Wi-Fi MAC address randomization vs Crowd Monitoring

ANDREI-ION COVACI, University of Twente, The Netherlands

Wi-Fi-based crowd monitoring is sensing crowds of people through the Wi-Fi probe requests broadcasted by their mobile devices. These probe requests contain valuable information that uniquely identifies a device (e.g. the MAC address) and, potentially, the person carrying it. Such risk represents a severe privacy issue, and the main countermeasure that mobile device and operating system manufacturers have against it is MAC address randomization. With the introduction of the General Data Protection Regulation (GDPR), many crowd-monitoring systems have been shut down due to the privacy-sensitive information they were storing. As a result, new crowd-monitoring systems had to be developed with people's privacy in mind. Such systems use various methods to anonymize the collected data before it is stored while providing ways to derive statistical counts from the anonymized data. This research focuses on the impact of MAC address randomization on this new generation of Wi-Fi based crowd-monitoring system that produce, as their only output, statistical counts on crowds. First, an analysis of how the various characteristics of MAC address randomization influence the statistical counts regarding crowd monitoring is given. Then, we implement a module for treating randomized MAC addresses and assess to what extent it can reduce their impact on the statistical counts.

Additional Key Words and Phrases: MAC address randomization, crowd monitoring, Wi-Fi, anonymization, footfall

1 INTRODUCTION

With the increasing popularity of smart mobile devices in recent times, Wi-Fi-based and Bluetooth-based crowd-monitoring systems have become standard practices for generating insights regarding the behavior of a crowd of people. For example, estimations such as the number of people present in a location or how a crowd moves in a given space proved to be very useful for analyzing mass events [2] or travel patterns in public transport [4]. Crowd-monitoring systems use the radio signals broadcasted by people's mobile devices close to a sensor. Such radio signals, also referred to as *'probe requests'* contain privacy-sensitive data, i.e. the MAC address of the device, which allows anyone to uniquely identify and track a mobile device and the person carrying it, thus infringing their privacy. Due to such severe privacy violations, several methods and regulations have been introduced to keep people safe.

In 2014, operating system and mobile device manufacturers introduced *MAC address randomization* to keep mobile devices anonymous when broadcasting probe requests in an unassociated state. This mechanism replaces the real MAC address of a device with a random address that periodically changes. If deployed correctly, it becomes significantly harder to identify or track a device, especially in a crowded environment, thus offering a better protection for individuals' privacy. Unfortunately, not having a well-defined standard for implementing this randomization process led manufacturers to develop separate processes for randomizing the MAC address. Because of this, different behaviours and inconsistencies of MAC address randomization appeared in practice. In the early days, Martin et al. [8] showed that some devices on the market are not using MAC address randomization at all. The ones that use randomization are still vulnerable to attacks or fingerprinting methods based on other fields that are part of the broadcasted probe requests, allowing eavesdroppers to track and identify mobile devices. Due to the increased awareness of this problem in recent years, manufacturers improved their randomization processes which had a positive impact with regards to users' privacy [7] [15]. However, these improvements have not made device fingerprinting impossible. Such fingerprinting methods have been leveraged in the past for crowd monitoring [14] [13] [10] and they still have the potential to bring a positive impact on the statistical counts of the new generation of crowd-monitoring systems.

To further regulate how people's data is used, various sets of rules have been introduced, such as the General Data Protection Regulation (GDPR) [1]. As a consequence, many Wi-Fi-based crowd-monitoring systems have been shut down due to their serious violations with regard to the privacy of the people. This led to the development of new crowd-monitoring systems that truly anonymize the information of the mobile devices being sensed while keeping a high accuracy of the estimations that the system is giving [12] [11]. The main idea of such systems is to only provide statistical counts about the data being sensed without storing any privacy-sensitive data or results that may trace back to a particular mobile device or person. Examples of such statistical counts include footfall, i.e. the size of a crowd present in a location, and crowd flow, i.e. the size of the flow of people traveling between several locations. In order to achieve this, several data structures, cryptographic techniques, and anonymization methods are used to hide the actual data that is being sensed while enabling the system to perform statistical counts on the anonymized data.

With a more extensive focus on individuals' privacy in crowd-monitoring, these state-of-the-art systems share certain particularities that heavily restrict the access to the captured data. Some of these particularities are: capturing data for a limited amount of time without leaving the sensor, discarding this data once the sensing period has ended, and encrypting any output that only trusted third parties can access. As a consequence, fingerprinting devices that use MAC address randomization becomes significantly harder. Thus, this research focuses on analyzing the impact of MAC address randomization on the statistical counts produced by the new generation of Wi-Fi-based crowd-monitoring systems. Moreover, we implement a module for treating randomized MAC addresses based on timing and Information Elements(IEs) fingerprints, and conduct experiments to assess to what extent this treatment can reduce the influence of randomized MAC addresses on the statistical

TScIT 37, July 8, 2022, Enschede, The Netherlands

counts. Our results indicate that, despite the restricted access to the sensed data, such fingerprints can significantly reduce the error of the footfall estimations.

