Participatory Location Fingerprinting through Stationary Crowd in a Public or Commercial Indoor Environment

A K M Mahtab Hossain
ISEC Research Group, University of Greenwich
London, United Kingdom
a.k.m.hossain@gre.ac.uk

George Loukas
ISEC Research Group, University of Greenwich
London, United Kingdom
g.loukas@gre.ac.uk

ABSTRACT

The training phase of indoor location fingerprinting has been traditionally performed by dedicated surveyors in a manner that is time and labour intensive. Crowdsourcing process is more efficient, but is impractical in public or commercial buildings because it requires occasional location fix provided explicitly by the participant, the availability of an indoor map for correlating the traces, and the existence of landmarks throughout the area. Here, we address these issues for the first time in this context by leveraging the existence of ***stationary crowd*** that have timetabled roles, such as desk-bound employees, lecturers and students. We propose a scalable and effortless positioning system in the context of a public/commercial building by using Wi-Fi sensor readings from its stationary occupants' smartphones combined with their timetabling information. Most significantly, the ***entropy*** concept of information theory is utilised to differentiate between good and spurious measurements in a manner that does not rely on the existence of known trusted users. Our analysis and experimental results show that, regardless of such participants' unpredictable behaviour, including not following their timetabling information, hiding their location or purposefully generating wrong data, our entropy-based filtering approach ensures the creation of a radio-map incrementally from their measurements. Its effectiveness is validated experimentally with two well-known machine learning algorithms.

CCS CONCEPTS

• **Networks** → **Location based services**; • **Human-centered computing** → **Ubiquitous computing**; • **Computing methodologies** → *Machine learning*; • **Mathematics of computing** → Information theory.

KEYWORDS

Indoor Localisation, Location Fingerprinting, Crowdsourcing, Entropy, Wi-Fi, Participatory Sensing

ACM Reference Format:
A K M Mahtab Hossain and George Loukas. 2019. Participatory Location Fingerprinting through Stationary Crowd in a Public or Commercial Indoor Environment. In *16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous), November 12–14, 2019, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3360774.3360791

MobiQuitous, November 12–14, 2019, Houston, TX, USA

ACM ISBN 978-1-4503-7283-1/19/11…$15.00
https://doi.org/10.1145/3360774.3360791

1 INTRODUCTION

A significant portion of a human's daily life is spent indoors. The emergence of smart ubiquitous applications generally requires access to a human's location information in such indoor environments too. Yet, despite having garnered tremendous interest in the research community, there is still no de-facto standard for indoor location determination (i.e., indoor localisation). Traditionally, two families of indoor localisation research have been pursued: one that requires specialised hardware (e.g., customised devices) and infrastructure setup within the localisation area [16, 20], and the other that utilises existing communication infrastructure such as Wi-Fi [1, 22] or Bluetooth [8]. The first family can enable centimetre-level accuracy with the help of specialised indoor infrastructure, but is extremely costly. Therefore, it is deemed impractical for a commercial or public indoor environment. Even though the second family provides coarser localisation accuracy (2 to 3 meter or sometimes even room-level granularity), it can be more practical and cost-effective for a public or commercial building facilitating location based services (LBSs), such as locating the nearest store or distributing electronic coupons in proximity to various business intelligence applications. Within this second family, location fingerprinting is a particularly popular approach, which involves one or more surveyors tasked with conducting a training phase by positioning themselves at several points of interest and collecting the signal strength samples. This process is time-consuming and labour intensive, hence suffers in terms of scalability in commercial and public building scenarios. Also, the surveyors need to be aware of the geometry of the building for explicitly indicating their position within an indoor map. Access to a map of a public or commercial building comprising of multiple owners or tenants can also be quite difficult for such purpose.

A newer trend of localisation techniques encourage implicit participation of users in such premises to achieve the same goal, with the main motivation of being the elimination of the surveyor's laborious training phase of fingerprinting. This approach generally involves crowdsourcing inertial sensor measurements (e.g., accelerometer, gyroscope, compass, etc.) from people's smartphones, followed by the application of Simultaneous Localisation and Mapping (SLAM) with dead-reckoning, sensor fusion and filtering techniques (e.g., Kalman) to compute the location [6, 17, 19]. These crowd-sourced localisation approaches have been shown to be impractical for a public or commercial building [11] because of the

requirement of an occasional location fix provided explicitly by the participant, the availability of an indoor map for correlating the traces, and the existence of landmarks throughout the area.

In this paper, we leverage the existence of a "stationary" crowd in an indoor environment for localisation purpose. In a public or commercial building, a number of people's positions can be considered stationary during a certain time of the day, and this is also supported by research findings such as [3]. For example, a salesperson in a shopping mall or a security guard of an office is expected to be at a certain location during working hours. This is generally true for many desk-bound employees in such public or commercial buildings. In a school or university, a student or teacher operates according to timetabling information. For example, a student might be scheduled to attend a tutorial session in room X at 2 pm on Tuesdays, and a teacher may deliver a lecture in room Y at 10 am on Wednesdays. In our approach, the crowd-sourced sensor readings (specifically Wi-Fi) from only these *stationary* personnel's smartphones are correlated with their expected position at certain times to formulate the location fingerprint. As a result, the need of a surveyor together with the aforementioned issues of the fingerprinting approaches are avoided.

The main contributions of the paper are as follows:

(1) We propose a scalable and effortless indoor positioning system in the context of a public/commercial building by utilising its *stationary* occupants' smartphones' Wi-Fi sensor readings combined with their timetabling information. We argue that this implicit participatory location fingerprinting radio-map creation will relieve the traditional laborious training phase.
(2) While the *stationary* crowd need not be aware of the underlying location-based data collection, there must be a provision for incorporating only the good quality sensor readings and filtering out spurious ones, for example if a user is not at his/her expected position (for legitimate or even malicious reasons). For this purpose, we utilise the ***entropy*** concept of information theory to differentiate between good and bad quality sensor measurements. To the best of our knowledge, no work has used entropy in the creation of a fingerprinting radio-map database before.
(3) Our approach has been experimentally validated using data collected from a floor of our university campus. A few lecturer volunteers participated in building the fingerprinting radio-map for the floor comprising of seven office rooms.

The rest of the paper is organised as follows. In Section 2, we discuss our idea of using entropy to differentiate between good and bad quality crowdsourced sensor measurements, and a resulting filtering algorithm for incorporating them into fingerprinting radio-map. We provide a brief description of related work in Section 3. In Section 4, we present our evaluation with experimental findings. Finally, we discuss in Section 5 the conclusions drawn, and our future work.

2 INFORMATION CONTENT IN LOCALISATION

2.1 Location Fingerprinting Principle

Suppose there is a set of l distinct rooms/locations where the i^{th} room is denoted by level L_i. According to the location fingerprinting principle, each location is expected to be uniquely identified in the signal domain. In other words, each fingerprint has one-to-one mapping to the set of locations, $L = \{L_1, L_2, \ldots, L_l\}$ where $|L| = l$. Let this set of fingerprints, F be denoted by, $F = \{F_1, F_2, \ldots, F_l\}$ where $|F| = l$. Traditionally, if n access points (APs) or anchors are observed at a particular location, L_i, the corresponding fingerprint of L_i in the signal domain can be represented as, $F_i = \{F_i^1, F_i^2, \ldots, F_i^n\}$. The quantity, F_i^j can take the form of a simple average received signal strength (RSS) indication [1] to a histogram representation of different signal levels [22] or even a much complex probabilistic measure [12] of the observed RSSs from AP j, where $j \in \{1, 2, \ldots, n\}$.

