



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Ren, Y;Salim, FD;Tomko, M;Bai, YB;Chan, J;Qin, KK;Sanderson, M

Title:

D-Log: A WiFi Log-based differential scheme for enhanced indoor localization with single RSSI source and infrequent sampling rate

Date:

2017-06-01

Citation:

Ren, Y., Salim, F. D., Tomko, M., Bai, Y. B., Chan, J., Qin, K. K. & Sanderson, M. (2017). D-Log: A WiFi Log-based differential scheme for enhanced indoor localization with single RSSI source and infrequent sampling rate. *Pervasive and Mobile Computing*, 37, pp.94-114. <https://doi.org/10.1016/j.pmcj.2016.09.018>.

Persistent Link:

<https://hdl.handle.net/11343/240875>

D-Log: A WiFi Log-based Differential Scheme for Enhanced Indoor Localization with Single RSSI Source and Infrequent Sampling Rate

Yongli Ren^a, Flora Dilys Salim^a, Martin Tomko^b, Yuntian Brian Bai^c, Jeffrey Chan^a, Kai Qin^a, Mark Sanderson^a

^a*School of Science, Computer Science and Information Technology, RMIT University, Victoria 3000, Australia*

^b*Department of Infrastructure Engineering, The University of Melbourne, Victoria 3010, Australia*

^c*School of Math and Geospatial Science, RMIT University, Victoria 3000, Australia*

Abstract

Currently, large amounts of Wi-Fi access logs are collected in diverse indoor environments, but cannot be widely used for fine-grained spatio-temporal analysis due to coarse positioning. We present a Log-based Differential (D-Log) scheme for post-hoc localization based on differentiated location estimates obtained from large-scale Access Point (AP) logs of WiFi connectivity traces, which can be used for data analysis and knowledge discovery of visitor behaviours. Specifically, the location estimates are calculated by utilizing a combination of Received Signal Strength Indicator (RSSI) records from two neighbouring APs. D-Log exploits real-world industry WiFi logs where RSSI data sampled at low rates from single AP sources are recorded in each connectivity trace. The approach is independent of device and network infrastructure type. D-Log is evaluated using WiFi logs collected from controlled environment as well as real-world uncontrolled public indoor spaces, which includes discrete single-AP RSSI traces of around 100,000 mobile devices over a one-year period. The experiment results indicate that, despite of the challenges with the infrequent sampling rate and the limitations of the data that only records RSSI from single AP sources in each instance, D-Log performs comparatively well to the state-of-the-art RSSI-based localization methods and presents a viable alternative for many application areas where high-accuracy positioning infrastructure may not be cost effective or where positioning applications are considered on legacy information infrastructure.

Keywords: RSSI, WiFi log, localization

1. Introduction

The use of a RSSI from multiple WiFi APs to estimate the position of mobile devices in a wireless networked environment is a well established procedure. Three main approaches are commonly used when RSSI traces are available: trilateration, scene analysis (WiFi fingerprinting), and proximity-based localization. Most of these methods aim to generate an accurate estimate of a mobile device's position in the networked environment. Furthermore, these approaches often demand either that the WiFi networks are configured for high sampling rates and continuous monitoring from multiple access points, or require users to install an app on their device for data collection. This leads to implementation barriers such as high setup, engineering, and calibration cost and the requirements for user participation. Hence, there is a need for approaches applicable to low sampling rates and single access point monitoring. Another source of data that has thus far been barely examined for enhancing localization: large volumes of WiFi AP logs of non-continuous WiFi connectivity traces that are normally stored in an external system, representing timestamped connections between a device and a single Access Point, along with the associated RSSI. With such data, a research

Email addresses: yongli.ren@rmit.edu.au (Yongli Ren), flora.salim@rmit.edu.au (Flora Dilys Salim), tomkom@unimelb.edu.au (Martin Tomko), yuntianbrian.bai@rmit.edu.au (Yuntian Brian Bai), jeffrey.chan@rmit.edu.au (Jeffrey Chan), kyleqin2008@gmail.com (Kai Qin), mark.sanderson@rmit.edu.au (Mark Sanderson)

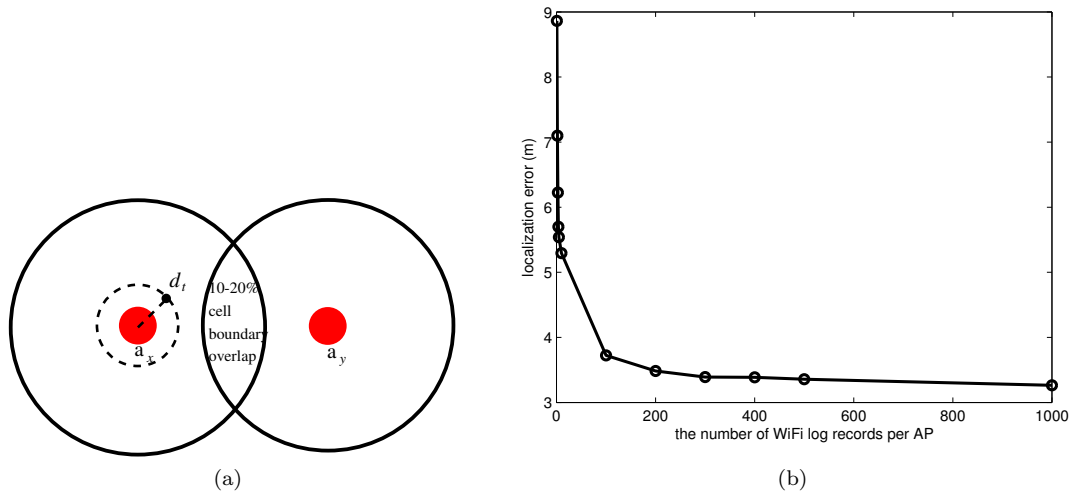


Figure 1: (a) Coverage areas of two adjacent APs and the cell boundary overlap. The overlapping area is 10 – 20% of a cell’s area. (b) An experimental illustration of the dependence of the accuracy of the method on the number of available RSSI observations during handover.

14 question emerges:

15

16 *How to perform accurate indoor localization using large-scale logs of discrete single-AP RSSI traces with*
 17 *low sampling rate?*

18

19 This problem opens a new direction for localization research. Specifically, we describe a robust WiFi log-
 20 based localization scheme which is:

- 21 1. non-intrusive: it expects nothing from the client mobile device, e.g. there is no need to install an app,
 22 turning-on of sensors other than WiFi;
- 23 2. generic: it is simple to deploy and applicable in any WiFi installation, which has an overlap between the
 24 coverage areas of adjacent APs and is capable of recording RSSI values when handovers occur between
 25 them. Additionally, the knowledge of the relative transmitter output power of the APs should be
 26 known by the operator;
- 27 3. light-weight: it uses algorithms that are simple to implement and maintain and do not overload existing
 28 computational infrastructures;
- 29 4. effective: as long as a mobile device connects to the WiFi network, the localization technique can be
 30 applied; and
- 31 5. accurate: the scheme delivers accuracy that is comparable to scene analysis, and exceeds the classical
 32 path loss model [1, 2], as demonstrated in the evaluation.

33 The D-Log positioning method is an enhancement over existing methods that roughly localize a device
 34 anywhere in the service area of an AP, by providing an estimate of the distance between the mobile device
 35 and the connected AP. This allows to further restrict the space in which the device is found. D-Log focuses
 36 on static localization, not the continuous tracking of people’s movement.

37 D-Log works by improving distance estimations from discrete single-AP RSSI traces of a mobile device.
 38 Specifically, D-Log applies the WiFi path loss model in combination with knowledge of the distance of
 39 neighbouring APs in a WiFi network and the probability distribution of each logged RSSI record to better
 40 estimate this distance. The key point is to utilize a combination of RSSI records from two neighbouring
 41 APs where handovers occur: a location that is known with some certainty [3, 4]. This information can be
 42 used to reduce errors introduced from the path loss model.

43 D-Log computes an enhanced distance estimate of a mobile device within the region served by a given AP
44 for each individual logged RSSI record. D-Log treats these estimates as independent instances drawn from
45 the same distribution. Applying probability theory, the average of these measurements allows estimation,
46 with greater accuracy, of the distance between the AP and the mobile device. The theoretical analysis is
47 provided to show D-Log’s performance in terms of localization accuracy (Section 3.5).

48 Consider two neighbouring APs a_x and a_y and the mobile device at the distance d_t , served by AP a_x ,
49 as shown in Fig. 1a. D-Log calculates one estimate of d_t by using each logged RSSI value when a handover
50 occurred. As there are a large number of WiFi log records for numerous devices, D-Log obtains a large
51 number of estimates of d_t at handover, and uses them to determine the average estimate as the distance to
52 the handover location d_t . This allows us to establish the empirical signal strength decay progression around
53 an AP, along which any non-handover locations can be interpolated for any observation of RSSI. Thus,
54 we use the knowledge of the handover to calibrate the signal strength decay function based on a path loss
55 model for each AP. Fig. 1b illustrates the dependence of the accuracy of the D-Log method on the number
56 of RSSI observations during handover, based on the experiments discussed later. As the number of logged
57 RSSI records increases, the average error of the position estimate decreases (Fig. 1b), converging towards a
58 limit value little above $3.0m$, achieved at around 300 observations.

59 Once a sufficient amount of WiFi AP logs has been recorded, they can be used to train the D-Log
60 algorithm. D-Log can then be used in (near) real-time, similar to other existing RSSI-based localization
61 methods. The D-Log scheme is, however, primarily meant to be deployed to improve the location estimate
62 in mobile device access records collected in a WiFi system in a post-processing step. Note that such logs are
63 collected at infrequent sampling rates from a single RSSI source to which the device is connected to. Most
64 existing RSSI-based methods are infeasible in such scenarios. Such enhancement of location estimation is
65 important for the improvement of indoor context estimation supporting a range of applications exploiting
66 indoor behaviour information mining and recommender systems [5, 6], in environments with free and publicly
67 available WiFi networks. Potential application areas include retail and advertising (e.g. shopping malls,
68 airports), leisure and tourism (e.g. attractions, entertainment areas), rich media consumption (e.g. smart
69 displays), teaching and learning support (e.g. in universities), and operational logistics (e.g. in airports,
70 transport hubs). Once accurate post-hoc localization of users within indoor spaces is possible, large-scale
71 Web activity and connectivity logs from the WiFi systems will enable extensive indoor information behaviour
72 mining and long-term prediction of user behaviours [7, 8].

73 The remainder of the paper is organized as follows. Section 2 presents the related work. The D-Log
74 scheme is detailed in Section 3, where a theoretical analysis is provided to show the performance benefit
75 of D-Log. Section 4 presents the data that we experiment with. Section 5 includes the evaluation of the
76 proposed method, and Section 6 concludes the paper and discusses possible future research.

77 2. Related Work

78 2.1. Indoor localization techniques

79 Existing research on indoor localization can be categorized into device-based [9, 10, 11, 12, 13], device-free
80 (passive) localization [14, 15], and infrastructure-based localization [16, 17, 18].

81 Device-based localization has gained popularity in recent years. This is due to the ability to integrate data
82 from multiple smartphone sensors (e.g. [19]) and thus allow for the combination of dead reckoning [12, 20, 21]
83 and particle filter estimation methods [22]. Although such a rich combination of signals improves indoor
84 localization, this is outside the scope of this paper, which is focused on post-hoc localisation based on
85 (sparse) WiFi AP logs of all the registered WiFi users. For device-based localization, it requires on-device
86 processing, typically via a mobile app, as well as continuous sampling of data. Given the requirement of
87 user participation and uptake with a mobile app, it limits the coverage of indoor monitoring. Full coverage
88 is often considered as a major requirement for indoor monitoring by facility owners and operators.

89 The most recent, albeit less common technique is device-free (passive) localization [14, 15]. Mobile
90 device-free localization does not require a device attached or carried by indoor visitors. But such methods
91 require high and continuous sampling rates and substantial post-processing efforts. They operate well

92 only in controlled environments, and multi-user tracking capability is often limited to small numbers of
93 simultaneously tracked objects. The most recent device-free (passive) localization method is capable of
94 tracking three users simultaneously [23]. Given the challenges with multi-user tracking and the need for
95 highly densed monitoring points and RSSI sampling, this is not applicable for tracking users in large-scale
96 public indoor spaces.