This research can be split up in two goals.

Goal 1: To get a clear overview of how the different MAC address randomization schemes influence the statistical counts of the Wi-Fi-based crowd-monitoring systems
Goal 2: To see how randomized MAC addresses can be treated in order to help in minimizing their impact on the statistical counts

To achieve these goals, the following research questions will be answered:

> **RQ1:** How do the different MAC address randomization schemes influence the counting process with respect to privacy-preserving Wi-Fi-based crowd monitoring?
> **RQ2:** How can a Wi-Fi based crowd-monitoring system decide what randomization scheme was used for a randomized MAC address?
> **RQ3:** To what extent understanding the MAC address randomization scheme can be leveraged into a positive impact on statistical counts?

The rest of the paper is structured as follows. Section 2 introduces some of the work related to anonymization techniques and MAC address randomization treatments for Wi-Fi crowd monitoring. Then, in section 3 we discuss how some of the most important factors linked to MAC address randomization influence the pedestrian counts. Section 4 describes the model of the proposed system for dealing with random MAC addresses. This is followed by section 5, which details the two experiments that were performed in order to answer the research questions. The results of the experiments are presented and discussed in section 6, and the research concludes with section 7.

2 RELATED WORK

Leveraging the data from the probe requests broadcasted by mobile devices for crowd monitoring represents a significant risk for individuals' privacy if it is not properly anonymized. In 2020 and 2021, Stanciu et al. [12] [11] proposed a Wi-Fi-based crowd-monitoring system that uses epochs, i.e. a fixed period in which a sensing infrastructure collects probe requests, that provides, as its only output, statistical counts on pedestrian dynamics. These papers focus on methods used in such a system for anonymizing the identity of people being sensed.

The first proposed method leverages the k-anonymity principles while addressing some of the flaws of these principles. The system guarantees the MAC address anonymity of the sensed mobile devices without negatively impacting the statistical counts through pseudonymization, truncation, and correction operations.

The second method uses an anonymization process based on Bloom Filters and Homomorphic encryption right at the sensor node, which hides the actual identifiers of the sensed devices before the data is sent anywhere else. The collected information is anonymized to allow the system to derive statistical counts under encryption, just like if the data was not anonymized in the first place.

Although the researchers of these papers are aware of MAC address randomization, they do not address it in their work, which could lead to potential problems such as overcounting footfall and undercounting crowd flows. Overcounting footfall could happen when a device changes its MAC address multiple times near the same sensor, making the system think it senses multiple devices when there is only one. Undercounting a crowd flow could happen when a device changes its MAC address from one sensor to another. The system would sense the device at one sensor, and when the individual moves in the proximity of another sensor, the system would sense the same device but consider it different and not count it in the crowd flow.

In order to overcome the negative impact of MAC address randomization on crowd analytics, many crowd-monitoring solutions focused on linking the probe requests to the source device through various techniques.

Espresso [13] shows how such a solution can make use of the information elements, sequence number, and received signal strength indicator present in probe requests to estimate the probability of associating a probe request to the source device. The main differences between Espresso and the system proposed by [12] and [11] is that Espresso does not explicitly use epochs when collecting probe requests, it sends the sensed data to a central server without fully anonymizing it, and it uses some of this data to constantly train the system. Moreover, Espresso uses a probabilistic model, which had to be trained on the data collected in a 24 hours window before it could be used in the evaluation of the system. Besides, the researchers only focused on the footfall of a crowd without addressing crowd flows.

Aforos [14] is another similar crowd-monitoring system that generates a fingerprint for each probe request based not only on the MAC address but also on the information elements. Then, the probe requests with the same fingerprint are grouped and linked to a single device. Just as in the case of Espresso, Aforos does not fully anonymize the data of the sensed devices due to the use of pseudonymization, and it only focuses on footfall. Moreover, their method required adjustments to improve its accuracy, and the estimations of the system were based on the data collected in a window of three hours. Furthermore, their sensing infrastructure was composed of only one sensor compared to the system presented in [12] and [11] that used multiple sensors.

3 INFLUENCE OF MAC ADDRESS RANDOMIZATION ON PEDESTRIAN COUNTS

A crowd-monitoring system that fully anonymizes the sensed data follows a strict set of rules in terms of data collection and processing.

