Li *et al. EURASIP Journal on Wireless Communications and Networking* (2016) 2016:123
DOI 10.1186/s13638-016-0624-2

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Privacy-preserving crowdsourced site survey in WiFi fingerprint-based localization

Shujun Li[1]*, Hong Li[2] and Limin Sun[2]

Abstract

Typically, site survey is an inevitable phase for WiFi fingerprint-based localization which is regarded as one of the most promising techniques for indoor localization. However, the site survey can cause potential location privacy leakage for the participants who contribute their WiFi fingerprint measurements. In this paper, we propose a privacy-preserving site survey scheme for WiFi fingerprint-based localization. In the proposed scheme, we use homomorphic encryption to protect the location privacy of the participants which get involved in the site survey. Further, we employ differential privacy model to ensure that the released data will not breach an individual's location privacy regardless of whether she is present or absent in the site survey group. We theoretically analyze the security of the proposed scheme and use simulation experiments on a real-world data to validate the efficiency of the proposed scheme.

Keywords: Privacy, Crowdsourcing, Site survey, Differential privacy

1 Introduction

Due to the increasing demand for location-based services (LBSs) and the lack of GPS signals in indoor environments, indoor localization has become more and more popular in recent years. Researchers have proposed a vast range of approaches, among which WiFi fingerprint-based indoor localization is one of the most promising technologies [1–4]. A typical WiFi fingerprint-based localization algorithm consists of two phases, *offline site survey* and *online operating*. In the offline site survey phase, the service provider collects WiFi signal strengths from multiple access points (APs) at every location of an interested area. Next, in the online operating phase, a to-be-localized client measures the signal strengths at a specific location from nearby APs, and then algorithms such as *k-nearest neighbors* [1, 3] or *probability-based algorithms* [5] are employed to infer the user's location based on the measured WiFi signal strengths.

Usually the site survey is conducted in a crowdsourced way [2, 6, 7]. Suppliers recruited by the service provider measure the WiFi signal strengths of nearby APs when they visit the places which the service provider is interested in, and then send the measured WiFi signal strengths and the corresponding locations to the service provider. The service provider aggregates the data contributed by the suppliers to estimate the parameters which will be used in the online operating phase. The parameters which need to be estimated depend on the algorithms used in the *online operating phase*. In *k-nearest neighbor*-based algorithms, the mean of the WiFi signal strength of every AP at every location needs to be estimated, while in the *probability*-based algorithms, both the mean and variance of the signal strength of every AP are required. Crowdsourcing is an efficient way to conduct the site survey, but the measurements contributed by the suppliers will inevitably leak their location privacy. The service provider can infer the locations that the suppliers visit based on the data they contribute. Existing research indicates that location traces can leak information about the individuals' habits, interests, activities, and relationships [8, 9]. Consequently, the loss of location privacy can expose the suppliers to unwanted advertisements and location-based spams/scams and may cause social reputation or economic damage to the suppliers and can make the victims of blackmails or even physical violence.

Several approaches have been proposed to address the privacy issues of indoor localization algorithms. In [10],

*Correspondence: jojo8086@126.com
[1] School of Information Engineering, Yancheng Teachers University, Yancheng, China
Full list of author information is available at the end of the article

Shu et al. studied the privacy issues in range-based localization algorithms and proposed a scheme to protect users' privacy during the localization process. In [11], Wang et al. developed a privacy-preserving fuzzy localization scheme with CSI (Channel State Information) fingerprint. These privacy-preserving schemes were not designed for WiFi fingerprint-based localization algorithms; thus, they cannot be used to address the privacy issues presented in this paper. The most closely related work to this paper is that of Li et al. [12] which proposed a privacy-preserving scheme to address the privacy issues of the online operating phase in WiFi fingerprint-based localization algorithms . However, they did not consider the privacy leaks of the offline site survey phase.

In this paper, we propose a privacy-preserving site survey scheme which can protect the suppliers' location privacy in crowdsourcing-based site survey for WiFi fingerprint-based localization and, at the same time, can ensure the usability of the aggregated result for the service provider. Under this scheme, all the suppliers involved in the site survey form a group and they cooperate with each other to hide their measurements from the service provider based on homomorphic encryption. Further, every supplier releases her measurements in a differential private manner to guarantee that the released data will not breach an individual's location privacy regardless of whether she is present or absent in the group. The contributions of this paper are summarized as follows:

- To the best of our knowledge, this work is the first to address the privacy issues of the site survey in WiFi fingerprint-based localization algorithms.
- We propose a privacy-preserving site survey scheme for WiFi fingerprint-based localization based on homomorphic encryption and differential privacy model.
- We theoretically analyze the security of the proposed scheme and carry out simulation experiments on a real-world dataset to evaluate the performance of our scheme.

The rest of the paper is organized as follows. We first discuss the related work and introduce the background. Then, we present the detailed design of our scheme and the security analysis. Finally, we report the evaluation results and conclude this paper.

2 Related work

Location privacy in LBSs has been widely studied in the literature. In general, all the existing works can be classified into two categories: *privacy-preserving service request* and *privacy-preserving localization*.