Majority of such location fingerprinting techniques utilise the already available wireless communication infrastructure indoors (e.g., Wi-Fi, Bluetooth) in order to build the radio-map, i.e., a collection of $< L_i, F_i >$ tuple obtained from the perceived RSS samples where $i \in \{1, 2, \ldots, l\}$. The conventional way of constructing such radio-map was to laboriously survey the whole localisation area, and collect the RSSs, i.e., F_i's at the points of interests L_i's. The location determination phase consists of first acquiring the fingerprint, by a client device at an "unknown" location. Subsequently, this perceived measurement will be compared against the fingerprints, F_i's of the stored radio-map $< L_i, F_i >$, and the best match will be returned as the corresponding location.

2.2 Probabilistic Localisation

Probabilistic localisation algorithms will return the most likely L_i among the set of training locations/rooms, $L == \{L_1, L_2, \ldots, L_l\}$ where $|L| = l$, given the perceived fingerprint, $S = \{S^1, S^2, \ldots, S^m\}$. The maximum a posteriori (MAP) algorithm is based upon the Naive Bayes classifier that computes $\text{argmax}_i\, P(L_i|S)$, where $P(L_i|S)$ is expressed by the formula,

$$P(L_i|S) = \frac{P(S|L_i)P(L_i)}{\sum_{i=1}^{l} P(S|L_i)P(L_i)}. \tag{1}$$

As commonly seen in the literature [12, 22], the perceived signal strength from a particular AP or anchor can be considered independent from other APs at a location. Subsequently, $P(S|L_i)$ is computed from the training radio-map database as,

$$P(S|L_i) = \prod_{j=1}^{m} P(S^j|L_i). \tag{2}$$

Without loss of generality, if all the locations/rooms are equally likely, then, $P(L_i) = \frac{1}{l}$. By choosing the normalising constant as $\sum_{i=1}^{l} P(S|L_i) = 1$ in (1), the MAP can equivalently be written as,

$$\begin{aligned} \underset{i}{\text{argmax}}\, P(L_i|S) &= \underset{i}{\text{argmax}}\, P(S|L_i) \\ &= \underset{i}{\text{argmax}} \prod_{j=1}^{m} P(S^j|L_i). \end{aligned} \tag{3}$$

2.3 Entropy and Information Content

Entropy expresses the measure of uncertainty. For a continuous probability distribution, $p(x)$ of a random variable x, its entropy is defined as, $H = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx$. Thus, minimising the maximum conditional probability distribution (3)'s entropy will correspond to: given the true measurement, the computed location estimation will be the least random. In other words, if we could reduce the uncertainty in RHS of (3), fingerprinting based algorithms are likely to produce more accurate estimations. Our filtering algorithm for crowdsourced measurements that we discuss in the next section is motivated by this. In order to add a crowdsourced measurement, we first compute the entropy of the resulting signal strength's probability distribution at the claimed location after its incorporation, and compare it with its existing entropy. We only accept the measurement if the resulting entropy is smaller. In other words, we discard any measurement, the incorporation of which increases the uncertainty in (2)'s modelling.

In order to derive the cumulative entropy of all the observed APs' signal strength distributions at a particular location, we first present the differential entropy expression considering only one. In localisation literature, a single AP j's signal strength distribution, $P(S^j|L_i)$ at a particular location, L_i is generally assumed to be normally distributed supported by experimental results [11, 12]. We also follow this claim. If $P(S^j|L_i) \sim N(\mu_j, \sigma_j)$, then a normal distribution's differential entropy expression can be directly used to represent the entropy of AP j's signal strength distribution as follows, $H = -\int_{-\infty}^{\infty} (2\pi\sigma_j^2)^{-\frac{1}{2}} e^{\frac{-(x-\mu_j)^2}{2\sigma_j^2}} \ln [(2\pi\sigma_j^2)^{-\frac{1}{2}} e^{\frac{-(x-\mu_j)^2}{2\sigma_j^2}}] dx$. By simplifying the RHS using the identities, $\int_{-\infty}^{\infty} (2\pi\sigma_j^2)^{-\frac{1}{2}} e^{\frac{-(x-\mu_j)^2}{2\sigma_j^2}} dx = 1$, and $\int_{-\infty}^{\infty} (2\pi\sigma_j^2)^{-\frac{1}{2}} (x-\mu_j)^2 e^{\frac{-(x-\mu_j)^2}{2\sigma_j^2}} dx = \sigma_j^2$, we obtain,

$$H = \frac{1}{2} \ln (2\pi\sigma_j^2) + \frac{1}{2} = \frac{1}{2} \ln (2\pi e \sigma_j^2). \quad (4)$$

Using (4), the differential entropy of n-dimensional Gaussian probability densities which is the entropy of RHS of (2) is computed as,

$$H_n = \frac{n}{2} \ln 2\pi e (\sigma_1^2 \sigma_2^2 \ldots \sigma_n^2)^{\frac{1}{n}}. \quad (5)$$

This is shown in [15] by McEliece. A simplified derivation of (5) can be provided based on the independence assumptions of the observed signal strengths from the APs at a location [12, 22], and the property that the differential entropy of n independent Gaussian variables is the sum of their individual entropy values, i.e.,

$$\begin{aligned} H_n &= \frac{1}{2} \ln (2\pi e \sigma_1^2) + \frac{1}{2} \ln (2\pi e \sigma_2^2) + \ldots + \frac{1}{2} \ln (2\pi e \sigma_n^2) \\ &= \frac{1}{2} \ln \left\{ (2\pi e)^n \sigma_1^2 \sigma_2^2 \ldots \sigma_n^2 \right\} \\ &= \frac{n}{2} \ln 2\pi e (\sigma_1^2 \sigma_2^2 \ldots \sigma_n^2)^{\frac{1}{n}}. \end{aligned}$$

2.4 Filtering Approach Based on Entropy

We conceptualise a filtering technique for crowdsourced measurements based on information content. Let us assume that incorporating a measurement,S received at a certain time, t results in the differential entropy, $H_n^{'}$ of the probability distribution of the observed signal strength at the claimed location. If the original differential entropy without this contribution is denoted by, $H_n = \frac{n}{2} \ln 2\pi e (\sigma_1^2 \sigma_2^2 \ldots \sigma_n^2)^{\frac{1}{n}}$, then the overall filtering algorithm works in two steps as follows: i) check whether a participant's measurement's input time, t is within the time constraint $t_s \leq t \leq t_f$, where t_s and t_f may be the starting and finishing time of his/her working hours, respectively, and ii) for all measurements satisfying the time constraint mentioned in (i), compute $H_n^{'}$, and

$$< L_i, S >= \begin{cases} \text{accept}, & \text{if } H_n^{'} < H_n \\ \text{reject}, & \text{otherwise} \end{cases} \quad (6)$$

The algorithm operates according to two constraints: i) a time constraint, and ii) an entropy constraint based on our previous section's discussion. The time constraint follows the idea that if an occupant's submitted measurement comes at a different time other than his/her expected location's timing, it is not accepted. The entropy constraint ensures that only the good quality crowdsourced measurements will be incorporated but the inappropriate ones will be discarded. In other words, only the measurements that reduces the uncertainty of the signal strength distribution at the claimed location inside the fingerprinting radio-map will be accepted.

Fig. 1 depicts our overall entropy based fingerprinting localisation approach. The crowdsourced measurements from participants' smartphones are collected and stored inside a central server. Each submitted measurement takes the form of an expected location at a time that may come from the participant's timetabling information, and the observed Wi-Fi signal strengths from the perceived APs during that time. The "Entropy-based filtering" entity consists of the algorithm that we discuss in this section. Its detailed algorithmic description that we implement is omitted for brevity. If the measurement is passed by this filtering entity, it is then fed into building the machine learning model's fingerprinting radio-map of the claimed location. During run-time or location determination phase, the collected measurement is used as input for the machine learning model's reasoning to obtain the location. Note that, for our evaluation of whether the filtering algorithm is efficient or not, we stored all the measurements irrespective of whether it is filtered or not. Hence, the "Entropy-based filtering" entity is followed by the central server storage in Fig. 1. For practical deployments, it should generally appear before the operation of storing the measurements once the effectiveness of the filtering algorithm is proven. Consequently, only the good quality measurements will be stored.

