 65261 answered requests, 41602 requests (63.75% of the user answered requests) have Wi-Fi and/or location information from a total of 82 participants. This data constitutes the target of analysis in this paper.

3.2 Installed Applications and Requested Permissions

To better understand the applications that have access to Wi-Fi data and, in this way, understand the relevance of this problem, we start by studying the context of installed applications and requested permissions. In Android, applications are divided into system (i.e. apps with system privileges) and non-system apps (i.e. apps with limited privileges). The COP-MODE dataset contains a total of 3926 distinct apps and 1737 non-system distinct apps. Regarding the requested permissions, we now analyze which apps request the permissions required to access Wi-Fi data and the respective grant/deny result.

Scanning nearby Wi-Fi devices have been changing with the release of new Android versions, with the first mandatory location-restrictions introduced in Android 9. Considering that the COP-MODE dataset was collected in Android 9 devices, an app would require the ACCESS_COARSE_LOCATION or ACCESS_FINE_LOCATION permission along with CHANGE_WIFI_STATE permission to start a scan of Wi-Fi devices and ACCESS_WIFI_STATE

permission to obtain the scanned Wi-Fi devices. In spite of the introduced constraints, where such access to scanned Wi-Fi devices also requires a location permission, the COP-MODE dataset shows that there are over 300 applications (≈10% of the total apps) that satisfy these conditions, which is corroborated by a recent work [43] that demonstrates that these permissions are among the most frequently requested in apps spanning multiple Android versions. Furthermore, Wi-Fi related permissions are classified as install-time permissions and, hence, are automatically granted when the app is installed and cannot be revoked, which still enables an app to access certain Wi-Fi information (e.g. RSSI) and/or change the Wi-Fi status without an explicit request.

From the performed analysis, the ACCESS_WIFI_STATE and the CHANGE_WIFI_STATE permissions were requested and automatically granted by 1499 (≈38%) apps and 581 (≈15%) apps, respectively. In addition, the location-related permissions were requested by ≈24% of the apps, with over 50% of granted permission requests. This is in line with recent analysis [16] that claims that full network access is still asked in 99% of the smartphone apps, view Wi-Fi connections in 72% of the apps, and that one third of the apps request location permissions. Such permissions are requested by diverse applications that can be categorized according to the Google Play Store. Figure 2 presents the percentage of distinct apps from each category in where each of the referred permissions is requested. Depending on the app category as well as the user expectation on the app objective, the permission decisions might be affected [29]. For instance, TRAVEL_AND_LOCAL category is expected to request the location permission and, consequently, users tend to grant it. On the other hand, since Wi-Fi related permissions are automatically granted independently of the app category and the users' preferences, privacy risks arise, leading to fingerprinting and identification of users.

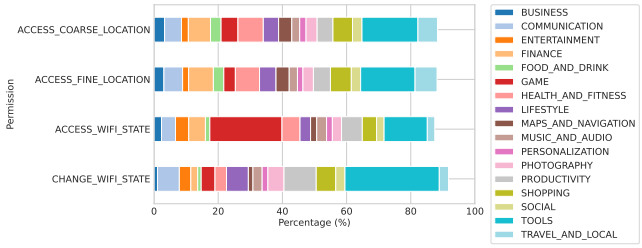


Figure 2: Percentage of distinct apps from each category in where the permission (y axis) was requested. Categories with a percentage of apps inferior to 1% were removed from the plot to simplify visualization.

4 Fingerprinting and Re-identification Risk from Wi-Fi Data

Building on the considerable number of applications that are able to collect either location and/or Wi-Fi data, this section performs an analysis on the privacy implications of collecting Wi-Fi data by studying the resulting fingerprinting and re-identification risk. Throughout this analysis, we shall use Wi-Fi SSID and Wi-Fi BSSID to refer respectively to Wi-Fi network name and MAC Address of the Wi-Fi Access Point (AP).

