A Privacy-Preserving Mechanism for Requesting Location Data Provider with Wi-Fi Access Points

Antoine Boutet, Mathieu Cunche

A Privacy-Preserving Mechanism for Requesting Location Data Provider with Wi-Fi Access Points

Antoine Boutet
Insa-Lyon, CITI, Inria, France
antoine.boutet@insa-lyon.fr

Mathieu Cunche
Insa-Lyon, CITI, Inria, France
mathieu.cunche@insa-lyon.fr

ABSTRACT

With the democratization of mobile devices embedding different positioning capabilities, the location of users is now collected to track the location of users. When used for behavioral profiling, this tracking for enhancing raises more and more privacy concerns.Depending on the permissions, mobile applications can get a fine-grained user's location from the GPS or a coarse-grained location by requesting location data provider with surrounding Wi-Fi access points for instance. While using the GPS does not rely on external untrusted party, requesting a location data provider clearly exposes the location of users. Whereas location privacy has been an active research field this last decade, most of the contributions are performed on GPS-based data, and it is not clear how to efficiently protect Wi-Fi-based positioning to preserve the users' privacy. In this paper, we propose a novel solution to preserve users' privacy from curious location data providers when requesting users' location from Wi-Fi while supporting high-utility. The key idea behind our online approach is to combine a *random sampling* (for controlling the quantity of revealed information) and a *obfuscation scheme* (for ensuring privacy-preserving information disclosure). We exhaustively evaluate our solution with a real dataset of mobility traces collected through multiple sensors. We show that the proposed approach provides a trade-off between privacy (i.e., avoiding to reveal its true location) and utility (i.e., still benefiting from services such as places recommendation) fully controllable by the users. Lastly, we also discuss the integration of our protection scheme in mobile operating systems.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; *Mobile and wireless security*; • **Networks** → *Location based services*;

KEYWORDS

Location privacy, Location data provider, Wi-Fi-based positioning

1 INTRODUCTION

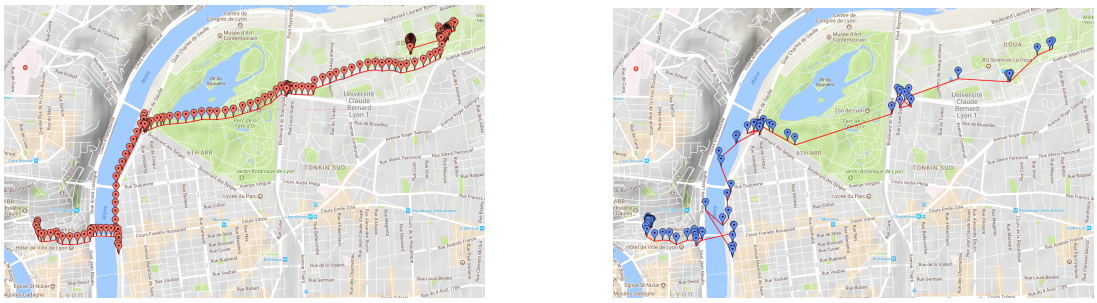
With the democratization of positioning capabilities on mobile devices, location-aware computing is now exploited in most of mobile applications. These applications are thus able to determine the location of users in real time and to provide them geolocated services, often called Location-Based Services (LBSs for short). These services provide a contextual and personalised information depending on the current users location. A multitude of LBSs have emerged these last years from venue finders (e.g., Foursquare [1]) to social games (e.g., Pokemon GO [2]) or crowd-sensing applications [3].

While these LBSs require users to disclose their location to make the application working as expected, some mobile applications also collect the location of users through different sensors without their explicit consent [1, 4, 9]. This intrusive and abusing tracking for behavioral profiling purpose raises important privacy concerns from users.

