Estimating real-time high street footfall from Wi-Fi probe requests

Balamurugan Soundararaj[1], James Cheshire[1] and Paul Longley[1]

[1] Department of Geography, University College London, United Kingdom

January 30, 2018

1 Abstract

In the past decade Wi-Fi has emerged as the most commonly used technology in providing internet access to mobile devices such as smartphones, tablets and laptops in public and private spaces. This has resulted in multiple Wi-Fi networks being available at almost every location in an urban environment. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi antennae regularly broadcast special type of signals known as probe requests to discover available Wi-Fi networks and switch seamlessly between them. This is a hardware level signal which was standardised by IEEE 802.11b/g specification relays information about the source mobile device to any Access Points (AP) available around it. Since this is the first step in establishing a connection between any two devices, it is is universal to any device which uses a Wi-Fi radio to communicate. This makes these probe requests an open, passive, continuous, and wireless source of data available at any urban location which can provide us with clues in understanding the number of people present in the immediate surroundings in real-time with high granularity [1, 3]. In this paper, from a set of probe requests collected at a high street location in London, along with manually counted data, we demonstrate that pedestrian footfall can be estimated with considerable accuracy without infringing on the privacy of the mobile users involved.

There have been numerous attempts at using Wi-Fi to measure the volume and movement of people in built environment for various applications [7, 6, 5]. Though most research observe feasible and favorable results, in recent years, one of the major challenges faced in such attempts has been the MAC address randomisation process. This process aims to protect the users' privacy by anonymising the only globally identifiable portion of the probe requests resulting in a set of probe requests generated by the same device with different random MAC addresses [2]. There have been various successful research in breaking this randomisation process to extract real MAC addresses [4] but this usually results in serious risk in infringement of privacy of the users of the mobile devices. There is a clear gap in the research for exploring methodologies which enables us to estimate the number of unique mobile devices from a set of anonymised probe requests without the need to reveal their original MAC addresses.

A pilot survey was conducted at Oxford street in London on 20 December 2017 from 12:30 to 13:00 where two sets of data were collected on pedestrian footfall through Wi-Fi sensing and manual counting in parallel. The Wi-Fi sensor collected all the probe requests that were broadcast around the area and recorded the timestamp at which they were collected, MAC address of the source mobile device (anonymised using a hashing algorithm), organisationally unique identifier (OUI) of the manufacturer of the device, total length of the signal in bits, the strength of the signal reported by the mobile device in dBm, the sequence number of the signal, duration for which the signal was transmitted, the service set identifier (SSID) of the access point targeted by the probe request and the length of the extra information (tags) embedded in the packets. The manual count was done by an android application on a mobile phone which records the time-stamp every time the surveyor touches the screen which corresponds to one pedestrian footfall on the sidewalk.

Being one of the busiest location in the United Kingdom, the location generated approximately 60,000 probe requests in the 30 minutes interval through the Wi-Fi sensor and 3,722 people were counted manually. When we aggregate the probe requests by their MAC address for every minute, the difference between the sensor counts and the manual counts is observed to be on average 425%. This shows that there is a

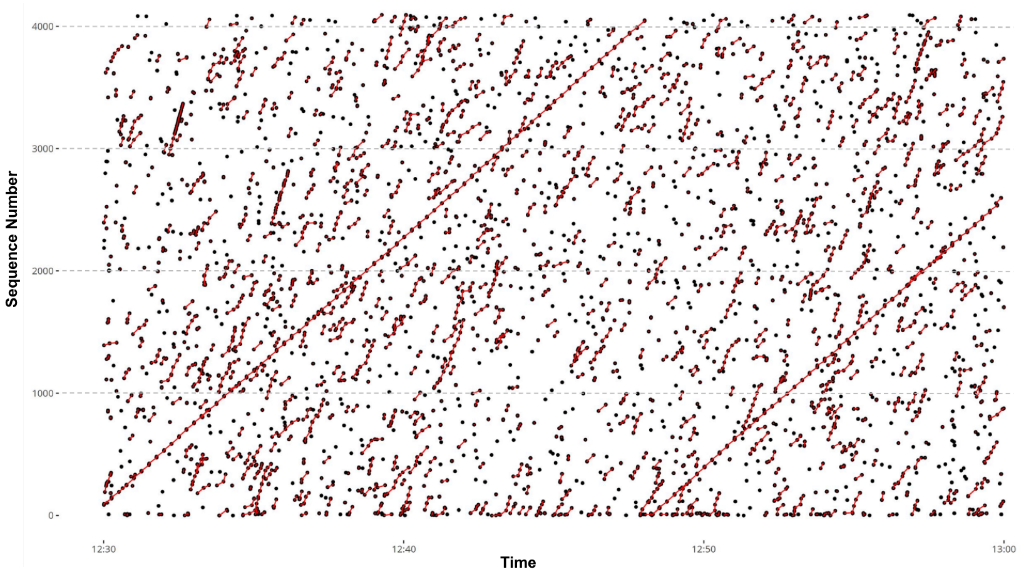

Figure 1: *Clustering probe requests based on increasing sequence numbers present in them.*