4.1 Fingerprinting Users from Wi-Fi Data

A fingerprint consists in a combination of features that uniquely identify individuals. While there might be benefits from this, such as personalized services, the fact that applications can uniquely identify their users might pose threats to user's privacy. In order to mitigate this problem, smartphone OSs have tightened the access to unique identifiers, as previously mentioned. This section addresses the construction of fingerprints from Wi-Fi data. An initial fingerprint can be composed of all the scanned Wi-Fi BSSIDs per participant during the COP-MODE field study. In this case, all users would have a unique fingerprint, since users have a unique set of scanned BSSIDs throughout the entire field study. However, generally it may not be possible to access all historical data of Wi-Fi BSSIDs because the full information about configured Wi-Fi networks is restricted to Device Owner (DO), Profile Owner (PO) and system apps since Android 10 (circa 2019) [5]. Therefore, we will consider the following more realistic fingerprinting setups:

(1) The attacker has access to a single snapshot of all scanned Wi-Fi BSSIDs, other than continuous access to scanned Wi-Fi networks. This corresponds to a weaker attacker model that has access to much fewer information (single snapshot) of scanned Wi-Fi networks. This can correspond to a situation in which an app is installed and immediately uninstalled;
(2) The attacker has access to the SSID (network name) of the scanned Wi-Fi network with the highest signal strength (RSSI), thus representing the Wi-Fi network the user would usually connect to.

In the first scenario, assuming an adversary model where the attacker has access to less information (e.g. a subset of data instead of whole data), we consider a random snapshot in time and the respective set of scanned Wi-Fi BSSIDs. For statistical significance, each selection was performed 100 times and, thus, the results will be presented with the confidence interval of 95%. From the results, a single snapshot of scanned Wi-Fi BSSIDs per user creates a unique fingerprint for about 99% of the participants, which means that collecting the set of scanned Wi-Fi APs once is enough to uniquely identify more than 99% of the users. This percentage goes up to 100% when considering three snapshots in time, as graphically represented in Figure 3.

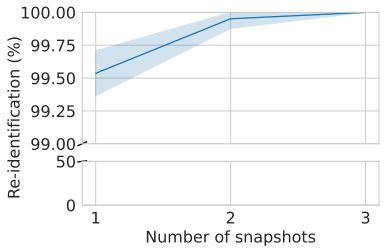
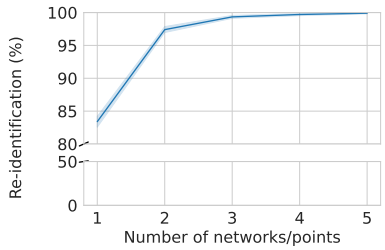


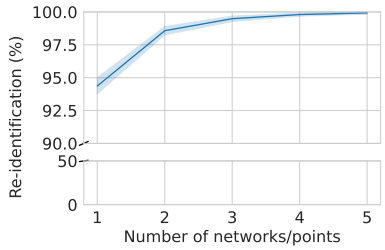
Figure 3: Percentage of identified users and respective confidence intervals of 95% when considering the selection of random snapshots (x axis) and the set of scanned Wi-Fi BSSIDs.

In the second scenario, we assume the access to the SSID of the network with the highest signal strength (RSSI), which represents the case when an adversary has access to the networks the

users usually connect to [35]. This information can be obtained in runtime from the current Wi-Fi connection. Figure 4a represents the results for this scenario, where the availability of a varying number of connected Wi-Fi SSIDs (network names) was considered for calculating the risk of unique fingerprinting. This rate goes from ≈83% when considering a single network name, ≈97% when considering 2 network names, up to ≈100% when considering 5 network names. If instead of the network name one has access to the BSSID/MAC address, its unicity makes the fingerprinting risk grows to 94% with a single instance as depicted in Figure 4b. This is justified by the multiple Wi-Fi APs that share the same SSID (e.g. eduroam (education roaming), that is, a world-wide roaming access service that provides Internet connectivity across University campus). These results show the relevance of information derived from Wi-Fi connections as a mean to fingerprint users, highlighting that 5 names of connected Wi-Fi networks suffice to re-identify all users in the dataset. The results discussed in this section emphasize the privacy risks of exposing either nearby or connected Wi-Fi AP devices and corresponding uniqueness.