First, the data is sensed in epochs. An *epoch* is defined as a fixed period of time in which the system's sensors collect probe requests from nearby devices. Once an epoch ends, the sensors compute the desired counts based on the MAC addresses extracted from the probe requests, then they discard all the data sensed during that epoch. Finally, the resulting counts are encrypted and sent to a central server. For our purpose, we differentiate between two types of counts:

(1) **Footfall**: the number of people present in one place during a particular period of time
(2) **Crowd flow**: the amount of people moving from one place to another

In the context of such a crowd-monitoring system, although not directly related to MAC address randomization, we identified that the device's characteristic with the most significant impact on the statistical counts is the inter-burst arrival time (IBAT). Most devices transmit probe requests in very short bursts of at most 500 ms across the different Wi-Fi channels [3]. The inter-burst arrival time (IBAT) is defined as the difference in arrival times between two consecutive bursts of probe requests from the same device. It was shown before that the IBAT of a device stays constant, regardless of MAC address randomization [3]. Thus, looking at this characteristic can help us better understand the frequency of random MAC addresses broadcasted by the same device.

Furthermore, Martin et al. discovered in [8] that mobile devices that use MAC address randomization keep the randomized MAC address for at least one burst before changing it. Therefore, another essential characteristic that influences the pedestrian counts is the time between two changes of a device's MAC address or the random MAC address lifetime.

We continue with an analysis of these two device properties with respect to a time characteristic of the crowd-monitoring system, namely the length of an epoch.

3.1 Analysis

We differentiate between six configurations of these parameters, depicted in Figure 3, Appendix B. There is a total of eight configurations possible. However, for scenarios *a* and *d* the size of the random MAC address lifetime does not matter.

In case *a*, we have a relatively big inter-burst arrival time and a long epoch. We observe that, regardless of the lifetime of the random MAC address, there is a burst transmitted every epoch with a different MAC address. This does not impact the counts for footfall since the source device sends only one burst per epoch that is counted once. However, this impacts the counts for crowd flow because any two consecutive bursts have different MAC addresses, which causes the system to treat them as coming from separate devices.

Case *b* depicts a scenario where the IBAT is short, but the random MAC address lifetime and the epochs are long. We see that there are multiple bursts with the same MAC address that are captured during the same epoch. As in the first scenario, the counts for footfall are not influenced since the MAC address is counted only once, but the counts for crowd flow are affected because the MAC address changes between epochs.

For case *c* we have a short IBAT, a short random MAC address lifetime, and a long epoch. This configuration causes the source device to transmit multiple bursts with different MAC addresses per epoch. Thus, the system treats all MAC addresses as different devices, which leads to overcounting the footfall and undercounting the crowd flow.

In case *d* we consider a long IBAT and a short epoch. Regardless of the random MAC address lifetime, we notice that if the epochs are too short, the sensing infrastructure may not be able to detect the presence of the source device during some epochs, which leads to undercounting the footfall and the crowd flow for that epoch. However, this issue occurs despite MAC address randomization, because it is caused by external factors that cannot be controlled by the system.

With a short epoch length and a short IBAT but a long lifetime of the random MAC address, case *e* shows that the source device can transmit a burst every epoch, keeping the same MAC address for a few consecutive epochs. Thus, the footfall is not affected in this scenario. However, the system undercounts the crowd flow for every two consecutive epochs when the device changes its MAC address.

Finally, for case *f*, we have a short IBAT, a short random MAC address lifetime, and a short epoch. We observe that the source device can transmit a burst every epoch but with a different MAC address. Thus, the system is able to count the footfall correctly but undercounts the crowd flow due to the different MAC addresses of the same device in every epoch.

The above scenarios are a simplified version of the most important patterns that are useful in understanding the impact on the statistical counts. In reality, there is a possibility to encounter different edge cases. For example, in any of the above situations, a burst of probe requests can be split between two epochs and, as a result, the device is sensed during both epochs. However, depending on the pattern, such edge case would still fall under one of the scenarios discussed above.

Furthermore, there are still many external factors that could impact the statistical counts of a crowd-monitoring system. For example, in 2015, Julien Freudiger showed in [6] how the different configurations of mobile devices affect the amount of broadcasted probe requests, as well as the frequency of probing. This has a direct impact on any Wi-Fi-based crowd-monitoring system regardless of MAC address randomization because the sensing infrastructure may not be able to detect a mobile device in range if its probe request frequency is reduced.

Given this overview, we propose a system model addressing case *c* for a couple of reasons. First, this scenario has the biggest impact on

the statistical counts, due to the multitude of random MAC addresses coming from the same device during one epoch. Depending on how short is the IBAT of the sensed devices, the footfall could be significantly overestimated. Second, working with a sufficiently long epoch allows the system to collect enough data that could offer sufficient insights to mitigate the impact of random MAC addresses.

4 SYSTEM MODEL

A crowd-monitoring system is usually deployed in a crowded space to get insights into the behavior of crowds of people. This is done by installing a sensing infrastructure in the chosen space. Then, various techniques are applied to the sensed data to derive statistical counts for pedestrian dynamics. This paper focuses on a component designed for crowd-monitoring systems that use Wi-Fi as sensing technology and anonymize any collected data at the sensor level. We start by describing the overview of such a system. Then, we discuss the component that deals with random MAC addresses in the context of the targeted crowd-monitoring system.

