Estimating real-time highstreet 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

June 20, 2018

Abstract - Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

In the past decade Wi-Fi has emerged as the most commonly used technology in providing high speed 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 dense urban environments. Traversing through this overlapping mesh of Wi-Fi networks, modern mobile devices with Wi-Fi antennae regularly broadcast a special type of signal known as 'Probe Requests', in order to discover Wi-Fi networks available to them. This helps these devices to connect and switch between the WiFi networks seamlessly.

Probe requests are low level signals standardised by IEEE 802.1b/g specification as the first step in establishing a Wi-Fi based connection between two devices and is implemented in any Wi-Fi capable device irrespective of the manufacturer or the model. This ubiquity and standardisation make them an excellent source of open, passive, continuous, and wireless data generated by Wi-Fi capable devices present at any given time and location. Considering the unprecendented levels of mobile device ownership in recent years, we can in turn use this data to understand the population distribution in highly dynamic urban environments with high spatial and temporal granularity [1, 3].

While a Wi-Fi based method to collect data offers us various advantages such as, easy scalability and efficiency in terms of cost and time, It also introduces few systematic biases, uncertainities in the collected data along with the serious risk of infringing on the privacy of the mobile users. In this paper, using a set of probe requests and manual counts collected at various high street locations across London, we demonstrate that pedestrian footfall at these locations can be estimated with considerable precision and accuracy while protecting the privacy of the pedestrians.

2 Previous Work

There have been numerous attempts at using Wi-Fi to measure the volume and movement of people in the built environment for various applications [7, 6, 5]. Though most research obtains 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, which results in a set of probe requests generated by the same device with different random MAC addresses [2]. There have been various successful attempts by researchers to breaking this randomisation process in order to extract real MAC addresses, [4] but this usually results in serious risk of infringement of the privacy of the users of the mobile devices. There is a clear gap in the research for exploring methodologies which enable 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.

3 Methodology

The overall methodology is to use a WiFi antenna to collect probe requests, assign a unique identification based on the device that created the request, aggregate them based on the unique identification for a specific time interval to create an estimate of number of people at the location. In this section we look at the characteristics of probe requests in detail, outline the methodology used to collect these probe requests, look at the uncertainities and biases in the process and device methods to overcome these issues.

3.1 Probe Requests

Probe request is a low level packet sent by a mobile device as a means of "scanning" for various access points available at a specific location. This packet consists information about the mobile device including but not limited to, Media Access Control address - A two part 12 bit identifier where the first part identifies the manufacturer of the device and the second part identifies the device itself. This MAC address is of two types, global - the real identifier of the device which doesnt change and local - virtual, random addresses used for temporary situations. Sequence number of the packet to keep track of the replies. The access points for which the packet is being sent to. Capabilities of the device. We can also infer other things about the packet such as time at which the packet has been received, total length of the packet time it took to transmit the packet Signal strength of the packet. All of this can help us identifying the packets and label them when they are transmitted by the same device.

3.2 Data Collection

The data collection involved two processes - A sensor based collection of probe requests and a manual count of pedestrians in the area next to the sensor. We have detailed the sensor design below,

The sensor consists of wireless transponder in a linux/ MacOSX based machine which is put into a monitor mode and wireshark software (tshark) is used to sniff, filter and save the relevant packets. The filters applied are management frame > probe requests > broadcast. This is then parsed, personal information is hashed, compressed and sent to a server via web socket protocol. The server receives, logs and stores the data in a postgresql database, for further analysis. The manual count was undertaken using an Android application on a mobile phone where timestamp of every individual pedestrian footfall was recorded by touch screen. A schematic is shown in ref figure.

3.3 Estimating Footfall

The next step after collecting the probe request is to estimate the footfall or pedestrian activity from them. There are three major challanges in the above data collection methodology. the first one is deliniating the area of interest. WiFi and its range depends on lots of factors. need a way to isolate the noise from the relevant data. depends a lot on microsite so we need a methodology which provides us with a general solution without much tinkering. Second one is the mac randomisation. modern phones randomise their MAC addresses to avoid detection and tracking. This leads to overcounting and explosion of randomised (local) mac addresses. We need a non MAC dependent method to overcome this. Finally, inherent mobile phone carrying bias. we need to account for the average mobile ownership. which is steadily increasing and is not 1:1.

3.3.1 WiFi range

One of the clues of the distance of the device generating the probe request is the strength of the signal recieved from it. The Signal strength varies in inverse square law over distance with a propagation constant. It also depends on lot of micro site, micro temporal factors. There cannot be a simple rule to fit and filter for all configuration. Our hypothesis is that in a specific setting and specific source of noise, there must exist a clear break in the data. for example, if there is a phone shop next to our sensor where hundreds of phones regularly send lots of probe requests we should be able to see a large increase in number of probe requests around a specific signal strength. we can identify this sharp change/ break using class interval algorithms such as k-means, jenks, quantile, etc.

