PILOT Dataset: A Collection of Multi-Communication Technologies in Different Mobility Contexts

Jana Koteich, Nathalie Mitton

PILOT Dataset: A Collection of Multi-Communication Technologies in Different Mobility Contexts

Jana Koteich[1] et Nathalie Mitton[1]

[1]*Inria, 40 Avenue Halley, 59650 Villeneuve-d'Ascq, France*

L'omniprésence des appareils mobiles équipés de technologies de communication radio a rendu courante la collecte de données à grande échelle sur la mobilité humaine [SSGL15]. Par conséquent, avec le manque d'ensembles de données, la collecte et la publication de données s'avèrent lever de nouvelles questions scientifiques. Dans cet article, nous présentons l'ensemble de données PILOT et sa méthodologie de collecte de données préservant la confidentialité des technologies de communication sans fil. L'ensemble de données est une collection de quatre éléments d'information collectés conjointement dans différents contextes de mobilité. Il comprend trois technologies de communication sans fil : balises WiFi, balises BLE (Bluetooth Low Energy) et paquets LoRa (Long Range Radio), ainsi que des informations supplémentaires qui comprennent : Accélération, Roulis et Tangage, toutes collectées en même temps. Nous fournissons les clés pour reproduire cette collecte de données. Le jeu de données est collecté pendant environ 60 heures, avec une taille de 150 Mo en utilisant les dispositifs FiPy de Pycom et il est téléchargé sur GitHub. Nous pensons que la combinaison des différents types d'informations collectées dans différents scénarios de mobilité constitue une nouvelle génération d'un riche ensemble de données qui offre une nouvelle perspective pour analyser le réseau et obtenir de nouvelles informations.

Mots-clefs : Dataset generation, WiFi, BLE, LoRa, Acceleration, IoT, FiPy

1 Introduction

These days, the rapidly evolving information technology and the development of wireless and mobile networks have promoted a new wave of information and industrial tide [GZPZ18]. Several promising applications like tracking smartphones, traffic monitoring, crowd dynamics monitoring, and other scientific research are based on the gathering and analysis of measurement data. This increases the urge of creating new forms of datasets that provide a new perspective to analyze data and bring out new measurements. The collected datasets are mainly characterized by the *Model* and *Parameters* recorded. Each type of dataset can serve a new form of analysis, that is why there are always newly generated datasets with different characteristics.

To this end, we noticed a lack of a collective dataset that includes several traces from wireless communication technologies and sensors recorded at the same time in different mobility contexts. Such dataset provides a new generation of collecting data that could help in studying human mobility. So, in this paper, we introduce a new approach to a collected dataset that is characterized by mainly two novel approaches: 1) Collecting different types of data from sensors (Acceleration, Roll and Pitch) and wireless communication technologies (WiFi beacons, BLE beacons, LoRa packets). 2) The data is collected in different mobility scenarios and mainly classified into *Static* and *Mobile* scenarios. The overall collected data till now is approximately 60 hours in total from different mobility scenarios with a size of 150 MB collected using a Micropython enabled microcontroller called FiPy device. The dataset is released as a collection of text files and comma-separated values (CSV) files with mainly the timestamp, a unique identifier of the emitting device, RSSI (Received Signal Strength), and other information dedicated to each wireless technology. This dataset is privacy-preserving since it fully meets the GDPR specification, where the MAC addresses and the device names are masked and it is uploaded to GitHub.

2 Related Work

In [FJG+14], Friesen et. al. present a complete data collection system developed at the University of Manitoba that uses a variety of wireless networking technologies and devices to collect inferred traffic data. They used XBee, GSM and Bluetooth modules for designing and implementing a slave probes network with the objective to collect Bluetooth device information. In [VNRG10], Vu et. al. introduce a new framework that collects location information and ad hoc contacts of humans at the University of Illinois campus. The Bluetooth MAC address is used to infer contact information and Wi-Fi MAC addresses are used to infer physical location of the phone. In [GMD20], a data collection campaign and a dataset of BLE beacons for detecting and analysing human social interactions is described, and it is collected in a High School. Although useful, these approaches are limited to contact tracing applications in a single environment (university campus or high school). To the best of our knowledge, the literature does not present any previous work that has generated a labeled dataset from multi-communication wireless technologies and sensors at the same time in different mobility contexts as we propose in this paper.

3 PILOT Dataset

PILOT dataset provides a group of collected packets from three different wireless communication technologies: WiFi, BLE and LoRa, and as additional useful information, the dataset includes the acceleration, roll, pitch and the battery voltage for collecting these data. These traces have been jointly collected all together in several mobility contexts using FiPy microcontrollers from pycom. The duration of each scanned data is between 10 min and 3 hours, and each recorded log file is labeled by its specific mobility scenario of scanning with a description of the conditions of scanning. The dataset is uploaded to github † as a collection of text files and CSV files.