97 Many infrastructure based techniques utilise trilateration, which requires RSSI from multiple nearby APs.
98 However, these techniques are expensive to implement, since the WiFi networks have to be deployed with a
99 data logging configuration allowing multiple access points to be monitored across each device connection for
100 passive localization. This is typically not the case with most indoor environments currently operating WiFi
101 networks. As such, the logs acquired cannot be mined for accurate indoor spatial behaviour estimation.

102 Some research employs fusion of techniques. In [21], in-device recorded RSSI from a single access point
103 is used, however, the technique relies on dead reckoning to provide a perceived triangulation on the device.
104 Khan et. al. improved the coverage of localization through active participation of users [24]. Other local-
105 ization techniques employ the use of ZigBee networks (e.g. [25, 26]), RFID tags [27], or propagation model
106 and autonomous crowdsourcing [28].

107 2.2. RSSI use in indoor localization

108 With regard to the use of RSSI from WiFi access points in localizing devices of a WiFi network, tradi-
109 tionally, there are three main methods that are widely employed: trilateration, scene analysis, and proximity
110 analysis [29, 30].

111 First is trilateration, which estimates the position of a device by calculating its distance from multiple
112 reference points [30]. When RSSI traces from multiple access points are available, the use of this path
113 loss based method is a more accurate approach to localize a device, rather than using Time-of-Arrival
114 or Time-Difference-of-Arrival calculation [30] to approximate a device location, as the latter two methods
115 require a clear Line-Of-Sight (LOS) between the transmitter and the receiver [30]. An example of the use
116 of trilateration is in [17], where WiFi RSSI traces from multiple reference (access) points were recorded in
117 order to monitor around 18,000 devices in a hospital. They used WiFi signals measured on mobile devices to
118 first localize users in the building, extracted the spatial and temporal features from the traces, analyzed the
119 flow of people from entrance to exit, and classified their behaviours based on the user roles [17]. However,
120 in our study, RSSI from multiple reference points are not available, hence, trilateration is not applicable.

121 The second established RSSI-based localization approach is Radio Frequency (RF) based scene analysis,
122 a method to use prior-collected features, or fingerprints, of a scene to determine the location [29]. The
123 most widely used scene analysis method is RSSI-based fingerprinting [30]. Swangmuang and Krishnamurthy
124 presented an analytical model to predict the performance of fingerprinting-based indoor localization systems
125 by applying proximity graphs [31]. A WiFi RSSI fingerprint for each location is used to match the monitored
126 (indoor) environment for accurate localization of the device [32]. In some cases, fingerprinting at the actual
127 site is not feasible, e.g., in a very large shopping mall or airport. Since fingerprinting requires a large amount
128 of time and resources and costly system calibration in the beginning [32], the real-world use of this approach
129 was difficult. For example, in a highly dynamic environment, where layouts and objects often change,
130 RF fingerprints could easily change due to alterations of the indoor environment, hence requires frequent
131 fingerprinting [12]. [33] used knowledge about the geometry of the environment and made assumptions
132 about continuous indoor movement tracking to address this problem, while [34] collected user feedback to
133 improve the fingerprinting process. Want et. al. proposed a combination of subarea fingerprinting and
134 gradient descent search to improve localization by probabilistic fitting [35], but this fingerprinting approach
135 requires high frequency sampling.

136 The third approach is proximity-based localization, which uses RSSI captured on users' devices to com-
137 pute approximate sets of devices that are located in proximity to each other to localize the position of a
138 device relative to another device [29]. This method does not apply in our study since we do not use apps or
139 device-based approach to localize a user.

140 In this paper, we propose the D-Log scheme as a new reference scheme for post-hoc localization, which
141 aims to be easy to implement and maintain, is independent of devices and network infrastructure, and is ef-
142 fective and reasonably accurate. In Table 1, we compare D-Log with existing schemes, including trilateration,

Table 1: Comparison of indoor localization schemes.

Schemes	Signal	Cost	Client Sensors /Apps	AP Place-ments	RSSI Source (No. of APs)	Sampling Rate	Comments
Trilateration Scene analysis	RSSI	Med	No	Normal	At least 3	Low (continuous)	Infrastructure-based
	RSSI& Sensors	High	Yes	Normal	Multiple	High (continuous)	Device-based
Proximity analysis Device free	RSSI	High	Yes	Dense	Multiple	High (continuous)	Device-based
	RSSI	High	Yes	Dense	Multiple	High (continuous)	Device-free/passive
D-Log	RSSI	Low	No	Normal	Single	Low (discrete)	Log-based

143 scene analysis, proximity analysis and device free approaches in terms of their deployment characteristics.
 144 The D-Log scheme is low cost, because it only requires infrequent RSSI sampling from single RSSI source,
 145 rather than continuous RSSI sampling from multiple RSSI sources like others (e.g. scene analysis).

146 3. Log-Based Differential Scheme

147 In this section, we formulate the targeted research question and present two D-Log algorithms to estimate
 148 the distance of the mobile device to the AP. Furthermore, the complexities of the D-Log algorithms are
 149 analysed, and a theoretical analysis is provided to show the performance benefit of the entire proposed
 150 D-Log scheme.

151 3.1. Problem Formulation

152 In this paper, the research question is the estimation of a mobile device location within the coverage
 153 area of several WiFi APs based on logs of discrete RSSI traces from single APs. We assume that the WiFi
 154 log includes discrete RSSI measurements relating to a single AP connection at any one time, in contrast
 155 to the trilateration and scene analysis methods requiring multiple parallel RSSI observations. Single RSSI
 156 records are recorded in most real-world Wi-Fi system data logs, where non-serving APs and their RSSI are
 157 not recorded. Although these single-AP RSSI traces are normally discrete and sampled at low frequency,
 158 the quantity of records obtained from different devices for each WiFi AP is large. For example, the real-
 159 world WiFi log we examined (as detailed in Section 4), was collected with a 5min sampling rate for each
 160 registered mobile device; logging only the RSSI values for currently connected APs. This resulted in 480,924
 161 connections distributed amongst 35 APs, with in average around 13,000 records per AP. This large volume
 162 of available records for each AP creates an opportunity to accurately estimate the distance of a mobile device
 163 from an AP given its RSSI value.

164 There are several techniques to calculate d_t given an RSSI value r_t for a mobile device when associating
 165 with an AP. The path loss model [1, 2] enables to determine the device distance based on the full set of
 166 inputs:

$$\hat{d}_t = 10^{\left(\frac{TX_{pwr} - r_t - L_{tx} - L_{rx} + G_{tx} + G_{rx} - PL - s}{10e}\right)} \quad (1)$$

167 where \hat{d}_t denotes the estimated distance between the transmitter and the receiver (the client mobile device)
 168 in meters; TX_{pwr} is the transmitter output power in dB; r_t is the detected RSSI in dB; L_{tx} is the sum of
 169 all transmitter-side cable and connector losses in dB; L_{rx} is the sum of all receiver-side cable and connector
 170 losses in dB; G_{tx} is the transmitter-side antenna gain in dBi; G_{rx} is the receiver-side antenna gain in dBi;
 171 PL is the reference path loss in dB for the desired frequency when the receiver-to-transmitter distance is one
 172 meter; s is the standard deviation associated with the degree of shadow fading present in the environment;
 173 e denotes the path loss exponent for the environment. Note, although Eq. 1 takes a range of factors into
 174 consideration, the estimation of \hat{d}_t is not accurate, as the RSSI values r_t at location p_x vary and can be
 175 affected by a large number of external factors, e.g. the people movement through the space, the layout of
 176 the walls and the materials used in the environment.

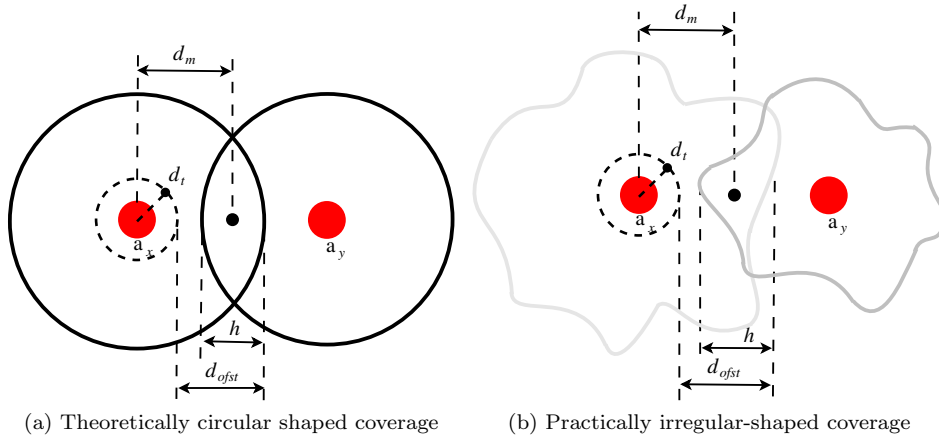


Figure 2: Illustration of d_m , h and d_{ofst} in D-Log algorithm with both theoretically circular shaped and practically irregular shaped coverage of several Wi-Fi APs. Here, the irregular shaped Wi-Fi AP coverage is obtained by following the study of wireless performance and coverage from Cisco Meraki [36].

177 Let us consider a general case: given two sets of sample RSSI values \mathcal{R}_x and \mathcal{R}_y , collected when the
 178 handover between two adjacent access points a_x and a_y happens, we denote $r_x^i \in \mathcal{R}_x$ a sample RSSI
 179 value observed when a mobile device is disassociating with a_x and then immediately associating with a_x 's
 180 topological adjacent AP a_y ; similarly, each $r_y^i \in \mathcal{R}_y$ denotes a sample RSSI value observed when a device is
 181 disassociating with a_y and then immediately associating with a_x . As there is only one observed RSSI value
 182 to the connected AP for the mobile device at any time, then other methods that rely on concurrent RSSI
 183 measurements from multiple APs are not applicable (e.g. trilateration and scene analysis). To address this
 184 problem, we propose the D-Log scheme to estimate d_t from the RSSI records $r_x^i \in \mathcal{R}_x$, not from r_t directly.
 185 Specifically, D-Log computes three other distances to interpolate d_t : 1) the distance d_m of mid-point of the
 186 overlapping coverage areas between a_x and a_y ; 2) the size, h of the handover area between a_x and a_y ; 3) the
 187 offset d_{ofst} between the mobile device and the handover boundary of a_x . As the two RSSI observations at
 188 handover have a number of inputs identical (assuming the transmitting power of the APs is either known
 189 or their proportions are known), this differential scheme allows to reduce the number of degrees of freedom
 190 influencing the distance determination. This indirect estimation enables D-Log to obtain a large number of
 191 distinct estimates for d_m , h and d_{ofst} , respectively, because there are a large number of $r_x^i \in \mathcal{R}_x$ in the log.
 192 As $r_x^i \in \mathcal{R}_x$ was collected independently in the log, the estimates from them are thus independent to each
 193 other. Then, from the aspect of probability theory, these observations can be used to estimate d_m , h and
 194 d_{ofst} , respectively. Take d_m as an example,

$$\hat{\mu}(d_m) = E(d_m|r_x) = E(\hat{d}_m^i) = \frac{1}{n} \sum_{i=1}^n d_m^i, \quad (2)$$

195 where d_m^i is the estimated distance of d_m based on a logged RSSI value r_x^i , and n is the number of log
 196 records. Moreover, this estimator has large practical application, as large datasets of RSSI logs are common
 197 and useful for a number of applications. Thus, the final interpolated d_t is accurate, and this will be detailed
 198 in the following sections.

199 3.2. D-Log Algorithm

200 Here, we propose the basic D-Log algorithm to estimate the location of a mobile device within the
 201 coverage area of an AP. The D-Log algorithm performs the localization using the following four steps:

- 202 • Step 1: Estimation of the distance d_m for the mid-point p_m of the overlapping coverage areas of two
 203 adjacent APs, a_x and a_y . Given a set of the RSSI values $r_x^i \in \mathcal{R}_x$ and $r_y^i \in \mathcal{R}_y$, obtained when the

204 handover happens between a_x and a_y , we define that

$$\hat{d}_m = E(\hat{d}_m^i) = \frac{1}{n} \sum_{i=1}^n \hat{d}_m^i = \frac{1}{n} \sum_{i=1}^n \frac{\hat{d}_x^i - \hat{d}_y^i + D}{2}, \quad (3)$$

205 where n denotes the number of sample RSSI values in \mathcal{R}_x and \mathcal{R}_y , D is the known distance between
 206 a_x and a_y , and \hat{d}_x^i and \hat{d}_y^i are the estimate distance from r_x^i and r_y^i by using Eq. 1, representing the
 207 distance from where the handover occurs to a_x and a_y , respectively.

