Localisation based on Wi-Fi Fingerprints: A Crowdsensing Approach with a Device-to-Device Aim

Patrice Raveneau, Stéphane D 'Alu, Hervé Rivano

Localisation based on Wi-Fi fingerprints: a crowdsensing approach with a device-to-device aim

Patrice Raveneau*, Stéphane D'Alu†, Hervé Rivano†

*Univ La Rochelle – L3i Lab – EA 2118, F-17000, La Rochelle, France
†Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, F-69621, Villeurbanne, France
Contact email: patrice.raveneau@univ-lr.fr

***Abstract*—Crowdsensing is for a few years a new way to gather information. Most smartphones and mobile operating systems provide applications which are able to sense and gather several data from the environment of the device. Thanks to this collected data, it is possible to combine information from several probes. A very common use case is the collection of network scans with location to help the localisation feature of these devices. Nevertheless, most users are not aware of this spying. The collected data might represent infringements of privacy. One possible solution to keep gathering these data while maintaining privacy would consist in device-to-device communications in order to break the links between data and users. In this article we propose an approach to test the feasibility of such a system. We collected data from mobile users to combine location and network scans data. With this data, we test the accuracy level we can reach while using Wi-Fi localisation. We analyse how a new measure should be pushed and how many scans should be realised to provide location-based Wi-Fi. We analyse the minimal dataset to cover the set of locations covered by users and prove that a multi-user gathering system can benefit the users.**

***Index Terms*—Crowdsensing, Wi-Fi fingerprint-based localisation, measurement data**

1. Introduction

Device-to-device communications is an aim of future communications for several applications including cellular offloading [1]. This type of communications can also be envisaged as a technique to preserve privacy while collecting data. One of the main drawbacks of most location-based services is the revelation of points of interests (POI) of the users [2]. Indeed, since GPS is energy-consuming and provides a localisation after a few tens of seconds with a cold start, several techniques of localisations based on association of locations with ambient fingerprints have been proposed [3]. The transmission of these data might reveal POI of users.

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

Most mechanisms aiming at preserving privacy rely on datasets obfuscation to preserve users' privacy or on homomorphic encryption, while transmitting requests on localisation datasets to maintain users and datasets provider's privacy [4]. Concerning the homomorphic solution, it is proposed in [4] for localisation-based services using Wi-Fi fingerprints. When a user transmits queries containing ambient Wi-Fi fingerprints, the dataset provider answers with an estimated location, then user's location is revealed. That is why mechanisms based on homomorphic encryption could handle this drawback. These two techniques which shall provide privacy to users do not protect the feeders of the datasets. Indeed, these datasets require that people carrying devices embedding Wi-Fi and localisation capabilities transmit this data. The points of interests of these people are revealed when they transmit it. This could be tackled by device-to-device communications.

Device-to-device communications could be used to mix data from several users so that users devices transmit small parts of their collected dataset to other users they meet. Such a scheme should determine which parts of the dataset should be transmitted to other users so that no point of interests is revealed to the other user. We can even imagine that users define areas where they estimate points of interests exist and they do not want the service enabled. Nevertheless, such a system shall handle long links disruptions since the encounters between the users are opportunistic. The system should use protocols following the Delay/Disruption Tolerant Network paradigm [5] to be able to discover neighbours. Then, data transmitted to the networks by devices would be obfuscated by data coming from other users. However, if the devices transmitted their whole dataset to every device they meet, they would transmit huge data volumes. It is then compulsory to transmit a minimal useful dataset. The devices should determine if a Wi-Fi fingerprint should be coupled to a location or if it is not required to store it.

Our aim is to test if a user-centric Wi-Fi localisation scheme is viable and accurate. We consider that it is important for an user to keep its privacy, then the best way for that is to transmit the least data. Devices would keep a connection between locations and Wi-Fi fingerprints. We focus on Wi-Fi fingerprints localisation scheme even if other methods provide more accurate results [6], they are

not suitable for deployment on smartphones, but on Wi-Fi Access Points (AP) which is out of the scope of this paper. We will need metrics to compare two fingerprints. We will use a common method to compute the euclidean distance between two fingerprints [7]. Nevertheless, most studies using the euclidean distance between two fingerprints consider fingerprints as vectors of same size. This assumption might be true in an indoor environment but not for outdoor contexts. We will adapt the computation of this distance to the intersection of Basic Service Set Identifiers (BSSID) of two fingerprints.