4.1 Current crowd-monitoring system

We assume that the model of the crowd-monitoring system used is similar to the one described in [11]. The main components of this system are the sensing infrastructure that collects probe requests from the mobile devices in range, and a central unit that handles the communication between sensors and the consumers that send queries to the system. Figure 1 depicts the architecture of the service.

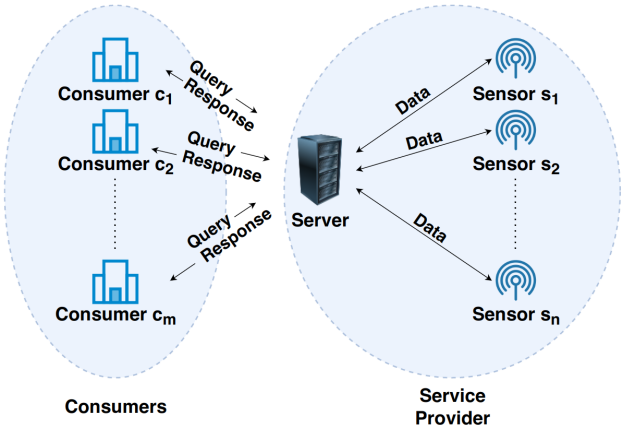

Fig. 1. Service Architecture

The whole system keeps track of time through epochs. If a sensor detects probe requests sent by a mobile device in range, it extracts the broadcasted MAC address and assigns it to the current epoch based on the timestamp of the reading. Because mobile devices can transmit multiple probe requests with the same MAC address during one epoch, a sensor will always remove any duplicate MAC addresses detected during that period of time. Moreover, at the end of an epoch, a sensor must discard any sensed data and only send back to the server the response to a query if the query targets that specific sensor. Figure 2 shows the workflow of such a sensing device.

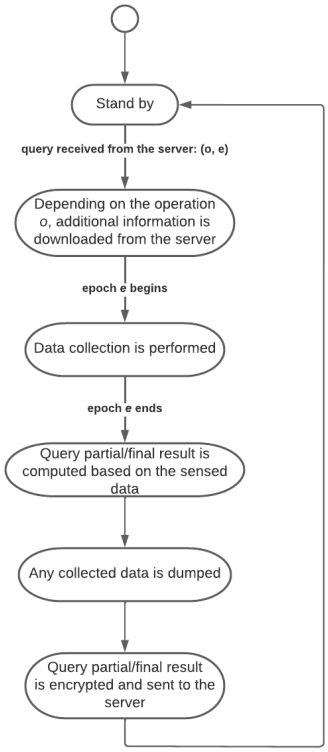

Fig. 2. Workflow of a sensor in the current crowd-monitoring system.

4.2 Dealing with random MAC addresses

In order to detect and treat random MAC addresses coming from the same device, we need access to the probe requests collected by the sensors. The design of the targeted crowd-monitoring system imposes that sensors discard any collected data at the end of an epoch, so probe requests will never leave the sensing device. Therefore, a module that deals with random MAC addresses can only be implemented at the sensor level, after the data collection period has ended and right before the query result is computed.

In the context of the current crowd-monitoring system, a module that mitigates the effects of MAC address randomization has to fulfill some requirements:

Reliably identify random MAC addresses

Process a variable volume of data - This is because the length of an epoch can vary from a few minutes to a few hours based on where the crowd-monitoring system is deployed, so the amount of probe requests available per epoch also varies.

Comply with the privacy requirements of the crowd-monitoring system - This means that the treatment module cannot store or learn any patterns that can be later used to

identify a particular device or group of devices. The component must gather insights only from the data collected during one epoch, let the sensor compute the result of the received query, then discard these insights and start over again for the next epoch.

Be computationally inexpensive - Usually, sensors are small, they run on batteries, and have limited resources. Furthermore, to preserve privacy as much as possible, the sensor already has to encrypt any data sent back to the server and apply various hashing algorithms to work with the anonymized data [12] [11].

Increase the accuracy of the statistical counts - The treatment component has to reduce the impact of random MAC addresses as much as possible, bringing the statistical counts as close to the real values as possible.

Figure 4 in Appendix A depicts the model of the treatment component. Once a sensor finishes collecting bursts of probe requests broadcasted in its proximity, it separates them into bursts with randomized MAC addresses and burst with non-randomized MAC addresses. In the latter case, the system simply extracts all unique MAC addresses. For the bursts with randomized MAC addresses, the system uses a "treatment station" to cluster them. Ideally, each cluster would represent a different source device that could be identified by a set of signatures. Combining these two outputs, a sensor can calculate the footfall by simply counting the unique non-randomized MAC addresses and the clusters formed by the treatment station. To calculate the crowd flow between two sensors, the system would perform the intersection of the two sets of unique non-randomized MAC addresses, and the intersection of the two sets of cluster signatures. The combined cardinality of the resulting sets represents the crowd flow.