208 • Step 2: Estimation of the size of the handover area of two adjacent APs:

$$\hat{h} = E(\hat{h}^i) = \frac{1}{n} \sum_{i=1}^n \hat{h}^i = \frac{1}{n} \sum_{i=1}^n (\hat{d}_x^i + \hat{d}_y^i - D). \quad (4)$$

209 • Step 3: Estimation of the offset between the mobile device at p_t and the handover boundary of the
 210 access point a_x .

$$\hat{d}_{ofst} = E(\hat{d}_{ofst}^i) = \frac{1}{n} \sum_{i=1}^n \hat{d}_{ofst}^i = \frac{1}{n} \sum_{i=1}^n (\hat{d}_x^i - \hat{d}_t^i), \quad (5)$$

211 where \hat{d}_t^i denotes the estimate distance from p_t to AP a_x based on the scanned RSSI value on the
 212 client device at p_t by using Eq. 1.

213 • Step 4: Calculation of the distance of the mobile device at p_t within the signal coverage area of a_x .

$$\hat{d}_t = \hat{d}_m + \frac{\hat{h}}{2} - \hat{d}_{ofst}. \quad (6)$$

214 Note, Eq. 6 differentiates the estimate of \hat{d}_t from each r_x^i and r_y^i via Eq. 3, 4, and 5 from Step 1, 2 and
 215 3. Thus, the D-Log algorithm can provide accurate localization of a mobile device within the coverage area
 216 of a_x . Once the distance to the mid point and the interpolation of RSSI values of a_x are determined, they
 217 can be applied to locate the mobile device at any distance from the serving AP as long as they are within
 218 the range. In addition, Fig. 2 shows an illustration of d_m , h and d_{ofst} in D-Log algorithm. Specifically,
 219 Fig. 2a shows these parameters when the Wi-Fi AP coverage shape is considered as circles theoretically,
 220 while Fig. 2b shows them when the coverage shape is irregular in practice.

221 3.3. Weighted D-Log Algorithm

222 The WiFi logs can be used to determine the distribution of the RSSI values when the handover happen
 223 between two adjacent APs a_x and a_y . Fig. 3 shows the distribution of these RSSI values collected in a real-
 224 world WiFi infrastructure in a large shopping mall in Australia (detailed in Section 4), and it is observed
 225 that they do not follow a uniform distribution. Highly frequent observations of the RSSI (here, around 2000
 226 RSSI observations with $r = -70dB$) bear higher impact on the final D-Log estimate than the less frequent
 227 ones (e.g. the 400 observations with $r = -90dB$). Commercial WiFi networks optimized for coverage often
 228 set $-70dB$ as a threshold value for received signal strength [37]. Following this, we propose a weighted
 229 D-Log algorithm by taking the RSSI sample frequency into consideration. Thus, we define the weighted
 230 version of the simple expectation location estimator (in Eq. 2) as:

$$\hat{\mu}(d_m) = E(d_m | r_x) = E(\hat{d}_m^i) = \frac{1}{\sum_i^u c_x^i} \sum_{i=1}^u c_x^i \hat{d}_m^i, \quad (7)$$

231 where c_x^i is the frequency of r_x^i , u denotes the number of unique r_x^i , and $\sum_i^u c_x^i = n$.

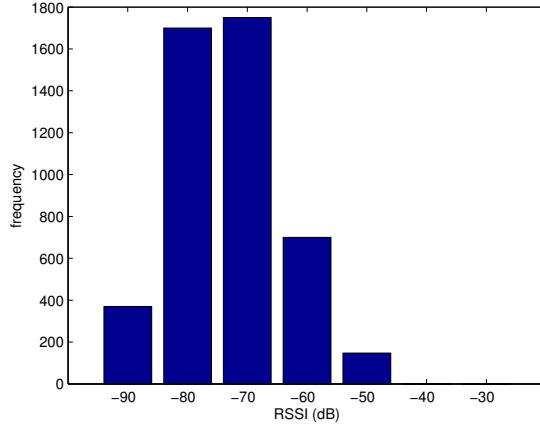


Figure 3: Distribution of RSSI values when handover happen between two adjacent APs in a real-world WiFi log, discussed in Section 4

232 Therefore, the corresponding weighted versions of \hat{d}_m , \hat{h} , \hat{d}_{ofst} and \hat{d}_t are defined as:

$$\hat{d}'_m = E(\hat{d}'_m) = \sum_{i=1}^u \frac{w_x^i (\hat{d}_x^i - \hat{d}_y^i + D)}{2}, \quad (8)$$

$$\hat{h}' = E(\hat{h}^i) = \sum_{i=1}^u w_x^i (\hat{d}_x^i + \hat{d}_y^i - D), \quad (9)$$

$$\hat{d}'_{ofst} = E(\hat{d}'_{ofst}) = \sum_{i=1}^u w_x^i (\hat{d}_x^i - \hat{d}_t^i), \quad (10)$$

$$\hat{d}'_t = \hat{d}'_m + \frac{\hat{h}'}{2} - \hat{d}'_{ofst}, \quad (11)$$

233 where $w_x^i = \frac{c_x^i}{\sum c_x^i}$, and c_x^i denotes the frequency of sample r_x^i .

234 3.4. Complexity Analysis

235 One advantage of the proposed D-Log scheme is its low computational complexity. The complexity of
 236 the D-Log algorithm is $O(n)$, where n denotes the average number of log records per AP; the complexity
 237 of the weighted D-Log algorithm is $O(u)$, where u denotes the number of unique RSSI values per AP. This
 238 indicates that D-Log scheme is efficient and only depends on the local log records for neighbouring APs,
 239 which enables the processing of large volume of records in parallel. In contrast, the complexity of the other
 240 RSSI based localization methods are often much larger than D-Log. For example, the complexity of machine
 241 learning based scene analysis (fingerprinting) models, is the same as that of the deployed machine learning
 242 methods, e.g, the complexity of SVM-based localization method is $O(\max(na, a) \cdot \min(na, a)^2)$ [38], where
 243 n is the number of training records, and a is the number of APs.

244 3.5. Performance Analysis

245 In this section, we provide a theoretical analysis of the performance of the unweighted D-Log algorithm.

246 The distance from where each r_x^i is observed to a_x can be estimated with Eq. 1, although there is an
 247 error ε caused by systematic and stochastic factors. For access point a_x , we define the estimation from r_x^i
 248 as

$$\hat{d}_x^i = d_x^i + \varepsilon_x^i, \quad (12)$$

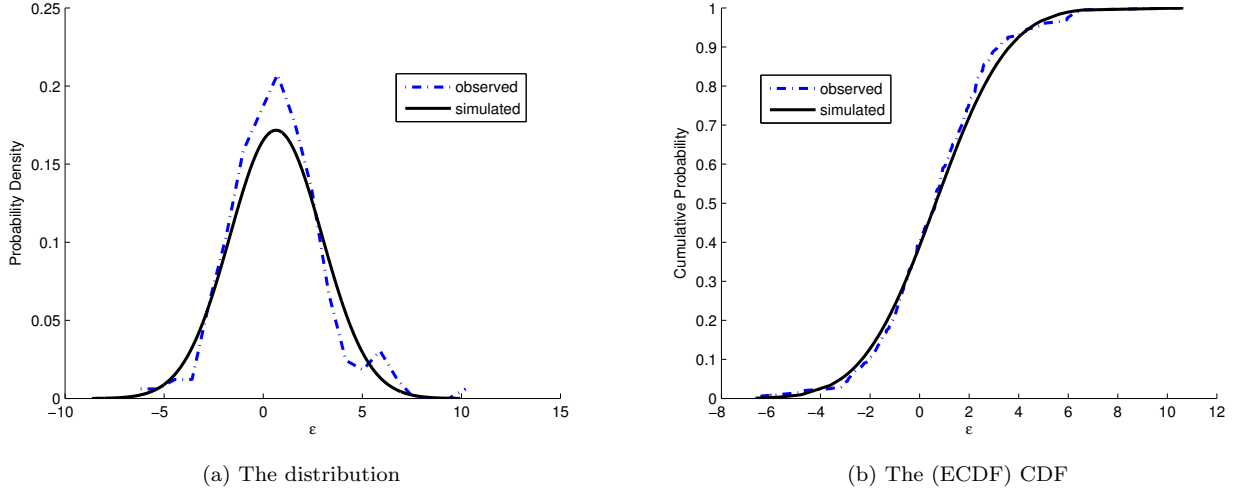


Figure 4: The distribution and (ECDF) CDF of ε and the reference Gaussian distribution

249 where \hat{d}_x^i is the distance estimation from r_x^i with Eq. 1, d_x^i is the real distance, and ε_x^i is the error for this
 250 estimation. Then, for access point a_y , we obtain

$$\hat{d}_y^i = d_y^i + \varepsilon_y^i. \quad (13)$$

251 We further assume that the estimation error ε from each sample RSSI value is independent and identically
 252 distributed (i.i.d), and we adopt the Gaussian distribution for theoretical analysis. This is motivated from
 253 the experimental results. Specifically, Fig. 4a shows the distribution of ε in our controlled experiment,
 254 which is detailed in Section 4. The dashed blue line depicts the observation empirical distribution of ε in the
 255 experiment, and the solid black line depicts the reference Gaussian distribution with the mean and standard
 256 deviation of ε . Fig. 4b shows the Empirical distribution function (ECDF) of ε (the dashed blue line) and
 257 the Cumulative Distribution Function(CDF) of the reference Gaussian distribution. It is observed that the
 258 reference Gaussian distribution fits the observation distribution of ε (with $D = 0.0558$, p -value = 0.5609 in
 259 Kolmogorov-Smirnov test), and it is thus a suitable model for the following theoretical analysis.

260 Consequently, the Probability Density Function (PDF) of ε is:

$$p(\varepsilon) \sim N(\mu_\varepsilon, \sigma_\varepsilon^2). \quad (14)$$

261 As stated in Eq. 2, we measure \hat{d} by applying the sample mean as the location estimator, and the distance
 262 on each observed RSSI can be considered as an observation. In the first step of D-Log algorithm, for the
 263 calculation of \hat{d}_m , according to Eq. 3 and Eq. 14, we obtain

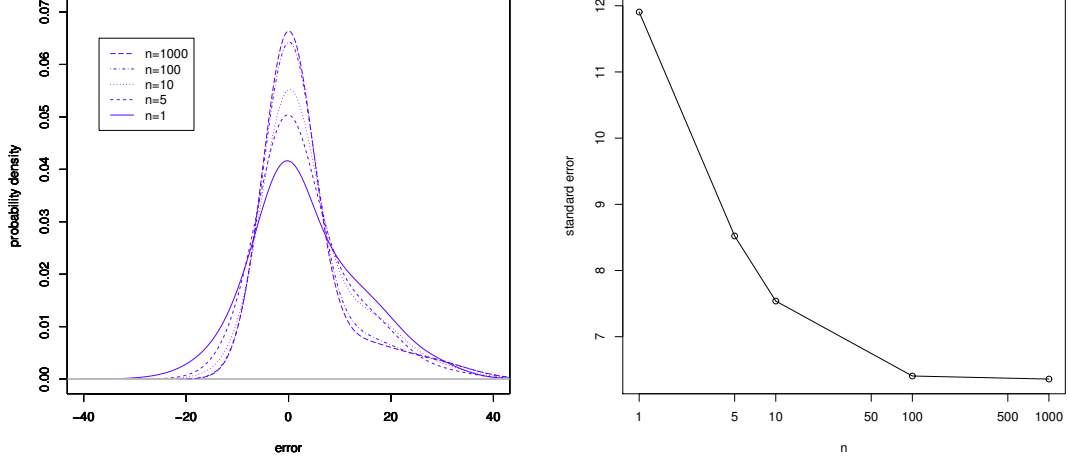
$$\hat{d}_m = E(d_m^i) = \frac{1}{n} \sum_{i=1}^n \frac{\hat{d}_x^i - \hat{d}_y^i + D}{2} = \frac{1}{2}(d_x - d_y + D) + \frac{1}{2n} \sum_{i=1}^n (\varepsilon_x^i - \varepsilon_y^i), \quad (15)$$

264 where d_x and d_y are the real distances of the handover boundary for a_x and a_y , respectively. Similarly,

$$\hat{h} = E(\hat{h}^i) = (d_x + d_y - D) + \frac{1}{n} \sum_{i=1}^n (\varepsilon_x^i + \varepsilon_y^i). \quad (16)$$

265 For the estimation of the offset between the mobile device at p_t and the handover boundary of the access
 266 point a_x , according to Eq. 5 and Eq. 14, we obtain

$$\hat{d}_{ofst} = E(\hat{d}_{ofst}^i) = \frac{1}{n} \sum_{i=1}^n (\hat{d}_x^i - \hat{d}_t^i) = (d_x - d_t) + \frac{1}{n} \sum_{i=1}^n (\varepsilon_x^i - \varepsilon_t^i), \quad (17)$$



(a) The distribution of $\varepsilon_{\hat{d}_t}$ with various n (b) the trend of the standard error $\sigma_{\hat{d}_t}$ with various n

Figure 5: The impact of n on $\varepsilon_{\hat{d}_t}$ and $\sigma_{\hat{d}_t}$

267 where d_t is the real distance between the test point p_t to a_x , and ε_t is the error when calculating \hat{d}_t .