2.5 Accept and Reject Scenarios

In this section, we will discuss a series of accept and reject scenarios for our filtering algorithm's entropy constraint (6). We provide proofs as to why measurements from certain scenarios should be accepted or rejected with intuitive explanation. They will later be supported by our experimental results in Section 4.

Lemma 2.1. *If $F_i = \{F_i^k\}$, $k = \{1, 2, \ldots, n\}$ denotes the existing signal strength distribution of n APs at the claimed location, L_i, the measurement $S = \{S^j\}$, $\forall_{j \in \{1,2,\ldots,m\}} (j \neq k)$ will always be rejected.*

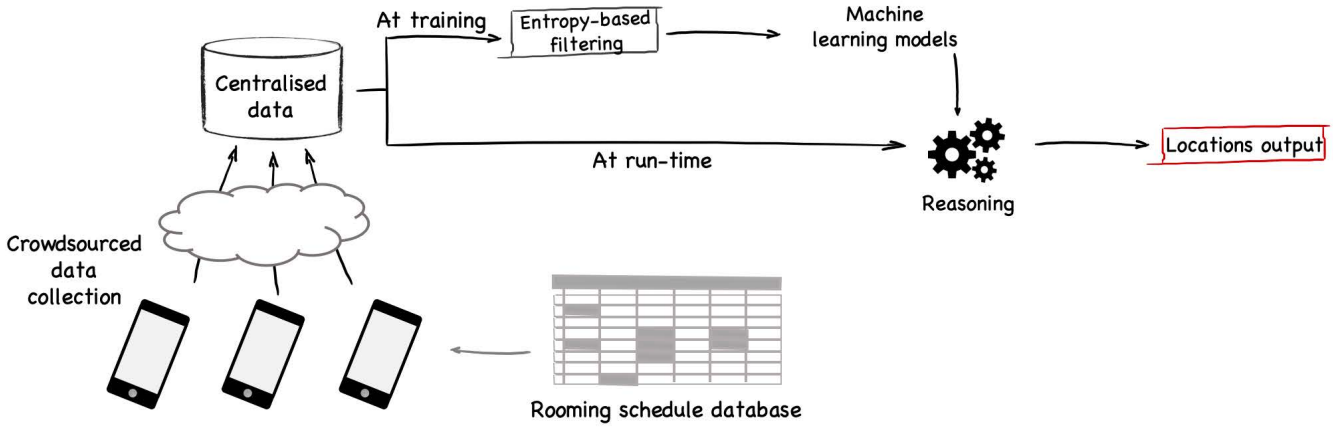


Figure 1: Our entropy-based fingerprinting localisation approach

Proof. In this scenario, none of the claimed observed APs of the measurement, $S = \{S^1, S^2, \ldots, S^m\}$ appears in the existing fingerprint of the location, L_i. The differential entropy after incorporating this measurement is represented as,

$$\begin{aligned}
H_n^{'} &= H_{n+m} \\
&= \frac{n+m}{2} \ln 2\pi e(\sigma_1^2\sigma_2^2 \ldots \sigma_n^2\sigma_{n+1}^2 \ldots \sigma_{n+m}^2)^{\frac{1}{n+m}} \\
&= \frac{n+m}{2} \ln 2\pi e(\sigma_1^2\sigma_2^2 \ldots \sigma_n^2)^{\frac{1}{n+m}} + \frac{1}{2} \ln (\sigma_{n+1}^2 \ldots \sigma_{n+m}^2) \\
&= \frac{n+m}{2} \ln 2\pi e(\sigma_1^2\sigma_2^2 \ldots \sigma_n^2)^{\frac{1}{n}} - \frac{m}{2n} \ln (\sigma_1^2\sigma_2^2 \ldots \sigma_n^2) \\
&\quad + \frac{1}{2} \ln (\sigma_{n+1}^2 \ldots \sigma_{n+m}^2) \\
&= H_n + \frac{m}{2n} \ln (2\pi e)^n (\sigma_1^2\sigma_2^2 \ldots \sigma_n^2) - \frac{m}{2n} \ln (\sigma_1^2\sigma_2^2 \ldots \sigma_n^2) \\
&\quad + \frac{1}{2} \ln (\sigma_{n+1}^2 \ldots \sigma_{n+m}^2) \\
&= H_n + \frac{m}{2} \ln (2\pi e) + \frac{1}{2} \ln (\sigma_{n+1}^2 \ldots \sigma_{n+m}^2) \\
&> H_n,
\end{aligned}$$

Consequently, the input measurement, S will be rejected by (6). □

Since the measurement may be produced in an automated and arbitrary manner, it is unlikely to include any AP that was observed at the same location inside the existing fingerprinting radio-map. Therefore, this type of measurement should not be accepted.

Lemma 2.2. *A measurement $S = \{S^1, S^2, \ldots S^m\}$ will be rejected (accepted) if after incorporation, at least one of the AP's signal strength's deviation (improvement) is more from its previously stored distribution, while the rest remains the same.*

Proof. Suppose, the j^{th} AP's signal strength's deviation is more than its stored distribution, i.e., $\sigma_j^{'2} > \sigma_j^2$, while for the rest, they remain the same, i.e., $\forall_{i \in \{1,2,\ldots n\} \setminus \{j\}} (\sigma_i^{'2} = \sigma_i^2)$. Subsequently, it can be proved that S will be rejected by (6) as follows,

$$\begin{aligned}
H_n^{'} &= \frac{n}{2} \ln 2\pi e(\sigma_1^2\sigma_2^2 \ldots \sigma_j^{'2} \ldots \sigma_n^2)^{\frac{1}{n}} \\
&> \frac{n}{2} \ln 2\pi e(\sigma_1^2\sigma_2^2 \ldots \sigma_j^2 \ldots \sigma_n^2)^{\frac{1}{n}}, \text{since, } \ln x \text{ is a} \\
&\quad \text{monotonically increasing function for } x > 0. \\
&= H_n.
\end{aligned}$$

The accept scenario's proof is exactly the same as above, where $\sigma_j^{'2} < \sigma_j^2$ which results in $H_n^{'} < H_n$. □

It was discussed in Section 2.3 that a measurement is accepted only if it reduces a particular location's overall signal strength distribution's uncertainty. Lemma 2.2 imposes a strict rejection constraint upon the measurement which takes into consideration that an intruder may snoop the signal strength, thereby gaining knowledge about the signal map of that particular indoor location. He/she may then submit tampered measurement to corrupt the fingerprinting radio-map. Incorporation of it will likely result in deviation from the previously stored fingerprints. This is prevented since the filtering approach rejects any measurement that causes deterioration in regard to even one AP's stored distribution while the rest remains the same.

In the above, we discussed a specific scenario where the intruder deliberately attempts to corrupt any particular AP's or a group of APs' fingerprints inside the training radio-map assuming the rest will remain the same. Next, we derive a generalised expression denoting the level of manipulation required by the intruder so that his/her malicious measurement is accepted. For this to happen, it can be shown using (5) that the impact of deterioration of a few APs' fingerprints should be offset by the improvement of a few others through manipulation of the measurement perceived at the location. Suppose among n APs' signal distribution model stored as a location fingerprint, I of them were improved, J were deteriorated, and the rest remained the same. In other words, $\forall_{i \in I} (\sigma_i^{'2} < \sigma_i^2)$, $\forall_{j \in J} (\sigma_j^{'2} > \sigma_j^2)$, and $\forall_{k \in \{1,2,\ldots n\} \setminus \{I \cup J\}} (\sigma_k^{'2} = \sigma_k^2)$.