4.3 Discussion

The biggest challenge that the proposed system has to overcome when calculating crowd flows is sharing the clusters of bursts with random MAC addresses from one sensor to another while preserving the privacy requirements of the current crowd-monitoring system. It is hard to suggest a general solution to this problem, because every crowd-monitoring system uses different techniques to anonymize and secure the data shared from the sensing infrastructure. For example, the crowd-monitoring system in [12] uses pseudonymization, truncation and correction operations to make the extracted MAC addresses k-anonymous, which ensures, under some condition, that the new identifiers cannot trace back exactly to the corresponding sensed devices. In our case, for the devices that broadcast random MAC addresses, we are not using only one MAC address to identify them, but a set of MAC addresses and signatures that can become very large, depending on how often a device changes its MAC address during an epoch. Thus, such operations may not even be applicable on the resulting signatures of the cluster that we obtain for a device.

For the crowd-monitoring system model [11] that we chose as the base for our treatment component, a possible solution could involve selecting the best signature of each cluster as an identifier for the source device represented by that cluster. In this way, the system could hash the signature using the selected hash functions, and mark it in a bloom filter that can safely be sent to another sensor under encryption. However, such solution involves a possible drop in accuracy, due to the lower amount of signatures shared per cluster. Furthermore, the solution assumes that the chosen hash functions addresses the fact that the signatures of two bursts coming from the same device could be different, but still sufficiently similar to be marked as having the same source. Investigating what would be the best hash functions to use in this context, or whether or not such hash functions exist is outside the scope of this paper.

5 EXPERIMENTS

In order to assess the functionality of the proposed system and to what extent it can improve the pedestrian counts, we implemented a proof of concept algorithm focused on counting the footfall, and conducted two types of experiments: an in the wild experiment and a dataset experiment.

5.1 Implementation

The implementation focuses only on the process that happens at the sensor level. The base of our program consists of a few modules that allow us to collect probe requests in epochs, discard any sensed data once the footfall is computed, group the probe requests in bursts using their MAC address and sequence number and split these bursts based on their MAC address type.

The probe request collector has two operation modes:

(1) **Sni□er**: In this mode our program sets the Wi-Fi interface of the device it runs on in monitor mode and starts hoping over the 14 different 2.4 GHz Wi-Fi channels. On each channel, the program collects probe requests for 5 seconds then it moves on to the next channel.
(2) **Dataset**: When this mode is selected, the program processes the dataset files based on a few parameters, such as the epoch length or a list of preferred device labels.

In order to group the bursts with randomized MAC addresses coming from the same source device, we use DBSCAN [5], a well known clustering algorithm, and two types of fingerprints. The first fingerprint consists of an ordered list of Information Elements (IEs) and certain bitmasks available in a probe request, which offer high entropy between device models and low entropy between the probe requests of the same source device [3] [8]. The second fingerprint is borrowed from [3], and it is based on the burst length and the arrival time difference between probe requests within that burst, or the Inter-Frame Arrival Time (IFAT). The main assumption behind this fingerprinting method is that mobile devices transmit probe requests regularly for a fixed amount of time. To generate this fingerprint, for every burst we split its length in equally sized bins, distribute the probe requests in bins depending on their arrival time, and calculate the mean IFAT and the percentage of probe requests for each bin. The list of all these values and the burst length represent the time signature of the burst.

5.2 Dataset

Due to the challenging task of obtaining the ground truth from wild data, we also use a real-world dataset [9] to evaluate how well our proof of concept can improve the statistical counts. This dataset was published in 2021 and consists of labelled probe requests with random and real MAC addresses, which were broadcasted by 22 devices in different modes. The probe requests were captured in a laboratory setting that made it possible to scan the devices individually. Such recent dataset offers us qualitative data, very close to real-world scenarios, that we can use as ground truth in our experiment.

5.3 Ethical consideration

In order to test the proposed system, both experiments involve using probe requests coming from real mobile devices. This kind of data represents a threat to individual's privacy if not handled properly.

The probe requests originating from the dataset have been collected in a laboratory setting and a set of rules have been followed to allow this dataset to be made public [9]. Thus, we can safely assume that this data does not pose any privacy risks. However, the probe requests captured in the wild could potentially trace back to certain individuals. Because of this issue, our program does not store any of the captured data. The collected probe requests are used only during the epoch in which they are sensed and they are discarded once the footfall is computed and the epoch ends.

