UJI Probes Revisited: Deeper Dive Into the Dataset of Wi-Fi Probe Requests

Tomas Bravenec, *Member, IEEE*, Joaquín Torres-Sospedra, Michael Gould, and Tomas Fryza, *Senior Member, IEEE*

***Abstract*—This article centers on the deeper presentation of a new and publicly accessible dataset comprising Wi-Fi probe requests. Probe requests fall within the category of management frames utilized by the 802.11 (Wi-Fi) protocol. Given the ever-evolving technological landscape and the imperative need for up-to-date data, research on probe requests remains essential. In this context, we present a comprehensive dataset encompassing a one-month probe request capture conducted in a university office environment. This dataset accounts for a diverse range of scenarios, including workdays, weekends, and holidays, accumulating over 1 400 000 probe requests. Our contribution encompasses a detailed exposition of the dataset, delving into its critical facets. In addition to the raw packet capture, we furnish a detailed floor plan of the office environment, commonly referred to as a radio map, to equip dataset users with comprehensive environmental information. To safeguard user privacy, all individual user information within the dataset has been anonymized. This anonymization process rigorously balances the preservation of users' privacy with the dataset's analytical utility, rendering it nearly as informative as raw data for research purposes. Furthermore, we demonstrate a range of potential applications for this dataset, including but not limited to presence detection, expanded assessment of temporal received signal strength indicator stability, and evaluation of privacy protection measures. Apart from these, we also include temporal analysis of probe request transmission frequency and period between Wi-Fi scans as well as a peak into possibilities with pattern analysis.**

***Index Terms*—Dataset, privacy, probe requests, received signal strength indicator (RSSI), WLAN, Wi-Fi, wireless communication.**

NOMENCLATURE

AIO	All-in-one.
AoI	Area of interest.
AP	Access point.
CDF	Cumulative distribution function.
GDPR	General Data Protection Regulation.
GNSS	Global navigation satellite systems.
HE	High efficiency.
HT	High throughput.
IPS	Indoor positioning system.
IoT	Internet of Things.
LBS	Location-based services.
MAC	Media access control.
OUI	Organization unique identifier.
PNL	Preferred network list.
RP	Reference position.
RM	Radio map.
RSSI	Received signal strength indicator.
RTC	Real-time clock.
SSID	Service set identifier.
ToA	Time of arrival.
UE	User equipment.
UUID-E	Universally unique identifier-enrollee.
VHT	Very high throughput.
WPS	Wi-Fi protected setup.

Manuscript received 2 November 2023; revised 20 November 2023; accepted 20 November 2023. Date of publication 22 November 2023; date of current version 20 December 2023. This work was supported by the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska Curie Grant 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints) and Grant 101023072 (ORIENTATE: Low-cost Reliable Indoor Positioning in Smart Factories). *(Corresponding author: Tomas Bravenec.)*

Tomas Bravenec is with the Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castellón de la Plana, Spain, and also with the Department of Radio Electronics, Brno University of Technology, Brno 616 00, Czechia (e-mail: bravenec@uji.es).

Joaquín Torres-Sospedra is with the ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal (e-mail: jtorres@algoritmi.uminho.pt).

Michael Gould is with the Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castellón de la Plana, Spain, and also with Esri, Inc., Redlands, CA 92373 USA (e-mail: gould@uji.es).

Tomas Fryza is with the Department of Radio Electronics, Brno University of Technology, Brno 616 00, Czechia (e-mail: fryza@vut.cz).

Digital Object Identifier 10.1109/JISPIN.2023.3335882

I. INTRODUCTION

ALONGSIDE the growing interest in indoor positioning and indoor navigation, the importance of privacy-related questions rises as well. Unlike GNSS, where the whole world uses the same coordinate systems and UEs can calculate their own location from the received information, IPSs in most cases require cooperation between the UE and the positioning infrastructure. This is due to variations in indoor environments, resulting in each indoor space being unique. In this process, the transmission of packets by the UE then permits possible breaching of user privacy. This holds particularly true when considering Wi-Fi networks for positioning. There, vulnerability is primarily associated with the management frames, specifically the probe request frames. These frames are known for their unencrypted nature, which makes them susceptible to potential interception and exploitation by unauthorized parties

The lack of encryption has been taking the attention of researchers for more than a decade. During the years before MAC address randomization, the probe request frames were in the spotlight of researchers for the ease of mobility tracking, which was demonstrated by Musa and Eriksson [1]. During the following year, the biggest probe request dataset to date was released by Barbera et al. [2], [3]. It was accompanied by a study focusing on the possibility of revealing social relations from probe requests. The vulnerability of probe requests to tracking by adversaries has also been proven by research done by Cheng et al. [4] and Cunche et al. [5].

The creation of locally assigned MAC addresses in order to preserve the privacy of users was introduced to the mainstream public by Apple as part of the iOS 8 update [6]. Since then, researchers around the world have been focused on bypassing it. Several studies were published, starting with the work published the following year by Freudiger [7], which focused on reverse engineering the algorithm generating random MAC addresses. In 2016, a large study was published by Luzio et al. [8]. The authors used an older dataset [2], [3] from the time before MAC address randomization to analyze the demographics of people participating in political events. The goal was to exploit the data from probe requests and reveal the origin of the participants. The results of the analysis proved to match the voting reports published by the Italian officials. The same year [9] looked into the time difference between consequent probes, which provided interesting results. The authors were able to breach MAC address randomization through this attack based on the timing of packets. During the next year, a very detailed study of MAC address randomization was published by Martin et al. [10]. Just a few years later, Fenske et al. [11] followed in the footsteps of Martin et al. [10] and published a study evaluating the successes of MAC address randomization in preserving the privacy of users.

This work extends our conference work presented at the 2023 International Conference on Indoor Positioning and Indoor Navigation (IPIN) [12] by providing a comparison of existing datasets on probe requests, description of information elements, description of temporal information and timing-related statistics, and results for baselines in four application domains.

A. Probe Requests

Within the framework of the IEEE 802.11 standard [13], probe requests fall under the category of management frames. They possess unique characteristics in comparison to other frames regarding their lack of encryption, allowing any device equipped with monitoring capabilities in their wireless interface driver to intercept and examine these frames. The primary intent behind probe requests was to identify and locate Wi-Fi networks in close proximity, facilitating their detection prior to association with a network. The visualization of Wi-Fi initialization, including the placement of probe requests, is in Fig. 1. Initially, the purpose of probe requests was to find a known AP to connect to. Following the rise in popularity of LBS, additional purposes for probe requests emerged. The list of identified APs in the proximity can be used to acquire a rough estimate of UE's location. It is achieved by comparison of the nearby APs to either a private database of APs maintained by the developer of the UE's operating system [14], [15] or a crowdsourced public database [16].