large amount of noise in the data which might include signals from devices outside the area where the manual count was conducted and the anonymised probe requests from the same devices with different MAC addresses. Before we go look at the data in detail, an initial analysis shows that the fields - SSID and tags are very sparse and doesn't provide much information for our cleaning process and the duration field is closely related to the length of the probe request and provides no new information. So we remove these fields from further analysis.

We remove the noise from devices outside the area of interest by removing all the probe requests which report a "low" signal strength. This classification of "high" vs "low" is done through "k-means" classification algorithm. The cutoff point for the collected data is -71dBm. This process of filtering is highly effective and reduces the difference between the sensor counts and manual counts to 30%. We observe that around 55% of all probe requests collected are anonymised. We assign the hashed MAC address the unique identifier for the rest of the 45% and investigate the anonymised probe requests further.

We then use the fields - OUI, lengths and sequence number, to tackle the noise from devices which anonymising the probe requests. OUI and length are used to split the dataset into groups probe requests from similar devices and each subset is classified further based on a graph based clustering algorithm where each cluster correspond to an unique device. The algorithm works by creating a graph where the nodes are the probe requests and the links are created between them based on the following rules, A link can go only forward in time. A link can exist between nodes with a maximum time difference of t (time threshold) A link can go from low to high sequence numbers. A link can exist between nodes with a maximum sequence number difference of s (sequence threshold) A node can have only one incoming link and outgoing link which is the shortest of all possible such links.

Then we turn our attention to other fields which can provide us information on the device which generated the probe request and hence enabling us to finger print them uniquely. From a preliminary investigation, we find that the data fields - SSID and length of the tags are very sparse and not useful in identifying the unique device while OUIs, length of the packet and sequence numbers can be crucial in doing the same. A pair of probes with the same OUI shows the device manufacturer and same length of the packet signifies that there is a high probability that it was generated by a similar devcie.

We first split the anonymised probe requests by the OUI and length then run a graph based partition algorithm on them based on the sequence numbers. This algortithm links probes with increasing sequence numbers in increasing time within a specific distance. It also enforces the rule that a probe request has only one incoming and outgoing links. It then returns the membership of the probe requests in distinct connected clusters in the resulting graph thus classifying probe requests with sequentially increasing numbers as ones generated from the same device as shown in Figure 1. Finally all the individual subsets are joined and cleaned for repeating probe requests with unique clusters. This filtering in-turn reduces the difference of the sensor count to the manual count to -18%.

The important detail to notice is that we have clustered and estimated unique device without the knowledge of the MAC address of the devices hence delivering output without compromising the users' privacy. The methodology can be applications in numerous footfall counting projects in retail, urban planning, facilities management etc.

References

[1] Julien Freudiger. “How talkative is your mobile device?: an experimental study of Wi-Fi probe requests”. In: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2015, p. 8.

[2] Ben Greenstein et al. “Improving wireless privacy with an identifier-free link layer protocol”. In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM. 2008, pp. 40–53.

[3] Constantine E Kontokosta and Nicholas Johnson. “Urban phenology: Toward a real-time census of the city using Wi-Fi data”. In: *Computers, Environment and Urban Systems* 64 (2017), pp. 144–153.

[4] Jeremy Martin et al. “A Study of MAC Address Randomization in Mobile Devices and When it Fails”. In: *arXiv preprint arXiv:1703.02874* (2017).

[5] Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. “LifeTag: WiFi-based continuous location logging for life pattern analysis”. In: *LoCA*. Vol. 2007. 2007, pp. 35–49.

[6] Piotr Sapiezynski et al. “Tracking human mobility using wifi signals”. In: *PloS one* 10.7 (2015), e0130824.

[7] Vasileios Zarimpas, Bahram Honary, and Mike Darnell. “Indoor 802.1 x based location determination and realtime tracking”. In: *The IET International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2006), Hang Zhou, China*. IET. 2006.