(a) Wi-Fi SSID



(b) Wi-Fi BSSID

Figure 4: Percentage of identified users and respective confidence intervals of 95% when considering the selection of random snapshots in time (x axis) and the Wi-Fi SSID/BSSID with the highest signal strength (RSSI).

4.2 Re-identification Risk through Top-N

Building on the fingerprinting analysis, this section assesses the re-identification risk through Wi-Fi data. The re-identification risk is a relevant privacy metric that evaluates the possibility of exposing private information of a certain individual. One of the biggest concerns in data privacy is related to identity disclosure, commonly mitigated by removing explicit identifiers. However, there are other attributes, also known as Quasi-Identifiers (QIDs), that can generate a unique combination and enable user re-identification. In order to assess the re-identification risk, we adapted the well-known "top-N" locations attack [51] to a "top-N" features/attributes attack. This attack consists in the selection of the top-N features/attributes of a user (i.e. the N most frequent) and an assessment of its uniqueness. If the selected top is unique, then the user is considered identified.

Figure 5 starts by presenting a semantic analysis of the top-N Wi-Fi AP BSSID per user, with N from 1 to 3. Considering the selected semantic location (*Home*, *Work*, *Traveling* or *Other*), we are able to categorize the most frequent location of the top-N Wi-Fi BSSIDs within an interval of 5 minutes. The baseline represented in the chart consists in the distribution of the scanned BSSIDs per semantic location, where ≈84% were at home, ≈9% at work, ≈4% in other location, and ≈3% while traveling. As expected, the top-1 (i.e. most frequent BSSID) corresponds to the home location in more than 90% of the situations. Similarly, top-2 and top-3 contain the *home* location in more than 85% of the cases, which can be explained by the fact that users might scan more than one Wi-Fi AP at home that will be in the most frequently scanned BSSIDs. Comparing with the baseline, the main difference is on the distribution of the remaining locations, where *other* location occurs with a higher percentage in the most frequent locations than in the baseline. The performed analysis stresses the need of protecting the end-points of users' trajectories, specifically home and work locations, due to their potential for user re-identification [22].

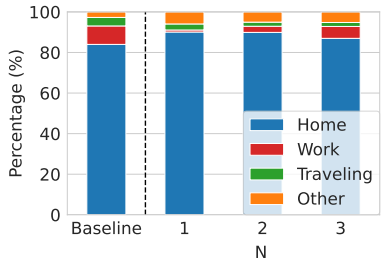


Figure 5: Semantic analysis of the scanned Wi-Fi BSSIDs (baseline) and the top-N Wi-Fi BSSID per user with N from 1 to 3 within a 5 minute interval between the collected data and the permission request prompt.

Figure 6 presents the percentage of re-identified users when considering the top-N attack (N from 1 to 7) for the following features: GPS location, Wi-Fi location, Wi-Fi SSID, and Wi-Fi BSSID. In line with the message from previous works [8, 12], four points are enough to uniquely identify over 95% of the users. From Figure 6, the lower value of re-identification occurs for the Wi-Fi SSID top-1, which can be explained by the number of public Wi-Fi hotspots with common names, such as those provided by Internet Service Providers (ISPs). On the other hand, over 90% of the users can be re-identified through the most scanned Wi-Fi BSSID and over 97% by the top-2. This is especially concerning since, as supported by the semantic analysis, the most frequent locations (i.e. top locations) are related to a private location: *home*.

While the re-identification risk assessment relied on the most frequent Wi-Fi APs per user in these results, the fingerprinting analysis presented in Section 4.1 considered two adversary scenarios, where the attacker has access to (1) a set of scanned BSSIDs per user and (2) the SSID with the highest signal strength. These approaches justify the differences in the re-identification percentages. For instance, the top-1 BSSID of Figure 6 considers the most

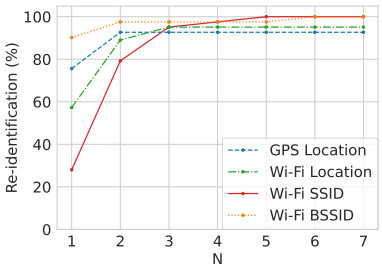


Figure 6: Percentage of identified users considering the top-N attack (N from 1 to 7) for the following features: GPS location, Wi-Fi BSSID, Wi-Fi SSID, and Wi-Fi location.