2.1 Privacy-preserving service request

In LBSs, users send their locations to the service provider to get the services, which will inevitably leak their privacy. Many schemes have been proposed to protect users' location privacy when they request the location-based services. k-anonymity [13, 14] provides a form of plausible deniability by ensuring that the client cannot be individually identified from a group of k clients. Mix zone-based schemes [15] divide the whole region into application and mix zones. Clients report their locations in application zones and receive new, unused pseudonyms at mix zones. Cryptography-based approaches [16, 17] protect users' location privacy based on secure multiparty computation protocols. Since all the above schemes focus on the privacy issues when users request the location services, thus they cannot be used to address the privacy issues we discuss in this paper.

2.2 Privacy-preserving localization

To address the location privacy issues in localization, Shu et al. [10] addressed the privacy leakage problem for range-based localization algorithms, thus preventing the leakage of the location information of both the target and the anchors. Wang et al. [11] developed a privacy-preserving fuzzy localization scheme with CSI fingerprint using homomorphic encryption and fuzzy logic. These privacy-preserving schemes were not designed for WiFi fingerprint-based localization; thus, they cannot be used to address the privacy issues presented in this paper. Li et al. [12] studied the privacy issues in WiFi fingerprint-based localization and proposed a privacy-preserving scheme to protect both the users' and the service provider's privacy during the online operating phase. However, they did not consider the privacy leaks during the site survey phase.

3 Background

3.1 WiFi fingerprint-based localization

The process of WiFi fingerprint-based localization can be divided into two phases: offline site survey phase and online operating phase. In the offline site survey phase, a supplier u_i recruited by the service provider measures the WiFi signal strengths V_s^i of nearby APs when they visit a place l_s and send (l_s, V_s^i) to the service provider which aggregates the measurements and estimates the parameters which will be used in the online operating phase. In the online operating phase, a to-be-localized user measures the WiFi signal strengths at her current location, denoted as $V' = \left(v'_1, v'_2, \ldots, v'_j, \ldots, v'_N\right)$. Then, the service provider uses *k-nearest neighbors* or *probability-based algorithms* to determine the location of the user.

In *k-nearest neighbor*-based algorithms [1, 3], the service provider estimates the average WiFi signal strengths $\overline{V_s}$ at every location l_s based on the suppliers' measurements and stores $(l_s, \overline{V_s})$ in the WiFi fingerprint database. In the online operating phase, k-nearest neighbors of V' are identified from the database to estimate the location of the user. In *probability-based* algorithms [18, 19], the location loc of the user is determined based on the Bayes' theorem

$$\begin{aligned} \text{loc} &= \arg\max_{l\in L} P\left(l|V'\right) \\ &= \arg\max_{l\in L} P\left(l|v'_1, v'_2, \ldots, v'_n\right) \\ &= \arg\max_{l\in L} \left(P\left(l\right) \cdot \prod_{i=1}^{n} P(v'_i|l)\right). \end{aligned} \tag{1}$$

A common assumption is that the signal strength of AP_i at location l follows a normal distribution parameterized with mean μ and variance δ. The parameters μ and δ are estimated by the service provider based on the measurements of the suppliers.

3.2 Differential privacy

The concept of differential privacy is originally introduced by Dwork [20]. Differential privacy ensures that a supplier is not at increasing risk of privacy when she participates in a certain statistical database. An algorithm $\mathcal{A}$ is ϵ-differential privacy, if for any datasets D_1 and D_2, where D_1 and D_2 differ in at most one record, and for all subsets of possible answers $S \subseteq \text{Range}(\mathcal{A})$,

$$\Pr(\mathcal{A}(D_1) \in S) \le e^{\epsilon} \cdot \Pr(\mathcal{A}(D_2) \in S). \tag{2}$$

The above equation indicates that the output of $\mathcal{A}$ is insensitive to the modification of any single user's data in the datasets (including its removal or addition). The parameter ϵ allows us to control the balance between the level of privacy and the data utility. A smaller ϵ implies stronger privacy. One common way to achieve differential privacy is to add Laplace noises to the original output of $\mathcal{A}$ according to the following theorem.

Theorem 1. *For all $f : D \rightarrow R^d$, the following mechanism $\mathcal{A}$ is ϵ-differential private: $\mathcal{A}(D) = f(D) + \mathcal{L}(\Delta(f)/\epsilon)$, where $\mathcal{L}(\Delta(f)/\epsilon)$ is an independently generated random variable following the Laplace distribution and $\Delta(f)$ denotes the global sensitivity of f, which is defined as follows:*

$$\Delta f = \max_{D_1,D_2} \left|f(D_1) - f(D_2)\right| \tag{3}$$

for all D_1 and D_2 differing in at most one record.

3.3 The Paillier cryptosystem

In this work, we employ the Paillier cryptosystem as our cryptographic primitive. Invented by Pascal Paillier [21], the Paillier cryptosystem is a probabilistic asymmetric algorithm based on the decisional composite residuosity problem. Paillier cryptosystem is summarized below to facilitate the understanding of our algorithm.

