A Privacy-preserving Scheme for Passive Monitoring of People’s Flows through WiFi Beacons

A Privacy-preserving Scheme for Passive Monitoring of People's Flows through WiFi Beacons

Kalkidan Gebru
Electronics and Telecommunications Department, Politecnico di Torino, Italy
E-mail: kalkidan.gebru@polito.it

***Abstract*—The proliferation of IoT-based services for smart cities, and especially those related to mobility, are ever becoming more relevant and gaining attention from a number of stakeholders. In our work, we tackle the problem of characterizing people movements in a urban environment by using WiFi sensors connected to the cellular network. In particular, we leverage WiFi probe requests transmitted by people's smartphones and a machine learning approach to detect people's flows, while preserving users' privacy. We validate our approach through a proof-of-concept testbed deployed in the proximity of our campus area. We consider two types of devices, namely, commercial, off-the-shelf WiFi scanners and ad-hoc designed scanners implemented with Raspberry PIs. They provide different levels of visibility of the captured traffic, preserving in different ways the privacy of the people's movements. In our current work, we investigate the different trade-offs between mobility tracking accuracy and the level of provided people's privacy.**

***Index terms*—People's flow detection, privacy-preserving data collection, proof-of-concept**

I. Introduction

Analysing people's movements in urban environments is central to several critical applications related to safety, as well as to a plethora of convenience services designed for mobile users (e.g., car sharing, use of public transports, and store recommendation systems). In particular, for many applications it is essential to detect the pattern taken by people's flows at different times of the day/week. One of the key technologies to achieve this goal is the Internet-of-Things (IoT) [1], [2], as IoT devices are becoming pervasive and most of them are equipped with a radio interface, such as WiFi, that can conveniently connect them with other devices as well as with the communication network infrastructure. Furthermore, they typically consume little energy, hence they contribute to creating sustainable communication systems, have low cost, and pose fewer privacy issues than other devices like smart city cameras.

In this work, we leverage the IoT technology and tackle the problem of characterizing both people's trajectories and the number of people following a given path, while *preserving users' privacy*. In particular, we focus on a urban environment and exploit both commercial sensors and such simple devices as Raspberry PIs, equipped with a WiFi interface. Such devices can scan the WiFi spectrum for probe requests, i.e., packets transmitted by user hand-handled devices towards nearby access points. Using the logs provided by these spectrum scanners, we develop techniques to ensure that the collection and processing of the data meets the General Data Protection Regulation (GDPR) [3]. Importantly, we aim at developing a solution that can cope with the serious limitations of commercial or Raspberry PI-based scanners, which demands for a new approach with respect to those proposed in prior art.

In particular, most of the existing schemes leveraging WiFi sensors only detect probes sent by users' smartphones, or infer the number of users but not the path they follow, or present experimental results rather than a full-fledged methodology for people's flow inference. In addition, they are typically quite complex, as it is often necessary to cope with implementation specifics in the periodicity of the transmissions of the probe requests sent by user devices.

Unlike existing work, we develop an approach that can effectively (i) cope with commercial sensors as well as simple ad-hoc designed devices that scan the spectrum for WiFi probes, and (ii) fully meet GDPR specifications. Further, applying an ML-based scheme, we show how the data collected through our privacy-preserving solution can be used to characterize people's flows. Our approach is then validated through a proof-of-concept testbed that we developed and a measurement campaign that we performed in the proximity of the campus area of the Politecnico di Torino, in the city of Turin, Italy.

The remainder of the paper is organized as follows. We introduce the scenario under study, which is also used for our proof-of-concept testbed, in Section II. The privacy-preserving solution that we develop to collect anonymized data meeting privacy constraints is described in Section III, while the ML approach for people's path detection is briefly described in Section IV. Finally, we discuss some related work in Section V, and draw some conclusions and our planned directions for future research in Sec. VI.