5.4 Metrics

The main goal of the proposed system is to improve the pedestrian counts as much as possible. Therefore, we evaluate how well our proof of concept can improve the statistical counts based on the relative error reduction (RER) percentage of the counts. This is calculated using the following formula:

$$RER = \left(1 - \frac{Relative\ error\ post\text{-}treatment}{Relative\ error\ pre\text{-}treatment}\right) \cdot 100$$

where the relative error is calculated as follows:

$$Relative\ error = \frac{|Estimated\ count - Real\ count|}{Real\ count}$$

5.5 Experimental setup

5.5.1 In the wild experiment. To test the functionality of our proof of concept, we setup a laptop in one of the canteens from the university campus during lunch time, which is one of the busiest times of the day. We set our program into *Sniffer* mode and we let it collect probe requests in epochs of 5 minutes for 1 hour.

5.5.2 Database experiment. For the database experiment, we set our program into *Dataset* mode, which extracts the probe requests from the *pcap* files of all devices that are in the same state. This allows us to better observe the impact of the different device states on the footfall. Then, the program normalizes the timestamps of the probe requests and groups these packets in 5 minutes epochs based on the new timestamps. For each epoch our system computes the real amount of devices that sent the probe requests and then it feeds each of these groups to the module responsible for calculating the footfall.

6 RESULTS

6.1 In the wild experiment results

Because we did not have access to the ground truth data of the in the wild experiment, i.e. the real amount of devices in the range of our sensor during every epoch, we cannot assess how accurate our footfall estimations are. However, in terms of the functionality of our proof of concept, we made the following observations:

- The sniffer can indeed collect probe requests from every Wi-Fi channel that the Wi-Fi interface is set to.
- The data collected was similar to the one from the dataset. This means that the wild probe requests contained randomized and non-randomized MAC addresses, and, most of the time, they could be grouped in short bursts of at most 500 ms. This was observed by printing on the screen whether or not the packets had a random MAC address and the timestamps when the packet was collected. No other data, such as the source address of the packet, have been observed.
- The program computed the footfall for every epoch, despite the volume of data that it had to process.
- Once the footfall of an epoch was computed, the data collected during that epoch was indeed discarded and could not be accessed anymore.

6.2 Dataset experiment results

The results of the dataset experiment are presented in Table 2. The values of the *Device mode* column are further explained in Table 1. For every device mode, we could split the available data into 4 epochs of 5 minutes each.

It can be observed that our proof of concept obtained a footfall error reduction of over 90% for counting the devices that had their Wi-Fi turned on, which shows an enormous improvement compared to the footfall estimations in which the randomized MAC addresses have not been treated. For counting the devices that did not have their Wi-Fi turned on, the error reduction stays below 90%. However, in those cases the volume of data was small and the estimations are still only 1 device away from the real count. These results are very promising and show that we can significantly mitigate the effects of MAC address randomization for the footfall count.

Furthermore, we also observed that our random MAC addresses treatment module does not perfectly cluster the bursts of probe requests by the source device, meaning that there is still a big chance that these devices cannot be identified, which offers some privacy protection. However, we noticed that even if this clustering error exists, the footfall estimations are still significantly improved, which is the actual goal of the proposed system.

7 CONCLUSIONS AND FUTURE WORK

Crowd monitoring is a domain that challenges individuals' privacy through the type of data it uses to produce insights into crowds of people. As more and more concerns regarding data privacy and

Table 1. Device modes ("X" means that the relevant mode is "on")

Mode	Active screen on	Wi-Fi on	Power saving on
A	X	X	
S		X	
PA	X	X	X
PS		X	X
WA	X		
WS			

Table 2. Dataset experiment results. The *RER* column represents the percentage of the relative error reduction of the counts. Each value from the *Device mode* column is explained in Table 1

Device mode	Epoch	Counts before treatment	Counts after treatment	Real count	RER %
S	0	398	18	22	98.936 %
	1	413	20	21	99.745 %
	2	333	16	20	98.722 %
	3	230	16	19	98.578 %
A	0	123	22	21	99.02 %
	1	103	23	20	96.386 %
	2	86	24	20	93.939 %
	3	76	21	18	94.828 %
PS	0	323	18	20	99.34 %
	1	303	14	20	97.88 %
	2	265	18	20	99.184 %
	3	104	16	18	97.674 %
PA	0	173	26	21	96.711 %
	1	102	27	21	92.593 %
	2	87	27	21	90.909 %
	3	90	19	20	98.571 %
WS	0	21	11	12	88.889 %
	1	12	7	7	100 %
	2	10	6	6	100 %
	3	7	4	4	100 %
WA	0	21	10	11	90 %
	1	14	5	6	87.5 %
	2	13	6	7	83.333 %
	3	11	5	5	100 %

usage are raised, various solutions had to be implemented to address these privacy issues in crowd monitoring. On one hand, new crowd-monitoring systems have been designed with people's privacy in mind. On the other, mobile device and operating system manufacturers introduced MAC address randomization to hinder device tracking. However, the latter solution proved to also affect the accuracy of the crowd-monitoring system's counts. Many studies have attempted to defeat this randomization mechanism with respect to crowd monitoring, but, to the best of our knowledge, not many considered individuals' privacy in such systems.