We developed an Android application which, once installed on a mobile terminal, was collecting data from several probes including Wi-Fi and localisation. Data was collected raw and we had to pre-process it to be able to test the feasibility of an user-centric Wi-Fi localisation service.

These deployments are described in Sec. 2 of the paper. In addition, we present the methodology to determine the strategy of nodes involved in a collaborative Wi-Fi fingerprint based localisation service and analyse the performance in terms of geographic error and storage space in Sec. 3 of the paper. Finally, Sec. 4 summarizes the conclusions of our work.

2. Dataset Presentation

We deployed our application on 100 devices of volunteers. We have gathered data from 94 users. The six other users might have not started their device. The data collected was stored in a database and was identified per user and per probe. For this study, we focus on Wi-Fi and localisation probes. From this raw data we had to apply some processing to be able to test the feasibility of an user-centric Wi-Fi-based localisation service. The localisation probe stores and transmits any location found by the system. This might be a problem because when the device is turned off, the GPS provides the last known location which might be far away from the current position. The Wi-Fi probe provides the results of a Wi-Fi scan pushed by the device. A Wi-Fi record is not a fingerprint but the result of a Wi-Fi beacon or Wi-Fi probe response.

Since each record has a timestamp, we computed for each user the time difference between every successive Wi-Fi record. We have analysed that most differences were less than one second. We have then decided to use a threshold of one second to determine whether a record were belonging to a Wi-Fi fingerprint or to the next one. After that we needed to link each Wi-Fi fingerprint with its location. For each user we have computed the time difference of each fingerprint with a location record of the same user. From this point we assumed that an user would not move a lot in a few seconds between the real location of the fingerprint and the location record. We have compared the results with a threshold of one second and a threshold of ten seconds. The latter provided a little bit more similarities. We considered that it was better to keep the generated datasets with the one second threshold because if the distance covered by an user walking would be small, it becomes great if the user is in a car.

From now on, we will use this reconstructed dataset as a reference for our tests. After this filtering process, only 44 datasets contain records of Wi-Fi fingerprints and linked locations.

3. Wi-Fi localisation from data collection

In this section, we explain how we evaluate the capability of a Wi-Fi based localisation service whose data would be collected by users.

As we explained earlier, our aim is to analyse if an autonomous or distributed localisation service running on users devices could work. Such a system already exist and is named Global Positioning System (GPS). However, a continuous use of GPS would drain the battery of the devices [8]. This is why we focus on a system providing localisations thanks to Wi-Fi fingerprints. Nonetheless, Wi-Fi alone does not provide any location information. Then, it is compulsory that devices have a strategy to determine whether they shall store a Wi-Fi fingerprint and an associated location. Indeed, to be efficient, this service shall keep the minimal set of fingerprints with linked locations. Then, when a mobile device has its Wi-Fi interface activated, it will receive messages from the ambient BSSIDs. The set of these BSSIDs is a Wi-Fi fingerprint. If in the storage of the device, there is no fingerprint close enough, then the device shall turn on GPS and links location to this fingerprint. We need to find a threshold based on the difference of fingerprints so that the device knows if it shall add this fingerprint with a location information or not.

3.1. Finding a Wi-Fi threshold

We consider as metrics the euclidean distance of Wi-Fi fingerprint. To compute this distance we represent fingerprint as a vector whose components are the Received Signal Strength Indication (RSSI). As we presented earlier, since the users are not in the same environment we will not have fixed length fingerprints as in [7]. Then when we compute the distance between two fingerprints, they shall share at least one BSSID. We calculate the euclidean distance by only considering the RSSI of BSSIDs existing in the two fingerprints.

We have then two metrics to consider, the number of shared BSSIDs and the RSSI distance. We are going to simulate the behaviour of a device which would embed the autonomous version of the Wi-Fi localisation service. For each user, we travel along the dataset from the first fingerprint to the last one. We have sorted the fingerprints per timestamp. We initialise a reference file with the first fingerprint then for each fingerprint we compute the RSSI distance, the number of shared BSSIDs and the geographic distance between the locations of these two fingerprints. After that, we are able to represent the values of geographic distances depending on the RSSI distance and the number of shared BSSIDs on Fig. 1.