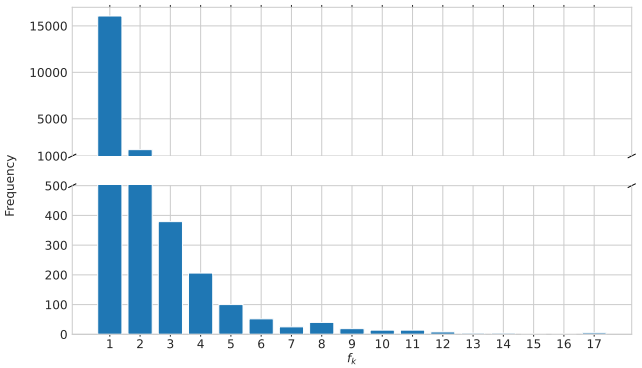


Figure 7: Frequency of the records f_k for the Wi-Fi BSSIDs.

frequent BSSID per user, whereas the random snapshot in time of Figure 3 contains all scanned Wi-Fi BSSIDs in that period of time, hence explaining the higher re-identification risk. The next section demonstrates the risk of re-identification through k-anonymity, a privacy principle that could be used in an attempt to minimize the risk of fingerprinting and re-identification.

4.3 Re-identification Risk through k-anonymity

A common approach to protect the user's privacy and mitigate the re-identification risk consists of reducing the data uniqueness. For instance, the k-anonymity principle guarantees that in a set of k individuals, the identity of each one cannot be disclosed from at least k-1 individuals in the same set [40, 41]. The achieved privacy level can be measured by the value of k, such that a higher value of k corresponds to a higher privacy level (i.e. it is harder to de-anonymize). Based on this concept, a user can be *singled out* if for a given number of combinations k of a set of Quasi-Identifiers (QIDs), the frequency of records f_k that have the same combination of QIDs is one. This section relies on this concept as a metric to assess the re-identification risk [17].

For k=1, where each set is composed by one of the 18613 unique scanned Wi-Fi BSSIDs in the COP-MODE dataset, we compute the frequency of records that have the same QID (i.e. the users that scanned the same BSSID). If $f_k = 1$, then the user is considered singled out and, hence, re-identified. Figure 7 presents the frequency of records f_k and the respective count. From these results, $f_k = 1$ for 16064 Wi-Fi BSSIDs, signifying that ≈86% of the Wi-Fi BSSIDs are scanned by only one user. In this case, 81 out of the 82 users single scanned one of these Wi-Fi APs at least once, which means that an individual could be re-identified by knowing a scanned BSSID within this set.

The assessment performed in this section concludes the analysis of this paper by emphasizing the uniqueness of Wi-Fi data (≈86% of the Wi-Fi BSSIDs are single scanned) and the resulting possibility of using such data as unique identifiers. This has serious implications for the user's privacy as discussed next.

5 Privacy Implications of Wi-Fi Fingerprinting

With the claimed purpose of enhancing personalized services aligned with users' preferences and behaviors, applications are allowed to collect diverse data. However, this is achieved at the cost of compromised users' privacy and lack of anonymity. In this paper, we have shown that a considerable amount of applications have the required permissions to collect location and/or Wi-Fi data, thus leading to a high fingerprinting and re-identification risk. This stresses the difficulty of not being uniquely identified by apps even with the existing restrictions to access explicit unique IDs in smartphones. The ability to uniquely identify users can be used to launch further attacks to compromise users' privacy, resulting in serious privacy implications that will be discussed in this section.

