Estimating real-time highstreet footfall from Wi-Fi probe requests

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

[a]Department of Geography, University College London, United Kingdom.

ARTICLE HISTORY
Compiled June 24, 2018

ABSTRACT
The accurate measurement human activity with high spatial and temporal granularity is crucial for us to understand the structure and function of built enviroment. With increasing mobile ownerships, the Wi-Fi probe requests generated by these devices can be an excellent source of data for such measursement. The major challanges in using Wi-Fi technology for such purpose are precisely delineating the field of measurement and arriving at accurate estimates without compromising the privacy of the users of the mobile devices. In this paper we demonstrate that, with the application of class intervals and a novel graph based technique, we can overcome the challanges and reliably measure real-time pedestrian footfall at retail highstreets.

KEYWORDS
Highstreet footfall; Wi-Fi Probe requests; Sensors; MAC Randomisation

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 (IEEE 2013) 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 (Freudiger 2015, Kontokosta and Johnson 2017).

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

CONTACT Balamurugan Soundararaj. Email: s.bala@ucl.ac.uk

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 (Zarimpas *et al.* 2006, Sapiezynski *et al.* 2015, Rekimoto *et al.* 2007). 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 (Greenstein *et al.* 2008). There have been various successful attempts by researchers to breaking this randomisation process in order to extract real MAC addresses, (Martin *et al.* 2017) 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 primary aim of this research is to enable us to collect a series of probe requests and process them into an usable pedestrian footfall count. We do this by using a Wi-Fi receiver to collect probe requests broadcasted by mobile devices, filtering out the background noise and aggregating them based on the device that generated them. We begin by looking at the characteristics of probe requests in detail, device a methodology to collect these probe requests in public areas, examine the systemic biases and uncertainties in the data collection method and device data processing methods to overcome these challanges. Finally we compare the processed footfall counts to the ground truth recorded by primary surveys.

Probe requests are a special type of management packets broadcast by Wi-Fi enabled devices as part of the various functions such as scanning for available access points (AP), quick geo-location by triangulation based known APs, etc. These are broadcast by all Wi-Fi enabled devices regardless of the manufacturer, type or model of the devices though there is some variation on the frequency and the information transmitted through them. In some cases, such as Android devices, these are broadcast even when the Wi-Fi functionality has been turned off by the user. Thus these signals can be used to reliably identify the presence of Wi-Fi enabled mobile devices.

Being a first step of connection initiated by the mobile device, these packets have information regarding the characteristics of the mobile device itself. Some of the key information we can infer from these requests are,

(1) **Media Access Control (MAC) address** which is an unique identifier for the wireless hardware of the mobile device,
(2) **Sequence number** of the request for the mobile device to keep track of the responses,

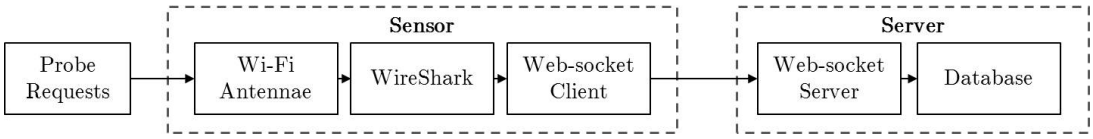


Figure 1. Schematic diagram showing the process of collecting probe requests using the sensor

(3) **Timestamp** at which the request was received by the AP,
(4) Total **length** of the request in number of bits, and
(5) The **strength of the signal** which transmitted the request.

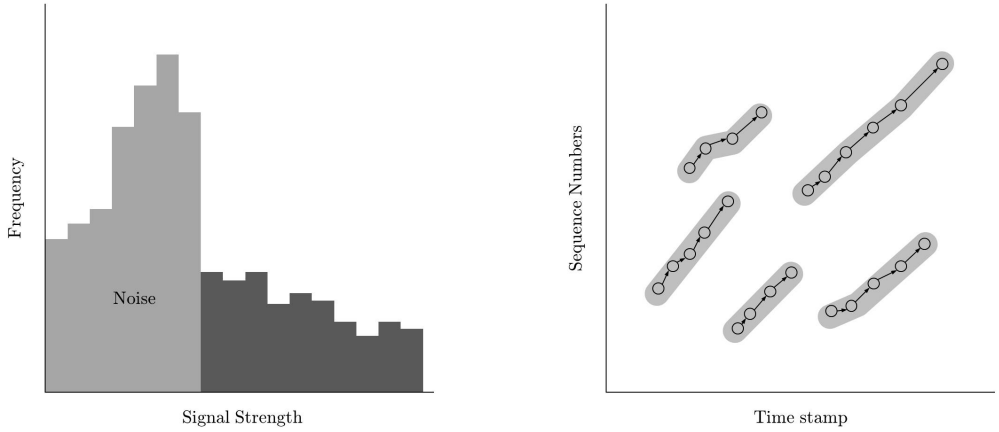
The MAC address is the primary unique identifier for the mobile device. It has two parts, first part is called an Organisation Unique Identifier (OUI) which gives information about the manufacturer of the device and the second is unique to the device. The MAC address can be randomised (hence non unique) and is marked as such. Though sequence number and length of the packet are not strictly unique, we hypothesize that we can use them to estimate unique devices.