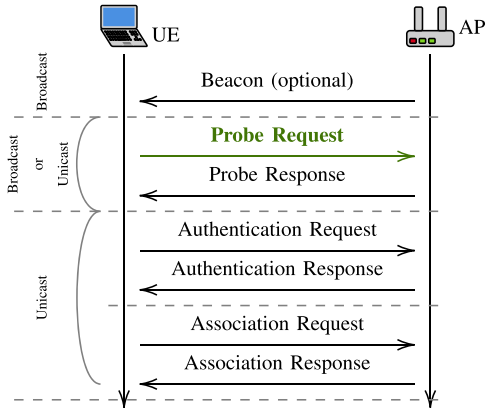

Fig. 1. Wi-Fi initialization sequence.

The structure of a probe request consists of two main sections: a header and an information element. The header, not exclusive to probe requests, contains essential details such as the frame control number (set to four for probe request frames), the MAC addresses of the destination and source devices (often with the destination address set as the broadcast address—*ff:ff:ff:ff:ff:ff*), and the frame's sequence number. When the broadcast address is used, it implies that the packet is intended for every device within the network or, in the case of wireless communications, for all devices in proximity. This is crucial to ensure that every nearby Wi-Fi AP receives the probe request.

The information element section of probe requests can carry a wealth of data. It may include the SSID, particularly when a device is actively searching for a specific network; otherwise, this field remains empty. In addition, other fields contain information about supported data transfer speeds, various capabilities of the UE, details about the manufacturer of the wireless interface, and, sometimes, even the WPS field.

Notably, the probe request frame does not include information regarding the transmission time, which means that the ToA is determined based on the RTC timer of the receiving device. Similarly, the radio information can be extracted from the wireless interface, depending on the capabilities of the capture device. Key insights include the channel on which the frame was received, the specific antenna used (if the system employs multiple antennas), and the RSSI of the captured frame. A visualization of the probe request's structure, inclusive of information collected by the receiver's wireless interface, is provided in Fig. 2.

B. Existing Probe Request Datasets

Numerous publicly accessible datasets of probe requests are readily available for examination and analysis. This compilation provides a comprehensive overview of these datasets, offering insights into their respective strengths and weaknesses. In this section, we delve into each dataset, outlining its distinctive characteristics, advantages, and disadvantages.

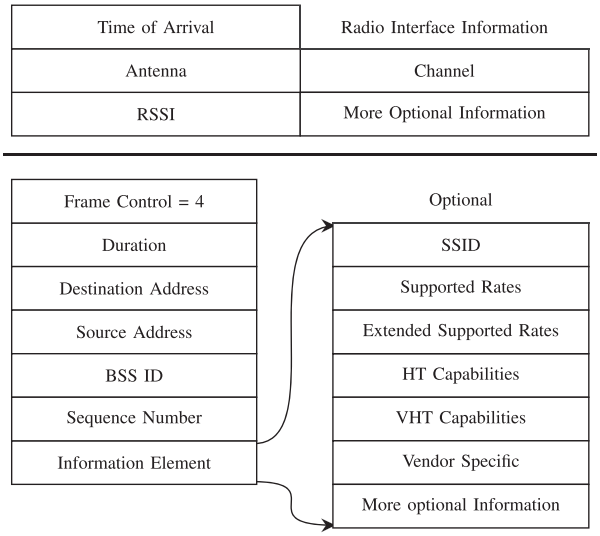

Fig. 2. Probe request structure.

1) Sapienza2013: This dataset was released in 2013 [2], [3] and is the largest dataset of probe requests to date. It contains over 11 million captured packets from varied places in Rome, including political events, shopping mall, train station, and university. The creators of the dataset also thought about the user privacy, and the possibilities of analysis as the SSIDs of directed probe requests were replaced by pseudonyms. The dataset was, however, captured before iOS 8 and other operating systems implemented MAC randomization, which unfortunately results in data that do not reflect the real world anymore.

2) Glimps2015: This dataset was collected in Ghent, Belgium, throughout the music festival called Glimps 2015 [17]. The dataset itself proves to be a challenge to analyze. Starting with the aggressive anonymization, which replaces all SSIDs with *Hidden* and sets all sequence numbers to 0. The rest of the fields in the information element section use hashes. The anonymization technique is not the only issue though, and the capture style provides even more challenges for analysis. The dataset contains only one probe request for each unique MAC address, which, combined with the hidden SSIDs, prevents researchers from analyzing the dataset through temporal means and from exploiting the PNL, respectively.

3) Nile2021: This dataset is one of the smallest datasets available [18]. The capture took place during nighttime inside of a shopping center. It is also important to note that the dataset does not incorporate any anonymization techniques. The point that stands out against most of the other available datasets is that it does not incorporate any anonymization of data. However, it is important to note that this dataset's data capture duration is rather limited, spanning just 40 min. This on its own removes any possibility of analyzing mobility patterns.

4) IPIN2021: It is a dataset that we collected throughout four days of an international conference IPIN 2021 [19], [20]. The conference took place in Lloret de Mar, Spain, in Evenia Olympic Congress Center. However, at that time, the capabilities of our capture device were limited. The main thing that was missing in the firmware was the ability to store the RSSI and other radio information alongside the captured probe requests. On the other hand, the anonymization was done in a way to preserve the ability to analyze the dataset, by employing hashing algorithm SHA512 over all sensitive fields to protect the privacy of users.

5) Pintor 2021: It is the one dataset that differs from the rest [21], [22]. The main point of difference is the inclusion of ground truth labels. The authors of the dataset achieved this by capturing the probe request frames inside an anechoic chamber in order to suppress signals from the environment. However, the dataset is small, with a capture of a single device taking only 20 min and, due to the synthetic nature of this dataset, it is unsuitable for pattern analysis.

To summarize the differences between existing datasets, none of them are perfect, be it due to the age and the lack of resemblance to current reality, short time period of data capture, or because of an aggressive implementation of anonymization that prevents any chance of analysis. To better visualize the differences, we present the most important information about each dataset in comparison to ours in Table I.

The motivation behind this work was to provide the research community with a new dataset without the shortcomings of the previous datasets. The dataset presented in this work covers a time period of an entire month while employing hashing as the anonymization technique of choice. Therefore, user privacy is preserved alongside the analysis possibilities. This includes possible pattern analysis, as the dataset presents data captured in a workplace, with the occupants being mostly the same.