This paper provides an analysis on how the pedestrian counts of this new generation of Wi-Fi based crowd-monitoring systems are affected by MAC address randomization. Furthermore, a system model that deals with random MAC addresses and improves the statistical counts, namely footfall and crowd flow, is presented in the context of a crowd-monitoring system that fully anonymizes the sensed data. A proof of concept for the proposed system was implemented focused on calculating the footfall, which attempts to cluster randomized MAC addresses coming from the same device using timing and Information Elements fingerprints. To test the functionality and accuracy of our implementation, we conducted an in the wild experiment and a dataset experiment. Results showed that the system can function properly in a real-world scenario and that it can reduce the error of the statistical counts by more than 90% when counting devices that have their Wi-Fi turned on. And even for those that do not have their Wi-Fi turned on, our estimations were only 1 device away from the real count. Such a significant error reduction clearly shows that MAC address randomization can be mitigated in the context of Wi-Fi based crowd-monitoring systems that produce, as their only output, pedestrian counts.

Although the results of this study are very promising with respect to calculating the footfall, our proof of concept have not addressed crowd flows yet. The biggest challenge here is passing the information about the clusters of bursts of probe requests with random MAC addresses from one sensor to another while still preserving the privacy requirements of the crowd-monitoring system. Thus, further research is required in order to determine how well the proposed system model can improve the crowd flow estimations.

REFERENCES

[1] 2016. General Data Protection Regulation. 2016. Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016. *Official Journal of the European Union* (2016).

[2] Anas Basalamah. 2016. Crowd mobility analysis using WiFi sniffers. *International Journal of Advanced Computer Science and Applications* 7, 12 (2016), 374–378.

[3] M. Célestin. 2017. *Wi-Fi Tracking: Fingerprinting Attacks and Counter-Measures.* Ph. D. Dissertation. Université de Lyon, Lyon, France. NNT: 2017LYSEI114..

[4] M. Dunlap, Z. Li, K. Henrickson, and Y. Wang. 2016. Estimation of origin and destination information from Bluetooth and Wi-Fi sensing for transit. *Transportation Research Record* 2595, 1 (2016), 11–17.

[5] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd* 96, 34 (1996), 226–231.

[6] Julien Freudiger. 2015. Short: How Talkative is your Mobile Device? An Experimental Study of Wi-Fi Probe Requests. *Proceedings of the 8th ACM Conference on Security and Privacy in Wireless and Mobile Networks* (2015).

[7] J. Martin, E. Fenske, D. Brown, T. Mayberry, P. Ryan, and E. Rye. 2021. Three Years Later: A Study of MAC Address Randomization In Mobile Devices And When It Succeeds. *Proceedings on Privacy Enhancing Technologies* 2021 (7 2021), 164–181. https://doi.org/10.2478/popets-2021-0042

[8] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E. C. Rye, and D. Brown. 2017. A Study of MAC Address Randomization In Mobile Devices And When It Fails. *Proceedings on Privacy Enhancing Technologies* 2017 (2017), 365–383. https://doi.org/10.1515/popets-2017-0054

[9] L. Pintor and L. Atzori. 2021. A dataset of labelled device Wi-Fi probe requests for MAC address de-randomization - 2021. *Mendeley Data, V1* (2021). https://doi.org/10.17632/j64btzdsdy.1

[10] R. Ribeiro, B. Rodrigues, C. Killer, L. Baumann, M. Franco, E. Scheid, and B. Stiller. 2021. ASIMOV: a Fully Passive WiFi Device Tracking. *2021 IFIP Networking Conference (IFIP Networking)* (2021). https://doi.org/10.23919/ifipnetworking52078.2021.9472786

[11] V. Stanciu, M. Steen, C. Dobre, and A. Peter. 2021. Privacy-Preserving Crowd-Monitoring Using Bloom Filters and Homomorphic Encryption. *Proceedings Of The 4Th International Workshop On Edge Systems, Analytics And Networking* (2021). https://doi.org/10.1145/3434770.3459735

[12] V. Stanciu, M. van Steen, C. Dobre, and A. Peter. 2020. k-Anonymous Crowd Flow Analytics. *Mobiquitous 2020 - 17Th EAI International Conference On Mobile And Ubiquitous Systems: Computing, Networking And Services* (2020). https://doi.org/10.1145/3448891.3448903

[13] J. Tan and S. Gary Chan. 2021. Efficient Association of Wi-Fi Probe Requests under MAC Address Randomization. *IEEE INFOCOM 2021 - IEEE Conference On Computer Communications* (2021). https://doi.org/10.1109/infocom42981.2021.9488769