Lemma 2.3. *The magnitude of allowed deviation of J APs' stored signal strengths' distributions is bounded by the I APs' achieved improvement by incorporating the same fingerprint, i.e.,* $\frac{\sigma'^2_{I+1}}{\sigma^2_{I+1}}\frac{\sigma'^2_{I+2}}{\sigma^2_{I+2}}\dots\frac{\sigma'^2_{I+J}}{\sigma^2_{I+J}} < \frac{\sigma^2_1}{\sigma'^2_1}\frac{\sigma^2_2}{\sigma'^2_2}\dots\frac{\sigma^2_I}{\sigma'^2_I}$.

Proof. According to our algorithm, an input measurement is accepted iff, $H'_n < H_n$.

$$\Rightarrow \frac{n}{2}\ln 2\pi e(\sigma'^2_1\dots\sigma'^2_I\sigma'^2_{I+1}\dots\sigma_{I+J}{}'^2\sigma'^2_{I+J+1}\dots\sigma'^2_n)^{\frac{1}{n}}$$
$$< \frac{n}{2}\ln 2\pi e(\sigma^2_1\dots\sigma^2_I\sigma^2_{I+1}\dots\sigma^2_{I+J}\sigma^2_{I+J+1}\dots\sigma^2_n)^{\frac{1}{n}},$$
$$\Rightarrow (\sigma'^2_1\dots\sigma'^2_I\sigma'^2_{I+1}\dots\sigma'^2_{I+J}) < (\sigma^2_1\dots\sigma^2_I\sigma^2_{I+1}\dots\sigma^2_{I+J}),$$
$$\Rightarrow \frac{\sigma'^2_{I+1}}{\sigma^2_{I+1}}\frac{\sigma'^2_{I+2}}{\sigma^2_{I+2}}\dots\frac{\sigma'^2_{I+J}}{\sigma^2_{I+J}} < \frac{\sigma^2_1}{\sigma'^2_1}\frac{\sigma^2_2}{\sigma'^2_2}\dots\frac{\sigma^2_I}{\sigma'^2_I}.$$

□

This implies that on one hand, it will require extensive knowledge of the existing radio-map database on the intruder's part, and on the other hand, it will limit the magnitude of deviation from the original fingerprint that can be caused. Additionally, even if the intruder was successful, the negative impact can be offset by subsequent good quality measurements by others at the same location.

Lemma 2.4. *A measurement* $S = \{S^1, S^2, \dots S^n, S^{n+1}\}$ *with a newer* $(n+1)^{th}$ *AP's reading at a location will be accepted under the following condition,* $\sigma^2_{n+1} < \frac{1}{2\pi e}\frac{\sigma^2_1}{\sigma'^2_1}\frac{\sigma^2_2}{\sigma'^2_2}\dots\frac{\sigma^2_n}{\sigma'^2_n}$.

Proof. Incorporating the measurement S, the resulting entropy is, $H_{n+1} = \frac{n+1}{2}\ln 2\pi e(\sigma'^2_1\sigma'^2_2\dots\sigma'^2_n\sigma_{n+1}{}^2)^{\frac{1}{n+1}}$. S is accepted iff,

$$\frac{n+1}{2}\ln 2\pi e(\sigma'^2_1\dots\sigma'^2_n\sigma_{n+1}{}^2)^{\frac{1}{n+1}} < \frac{n}{2}\ln 2\pi e(\sigma^2_1\sigma^2_2\dots\sigma^2_n)^{\frac{1}{n}},$$
$$\Rightarrow \sigma^2_{n+1} < \frac{1}{2\pi e}\frac{\sigma^2_1}{\sigma'^2_1}\frac{\sigma^2_2}{\sigma'^2_2}\dots\frac{\sigma^2_n}{\sigma'^2_n}. \quad (7)$$

□

Eq. (7) gives an idea of the initial sample's variance to be set which is influenced by the improvement achieved from other n APs' distributions. For example, if $\sigma^2_{n+1} = 1$, the required improvement should be greater than $2\pi e$ for the measurement to be accepted. We need to carefully consider this scenario as it influences how missing APs from the stored distribution can be part of the actual fingerprint. Following (7), it is straightforward to show that for m new APs to be integrated through a measurement, the improvement required is greater than $(2\pi e)^m$. This also ensures the crowdsourcing mechanism of creating fingerprinting radio-map evolves over time while still being adaptable to environmental changes.

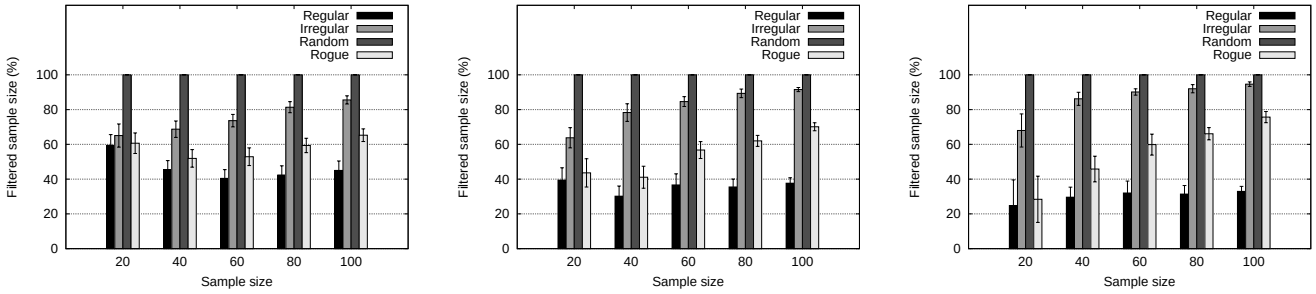
3 RELATED WORK

The field of crowdsourced indoor positioning has received considerable attention over the last few years. Most of the related research focuses on increasing accuracy by optimising the reasoning approach, for example through ensemble learning [21], collaborative sensing between nearby devices [13] or activity detection [23], as well as on increasing efficiency [5] and reducing computational complexity [24].

Here, our focus instead is on filtering out unreliable data at the labelling stage. The handling of unreliable data labelling is a key challenge not only in crowdsourced indoor positioning but more generally in all participatory sensing applications. For example, Barnwal et al. [2] have followed a Bayesian approach to enhancing the reliability of a vehicular participatory sensing system. The rationale is that confidence can be estimated based on the conditional probability of occurrence of a particular traffic event at a particular location given that supporting reports have been generated. Also, Gisdakis, Giannetsos and Papadimitratos [9] have proposed a comprehensive framework that is agnostic of the cause of a faulty measurement. Each report is transformed into a probability mass, so as to compute the hypothesis with the maximum belief; the belief corresponding to this hypothesis; and the local conflict of the probability mass, as per Dempster-Shafer Theory. Its output is a partitioning into inliers and outliers, which is dependent on the existence of an 'honest majority'. The system then compares the similarity between the inlying reports of two neighbouring units with a two-sample Kolmogorov-Smirnov test. It uses a merging and training phase, followed by an ensemble of machine learning classifiers to characterise incoming reports as inliers or outliers, and a concept drift detection module to detect changes in the statistical properties of the sensed phenomenon. The framework has been evaluated on environmental monitoring.