Depending on the permissions, the mobile operating system will serve to a mobile application a fine or a coarse-grained location information. The user's location can be retrieved by the operating system from the GPS (fine-grained) or by requesting a location data provider (coarse-grained) to covert surrounding Wi-Fi access points (APs for short), nearby cellular antennas, or an IP address into location. While GPS provided a finer-grained location, relying on location data provider becomes increasingly used as GPS signals are not reliably detectable in indoor environment and takes more time for response and causes delay in location determination. Due to the high density of detectable access points (APs for short) in many areas, both outdoor and indoor, Wi-Fi is an excellent communication technology on which to base a location system. Consequently, many online location data providers are available to request in real time the location of users according to the surrounding Wi-Fi access points (e.g., WiGLE [3], Google [4], or Skyhook [5]). To illustrate this positioning capability from Wi-Fi, Figure 1 depicts both the location of a user in Lyon, France, collected from the GPS (Figure 1a) and approximated from the Wi-Fi (Figure 1b) using Google Maps Geolocation API. This illustration clearly demonstrates that the user's location can be precisely revealed to the location data provider from Wi-Fi.

While relying on the GPS to retrieve the location is local and does not reveal the location of user to any third parties, requesting location data provider with surrounding Wi-Fi APs points obviously exposes its location to the location data provider. Location privacy has generated an important literature this last decade [2, 6, 7]. However, most of the proposed works address GPS-based location remaining the protection of Wi-Fi-based positioning not clear. As far as we know, only very few contributions address this issue. For instance, [11, 12] introduces the protection of Wi-Fi-based positioning information when releasing a dataset for a challenge. However, this privacy preserving operation includes many

[1] Foursquare: https://www.foursquare.com
[2] Pokemon GO: http://pokemongo.nianticlabs.com
[3] WiGLE: Wireless Network Mapping, http://wigle.net
[4] Google Maps Geolocation API: https://developers.google.com/maps/
[5] Skyhook: https://www.skyhookwireless.com



(a) The user's locations captured by the GPS

(b) The locations inferred from the Wi-Fi

Figure 1: The location of the users can be precisely inferred from the Wi-Fi.

manual operations which are difficult to fully understand and reproduce. In addition, the evaluation of both the privacy and the utility is not reported and this protection scheme is not adapted in our considered online scenario. Indeed, Location Privacy-Preserving Mechanisms (LPPMs for short) can be classified in two categories, the offline and the online mechanisms. While offline mechanisms aim to protect entire datasets at once before they are published, the online ones aim at protecting on the fly the generated information before being sent to a LBS. Protecting requests sent to location data providers falls into the online category. This online scenario has been considered by Li et al. [13], who introduced a scheme based on homomorphic encryption to protect both the client's location privacy and the service provider's data privacy. However, the use of homomorphic encryption induces a computation and communication cost that makes it impractical for large scale application. Regardless the category, protecting (or sanitizing) location information improves the privacy but have also an inherent harmful impact of the utility of the protected information. For instance, introducing spatial noise obfuscates the real location of user (i.e., privacy gain), however this obfuscation reduces the accuracy of recommendations of places based on the protected data (i.e., utility loss). Privacy and utility metrics are very often dependent on the considered application.

In this paper, we propose a novel online solution to preserve users' privacy from location data providers when requesting the location of users from surrounding Wi-Fi access points, while supporting high-utility. To achieve our goal we combine a *random sampling* for controlling the quantity of revealed information and a *obfuscation scheme* for ensuring privacy-preserving information disclosure. More precisely, our protection mechanism picks at random a limited number of Wi-Fi APs in all surrounding APs to be part of the request. The number of samples included in the request impacts the precision of the location approximation, the more samples, the most accurate. Additionally, we employ an obfuscation scheme to add noise to the request. Specifically, we locally maintain on each user device a graph of collocated APs (i.e., two APs are linked together if they are visible simultaneously in the same neighborhood by the user device). Then for each sample, we include the true MAC address of the AP with a certain probability or the MAC address of a random one at a certain distance in the graph otherwise.

We exhaustively evaluate our obfuscation scheme with a real dataset of mobility traces collected through multiple sensors. We show that our solution can improve the privacy of user by reducing the precision of its location while limiting the utility loss of the protected information. We show that the proposed approach provides a trade-off between privacy and utility that can be controllable by the users. Lastly, we also discuss the integration of our protection mechanism in mobile operating systems.

The remaining of this paper is organised as follow. We first describe the problem statement associated to revealing Wi-Fi information and review related works in Section 2. We then present our protection mechanism in Section 3. Finally, we introduce the experiment setup and the evaluation we consider to assess our protection mechanism in Section 4 and 5, respectively. Lastly, we discuss the integration of this mechanism in mobile operating system and conclude this paper in Section 6.