Regardless of the innumerous opportunities and benefits for both users and service providers, fingerprinting and profiling are examples of severe privacy concerns in the current digital society that are worsened by the high data uniqueness. Profiling consists of creating detailed and accurate models of users based on their data, making it possible to identify and track them. The resulting privacy risks are exacerbated when data correlations or other linkable data are present. Building on the analysis performed in this paper, where we emphasized the uniqueness of Wi-Fi data, an attacker would be able to create users' profiles based on Wi-Fi data that could be enriched with other contextual information (e.g. hour of the day or location) to fingerprint or even track users. For illustration, tracking users would allow entities to know whether the user is at home or not, which poses privacy risks that go beyond physical safety. This is a real and increasingly worrying problem, as demonstrated in recent news [23, 24]. The privacy breach presented in [23] reports a real-world battery monitor app that was able to collect and share GPS coordinates, nearby cell phone towers, Wi-Fi access points, and also the user's street address. This is especially critical since location data (even if anonymized) can be used to profile users and infer sensitive information (e.g. religion, as exposed in the news article [19]). These real-world examples accentuate the insights of this paper as enablers of potential privacy issues.

Due to the subtle and pervasive way in which fingerprinting can be performed, finding a trade-off between effective fingerprinting and user's privacy is still a major challenge in the current privacy threats' landscape. Although users are often unaware of the large amounts of data that are collected about them and how this information can be used to identify their profiles, the pervasiveness of mobile devices has fostered a rich opportunity to collect personal data. As discussed in this paper, despite the efforts to restrict the access to unique IDs, fingerprinting users is still possible through Wi-Fi data. This is a transversal problem for mobile devices in general, not only present in smartphones.

5.1 Privacy-Preserving Strategies

To mitigate the discussed privacy implications, it is crucial to implement data protection mechanisms and strategies that further safeguard users' privacy, as summarized below.

- Limiting data collection: Limiting the collection of unnecessary data that may never be used, specifically preventing the collection and retention of personal data to reduce users' exposure to privacy risks.
- Risk analysis and re-identification risk assessment: The risk analysis and re-identification risk assessment should be performed in a more systematic manner, and be part of the procedure for warranting access to any data types.
- Usage of anonymization methods: Upon identification of privacy risks, usage of anonymization techniques should be considered to provide adequate privacy-utility trade-offs. In particular, except for user location (for which obfuscation is already available in recent mobile OSes), current permission systems typically operate on an all-or-nothing basis, meaning that either users have access to the data (full utility, no privacy) or do not have access at all. A middle ground through anonymization methods should be explored, yet through automated privacy protection mechanisms, since the average user is not available/knowledgeable enough to setup and configure such techniques.
- Promoting data transparency: Companies and organizations should have a clear understanding of who they are interacting with, warranting that the users' digital identity is secure, and privacy is respected. In addition, methods that allow users to exercise their privacy rights should be provided, such as controlling the use and disclosure of sensitive personal data.
- Enhancing user awareness and education: Users should be aware of the privacy implications of their online activities and identities, but also educated about privacy to make informed decisions about data sharing and protection.

These privacy-preserving strategies aim at protecting users from potentially unauthorized access to their data. As one of the most widely used devices, smartphones have been at the forefront of introducing restrictions that safeguard users' privacy, however, much still remains to be done, particularly when other mobile devices are concerned.