268 Consequently, in the last step of D-Log, according to Eq. 6, Eq. 15, Eq. 16, Eq. 17 and Eq. 14, we obtain:

$$\hat{d}_t = \hat{d}_m + \frac{\hat{h}}{2} - \hat{d}_{ofst} = d_t + \frac{1}{2n} \sum_{i=1}^n (\varepsilon_x^i - \varepsilon_y^i) + \frac{1}{n} \sum_{i=1}^n (\varepsilon_x^i + \varepsilon_y^i) - \frac{1}{n} \sum_{i=1}^n (\varepsilon_x^i - \varepsilon_t). \quad (18)$$

269 Thus, according to Eq. 18 and Eq. 14, we obtain the $100(1 - \alpha)\%$ confidence interval $CI(\hat{d}_t)$ for the
 270 estimation of \hat{d}_t , which has been widely used to indicate the reliability of an estimation [39],

$$CI(\hat{d}_t) = d_t \pm z_{\frac{\alpha}{2}} \sqrt{\frac{5\sigma_\varepsilon^2}{n}}, \quad (19)$$

271 where $z_{\frac{\alpha}{2}}$ is a standard normal variate which exceeded with a probability of $\frac{\alpha}{2}$. Therefore, the standard
 272 error of \hat{d}_t is:

$$\sigma_{\hat{d}_t} = \sqrt{\frac{5\sigma_\varepsilon^2}{n}}, \quad (20)$$

273 where n denotes the sample size.

274 **Theorem 1.** *The standard error $\sigma_{\hat{d}_t}$ of D-Log scheme is bounded to be no more than $\sqrt{5\sigma_\varepsilon^2}$, with equality
 275 if and only if $n = 1$.*

276 **PROOF 1.** *As the sample size $n \geq 1$, based on Eq. 20, we obtain:*

$$\sigma_{\hat{d}_t} \leq \sqrt{5\sigma_\varepsilon^2}, \quad (21)$$

277 where the equality is satisfied when $n = 1$.

278 Fig. 5 shows the distribution of D-Log's localization error, $\varepsilon_{\hat{d}_t}$, and the trend of $\sigma_{\hat{d}_t}$, with various n values
 279 in our real-world indoor experiment environment, which is detailed in Section 4. Specifically, where $n = 1$,
 280 $\sigma_{\hat{d}_t}$ meets the worst case with the value of 11.9, as there is only 1 row of RSSI logs available. However, when
 281 more logs are available as shown in Fig. 5b, $\sigma_{\hat{d}_t}$ starts to decrease as n increases. It indicates that 1) as n
 282 increases, $\sigma_{\hat{d}_t}$ decreases; 2) D-Log has a floor level, which is influenced by the localization environment.

Table 2: Aggregate statistics of the WiFi log collected in a real-world large indoor retail environment

Number of user devices:	94,396
Number of AP association:	480,924
Number of Visits:	183,745
Number of WiFi APs:	35
Average of AP association per AP :	13,741

4. Data

In this section, we present the data used for the evaluation of the performance of the proposed D-Log scheme. We evaluate the performance of the D-Log scheme in two environments: a controlled environment and a real-world large indoor environment. The complexity of the two environments is different, and so is the evaluation setup. While in the simulated environment, the mobile devices used in the training and testing set of the controlled environment are identical and therefore the variability of the used WiFi is controlled, this is not the case in the real-world large indoor environment.

4.1. Experiment Data

Here, we describe the experiment data from the two experimental environments: the controlled environment and the real-world large indoor environment.

For the controlled environment, we set up an experimental WLAN with 4 access points in a university meeting room (dimension: $7m$ by $5m$). We have partitioned the room into 35 ($1m \times 1m$) square grids, and used 16 of them as the test locations. These test locations were located along walls and in locations not occupied by furniture. Then, we recorded the RSSI values during handover of the carried mobile device (a smartphone) from one test AP to another. These recordings supply the training RSSI logs for D-Log scheme. For testing purposes, we collected around 6000 sample RSSI records (about 360 per location) from all detected APs, which will be used to evaluate the performance of D-Log scheme and the compared state-of-the-art localization methods.

Additionally, we have conducted real-world experiments in a large inner-city shopping mall in Sydney, Australia, covered by 67 WiFi APs across 90,000 square meters. We used three levels of the mall to conduct our experiments, in an area of around 35,000 square meters covered by 35 WiFi APs. The WiFi log were collected from September 2012 to October 2013, and were stored in an external system. It contains around half a million AP access records from around 100,000 mobile devices. Specifically, the log includes the WiFi access point associated with the user’s mobile device sampled at every 5 minutes, and the respective RSSI value for each association. These data are used as training data for the D-Log scheme with some preprocessing that is detailed in Section 4.2. Table 2 shows the statistics of the log. Note, all user identifiable information (registration details and WiFi MAC addresses) were replaced by a hash key in a non-reversible way. To examine the localization performance in this real industry environment, we selected 43 test locations across the three floors of the mall, and collected around 4000 sample RSSI records (around 100 per location) from all detected APs. Fig. 6 shows the floor maps and the test locations. Specifically, we collected 10 test locations on the 1st floor, 15 on the 2nd floor, and 18 on the 3rd floor. Moreover, note this real-world RSSI log contains much complexities, which may influence all RSSI based localization methods, e.g. the variance mobile devices/antenna/Wi-Fi chipsets. There are 694 different mobile models from 53 manufacturers in our collected WiFi logs, and Table 3 and 4 show the most common manufacturers and models of the used devices in the log, respectively.

4.2. Pre-processing the WiFi AP Log

The real-world industry WiFi log we used was sampled at 5 minutes frequency for each user visit, and for each device, only the RSSI values for current connected AP were logged. Table 5 shows a sample of the log for a specific user.

Table 3: Most common manufacturers of used mobile devices

Manufacturer	#	Manufacturer	#	Manufacturer	#
Apple	66921	Unidentified	187	Xiaomi	22
Samsung	10587	Huawei	114	Toshiba	16
Generic (<i>Android</i>)	9018	Amazon	106	ZTE	13
HTC	1861	Sony	90	Fujitsu	12
RIM	1284	Microsoft	82	Opera	11
SonyEricsson	697	Asus	53	KDDI	11
Nokia	585	Pantech	41	NEC	9
Google	401	Sharp	35	Alcatel	8
LG	347	DoCoMo	32	HP	7
Motorola	240	Acer	26	Lenovo	7

Table 4: Most common models of used mobile devices

Model	#	Model	#	Model	#
iPhone (<i>Apple</i>)	54873	Galaxy Nexus (<i>Samsung</i>)	420	BlackBerry 9780 (<i>RIM</i>)	177
iPad (<i>Apple</i>)	7523	GT-I9305 (<i>Samsung</i>)	414	Desire HD (<i>HTC</i>)	173
iPod Touch (<i>Apple</i>)	4525	GT-I9000 (<i>Samsung</i>)	407	Desire (<i>HTC</i>)	159
Android 4.1 (<i>Generic</i>)	4173	GT-N7000 (<i>Samsung</i>)	358	PJ83100 (<i>HTC</i>)	145
GT-I9300 (<i>Samsung</i>)	2791	Fennec (<i>Generic</i>)	291	LT26i (<i>SonyEricsson</i>)	142
GT-I9100 (<i>Samsung</i>)	2602	BlackBerry Bold Touch 9900 (<i>RIM</i>)	261	BlackBerry 9800 (<i>RIM</i>)	139
Android (<i>Generic</i>)	1989	Nexus 4 (<i>Google</i>)	231	Nexus S (<i>Google</i>)	130
Android 4.0 (<i>Generic</i>)	1801	GT-S5830 (<i>Samsung</i>)	220	BlackBerry 9700 (<i>RIM</i>)	127
GT-N7100 (<i>Samsung</i>)	849	GT-I9305T (<i>Samsung</i>)	214	A510 (<i>HTC</i>)	126
Android 2.3 (<i>Generic</i>)	452	Unidentified (<i>Generic</i>)	199	S710E (<i>HTC</i>)	124

322 This infrequent sampling rate from single RSSI source makes it infeasible to apply existing localization
323 methods, including trilateration, scene analysis, proximity analysis and device free method. This is because
324 all of these existing methods require RSSI traces from multiple sources with frequent continuous sampling.
325 So, doing localization based on this sort of data is not trivial. We conducted some data pre-processing as
326 follows: 1) We carry two mobile devices (one IOS iPhone 4 and one Android Sumsung S4) to the mall to
327 record the RSSI values when a handover happens between neighbouring APs, and treat these RSSI values
328 as the handover boundaries of corresponding APs; then 2) for each AP, we extract all the RSSI values
329 that are less than those identified handover boundaries from the real-world WiFi log, so as to estimate the
330 distribution of the RSSI values when handovers happen. Finally, these extracted subset of RSSI values are
331 used as training samples for the D-Log scheme.

Table 5: Examples of the WiFi log for user *E154GCHIJDSPMLX5KFJC*

Hashed MAC address	WiFi AP	RSSI	association time	disassociation time	Duration (sec)
E154GCHIJDSPMLX5KFJC	AP 1	-76	2013-02-04 14:16:24	2013-02-04 14:21:24	300
E154GCHIJDSPMLX5KFJC	AP 3	-72	2013-02-04 14:21:24	2013-02-04 14:26:24	300
E154GCHIJDSPMLX5KFJC	AP 7	-75	2013-02-04 14:26:24	2013-02-04 14:31:24	300
...