2 PROBLEM STATEMENT AND BACKGROUND

Most of the mobile phones nowadays embed a Wi-Fi interface. By regularly performing Wi-Fi scans, those Wi-Fi enabled devices maintain an up-to-date list of nearby APs. Consequently, through these network discovery operations, a mobile phone is always aware off the surrounding Wi-Fi APs available.

The collected information about the surrounding Wi-Fi APs can be used to locate the user. Indeed, many Location Data Providers offer online API to convert this information into location. Specifically, these providers collect and maintain a database with the physical location of a large amount of Wi-Fi APs and use position estimator [10] to translate a list of Wi-Fi APs into a location. This location is then used by the mobile system and transmitted to permitted mobile applications to provide a geolocated service.

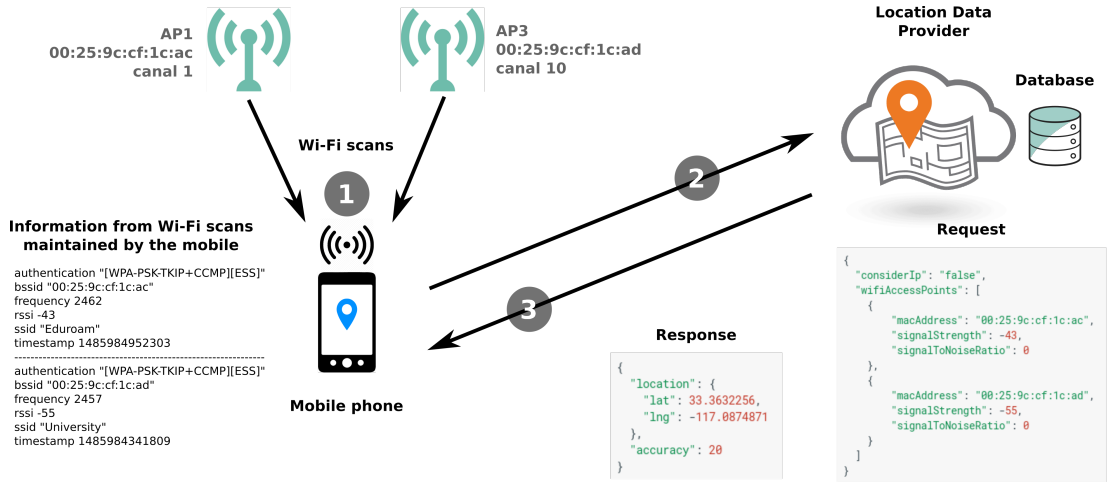


Figure 2: The user's location is exposed to the location data provider via the list of surrounding Wi-Fi APs.

Figure 2 gives an overview of the process. First, the mobile system performs regularly Wi-Fi scans to discover and maintain an up-to-date list of nearby APs (❶). From these Wi-Fi scans the mobile system gathers several pieces of information such as the authentication mode, the MAC address of the AP (BSSID), the operating channel, the Service Set Identifier (SSID), a timestamp and a Received Signal Strength Indication (RSSI), the higher the stronger. When the mobile system decides to update the location, it requests the API of a location data provider (❷). This request contains the MAC address of a list of Wi-Fi APs and may include the RSSI. Obviously, requesting this service is a privacy threat as it reveals to the location data provider information related to the location of the users. Lastly, this service provider responds to the mobile system by providing an estimation of the location from the request (❸) which stores and delivers this information to permitted applications. Note that in order to avoid to maliciously collect the location of one specific AP, those services requires that at least two Wi-Fi APs are provided in the request. Location data providers usually can also provide an estimation of the location from the IP address and other wireless networks such as surrounding Cell Towers or Bluetooth networks. In this paper, we consider the protection of this request only including information from Wi-Fi. We consider an adversary inferring the location of the user from a single API request. We let the problem of an adversary collecting and processing a history of requests and thus inferring more information about the users for future works.

