Privacy-aware Publication of Wi-Fi Sensor Data for Crowd Monitoring and Tourism Analytics

LEONIE ACKERMANN, University of Bamberg, Germany
CHRISTOPH BAUM, University of Bamberg, Germany
SYED IBRAHIM KHALIL, University of Bamberg, Germany
ALEKSANDR LITVIN, University of Bamberg, Germany
DANIELA NICKLAS, University of Bamberg, Germany

Estimating visitor frequency in urban areas often relies on camera-based solutions or the active participation of individuals using smartphone applications or devices with RFID or Bluetooth technology. This paper presents the results of a preliminary study on anonymous data collection as a basis for data-driven visitor guidance in Bamberg's Old Town, using passive, non-intrusive and low-cost Wi-Fi sensors. The study includes a field test installation to evaluate data quality under robust anonymization measures. The data collected as part of the project will be made available to the public in the Mobilithek of the Federal Ministry of Digital Affairs and Transport (BMDV). Still, the collection of Wi-Fi probe requests raises legitimate privacy concerns. We address potential attack models for identifying and tracking devices based on these requests and explain the data collection architecture and Technical Data Protection Concept implemented within CrowdAnym to mitigate these risks. We evaluate the impact of our approach on the quality of the CrowdAnym dataset by comparing the data from one sensor location with the number of people counted by a nearby laser scanner. We are able to show that our approach approximates visitor density well, however the deviation of our data from ground truth increases as visitor frequency increases.

CCS Concepts: • **Information systems → Mobile information processing systems**; • **Computer systems organization → Sensor networks**; • **Security and privacy → Pseudonymity, anonymity and untraceability**.

Additional Key Words and Phrases: MAC address detection, datasets, anonymization

ACM Reference Format:
Leonie Ackermann, Christoph Baum, Syed Ibrahim Khalil, Aleksandr Litvin, and Daniela Nicklas. 2023. Privacy-aware Publication of Wi-Fi Sensor Data for Crowd Monitoring and Tourism Analytics. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Local congestion occurs frequently in Bamberg's Old Town, which is popular with tourists and locals alike. In the future, a data- and sensor-based system is intended to provide smart recommendations and to monitor the effects of measures that aim to improve the situation. The preliminary study CrowdAnym explored the potential of anonymous data collection for data-driven visitor guidance in Bamberg's Old Town. It involved a field test installation to assess data quality, acceptance levels among the population, and which further applications are possible with an OpenData

Manuscript submitted to ACM

provision of the data. Since it is our goal to publish a curated data set in the Mobilithek[1] of the Federal Ministry of Digital Affairs and Transport (BMDV) at the end of the project, we focus on approaches that ensure that the collected data can be made available as OpenData without allowing re-identification and tracking of individuals, while keeping the data usable for analysis.

There are several approaches for crowd size estimation in urban areas. Camera-based solutions [10] [12] [16] are a promising alternative to manual estimation. However, they raise privacy concerns and are also very energy-intensive. Non-image-based localization solutions usually require the active participation of humans. Here, either a smartphone application [9] must be installed or self-built devices with RFID tags [7] or Bluetooth [15] must be carried voluntarily. 3D distance sensors, such as laser scanners [3] or LIDARs are a passive alternative; however, these options can be cost-prohibitive and pose installation challenges, particularly when dealing with historic listed buildings within Bamberg's Old Town. However, it is a suitable method as a source of ground truth.

The system we implemented uses Wi-Fi sensors as a passive, non-intrusive, and low-cost alternative [14] to estimate visitor frequency. As you can see in Figure 1, sensors (called bz2463, bz2454, etc.) are installed in several locations (Points of Interests, like Dom, Obere Brücke or the tourist information). They collect Wi-Fi probe requests within a range between 30 and 60 meters, depending on the conditions of the installation site. The ranges of most sensors do not overlap. They transmit the data as a JSON string via https to a data endpoint provided by the University of Bamberg. The endpoint authenticates the data to ensure that only the sensors installed by the project can send data. The Datastream Management System Odysseus [2] extracts and structures the data so that it can be stored in a relational Postgres database on which the data analysis is performed. Figure 1 shows a model of the workflow of data collection and processing.