C. Outline

The rest of this article is organized as follows. Section II explains the creation of the dataset, the hardware used, and the format in which the dataset is stored. It also describes all of the aspects of the dataset starting with MAC addresses and SSIDs, which is followed up by the evaluation of the information element and both radio and temporal information. The possible use case examples ranging from presence detection, through signal stability, to privacy evaluation are available in Section III; we have also included in this section a look into possibilities of pattern analysis. Section IV contains the discussion on ethics of dealing with real-life data, which is then followed up by Section V with the data availability statement. Finally, Section VI concludes this article.

II. Dataset

For the duration of the entire month of March 2023, every probe request transmitted in our office at University Jaume I, Spain, and in its close proximity was captured. The capture was done using an ESP32-based sniffer created for our previous works [20], [24], [25].

The sniffer functions by capturing raw packets and saving them into a standardized packet capture file on a micro SD card. These files are designed to be seamlessly compatible with various network analysis tools, including but not limited to Wireshark and Python packages like Scapy.

TABLE I
COMPARISON OF EXISTING DATASETS

	Sapienza	Glimps	Nile	IPIN 2021	Pintor 2021	**UJI Probes**
Year	2013	2015	2021	2021	2021	**2023**
Time period	February to May	3 days	40 minutes	4 days	20 minutes	**1 month**
Number of packets	11 136 711	122 989	81 635	390 810	69 701	**1 410 834**
Location	Varied in Rome	Music festival	Shopping centre	Conference venue	Anechoic chamber	**University office**
Anonymized	✓	✓	✗	✓	✗	✓
Anonymization	SSIDs use pseudonyms	SSIDs are hidden, Sequence numbers = 0	✗	Using SHA512 on every field	✗	**Using SHA512 on every field**
Labeled	✗	✗	✗	✗	✓	✗
Available	[2], [3]	[17]	[18]	[19], [20]	[21], [22]	[23]
Notes	From before MAC address randomization. Nowadays the data is unrealistic.	Deep analysis impossible due to aggressive anonymization technique.	Very short, not useful for pattern analysis.	Not possible to analyse patterns, every conference day was different. No RSSI information.	Very short, not continuous. Not usable for pattern analysis. Only labeled dataset.	**Workdays, one week of holidays, summertime change, all fields (hashed).**

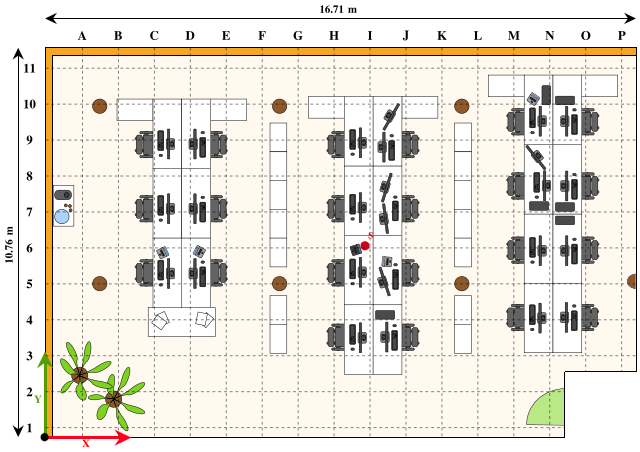

Fig. 3. Floor plan and location of sniffer in the office space of GEOTEC Department, Universitat Jaume I, Spain.

It is important to note that the radio-specific data are not inherently included in the official 802.11 frames. To address this, we have implemented the practice of saving Radiotap headers [26] in conjunction with the 802.11 frames. This approach ensures that our packet capture files are fully compatible with common network analysis tools while containing additional radio-related information.

A. Data Description

The dataset was collected within the GEOTEC group offices at University Jaume I, Spain. The office space itself is rectangular measuring approximately 16.71 m in length and 10.76 m in width. It is designed as an open-space office, but it is further divided into three distinct sections by bookshelves, with each section capable of comfortably housing between six and eight people. For reference, the office's floor plan is in Fig. 3.

The creation of the dataset took place throughout the month of March 2023. The reason we chose this time of the year is to capture both regular work weeks, weekends, and special events. A local holiday week, Magdalena 2023, also took place during March, which we captured in addition to the ordinary days at the office. During Magdalena holidays, starting on Saturday, March 11, 2023 and ending a week later on Sunday, March 19, 2023, the university was mostly closed. To visualize it and clearly highlight the difference from the normal work weeks, we visualized the Magdalena holiday week in Fig. 4, where it uses a light green color.

In a similar way to the holiday week, the dataset presents even more trends. One of these is the constant flow of probe requests. From Fig. 4, it can be seen that the probe requests are transmitted regardless of the time and day. The only time there are no probe requests at all is at night of Sunday, March 26, 2023. This is caused by the change of Central European Time to Central European Summer Time when the time changed from 02:00 to 03:00. The probe requests captured during the nighttime hours may originate from static devices in the office like AIO computers connected through Wi-Fi instead of a wired connection, IoT devices, or mobile phones left in the office for experiments. The next clearly noticeable trend is a peak in captured probe requests in the morning hours. This peak happens every day at 05:00 in the morning (or at 06:00 after the change to the Summer Time), which means that it happens regardless of the current time zone. The cause is a scheduled reboot of the Wi-Fi AP in the office. The restart of the AP causes all connected devices to disconnect and start actively looking for a new network to connect to, resulting in a peak in transmitted probe request frames.

During the week of the Magdalena holiday and during the first weekend of March, it can also be observed from the rise in collected probe requests that few of our colleagues went to the office. These visits to the office explain the noticeable increase in captured probes.

B. MAC Addresses and SSIDs

A crucial factor when it comes to the analysis of Wi-Fi probe requests are their MAC addresses. Since many devices are still not using MAC address randomization, tracking such UEs becomes pretty straightforward. Even the identification of a device using randomized address is simple, due to the characteristics of the address's first byte, particularly the second least significant