3 PROTECTION MECHANISM

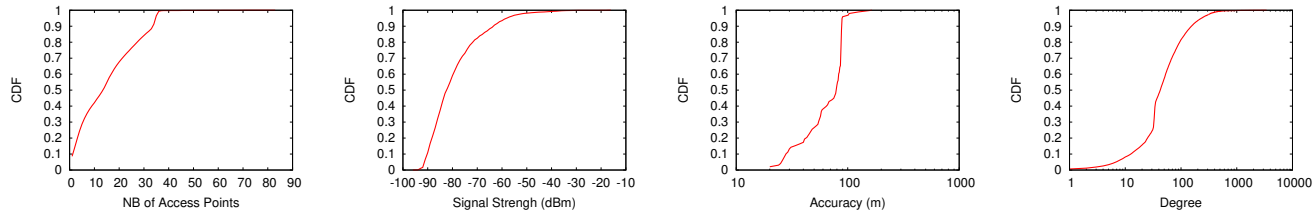
Our mechanism aims at protecting the requests sent to location data providers to get the position of the users from the surrounding Wi-Fi access points. Specifically, to avoid revealing a fine-grained information about the location of users while maintaining a high utility, our protection mechanism combines two techniques: a *random sampling* and *obfuscation scheme*.

3.1 Random Sampling

To limit and to control the precision of the users' location revealed to the service provider, we exploit a random sampling. More precisely, we select a sample of size s from the set of all surrounding Wi-Fi APs available, where each AP in the list has the same chance of being included in the sample. Obviously, the larger s, the most accurate will be the estimation of the location provided by the location data provider. We empirically define three different values for the size s, 2 (the smallest accepted size by the service provider for the list of APs informed in the request), 5, and 10 for a high, medium, and low protection of the location, respectively. Our approach is user-driven, according to the expected privacy level, users define the size of the sample among the three different values. If the number of available Wi-Fi APs is smaller than the expected threshold, we use all available Wi-Fi APs.

3.2 Obfuscation Scheme

To avoid revealing to the service provider only APs actually located nearby the user, we exploit an obfuscation scheme. This obfuscation protects users privacy by allowing individuals to prepare a request without providing truthful list of surrounding Wi-Fi APs all the time, yet it allows location data provider to provide an approximation of the user location. Our obfuscation scheme works as follows: suppose a user willing to send a request to a location data provider to retrieve its location. Once, the sample of APs is defined, for each considered AP in the sample, our protection mechanism flips a coin, if it comes up heads (with probability of p), then the mechanism keeps the truthful MAC address associated to the AP; otherwise with probability of $(1-p)$, the mechanism set another MAC address. In this second case, to define which MAC address will replace the true one, our mechanism



(a) Distribution of the number of surrounding APs **(b) Distribution of the signal strength** **(c) Distribution of the accuracy reported by Google** **(d) Distribution of the degree of the collocated APs graph**

Figure 3: Cumulative Distribution Functions.

exploits a local graph structure maintaining information collected from past Wi-Fi scans. This undirected graph $G(A, E)$ maintains the past encountered APs represented by vertices, $A = a_1, u_2, u_3, ...u_n$ where edges, $E = e_1, e_2, e_3, ...e_n$, connect APs that were part of the same Wi-Fi scan (i.e., actually nearby located). With a probability of $(1 - p)$, our scheme performs a 2-hop random walk from the truthful AP in the graph to select the MAC address that will be used instead of the real one. Using this graph to control the obfuscation of the request, makes the approximation of the location with more uncertainty while maintaining a plausible list of APs in the request. Indeed, if the list of APs present in the requests are unrealistic (i.e., actually not nearby located), the location data provider is not able to provide any location and returns an error message [6].

Therefore, this obfuscation scheme ensures privacy-preserving information disclosure by bounding the amount of information the location data provider gets when receiving a request from a user. Instead of knowing with certainty that the user is located nearby all the Wi-Fi APs informed in the request, the service provider only knows that the user is located nearby each AP of the request with a probability of $(1 - p)$.

Intuitively, the modification of the signal strength could be also leveraged to obfuscate the location of the users. We tried this approach by replacing the signal strengths with random values and by removing altogether the signal strengths from the requests. However we obtained inconclusive results and we thus decided to discard this method. Consequently, our solution only leverages the list of Wi-Fi APs in the request to protect the real users' location and does not use other fields such as signal strength.