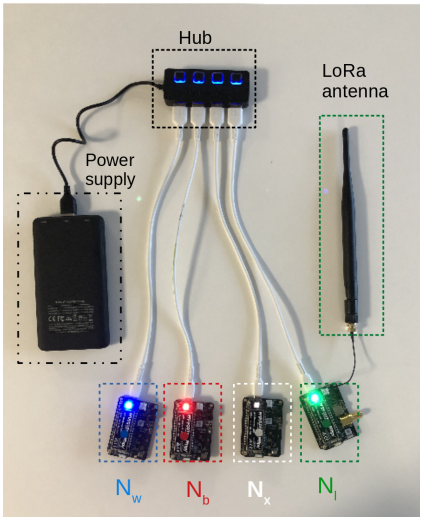

FIGURE 1: Data Collection Setup

3.1 Experimental Design

The experimental setup described in this section is the basis of the collected data and the formation of the desired dataset. The aim is to scan the different wireless communication activities in the range of the scanning device. To do so, we used FiPy devices since they support the three wireless communication technologies: WiFi, Bluetooth, and LoRa. The FiPy is connected to a Pytrack or Pysense expansion board that includes Accelerometer and an SD card module. The development boards of the FiPy include an onboard WiFi and Bluetooth antenna, and for LoRa, an external antenna is connected. To ensure all needed information is gathered at the same time, the scanning process is distributed on four devices as shown in Figure 1. To get the actual time, the FiPy is connected to an access point, to be able to connect to the Network Time Protocol (NTP) server that will help to synchronize the real-time clock (RTC) and get the current timestamp. The collected data by each device is saved on an SD card.

3.2 Configuration and Received Data

The configuration of each wireless technology is as follows:

WiFi Node (N_W): The WiFi node is configured to start active scanning, as the device radio transmits a probe request and listens for a probe response from an AP or active devices such as phones or laptops. The FiPy device supports 802.11b/g/n so the radio scanning is in the 2.4-GHz to 2.4835-GHz spectrum. Upon detecting a probe request, scanners log several pieces of information related to that probe, as it indicates

†. https://github.com/Janakoteich/PILOT-Dataset-Collection-of-Multi-communication-Technologies

the following named tuples: (SSID, BSSID, sec, channel, RSSI). The saved log mainly includes 1) the timestamp related to the probe request detection, 2) the service set identifier (SSID) which is the name of the device, 3) the basic service set identifier (BSSID) which is the MAC address of the sender, 4) the *sec* that stands for security, 5) the channel number which is in the range of 1 to 11, and 6) the received signal strength (RSSI). The value of *sec* attribute defines the type of security where each value means the following: '0' is open, '1' is WEP, '2' is WPA-PSK, '3' is WPA2-PSK, and '4' is for WPA/WPA2-PSK.

Bluetooth Node (N_B): The BLE node is configured for passive scanning to receive the advertising packets (PDUs) that are retrieved every second. The following named tuple with the advertisement data is received during the scanning: $(MAC, addr\_type, adv\_type, rssi, data)$, where *data* contains the complete 31 bytes of the advertisement message. Then this *data* is parsed to get the following information: adv_flag, $scan\_tx\_pwr$, $conn\_tx\_pwr$, tx_range and adv_tx_pwr. The log is first saved with a timestamp since the advertisement is received, with the MAC address of the sender and the RSSI. Other information is retrieved like the advertising flag, the *name* of the device, scanning transmission power, connection transmission power, transmission range, and advertising transmission power.

LoRa Node (N_L): The LoRa Alliance has defined two frequency bands for the usage of LoRa technology in Europe. In our location, only EU868 is supported. The microcontroller will be listening to different frequencies and switch between them every second. In this way, we can receive some LoRa packets that are operating on the following frequencies: 863000000, 864000000, 865000000, 866000000, 867000000, 868000000, 869000000, 864862500, 865062500, 865402500, 865602500, 865985000, 866200000, 866400000, 866600000. These frequencies are defined randomly in the configuration of the FiPy. As we increase the number of frequencies to listen on, the delay will increase for switching between selected frequencies, and as a consequence, the chance of receiving packets will decrease.So, if by chance we received data while listening on a specific frequency, the data will be saved in the file with mainly the following information: The timestamp, the spreading factor, the data itself (which will be masked for privacy issues), the frequency, RSSI, the signal-to-noise ratio (SNR), and other information as shown in the log sample in Figure.

Acceleration Node (N_x): Is the node that collects the acceleration information, roll and pitch.

