DP3: A Differential Privacy-Based Privacy-Preserving Indoor Localization Mechanism

Yufeng Wang, Minjie Huang, Qun Jin, and Jianhua Ma

***Abstract*— Wi-Fi fingerprint-based indoor localization is regarded as one of the most promising techniques for location-based services. However, it faces serious problem of privacy disclosure of both clients' location data and provider's fingerprint database. To address this issue, this letter proposes a differential privacy (DP)-based privacy-preserving indoor localization scheme, called DP3, which is composed of four phases: *access point* (AP) *fuzzification* and *location retrieval* in client side and *DP-based finger clustering* and *finger permutation* in server side. Specifically, in *AP fuzzification*, instead of providing the measured full finger (including AP sequence and the corresponding received signal strength), a to-be-localized (TBL) client only uploads the AP sequence to the server. Then, the localization server utilizes the DP-enabled clustering to build the fingerprints related to the AP sequence into k clusters, permutes these reference points in each cluster with exponential mechanism to mask the real positions of these fingerprints, and sends the modified data set to the TBL client. At client side, *location retrieval* phase estimates the location of the client. Theoretical and experimental results show that DP3 can simultaneously protect the location privacy of the TBL client and the data privacy of the localization server.**

***Index Terms*— Indoor localization, WiFi fingerprint, differential privacy, privacy-preserving.**

I. Introduction

In WiFi fingerprint-based indoor localization, the problem of privacy leakage including both clients' location privacy and data privacy of fingerprint database has attracted great attention. Specifically, for the former, the attacker could be an untrusted LBS provider owning the fingerprints database who collects clients' location and sells them for profit; for the latter, attackers may infer the valuable WiFi fingerprint database of provider through randomly generating abundant localization queries. To address this issue, Privacy-Preserving WiFi Fingerprint Localization scheme (PriWFL) was proposed in [1], which adopted full homomorphic encryption to protect the privacy of both entities, but this scheme employed a fixed number of APs, which may restrict the application scenario of PriWFL. Similarly, [2] proposed a cryptographic scheme with additive homomorphic encryption. Homomorphic encryption is widely used to ensure the data privacy in indoor localization system, but the application is limited for its large computational overhead.

Recently, DP has become an effective way to protect the sensitive information of entities by adding suitable noise in dataset. Considering the issue that, in cognitive radio networks, sharing sensing data among secondary users (SUs) may leak their locations, a privacy preservation framework called PrimCos with DP was proposed [8], which provided privacy guarantee for each SU. In a sense, the work above is similar with the site survey phase in WiFi fingerprint indoor localization, in that both schemes aim to protect the data contributors' location privacy. Our proposed scheme DP3 is intentionally designed for the online operating phase.

Reference [3] used homomorphic encryption to protect the location privacy of the participants in site survey phase, and DP was used to ensure that the released data will not breach an individual's location privacy. Reference [4] focused on the application of local DP (i.e. instead of protecting the privacy of dataset provider, ensure the privacy of the individual data contributor) to the collection of indoor positioning data. In brief, DP has been widely used in site survey for data collection, but not been used in online real-time operating phase. Intuitive, the challenge of directly employing DP into operating phase is the poor localization accuracy, due to the large amount of added noise. In literature, k-means clustering and DP are combined in data release to mask each user's exact locations as well as the frequency of visiting locations within a given privacy budget [5], [6]. Inspired by this idea, we propose a DP-based privacy-preserving indoor localization scheme (DP3) to balance the trade-off between data privacy and data utility in online real-time operating phase.

II. Proposed Scheme

A. DP3 Architecture

As shown in Fig.1, DP3 involves two entities, to-be-localized (TBL) client and localization server, and is composed of four phases, *AP Fuzzification* and *Location Retrieval* at client side, and *DP-based Finger Clustering* and *Finger Permutation* at server side. Detailed processes are described as follows.

B. Detailed Phases

In a fingerprint database $D_1 =< i, t_i : (x_i, y_i), \{rss_{ij}\}_{j\in n_i} >_{i=1}^{N}$, there are totally N reference points. $t_i : (x_i, y_i)$ is the coordinate of the i-th fingerprint (reference point i), n_i is the APs set associated with i, and rss_{ij} is the Received Signal Strength (RSS) from AP_j.