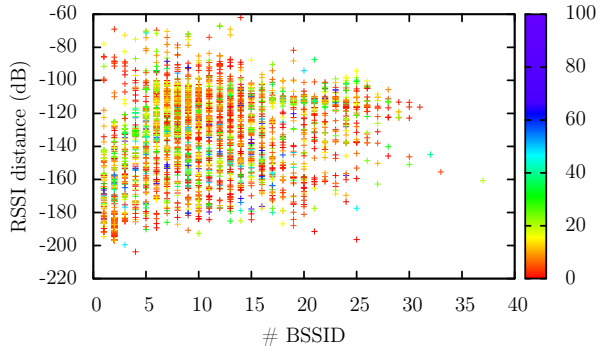

Figure 1: Geographic distance error depending on RSSI distance and number of shared BSSIDs

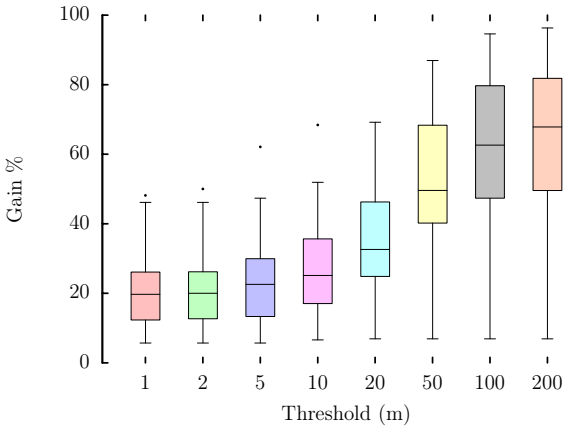

Figure 2: Gain as percentage of the dataset depending on the geographic distance threshold

We observe on Fig. 1 that most geographic distances are represented with red cross and are very close to zero meter, which is very accurate. Nonetheless, we expected that we would have seen a split or a border based on number of BSSIDs or RSSI distance. We thought that the greater the number of BSSIDs and the lower the RSSI distance, the lower the geographic distance. But here we can notice that we have geographic distances close to one hundred meters while we have 14 BSSIDs and a RSSI distance between -160 and -180 dB. And on the other side, we have several points with close to zero meter distance which have less than five BSSIDs and between -60 and -100 dB.

From that point we can only say that we should add a new record when we have no BSSIDs in common or when the RSSI distance is greater than -60 dB.

For the remainder of this study, we will keep as a threshold to have more than one BSSID in common. We will have from time to time errors greater than one hundred meters but we will be able to reconstruct a bigger portion of the set of locations.

3.2. Evaluation of the gain on dataset size

Before using the threshold that we defined on RSSI distance, we want to evaluate the gain on the dataset size based on the accuracy of the geographic distance. Our user-centric localisation system shall keep the size of the dataset as small as possible while covering the biggest set of locations where the user went.

We start by evaluating on Fig. 2 the gain of each user on the dataset size depending on a geographic distance threshold. When we compare a new fingerprint to the set of stored fingerprints, if the smallest geographic distance between the new fingerprint and any fingerprints from the dataset is greater than a pre-defined threshold, we add this new fingerprint to the dataset. We compute the size of datasets before and after thresholding. On Fig. 2, we observe that when we increase the threshold, we increase the gain on the dataset size. Since the accuracy level is less strict,

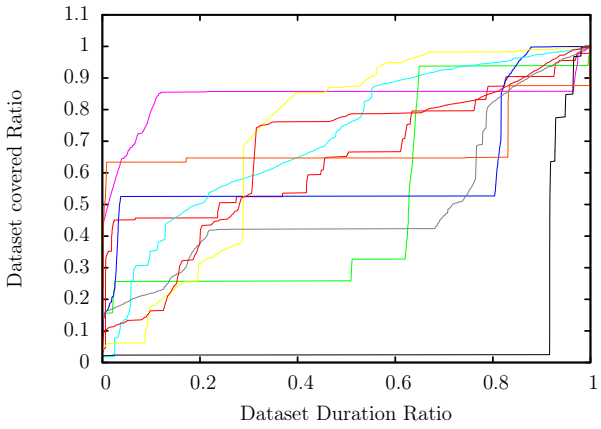

Figure 3: Covered dataset versus time using 1 meter threshold

we can tolerate bigger errors and then we need less records to cover the set of locations of an user. We notice that even with one meter accuracy, we get half users having 20% gain on their dataset size. We can also see that the lower limit is always at the same gain, around 5%. This can be explained by the fact that we have users with very small datasets; then their dataset might have very few records which are close enough to be replaced by another one.