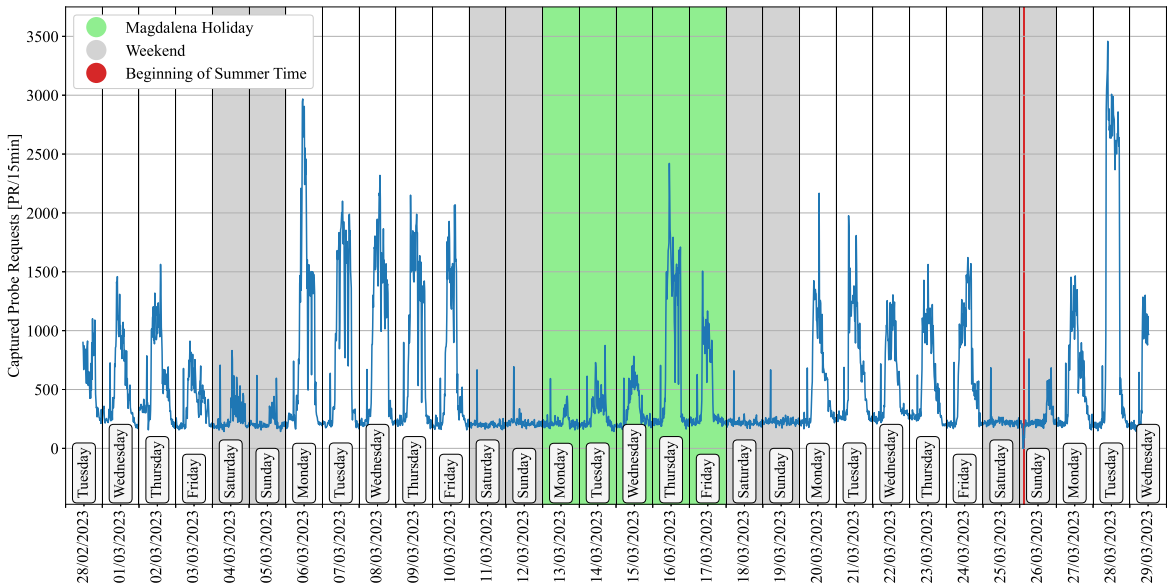

Fig. 4. Density of captured probe requests over the course of capture (amount of probe requests grouped in 15-min clusters).

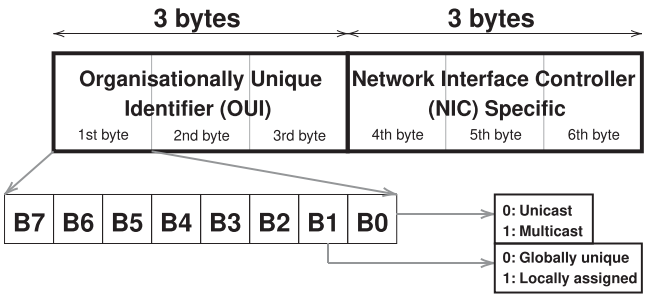

Fig. 5. Structure of MAC address with the functional bits.

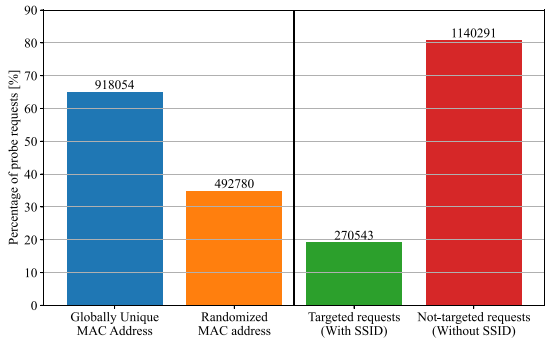

Fig. 6. Split between randomized and globally unique MAC addresses in the captured probe requests and split between targeted and not targeted probe requests.

bit, denoted as **B1**. When this bit is set, it is an indication that the address was randomized by the network controller of the UE in question. In addition, the least significant bit of the first byte, **B0**, serves to distinguish individual devices and device groups. The structure of the MAC address is visually depicted in Fig. 5.

Taking into account the constraints posed by these two least significant bits, the second digit of locally assigned MAC addresses in hexadecimal format has only four possible values: 2 (0010), 6 (0110), A (1010), or E (1110). Within the collected probe requests, it was observed that approximately 35% of them utilized randomized MAC addresses, as illustrated in Fig. 6. Throughout all probe requests, there are in total 129 330 unique MAC addresses, out of which 2437 did not use randomization of the MAC address.

Another critical parameter is the PNL, which comprises a list of network SSIDs to which the UE frequently connects. This information can potentially be leaked when a device sends probe requests specifically targeted at particular networks. In the dataset, 19% of the captured probe requests include an SSID, as shown in Fig. 6. Although this percentage might initially seem relatively low, it is important to note that the dataset encompasses a total of 2030 unique SSIDs.

C. Information Elements

This optional section within probe requests proves to be invaluable for fingerprinting purposes. It provides a stable set of information about the capabilities and supported functions that remain consistent over time for a given UE. The range of information encompasses various aspects, including supported data rates, capabilities associated with specific Wi-Fi standards (such as HT capabilities for 802.11n, VHT capabilities for 802.11ac, and HE capabilities for 802.11ax), vendor-specific elements (with some devices featuring multiple), and WPS fields.

The WPS fields, in particular, can be quite revealing. They often contain a wealth of user-identifying information, extending from device manufacturer details to the device's name. Notably, many users personalize their device names, such as "Julia's iPhone," inadvertently disclosing their own names. Moreover, WPS fields also carry UUID-E (universally unique identifier), serving as a persistent and unchanging identifier that can compromise anonymity, even when randomized MAC addresses are used.

TABLE II
PROBE REQUEST FIELDS USED TO CREATE DEVICE FINGERPRINT AND FREQUENCY OF OCCURRENCE IN DATA COLLECTED IN OUR LABORATORY

Information Element	Included in Probes	[%]
Supported rates	1 410 832	100.00
Extended Supported rates	1 405 288	99.61
HT Capabilities	1 093 575	77.51
HE Capabilities	394 347	27.95
VHT Capabilities	262 365	18.60
Extended Capabilities	1 182 014	83.78
Vendor - Specific elements	315 815	22.39
1 Vendor-Specific element	135 973	9.64
2 Vendor-Specific elements	110 071	7.80
3 Vendor-Specific elements	9607	0.68
4 Vendor-Specific elements	257	0.02
5+ Vendor-Specific elements	146	0.01
WPS - UUID-E	16 596	1.18
WEP Protected	2	0.00
Total Collected Probe Requests	1 410 834	

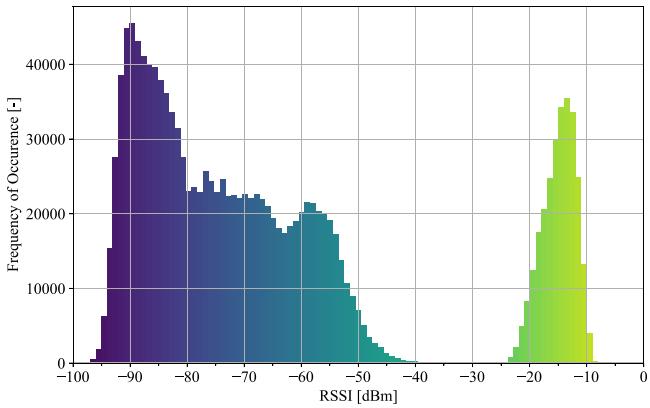

Fig. 7. Frequency of occurrence of RSSI in the captured probe requests.

A detailed breakdown of the occurrences of each field throughout the dataset is present in Table II.