1) AP Fuzzification: In this phase, the TBL client measures the real-time WiFi fingerprint, including names of

Manuscript received August 4, 2018; revised September 7, 2018; accepted October 3, 2018. Date of publication October 16, 2018; date of current version December 10, 2018. This work was supported in part by NSFC under Grant 61801240, in part by Jiangsu Educational Bureau Project under Grant 14KJA510004, and in part by State Key Laboratory of Novel Software Technology under grant KFKT2017B14. The associate editor coordinating the review of this letter and approving it for publication was M. Khabbazian. *(Corresponding author: Yufeng Wang.)*

Y. Wang and M. Huang are with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: wfwang@njupt.edu.cn; huangminjie1028@163.com).

Q. Jin is with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokorozawa 359-1192, Japan (e-mail: jin@waseda.jp).

J. Ma is with the Digital Media Department, Faculty of Computer and Information Sciences, Hosei University, Tokyo 184-8584, Japan (e-mail: jianhua@hosei.ac.jp).

Digital Object Identifier 10.1109/LCOMM.2018.2876449

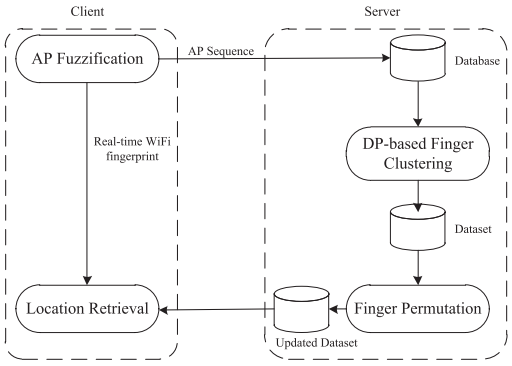

Fig. 1. The architecture of DP3 sheme.

APs and the corresponding RSS values, denoted as $< AP_j, RSS_{AP_j} >_{AP_j \in Sample}$. Then client only sends the AP sequence *Sample* to the server for the localization request.

Definition 1 (ϵ-Differential Privacy): A randomized mechanism M gives ϵ-DP if for both datasets D and D' differing in at most one record, and for all output dataset $S \subseteq Range(M)$,

$$Pr[M(D) \in S] \leq exp(\epsilon) \cdot Pr[M(D') \in S] \tag{1}$$

2) DP-Based Finger Clustering: After receiving the AP sequence, the records in fingerprint database D_1 pertaining to the AP sequence are extracted to form a specific dataset D. Then *DP-based Finger Clustering* groups the reference points into k clusters. Specifically, DP is used to mask the real center of each cluster. According to Definition 1, with the output of M, deleting or adding a location (D and D') will hardly influence the clustering results (S), which means that an attacker cannot infer which cluster a location belongs to in clustering output. Here ϵ is the privacy budget, used to control the balance between the privacy level and data utility. A smaller ϵ implies a stronger privacy guarantee. Two mechanisms are usually employed to achieve DP, which are Laplace mechanism and Exponential mechanism [7]. These mechanisms are concerned with the sensitivity, and in our *DP-based Finger Clustering* phase, Laplace mechanism is applied.

Definition 2 (Sensitivity): For a function f, the sensitivity of f is defined as: $\Delta f = max_{D,D'} ||f(D) - f(D')||_2$.

Laplace noise based on ϵ-DP mechanism is described as follows. Given a numeric function f working on the dataset D, the mechanism $M(D) = f(D) + Laplace(\frac{\Delta f}{\epsilon})$ provides the ϵ-DP, where $Laplace(\frac{\Delta f}{\epsilon})$ is consistent with Laplacian distribution. Moreover, in our *DP-based Fingerprint Clustering*, the sensitivity is calibrated by the maximal distance between any pair of locations, that is, the sensitivity $GS = max||(x_i, y_i) - (x_j, y_j)||_2$.

The pseudocode of iterative clustering process using k-means and Laplace mechanism is as Algorithm 1.