- *Key generation:* To construct the public and private keys, one first chooses two large primes p, q of equivalent length and computes $N = pq$, $\lambda = lcm(p-1, q-1)$, $g = N+1$, and $\mu = \varphi(N)^{-1} \mod n$, where $\varphi(N) = (p-1)(q-1)$. The public key PK and private key PR are (N, g) and (λ, μ), respectively.
- *Encryption:* Let m be the plaintext to be encrypted. We denote the ciphertext of m by $E(m)$, which is given by
  $$E(m) = g^m r^N \mod N^2, \tag{4}$$
  where $r \in \mathbb{Z}_N$ is a random number.
- *Decryption:* Let c be the ciphertext, the plaintext $D(m)$ is obtained by
  $$D(m) = L(c^{\lambda} \mod N^2)\mu \mod N. \tag{5}$$

The Paillier cryptosystem is additively homomorphic. Given only the public key, one can compute $E(m_1 + m_2)$ from $E(m_1)$ and $E(m_2)$ as follows:

$$E(m_1 + m_2 \mod N) = E(m_1) \cdot E(m_2) \mod N^2. \tag{6}$$

4 System model and problem formulation

4.1 System model

A typical scenario of crowdsourcing-based site survey in WiFi fingerprint-based localization is depicted in Fig. 1. In general, there are n suppliers and an aggregator (i.e., the service provider). The suppliers could be volunteers or workers recruited by the service provider. Every

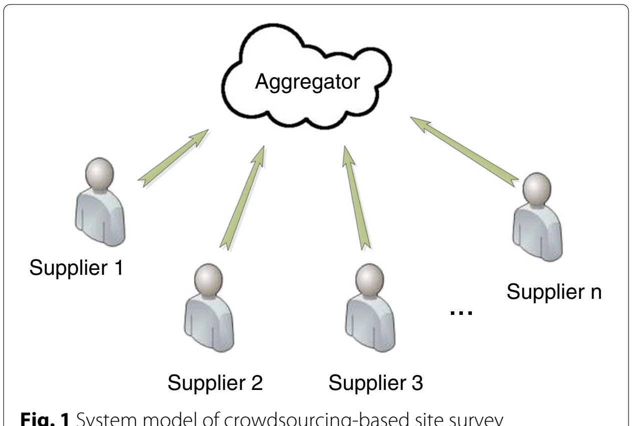

Fig. 1 System model of crowdsourcing-based site survey

supplier u_i records the WiFi signal strengths $V_i^s = (v_i^{s1}, v_i^{s2}, \ldots, v_i^{sj}, \ldots)$ when she visits location l_s and stores (l_s, V_i^s) in her local database V_i, where v_i^{sj} is the measured WiFi signal strength of the jth AP, $1 \leq i \leq n$, $l_s \in L$, and L is a location set defined by the service provider. The aggregator collects the measurements from the suppliers and would like to estimate the mean and variance of the WiFi signal strengths of every AP at every specific location in L based on the measurements of the suppliers.

4.2 Design goal

In the crowdsourcing-based site survey, the aggregator estimates the parameters based on the data (i.e., the measured WiFi signal strengths and the corresponding locations) contributed by the suppliers. However, the released data inevitably leak the location privacy of the suppliers. The aggregator can learn the locations that the suppliers visit. The goal of this paper is to ensure that the aggregator can estimate the mean and variance of WiFi signal strengths of every AP at every location in L, and at the same time, the location privacy of the suppliers is not compromised. In detail, we want to achieve the following privacy goals:

- *Location privacy:* Our scheme should ensure that the aggregator cannot learn the locations that the suppliers visited before. Also, the WiFi signal strengths collected by the suppliers should not be revealed, since the aggregator can infer their location privacy based on their measured WiFi signal strengths.
- *Differential privacy:* In the crowdsourcing-based site survey, even though the measurements of every supplier are completely hidden from the aggregator, it still can infer the location privacy of a supplier u_i by comparing the aggregating result when the u_i is in the site survey group and that when u_i is not in the site survey group.[1] Therefore, our scheme should achieve differential privacy which has been accepted as a standard for privacy preservation [20, 22]. Differential privacy can guarantee that the aggregator can retrieve information about any supplier only up to a predefined threshold, no matter what auxiliary information it knows about that supplier.

In this paper, we adopt the "honest-but-curious" model which assumes that each player honestly follows the designated protocols and procedures while it intends to disclose the other's private information.

5 Privacy-preserving site survey

In this section, we present a novel privacy-preserving crowdsourcing-based site survey scheme which can estimate the distribution of the WiFi signal strengths at each specific location without leaking the privacy of each supplier. The proposed scheme consists of four phases which are detailed as follows.

5.1 Preparation and initiation

In this phase, n suppliers and an aggregator form a site survey group. Within this site survey group, every supplier u_i generates its public key PK_i and private key PR_i using the Paillier cryptosystem and then sends the public key PK_i to other suppliers and the aggregator. The above process can be executed offline and only needs to be performed once. If the aggregator wants to estimate the mean and variance of the WiFi signal strengths from the jth AP at location l_s, it sends a request with $< \mathrm{AP}_j, l_s >$ to every supplier in this group.

5.2 Adding noises

After receiving the aggregator's request, every supplier u_i first queries its local dataset V_i to get a tuple (m_i, v_i^{sj}), where m_i indicates whether the supplier u_i visited location l_s before and v_i^{sj} is the measured WiFi signal strength of the jth AP at location l_s. If the supplier u_i visited location l_s before, m_i is set to 1 and v_i^{sj} is set to the measured WiFi signal strength. If the supplier u_i never visited location l_s before, m_i and v_i^{sj} are both set to 0.