3.3.2 MAC randomisation

This is a recent problem. ref in figure. how mac randomisation has caused problem MAC address has been our unique identifier. Now we need to look for others. the contenders are length, duration which seem to be unique for device sets of known wlans and capabilites which can give us unique finger print and finally sequence numbers in the packets. This is a tricky one since it is neither unique nor aggregatable. we need a method to seperate sequences shown in fig. We propose a graph based clustering algorithm where each cluster corresponded to a unique device. The algorithm creates a graph where the probe requests represented the nodes, and links are created between them based on the following rules:

1. A link could go only forward in time.

2. A link could exist between nodes with a maximum time difference of α (time threshold).
3. A link could go from low to high sequence numbers.
4. A link could exist between nodes with a maximum sequence number difference of β (sequence threshold).
5. A node could have only one incoming link and one outgoing link, which is the shortest of all such possible links.

The nodes were then classified based on the unique connected component they belonged to. This classification was assigned as the unique identifier for the anonymised probe requests this unique identifier is used instead of MAC to aggregation.

3.3.3 Mobile phone ownership

This is both a long term change and result of demographic factors. Phone ownership is not 1:1. It changes slowly overtime. It also changes with place to place. It can be more in dense urban centers and can be low in rural areas. We propose a adjustment factor methodology we use periodically carried out manual counts to adjust the numbers to what is reported on ground. The adjustment is as strong as the amount of ground truth we collected.

4 Pilot Study

We conducted a pilot study to get a feel for the data. See which fields are relevant and which ones are not. to check if the sequencing algorithm holds true. to findout which one of the classification algorithm works. if these methodology works in real world scenario scalably. A pilot survey was conducted on Oxford Street in London in December 2017, where two sets of data were collected on pedestrian footfall with the aim of establishing merit in measuring pedestrian footfall as a function of the number of wifi probe requests collected at a given location. These datasets were collected through Wi-Fi sensing and manual counting in parallel.

Being located at one of the busiest retail locations in the United Kingdom, the WiFi sensor captured approximately 60,000 probe requests over a 30 minutes interval, and 3,722 people were counted manually.

When we aggregated the probe requests by their MAC address for every minute, the difference between the sensor counts and the manual counts was observed to be on average 425%. This suggested that there was a large amount of noise in the data which might have included signals from devices outside the area where themanual count was conducted, as well as anonymised probe requests from the same devices but with different MAC addresses. This process of filtering was highly effective and reduced the difference between the sensor counts and manual counts to 30%. We observed

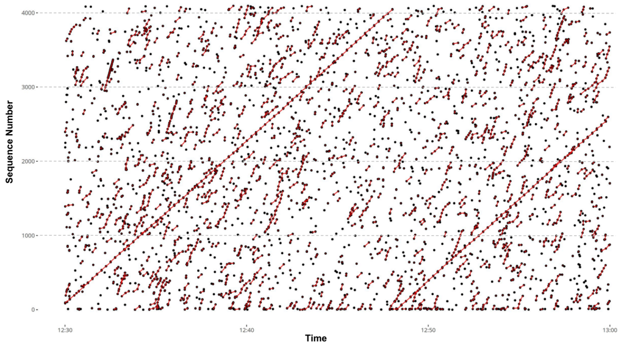

Figure 1: *Clustering probe requests based on increasing sequence numbers present in them.*

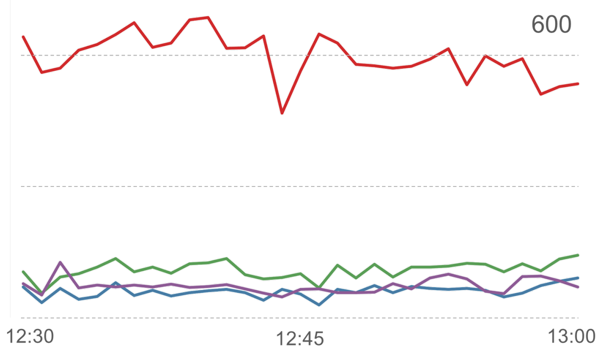

Figure 2: *Comparision of counts after filtering with manual counts*

that around 55% of all probe requests collected were anonymised. We assigned the hashed MAC address the unique identifier for the remaining 45% and investigated the anonymised probe requests further.

An initial analysis revealed that the fields - SSID and tags - were very sparse and did not provide much information for our cleaning process. In addition, the duration field was closely related to the length of the probe request and provides no new information. Therefore, we removed these fields from further analysis. We eliminated the noise from devices outside the area of interest by removing all the probe requests which reported a "low" signal strength. This classification of "high" vs "low" was performed using a k-means classification algorithm. The cut-off point for the collected data was -71 dBm.

Figure 1 shows the clustering process: the black dots show the probe requests and the red lines connect them into clusters representing those which were generated by the same device. We finally combined both normal and anonymised probe requests, aggregated them based on their unique identifier, and removed repeating probe requests which reduced the difference between the sensor counts and the manual counts to -18%.

5 Main Study

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

5.1 Aims and Objectives

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5.2 Data Collection

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

5.3 Results

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

6 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

7 Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

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.