D. Radio Information

The ESP32 sniffer was set up to collect probe requests via an all-channel scan. However, it is important to note that the ESP32 lacks support for the 5-GHz band of Wi-Fi. Consequently, it only captured probe requests transmitted in the 2.4-GHz band.

The distribution of RSSI values is represented as a histogram in Fig. 7, which reveals a clear pattern: the majority of frames exhibit RSSI readings falling within the range of −100 to −40 dBm. Interestingly, there is a noteworthy observation to be made. Quite surprisingly, a substantial number of probe requests display RSSI values higher than −30 dBm. It is worth mentioning that every one of these probes originated from a single device. The cause for this abnormally high RSSI is the placement of the ESP32-based packet sniffer right next to an AIO desktop computer (the MAC address of the AIO in the anonymized dataset is *48:5f:99:07:10:21*). The close proximity between the AIO and the sniffer resulted in much higher RSSI than would be normal otherwise. Another noteworthy aspect to consider is the significant volume of probe requests captured with RSSI levels

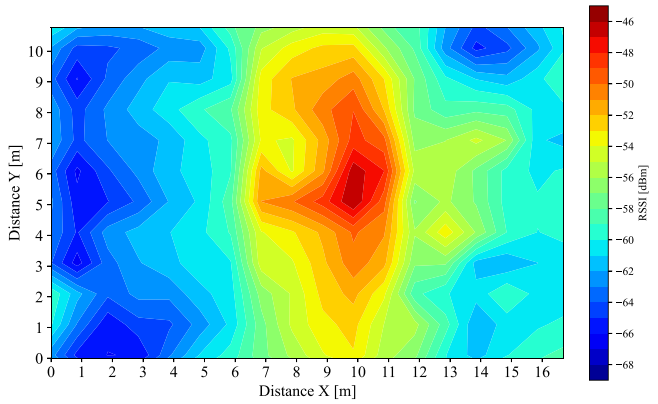

Fig. 8. RM of the RSSI in different locations of the office mapped with ESP32 microcontrollers.

hovering around −90 dBm. It is highly probable that these probe requests originate from devices in neighboring offices.

We have also mapped the radio signal propagation throughout our office for another study [27]. For this purpose, we used another ESP32 with firmware modified to transmit probe requests. We walked around the office while stopping on RPs placed throughout the AoI. In total, we gathered 50 probe requests at 143 RPs. The dataset with the RM data is available to download from Zenodo repository [28]. The RM of the office mapped using another ESP32 microcontroller is in Fig. 8. The captured RM can be used as a rough distance filter, as it can be used for the selection of RSSI threshold.

E. Temporal Information

Apart from radio information, the dataset can be also described through temporal means. As presented in [9], the time between subsequent probe requests depends on the wireless interface of each UE, and it can be used to defeat MAC address randomization.

There are several timing-related statistics of the dataset that are worth mentioning. To begin with, as a single Wi-Fi scan, we considered probe requests with time difference between two consecutive receivals shorter than 1 s [29]. In case the time between two probes is higher, it is considered as a new scan instance. The shortest time between two subsequent probe requests from the same device during a single scan instance is just 67 μs, while the median is 19.87 ms. The 95th percentile of the time between subsequent probe requests is 242.19 ms, which means 95% of all consecutive probe requests in one scan instance arrive less than a quarter of a second apart. The distribution of time difference between receiving two probes from the same device in the same Wi-Fi scan instance is presented through the CDF in Fig. 9, highlighting some key metrics.

A different temporal statistic of interest is the time period between Wi-Fi scans from a single device. This we illustrate by utilizing CDF, as depicted in Fig. 10. It is important to note that the CDF is limited to 2 h between consecutive scans from the same device. The reason for that is to preserve the clarity of the plot because the maximum time between two identified scans from the device using the same MAC address is almost 28 days.

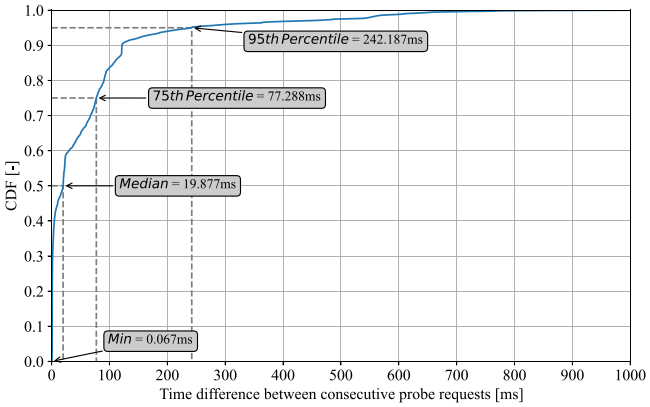

Fig. 9. CDF plot of time differences between consecutive arrivals of probe requests from the same device during a single scan.

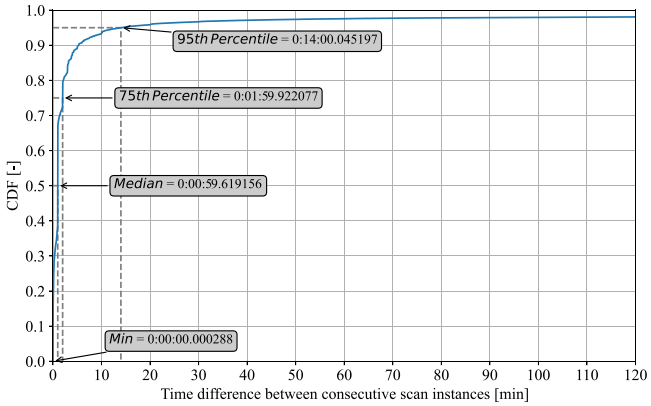

Fig. 10. CDF plot of time differences between scans of probe requests from the same device. (A time difference larger than 2 h is considered as one of the users leaving the office with the device, or turning the device OFF.)

With the 95th percentile being 14 min between scan instances, we assume that in case this time period exceeds 2 h, the person has left the office during this time. There are two statistics that stand out from the CDF plot; the median is approximately 1 min and the 75th percentile is 2 min, which tells us that for majority of devices, the interval between scans for nearby Wi-Fi networks is shorter than 2 min. The minimum interval we detected is 288 μs; however, this is from a single scan instance that has been interrupted by an incoming probe request transmitted by a different device.

III. Baselines for Applications

The dataset offers a multitude of possible use cases. Within this section, we provide a few possibilities to showcase the potential use. In addition to the presented use cases, we believe that the dataset has many more potential use cases, including analysis through machine learning.