Adding Laplace noise changes the center of each cluster in each iterative round, which, in turn leads to further clustering process. Note that, the privacy budget $\frac{\epsilon}{2}$ is allocated to this *DP-base clustering* phase and correspondingly privacy budget allocated to each iterative round is $\frac{\epsilon}{2T}$.

3) Finger Permutation: This phase aims to guarantee that the fingerprint in D will not be re-identified. The replacement of the reference points utilizes the exponential mechanism. For a particular t_i in the cluster C_l, all locations of the reference points in C_l are included into the candidate set I_l, with different probabilities.

Algorithm 1 DP-Based Clustering Algorithm

Input: dataset D, number of cluster k, privacy budget $\frac{\epsilon}{2}$, iterative rounds T, the sensitive GS

Output: k DP-enabled cluste $C_1, C_2, \ldots, C_k$

1. Initially, pick k locations randomly as centers $S_1^{(0)}$, $S_2^{(0)}$, …, $S_k^{(0)}$ of clusters $C_1^{(0)}, C_2^{(0)}, \ldots, C_k^{(0)}$;

for $p = 1$ **to** T **do**

2. Assign all the reference points $t_1, t_2, \ldots, t_{|D|}$ in D to their nearest cluster;

3. For each cluster $C_j^{(p)}$, add noise to the sum of coordinates $ns_j^{(p)}$ and the number of reference points $nc_j^{(p)}$ as:

$$ns_j^{(p)} = \sum_{t_i \in C_j^{(p)}} (x_i, y_i) + Laplace(\frac{2T \cdot GS}{\epsilon})$$

$$nc_j^{(p)} = |C_j^{(p)}| + Laplace(\frac{2T \cdot GS}{\epsilon})$$

4. Update centers $S_1^{(p+1)}, S_2^{(p+1)}, \ldots, S_k^{(p+1)}$ of the k clusters as:

$$S_j^{(p+1)} = \frac{ns_j^{(p)}}{nc_j^{(p)}}$$

end for

5. After T iterative rounds, the clusering output is $C_1^{(T)}$, $C_2^{(T)}$, …, $C_k^{(T)}$, briefly denoted as $C_1, C_2, \ldots, C_k$.

Definition 3 (Exponential Mechanism): Given a function q working on the dataset I, the mechanism M gives ϵ-DP if $Pr[M(I) = t_j] \propto exp(\frac{\epsilon \cdot q(I, t_j)}{2\Delta q})$.

Here, $q(I, t_j)$ is a score function used to evaluate the usability of t_j as the output from I and Δq is the sensitivity of q.

According to Definition 3, in cluster C_l, each location queried from I_l is associated with a probability based on the score function and its related sensitivity. The probability of substituting t_j for $t_i (t_j, t_i \in I_l)$ is given with $\frac{\epsilon}{2}$ privacy budget as follows:

$$Pr_{t_j, t_i \in I_l}(t_j) = \frac{exp(\frac{\epsilon \cdot q_i(I_l, t_j)}{4 \cdot GS})}{\sum_{t_j, t_i \in I_l} exp(\frac{\epsilon \cdot q_i(I_l, t_j)}{4 \cdot GS})} \tag{2}$$

The score function $q_i(I_l, t_j)$ is defined by the distance between reference points.

$$q_i(I_l, t_j) = GS - ||(x_i, y_i) - (x_j - y_j)||_2 \tag{3}$$

Here, the sensitivity Δq is measured by the maximal change in the distance between t_i and t_j, i.e., GS. The location t_j is selected to replace the reference point t_i according to the probability and when all the locations of reference points are permutated, the localization server sends the updated dataset S to the TBL client.

4) Location Retrieval: After receiving the dataset S, a typical localization algorithm is used to infer the location of TBL client. For illustration, without loss of generality, the popular KNN algorithm is used in our letter.