To ensure that the presence or absence of the supplier u_i in the site survey group will not significantly increase her chance of being compromised (i.e., to achieve ϵ-differential privacy), every supplier u_i adds appropriately chosen random noises to v_i^{sj} and m_i as follows:

$$v_i' = v_i^{sj} + \mathcal{G}_1(n, \lambda_1) - \mathcal{G}_2(n, \lambda_1), \tag{7}$$

$$m_i' = m_i + \mathcal{G}_3(n, \lambda_2) - \mathcal{G}_4(n, \lambda_2), \tag{8}$$

where $\mathcal{G}_1(n, \lambda_1)$ and $\mathcal{G}_2(n, \lambda_1)$ are independent and identically distributed (i.i.d.) random variables having gamma distribution with probability density function (PDF)

$$g(x, n, \lambda_1) = \frac{(1/\lambda_1)^{1/n}}{\Gamma(1/n)} x^{\frac{1}{n}-1} e^{-x/\lambda_1}, \tag{9}$$

and $\mathcal{G}_3(n, \lambda_2)$ and $\mathcal{G}_4(n, \lambda_2)$ are i.i.d. random variables having gamma distribution with PDF

$$g(x, n, \lambda_2) = \frac{(1/\lambda_2)^{1/n}}{\Gamma(1/n)} x^{\frac{1}{n}-1} e^{-x/\lambda_2}. \tag{10}$$

$\lambda_1 = \Delta f_1/\epsilon$ and $\lambda_2 = \Delta f_2/\epsilon$, where Δf_1 and Δf_2 are the global sensitivity of the WiFi signal strength and m, respectively. Since the WiFi signal strength ranges from -90 to 0 dbm and $m \in \{0, 1\}$, we set $\Delta f_1 = 90$ and $\Delta f_2 = 1$. The parameter ϵ controls the trade-off between the desired privacy level and the data utility. A smaller ϵ yields a stronger privacy guarantee but generates more noises. In the evaluation section, we will investigate the impact of ϵ on the data utility. We will prove that the proposed scheme can achieve ϵ-differential privacy in the next section.

5.3 Encrypting data

After adding noises to her data, every supplier employs secret sharing [23] and Paillier cryptosystem to hide her data from the aggregator. For simplicity, we only demonstrate how to hide v'_i. The way to hide m'_i is the same. Each supplier u_i first splits v'_i into n random shares as follows:

$$\begin{aligned}
&u_1 : v'_1 = v'_{11} + v'_{12} + \dots v'_{1i} + \dots v'_{1n} \mod \eta \\
&u_2 : v'_2 = v'_{21} + v'_{22} + \dots v'_{2i} + \dots v'_{2n} \mod \eta \\
&\dots \\
&u_i : v'_i = v'_{i1} + v'_{i2} + \dots v'_{ii} + \dots v'_{in} \mod \eta \\
&\dots \\
&u_n : v'_n = v'_{n1} + v'_{n2} + \dots v'_{ni} + \dots v'_{nn} \mod \eta,
\end{aligned} \tag{11}$$

where η is a large integer. Then, each supplier u_i keeps v'_{ii} for herself, encrypts v'_{ij} using the public key of supplier u_j, and then sends $E_{\mathrm{PK}_j}(v'_{ij})$ to the aggregator. After the aggregator receives the encrypted shares from all the suppliers, she adds the shares which are encrypted by the same public key based on the additively homomorphic property of the Paillier cryptosystem as follows:

$$\begin{aligned}
&V_1 = E_{\mathrm{PK}_1}(\sum_{j\neq 1} v'_{j1}) = \prod_{j\neq 1} E_{\mathrm{PK}_1}(v'_{j1}) \\
&V_2 = E_{\mathrm{PK}_2}(\sum_{j\neq 2} v'_{j2}) = \prod_{j\neq 2} E_{\mathrm{PK}_2}(v'_{j2}) \\
&\dots \\
&V_i = E_{\mathrm{PK}_i}(\sum_{j\neq i} v'_{ji}) = \prod_{j\neq i} E_{\mathrm{PK}_i}(v'_{ji}) \\
&\dots \\
&V_n = E_{\mathrm{PK}_n}(\sum_{j\neq n} v'_{jn}) = \prod_{j\neq n} E_{\mathrm{PK}_n}(v'_{jn}).
\end{aligned} \tag{12}$$

Then, the aggregator sends V_i to the supplier u_i. Every supplier u_i decrypts V_i using her secret key PR_i, and adds her share v'_{ii} to $D_{\mathrm{PK}_i}(V_i)$ to get $V'_i = \sum_{j=1}^{n} v'_{ji}$ in plaintext and sends V'_i to the aggregator. Adding all $V'_i (1 \leq i \leq n)$ together, the aggregator can get $V' = \sum_{i=1}^{n} V'_i$ which is equal with $\sum_{i=1}^{n} v'_i$. We will prove its correctness in next section. In the same way, every supplier can hide m'_i from others, but the aggregator can get $M' = \sum_{i=1}^{n} m'_i$.

5.4 Estimating the parameters

After getting $V' = \sum_{i=1}^{n} v'_i$ and $M' = \sum_{i=1}^{n} m'_i$, the aggregator can estimate the mean (denoted as μ') of the WiFi signal strengths of AP_t at location l_s as follows:

$$\mu' = \frac{V'}{M'} = \frac{\sum_{i=1}^{n} v'_i}{\sum_{i=1}^{n} m'_i}. \tag{13}$$

Since every supplier adds controlled noises to her data, the estimated mean μ' is not exactly the same as $\mu = \sum_{i=1}^{n} v_i^{sj} / \sum_{i=1}^{n} m_i$. The estimation error is controlled by the parameter ϵ. We will investigate the impact of ϵ on the estimation errors and show that the localization accuracy when we use μ' is comparable with that when we use μ in most cases.