Before moving to the analysis of the system with a Wi-Fi based threshold, we evaluate if with the best accuracy that we defined we are able to reconstruct the original dataset locations or if some were missing. We also analyse within how much time which percentage of the dataset locations is covered. We analyse this behaviour per user. Since the durations of datasets are not the same for the users, we use a ratio of the dataset duration of each user. We present the results on Fig. 3 for the ten users with biggest dataset size. First of all, we see that for each user we reach the 100% dataset covered. The second important information is that we observe big steps for each user. This indicates that when

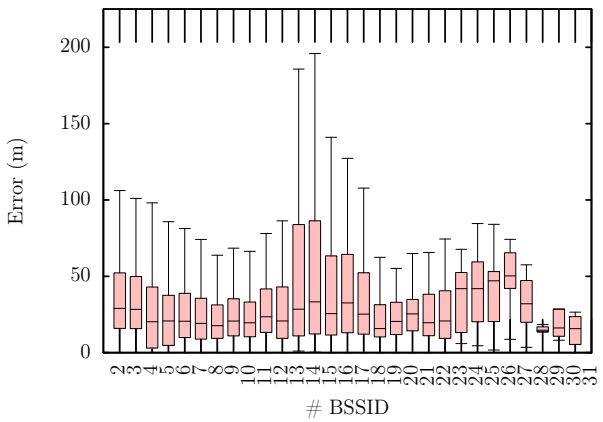

Figure 4: Distribution of errors per number of shared BSSIDs

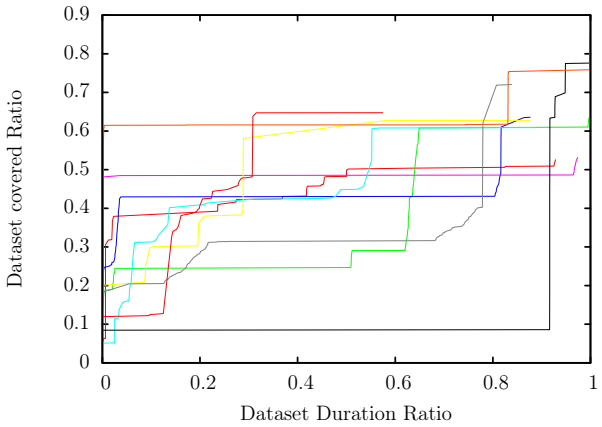

Figure 5: Covered dataset versus time using Wi-Fi threshold

an user gets to a new place, the system stores new records and can reuse it till the moment when the user moves to a new place.

3.3. Evaluation of Wi-Fi based localisation technique

We are now going to evaluate the performance of the localisation when relying on Wi-Fi threshold we defined earlier, more than one BSSID in common within two fingerprints. We represent on Fig. 4 the distribution of errors depending on the number of BSSIDs in common. We would expect that errors would decrease as the number of BSSIDs shared between two fingerprints increase. We see that the lower limit is always very close to zero meter error, then even with very few BSSIDs in common, we can reach the best accuracy for some records. We also notice that the median, quartiles and upper limit are decreasing when the number of BSSIDs increase from two to ten. Nevertheless, we observe a huge increase on third quartile and upper limit when the number of BSSIDs is in the range of 12 to 17. Then these values decrease again while the number of BSSIDs increase. We assume that a possible explanation might come from the imprecision of the localisation probe in some areas and more specifically in indoor environments. Indeed, when we are inside a building, a device receives generally Wi-Fi signals coming from more APs in particular in an office environment. Well, when a device is indoor, GPS signals are harder to receive and then the accuracy of GPS decreases. We think that this particular range might be linked to offices buildings and this is why we get poor 3^{rd} quartiles and upper limits in this range. Nonetheless, we notice that the median and the 3^{rd} are respectively, most of the time, less than 25 and 50 meters, which is a good accuracy error. In order to improve these results, we need to get a reliable localisation reference.