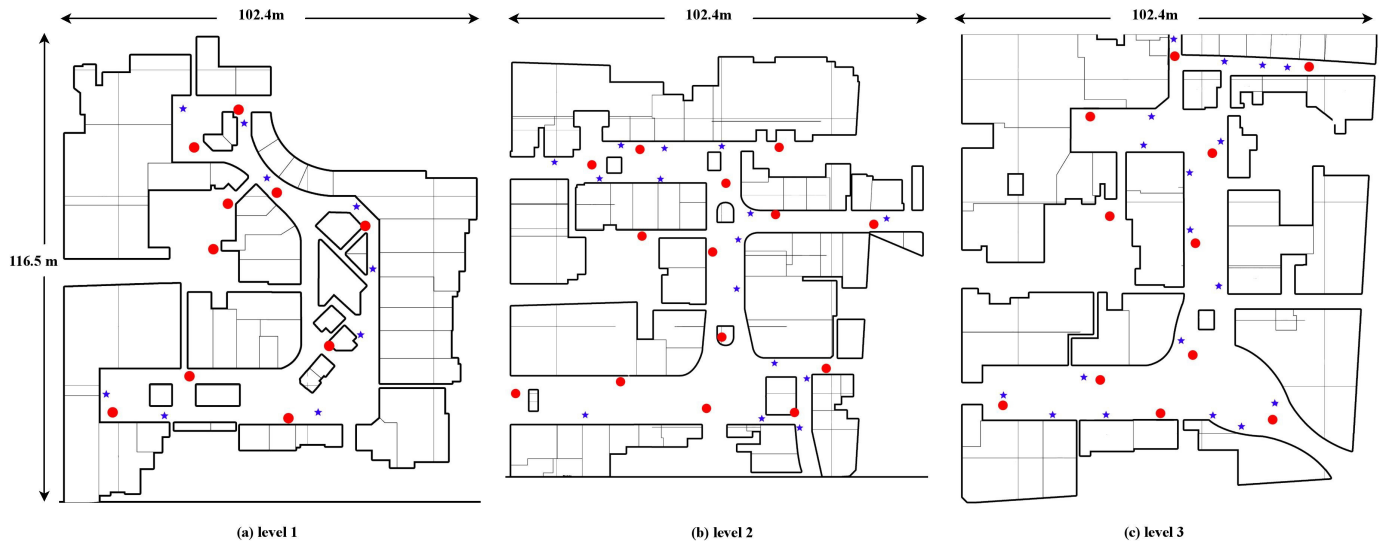


Figure 6: The floor maps in the mall where the experiments are conducted. The red dots represent the Wi-Fi APs, and the blue stars denote the test locations where ground-truth RSSI information were collected.

332 5. Experiment Results

333 In this section, we present the experimental configuration and the performance of the proposed D-
 334 Log scheme in terms of localisation accuracy achieved by D-Log. Note, this localization relates to the
 335 determination of the distance of the mobile device from the AP and therefore the reported error indicate
 336 the width of the band in which the mobile device is located.

337 5.1. Experiment Baselines

338 To examine the performance of D-Log scheme thoroughly, we compare the proposed D-Log scheme with
 339 two state-of-the-art RSSI-based localization methods: scene analysis methods, and Path Loss model [1, 2].
 340 There are two reasons we choose these two baselines: 1) By comparing with scene analysis, we demonstrate
 341 how closely the D-Log scheme performs comparing to the state-of-the-art, because scene analysis is one of
 342 the most accurate and most popular RSSI-based localization methods; 2) the path loss model is selected
 343 to perform a fair comparison because it also makes an estimate of the radius of the receiver like the D-
 344 Log scheme. For the scene analysis methods, we choose two algorithms: SVM-based method [40] and the
 345 Bayesian Network-based method [30], given that these two are among the state-of-the-art learning techniques
 346 applied for fingerprint-based indoor localization.

347 5.2. Experimental Configuration

348 5.2.1. Evaluation Metrics

349 The experiments were conducted on a PC running the Windows 7 Operating System with 8 GB RAM
 350 and Intel Core i7 CPU, and we conducted a 10-fold cross validation and report the results. Note that We
 351 deployed the well-known LibSVM¹ package to perform the SVM-based method, and Weka (Data mining
 352 Software in Java²) to perform the Bayesian Network-based method; For the proposed D-Log scheme and
 353 the state-of-the-art Path Loss model, we implemented them in Java.

¹<https://www.csie.ntu.edu.tw/~cjlin/libsvm>

²<http://www.cs.waikato.ac.nz/ml/weka/>

354 Following literature [31], we apply the mean precision $P(\mathcal{T})$ and the mean absolute error (ε , localization
355 accuracy) as the measurement metric:

$$P(\mathcal{T}) = \frac{|\mathcal{T}_c|}{|\mathcal{T}|}, \quad (22)$$

$$\varepsilon = \frac{\sum |d_t - \hat{d}_t|}{|\mathcal{T}|}, \quad (23)$$

356 where \mathcal{T} is the test set, $|\mathcal{T}_c|$ denotes the number of test locations that are correctly assigned to its true
357 location, $|\mathcal{T}|$ denotes the size of \mathcal{T} , including both correctly assigned and incorrectly assigned test locations,
358 d_t is the true distance, and the \hat{d}_t is the estimated distance. For D-Log and Path Loss model, while
359 calculating $|\mathcal{T}_c|$, if $d_t - \sigma_\varepsilon < \hat{d}_t < d_t + \sigma_\varepsilon$, \hat{d}_t is considered as the true location, otherwise false location. For
360 SVM-based method [40] and the Bayesian Network-based method [30], they output the labels of each test
361 location. While calculating ε , if the output label is the real label of the test location, ε for this test location
362 is 0, otherwise the difference between the true distance d_t and the distance from the AP to the output label
363 location, which is \hat{d}_t .

364 5.2.2. Parameter Estimation

365 Like other localization methods, there are parameters in the proposed D-Log scheme, which are the
366 parameters in the path loss model as shown in Eq. 1. Some of these parameters are known (e.g. the
367 transmitter output power), or can be measured by site surveying process (e.g. path loss exponent e),
368 but some others are hard to measure or measure accurately in practice. For example, in the investigated
369 mall, a large variety of different brands and models of receivers (mobile phones) are involved, which makes it
370 infeasible to measure the receiver-side related parameters; the presence of obstructions and people movement
371 is changing frequently, which makes it hard to accurately measure other parameters, e.g. the path loss
372 exponent e and the standard deviation of shadow fading s [2].

373 Thus, similar to other localization methods again, some data mining techniques can be applied to es-
374 timate these parameters. For example, Durgin et. al. applied linear regression to estimate the path loss
375 exponent e and the reference path loss PL at $1m$ transmitter-receiver separation by using pairwise RSSI
376 measurements and log distances [1]. Recently, cross validation has been widely used to estimate parameters
377 of indoor localization models, e.g. kernel-based indoor localization algorithms [41], machine learning based
378 algorithms [42], and powerline positioning algorithms [43]. Following this, we deploy cross validation to
379 estimate the parameters of D-Log scheme by using pairwise RSSI measurements and log distances.

380 Specifically, because we used the collected experimental data to both estimate the parameters of the
381 models and evaluate them, we deployed a nested cross validation to ensure the final model evaluation is
382 unbiased [44]. Note that, there are two disjoint datasets in D-Log scheme, the RSSI logs, and the pairwise
383 RSSI records and distances collected at test locations. We call the RSSI logs the *training* set, and divide
384 the pairwise RSSI records and distances collected at test locations into another two disjoint subsets: the
385 *validation* set and the *test* set. Therefore, the *training* set, the *validation* set and the *test* set are independent
386 to each other. Consequently, the learnt parameters will not overfit the data, and the final localization results
387 are unbiased [44, 45].

388 Although theoretically the nested cross validation strategy can search and estimate the parameters in
389 anyway, it is practically helpful to obtain the ranges of these parameters as accurate as possible. To estimate
390 the ranges of these parameters accurately and to not disturb the investigated mall's daily business (running
391 7 days), we set up a shopping mall like simulation environment in the RMIT Indoor Positioning Lab.
392 Specifically, we set up a Wi-Fi network in the simulation environment with the same configurations of that
393 in the investigation mall, e.g. the wireless networking standard 802.11n(2.4GHz) and the model of access
394 points; and we used three different phones (one IOS iPhone 4, one Android Sumsung S4 and one HTC ONE)
395 with a Java program installed to measure the receiver-side related parameters. Then, an expert, one author
396 of this paper, measured the ranges of all parameters, which are used to determine the possible candidate
397 values for each parameter. The detailed procedure of the deployed nested cross validation strategy is shown

Table 6: Comparison of localization precision in controlled environment. Note, weighted D-Log, D-Log and path loss model used logs of single-AP traces; SVM-based method and Bayesian Network-Based Method used the RSSI records from multiple APs.

	Weighted D-Log	D-Log	SVM-Based	Bayesian Network-Based	Path Loss
$P(\mathcal{T})$	61.3%	60.1%	69.1%	66.9%	32.9%
ε (m)	0.93	1.01	0.91	1.03	1.82

398 in Algorithm 1.

```

1 randomly divide the pairwise RSSI records and distances collected at test locations into  $k$  equal sized
  subsets;
2 for each subset do // outer loop
3   use this subset as test set, and the rest  $k - 1$  subsets as validation set;
4   for each candidate value of the parameters in the measured ranges do // inner loop
5     use this candidate parameter to build D-Log model on the training RSSI Logs;
6     validate the model on the validation set and calculate localization error for each pair of RSSI
399     records and distances;
7     average the localization error of all pairs to get  $\varepsilon_{validation}$  on the validation set;
8   end
9   select parameters that minimize  $\varepsilon_{validation}$ ;
10  build model with the learnt parameters, and calculate  $P(\mathcal{T})$  and  $\varepsilon$  on the test set;
11 end
12 average  $P(\mathcal{T})$  and  $\varepsilon$  on all test set as the final result;

```

Algorithm 1: Nested cross validation

400 Note that, the *training* set, the *validation* set and the *test* are disjoint to each other. The deployed nested
401 cross validation includes two loops: *inner* loop and *outer* loop. The inner loop is designed to estimate the
402 parameters, which is a loop of a variant leave-one-out cross validation in D-Log scheme due to the following
403 two factors: 1) the *training* set is always the same and is always disjoint with the *validation* set and the
404 *test* set; 2) $\varepsilon_{validation}$ is obtained by repeating and averaging the calculation of localization error on each
405 pair of RSSI records and distances in the *validation* set with current parameters. The outer loop is used to
406 evaluate the performance of the model, which is a standard k -fold cross validation, and we set $k = 10$ in
407 this study.

408 5.3. Controlled Environment

409 Here, we present the experiment results in the controlled environment, including the localization accuracy
410 and the impact of sample size.

411 5.3.1. Localization Accuracy

412 Table 6 shows the results of localization precision $P(\mathcal{T})$ and ε in the controlled environment. It is
413 obtained that, for $P(\mathcal{T})$, the *chi*-squared test shows that there is no statistical significant difference (with
414 *chi*-squared = 0.6735, *p*-value = 0.7141) between D-Log, SVM-based method, and Bayesian Network-based
415 method. This indicates that the D-Log scheme performs well in comparison to the high-cost high-complexity
416 scene analysis methods, SVM-based method and Bayesian Network-based method. Furthermore, the D-Log
417 scheme performs significantly better than the path loss model. More importantly, D-Log scheme achieves
418 similar performance to SVM-based method, Bayesian Network-based method in terms of ε . The weighted
419 D-Log algorithm achieves a localization error of 0.93 meters, which is only slightly higher than that of
420 the SVM-based method (0.91 meters); at the same time, it outperforms both Bayesian Network-based
421 method (1.03 meters) and the Path Loss model (1.82 meters). Overall, D-Log scheme achieves comparable
422 localization accuracy to the high-cost high-complexity localization methods.

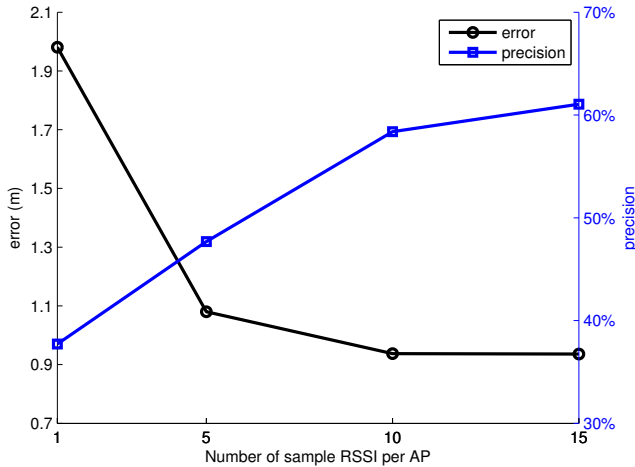


Figure 7: The impact of sample size in the controlled environment

Table 7: Single-floor localization performance in the real-world mall environment. Note, weighted D-Log, D-Log and path loss model used logs of single-AP traces; SVM-based method and Bayesian Network-Based Method used the RSSI records from multiple APs.