To estimate the variance δ', the aggregator send μ' back to every supplier u_i which can then get δ_i as follows:

$$\delta_i = \begin{cases} (v_i^{sj} - \mu')^2, & \text{if } m_i = 1 \\ 0, & \text{if } m_i = 0. \end{cases} \tag{14}$$

Following the same rules described above, every supplier u_i adds a random noise $\mathcal{G}_5(n, \lambda_1) - \mathcal{G}_6(n, \lambda_3)$ to δ_i to get $\delta'_i = \delta_i + \mathcal{G}_5(n, \lambda_1) - \mathcal{G}_6(n, \lambda_3)$, and then sends δ'_i to the aggregator in a secret way. $\mathcal{G}_5(n, \lambda_1)$ and $\mathcal{G}_6(n, \lambda_3)$ are i.i.d. random variables having gamma distribution with PDF

$$g(x, n, \lambda_3) = \frac{(1/\lambda_3)^{1/n}}{\Gamma(1/n)} x^{\frac{1}{n}-1} e^{-x/\lambda_3}, \tag{15}$$

where $\lambda_3 = 90^2/\epsilon$. Following the same rules above, the aggregator computes $\sum_{i=1}^{n} \delta'_i$ without knowing every δ'_i, and the variance of the WiFi signal strengths of AP_t at location l_s can be estimated by $\delta' = \sum_{i=1}^{n} \delta'_i / M'$.

6 Theoretical analysis of the proposed scheme

In this section, we will theoretically analyze the correctness of the proposed scheme and prove that the proposed scheme can achieve the desired privacy goals.

6.1 The correctness of the scheme

In our scheme, we employ secret sharing and homomorphic encryption to protect the privacy of every supplier. Every supplier u_i splits her data v'_i into n shares and submits V'_i to the aggregator. Adding all $V'_i (1 \leq i \leq n)$ together, the aggregator can get $V' = \sum_{i=1}^{n} V'_i$. We claim that $V' = \sum_{i=1}^{n} V'_i$ is equal to $\sum_{i=1}^{n} v'_i$, which is supported by the following theorem.

Theorem 2. *Only given $V'_i (1 \leq i \leq n)$, the aggregator can correctly compute $\sum_{i=1}^{n} v'_i$ by adding $V'_i (1 \leq i \leq n)$ together.*

Proof. As described above, $V_i' = D_{\mathrm{PK}_i}(V_i) + v_{ii}'$ thus, we have

$$\sum_{i=1}^{n} V_i' = \sum_{i=1}^{n} \left(D_{\mathrm{PK}_i}(V_i) + v_{ii}'\right)$$
$$= \sum_{i=1}^{n} \left(D_{\mathrm{PK}_i} \left(\prod_{j\neq i} E_{\mathrm{PK}_i}(v_{ji}') \right) + v_{ii}' \right)$$

Applying the additively homomorphic property of Paillier cryptosystem, we have $\prod\limits_{j\neq i} E_{\mathrm{PK}_i}(v_{ji}') = E_{\mathrm{PK}_i}(\sum\limits_{j\neq i} v_{ji}')$, thus

$$\sum_{i=1}^{n} V_i' = \sum_{i=1}^{n} \left(D_{\mathrm{PK}_i} \left(E_{\mathrm{PK}_i}(\sum_{j\neq i} v_{ji}') \right) + v_{ii}' \right)$$
$$= \sum_{i=1}^{n} \left(\sum_{j\neq i} v_{ji}' + v_{ii}' \right)$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ji}' = \sum_{i=1}^{n} v_i'$$

Then, we have $\sum_{i=1}^{n} V_i' = \sum_{i=1}^{n} v_i'$, which proves Theorem 2. □

Following the same rules, we can also prove that the aggregator can correctly compute $\sum_{i=1}^{n} m_i'$ and $\sum_{i=1}^{n} \delta_i'$. Therefore, the correctness of the proposed scheme is proved.

6.2 The security of the scheme

In the proposed scheme, every supplier adds random noises having gamma distribution to achieve differential privacy and further employs secret sharing to hide her data. We claim that the proposed scheme can achieve the desired privacy goals, which is supported by the following two Theorems:

Theorem 3. *The proposed scheme satisfies ϵ-differential privacy.*

Proof. In the proposed scheme, $v_i' = v_i + \mathcal{G}_1(n, \lambda_1) - \mathcal{G}_2(n, \lambda_1)$, where $\mathcal{G}_1(n, \lambda_1)$ and $\mathcal{G}_2(n, \lambda_1)$ are i.i.d. random variables having gamma distribution with PDF

$$g(x, n, \lambda_1) = \frac{(1/\lambda_1)^{1/n}}{\Gamma(1/n)} x^{\frac{1}{n}-1} e^{-x/\lambda_1} \tag{16}$$

thus, we have

$$\sum_{i=1}^{n} v_i' = \sum_{i=1}^{n} \left(v_i + \mathcal{G}_1(n, \lambda_1) - \mathcal{G}_2(n, \lambda_1)\right)$$
$$= \sum_{i=1}^{n} v_i + \sum_{i=1}^{n} \left(\mathcal{G}_1(n, \lambda_1) - \mathcal{G}_2(n, \lambda_1)\right).$$