6 Conclusion

The proliferation of mobile devices has fostered a multitude of services and applications, but also a rich opportunity to collect large amounts of data. In particular, the high demand for personalized services has been leading to a special interest on uniquely identifying users. Due to the privacy risks that advent from the data uniqueness, smartphones have tightened the access to unique identifiers (IDs). In this paper, resorting to a dataset that collected user data for at least one week, we demonstrate the possibility of fingerprinting users through location and Wi-Fi data, showing that it is still possible to fingerprint users by relying on available data despite the efforts to avoid unique IDs. From the performed analysis, we concluded that *a single snapshot of a set of scanned Wi-Fi BSSIDs (MAC addresses) per user is enough to uniquely identify about 99% of the users. The most frequent Wi-Fi BSSID is sufficient to re-identify more than 90% of the users and goes up to 97% of the users with the top-2 scanned BSSIDs.* Moreover, having access to the strongest Wi-Fi AP, that the users would connect to, leads to a re-identification risk of about 83% and 97% with 1 and 2 of the strongest SSIDs (network names), respectively. Thus, based on our results and according to the COP-MODE dataset, *more than 300 applications are able to use location and/or Wi-Fi data as unique identifiers, even when other types of identifiers are blocked from access.* While fingerprinting can be beneficial to provide personalized services that are aligned with users' preferences and behaviors, the consequences and resulting privacy risks cannot be ignored. Users are often unaware of the data collected through their mobile devices and, specifically, through the installed smartphone applications. In fact, due to the subtle and pervasive ways in which information is collected, there is a difficulty to control and/or opt-out of such collection, as well as a lack of techniques to mitigate the users' tracking. For illustration, tracking users would allow entities to know whether the user is at home or not, which poses privacy risks that go beyond physical safety. Therefore, finding a trade-off between effective fingerprinting and user's privacy protection is still a critical challenge in the current landscape of always connected mobile devices.

Acknowledgment

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020). The authors wish to acknowledge the support of the project CISUC - UID/CEC/00326/2020 and the European Social Fund through the Regional Operational Program Centro 2020 of FCT, and the project PRIVATEER funded by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101096110. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or SNS JU. Mariana Cunha wishes to acknowledge financial support by the Portuguese funding institution Fundação para a Ciência e a Tecnologia (FCT) under the grant 2020.04714.BD (DOI 10.54499/2020.04714.BD).

References

[1] Jagdish Prasad Achara, Gergely Acs, and Claude Castelluccia. 2015. On the Unicity of Smartphone Applications. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society (WPES '15)*. ACM, 27–36.

[2] Jagdish Prasad Achara, Mathieu Cunche, Vincent Roca, and Aurélien Francillon. 2014. Short paper: Wifileaks: Underestimated privacy implications of the ACCESS_WIFI_STATE Android permission. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks*. ACM, 231–236.

[3] Md Sabbir Ahmed and Nova Ahmed. 2021. Exploring unique app signature of the depressed and non-depressed through their fingerprints on apps. In *International Conference on Pervasive Computing Technologies for Healthcare*. Springer, 218–239.

[4] Efthimios Alepis and Constantinos Patsakis. 2017. There's Wally! Location Tracking in Android without Permissions. In *Proceedings of the 3rd International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,*. INSTICC, SciTePress, 278–284.

[5] Android. 2024. Android 10. https://developer.android.com/guide/topics/connectivity/wifi-scan. Accessed: 2024-07-05.

[6] John S. Atkinson, John E. Mitchell, Miguel Rio, and George Matich. 2018. Your WiFi is leaking: What do your mobile apps gossip about you? *Future Generation Computer Systems* 80 (2018), 546–557. https://doi.org/10.1016/j.future.2016.05.030

[7] Benjamin Baron and Mirco Musolesi. 2020. Where you go matters: a study on the privacy implications of continuous location tracking. *Proceedings of the ACM*

on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 4 (2020), 1–32.

[8] Antoine Boutet and Sonia Ben Mokhtar. 2021. Uniqueness assessment of human mobility on multi-sensor datasets. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*. ACM, 1–10.

[9] Federal Trade Commission. 2016. Mobile Advertising Network InMobi Settles FTC Charges It Tracked Hundreds of Millions of Consumers' Locations Without Permission. https://www.ftc.gov/news-events/news/press-releases/2016/06/mobile-advertising-network-inmobi-settles-ftc-charges-it-tracked-hundreds-millions-consumers. Accessed: 2024-07-17.

[10] Mathieu Cunche, Mohamed Ali Kaafar, and Roksana Boreli. 2012. I know who you will meet this evening! linking wireless devices using wi-fi probe requests. In *2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 1–9.

[11] Mariana Cunha, Ricardo Mendes, and João P Vilela. 2021. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer science review* 41 (2021), 100403.

[12] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3, 1 (2013), 1–5.

[13] Android Developers. 2023. Get the last known location. https://developer.android.com/training/location/retrieve-current. Accessed: 2024-07-05.