Floor	Metric	Weighted D-Log	D-Log	SVM-Based	Bayesian Network-based	Path Loss
1 st	$P(\mathcal{T})$	92.3%	92.3%	96.1%	91.0%	10.3%
	ε (m)	1.53	1.53	0.44	1.46	7.74
2 nd	$P(\mathcal{T})$	81.6%	81.6%	89.5%	81.6%	21.1%
	ε (m)	2.93	2.93	1.54	4.09	8.98
3 rd	$P(\mathcal{T})$	74.3%	74.3%	84.4%	77.9%	44.9%
	ε (m)	4.07	4.07	3.38	6.24	8.14

5.3.2. Impact of Sample Size

D-Log scheme uses the RSSI values measured during handover between two neighbouring APs, so it is important to examine the impact of the size of these sample RSSI values. Fig. 7 shows the performance of the D-Log scheme over the number of RSSI values per AP in terms of both localization precision $P(\mathcal{T})$ and error ε . It is observed that, as the size of training RSSI values increases, $P(\mathcal{T})$ consistently increases and ε consistently decreases. This is as what we have analysed in Eq. 20 in Section 3.5, because the confidence interval of D-Log’s estimation is proportional to the size of the sample observations. When only several sample observations are available, the performance is inferior, but improves and stabilizes when the sample size is greater than 10 observations in the controlled environment.

5.4. Large Real-World Environment

Here, we evaluate the proposed D-Log scheme in a real-world large indoor retail environment, an inner-city shopping mall in Sydney, Australia, by using the anonymized real-world WiFi log of an opt-in free WiFi network operated by the mall owner. Note that this real-world mall environment is different from the environment of the department meeting room in the controlled environment, especially in terms of environment complexity, which may affect the values of RSSI readings, including brands/models of mobile devices, antenna models, Wi-Fi chipsets [46], and people movement [47] etc.

5.4.1. Localization Accuracy

Table 7 shows the localization accuracy in both $P(\mathcal{T})$ and ε within specific single floor. Here, all compared algorithms assume the training set is restricted to the data collected on the same floor as the test location,

Table 8: Multi-floor localization performance in the real-world mall environment. Note, weighted D-Log, D-Log and path loss model used logs of single-AP traces; SVM-based method and Bayesian Network-based method used the RSSI records from multiple APs.

	Weighted D-Log	D-Log	SVM-Based	Bayesian Network-Based	Path Loss
$P(\mathcal{T})$	81.1%	81.1%	84.3%	82.3%	28.4%
ε (m)	3.07	3.07	2.89	4.3	8.34

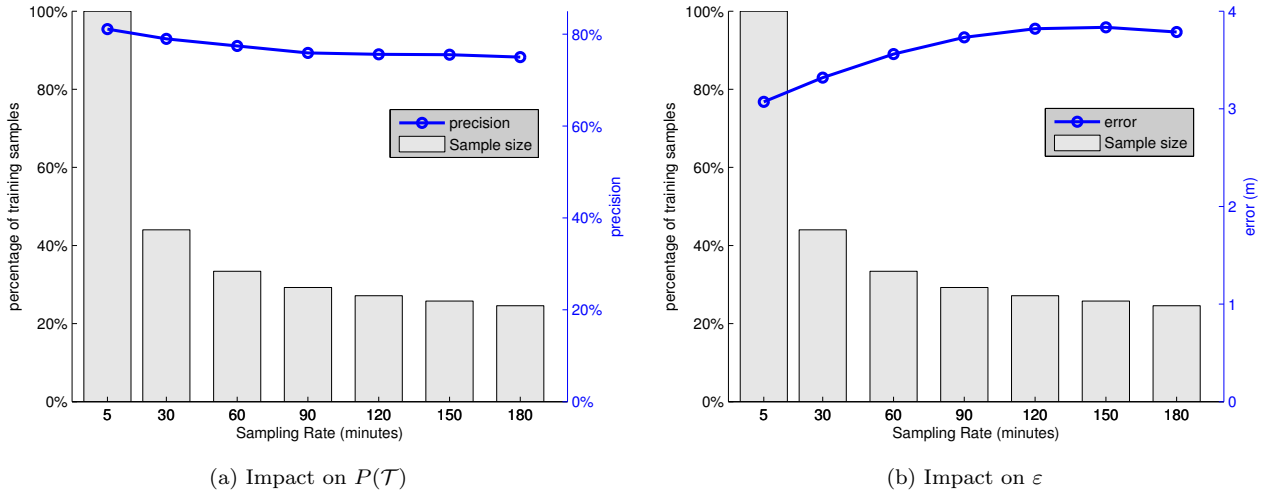


Figure 8: The impact of sampling rate in the real-world mall environment

442 an approaches replicated from a similar experimental environment [48]. For $P(\mathcal{T})$, it is observed that D-
 443 log scheme performs comparatively to SVM-based method and Bayesian Network-based method across all
 444 three tested floors, and the *chi*-squared test results confirm that there is no significant difference in their
 445 performance: the 1st floor (*chi*-squared = 0.1508, *p*-value = 0.9274), the 2nd floor (*chi*-squared = 0.4939,
 446 *p*-value = 0.7812), the 3rd floor (*chi*-squared = 0.6645, *p*-value = 0.7173). For ε , when the complexity of the
 447 test location increases from the 1st floor to the 3rd floor, D-Log scheme starts outperforming the Bayesian
 448 Network-based method. This indicates that in the complex environment, some scene analysis methods
 449 will be limited to the capability of the deployed data mining method. In contrast, D-Log exhibits strong
 450 robustness in these complex environments.

451 Table 8 shows the results of $P(\mathcal{T})$ and ε across multiple floors. To illustrate the performance of algo-
 452 rithms in this scenario, following [48], we remove the floor information by projecting the training points
 453 collected on different floors to a single plane, and execute all the compared algorithms. Again, the D-Log
 454 scheme significantly outperforms the Path Loss model and Bayesian network-based method, and performs
 455 comparably well to the SVM-based method.

456 Overall, the D-Log scheme performs comparatively to the state-of-the-art localization algorithms while
 457 utilizing less resources and being computationally less complex. In addition, we observe that in both
 458 single-floor and multiple-floor environments, weighted D-Log algorithm performs equivalently to the D-Log
 459 algorithm. This is due to the large size of the WiFi log, enabling the two methods to converge in performance.

460 5.4.2. Impact of Sampling Rate

461 As analysed in Section 3.5, D-Log scheme can provide accurate localization accuracy by utilizing large
 462 RSSI logs, and it is independent of the sampling rate when logging the WiFi RSSI traces. Fig. 8 shows the
 463 sample size and the $P(\mathcal{T})$ and ε performance of D-Log algorithm when the sampling rate of our real-world
 464 WiFi logs varies from 5 minutes to 3 hours. The sample size is presented as the fraction of the sampling

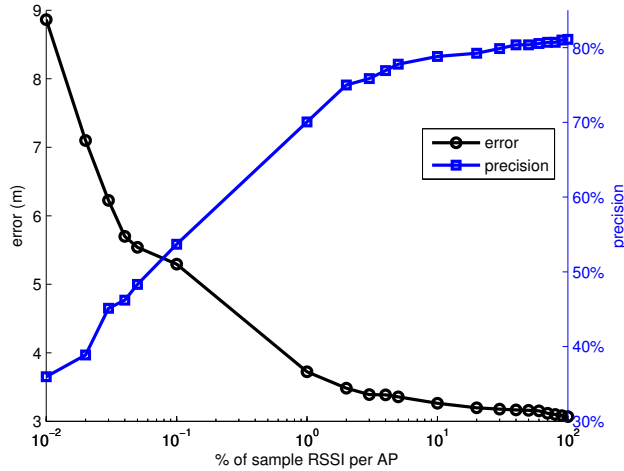


Figure 9: The impact of sample size in the real-world environment.

Table 9: Comparison of possible handover RSSI values

	Pre-processing	Average	Fixed $-70dB$ [37]	30% of least RSSIs	Path Loss
$P(\mathcal{T})$	81.1%	61.9%	59.5%	63.1%	28.4%
ε (m)	3.07	4.21	4.41	4.23	8.34

465 rate at the default 5 minutes. While the sampling rate drops from 5 minutes to 3 hours, $P(\mathcal{T})$ drops from
 466 81.1% to 75.0%, and ε increases from 3.07 meters to 3.78 meters. In other words, while the sampling rate
 467 drops 18 times, there is no corresponding reduction in $P(\mathcal{T})$ and ε . This indicates that the sampling rate of
 468 the WiFi logs has little impact on the performance of D-Log scheme.

469 The small decrease of localization accuracy when sampling rate drops is caused by the drop of corre-
 470 sponding sample sizes. Specifically, when sampling rate varies from 5 minutes to 3 hours, the size of the
 471 corresponding RSSI samples drops by 75.4%. A detailed discussion of the impact of sample size in this
 472 real-world environment is discussed in the following section.

473 5.4.3. Impact of Sample Size in Real-World Environment

474 In the real-world environment, the collected Wi-Fi logs capture heterogeneous mobile devices, thus
 475 impacting on localization. We therefore examine the impact of this noisy training sample on the performance
 476 of the D-Log scheme. Fig. 9 shows the $P(\mathcal{T})$ and ε performance of D-Log in function of the training sample
 477 proportion used in the D-Log scheme, where each result in the figure is executed 10 times and then averaged.
 478 We observe that $P(\mathcal{T})$ increases proportionally with number of training samples, while ε decreases, which
 479 is consistent with the findings from the controlled environment in Section 5.3. Specifically, the first several
 480 samples can largely boost the performance of the D-Log algorithm, and makes it outperform the classic
 481 path loss model; the elbow-point is achieved at around 2% of training samples, which is around 250 training
 482 samples. This indicates that in large complex environments, D-Log scheme is also robust to the noises of
 483 the training data, and can achieve accuracy comparable with competing methods with a limited number of
 484 training samples. Recall that the accuracy of the positioning relates to the determination of the distance of
 485 the mobile device from the AP, not to an exact point in 2D space.

486 5.4.4. Impact of Handover RSSI

487 To accurately estimate the distance between a mobile device and the servicing AP, D-Log scheme requires
 488 the RSSI values when handover happens between adjacent APs in the WiFi network. However, in some

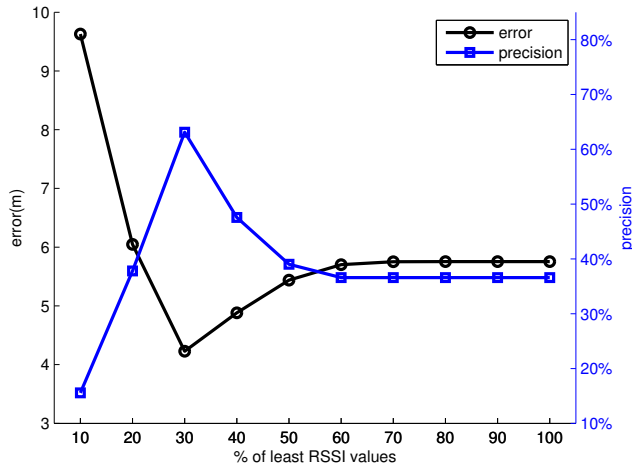


Figure 10: The impact of possible handover RSSI values

existing logs the RSSI values may be collected at very coarse frequency, e.g. the 5 minutes sampling interval in the WiFi log we experimented with. To test the applicability of such a coarsely sampled log, we have collected the accurate RSSI values at exact handover moments as a baseline (see Sec 4.2), and compared it to the subset of records estimated to have happened at, or close to, the handover. Here, we discuss the impact of the uncertainty of the handover identification on the calibration of the D-Log scheme. The baseline D-Log accuracy achieved based on the pre-processed input is compared to the following three methods:

- Average: uses the average of RSSI values of each AP in the log as the handover threshold;
- Fixed $-70dB$: applies a fixed value of $-70dB$ as the handover threshold. This RSSI value is commonly suggested by commercial WiFi network installation manuals, e.g. Cisco [37];
- Least RSSI: for this method, it is assumed that the potential handovers happened when the disassociation time of a_x is the same as the association time of a_y , which is a_x 's adjacent AP in the WiFi network (recall, that our logs have a sampling frequency of 5mins). Then, a limited fraction of the least of these RSSI values is used to select records assumed to relate to handover RSSIs. Fig. 10 shows the performance of this method as a function of the fraction of least RSSI values. Initially, when only a small proportion (no more than 30%) of the least RSSI values are selected, the performance increases steeply; beyond 30%, the performance deteriorates.