Let $\mathcal{L}(\lambda)$ denote a random variable which has a Laplace distribution with PDF $f(x, \lambda_1) = \frac{1}{2\lambda_1} e^{\frac{|x|}{\lambda_1}}$. According to [24], the distribution of $\mathcal{L}(\lambda_1)$ is infinitely divisible. Furthermore, for every integer $n \geq 1$, $\mathcal{L}(\lambda_1) = \sum_{i=1}^{n} [\mathcal{G}_1(n, \lambda_1) - \mathcal{G}_2(n, \lambda_1)]$, where $\mathcal{G}_1(n, \lambda_1)$ and $\mathcal{G}_2(n, \lambda_1)$ are i.i.d. random variables having gamma distribution with PDF $g(x, n, \lambda_1) = \frac{(1/\lambda_1)^{1/n}}{\Gamma(1/n)} x^{\frac{1}{n}-1} e^{-x/\lambda_1}$, where $x \geq 0$. Thus, we have

$$\sum_{i=1}^{n} v_i' = \sum_{i=1}^{n} v_i + \sum_{i=1}^{n} \mathcal{L}(\lambda_1).$$

Similarly, we can have

$$\sum_{i=1}^{n} m_i' = \sum_{i=1}^{n} m_i + \sum_{i=1}^{n} \mathcal{L}(\lambda_2)$$

$$\sum_{i=1}^{n} \delta_i' = \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} \mathcal{L}(\lambda_3),$$

where $\mathcal{L}(\lambda_2)$ and $\mathcal{L}(\lambda_3)$ are two random variables following the Laplace distribution with PDF $f(x, \lambda_1) = \frac{1}{2\lambda_2} e^{\frac{|x|}{\lambda_2}}$ and $f(x, \lambda_3) = \frac{1}{2\lambda_3} e^{\frac{|x|}{\lambda_3}}$, respectively. According to Theorem 1, the proposed scheme achieves ϵ-differential privacy. □

Theorem 4. *The proposed scheme can protect every supplier's location privacy.*

Proof. In the proposed scheme, every supplier u_i splits its data v_i', m_i', and δ_i' into n random shares and sends the other encrypted $n-1$ shares to the aggregator. Even the aggregator gets the plaintexts of the other $n-1$ shares, it still cannot know v_i', m_i', and δ_i' since u_i keeps one share after splitting the data. Therefore, the aggregator cannot figure out whether u_i visited l_s and the measured WiFi signal strength, which proves that the proposed scheme can protect every supplier's location privacy. □

7 Evaluation

In this section, we evaluate the performance of the proposed scheme. We focus on two important metrics in the evaluation: the utility of the aggregated data and the efficiency of the proposed scheme.

7.1 Experiment setup

We implement the supplier side of the proposed scheme on a Android platform with a Qualcomm Snapdragon600 Quad-Core 1.7 GHz CPU and 2 G RAM, and the aggregator side of the proposed scheme on a 32-bit computer with Intel i7 CPU of 3.4 GHz and 4 G memory. The Paillier modulus used in this work is set to 1024. In the experiments, we use a real-world WiFi fingerprint dataset to evaluate the performance of our algorithm. The dataset has total 1000 records which are collected in a typical indoor environment. Each record contains the WiFi signal strengths from nearby APs. The total number of APs used in the experiments is 10 (i.e., $n = 10$) and the total number of locations in the indoor environment is 76 (i.e., $|L| = 76$). In the simulations, the data are randomly distributed to the suppliers and then the aggregator tries to estimate the WiFi fingerprint at every location.

7.2 Utility evaluation

In the experiments, the aggregator estimates the WiFi fingerprint at every location based on the data with noises provided by the suppliers, and then uses the estimated data to offer localization service. In this section, we evaluate the impact of the added noises on the aggregated results and the accuracy of the localization.

In the proposed scheme, every supplier adds two random noises with gamma distribution to her measurements to achieve ϵ-differential privacy. In the experiments, we employ the Euclidean distance as the metric to evaluate the usability of the data with noises. Figure 2 presents the cumulative distribution function (CDF) of the Euclidean distance between the estimated WiFi fingerprint without noises and the estimated WiFi fingerprint with noises. It is observed that the accuracy of estimation increases when ϵ increases from 0.4 to 2.0, and 80 % of the Euclidean distances between the noisy WiFi fingerprint and the original WiFi fingerprint are smaller than 6. A smaller ϵ yields larger noises, which, on the other hand, provides a stronger privacy preservation. It is a trade-off between the utility of the data and the privacy.

Further, we investigate the impact of our privacy-preserving scheme on the localization accuracy. In this paper, we employ *k-nearest neighbors* to determine the unknown locations in the online operating phase. Figure 3 shows the CDF of the localization errors when ϵ is set to different values and when no noises are added to the data. It is observed that the localization accuracy increases when ϵ increases from 0.4 to 2.0 and the localization accuracy is the highest when no noises are added to the data. We also observed that 80 % of the localization errors are within 5 m, which implies a high data usability.