As there are privacy risks associated with the collection of Wi-Fi probe requests [1][4][13], we discuss attack models that can be used to identify and track devices based on their Wi-Fi probe requests in Section 2. In Section 3, we introduce the Technical Data Protection Concept we implemented to prevent identification and tracking of individuals. MAC randomization, a privacy protection feature that has increased in prevalence in recent years and is used by many modern devices (e.g., Apple, Android, Windows) [6], helps prevent identification and tracking. The impact of this technique on data quality was evaluated by Rütermann et al. [8]. In Section 4, we evaluate the data quality and utility of the CrowdAnym dataset, and conclude in Section 5 with a summary of our results and an outlook regarding future work.

2 ATTACK MODELS

When Wi-Fi probe requests are collected across a wide area, and an attacker gains access to the dataset, there is a risk that individuals could be tracked, potentially re-identified, and their personal trajectories exposed. Wi-Fi management frames included in probe requests contain a rich set of parameters from which a highly specific device signature can be derived [5]. These signatures allow a passive adversary to perform device fingerprinting. This is effective even when attempts are made to anonymize Wi-Fi probe data, e.g., using hash values, or to mask device identity using the MAC randomization technique. As several researchers have demonstrated [13][5][4] signatures for most devices allow fairly granular device binning, enabling adversaries to determine a small class of devices that could have produced any given probe request.

In addition, probe requests can contain identifying information about the device owner depending on the age of the device and its OS, like the preferred network list (PNL), which includes networks identified by their SSIDs. According

[1] https://mobilithek.info/ [accessed 2023-09-14]

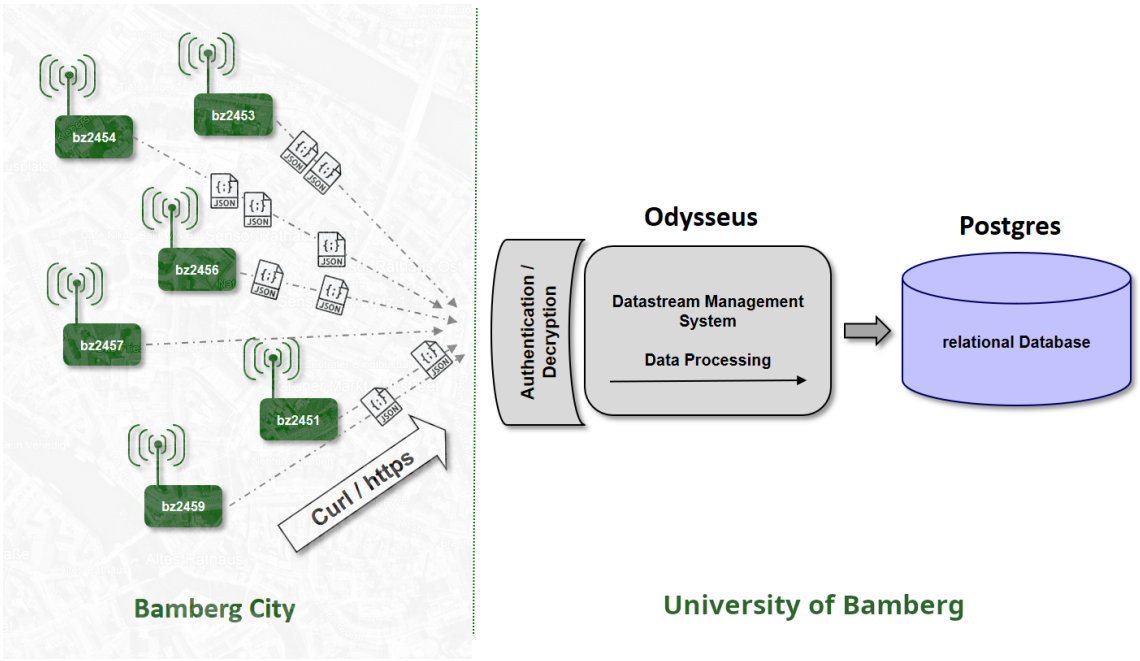

Fig. 1. System Model of Sensor Installation

to measurements by McDougall et al. [1], around 23% of probe requests contain SSIDs of networks the devices were connected to in the past. Thereby a SSID could potentially expose not only the device owner's home or workplace but also other visited locations where Wi-Fi connections were established, thus revealing highly sensitive personal details. Furthermore, they observed that 11.8% of probe requests containing SSIDs may disclose passwords within the SSID field. Since the SSIDs are fixed, they enable tracking of devices even if their MAC address is randomized.