Specifically for indoor positioning, Li et al. [14] have proposed defences for different adversary models and attacks. Their logic is that an initial set of measurements from trusted users can be used to infer the trustworthiness of the fingerprints submitted by unknown users. The authors have used two metrics to evaluate trustworthiness and a corresponding iterative algorithm to build a reliable fingerprint radio-map in the presence of unreliable reports. The first metric is the temporal correlation within an RSS trace, as fingerprints collected by different users tend to exhibit a similar RSS trend (e.g., when the user walks towards an AP, the RSS increases, and when the user walks away, it decreases). The second is the spatial likelihood, which captures the spatial RSS correlation between the fingerprints from the same position in different traces. However, the defences proposed assume that there are always some users that can be trusted, such as the employees in a shopping mall, and this may not always be the case. For example, an employee may have reasons to want to hide their location at a specific point in time.

Cheng et al. [4] have proposed a technique for addressing the challenge of missing values in participatory sensing. The key idea is to employ the spatio temporal compressive technique originally proposed in [18] to reconstruct the sensory data given an incomplete and partially inaccurate dataset if the sensory data being reconstructed exhibit low-rank structure and spatio-temporal properties. One of the two case studies evaluated is crowdsourced Wi-Fi fingerprinting. 10 users equipped with smartphones were asked to walk through a university campus for two hours. Their technique



(a) 25%-75% split. (b) 50%-50% split. (c) 75%-25% split.

Figure 2: Performance of filtering approach while incorporating different types of participants' measurements

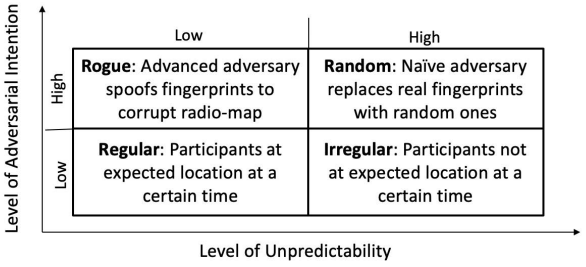


Figure 3: The four types of participants considered

involves inferring the smartphones' proximity based on other multidimensional sensor readings, and to derive a corresponding spatial constraint. This was feasible because the users are non-stationary and specifically tasked with the work of fingerprinting, so that a sensor node could obtain substitute sensor readings from the next time slot.

Zhou et al. [25] have proposed a minimax conditional entropy principle to infer ground truth from noisy crowdsourced labels. Based on this, they derived a unique probabilistic labelling model jointly parameterised by worker ability and item difficulty. This is the only known example of work in the literature that has proposed to benefit from entropy for identifying unreliable crowdsourced labels. However, it has not been evaluated on dataset related to location fingerprinting.

The above solutions proposed in the literature for handling unreliable data in participatory sensing either have not been designed and evaluated for indoor positioning applications or assume that volunteers are tasked with walking through areas with the purpose to collect series of spatio-temporal data that can be cross-checked for their veracity, or that there exist users whose measurements can always be considered as trusted. In our work, we do not need to record users' movement across different locations other than their destination as expected by their pre-defined timetable, and we also do not assume the trustworthiness of a select set of users. Next, we present the experimental evaluation of our approach.

4 EVALUATION

4.1 Experimental Setup and Participant Groups

We collected measurements from seven rooms of a building of our university campus where four rooms are on one side and the rest are on the other side divided by a corridor. Each room has the dimension of 7.85m×3.8m. We involved lecturer volunteers who are the users of those rooms. Their timetabling information were preloaded in a smartphone application that was given to them. The smartphone's application perceives the Wi-Fi signal strength, and correlates it with the location retrieved from the timetabling information by the software running inside the particular volunteer's smartphone, and sends it to a central server. All measurements satisfying the time constraint as discussed in Section 2.4 are stored.

The crowdsourced measurements do not require the participants to explicitly indicate their locations where they are taken, and they can be oblivious of the data collection procedure. In order to provide supporting results for the proofs of Section 2.5, we first discuss four different types of participants based on the scenarios (Fig. 3), and then describe how we emulate their measurements:

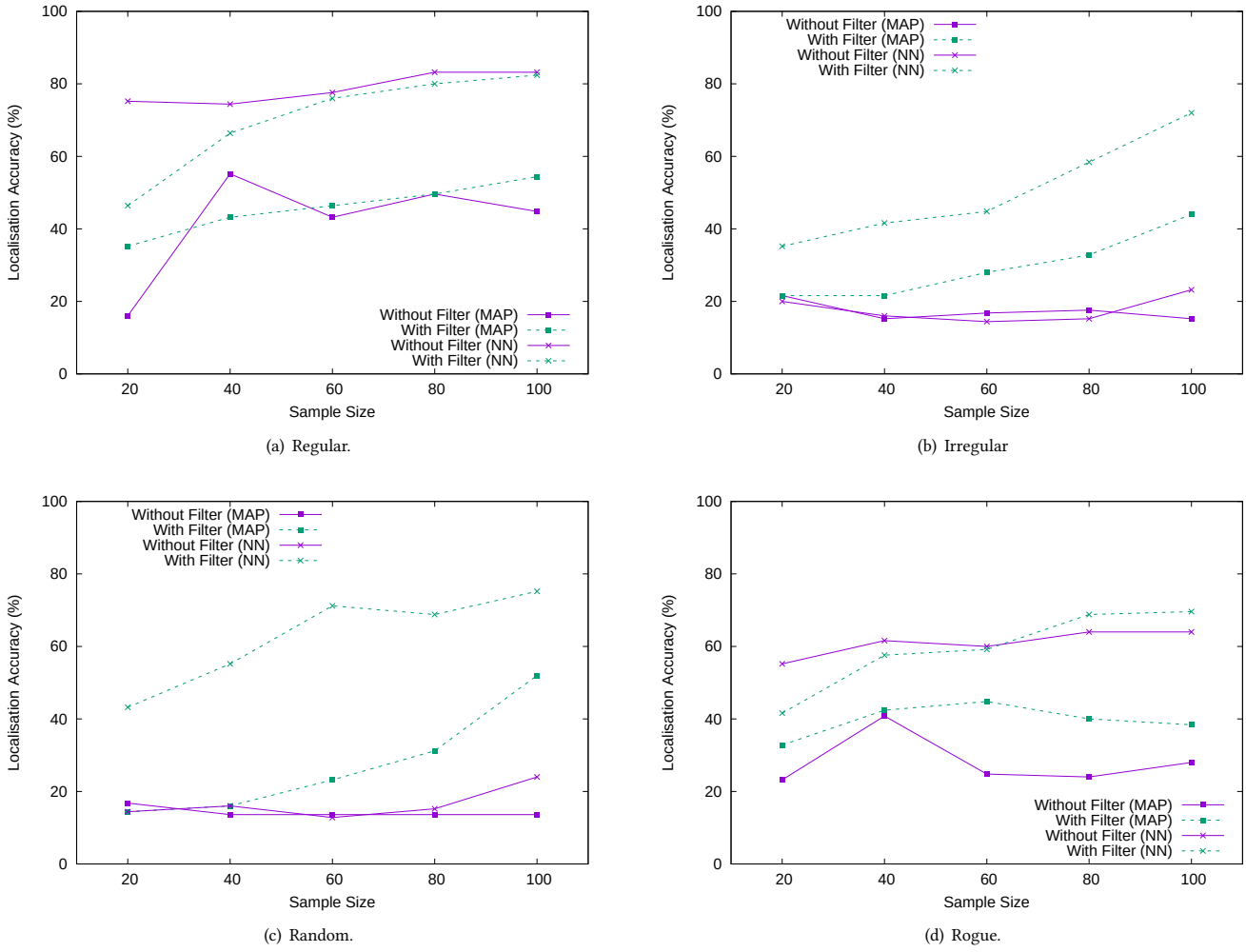
i) **Regular:** participants who remain at their expected locations at the time their devices submit the measurements,

ii) **Irregular:** participants who are not at their timetabled locations during submission,

iii) **Random:** adversarial participants who wish to hide their location by generating automated or arbitrary measurements that do not correlate with the indoor environment's geometry and communication infrastructure, and

iv) **Rogue:** adversarial participants who intentionally try to corrupt the radio-map database through tampered measurements.