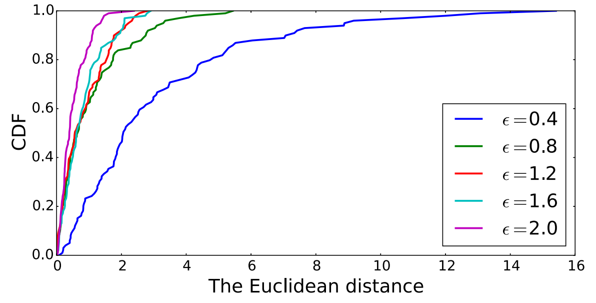

Fig. 2 CDF of the Euclidean distances

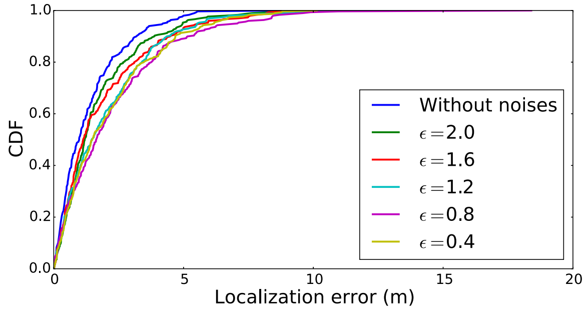

Fig. 3 CDF of the localization errors

7.3 Computational and communication overhead

In this work, we employ Paillier cryptosystem as our cryptographic primitive to protect the suppliers' privacy, which inevitably brings more computational and communication overhead. In this section, we evaluate the computational and communication cost of the proposed scheme. In the experiments, we set $\epsilon = 0.4$ and investigate the impact of the number of suppliers (i.e., n) on the computational time and communication overhead.

Figure 4 shows the time cost on the supplier side and the aggregator side for estimating the WiFi signal strength of one AP at every location. We can see that the computational time on the supplier side is proportional to the number of the suppliers. When the number of the suppliers is set to 10, the time cost on the supplier side is 1.6 s. When the number of the suppliers reaches 100, the time cost on the supplier side is 16 s. The computational time

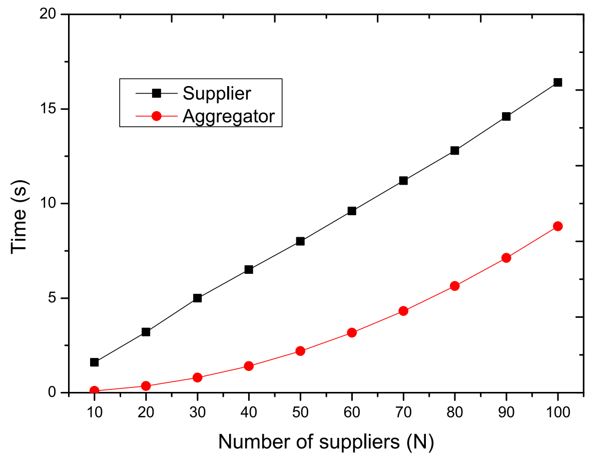

Fig. 4 The time cost of the suppliers and the aggregator

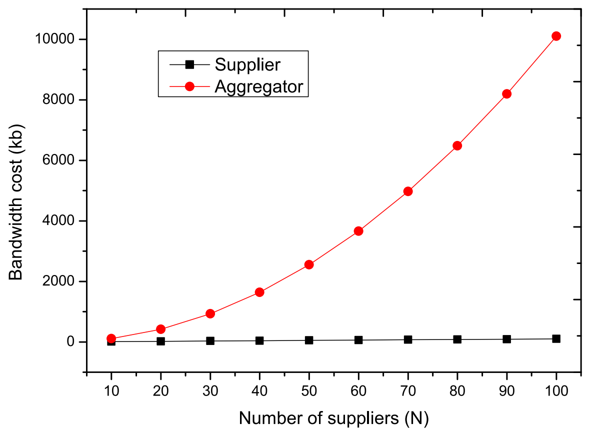

Fig. 5 The bandwidth cost of the suppliers and the aggregator

on the aggregator side is proportional to the square of the number of the suppliers. When the number of the suppliers is 10, the computational time on the aggregator side is only 0.08 s. However, when the number of the suppliers reaches 100, the computational time on the aggregator side becomes 8.8 s.

Figure 5 shows the impact of the number of the suppliers on the bandwidth cost of every supplier and the aggregator. It is observed that the bandwidth cost of the supplier is proportional to the number of the suppliers. The bandwidth cost is 10 kb for every supplier when the number of the supplier is 10. The bandwidth cost increases with the increase of the number of the supplers. The bandwidth cost reaches 101 kb for every supplier when the number of the supplier is 100. The bandwidth cost on the aggregator side is proportional to the square of the number of the suppliers. When the number of the suppliers is 10, the bandwidth cost of the aggregator is only 110 kb. When the number of the suppliers is 100, the bandwidth cost becomes 10100 kb.

8 Conclusions

In this work, we propose a privacy-preserving site survey scheme for WiFi fingerprint-based localization. The proposed scheme uses homomorphic encryption and differential privacy model to protect the location privacy of the participants which get involved in the site survey process of the WiFi fingerprint localization. We theoretically analyze the security of the scheme and use simulation experiments on real-world data to validate the efficiency of the proposed scheme.