III. Analysis and Evaluation

A. Security Analysis

1) Location Privacy Preservation of TBL Client: In *AP Fuzzification* phase, the TBL client measures the real-time

fingerprint in a specific spot, including APs and the corresponding RSS. Nevertheless, the client only sends the AP sequence to the server instead of the full WiFi fingerprint. Therefore, any attackers (including localization server) cannot get the client's real-time fingerprint and further query his/her location by the localization system. In addition, *Location Retrieval* phase is executed only in the TBL client side. Hence, attackers cannot get the client's location directly.

2) Data Privacy Preservation of Database Provider: In order to control the privacy budget within a given range, two important properties of DP can be used: the sequential and the parallel composition. The sequential composition will accumulate privacy budget of each step when a series of functions are performed sequentially, and the parallel composition can guarantee privacy decided by the function with the maximal privacy budget.

Similarly as [5], our scheme DP3 employs *DP-based Finger Clustering* with Laplace noise and *Fingerprint Permutation* with exponential probability to protect the leakage of fingerprint dataset. In our work, privacy budget ϵ is evenly divided into two parts. In *DP-based Finger Clustering* phase privacy budget allocated to each iterative round is $\frac{\epsilon}{2T}$. According to sequential composition [9], this operation preserves $\frac{\epsilon}{2}$-differential privacy. *Finger Permutation* phase utilizes exponential mechanism and each selection is performed on the individual reference point. According to parallel composition [9], this operation preserves $\frac{\epsilon}{2}$-differential privacy. Actually, according to sequential and parallel compositions, DP3 can preserve ϵ-differential privacy, regardless of how the privacy budget is allocated into these two phases. However, the allocation rule may affect the localization accuracy. Intuitively, if privacy budget allocated in *DP-based Finger Clustering* phase is reduced, noise added in this phase would increase and Distance Error (DE) would increase; While privacy budget distributed in *Finger Permutation* phase would correspondingly increase, which would reduce the noise added in this phase, that is, the possibility of exchanging two reference points is reduced and DE would decrease. In brief, these two alternative effects jointly determine the allocation of privacy budget, that is, the key principle is to appropriately allocate more privacy budget to the phase that has a greater impact on DE. Actually, the issue of properly allocating privacy budget is another interesting problem.

B. Utility Analysis

1) Utility Analysis of Dataset: The data utility is measured by the average distance error DE between the replaced location in permutated dataset S and the corresponding location in original dataset D.

$$DE = \frac{\sum_{t_i \in D, t_i' \in S} ||(x_i, y_i) - (x_i', y_i')||_2}{GS \cdot |S|} \tag{4}$$

Here, t_i' is the location in dataset S replacing the corresponding original location t_i. Theorem 1 shows that DE of the dataset processing is bounded.

Theorem 1: For all $\delta > 0$, with probability higher than $1-\delta$, the semantic loss of reference points is less than α, that is $Pr(DE < \alpha) > 1 - \delta$.

Here, $\alpha \geq \sum_{t_i \in D, t_i' \in S} \frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{|S| \cdot \rho_i \cdot \delta}$ and $\rho_i = \sum_{t_i' \in S} exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2}{4 \cdot GS})$ is the normalization factor depending on the cluster which t_i belongs to. The processing of dataset is (α, δ)-usefulness.

Proof: DE is proportional to the distance between the original location and the corresponding replaced location. For each original location t_i in cluster C_l, the probability of being substituted by t_i' is $\frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{\rho_i}$.

In probability theory, Markov's inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. In other words, If X is a non-negative random variable and $a > 0$, then the probability that X is greater than a is less than the expectation of X divided by a, which is described with the equation (5).

$$Pr(X \geq a) \leq \frac{E(X)}{a} \tag{5}$$

According to Markov's inequality with the equation (5), we have:

$$Pr(DE < \alpha) > 1 - \frac{E(DE)}{\alpha} \tag{6}$$

Moreover, the expectation $E(DE)$ is evaluated by

$$E(DE) = \sum_{t_i \in D, t_i' \in S} \frac{||(x_i, y_i) - (x_i', y_i')||_2}{GS \cdot |S|} \cdot \frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{\rho_i} \tag{7}$$