[14] Android Developers. 2024. Best practices for unique identifiers. https://developer.android.com/identity/user-data-ids. Accessed: 2024-07-05.

[15] Android Developers. 2024. Package visibility filtering on Android. https://developer.android.com/training/package-visibility. Accessed: 2024-07-05.

[16] Edvardas Mikalauskas. 2022. Android apps are asking for too many dangerous permissions. Here's how we know. https://cybernews.com/privacy/android-apps-are-asking-for-too-many-dangerous-permissions-heres-how-we-know/. Accessed: 2024-07-05.

[17] Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association* 15, 5 (2008), 627–637.

[18] European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council.* http://data.europa.eu/eli/reg/2016/679/oj

[19] Lorenzo Franceschi-Bicchierai. 2015. Redditor cracks anonymous data trove to pinpoint Muslim cab drivers. https://mashable.com/archive/redditor-muslim-cab-drivers. Accessed: 2024-11-11.

[20] Julien Freudiger. 2015. How talkative is your mobile device? an experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks (WiSec '15)*. ACM, Article 8, 6 pages.

[21] Julien Gamba, Álvaro Feal, Eduardo Blazquez, Vinuri Bandara, Abbas Razaghpanah, Juan Tapiador, and Narseo Vallina-Rodriguez. 2023. Mules and Permission Laundering in Android: Dissecting Custom Permissions in the Wild. *IEEE Transactions on Dependable and Secure Computing* (2023), 1–18.

[22] Philippe Golle and Kurt Partridge. 2009. On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing*. Springer, 390–397.

[23] haxrob. 2023. Discovering that your car battery monitor is siphoning up your location data. https://haxrob.net/discovering-that-your-bluetooth-car-battery-monitor-is-siphoning-up-your-location-data/. Accessed: 2024-11-11.

[24] Misha Rykov Jen Caltrider and Zoë MacDonald. 2023. Discovering that your car battery monitor is siphoning up your location data. https://foundation.mozilla.org/en/privacynotincluded/articles/its-official-cars-are-the-worst-product-category-we-have-ever-reviewed-for-privacy/. Accessed: 2024-11-11.

[25] Kathryn Zickuhr. 2013. Main Report. https://www.pewresearch.org/internet/2013/09/12/location-based-services-2/. Accessed: 2024-07-05.

[26] Tong Li, Tong Xia, Huandong Wang, Zhen Tu, Sasu Tarkoma, Zhu Han, and Pan Hui. 2022. Smartphone app usage analysis: datasets, methods, and applications. *IEEE Communications Surveys & Tutorials* 24, 2 (2022), 937–966.

[27] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *12th Symposium on Usable Privacy and Security (SOUPS)*. 27–41.

[28] Allan Lyons, Julien Gamba, Austin Shawaga, Joel Reardon, Juan Tapiador, Serge Egelman, Narseo Vallina-Rodriguez, et al. 2023. Log: It's Big, It's Heavy, It's Filled with Personal Data! Measuring the Logging of Sensitive Information in the Android Ecosystem. In *Usenix Security Symposium*.

[29] Ricardo Mendes, André Brandão, João P Vilela, and Alastair R Beresford. 2022. Effect of user expectation on mobile app privacy: a field study. In *2022 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 207–214.

[30] Ricardo Mendes, Mariana Cunha, João P Vilela, and Alastair R Beresford. 2022. Enhancing User Privacy in Mobile Devices Through Prediction of Privacy Preferences. In *Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part I*. Springer, 153–172.

[31] Mark Huasong Meng, Qing Zhang, Guangshuai Xia, Yuwei Zheng, Yanjun Zhang, Guangdong Bai, Zhi Liu, Sin G Teo, and Jin Song Dong. 2023. Post-GDPR Threat Hunting on Android Phones: Dissecting OS-level Safeguards of User-unresettable Identifiers. In *30th annual Network and Distributed System Security Symposium*.

[32] ABM Musa and Jakob Eriksson. 2012. Tracking unmodified smartphones using wi-fi monitors. In *Proceedings of the 10th ACM conference on embedded network sensor systems*. Association for Computing Machinery, 281–294.