To summarize, raw Wi-Fi probe requests contain sensitive information that can be used to re-identify users. Hence, we need measures to convert this data into a publishable dataset.

3 TECHNICAL DATA PROTECTION CONCEPT

As discussed in the Section 2, Wi-Fi probe requests are vulnerable to attacks such as fingerprinting, can reveal personal information, and enable involuntary tracking of device owners. Moreover, since MAC addresses are personal data according to Art. 4 No. 1 GDPR, appropriate anonymization techniques must be applied to these data before they are stored and published.

In the following sections, we outline our approach to privacy-sensitive detection of Wi-Fi probe requests for the purpose of estimating visitor frequency, as well as additional measures to prevent identification and tracking of individuals. A two-stage approach is followed here: The basic measures take effect directly at the edge during data collection, and application-specific anonymization measures are then applied to the data before it is published.

3.1 Basic Measures

First of all, we installed the sensors only in touristic relevant locations. Most of the sensor coverage areas do not overlap, there are large gaps between the sensors. This prevents extensive tracking from the outset. Secondly, we hash the detected MAC addresses directly on the sensor using the SHA 224 algorithm. The controller never uses the MAC address in plain text and only transmits it hashed to the data endpoint. The salt for the hashes is recreated daily on each controller using the date and a seed specified via XML parameters. Optionally, the salt can be sent along with a data endpoint and is then appended to the rest of the dataset. The seed is never passed on to third parties. Finally, we do not transmit or store of probe request frames and SSIDs. As stated in Section 2, these data could be used for fingerprinting and leak sensitive information. Therefore they are not part of the CrowdAnym data collection.

3.2 Application-specific Anonymization Measures

In addition to these basic measures, we remove MAC addresses from static devices. The sensors detect all Wi-Fi probes within their range, including those sent by devices such as printers, laptops and smart home devices. Detecting devices from residents or nearby businesses and their employees is not relevant for estimating visitor frequency. By analyzing individual MAC addresses and their occurrence within a day, it became clear that MAC addresses with a frequency of 10 or more have a long dwell time at the sensor location. Below this limit, devices mostly appear temporarily in the data. Therefore MAC addresses that appear in the data more than 10 times per day are removed from the published data. Finally, we remove periods with low frequencies. Since the purpose of the data collection is to monitor and react on overcrowding, data is removed when there is little activity. Such periods include nights when streets are less busy. Periods when fewer than 10 unique MAC addresses are captured in a 10-minute period are not published.

4 EVALUATION

The dataset for publication in the Mobilithek includes data collected between July 10, 2023 and August 20, 2023. It was anonymized as described in Section 3. The dataset includes a timestamp, sensor location, hashed MAC address, and RSSI[2] value. To evaluate the quality of the CrowdAnym data, we compared the number of unique MAC addresses collected by the Wi-Fi sensor at the Gabelmann location with visitor frequency data collected by a laser scanner installed by hystreet.com GmbH in close proximity. We define visitor frequency as the number of people counted on site, in this case the number of people passing through the light curtain of the laser scanner.

The data from the hystreet and the Gabelmann sensor were compared at ten-minute intervals, and the deviation of the Gabelmann sensor was calculated as a percentage. In Figure 2 the deviation is aggregated for each week. It is noticeable that more and larger deviations occur in calendar week 28. We assume that this is related to the popular street artist event *Bamberg zaubert*, which took place in the city center from July 14 to 16. As Figure 3 shows, the higher the visitor frequency, the higher the deviations of the Gabelmann from the hystreet data.

Our dataset generally matches visitor frequency data well, with minimal deviations when visitor frequency is low, while higher visitor frequencies lead to larger deviations in Wi-Fi sensor data. This insight underscores the complexity of data collection in real-world settings and the challenges introduced by MAC randomization. While it is not surprising that the number of unique MAC addresses in the CrowdAnym dataset may deviate from the actual number of pedestrians in the hystreet dataset due to the presence of fake devices[8], this observation underscores the need for robust device-person mapping techniques in data analysis.