Data collection was done with the help of custom sensors built from modifing the Smart street sensor (CDRC 2016) hardware and updating them with custom software. The sensor is essentially a Raspberry Pi connected with Wi-Fi and 3G antennae. It keeps the Wi-Fi module in 'Monitor' mode and uses the open source software - wireshark cite to passively collect all packets sent to 'broadcast', marked with type - 'management' and subtype - 'probe requests'. The MAC address in these probe requests is anonymised using a cryptographic hashing algorithm and transmitted through 3G connection to a central database via web-sockets protocol, where it is stored in a PostgreSQL database for further analysis. A overall schematic of the data collection process is shown in Figure 1. The ground truth on number of pedestrian footfall was recorded using a purpose built Android application cite.

The next step after collecting data was to estimate the footfall or pedestrian activity from them. We identified the following major challanges which arise from our collection methodology.

(1) **Background noise** - since the extent to which Wi-Fi signals travel differs subject to various factors such as interference and humidity, it is close to impossible to restrict our data collection to a finite area of interest. This can lead to a signicant background noise at certain locations. E.g. a phone shop or a bus stop located next to the study area can increase the number of probe requests received by the sensor.
(2) **MAC randomisation** - The mobile devices in the past few years have been using randomised 'local' MAC addresses for probe requests to protect the users from being tracked. This makes it impossible to tell if the probe requests are being sent by the same mobile device which is being stationed next to the sensor. This along with the previous problem can further increase the magnitude of error by several fold.
(3) **Mobile ownership** - Since the rate of mobile ownership can vary widely across geography and demography, we cannot assume that every mobile device translates to one pedestrian footfall. In addition to this, there is a long term overall increase in mobile ownership which may lead to the number of probe requests collected overtime.

We propose the following methods to tackle each of these challages.



(a) Distribution of signal strengths showing the filtering of background noise

(b) Clustering probe requests as nodes in a graph using increasing sequence numbers

Figure 2. Schematic diagrams explaining the methods for filtering by signal strength and clustering using sequence numbers

3.1. *Filtering with Signal Strength*

One of the clues that we can use to estimate the distance between the mobie device and the sensor is the strength of the signal received by the sensor. The obvious approach here is to try and establish a relationship between the signal strength and distance first and use this to filter out the unwanted probe requests. This approach has numerous pitfalls and uncertainities since the decay of signal strength with distance is not always constant. It varies with atmospheric conditions, presence of obstructions between the source and target, the nature of these obstructions and the strength (power level) of the source transponder. This severely limits our ability to establishing a simple conversion between reported signal strength and distance. There is a need for a method which takes in to account these variables across varous locations.

We hypothesise that in configurations where a specific source of background noise is at a constant distance, there must be a distinct break in the number of probe requests reporting signal strength corresponding to that distance. For example, if there is a phone shop next to our sensor where hundreds of phones regularly send probe requests there should be a sharp rise of number of probe requests with reported singal strength corresponding to the distance between the sensor and the phone shop at any given set of conditions as shown in Figure 2. We could identify these breaks in the data using traditional one diamensional classification algorithms such as Jenks natural breaks, k-means, quantile and hierarchical clustering, etc. Since we are only looking for the break in the data and not for absolute values, the methodology should apply for all the variations due micro site conditions thus reducing the overall noise in the collected data.

3.2. *Clustering with sequence numbers*

Since our primary unique identifier - MAC adddress, is being anonymised by new devces, we need to find other information present in the probe request for a unique identifier. Obvious approach here is to establish a factor of randomisation and adjust

the counts for these probe requests based on this factor. We found this aproach not feasible, since the proportion of devices which randomise the MAC addresses increases over time. There is also a wide variation in the frequency at which the devices randomise the MAC addresses and the method used for the process. This leads us to look for a more generalisable approach which is independent of the device model.

From our initial analysis we found that OUI, length of the packet and sequence number of the packet being the most promising information to acheive this. First we divide our dataset into sets of probe requests with randomised and non-randomised MAC addresses and keep the MAC address as the unique identifier for the latter set. For randomised ones we futher divide them in to sub categories based on their OUIand length of the packet. Since the length tends to stay unique to specific models of devices we are left with the task of identifying the unique mobile devices from within these distinct models.