All the collected measurements in our experimental setup are considered to be input by regular participants. The measurements from the other participants are emulated by manipulating a regular participant's measurement as follows. Suppose, $< L_i, F_i, S >$ represent the <location, stored fingerprint, measurement> at the claimed location L_i where $i \in \{1, 2, \ldots l\}$. An irregular participant's measurement $< L_j, S >$ is emulated by selecting a location L_j from a uniform distribution of the available locations, $\{L_j\}$, $j \in \{1, 2, \ldots l\} \setminus \{i\}$, that is different from L_i. In order to create a random participant's measurement, we first select the number of arbitrary APs, x from $U(1, m)$ where $U(.)$ denotes a uniform distribution over the range. We chose $m = n$ where n is the number of APs observed at L_i, and $\forall_{j \in \{1,2,\ldots x\}}(j \notin \{1, 2, \ldots n\})$. Then, each of the x signal strengths is generated from $U(\mathrm{RSS}_{\min}, \mathrm{RSS}_{\max})$. In our experiments, we set $\mathrm{RSS}_{\min} = -90$ dBm, and $\mathrm{RSS}_{\min} = -30$ dBm. To



(a) Regular.

(b) Irregular

(c) Random.

(d) Rogue.

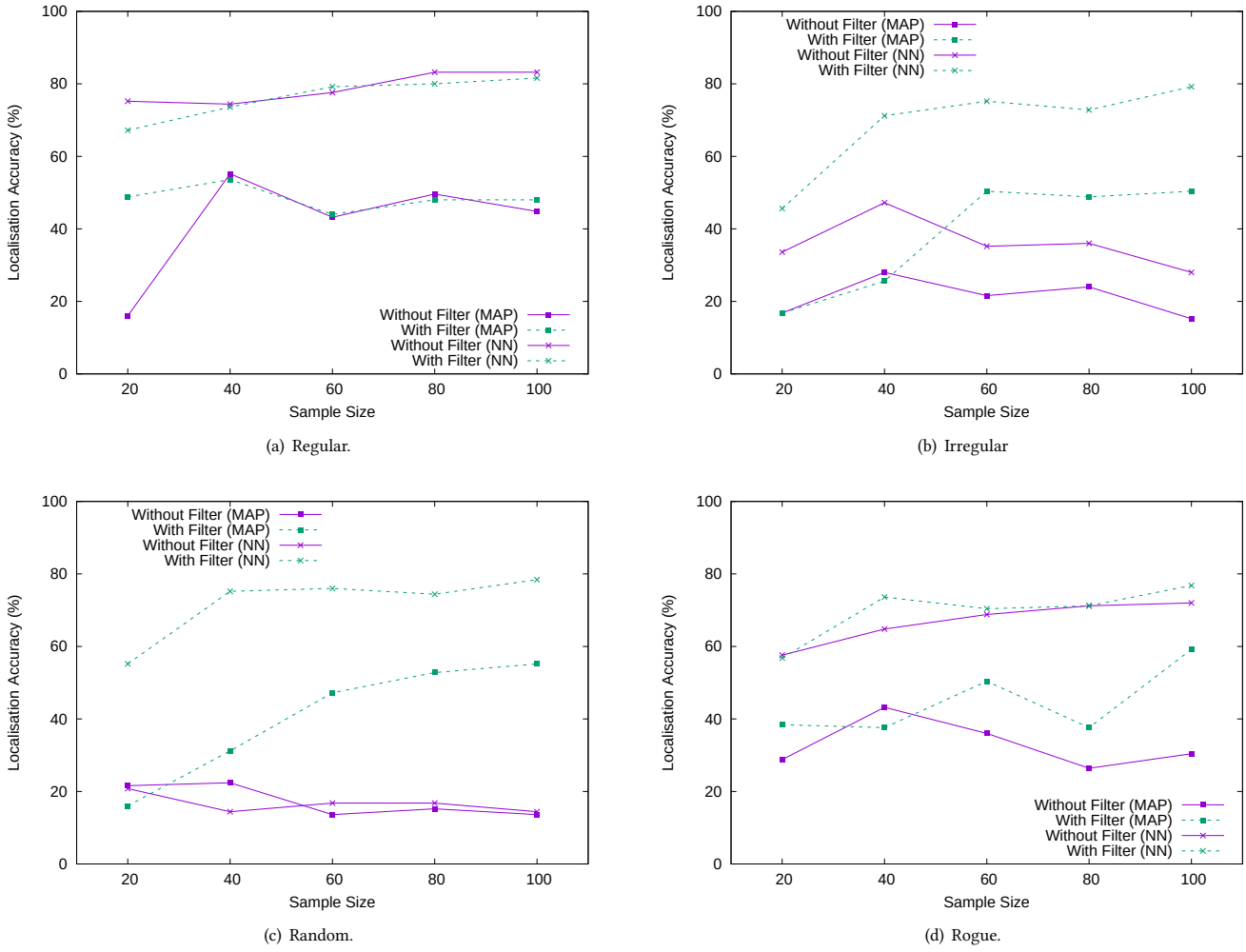
Figure 4: 25%-75% split between already existing and incorporating different types of participants' measurements

emulate a rogue participant's measurement, we first pick a number x form $U(1, n)$, and then choose a set of x indices again from $U(1, n)$. Then, each $j \in \{1, 2, \ldots x\}$ AP's signal strength is selected from a Gaussian distribution $N(S^j, \sigma^2)$ where we change σ^2 to control the deviation of noise.

4.2 Results and Discussion

125 measurements were recorded from our volunteers across seven rooms. We randomly divided them into 5 sets of 25 measurements each. 5-fold cross validation was used where 4 sets (100 measurements in total) were used as fingerprinting radio-map (training) in each fold, and the remaining (25 measurements) were used as testing samples.

With the first set of experiments, we aim to show the effectiveness of our filtering approach discussed in Section 2. For this purpose, we assume that there are already some existing measurements inside the fingerprinting radio-map. We consider three cases where the training samples were separated between already existing and the participants' contributions. 25%–75%, 50%–50% and 75%–25% depict the separation between already existing and participants' contributions, respectively. Five different training sample points, 20, 40, 60, 80 and 100 are considered for each separation. The different participants' measurements were modelled following the previous section's discussion. Fig. 2 is constructed as the average of 10 experimental runs with 95% confidence interval. Each run constitutes an instance of 5-fold cross validation. Our filtering approach's effectiveness can be seen from the results of Fig. 2. 100% of the Random participants' measurements were filtered. This directly follows Lemma 2.1. Another observation is that the filtering approach's performance improves as the sample size increases concerning both Irregular and Rogue measurements. This can be perceived for all three separations. This is intuitive since a larger sample size is expected to model the fingerprinting radio-map with less uncertainty which in turn will improve the filtering performance. Furthermore, the incorporation of regular measurements remain steady across various sample sizes and different separations. This is an important characteristics, because the filtering approach ensures good quality measurements are accepted, and also, it is not



(a) Regular.