4 EXPERIMENT SETUP

In this section, we present the dataset, the methodology and the evaluation metrics we used to conduct our experiments.

4.1 Dataset

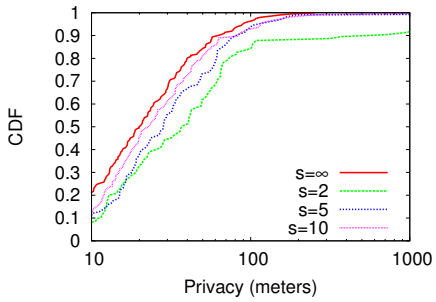
The PRIVAMOV dataset [5] involves 100 students and staff from various campuses in the city of Lyon equipped with smartphones running a data collection software. The data collection took place from October 2014 to January 2016 and gathers information from many sensors such as GPS, Wi-Fi, GSM, accelerometer to name a few. In this paper, we use the records from the GPS periodically collected and the logs from the Wi-Fi scan as presented in Section 2. These two data collection gather 156M and 25M of records, respectively.

To compare both the location of the user inferred from the Wi-Fi and the location measured from the GPS, we first identify the information from the Wi-Fi scan that are combined to a GPS record with less than 2 seconds difference. We identify $29,405$ Wi-Fi scans (i.e., a list of surrounding Wi-Fi APs) associated to a GPS coordinates. Figure 3 presents the distribution (through CDF) of different properties of this dataset such as the number of APs in each Wi-Fi scan (Figure 3a), the signal strength of each Wi-Fi AP (Figure 3b), the accuracy of the location estimated by Google Maps Geolocation API (Figure 3c), and the degree of APs in a graph structure reflecting the nearby APs capturing through all Wi-Fi scans (Figure 3d). The distributions show that on average a Wi-Fi scan contains 15 APs and have a signal strength smaller than -80dbm (the closer to 0, the strengthen). Distribution Figure 3c also shows that on average the accuracy of the location reported by the Google Geolocation service is around 40 meters when we inform the signal strength in the request. Without this information, the accuracy reported by Google is coarse-grained and is 150 meters for all requests. However, as discussed previously, our analysis reports opposite results where requests without signal strength provides on average a slightly better accuracy. Lastly, the distribution of the degree of APs in the collocated graph is around 40.

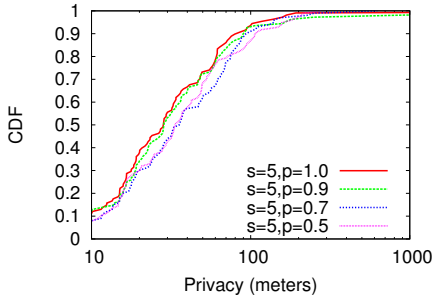
4.2 Utility and privacy metrics

LPPMs improve the user privacy but inherently impact the utility of the resulting protected data [8]. Many utility and privacy metrics have been proposed. In this paper, the considered privacy assessment is performed by measuring the spatial distortion between the real location of the user (i.e., the location collected by the GPS) and the location retrieved from the Google Maps Geolocation API with a protected request. The utility, in turn, is evaluated by quantifying the completeness (i.e., the recall) of the recommendations list provided by Google Places API associated to the real location of the user compared to the recommendations list provided

[6]For instance, using a length of random walk larger than 2 drastically increases the number of queries that do not result to a retrieval of a location from the location data provider.

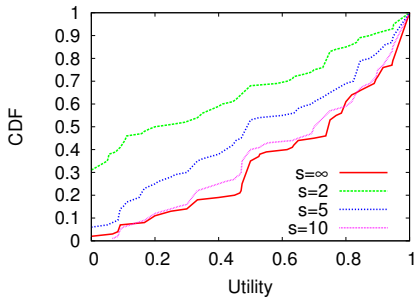


(a) Impact of random sample

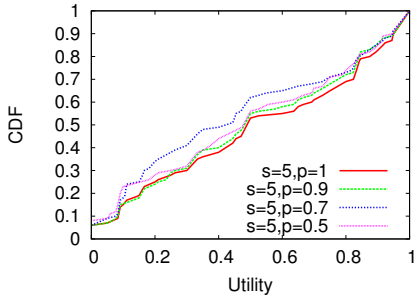


(b) Impact of obfuscation scheme (s=5)

Figure 4: Privacy distribution.