The proposed algorithm creates a graph where the probe requests represented the nodes, and links are created between them based on the following rules:

- A link could go only forward in time.
- A link could go from low to high sequence numbers.
- A link could exist between nodes with a maximum time difference of α - time threshold.
- A link could exist between nodes with a maximum sequence number difference of β - sequence threshold.
- 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 belong to as shown in Figure 2. This classification was assigned as the unique identifier for the anonymised probe requests in the place of MAC address. Though the recycling of sequence number after 4000 leads to multiple classifications reported on single device, the maginitude of error is greatly reduced.

3.3. *Calibrating with ground truth*

Since mobile phone ownership is an external uncertainty to our study and could arise from variety of spatio, temporal and demographic factors, we propose to solve this using external source of information. We hypothesize that an adjustment factor could be arrived at for each location of data collection, comparing the sensor based counts and ground truth and it can be used to adjust the data reliably to reflect the ground truth in absolute numbers for the future. This calibration can be carried over periodically and the frequecy of which will improve the quality of the estimation.

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

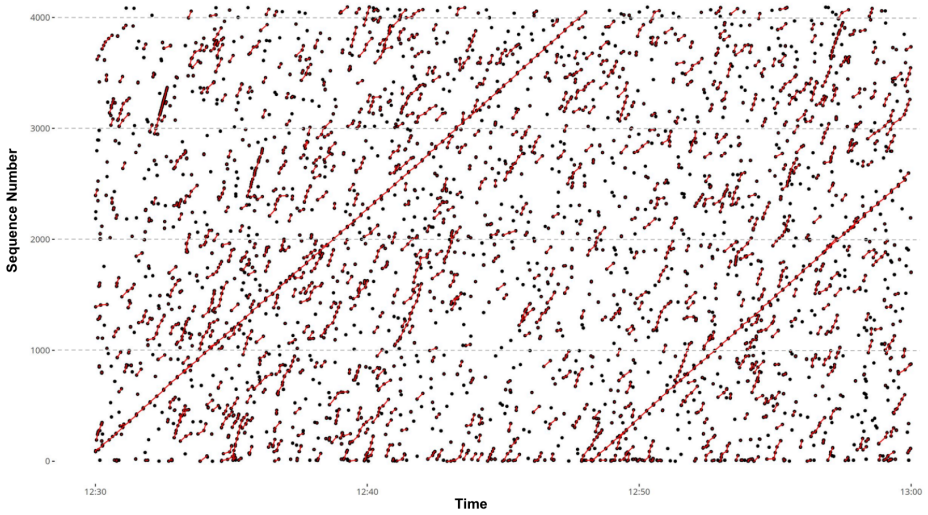


Figure 3. Clustering probe requests based on increasing sequence numbers present in them.

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 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 3 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%.

we find that the methodology works as shown by the reduction in the mean error We find kmeans and quantile best algorithms for filtering sequencing works. Even in densely populated areas.

Though this is promising we need to know more to generalise the methodology.

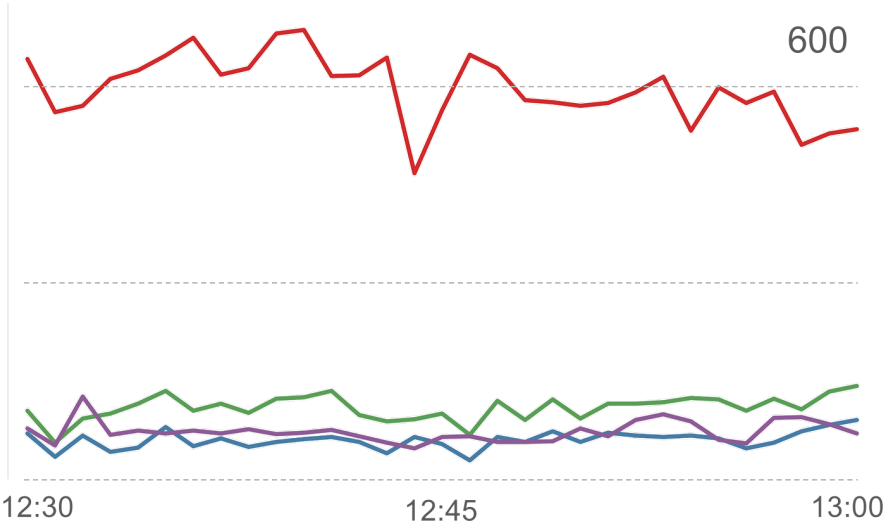


Figure 4. Comparision of counts after filtering with manual counts

we do a detailed multi location, longer term study with multiple manual verification which is the main study.