II. System and Proof-of-concept Scenario

As depicted in Figure 1, we consider an urban area where, e.g., pedestrians and cyclists transit from a location to another,

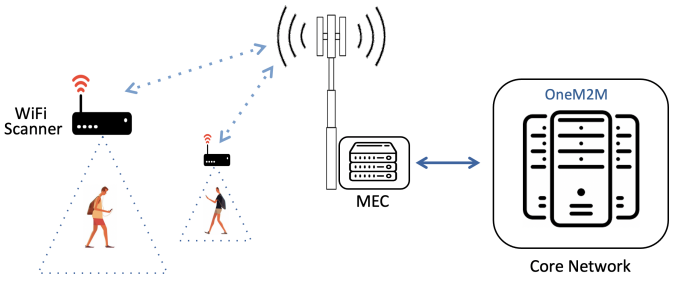

Fig. 1. System scenario and proof-of-concept setting.

such as from a railway/subway station to a university campus. WiFi probe-detection scanners are deployed along streets, with each scanner being connected to the cellular network. Scanners receive WiFi probe request packets that are transmitted by user devices nearby, with the purpose of swiftly discovering access points with which they had been previously connected. Scanners collect data carried by such probes, specifically, the sender's hashed MAC address and the probe timestamp.

Such a scenario poses a number of challenges. First, one needs to cope with randomized MAC addresses in probe requests, which makes user count hard since MAC addresses across probes generated by the same user in a flow may be different. Notice that most of the existing solutions addressing this issue cannot be applied here, as they would need to process the entire probe MAC header, which is not available in commercial WiFi scanners. Second, the timing of probe request generation may vary across different devices [4], and, in particular in commercial scanners, they are not frequently transmitted, which impairs the use of many of the existing solutions. Third, scanners may have not been placed so as to optimize geographical coverage, which may thus result into missing or fleeting connectivity, or on coverage overlaps. The latter, in particular, combined with the coarse time granularity with which probe requests are reported, makes multiple scanners detect the same user at the same time.

With regard to our proof-of-concept scenario, this matches the one described above and exemplified in Figure 1. It has been deployed nearby the Politecnico di Torino campus area, in the city of Turin, Italy, with the help of the Turin municipality. It includes Libellium Meshliums [5] and ad-hoc designed sensors using WIP Raspberry PI, both scanning the WiFi spectrum at 2.4 and 5 GHz. All the scanners are synchronized via NTP. Libellium scanners log the collected probes just once every 50 seconds, while in the case of the ad-hoc designed sensors the log periodicity had to be set to at least 5 seconds, in order to let the capturing and processing software properly run. Further, scanners are connected to the 5G-EVE platform [6] via the cellular network, and, through such a platform, they upload their log to an OneM2M server [7] (see Figure 1) every two minutes.

III. Data Collection

As first, fundamental step in our methodology, we develop a privacy-preserving technique to extract data from users' probes. Upon detecting a probe request, scanners log several pieces of information related to that (see Figure 2). Specifically, (i) the received signal strength index (RSSI), which however may lead to inaccurate estimations in outdoor scenarios, (ii) the interface vendor, which unfortunately does not uniquely identify the interface and it is often set to "unknown", (iii) the timestamp related to the probe request detection, and (iv) the sender's MAC address. Given the above issues, RSSI and interface vendor cannot be exploited.


```
{"data":[
    {"RSSI":"-68","Vendor":"Samsung",
        "TimeStamp":"2020-02-15 12:32:45","MAC":"B7...BA"},
    {"RSSI":"-64","Vendor":"Apple",
        "TimeStamp":"2020-02-15 12:32:45","MAC":"9E...01"},
    {"RSSI":"-86","Vendor":"Unknown",
        "TimeStamp":"2020-02-15 12:32:45","MAC":"3F...FA"}
]}
```


Fig. 2. Example of the log of a probe request detection provided by a WiFi sensor.

With regard to the timestamp, due to the coarse periodicity with which such information is logged, it is impossible to infer an accurate probe time sequence, thus we cannot exploit the probe arrival times to infer the heading of the movement. Additionally, several probe requests generated by the same user device during a 1-second sampling interval may be collapsed by scanners into a unique record.

As for the MAC address, this is considered personal data by GDPR [8]. Thus, in the following, we describe the privacy-preserving techniques that we designed and implemented, in order to leverage the MAC address information while meeting privacy constraints, in both the cases where commercial WiFi sensors and ad-hoc designed scanners are used.