Table 9 shows the performance of these methods in terms of $P(\mathcal{T})$ and ε . We observe that: 1) the D-Log scheme with the proposed pre-processing steps in Sec 4.2 achieves the best performance; 2) D-Log scheme with possible handover RSSIs, including Average, Fixed $-70dB$ [37] and 30% of least RSSI, outperforms significantly the path loss model. This indicates that even when accurate handover RSSIs are not available, D-Log scheme still outperforms the state-of-the-art path loss model. Furthermore, with minimal environment fingerprinting that is substantially simpler than fingerprinting required by other methods, D-Log is able to achieve very good performance.

5.5. Discussion

The proposed D-Log scheme fulfils the five requirements introduced in the introduction of the paper:

1. non-intrusive: D-Log scheme works on the logs of discrete single-AP RSSI traces collected on the AP side, and does not need any information related with the client mobile devices, e.g. no need to install apps, or turning-on of phone sensors;

- 517 2. generic: as long as there is an overlap between the signal coverage areas of two adjacent APs, a valid
518 localization can be performed. Note this is generally a priority in WiFi network design. Similarly, the
519 transmitting power of all APs is typically standard and identical for large-scale deployments and can
520 be found in manufacturer’s manuals [37];
- 521 3. light-weight: the proposed D-Log scheme is composed of simple computational components with only
522 basic computational requirements;
- 523 4. effective: as long as a mobile device connects to the WiFi network, its RSSI value can be identified.
524 Thus, D-Log can make a valid estimate of the radius to the connected AP;
- 525 5. accurate: the accuracy of the D-Log scheme is comparable to other state-of-the-art RSSI-based local-
526 ization methods as shown by our analysis in Section 3.5, with values sufficient for applications requiring
527 an estimate of the immediate spatial context of the user.

528 One limitation of the D-Log scheme is that it builds on the Path Loss model which requires certain
529 parameters of the WiFi network to be known, as shown in Eq. 1. These parameters are known or can be
530 measured by site surveying process, or can be learnt by using cross validation as shown in Section 5.2.2.

531 Due to the above discussed characteristics, D-Log can be applied in a range of applications, e.g. fine-
532 grained spatio-temporal analysis, spatial data management and indoor behaviour analysis [8]. For example,
533 Fig. 11 shows how D-Log scheme can help when only discrete single AP-traces are available. Specifically,
534 the figure on the left shows the D-Log’s positioning of two particular mobile devices (the two purple stars).
535 Namely, for each mobile device, the red line denotes the mean of the distribution of the estimated distance
536 between the mobile device and its serving AP, and the pink region corresponds to the standard deviation
537 around the estimated distance. Note that theoretically both the red line and the corresponding pink region
538 are circular rings, but in practice this region’s geometry may not resemble a circle due to some reasons,
539 e.g. the varying signal strength distribution. The path loss model can also position the device in a similar
540 way, but with much worse accuracy than that of D-Log, which is theoretically analysed in Section 3.5. The
541 application of D-Log is highlighted in inset (right), showing localization improvement (the dark cyan line
542 and the corresponding light cyan region) over simple service area positioning approximated by a Voronoi
543 polygon [49] (thick blue line) and adjusted Voronoi regions (orange line), each centered on a single AP, that
544 encompass all the points that are closest to that AP and accessible to the visitors based on the floorplan
545 layout data [50]. Specifically, take the test mobile device near the bottom as an example. The corresponding
546 adjusted Voronoi region covers around $319 m^2$, and D-Log positions it in a circular region of approximately
547 $57 m^2$. By overlapping the D-Log positioning results with the adjusted Voronoi region, the localization
548 of the device is improved to a more accurate region of approximately $33 m^2$ as shown in Fig. 11 (right).
549 The computational cost of determining this enhanced region is only linearly proportional to the number of
550 locations considered.

551 Finally, like other RSSI based localization methods, the layouts of the environment or the configurations
552 of APs affect the proposed D-Log scheme. If they change, new AP logs need to be collected before positioning.
553 However, the layout does not change frequently, hence data collection and model re-training will occur only
554 as required.

555 6. Conclusions

556 In this paper, we investigated the following problem: *How to perform accurate indoor localization using*
557 *large-scale logs of discrete single-AP RSSI traces with low sampling rate?* We have provided a novel means
558 of post-hoc localization scheme, which is based on WiFi logs only, named the *D-Log* scheme, and proposed
559 two algorithms: the D-Log algorithm and the weighted D-Log algorithm, with D-Log focusing on accuracy
560 and weighted D-Log focusing on efficiency. While D-Log does not allow for the exact computation of
561 the coordinates of the user’s position, our contribution is to enhance the position estimation of post-hoc
562 localization based on logs of single-AP traces with infrequent sampling rates. D-Log emerges as a novel means
563 of localization enhancement which is simple and allows for improved estimation of the spatial context of the
564 device in an indoor environment. In addition, high absolute accuracy is not always necessary. Approaches
565 enabling contextual reasoning based on topological relationships of objects with approximate boundaries,

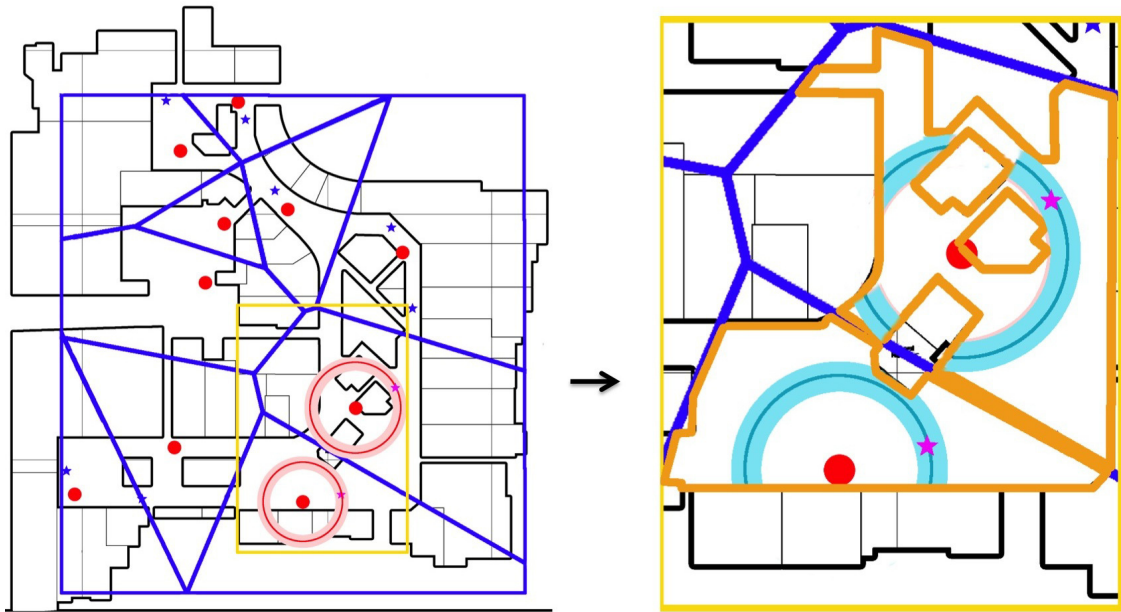


Figure 11: Illustration of the aim of D-Log (left), and how D-Log can help in reasoning about the tracked device location in spatial data management (right). The band around the ring indicates the accuracy of the D-Log positioning.

566 such as the egg-yolk model [51, 52] can be used to improve the estimate of the spatial context in which a user
 567 is active. We suggest that, by analysing spatial relations of vague regions [53], we can improve our estimates
 568 of spatial indoor behaviour of users and thus improve our estimates and predictions of indoor information
 569 needs [54, 5]. Coupled with detailed knowledge of the environmental layout, D-Log enables a substantially
 570 improved estimation of the likely space in which a user may be located. Together with other signal about
 571 the users behaviour (movement history, web browsing logs), D-Log enables sophisticated reasoning about
 572 the users' location. Accurate estimates of the indoor context (e.g., proximity to a specific shopping mall)
 573 are critical for the improvement of indoor services and have great economical potential in the near future.
 574 In the future, we plan to combine D-Log scheme with trilateration to get better localization performance.

575 Acknowledgement

576 This research is supported by a Linkage Project grant of the Australian Research Council (LP120200413).

577 References

- 578 [1] G. Durgin, T. S. Rappaport, H. Xu, Measurements and models for radio path loss and penetration loss in and around
 579 homes and trees at 5.85 GHz, *IEEE Transactions on Communications* 46 (11) (1998) 1484–1496. doi:10.1109/26.729393.
 580 [2] Cisco Systems Inc., Location Tracking Approaches, in: *Wi-Fi Location-Based Services 4.1 Design Guide*, 2008, Ch. 2.
 581 URL <http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Mobility/WiFiLBS-DG.pdf>
 582 [3] J. A. Santana, E. Macías, Á. Suárez, D. Marrero, V. Mena, Adaptive Estimation of WiFi RSSI and Its Impact Over
 583 Advanced Wireless Services, *Mobile Networks and Applications* (2016) 1–13. doi:10.1007/s11036-016-0729-1.
 584 [4] A. Suárez, J. A. Santana, E. M. Macias-Lopez, V. E. Mena, J. M. Canino, D. Marrero, RSSI Prediction in WiFi Considering
 585 Realistic Heterogeneous Restrictions, *Network Protocols and Algorithms* 6 (4) (2014) 19. doi:10.5296/npa.v6i4.6066.
 586 URL <http://www.macrothink.org/journal/index.php/npa/article/view/6066>
 587 [5] Y. Ren, M. Tomko, K. Ong, B. Yuntian, M. Sanderson, The influence of indoor spatial context on user information
 588 behaviours, in: M.-D. Albakour, C. Macdonald, I. Ounis, C. L. A. Clarke, V. Bicer (Eds.), *Workshop on Information
 589 Access in Smart Cities*, held in conjunction with the 36th European Conference on Information Retrieval ECIR 2014,
 590 ACM, 2014.