(a) Impact of random sample



(b) Impact of obfuscation scheme (s=5)

Figure 5: Utility distribution.

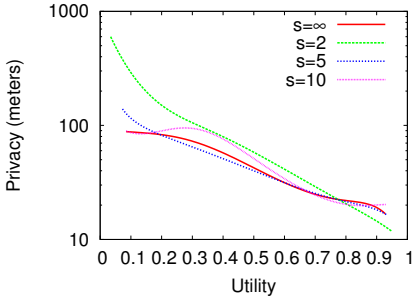


Figure 6: Privacy and utility trade-off (p=1).

by the approximation of the user location (i.e., from the protected request). We consider a radius of 50 meters when we request the recommendations from Google Places API. Both the privacy and the utility metrics are defined as follow:

$$Privacy = \Delta(coord_{gps}, coord_{Wi-Fi}) \quad ; \quad Utility = \frac{|R_{gps} \cap R_{Wi-Fi}|}{|R_{gps}|}$$

where, $\Delta(a, b)$ provides the distance between the coordinate a and b, and R_{gps} and R_{Wi-Fi} the list of recommendations associated to the real location of the user (i.e., the GPS coordinates) and the approximation retrieved from the protected requested sent to Google Places API.

5 EVALUATION

In this section we evaluate the capacity of our mechanism to preserve the privacy of user by protecting the requests sent to location data providers (Section 5.1) while limiting the associated utility loss (Section 5.2). Finally, we analyse the trade-off between privacy and utility (Section 5.3).

5.1 Privacy evaluation

In this section, we evaluate the gain of privacy provided by our protection mechanism. As described in Section 3, our protection scheme modifies on the fly the requests sent to location data providers in order to both include only a sub sample and uncertainly on the surrounding Wi-Fi APs.

We start by evaluating the impact of the random sampling. Figure 4a plots the distribution of the distance between the real location of users and the approximation inferred from Wi-Fi provided by Google Maps Geolocation API for different sizes of sample (parameter s where $s = \infty$ represents no sampling). Firstly, results show that the average accuracy of the location retrieved from Wi-Fi from this service provider is around 30 meters. For comparison purpose, the accuracy of the location retrieved from the GPS is around 3 meters, 10 times for precise. Secondly, results show that a sample with 10 APs provide a similar privacy level, and reducing the size of this sample improves the privacy level (an accuracy of 40 and 50 meters for a size of 5 and 2, respectively).

We then evaluate the impact of the obfuscation scheme. Figure 4b depicts the distribution of the privacy for a size of sample fixed at 5 with varying probability of obfuscation. Results show that increasing the probability of obfuscation (parameter p) improves the privacy (up to 40 meters for a probability of 0.5).

Combining random sample and obfuscation scheme are complementary and improves the privacy. For instance, combining both with a size of sample of 5 provides a privacy level from 30 to 90 meters.

5.2 Utility evaluation

We now evaluated the utility loss introduced by our protection mechanism. Figure 5a starts by evaluating the impact of the

random sampling by depicting the distribution of the utility for varying size of sample (parameter s), namely 2, 5, 10, and without sampling ($s = \infty$). Similarly to the privacy, retrieving the location of users from the Wi-Fi even without sampling inherently reduces the utility compared to a location retrieved from the GPS. For instance, 80% of the answers have more than 0.5 of utility, and 50% of the answers have more than 0.95 of utility. In addition, results show that reducing the size of the sample have an important impact of the utility, the smaller size, the smaller utility. For instance, 50% of the answers have more than 0.7 of utility for a size of sample of 10 while this value drastically drops to 0.05 for a size of 2.

Figure 5b then assesses the impact of the obfuscation scheme and depicts the distribution of the utility for different probability p with a size s=2. results show that the utility is a function of the probability p, the larger probability, the greater utility. For instance, 25% of the requests provide a utility at 0.5 for a probability at 0.7 while only 0.05% of the requests reaches this utility level with a probability p at 0.5.