A. Dealing with anonymized MAC addresses in commercial WiFi sensors

Since the sender's MAC address is considered personal data [8], in commercial scanners it is anonymized by digesting the device MAC address with an `SHA-224` function.

Given such collected information, we tried to identify the most common paths followed by the users in the considered area, and the effect of MAC randomization on the accuracy of the performed analysis. To do so, we considered a trace taken in our proof-of-concept testbed area, during one week in October 2019. This trace includes over 190,000 distinct MAC addresses. Due to MAC randomization, hence the fact that a user device may send multiple probes with different MAC addresses, such a number is just an upper bound on the number of users traversing the considered area. To evaluate the impact of MAC randomization, we leveraged our knowledge of over 34,000 MAC addresses of WiFi devices owned by students, professors and administrative employees of Politecnico di Torino. Notice that such MAC addresses are not randomized, as they are collected after a device had associated with an AP on campus; however, for the sake of privacy, they were anonymized with the same hash function used by the scanners. In so doing, we could identify the subset of MAC addresses of Politecnico di Torino users that appeared in the trace.

Next, to identify the different paths, we looked at the temporal sequence of probe captures: $[(t_i, s_i)]_i$, $i = 0, 1, \ldots$, where t_i denotes the probe detection time and s_i the scanner that performed it. We then created sub-sequences thereof, by grouping consecutive samples within a time span shorter than or equal to 4 minutes. Each sub-sequence models a different mobility pattern and has been translated into a representative string reporting the sequence of scanner identifiers. For instance, a string "s_1" means that the user was detected only by scanner s_1 within a time span of 4 minutes, while "$s_1\, s_2\, s_1$" corresponds to a user that in the same time span has been detected by s_1, s_2, and then s_1 again. All cases where a user was simultaneously under the coverage of (hence it has been detected by) multiple scanners have been denoted with a special symbol.

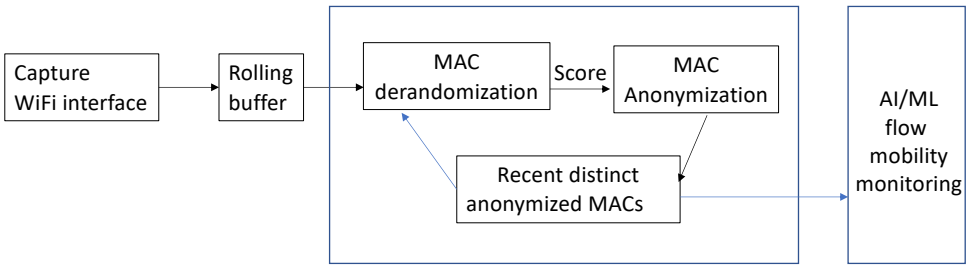

Fig. 3. Process pipeline implemented in the ad-hoc designed WiFi sensors using Raspberry PI devices.

B. A Privacy-preserving technique in ad-hoc designed WiFi sensors

We now consider ad-hoc designed WiFi scanners implemented using Raspberry PI devices. In this case, it is possible to get detailed information about the captured WiFi probe request packets, but clearly this violates user's privacy since the MAC address is considered to be personal information. At the same time, a MAC address cannot be anonymized as soon as the probe request packet is received, otherwise it will be impossible to understand if the address belongs to the range of randomized addresses.

Thus, we apply a de-randomization scheme to the received probe request, just for the sake of local processing. We consider the de-randomization scheme proposed in [9], which was designed to count people on public transportation systems. Such scheme exploits the temporal correlation of the data included in the WiFi probe request headers, e.g., increasing sequence numbers, to identify MAC addresses that would likely correspond to the same device.

Figure 3 describes the processing pipeline which we have implemented on Raspberry PI. The traffic is captured on the WiFi interface through `tshark` and stored in a temporary buffer through the "rolling capture" feature. Batches of traces are removed from the rolling buffer after being completely processed. The original MAC address is thus stored just for few seconds on the local storage of the Raspberry PI. The original WiFi header is then processed by the de-randomization algorithm, which computes a score of the MAC address to understand if it can be associated with a recently seen address. The score is proportional to the probability that two or more randomized MAC addresses correspond to the same device, based on the approach proposed in [9]. The higher the score, the more likely that the two MAC addresses relate to the same device. Specifically, we adopt a threshold-based scheme: whenever the score is above a fixed threshold, the MAC corresponds to a recently seen MAC and it must be associated to a same device.