A. Wi-Fi Signal Stability Evaluation

The inclusion of RSSI values in the capture of each probe request opens up an opportunity to examine the signal strength

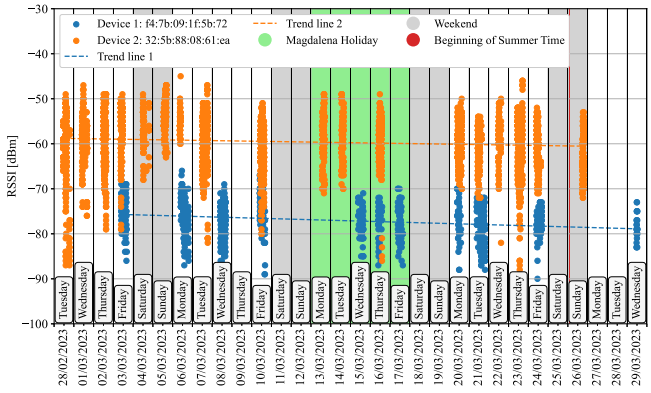

Fig. 11. Example of RSSI long-term stability evaluation for one MAC address from the dataset for two devices, both with and without MAC address randomization. (Device 1: *f4:7b:09:1f:5b:72*, device 2: *32:5b:88:08:61:ea*) including trend lines.

aspect, especially when assuming that individuals primarily remain stationed at their desks. In this illustrative example, we investigate signal stability over a span of time. As an example, we have randomly chosen two MAC addresses that make repeated appearances in the dataset. The MAC addresses for these selected devices in the anonymized dataset are a globally unique address *f4:7b:09:1f:5b:72* and a locally assigned address *32:5b:88:08:61:ea*. We have then generated a visual representation of the RSSI values for all probe requests across the temporal dimension.

Through this visual depiction, we can observe evident changes in the signal strength over time. To better highlight these changes, we have introduced trend lines in Fig. 11, providing a clear illustration of the declining RSSI values as the data capture progresses. This can shed light on how signal strength evolves over time for specific devices.

B. Presence Detection and Room Occupancy Estimation

The network traffic data can also serve as a valuable resource for presence detection. This can be observed in Fig. 4, which offers insights into the fluctuations in activity around the sniffer's vicinity. Notably, it enables us to identify periods of increased activity, as well as discern the declines occurring on weekends and holidays.

In addition, the RM in Fig. 8 can be employed to provide a rough estimate of the distance between the UE and the sniffer. The RM was created by collecting RSSI data from probe requests throughout the office at a 1-m grid resolution. This information can prove invaluable for presence detection and room occupancy assessment. On a smaller scale, we have employed similar methodologies to estimate room occupancy on the university campus, as detailed in [25].

Regrettably, capturing the ground truth of room occupancy within our office presented significant challenges. With a dynamic environment featuring up to 30 individuals frequently entering and leaving during office hours, gathering precise ground truth data was infeasible with our existing resources. However, it is plausible to categorize occupancy in general terms, such

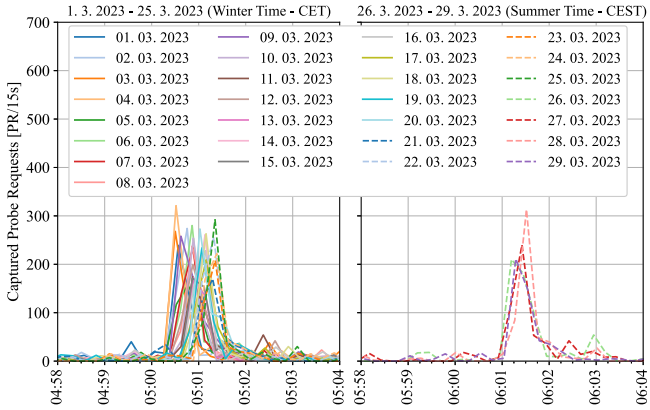

Fig. 12. Recognized pattern of daily recurrence of a peak in received probe requests.

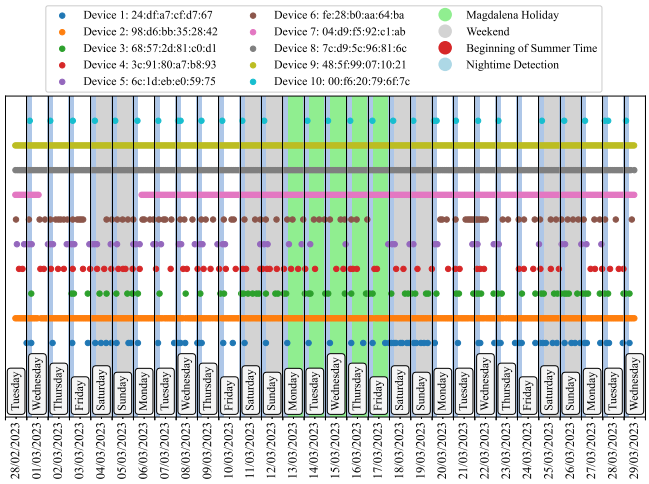

Fig. 13. Devices exhibiting a recurring pattern of probe request transmission during 50% of nights between 23:00 and 05:00.

as *empty, low, medium*, and *high*. Such a classification system aligns with previous occupancy estimation work, as evidenced by Ciftler et al. [30], who used a scale ranging from *low* to *high* to assess occupancy levels.

C. User Privacy Exploitation

The anonymization process applied to this dataset was selected with preserving the ability to analyze it in mind. Notably, no data removal occurred during this process. Instead, an anonymization technique utilizing hashing was employed. This approach offers distinct advantages, as it allows for an in-depth analysis of potential privacy leaks. The absence of data removal ensures that the dataset retains its integrity, while the use of pseudonyms via hashing enhances privacy protection.

Furthermore, the ToA of packets within the dataset remains entirely unaltered. This aspect is particularly valuable, as it opens possibilities for temporal analysis similar to the one done by [9].

The dataset is also ideal for analysis of the recurring appearance of randomized MAC addresses and as such identification of users despite the implemented MAC randomization. Moreover, the dataset is a valuable resource for uncovering potential vulnerabilities within the probe request mechanism.

D. Pattern Analysis

The length of the dataset also enables the analysis of repeating patterns in the collected probe requests. As examples, we are going to present an exploration of repeating peaks in received probe requests as well as an analysis of nighttime.

1) Peak Analysis: In the dataset, there is a sudden increase in the number of transmitted probe requests that happens every day at 05:00 in the morning (or at 06:00 following the switch to the Summer Time at night of the Sunday, March 26, 2023).

To gain a deeper understanding of the recurrent peak, we created a plot that displays the number of probe requests grouped in 15-s intervals within a time window spanning ± 3 min around both 05:00 and 06:00. This visualization is presented in Fig. 12. From the figure, it is clearly visible that the higher frequency of transmitted probe requests happens every single day at exactly the same time and takes less than a minute to pass.