We now check the effect of using the defined Wi-Fi threshold on the covered dataset ratio in Fig. 5. In this figure,

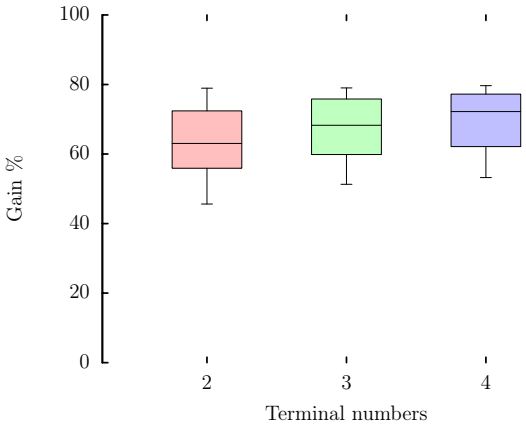

Figure 6: Gain on datasets sizes while using multiple users

we use the same users whose results were presented on Fig. 3. We observe that the shapes present the same trend as earlier. This indicates that we are able to cover almost the same percentage of the dataset with the same records within almost the same duration. However, we also notice that we do not reach 100% of covered dataset. This can be explained by the fact that, even when we have several BSSIDs in common, we can get errors above 100 meters. Then, it is possible that in the original dataset we have a fingerprint at a location which is close enough to another fingerprint from the learning dataset at another location. Then the location covered by the fingerprint from the original dataset is lost. To improve this result, we would have to define a more discriminant threshold providing more accurate results on geographic distance.

All the presented results till now deal with the autonomous version of the localisation service. We are now evaluating the impact of using multiple users. We present on Fig. 6 the distributions of the best gain per user when using combined datasets from two, three and four users. We calculate the ratio of the size of the combined dataset by the sum of the datasets of the implied users. We observe that

in the worst case, we have a 45% gain on the dataset size and that half the datasets are more than 60% reduced. This proves that there is an interest to use a collaborative Wi-Fi localisation service.

4. Conclusion

In this paper we studied the feasibility and the performance of an user-centric Wi-Fi based localisation system. We defined from the analysis of traces of several tens of users a threshold on Wi-Fi so that a device knows if it shall add a record to its database or if a record in the dataset already handles this location. We have also analysed that a collaborative version of this localisation service, using several users to create the dataset, would benefit all users of the system.

As a future work, we plan to develop an application embedding the opportunistic communications capabilities, to test if the collaborative version is efficient and can exist in a fully distributed way without any server. We also plan to analyse if the building of the dataset converges faster in the collaborative version by exploiting the encounters with other users of the system. We will also analyse to which extent the privacy would be protected by such a system with and without storing server.

References

[1] F. Rebecchi, L. Valerio, R. Bruno, V. Conan, M. D. de Amorim, and A. Passarella, "A joint multicast/d2d learning-based approach to lte traffic offloading," *Computer Communications*, vol. 72, pp. 26 – 37, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366415003679

[2] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie, "Adaptive Location Privacy with ALP," in *35th Symposium on Reliable Distributed Systems*, ser. Proceedings of the 35th Symposium on Reliable Distributed Systems, Budapest, Hungary, Sep. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01370447

[3] M. Azizyan, I. Constandache, and R. Roy Choudhury, "SurroundSense: Mobile Phone Localization via Ambience Fingerprinting," in *Proceedings of the 15th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '09. New York, NY, USA: ACM, 2009, pp. 261–272. [Online]. Available: http://doi.acm.org/10.1145/1614320.1614350

[4] H. Li, L. Sun, H. Zhu, X. Lu, and X. Cheng, "Achieving privacy preservation in WiFi fingerprint-based localization," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, Apr. 2014, pp. 2337–2345.

[5] K. Scott and S. Burleigh, "Bundle Protocol Specification," RFC 5050 (Experimental), Internet Engineering Task Force, Nov. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc5050.txt

[6] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter Level Localization Using WiFi," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, ser. SIGCOMM '15. New York, NY, USA: ACM, 2015, pp. 269–282. [Online]. Available: http://doi.acm.org/10.1145/2785956.2787487

[7] Z. Yang, C. Wu, and Y. Liu, "Locating in Fingerprint Space: Wireless Indoor Localization with Little Human Intervention," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 269–280. [Online]. Available: http://doi.acm.org/10.1145/2348543.2348578

[8] J. G. Krieg, G. Jakllari, H. Toma, and A. L. Beylot, "Unlocking the smartphone's senses for smart city parking," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.