After being processed by the de-randomization module, the MAC is anonymized through the `SHA-224` hash function and, if the score is above threshold, it is stored locally on a temporary list of MACs, with the remaining part of the original header. This list reports the set of MACs (now anonymized) that have been considered estimated as "distinct" based on the score and each MAC is associated univocally with a device. Notably, storing the original WiFi header with the anonymized MAC address allows us to compute the score for the address of new incoming probe request packets, and, at the same time, keep the original one confidential.

Finally, the list of recent anonymized MAC addresses is used to compute the total number of distinct MAC addresses, as an estimation of the number of devices currently under coverage, and it is fed to the ML scheme described next, for the detection of the flow's path.

IV. ML-BASED PATH DETECTION

Besides detecting the number of users belonging to a flow, we leverage the collected anonymized data to identify the flow trajectory, i.e., the path they follow. To do so, we draw on our previous work in [10] and apply an ML-based technique.

We start by classifying paths based on some preliminary experiments, thus obtaining a catalogue of possible paths that will represent our ground-truth. For each path, the catalogue also includes a set of possible footprints, i.e., a sequence of probe requests sent by a user following that path. Then, we compare the sequence of probe requests transmitted by a user (and captured by the scanners) against the paths' fingerprints, and match it with the path for which the vectors' Euclidean distance is minimized. In case of ties, we select the path with the maximum number of minimum distance footprints. If no unique path can yet be identified, then the sequence is labelled as not traceable.

More in details, given a physical path, we compute a possible corresponding footprint as follows. We assign to every scanner covering that path a numerical weight. Then we let a user follow the path and we collect the sequence of probe requests captured by the scanners. We divide the time taken to go through the entire path and divide it into n intervals. For each interval, we look at the probes detected by the nearby scanner(s) and calculate the average weight across the scanners that reported a user probe in that interval. This allows us to characterize the geographical trajectory of the user. Finally, to better account for the user's movement direction, we compute the slope of the best fitting linear interpolating function of the probe samples over the path. Experimental results validating this approach can be found in [10].

V. RELATED WORK

The problem of monitoring user mobility using WiFi probes has been widely addressed in the literature. Some of the existing studies focus on specific scenarios, with, e.g., [11] targeting a university campus, and [12] a railway station. Interestingly, the methodology proposed in [12] exploits the knowledge of the number of people with an active WiFi

interface. Pedestrian users in a more general urban scenario are considered in [13], where a methodology leveraging the time difference between probes and that between probe sequence numbers is proposed. On the contrary, GPS is used in [14] to characterize locations along the paths followed by users as well as their mobility. We remark that our goal is to use WiFi commercial, or simple ad-hoc designed, sensors instead of GPS, and that such sensors are unable to provide the type of information used in [13]. Further, it is worth noting that WiFi probes allow for the measurement of the received signal strength indicator, which can be exploited for detecting the user distance from the WiFi sensor [15]. However, this approach has been proved to work well indoors, while it may exhibit low accuracy in outdoor scenarios [16].

Relevant to our work is also the experimental study in [2], which presents a system design as well as extensive experiments, and shows heat maps describing people's density. We stress that, unlike [2], our focus is on a privacy-preserving mechanism to collect data so as to implement an ML-based approach for people's flows detection.

Finally, we mention that a first step towards the development of our proof-of-concept testbed can be found in [10], where we also detailed the ML-approach for path detection that complements the privacy-preserving methodology presented in the present paper.

VI. Conclusions and Future Work

We tackled privacy-preserving people's flows detection in urban environments, using commercial/off-the-shelf and ad-hoc designed WiFi scanners. We designed a methodology to guarantee user privacy, while collecting and processing the information carried by probe requests generated by WiFi user devices. In particular, we designed techniques specifically tailored for the considered WiFi scanners, namely, commercial/off-the-shelf sensors and ad-hoc designed ones implemented in Raspberry Pi. Such collected data allowed us to count the number of people belonging to a flow, and to apply a machine learning-based approach for identifying the flow trajectories. We implemented our solution through the proof-of-concept testbed we developed, thus validating the approach and its ability to cope with the serious challenges posed by practical scenarios.