[2]Received Signal Strength Indicator

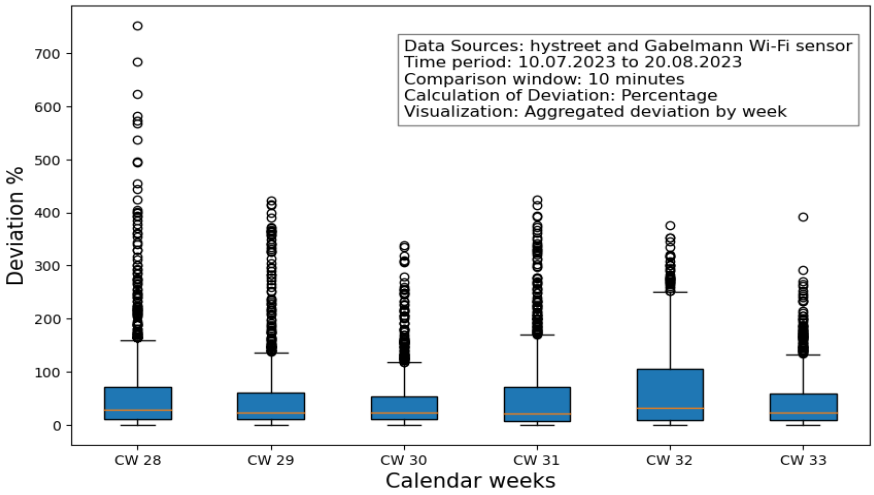

Fig. 2. Deviation by week between laser scanner and Wi-Fi sensor

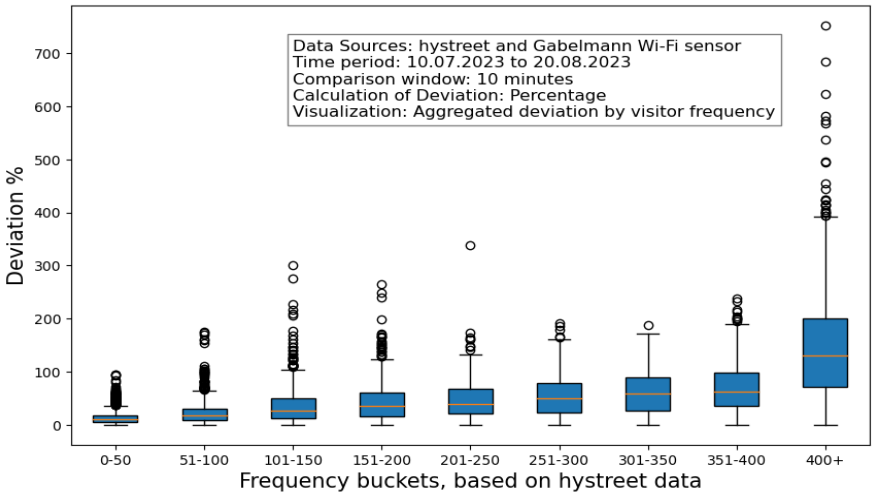

Fig. 3. Deviation by visitor frequency between laser scanner and Wi-Fi sensor

In summary, our evaluation provides insight into the CrowdAnym dataset's strengths and limitations, as well as the intricacies of working with real-world sensor data.

5 CONCLUSION AND FUTURE WORK

Using Wi-Fi trackers to estimate visitor frequency has proven effective. However, additional ground truth measurements are needed for a more precise estimate of actual crowd size. Another avenue for future research is to explore advanced processing techniques. Investigating the possible elimination of fake devices and implementing error estimation models directly at the sensor level could be a promising endeavor. In future work, we plan to expand the installation to include

laser scanners at selected locations. This might allow us to develop online calibration methods for Wi-Fi probe based monitoring, similar to Solmaz et al. [11].

ACKNOWLEDGMENTS

To safectory GmbH, our industry project partner in CrowdAnym, for providing their expertise and building and maintaining the sensor infrastructure. To hystreet.com GmbH for providing sensor data that was invaluable to our data quality assessment. CrowdAnym was funded by the innovation initiative "mFUND" of the BMDV (Federal Ministry of Digital Affairs and Transport).