5. Main Study

The aims of the main study are, test the validity of the signal strength algorithm in different micro site conditions. Test that the sequence number algorithm works in real world for different locations and different times. check if the thresholds are consistent. Test if the calibration works over intervals Finally conclude if we can estimate footfall confidently with just probe requests.

five locations were selected across central london which had different types of configurations and specific problems configuratons are shown in fig. map is shown in figure.

(1) Phone Shop Camden - has phones and bus stops.
(2) Restaurant TCR - has seating area on either side.
(3) Holborn Information Kiosk - High volumne station entry
(4) Restaurant Russell Square - seating on one side and side walk on other
(5) Shop Charring Cross - sidewalk on one side and phone shop next door

Installations were carried out over the time period from xxxx to xxxx. the data collection happened from xxxx to xxxx. Manual counting was carried out with high precision on dates xxxx and aggregated five minutes on xxxx. The difference in methods could lead to some inaccuracies in data. The overall schedule is shown in half page graphic.

The overall statistics of data collected. How many probe requests. Comparision of location in terms of volume, patterns in daily footfall etc. The comparision between global and local. comparision between different types of vendors. Specifics on top 5 manufacturers.

We do a daily analysis of distribution of signal strengths. The thresholds are shown in the table. The average is xxxx and standard deviation is xxx. we notice that the variation is lot. There is a definite change with the micro site locations.

We see how the signal strength filtering affects the counts. compared to manual counts we look at the average mean errors (daily) per 5 minutes. The counts go as follows. Shown as a red line in the Figure.

We can conclude that even though it has variations, this is a good method to reduce the overall level of error.

This is also done for different locations hourly for all the data we had. we compare it to the manual counts and see that the average mean error has been reduced/increased. The finger print works well for all the locations. It also works over a period of time and gives us a comparable and close footfall count to the manual count. The thresholds found in the pilot study works as well.

Finally we normalise the sensor counts to match the manual count using a fraction/ adjustment factor calculated from the know manual counts. we have three sets of counts. We check if the adjustment factor holds the same in all three counts across locations. It does with a variations from xxx to xxx. The results are shown in the table.

we see that the signal strength filtering works and reduces error by xxxxx. there is variation by locations. we see that sequence number algorithm works as well. The threshold stays constant well and works well across locations. The calibration also works and ajustment factor stays consistent short term. This needs more work long term.

6. Conclusion

We have established methodology for collection and estimation of footfall from Wifi data. It is important to note that the filtering process was done based soley on the information present in the probe requests and their temporal distribution. This ensured that although the mobile devices were uniquely identified, there was no further personal data generated by linking the probe requests to the users of the mobile devices. This method essentially gave us a way to estimate the footfall in real-time without identifying or tracking the mobile devices themselves. A real time footfall in these locations are shown in the figure. A dashboard can be found at xxxxx.

This Wi-Fi based footfall counting methodology offers a large number of applications and benefits for real time spatial analysis. Since Wi-Fi based sensors are inexpensive and the data model is scalable, it is possible to use this methodology for a large network of sensors to gather granular data on pedestrian footfall. Projects such as SmartStreetSensors (CDRC 2016), may utilise this methodology to overcome the challenges introduced by the implementation of MAC address randomisation. Such precise and granular data also enables us to confidently model the pedestrian flow in urban road networks, and will be an indispensable tool in the smart city framework. It can also be used to understand and classify geographical areas based on the spatio-temporal distribution of the volume of activity in them.

Acknowledgement

Authors would like to acknowledge the contributions of the Local Data Company in facilitating the data collection process and Dr Roberto Murcio for his inputs.

References

CDRC, 2016. Smart street sensor project. goo.gl/E4tR8o. [Online; accessed 31-January-2018].

Freudiger, J., 2015. 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, 8.

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

IEEE, 2013. Part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications; amendment 4: Enhancements for very high throughput for operation in bands below 6 ghza.

Kontokosta, C.E. and Johnson, N., 2017. Urban phenology: Toward a real-time census of the city using wi-fi data. *Computers, Environment and Urban Systems*, 64, 144–153.

Martin, J., *et al.*, 2017. A study of mac address randomization in mobile devices and when it fails. *arXiv preprint arXiv:1703.02874*.

Rekimoto, J., Miyaki, T., and Ishizawa, T., 2007. Lifetag: Wifi-based continuous location logging for life pattern analysis. *In*: *LoCA*. vol. 2007, 35–49.

Sapiezynski, P., *et al.*, 2015. Tracking human mobility using wifi signals. *PloS one*, 10 (7), e0130824.

Zarimpas, V., Honary, B., and Darnell, M., 2006. 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.