Substituting equation (7) into the equation (6), we can get

$$Pr(DE < \alpha) > 1 - \sum_{t_i \in D, t_i' \in S} \times \frac{||(x_i, y_i) - (x_i', y_i')||_2 \cdot exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{GS \cdot |S| \cdot \rho_i \cdot \alpha} \tag{8}$$

Since $GS = max_{t_i, t_i' \in D} ||(x_i, y_i) - (x_i', y_i')||_2$, $||(x_i, y_i) - (x_i', y_i')||_2 \leq GS$. Thus,

$$Pr(DE < \alpha) > 1 - \sum_{t_i \in D, t_i' \in S} \frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{|S| \cdot \rho_i \cdot \alpha} \tag{9}$$

Let $1 - \sum_{t_i \in D, t_i' \in S} \frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{|S| \cdot \rho_i \cdot \alpha} \geq 1 - \delta$, which can guarantee $Pr(DE < \alpha) > 1 - \delta$. In brief, we can get $\alpha \geq \sum_{t_i \in D, t_i' \in S} \frac{exp(\frac{\epsilon \cdot (GS - ||(x_i, y_i) - (x_i', y_i')||_2)}{4 \cdot GS})}{|S| \cdot \rho_i \cdot \delta}$, which implies the minimal distance error of our dataset, namely if a certain level of privacy needs to be achieved, α indicates the least utility loss that needs to be sacrificed.

2) Experimental Results: To evaluate the performance of our scheme, we employ the real dataset of indoor localization system: PosData (available at http://www.cs.rtu.lv/jekabsons). The area of the testbed is approximately $860m^2$, and it contains 328 records considering device's orientation information. Fig.2 shows DE with different clustering iterative rounds (i.e., $T \in \{2, 5, 10\}$). Obviously, DE decreases with the increment of ϵ. Interestingly, under the same privacy budget ϵ, DE varies with different iterative rounds. For example, DE achieves the smallest value with the parameters $T = 2$

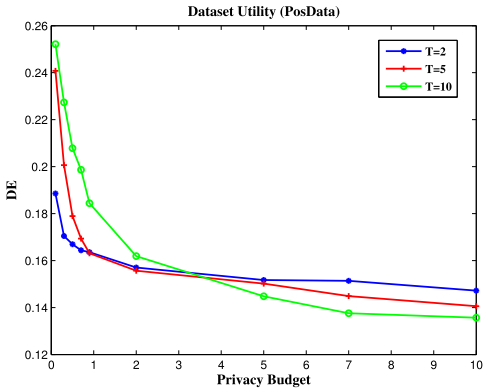

Fig. 2. Utility of PosData at different iterative rounds.

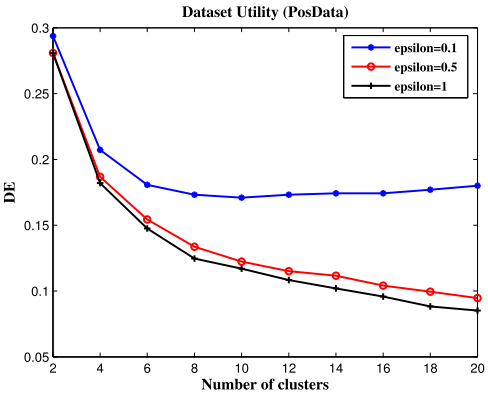

Fig. 3. Utility of PosData with different clusters.

and the privacy budget $\epsilon < 1$ compared with other two kinds of iterative rounds. The reason for this phenomenon lies in that when ϵ is small, the noise added in each iterative round is large, which causes a large DE. It implies that, with a smaller privacy budget, the number of clustering rounds should be less than the case of larger privacy budget to prevent the results from being overwhelmed by too much noise. However, a small number of rounds may be insufficient for clustering convergence, therefore, it is necessary to decide the number of iterative rounds dynamically based on the number of reference points N and the selected ϵ.