Notably, during this recurring activity, we noticed that out of the total of 2030 SSIDs that can be found in the dataset, the targeted probe requests sent during the morning peaks were in total pointed toward only ten different APs.

Similar to the SSIDs, the probe requests captured at this time came from at most 102 devices. Out of these, only 33 were employing MAC address randomization. The rest 69 MAC addresses were not randomized, indicating those were distinct devices.

2) Nighttime Analysis: Another noticeable pattern from Fig. 4 is the consistent transmission of probe requests during nighttime. These devices send out probe requests regardless of the office's occupancy.

To investigate this further, we focused on a nighttime interval when the presence of people in the office was unlikely, spanning from 23:00 at night to 05:00 in the morning (as supported by Fig. 4). During this time frame, probe requests were consistently sent by devices with a limited set of MAC addresses. We applied a filtering process to the probe requests collected during these nighttime intervals, retaining only those transmitted by devices that appeared during 50% of the nights within the dataset.

This selection resulted in just ten MAC addresses meeting these criteria. In Fig. 13, we visualized the capture periods for each of these addresses. From this figure, we can infer various patterns. For instance, Devices 1, 3, and 4 remained continuously active, never turning OFF, as they transmitted probe requests throughout the entire data capture period. Device 2, on the other hand, was turned OFF for four days at the beginning of the month. Notably, Devices 7 and 8 exhibit distinct patterns, with Device 7 appearing to be active primarily during the nighttime, while Device 8 sends probe requests predominantly in the evenings and at night. The remaining four devices, although active both during the day and night, lack a consistent pattern like Devices 1, 3, and 4, as their activity is not continuous.

Moreover, it is crucial to explore and analyze other patterns within the dataset. Discovering these patterns can offer valuable

insights into the behavior of Wi-Fi-enabled devices and the mobility of users.

IV. Ethics and Sensitive Information

The probe requests captured by the ESP32-based sniffer are stored in a packet capture file, which follows the same structure as packet capture files created by network analysis tools like Wireshark. The packet capture files created by the sniffer contain the packets in the same way in which they were transmitted. That includes all sensitive or private information. The level of information can vary significantly on the device; it can contain only the globally unique MAC address, network names from the PNL, manufacturer of the network interface, or the device itself; and, in some cases, it is possible to find a device name as well.

Following the data capture, we ran an anonymization script to hide any user information in order to preserve user privacy in accordance with GDPR. For this purpose, a selection of bytes from the SHA512 hashes was used. The use of the selection of bytes from hashes reduces the size of the dataset and memory requirements for further analysis while protecting sensitive data even further. That is because the use of partial hashes reduces the chance of finding the original values of fields by using a reverse hash lookup. The hashing of fields is employed to maintain the data's analyzability, allowing for analysis in a manner similar to data that have not undergone anonymization.

To preserve information related to both randomized and globally unique MAC addresses, only the last 3 B of MAC addresses are subjected to hashing. This strategy preserves the least significant bits of the first byte of the MAC address, as well as the OUI [31]. Importantly, anonymization through hashing ensures that it is impossible to establish a direct link between the actual identities of individuals and the captured probe requests.

V. Data Availability

To ensure reproducibility, all of the variations of the packet sniffer firmware for the ESP32 are available to the public from GitLab repository [32] under the Public Domain license. The firmware is targeted at the ESP32-CAM variant written in C using the ESP-IDF, but is suitable for other ESP32 board variants as well; however, in such cases, modifications to ensure correct connections of the button, LED, and the micro SD slot might be necessary.

The anonymized version of the created dataset in the standardized packet capture format is readily accessible from the Zenodo repository [23].

VI. Conclusion

This article primarily focused on the review of existing probe request datasets as well as an introduction of a new Wi-Fi dataset. The new dataset overcame a majority of the shortcomings affecting the existing datasets. This included that the implementation of aggressive anonymization, noncontinuous data capture, or the length of the capture was insufficient for the purposes of pattern analysis. In addition to the dataset, we also published the firmware for the ESP32 microcontroller, which allows people to create more datasets, be it for personal use or for release to the public.

Throughout this article, we introduced the methodology for the capture of probe requests as well as went over the important features of the dataset. Apart from describing the dataset, we also presented some of the applications, such as detection presence detection, estimation of room occupancy, evaluating security and possibilities of privacy breaches, and the stability of Wi-Fi over time, and a brief look into the pattern analysis, which is all possible with the dataset.

In future research, we will continue our focus on indoor positioning applications. We will further explore the possibilities of noncooperative tracking of users as well as other privacy-related issues and possibilities inside the IEEE 802.11 communication protocol. We plan to achieve this by employing several ESP32-based probe request sniffers working together.

Acknowledgment

This work does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

References

[1] A. Musa and J. Eriksson, "Tracking unmodified smartphones using Wi-Fi monitors," in *Proc. 10th ACM Conf. Embedded Netw. Sens. Syst.*, 2012, pp. 281–294.

[2] M. V. Barbera, A. Epasto, A. Mei, V. C. Perta, and J. Stefa, "Signals from the crowd: Uncovering social relationships through smartphone probes," in *Proc. Conf. Internet Meas. Conf.*, 2013, pp. 265–276.

[3] M. V. Barbera, A. Epasto, A. Mei, S. Kosta, V. C. Perta, and J. Stefa, "CRAWDAD sapienza/probe-requests," 2013, doi: 10.15783/C76C7Z.

[4] N. Cheng, X. O. Wang, W. Cheng, P. Mohapatra, and A. Seneviratne, "Characterizing privacy leakage of public WiFi networks for users on travel," in *Proc. IEEE INFOCOM*, 2013, pp. 2769–2777.

[5] M. Cunche, M.-A. Kaafar, and R. Boreli, "Linking wireless devices using information contained in Wi-Fi probe requests," *Pervasive Mobile Comput.*, vol. 11, pp. 56–69, 2014.

[6] L. Hutchinson, "iOS 8 to stymie trackers and marketers with MAC address randomization," *Ars Technica*, 2014. [Online]. Available: https://arstechnica.com/gadgets/2014/06/ios8-to-stymie-trackers-and-marketers-with-mac-address-randomization/

[7] J. Freudiger, "How talkative is your mobile device? An experimental study of Wi-Fi probe requests," in *Proc. 8th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, 2015, pp. 1–6.

[8] A. D. Luzio, A. Mei, and J. Stefa, "Mind your probes: De-anonymization of large crowds through smartphone WiFi probe requests," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[9] C. Matte, M. Cunche, F. Rousseau, and M. Vanhoef, "Defeating MAC address randomization through timing attacks," in *Proc. 9th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, 2016, pp. 15–20.