REFERENCES

[1] Johanna Ansohn McDougall, Christian Burkert, Daniel Demmler, Monina Schwarz, Vincent Hubbe, and Hannes Federrath. 2022. Probing for Passwords – Privacy Implications of SSIDs in Probe Requests. In *Applied Cryptography and Network Security (Lecture Notes in Computer Science)*, Giuseppe Ateniese and Daniele Venturi (Eds.). Springer International Publishing, Cham, 376–395. https://doi.org/10.1007/978-3-031-09234-3_19

[2] H.-Jürgen Appelrath, Dennis Geesen, Marco Grawunder, Timo Michelsen, and Daniela Nicklas. 2012. Odysseus: A Highly Customizable Framework for Creating Efficient Event Stream Management Systems. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems (DEBS '12)*. Association for Computing Machinery, Berlin, Germany, 367–368. https://doi.org/10.1145/2335484.2335525

[3] Drazen Brscic, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. 2013. Person Tracking in Large Public Spaces Using 3-D Range Sensors. *IEEE Trans. Hum. Mach. Syst.* 43, 6 (2013), 522–534. https://doi.org/10.1109/THMS.2013.2283945

[4] Ellis Fenske, Dane Brown, Jeremy Martin, Travis Mayberry, Peter Ryan, and Erik Rye. 2021. Three Years Later: A Study of MAC Address Randomization In Mobile Devices And When It Succeeds. *Proceedings on Privacy Enhancing Technologies* 2021, 3 (July 2021), 164–181. https://doi.org/10.2478/popets-2021-0042

[5] Denton Gentry and Avery Pennarun. 2017. Passive Taxonomy of Wifi Clients using MLME Frame Contents. https://doi.org/10.48550/arXiv.1608.01725 arXiv:1608.01725 [cs].

[6] Carlos Andres Gomez, Laura Juliana Guerrero, and Luis Fernando Pedraza. 2022. Evolution of the Use of Random MAC Addresses in Public Wi-Fi Networks. *Journal of Engineering Science and Technology Review* 15, 3 (2022), 147–152. https://doi.org/10.25103/jestr.153.15

[7] Christian Oberli, Miguel Torres-Torriti, and Dan Landau. 2010. Performance Evaluation of UHF RFID Technologies for Real-Time Passenger Recognition in Intelligent Public Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems* 11, 3 (Sept. 2010), 748–753. https://doi.org/10.1109/TITS.2010.2048429 Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[8] Tim Rütermann, Aboubakr Benabbas, and Daniela Nicklas. 2019. Know Thy Quality: Assessment of Device Detection by WiFi Signals. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 639–644. https://doi.org/10.1109/PERCOMW.2019.8730828

[9] Rijurekha Sen, Youngki Lee, Kasthuri Jayarajah, Archan Misra, and Rajesh Krishna Balan. 2014. GruMon: fast and accurate group monitoring for heterogeneous urban spaces. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys '14)*. Association for Computing Machinery, New York, NY, USA, 46–60. https://doi.org/10.1145/2668332.2668340

[10] Chong Shang, Haizhou Ai, and Bo Bai. 2016. End-to-end crowd counting via joint learning local and global count. In *2016 IEEE International Conference on Image Processing (ICIP)*. 1215–1219. https://doi.org/10.1109/ICIP.2016.7532551 ISSN: 2381-8549.

[11] Gürkan Solmaz, Pankaj Baranwal, and Flavio Cirillo. 2022. CountMeIn: Adaptive Crowd Estimation with Wi-Fi in Smart Cities. In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 187–196. ISSN: 2474-249X.

[12] Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. 2012. Counting People in the Crowd Using a Generic Head Detector. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. 470–475. https://doi.org/10.1109/AVSS.2012.87

[13] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S. Cardoso, and Frank Piessens. 2016. Why MAC Address Randomization is not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIA CCS '16)*. Association for Computing Machinery, New York, NY, USA, 413–424. https://doi.org/10.1145/2897845.2897883

[14] Mario Vega-Barbas, Manuel Álvarez Campana, Diego Rivera, Mario Sanz, and Julio Berrocal. 2021. AFOROS: A Low-Cost Wi-Fi-Based Monitoring System for Estimating Occupancy of Public Spaces. *Sensors* 21, 11 (2021). https://doi.org/10.3390/s21113863

[15] Jens Weppner and Paul Lukowicz. 2013. Bluetooth based collaborative crowd density estimation with mobile phones. In *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 193–200. https://doi.org/10.1109/PerCom.2013.6526732

[16] Qi Zhang and Antoni B. Chan. 2019. Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs. 8297–8306. https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Wide-Area_Crowd_Counting_via_Ground-Plane_Density_Maps_and_Multi-View_Fusion_CVPR_2019_paper.html

Received 15 September 2023