Fixing the number of the iterative rounds $T = 2$, Fig.3 shows the utility varying with different number of clusters at different privacy budget (i.e., $\epsilon \in 0.1, 0.5, 1$). Generally, a larger k can achieve higher accuracy, but DE will be stable or rise slightly when k reaches a threshold. For example, with privacy budget $\epsilon = 0.1$, DE gets the minimum 0.1709 when $k = 10$, while increases a little when $k = 12$, and then DE is almost stable. When $\epsilon = 0.5$ and 1, DE keeps dropping as k increases, but when $k > 10$, the trend of decreasing tends to be slow. The reason lies in that DE is affected by both *DP-based Finger Clustering* and *Finger Permutation* phases. When k becomes larger, each randomized position permutation domain will be smaller and *Finger Permutation* can exchange pair of positions with less error (i.e., incur small DE), but the added noise at each iterative round in *DP-based clustering* increases (i.e., lead to large DE). Therefore, the choice of k relies on the trade-off between these two phases.

In brief, those experiments indicate various parameters ϵ, k and T will influence the utility of the fingerprint dataset S

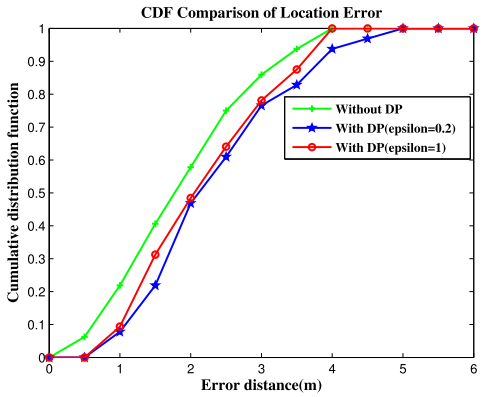

Fig. 4. Cumulative error distribution of the scheme.

and can give us a hand to control the randomized domain and allocate the privacy budget.

3) Experimental Analysis of Localization Error With DP3: We randomly select 64 TBL clients in PostData to illustrate the localization error of DP3-enabled localization method. The experiments on three scenarios are conducted: without DP, with DP ($\epsilon = 0.2$ and $\epsilon = 1$) under $k = 10$ and $T = 2$. Fig.4 illustrates that the maximal localization error without DP3 is almost 4m. After employing DP3 in the localization system, localization error just increases a little, and maximum is 5m. Moreover, the smaller ϵ is set, the larger the localization error is. In brief, we can conclude that our method DP3 can obtain relatively good positioning results while protecting the privacy.

IV. Conclusion

In this letter, we propose DP3 to solve the privacy problem of both TBL clients and localization server in WiFi fingerprint based indoor localization system. The security analysis confirms that DP3 can protect both location privacy of TBL clients and data privacy of the fingerprint database. In addition, the utility analysis and the experimental results illustrate that, compared with without DP3 scheme, DP3 can achieve comparable localization result while protecting the privacy.

References

[1] H. Li, L. Sun, H. Zhu, X. Lu, and X. Cheng, "Achieving privacy preservation in WiFi fingerprint-based localization," in *Proc. IEEE INFOCOM*, Apr./May 2014, pp. 2337–2345.

[2] T. Zhang, S. M. Chow, Z. Zhou, and M. Li, "Privacy-preserving Wi-Fi fingerprinting indoor localization," in *Proc. Int. Workshop Secur.*, 2016, pp. 215–233.

[3] S. Li, H. Li, and L. Sun, "Privacy-preserving crowdsourced site survey in WiFi fingerprint-based localization," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, p. 123, Dec. 2016.

[4] J. W. Kim, D. H. Kim, and B. Jang, "Application of local differential privacy to collection of indoor positioning data," *IEEE Access*, vol. 6, pp. 4276–4286, Jan. 2018.

[5] P. Xiong, T. Zhu, W. Niu, and G. Li, "A differentially private algorithm for location data release," *Knowl. Inf. Syst.*, vol. 47, no. 3, pp. 647–669, 2016.

[6] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private K-means clustering," in *Proc. 6th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, 2016, pp. 26–37.

[7] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[8] W. Wang and Q. Zhang, "Privacy-preserving collaborative spectrum sensing with multiple service providers," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1011–1019, Feb. 2015.

[9] F. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," *Commun. ACM*, vol. 53, no. 9, pp. 89–97, Sep. 2010.