- 591 [6] A. Misra, R. K. Balan, LiveLabs: Initial reflections on building a large-scale mobile behavioral experimentation testbed,
592 SIGMOBILE Mob. Comput. Commun. Rev. 17 (4) (2013) 47–59. doi:10.1145/2557968.2557975.
593 URL <http://doi.acm.org/10.1145/2557968.2557975>
- 594 [7] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, A. T. Campbell, Nextplace: A spatio-temporal prediction framework for
595 pervasive systems, in: Proceedings of the 9th International Conference on Pervasive Computing, Pervasive'11, Springer-
596 Verlag, Berlin, Heidelberg, 2011, pp. 152–169.
- 597 [8] Y. Ren, M. Tomko, F. Salim, K. Ong, M. Sanderson, Analyzing Web Behavior in Indoor Retail Spaces, Journal of the
598 Association for Information Science and Technology.
- 599 [9] C. Luo, H. Hong, M. C. Chan, PiLoc: A self-calibrating participatory indoor localization system, in: IPSN-14 Proceedings
600 of the 13th International Symposium on Information Processing in Sensor Networks, 2014, pp. 143–153. doi:10.1109/
601 IPSN.2014.6846748.
- 602 [10] K. P. Subbu, B. Gozick, R. Dantu, LocateMe: Magnetic-fields-based indoor localization using smartphones, ACM Trans.
603 Intell. Syst. Technol. 4 (4) (2013) 73:1–73:27. doi:10.1145/2508037.2508054.
604 URL <http://doi.acm.org/10.1145/2508037.2508054>
- 605 [11] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, R. Sen, Zee: Zero-effort crowdsourcing for indoor localization, in:
606 Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, Mobicom '12, ACM,
607 New York, NY, USA, 2012, pp. 293–304. doi:10.1145/2348543.2348580.
608 URL <http://doi.acm.org/10.1145/2348543.2348580>
- 609 [12] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, R. R. Choudhury, No need to war-drive: Unsupervised indoor
610 localization, in: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, MobiSys
611 '12, ACM, New York, NY, USA, 2012, pp. 197–210.
- 612 [13] J. T. Biehl, M. Cooper, G. Filby, S. Kratz, LoCo: A ready-to-deploy framework for efficient room localization using wi-fi,
613 in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14,
614 ACM, New York, NY, USA, 2014, pp. 183–187. doi:10.1145/2632048.2636083.
615 URL <http://doi.acm.org/10.1145/2632048.2636083>
- 616 [14] M. Youssef, M. Mah, A. Agrawala, Challenges: Device-free passive localization for wireless environments, in: Proceedings
617 of the 13th Annual ACM International Conference on Mobile Computing and Networking, MobiCom '07, ACM, New
618 York, NY, USA, 2007, pp. 222–229. doi:10.1145/1287853.1287880.
619 URL <http://doi.acm.org/10.1145/1287853.1287880>
- 620 [15] A. S. Paul, E. A. Wan, F. Adenwala, E. Schafermeyer, N. Preiser, J. Kaye, P. G. Jacobs, MobileRF: A robust device-
621 free tracking system based on a hybrid neural network HMM classifier, in: Proceedings of the 2014 ACM International
622 Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, ACM, New York, NY, USA, 2014, pp. 159–170.
623 doi:10.1145/2632048.2632097.
624 URL <http://doi.acm.org/10.1145/2632048.2632097>
- 625 [16] P. Castro, P. Chiu, T. Kremenek, R. R. Muntz, A probabilistic room location service for wireless networked environments,
626 in: Proceedings of the 3rd International Conference on Ubiquitous Computing, UbiComp '01, Springer-Verlag, London,
627 UK, UK, 2001, pp. 18–34.
628 URL <http://dl.acm.org/citation.cfm?id=647987.741335>
- 629 [17] A. Ruiz-Ruiz, H. Blunck, T. Prentow, A. Stisen, M. Kjaergaard, Analysis methods for extracting knowledge from large-scale
630 WiFi monitoring to inform building facility planning, in: 2014 IEEE International Conference on Pervasive Computing
631 and Communications (PerCom), 2014, pp. 130–138. doi:10.1109/PerCom.2014.6813953.
- 632 [18] S. Bell, W. R. Jung, V. Krishnakumar, Wifi-based enhanced positioning systems: Accuracy through mapping, calibration,
633 and classification, in: Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness,
634 ISA'10, ACM, New York, NY, USA, 2010, pp. 3–9.
- 635 [19] M. Azizyan, I. Constandache, R. Roy Choudhury, SurroundSense: Mobile phone localization via ambience fingerprinting,
636 in: Proceedings of the 15th Annual International Conference on Mobile Computing and Networking, MobiCom '09, ACM,
637 New York, NY, USA, 2009, pp. 261–272. doi:10.1145/1614320.1614350.
638 URL <http://doi.acm.org/10.1145/1614320.1614350>
- 639 [20] H. Bao, W.-C. Wong, A novel map-based dead-reckoning algorithm for indoor localization, Journal of Sensor and Actuator
640 Networks 3 (1) (2014) 44–63. doi:10.3390/jsan3010044.
641 URL <http://www.mdpi.com/2224-2708/3/1/44>
- 642 [21] A. T. Mariakakis, S. Sen, J. Lee, K.-H. Kim, Sail: Single access point-based indoor localization, in: Proceedings of the
643 12th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '14, ACM, New York, NY,
644 USA, 2014, pp. 315–328. doi:10.1145/2594368.2594393.
645 URL <http://doi.acm.org/10.1145/2594368.2594393>
- 646 [22] J. Hightower, G. Borriello, Particle filters for location estimation in ubiquitous computing: A case study, in: In Proceedings
647 of International Conference on Ubiquitous Computing (UbiComp), 2004, pp. 88–106.
- 648 [23] I. Sabek, M. Youssef, A. Vasilakos, ACE: An Accurate and Efficient Multi-Entity Device-Free WLAN Localization System,
649 IEEE Transactions on Mobile Computing 14 (2) (2015) 261–273. doi:10.1109/TMC.2014.2320265.
- 650 [24] A. Khan, S. K. A. Imon, S. K. Das, A novel localization and coverage framework for real-time participatory urban
651 monitoring, Pervasive and Mobile Computing 23 (2015) 122–138. doi:10.1016/j.pmcj.2015.07.001.
- 652 [25] F. Salim, M. Williams, N. Sony, M. Dela Pena, Y. Petrov, A. A. Saad, B. Wu, Visualization of wireless sensor networks
653 using ZigBee's received signal strength indicator (RSSI) for indoor localization and tracking, in: 2014 IEEE International
654 Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014, pp. 575–580. doi:
655 10.1109/PerComW.2014.6815270.

- 656 [26] A. Savioli, E. Goldoni, P. Savazzi, P. Gamba, Low complexity indoor localization in wireless sensor networks by UWB
657 and inertial data fusion, arXiv e-print 1305.1657 (may 2013).
- 658 [27] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, M. Philipose, Mapping and localization with rfid technology, in: ICRA'04.
659 2004 IEEE International Conference on Robotics and Automation, Vol. 1, IEEE, 2004, pp. 1015–1020.
- 660 [28] Y. Zhuang, Z. Syed, J. Georgy, N. El-Sheimy, Autonomous smartphone-based wifi positioning system by using access points
661 localization and crowdsourcing, *Pervasive and Mobile Computing* 18 (2015) 118–136. doi:10.1016/j.pmcj.2015.02.001.
- 662 [29] J. H. Hightower, G. Borriello, Location systems for ubiquitous computing, *Computer* 34 (8) (2001) 57–66.
- 663 [30] H. Liu, H. Darabi, P. Banerjee, J. Liu, Survey of wireless indoor positioning techniques and systems, *IEEE Transactions*
664 *on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37 (6) (2007) 1067–1080. doi:10.1109/TSMCC.
665 2007.905750.
- 666 [31] N. Swangmuang, P. Krishnamurthy, An effective location fingerprint model for wireless indoor localization, *Pervasive and*
667 *Mobile Computing* 4 (6) (2008) 836–850. doi:10.1016/j.pmcj.2008.04.005.
- 668 [32] E. Mok, G. Retscher, Location determination using WiFi fingerprinting versus WiFi trilateration, *Journal of Location*
669 *Based Services* 1 (2) (2007) 145–159. doi:10.1080/17489720701781905.
670 URL <http://dx.doi.org/10.1080/17489720701781905>
- 671 [33] M. Werner, L. Schauer, A. Scharf, Reliable trajectory classification using wi-fi signal strength in indoor scenarios, in:
672 *Position, Location and Navigation Symposium - PLANS 2014, 2014 IEEE/ION, 2014*, pp. 663–670. doi:10.1109/PLANS.
673 2014.6851429.
- 674 [34] A. K. M. Mahtab Hossain, H. Nguyen Van, W.-S. Soh, Utilization of user feedback in indoor positioning system, *Pervasive*
675 *and Mobile Computing* 6 (4) (2010) 467–481. doi:10.1016/j.pmcj.2010.04.003.
676 URL <http://www.sciencedirect.com/science/article/pii/S1574119210000416>
- 677 [35] B. Wang, S. Zhou, L. T. Yang, Y. Mo, Indoor positioning via subarea fingerprinting and surface fitting with received
678 signal strength, *Pervasive and Mobile Computing* 23 (2015) 43–58. doi:10.1016/j.pmcj.2015.06.011.
- 679 [36] Cisco Meraki, Understanding Wireless Performance and Coverage. URL [https://documentation.meraki.com/MR/WiFi](https://documentation.meraki.com/MR/WiFi_Basics_and_Best_Practices/Understanding_Wireless_Performance_and_Coverage)
680 [_Basics_and_Best_Practices/Understanding_Wireless_Performance_and_Coverage](https://documentation.meraki.com/MR/WiFi_Basics_and_Best_Practices/Understanding_Wireless_Performance_and_Coverage), Tech. rep.
- 681 [37] Cisco Systems Inc., Voice over Wireless LAN 4.1 Design Guide, Tech. Rep. Cisco Validated Design I (2010).
- 682 [38] O. Chapelle, Training a support vector machine in the primal., *Neural computation* 19 (5) (2007) 1155–1178. doi:
683 10.1162/neco.2007.19.5.1155.
- 684 [39] G. K. B. Richard A. Johnson, *Statistics: Principles and Methods*, 6th Edition, John Wiley and Sons, 2009.
- 685 [40] C. L. Wu, L. C. Fu, F. L. Lian, WLAN location determination in e-home via support vector classification, in: *IEEE*
686 *International Conference on Networking, Sensing and Control, 2004*, Vol. 2, 2004, pp. 1026–1031. doi:10.1109/ICNSC.
687 2004.1297088.
- 688 [41] P. Agrawal, N. Patwari, Kernel Methods For RSS-Based Indoor Localization, in: S. A. R. Zekavat, R. M. Buehrer (Eds.),
689 *Handbook of Position Location: Theory, Practice, and Advances*, first edit Edition, John Wiley & Sons, Inc., 2012, Ch. 14,
690 pp. 457–486.
- 691 [42] H. Zou, X. Lu, H. Jiang, L. Xie, A Fast and Precise Indoor Localization Algorithm Based on an Online Sequential Extreme
692 Learning Machine, *Sensors* 15 (1) (2015) 1804–1824. doi:10.3390/s150101804.
- 693 [43] S. N. Patel, K. N. Truong, G. D. Abowd, PowerLine Positioning : A Practical Sub-Room-Level, in: *UbiComp'06 Proceed-*
694 *ings of the 8th international conference on Ubiquitous Computing, 2006*, pp. 441–458.
- 695 [44] B. D. Ripley, *Pattern Recognition and Neural Networks*, 1st Edition, Cambridge University Press, 1996.
- 696 [45] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, *Statistics Surveys* 4 (2010) 40–79.
697 arXiv:0907.4728, doi:10.1214/09-SS054.
698 URL <http://eprints.pascal-network.org/archive/00006812/>
- 699 [46] G. Lui, T. Gallagher, B. Li, A. G. Dempster, C. Rizos, Differences in RSSI readings made by different Wi-Fi chipsets: A
700 limitation of WLAN localization, in: *2011 International Conference on Localization and GNSS, ICL-GNSS 2011, 2011*,
701 pp. 53–57. doi:10.1109/ICL-GNSS.2011.5955283.
- 702 [47] J. S. C. Turner, M. F. Ramli, L. M. Kamarudin, A. Zakaria, a. Y. M. Shakaff, D. L. Ndzi, C. M. Nor, N. Hassan, S. M.
703 Mamduh, The study of human movement effect on Signal Strength for indoor WSN deployment, in: *IEEE Conference on*
704 *Wireless Sensor (ICWISE), 2013*, pp. 30–35. doi:10.1109/ICWISE.2013.6728775.
- 705 [48] V. Otsason, A. Varshavsky, A. Lamarca, E. D. Lara, Accurate GSM Indoor Localization, in: *Proceeding UbiComp'05*
706 *Proceedings of the 7th international conference on Ubiquitous Computing, 2005*, pp. 141–158. doi:10.1007/11551201_9.
- 707 [49] A. Okabe, B. Boots, K. Sugihara, S. N. Chiu, *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams*, 2nd
708 Edition, *Wiley Series in Probability and Statistics*, John Wiley and Sons, Ltd., Chichester, UK, 1999.
- 709 [50] Y. B. Bai, S. Wu, Y. Ren, K. Ong, G. Retscher, A. Kealy, M. Tomko, M. Sanderson, H. Wu, K. Zhang, A new approach
710 for indoor customer tracking based on a single wi-fi connection, in: *Fifth International Conference on Indoor Positioning*
711 *and Indoor Navigation IPIN2014, IEEE, 2014*.
- 712 [51] A. G. Cohn, N. M. Gotts, The egg-yolk representation of regions with indeterminate boundaries, *Geographic objects with*
713 *indeterminate boundaries* 2 (1996) 171–187.
- 714 [52] T. Beaubouef, F. Petry, Vagueness in spatial data: Rough set and egg-yolk approaches, in: L. Monostori, J. Vancza, M. Ali
715 (Eds.), *Engineering of Intelligent Systems*, Vol. 2070 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg,
716 2001, pp. 367–373. doi:10.1007/3-540-45517-5_41.
- 717 [53] A. J. Roy, J. G. Stell, Spatial relations between indeterminate regions, *International Journal of Approximate Reasoning*
718 *27* (3) (2001) 205–234.
- 719 [54] Y. Ren, K. Ong, M. Tomko, M. Sanderson, How people use the web in large indoor spaces, in: *2014 ACM International*
720 *Conference on Information and Knowledge Management CIKM 2014, ACM, 2014*, pp. 1879–1882.