5.3 Privacy and Utility Trade-off

Finally, we analyse the trade-off between privacy and utility. Figure 6 presents this trade-off for a sample size of 2, 5, and 10, as well as without sampling ($s = \infty$). Results shows that when results provide an important utility, the privacy is however small (privacy of 10 meters for an utility of 0.95). Inversely, when the privacy level is important, the utility is small (e.g., a utility of 0.3 gives a privacy of around 75 meters). These curves illustrate the well-known conflicting trade-off between utility and privacy.

It is interesting to note that each curve have to be correlated with the distribution presented in Section 5.1 and 5.2. Indeed, as shown Figure 4 and Figure 5, the privacy and utility do not follow a uniform distribution over their range of values. Consequently, for instance for r=2, most of the values are represented for an utility inferior to 0.4 However, only watching the trade-off do not provide enough information to appreciate the range of values.

6 DISCUSSIONS AND CONCLUSIONS

We proposed in this paper, a practical mechanism to preserve the privacy of users against location data providers from Wi-Fi. We show that our protection mechanisms can improve the privacy while maintaining a high utility. This mechanism could be included by vendors inside their mobile operating system. In this case, the user would then have the choice of using the original service or to activate our mechanism and choose its level of obfuscation. We are currently implementing and testing this solution to measure the quality of some services and mobile applications in real condition. As future works, we plan to provide differential privacy guaranty by using Randomized Response in our protection scheme.

It is interesting to note that the perturbed data requests might have a negative impact on the database maintained by the location data provider. Indeed, those requests are used by the service provider to keep their database up-to-date and to correct it; thus sending inaccurate requests might damage the quality of the location data.

REFERENCES

[1] Jagdish Prasad Achara, Franck Baudot, Claude Castelluccia, Geoffrey Delcroix, and Vincent Roca. 2013. Mobilitics: Analyzing Privacy Leaks in Smartphones. *ERCIM News* 2013, 93 (2013).

[2] Berker Ağır, Kévin Huguenin, Urs Hengartner, and Jean-Pierre Hubaux. 2016. On the Privacy Implications of Location Semantics. In *PETS*, Vol. 2016. 165–183.

[3] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and Shaping Social Mechanisms in the Real World. *Pervasive Mobile Computing* 7, 6 (Dec. 2011), 643–659.

[4] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. 2015. Your Location Has Been Shared 5,398 Times!: A Field Study on Mobile App Privacy Nudging. In *CHI*. 787–796.

[5] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D 'alu, Vincent Primault, Patrice Raveneau, Herve Rivano, and Razvan Stanica. 2017. PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets. In *NetMob 2017*. Milan, Italy.

[6] V. Bindschaedler and R. Shokri. 2016. Synthesizing Plausible Privacy-Preserving Location Traces. In *S&P*. 546–563.

[7] Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2014. Optimal Geo-Indistinguishable Mechanisms for Location Privacy. In *CCS*. 251–262.

[8] Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Robert Birke, Sara Bouchenak, Lydia Y Chen, Nicolas Marchand, and Bogdan Robu. 2017. PULP: Achieving Privacy and Utility Trade-off in User Mobility Data. In *SRDS*. Hong Kong, Hong Kong SAR China.

[9] Mojtaba Eskandari, Maqsood Ahmad, Anderson Santana de Oliveira, and Bruno Crispo. 2017. Analyzing Remote Server Locations for. Personal Data Transfers in Mobile Apps. In *PETS*. 118–131.

[10] D. Kelly, R. Behan, R. Villing, and S. McLoone. 2009. Computationally tractable location estimation on WiFi enabled mobile phones. In *ISSC*. 1–6.

[11] N. Kiukkonen, Blom J., O. Dousse, Daniel Gatica-Perez, and Laurila J. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. In *ICPS*.

[12] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. 2013. From Big Smartphone Data to Worldwide Research: The Mobile Data Challenge. *Pervasive Mob. Comput.* 9, 6 (Dec. 2013), 752–771.

[13] Hong Li, Limin Sun, Haojin Zhu, Xiang Lu, and Xiuzhen Cheng. 2014. Achieving privacy preservation in WiFi fingerprint-based localization. In *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2337–2345.