[33] Suman Nath. 2015. MAdScope: Characterizing Mobile In-App Targeted Ads. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services* (Florence, Italy) *(MobiSys '15)*. Association for Computing Machinery, New York, NY, USA, 59–73. https://doi.org/10.1145/2742647.2742653

[34] Le Nguyen, Yuan Tian, Sungho Cho, Wookjong Kwak, Sanjay Parab, Yuseung Kim, Patrick Tague, and Joy Zhang. 2013. Unlocin: Unauthorized location inference on smartphones without being caught. In *2013 International Conference on Privacy and Security in Mobile Systems (PRISMS)*. IEEE, 1–8.

[35] Changhua Pei, Zhi Wang, Youjian Zhao, Zihan Wang, Yuan Meng, Dan Pei, Yuanquan Peng, Wenliang Tang, and Xiaodong Qu. 2017. Why it takes so long to connect to a WiFi access point. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[36] The Associated Press. 2021. Priest outed via Grindr app highlights rampant data tracking. https://www.nbcnews.com/tech/security/priest-outed-grindr-app-highlights-rampant-data-tracking-rcna1493. Accessed: 2024-07-05.

[37] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, Phillipa Gill, et al. 2018. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *The 25th annual Network and Distributed System Security Symposium (NDSS)*.

[38] Joel Reardon, Álvaro Feal, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, and Serge Egelman. 2019. 50 ways to leak your data: An exploration of apps' circumvention of the android permissions system. In *28th USENIX security symposium (USENIX security 19)*. USENIX Association, Santa Clara, CA, 603–620.

[39] Fergus Ryan and Michael Schukat. 2019. Wi-fi user profiling via access point honeynets. In *2019 30th Irish Signals and Systems Conference (ISSC)*. IEEE, 1–4.

[40] Pierangela Samarati and Latanya Sweeney. 1998. Generalizing data to provide anonymity when disclosing information. In *PODS*, Vol. 98. Citeseer, 188.

[41] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical Report. technical report, SRI International.

[42] Vedran Sekara, Laura Alessandretti, Enys Mones, and Håkan Jonsson. 2021. Temporal and cultural limits of privacy in smartphone app usage. *Scientific reports* 11, 1 (2021), 1–9.

[43] Yash Sharma and Anshul Arora. 2024. A comprehensive review on permissions-based Android malware detection. *International Journal of Information Security* (2024), 1–36.

[44] Vincent Toubiana and Mathieu Cunche. 2021. No need to ask the Android: Bluetooth-Low-Energy scanning without the location permission. In *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks* (Abu Dhabi, United Arab Emirates) *(WiSec '21)*. ACM, 147–152.

[45] Martin W Traunmueller, Nicholas Johnson, Awais Malik, and Constantine E Kontokosta. 2018. Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems* 72 (2018), 4–12.

[46] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (2018), 23 pages.

[47] Imdad Ullah, Roksana Boreli, and Salil S Kanhere. 2023. Privacy in targeted advertising on mobile devices: a survey. *International Journal of Information Security* 22, 3 (2023), 647–678.

[48] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S. Cardoso, and Frank Piessens. 2016. Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 413–424.

[49] Gabriella Verga, Salvatore Calcagno, Andrea Fornaia, and Emiliano Tramontana. 2019. Smart Cities and Open WiFis: When Android OS Permissions Cease to Protect Privacy. In *Internet and Distributed Computing Systems*. Springer International Publishing, Cham, 457–467.

[50] Qiang Xu, Rong Zheng, Walid Saad, and Zhu Han. 2015. Device fingerprinting in wireless networks: Challenges and opportunities. *IEEE Communications Surveys & Tutorials* 18, 1 (2015), 94–104.

[51] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 145–156.

[52] Wanyi Zhang, Qiang Shen, Stefano Teso, Bruno Lepri, Andrea Passerini, Ivano Bison, and Fausto Giunchiglia. 2021. Putting human behavior predictability in context. *EPJ Data Science* 10, 1 (2021), 42.