(b) Irregular

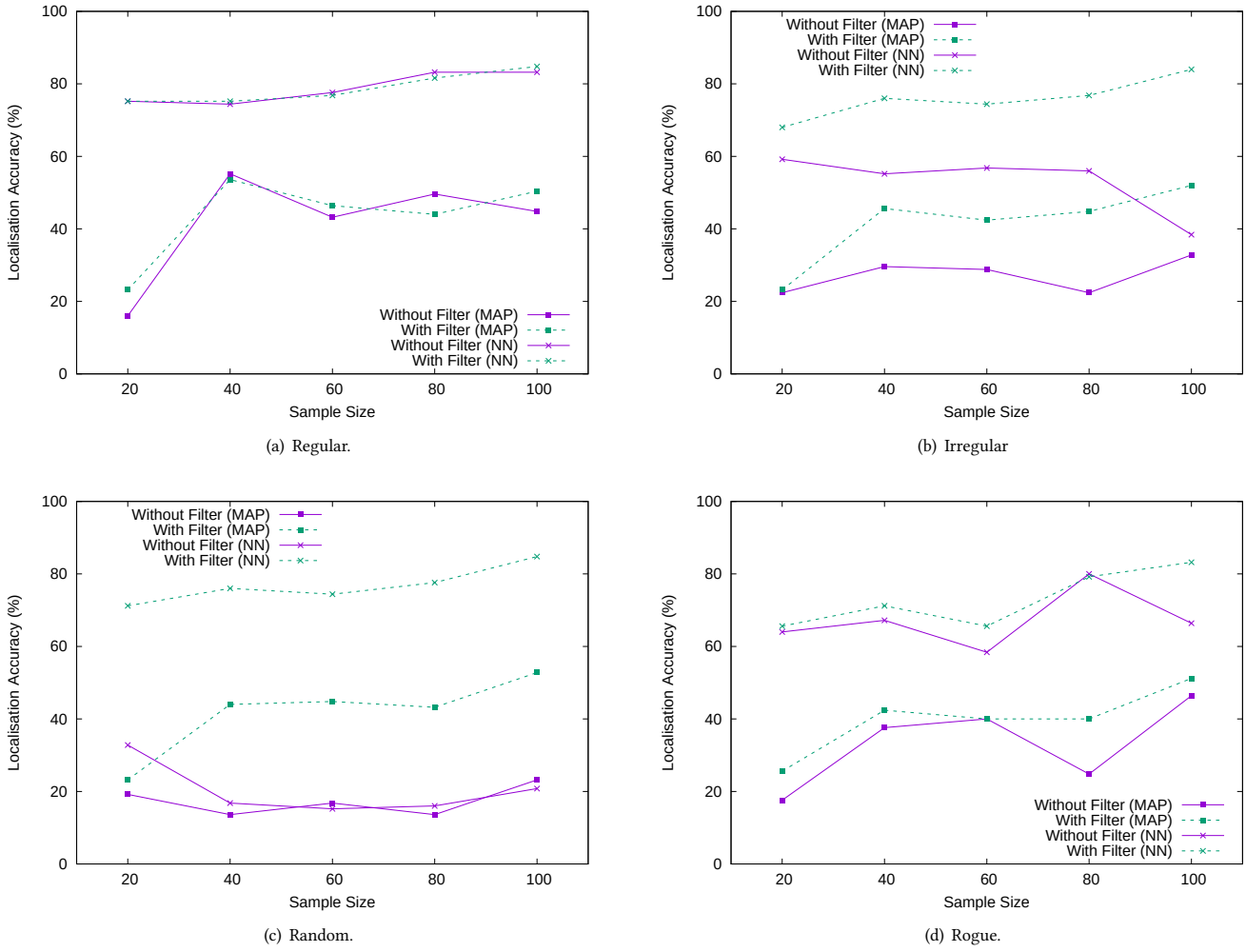
(c) Random.

(d) Rogue.

Figure 5: 50%-50% split between already existing and incorporating different types of participants' measurements

overly restrictive. This follows the property of Lemma 2.2 and 2.3. One may argue that why 100% regular measurements were not accepted by the filtering approach. It is a well-known phenomenon in localisation literature that even at the same location, the perceived signal strength may vary due to environmental factors, device heterogeneity, and also the time of the day [11]. This justifies a proportion of regular measurements being filtered. We argue that as long as a steady stream of good quality of regular measurements are ensured to be incorporated, the fingerprinting radio-map will evolve over time. This is the case as can be seen in Fig. 2. With 75% existing training samples, the least number of regular measurements are discarded which is again intuitive since the filtering approach's modelling is based upon a larger sample size compared to the 25% and 50% ones.

For the second set of experiments, we retained the same separation across similar training sample points as the previous one. Two well-known machine learning algorithms such as Nearest Neighbour (NN) and maximum a posteriori (MAP) are then applied. For comparison, we considered both 'with filter' and 'without filter' training dataset, where one results from applying the filtering approach, and the other consists of all the measurements with complete trust. The testing dataset comprises of 25% measurements in each fold of 5-fold cross validation as described in the beginning of this section. The results of the two algorithms' are presented in Fig. 4, 5, and 6. In general, for all combinations, both algorithms' localisation accuracy is better for 'with filter' variant than its 'without filter' counterpart. This is evident more when the sample size increases. This directly follows from our previous experiment's results too since the filtering approach performed better with larger sample size, and also for 75%–25% separation which consequently gave rise to a more accurate radio-map for the machine learning algorithms. This leads to another observation that irrespective of the different types of participants' measurements, the localisation accuracy reached similar levels for both algorithms (see 100 training sample points' results for the four different types of participants' measurements of Fig. 6). Also, we expect the results based on Regular participants' measurements for 'with filter' variant should generally follow the trend of its 'without filter' counterpart which is



(a) Regular.

(b) Irregular

(c) Random.

(d) Rogue.

Figure 6: 75%-25% split between already existing and incorporating different types of participants' measurements

generally observed in Fig. 4(a), 5(a) and 6(a). These findings validates our claim that our entropy based fingerprinting approach can result in an effortless and scalable IPS for a public or commercial building.

We conclude this section with a few more details. While it can be argued that NN is another form of MAP, we have utilised deterministic average RSS as NN's fingerprint, and applied Euclidean distance between fingerprints for location estimation decision. MAP is implemented following Section 2.2's model. This might be the reason for inferior performance of MAP compared to NN in our experiments, where MAP generally requires a significant number of samples for its fingerprint modelling. The number of samples per room ($\approx$ 18) in our experiments was relatively small. Both algorithms were implemented with efficient data structure, and have run-time complexity of $O(nl)$ where l is the number of locations, and n is the dimension of the fingerprint at each location. NN provided better localisation accuracy with almost 85% correct detection of rooms. We observed more than 300 different Wi-FI APs in total within just one premise during our data collection process. Our university wireless network providers are only considered which is a natural localisation choice for any particular commerical or public building that reduces this number to 125. However, only on-demand availability for energy conservation purpose, heterogeneity of mobile devices with varying capability to scan the nearby APs, and spatio-temporal factor result in variability in the number of APs observed at a certain location. This can give rise to missing RSS phenomenon of fingerprinting techniques [7, 10] that we perceive in our radio-map as well. We believe this also has adverse impact on both the algorithms' offered localisation accuracy since we adopt an elementary imputation practice that substitutes the missing value with the minimum RSS (e.g., -96 dBm). There are multiple research work as in [7, 10] that try to resolve this missing RSS phenomenon, which we consider to be out of scope for our work. Because, our crowdsourced fingerprinting approach's benefit is independent of the choice of the machine learning algorithm, and any other improvements that they may be incorporated. This claim follows from our observation that in all our experiments,

the 'with filter' variant is generally better than its 'without filter' counterpart. We contend that by adopting an appropriate missing value resolution techinque, and considering more advanced machine learning algorithms is likely to offer better localisation accurary compared to the two simplistic ones that we considered here. For the presented results concerning the Rogue participants, we fixed the Gaussian noise deviation, σ of its modelling (see Section 4.1) to be 20. We observed that for lower values than 20, it performs almost like the Regular participants' contributions that were even better. This is quite intuitive looking at Fig. 3 since lower σ will shift the participants' level of adversarial tendency from high to low. For higher values of σ, more measurements are required to achieve the similar accuracy as the presented ones since higher proportion of them are filtered by our algorithm. Those results are omitted for brevity.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a scalable and effortless fingerprinting-based IPS in the context of a public/commercial building by leveraging the existence of a "stationary" crowd, and correlating their smartphones' Wi-Fi sensor readings with their timetabling information. Both our analysis and experimental results have shown that, regardless of such participants' unpredictable behaviour, i.e., not following their timetabling information, our entropy based filtering approach ensures the creation of fingerprinting radio-map incrementally from their measurements. We considered four types of participants' behaviours to support our claim. The localisation performance of two machine learning algorithms was evaluated based on the created fingerprinting radio-map which has shown our approach's effectiveness.