[10] J. Martin et al., "A study of MAC address randomization in mobile devices and when it fails," *Proc. Privacy Enhancing Technol.*, vol. 2017, no. 4, pp. 365–383, 2017.

[11] E. Fenske, D. Brown, J. Martin, T. Mayberry, P. Ryan, and E. Rye, "Three years later: A study of MAC address randomization in mobile devices and when it succeeds," *Proc. Privacy Enhancing Technol.*, vol. 2021, no. 3, pp. 164–181, 2021.

[12] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Fryza, "UJI probes: Dataset of Wi-Fi probe requests," in *Proc. 13th Int. Conf. Indoor Position. Indoor Navigat.*, 2023, pp. 1–6.

[13] *IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks—Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 5: Preassociation Discovery*, IEEE Standard 802.11aq-2018, 2018.

[14] Google, "Google developers: Geolocation API," Google, 2022. [Online]. Available: https://developers.google.com/maps/documentation/geolocation/
[15] Google, "Google support: Control access point inclusion in Google's location services," 2022. [Online]. Available: https://support.google.com/maps/answer/1725632?hl=en
[16] WiGLE, "Wigle: Wireless network mapping," 2023. [Online]. Available: https://wigle.net/
[17] P. Robyns, B. Bonné, P. Quax, and W. Lamotte, "CRAWDAD hasselt/glimps2015," 2015, doi: 10.15783/c7sd19.
[18] B. Abdulrahem, "CRAWDAD nile/probe-requests," 2021, doi: 10.15783/s9mm-z673.
[19] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Frýza, "Supplementary materials for "what your wearable devices revealed about you and possibilities of non-cooperative 802.11 presence detection during your last IPIN visit"," 2022, doi: 10.5281/zenodo.6798302.
[20] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Fryza, "What your wearable devices revealed about you and possibilities of non-cooperative 802.11 presence detection during your last IPIN visit," in *Proc. IEEE 12th Int. Conf. Indoor Position. Indoor Navigat.*, 2022, pp. 1–7.
[21] L. Pintor and L. Atzori, "A dataset of labelled device Wi-Fi probe requests for MAC address de-randomization," *Comput. Netw.*, vol. 205, 2022, Art. no. 108783.
[22] L. Pintor and L. Atzori, "A dataset of labelled device Wi-Fi probe requests for MAC address de-randomization," 2021, doi: 10.17632/j64btzdsdy.1.
[23] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Frýza, "Supplementary materials for "UJI probes: Dataset of Wi-Fi probe requests"," 2022, doi: 10.5281/zenodo.7801798.
[24] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Fryza, "Exploration of user privacy in 802.11 probe requests with MAC address randomization using temporal pattern analysis," 2022, *arXiv:2206.10927*.
[25] T. Fryza, T. Bravenec, and Z. Kohl, "Security and reliability of room occupancy detection using probe requests in smart buildings," in *Proc. 33rd Int. Conf. Radioelektronika*, 2023, pp. 1–6.
[26] "Radiotap," Dec. 2023. [Online]. Available: http://www.radiotap.org/
[27] T. Bravenec, M. Gould, T. Fryza, and J. Torres-Sospedra, "Influence of measured radio map interpolation on indoor positioning algorithms," *IEEE Sens. J.*, vol. 23, no. 17, pp. 20044–20054, Sep. 2023.
[28] T. Bravenec, J. Torres-Sospedra, M. Gould, and T. Frýza, "Supplementary materials for "influence of measured radio environment map interpolation on indoor positioning algorithms"," 2022, doi: 10.5281/zenodo.7193602.
[29] D. Jaisinghani, "Understanding the role of active scans for their better utilization in large-scale WiFi networks," Ph.D. dissertation, Dept. Comput. Sci. Eng., Indraprastha Inst. Inf. Technol. Delhi, New Delhi, India, 2019.
[30] B. S. Ciftler, S. Dikmese, I. Güvenç, K. Akkaya, and A. Kadri, "Occupancy counting with burst and intermittent signals in smart buildings," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 724–735, Apr. 2018.
[31] IEEE, "IEEE Standards-OUI/MA-L," Jun. 2023. [Online]. Available: https://standards-oui.ieee.org/
[32] T. Bravenec, "ESP32 probe sniffer," 2022. [Online]. Available: https://gitlab.com/tbravenec/esp32-probe-sniffer

Tomas Bravenec (Member, IEEE) received the M.S. degree in electronics and communications from the Brno University of Technology, Brno, Czechia, in 2019. He is currently working toward the joint Ph.D. degree in computer science with University Jaume I, Castellón de la Plana, Spain, and the Brno University of Technology, Czechia.

He is working on the A-WEAR project as an Early Stage Researcher. His research interests include machine learning, indoor localization, and privacy and security issues related to wearable applications.

Joaquín Torres-Sospedra received the Ph.D. degree in computer science from Universitat Jaume I, Castellón de la Plana, Spain, in 2011.

He is currently an MSCA Postdoctoral Fellow with the University of Minho, Guimarães, Portugal, where he works on indoor positioning (Wi-Fi, Bluetooth low energy, visual light communication, and machine learning) for industrial applications. He has authored more than 170 articles in journals and conferences and has supervised 21 master's and five Ph.D. students. Currently, he is supervising five Ph.D. students.

Dr. Torres-Sospedra is the Chair of the Indoor Positioning and Indoor Navigation (IPIN) International Standards Committee and IPIN Smartphone-Based Off-Site Competition.

Michael Gould received the Ph.D. degree in geographic information systems from the University at Buffalo—State University of New York, Buffalo, NY, USA.

He is currently an Associate Professor of Information Systems with Universitat Jaume I, Castellón de la Plana, Spain. He has been Principal Investigator and Researcher on several European and global projects. Since 2009, he has been a Global Education Manager with the software company Esri, Inc., Redlands, CA, USA, where he works on capacity development projects around the world. His research and teaching focuses on geospatial information, smart cities, and spatial data infrastructures.

Dr. Gould was the Chair of the Association of Geographic Information Laboratories in Europe.

Tomas Fryza (Senior Member, IEEE) received the Ph.D. degree in electronics and communication technologies from the Brno University of Technology, Brno, Czechia, in 2006.

Since 2010, he has been an Associate Professor with the Department of Radio Electronics, Brno University of Technology, where he was a Vice Head of the Department of Radio Electronics from 2013 to 2021. He visited the Simula Research Laboratory, Oslo, Norway; CSC—IT Center for Science, Espoo, Finland; and Umeå University, Umeå, Sweden, in 2012, 2018, and 2022, respectively. His research interests include the development of optimized codes for embedded systems, intelligent data analysis, machine learning, and signal processing.

Dr. Fryza was an Executive Editor for *Radioengineering Journal*.