4 Data Collection Methodology

After scanning, each log is saved with a real timestamp in a text file on the SD card. The goal is to observe and record the variations of the wireless technologies in different mobility contexts, which are mainly categorized into two: *Static* and *Mobile* scenarios. For *Static* we defined the following cases: Home, Office, Restaurant, Bus_station, University and Meetings, and for *Mobile* we have the following scenarios: Pedestrian, Car, Bus, Metro, and Trains. We took into consideration the different conditions of each scenario, as we have rural areas (auto-route), urban areas like cities, and less crowded places like villages.

4.1 Dataset Description

One of the main characteristics of the dataset is that it is labeled. This is achieved by tracking the places and the time of scanning, then the data is retrieved from the four devices and saved in a file with a label based on the scanning conditions that were noted. In the first step we get four text files, then using python programming language, the MAC addresses, and SSIDs' are encrypted, then each text file is transferred to a CSV file. The data is classified and uploaded on GitHub. For each scenario, we have sub-folders named by the prefix of the name of this scenario, and each of them includes the scanned data. At the time this paper is written, 72 datasets are made available with different scanning time slots and intervals for 11 mobility contexts.

5 Dataset Insights

In this section, a comparison between *static* and *mobile* scenarios is presented. Figures 2 and 3, represent the received beacons from each access point over time, with the corresponding RSSI. In Figure 2 which displays the WiFi data collected from an office (static scenario), it can be seen that the beacons are detected

over the whole scanning time. Knowing that the scanner is fixed, would indicate that the access points detected by the scanner are also fixed. While in Figure 3 which is related to the data collected from a bus (mobile scenario) in an urban area, the beacons are appearing only for a very short duration. This is because we are losing connection with the access point because of the mobility of the bus. From these observations, we can see how each scenario has its unique pattern of received beacons. What is more interesting is that several datasets for the same scenario, show the same behavior of received beacons. Such results could be important for building new knowledge and bringing a new perspective to analyzing human mobility.

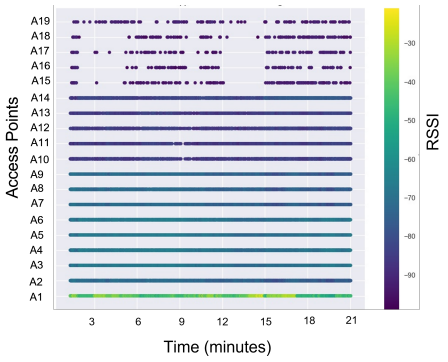

FIGURE 2: WiFi beacons - Office monitoring

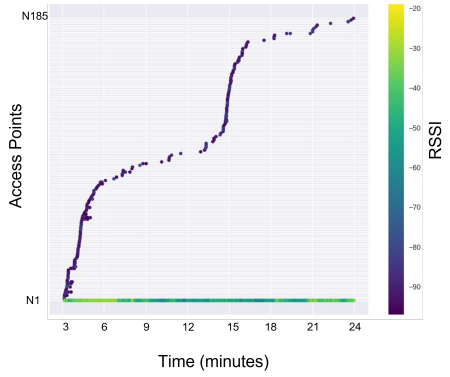

FIGURE 3: WiFi beacons - Bus monitoring

6 Conclusion and Future Work

In this work, we have provided a rich dataset called PILOT dataset that includes four jointly collected information from multi-communication wireless technologies and sensors. The traces are collected in different mobility scenarios mainly categorized as *Static* and *Mobile*. The dataset is a collection of WiFi beacons, BLE beacons, LoRa packets, acceleration, roll, and pitch information. The methodology to reproduce the dataset is illustrated and provided on GitHub. This dataset is reproducible and will be enriched with more data with time. In the meantime, we are still collecting data for an average of 4 hours per week. In future work, we will add new scenarios for collecting data, and we will add a new category for the collected dataset related to predefined mobility models that meet the expectation of more research work.

References

[FJG+14] Marc Friesen, Rory Jacob, Paul Grestoni, Tyler Mailey, Marcia R. Friesen, and Robert D. McLeod. Vehicular traffic monitoring using bluetooth scanning over a wireless sensor network. *Canadian Journal of Electrical and Computer Engineering*, 37(3):135–144, 2014.

[GMD20] Michele Girolami, Fabio Mavilia, and Franca Delmastro. A bluetooth low energy dataset for the analysis of social interactions with commercial devices. *Data in Brief*, 32:106102, 2020.

[GZPZ18] Ruichun Gu, Hongyu Zhang, Dong Pei, and Junxing Zhang. A scalable and virtualized testbed for iot experiments. In *TRIDENTCOM 2018*, 2018.

[SSGL15] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PLOS ONE*, 10(7):1–11, 07 2015.

[VNRG10] Long Vu, Klara Nahrstedt, Samuel Retika, and Indranil Gupta. Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus. MSWIM '10, page 257–265, New York, NY, USA, 2010. Association for Computing Machinery.