By having provided a practical means for introducing participatory location fingerprinting through the stationary crowd of a commercial or public building, we anticipate the generation of several future work directions. For example, we have assumed a few measurements to exist inside the fingerprinting radio-map in all scenarios (i.e., the 25%-75%, 50%-50% and 75%-25% separations) of our filtering approach's experimental evaluation. The creation of radio-map from scratch with no existing fingerprint will require modifications to our current filtering approach so that the few initial measurements are integrated only after careful consideration, i.e., imposing additional constraints. More experiments with different public or commercial building setup and size other than a university campus can be conducted to establish applicability in very large indoor areas and involving large crowd. Also, the radio-map created following our approach could easily be applied to train different families of machine learning models, and subsequently compare their localisation performance with finer granularity. Finally, a rigorous theoretical framework can be pursued to show that the entropy based filtering approach can incrementally create the training radio-map. In this paper, experimental validation was provided together with the relevant lemmas with proofs.

ACKNOWLEDGMENTS

We would like to thank Malik Shams for the creation of the smartphone application, and the server-side program for data collection, and storage, and Sameet Sidhu for contributing to the data collection process.

REFERENCES

[1] P. Bahl and V. N. Padmanabhan. 2000. RADAR: An In-Building RF-Based User Location and Tracking System. In *Proc. IEEE INFOCOM.* Tel Aviv, Israel, 775–784.
[2] Rajesh P Barnwal, Nirnay Ghosh, Soumya K Ghosh, and Sajal K Das. 2016. Enhancing Reliability of Vehicular Participatory Sensing Network: A Bayesian Approach. In *Proc. of IEEE SMARTCOMP.* 1–8.
[3] C. Cai, X. Ma, M. Hu, Y. Yang, Z. Li, and J. Liu. 2018. SAP: A Novel Stationary Peers Assisted Indoor Positioning System. *IEEE Access* 6 (2018), 76475–76489.
[4] Long Cheng, Jianwei Niu, Linghe Kong, Chengwen Luo, Yu Gu, Wenbo He, and Sajal K Das. 2017. Compressive sensing based data quality improvement for crowd-sensing applications. *Journal of Network and Computer Applications* 77 (2017), 123–134.
[5] Ka-Ho Chow, Suining He, Jiajie Tan, and S-H Gary Chan. 2019. Efficient Locality Classification for Indoor Fingerprint-Based Systems. *IEEE Transactions on Mobile Computing* 18, 2 (2019), 290–304.
[6] I. Constandache, R.R. Choudhury, and I. Rhee. 2010. Towards Mobile Phone Localization without War-Driving. In *Proc. of IEEE INFOCOM.* 1–9.
[7] C. Feng, W. S. A. Au, S. Valaee, and Z. Tan. 2012. Received-Signal-Strength-Based Indoor Positioning Using Compressive Sensing. *IEEE Transactions on Mobile Computing* 11, 12 (Dec 2012), 1983–1993.
[8] Avgoustinos Filippoupolitis, William Oliff, and George Loukas. 2016. Bluetooth low energy based occupancy detection for emergency management. In *15th International Conference on Ubiquitous Computing and Communications and International Symposium on Cyberspace and Security (IUCC-CSS).* IEEE, 31–38.
[9] Stylianos Gisdakis, Thanassis Giannetsos, and Panos Papadimitratos. 2015. SHIELD: A data verification framework for participatory sensing systems. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks.* ACM, 16.
[10] S. He, W. Lin, and S. . G. Chan. 2017. Indoor Localization and Automatic Fingerprint Update with Altered AP Signals. *IEEE Transactions on Mobile Computing* 16, 7 (July 2017), 1897–1910.
[11] A-K-M-Mahtab Hossain and Wee-Seng Soh. 2015. A survey of calibration-free indoor positioning systems. *Computer Communications* 66 (2015), 1–13.
[12] K. Kaemarungsi and P. Krishnamurthy. 2004. Properties of indoor received signal strength for WLAN location fingerprinting. In *Proc. MobiQuitous'04.* San Diego, CA, 14–23.
[13] Syed Khandker, Joaquín Torres-Sospedra, and Tapani Ristaniemi. 2019. Improving RF Fingerprinting Methods by Means of D2D Communication Protocol. *Electronics* 8, 1 (2019), 97.
[14] T. Li, Y. Chen, R. Zhang, Y. Zhang, and T. Hedgpeth. 2018. Secure crowdsourced indoor positioning systems. In *Proc. of IEEE INFOCOM.* 1034–1042.
[15] R. J. McEliece. 1977. *The Theory of Information and Coding: A Mathematical Framework for Communication.* Addison-Wesley.
[16] N. Priyantha, A. Chakraborty, and H. Balakrishnan. 2000. The Cricket Location-Support System. In *Proc. ACM MobiCom'00.* Boston, MA, 32–43.
[17] P. Robertson, M. Angermann, and M. Khider. 2010. Improving Simultaneous Localization and Mapping for pedestrian navigation and automatic mapping of buildings by using online human-based feature labeling. In *IEEE/ION Position Location and Navigation Symposium (PLANS).* 365–374.
[18] Matthew Roughan, Yin Zhang, Walter Willinger, and Lili Qiu. 2012. Spatio-temporal compressive sensing and internet traffic matrices. *IEEE/ACM Transactions on Networking (ToN)* 20, 3 (2012), 662–676.
[19] Guobin Shen, Zhuo Chen, Peichao Zhang, Thomas Moscibroda, and Yongguang Zhang. 2013. Walkie-Markie: Indoor Pathway Mapping Made Easy. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation.* Berkeley, CA, USA, 85–98.
[20] A. Ward, A. Jones, and A. Hopper. 1997. A new location technique for the active office. *IEEE Personal Communications* 4, 5 (Oct. 1997), 42–47.
[21] Jianghong Yang, Xiaohui Zhao, and Zan Li. 2019. Crowdsourcing Indoor Positioning by Light-Weight Automatic Fingerprint Updating via Ensemble Learning. *IEEE Access* (2019).
[22] Moustafa Youssef and Ashok Agrawala. 2005. The Horus WLAN Location Determination System. In *Proc. of ACM MobiSys.* 205–218.
[23] WenPing Yu, JianZhong Zhang, JingDong Xu, and YuWei Xu. 2019. An accurate indoor map matching algorithm based on activity detection and crowdsourced Wi-Fi. *Science China Technological Sciences* (2019), 1–10.
[24] Caifa Zhou and Andreas Wieser. 2019. Modified Jaccard index analysis and adaptive feature selection for location fingerprinting with limited computational complexity. *Journal of Location Based Services* (2019), 1–30.
[25] Dengyong Zhou, Qiang Liu, John C Platt, Christopher Meek, and Nihar B Shah. 2015. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240* (2015).