[14] M. Vega-Barbas, M. Álvarez Campana, D. Rivera, M. Sanz, and J. Berrocal. 2021. AFOROS: A Low-Cost Wi-Fi-Based Monitoring System for Estimating Occupancy of Public Spaces. *Sensors 2021, 21* 3863 (2021). https://doi.org/10.3390/s21113863

[15] Z. Zhu, S. Chen, and L. Lu. 2021. Research on Real MAC Address Acquisition Technology for WIFI Connection. *2021 2Nd International Seminar On Artificial Intelligence, Networking And Information Technology (AINIT)* (2021). https://doi.org/10.1109/ainit54228.2021.00115

A ANALYSIS DIAGRAM

B SYSTEM MODEL DIAGRAM

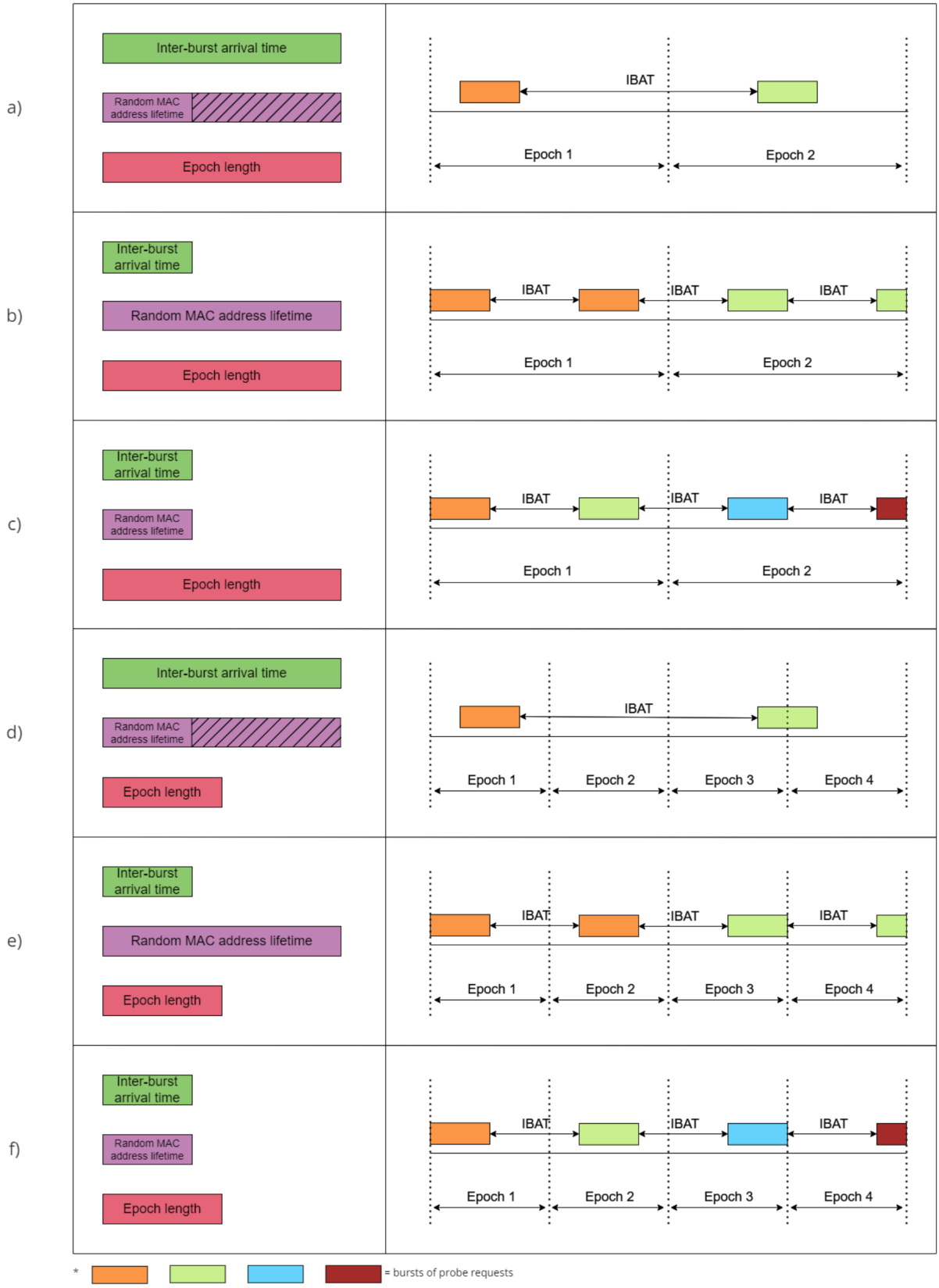

Fig. 3. Effects of different IBAT, random MAC address lifetime and epoch length configurations on bursts detection. First column shows the length of each parameter, which can be short or long, and the second column shows a detection scenario based on the corresponding configuration of the parameters. The different colors of the bursts represent a different MAC address for all probe requests within that burst.