Endnote

[1] For example, assume that the mean of the WiFi signal strengths estimated by the aggregator is μ_1 when u_i is in the group, and the mean of the WiFi signal strengths is μ_2 when u_i is not in the group. The aggregagor can get the measured WiFi signal strength of u_i by the formula $\mu_1 \cdot n - \mu_2 \cdot (n-1)$, where n is the number of suppliers in the group.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61472418).

Author details

[1] School of Information Engineering, Yancheng Teachers University, Yancheng, China. [2] Beijing Key Laboratory of IOT Information Security Technology, Institute of Information Engineering, CAS, Beijing, China.

Received: 18 December 2015 Accepted: 24 April 2016
Published online: 04 May 2016

References

1. P Bahl, VN Padmanabhan, in *Proc. of IEEE INFOCOM*. Radar: an in-building RF-based user location and tracking system, (2000), pp. 775–784
2. Z Yang, C Wu, Y Liu, in *Proc. of ACM MobiCom*. Locating in fingerprint space: wireless indoor localization with little human intervention, (2012), pp. 269–280
3. H Liu, Y Gan, J Yang, S Sidhom, Y Wang, Y Chen, F Ye, in *Proc. of ACM MobiCom*. Push the limit of WiFi based localization for smartphones, (2012), pp. 305–316
4. W Cheng, D Wu, X Cheng, D Chen, in *WASA*. Routing for information leakage reduction in multi-channel multi-hop ad-hoc social networks, (2012), pp. 31–42
5. D Milioris, L Kriara, A Papakonstantinou, G Tzagkarakis, P Tsakalides, M Papadopouli, in *Proceedings of the 13th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems*. Empirical evaluation of signal-strength fingerprint positioning in wireless LANs (ACM, 2010), pp. 5–13
6. J Niu, B Wang, L Cheng, JJ Rodrigues, in *Communications (ICC) 2015 IEEE International Conference on*. WicLoc: an indoor localization system based on WiFi fingerprints and crowdsourcing (IEEE, 2015), pp. 3008–3013
7. J Li, Z Cai, M Yan, Y Li, in *INFOCOM, year=2016 Proceedings IEEE*. Using crowdsourced data in location-based social networks to explore influence maximization (IEEE, 2016)
8. R Shokri, G Theodorakopoulos, J Le Boudec, J Hubaux, in *IEEE Symposium on Security and Privacy*. Quantifying location privacy, (2011), pp. 247–262
9. Y He, L Sun, W Yang, H Li, A game theory-based analysis of data privacy in vehicular sensor networks. Int. J. Distrib. Sens. Networks. **2014** (2014)
10. T Shu, Y Chen, J Yang, A Williams, in *INFOCOM, 2014 Proceedings IEEE*. Multi-lateral privacy-preserving localization in pervasive environments (IEEE, 2014), pp. 2319–2327
11. X Wang, Y Liu, Z Shi, X Lu, L Sun, *A privacy-preserving fuzzy localization scheme with CSI fingerprint*, (2016)
12. H Li, L Sun, H Zhu, X Lu, X Cheng, in *INFOCOM, 2014 Proceedings IEEE*. Achieving privacy preservation in WiFi fingerprint-based localization (IEEE, 2014), pp. 2337–2345
13. D Yang, X Fang, G Xue, in *Proc. of IEEE INFOCOM*. Truthful incentive mechanisms for k-anonymity location privacy, (2013), pp. 3094–3102
14. X Liu, K Liu, L Guo, X Li, Y Fang, in *Proc. of IEEE INFOCOM*. A game-theoretic approach for achieving k-anonymity in location based services, (2013), pp. 3085–3093
15. AR Beresford, F Stajano, in *Proc. of the IEEE PerSec*. Mix zones: user privacy in location-aware services, (2004), pp. 127–131
16. J Shao, R Lu, X Lin, in *INFOCOM, 2014 Proceedings IEEE*. FINE: a fine-grained privacy-preserving location-based service framework for mobile devices (IEEE, 2014), pp. 244–252
17. I Bilogrevic, M Jadliwala, K Kalkan, J-P Hubaux, I Aad, in *Privacy Enhancing Technologies*. Privacy in mobile computing for location-sharing-based services (Springer, 2011), pp. 77–96
18. A Chen, C Harko, D Lambert, P Whiting, An algorithm for fast, model-free tracking indoors. ACM SIGMOBILE Mob. Comput. Commun. Rev. **11**(3), 48–58 (2007)

19. D Milioris, G Tzagkarakis, A Papakonstantinou, M Papadopouli, P Tsakalides, Low-dimensional signal-strength fingerprint-based positioning in wireless LANs. Ad Hoc Netw. **12**, 100–114 (2014)
20. C Dwork, in *Encyclopedia of Cryptography and Security*. Differential privacy (Springer, 2011), pp. 338–340
21. P Paillier, in *Proc. of ACM EUROCRYPT*. Public-key cryptosystems based on composite degree residuosity classes, (1999)
22. ME Andrés, NE Bordenabe, K Chatzikokolakis, C Palamidessi, in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. Geo-indistinguishability: differential privacy for location-based systems (ACM, 2013), pp. 901–914
23. FD Garcia, B Jacobs, in *Security and Trust Management*. Privacy-friendly energy-metering via homomorphic encryption (Springer, 2011), pp. 226–238
24. S Kotz, T Kozubowski, K Podgorski, *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. (Springer Science & Business Media, 2012)