A. Discussion on future work

The detection of flow mobility patterns, as described in Sec. IV, requires to correlate spatially the MAC addresses observed at different WiFi scanners. While this is already implemented and fully working with Libellium Meshliums sensors, in the case of our solution based on Raspberry PI, the MAC addresses are stored locally as anonymized. The actual stored identifier depends on the de-randomization process. Thus, it may happen that the same MAC address can be stored with different anonymized MAC addresses at different scanners. This would prevent to correlate spatially the detection of the same user device at different scanners, limiting the accuracy of the proposed approach.

To address this issue, we are devising a centralized solution in which the table with the recent distinct anonymized MAC addresses is shared across all the scanners. In this way, all the scanners can access this table and de-randomize the MAC addresses coherently with the others. Thanks to a consistent identifier across all the scanners, the MAC addresses would be associated to the same device, enabling the analysis of the spatial correlation across multiple sensors. Notice that this approach requires to send anonymized data from a central server to the WiFi scanners, and, hence, it does not violate users' privacy.

VII. Acknowledgement

The author wishes to thank Prof. Claudio Casetti, Prof. Carla Fabiana Chiasserini, and Prof. Paolo Giaccone, for their help and useful advice.

References

[1] C. Badii, P. Bellini, A. Difino, and P. Nesi, "Sii-mobility: An IoT/IoE architecture to enhance smart city mobility and transportation services," *Sensors*, vol. 19, no. 1, 2019.

[2] M. Uras, R. Cossu, and L. Atzori, "PmA: a solution for people mobility monitoring and analysis based on WiFi probes," in *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2019, pp. 1–6.

[3] "GDPR," https://www.gdpr.net/.

[4] Y. Durmus and K. Langendoen, "WiFi authentication through social networks: A decentralized and context-aware approach," in *IEEE PERCOM*, 2014, pp. 532–538.

[5] Libelium Meshlium. [Online]. Available: http://www.libelium.com/products/meshlium/

[6] European 5G validation platform for extensive trials. [Online]. Available: https://www.5g-eve.eu/

[7] OneM2M. [Online]. Available: http://www.onem2m.org

[8] EU General Data Protection Regulation. [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

[9] M. Nitti, F. Pinna, L. Pintor, V. Pilloni, and B. Barabino, "iABACUS: A wi-fi-based automatic bus passenger counting system," *Energies*, vol. 13, no. 6, p. 1446, 2020.

[10] K. Gebru, C. Casetti, C. F. Chiasserini, and P. Giaccone, "IoT-based mobility tracking for smart city applications," in *2020 European Conference on Networks and Communications (EuCNC)*, 2020, pp. 326–330.

[11] E. Kalogianni, R. Sileryte, M. Lam, K. Zhou, M. Van der Ham, S. Van der Spek, and E. Verbree, "Passive WiFi monitoring of the rhythm of the campus," in *AGILE International Conference on Geographic Information Science*, 2015, pp. 1–4.

[12] P. Reichl, B. Oh, R. Ravitharan, and M. Stafford, "Using WiFi technologies to count passengers in real-time around rail infrastructure," in *IEEE ICIRT*, 2018, pp. 1–5.

[13] B. Soundararaj, J. Cheshire, and P. Longley, "Estimating real-time high-street footfall from Wi-Fi probe requests," *International Journal of Geographical Information Science*, vol. 34, no. 2, pp. 325–343, 2020.

[14] C. Chilipirea, C. Dobre, M. Baratchi, and M. van Steen, "Identifying movements in noisy crowd analytics data," in *IEEE MDM*, 2018.

[15] G. Pipelidis, N. Tsiamitros, M. Kessner, and C. Prehofer, "HuMAn: Human movement analytics via WiFi probes," in *IEEE PERCOM*, 2019.

[16] L. Zhu, H. Tong, L. Lou, and Y. Xiong, "A passenger flow monitoring method in Hongqiao hub area based on gridded Wi-Fi sniffing," in *IEEE IMCEC*. IEEE, 2018, pp. 1052–1057.