Computers, Materials & Continua CMC, vol.63, no.1, pp.223-241, 2020

A Differentially Private Data Aggregation Method Based on Worker Partition and Location Obfuscation for Mobile Crowdsensing

Shuyu Li[1] and Guozheng Zhang[1, *]

Abstract: With the popularity of sensor-rich mobile devices, mobile crowdsensing (MCS) has emerged as an effective method for data collection and processing. However, MCS platform usually need workers' precise locations for optimal task execution and collect sensing data from workers, which raises severe concerns of privacy leakage. Trying to preserve workers' location and sensing data from the untrusted MCS platform, a differentially private data aggregation method based on worker partition and location obfuscation (DP-DAWL method) is proposed in the paper. DP-DAWL method firstly use an improved K-means algorithm to divide workers into groups and assign different privacy budget to the group according to group size (the number of workers). Then each worker's location is obfuscated and his/her sensing data is perturbed by adding Laplace noise before uploading to the platform. In the stage of data aggregation, DP-DAWL method adopts an improved Kalman filter algorithm to filter out the added noise (including both added noise of sensing data and the system noise in the sensing process). Through using optimal estimation of noisy aggregated sensing data, the platform can finally gain better utility of aggregated data while preserving workers' privacy. Extensive experiments on the synthetic datasets demonstrate the effectiveness of the proposed method.

Keywords: Mobile crowdsensing, data aggregation, differential privacy, K-means, kalman filter.

1 Introductions

The market of hand-held mobile devices (e.g., smartphones and wearable devices) is proliferating rapidly in recent years. These devices possess powerful computation and communication capabilities, and are equipped with various functional built-in sensors. Along with users round-the-clock, mobile devices have become an important information interface between users and environments. These advances have enabled and stimulated the development of mobile sensing technologies, among which mobile crowdsensing catches more and more attention owing to its capability of completing complex social and geographical sensing applications. Mobile crowdsensing (MCS) is defined as a new sensing paradigm that empowers ordinary citizens to contribute data sensed or generated from their

[1] School of Computer Science, Shaanxi Normal University, Xi'an, 710119, China.

[*] Corresponding Author: Guozheng Zhang. Email: guozhengzhang@snnu.edu.cn.

Received: 28 May 2019; Accepted: 19 June 2019.

mobile devices and aggregates and fuses the data in the cloud for crowd intelligence extraction and human centric service delivery [Guo, Zhang and Zhou (2014)]. Different from traditional physical sensors based sensing paradigm, MCS needs large number of participants with smart phones to sense the surrounding environment. The smart phone can collect the useful information about the user such as location or GPS. Users can use the smart phone to collect the useful information of the environment such as images. Through using and analyzing the multi-modal sensing information, it is possible to update the development of public security, smart cities, location based services, etc. Overall, MCS is an emerging computing paradigm that tasks everyday mobile devices to form participatory sensor networks. However, every coin has its two sides. Although with the above advantages, the new sensing paradigm also encounters new challenges such as privacy concerns. In MCS application scenario, participants contribute their sensing data to the MCS platform for further aggregation and analysis, which may carry sensitive information related to users and expose users to the risk of personal privacy leakage. As people pay more attention to personal privacy issues, this becomes a key challenge hindering individuals (workers) from participation, more than the consumption of the limited system resources (e.g., battery and computing power) of their mobile devices. Therefore, the success of MCS hinges upon the design of efficient privacy preserving mechanisms to protect workers' privacy and stimulate workers' participation.

In many MCS applications such as urban road planning and making business decision, platform needs participants' precise locations for optimal task allocation and execution. However, the exposure of their locations raises privacy concerns. Especially for those participants who are not eventually selected for any task, their location privacy is sacrificed in vain. Besides, the MCS platform that collects the sensing data is not trustworthy, it may be curious or even malicious to the participants' location or sensing data for special purpose. The participants may get discouraged and leave the MCS platform, downsizing the candidate worker pool and impairing the performance of the whole platform. Therefore, location and sensing data privacy need to be carefully considered in the phase of data aggregation.

The main contributions of this paper are summarized as follows:

(i) Trying to preserve workers' location and sensing data from the untrusted MCS platform, a differentially private data aggregation method based on worker partition and location obfuscation (DP-DAWL method) is proposed in the paper. Considering that the number of workers in different regions can be various and workers may have diverse privacy requirements, DP-DAWL method adopts an improved K-means algorithm to divide workers into groups. The K-means clustering algorithm is improved by normalizing worker location and adaptively selecting the optimal number of clusters by contour coefficients. Each group will be assigned privacy budget according to the group size (the number of workers), worker intensive group will be assigned more privacy budget and vice versa.

(ii) Since each worker's location is obfuscated and his/her sensing data is perturbed by adding Laplace noise before uploading to the platform, DP-DAWL method adopts an improved Kalman filter algorithm to filter out the added noise (including both added noise of sensing data and the system noise in the sensing process) during the data

aggregation phase, the added noise. Improve process noise of system model based on system process noise affected by population density. The greater the population density, the greater the process noise of the sensing system. Through using optimal estimation of noisy aggregated sensing data, the platform can finally gain better utility of aggregated data while preserving workers' privacy. Privacy analysis proves DP-DAWL method satisfy differential privacy. Besides, Computational Complexity of DP-DAWL method is also provided.

(iii) To validate the effectiveness of DP-DAWL method, synthetic datasets with various size (the number of workers) are generated and compared. The results of experiment show that DP-DAWL method achieves good data utility in condition of preserving worker' privacy.

2 Preliminary and related works

2.1 Differential privacy

Differential privacy is a privacy protection method proposed by Dwork [Dwork (2011)]. Simply to say, a mechanism is differentially private if its outcome is not significantly affected by the removal or addition of a single user. An adversary thus learns approximately the same information about any individual user, irrespective of his/her presence or absence in the original database.

Definition 2-1 (Differential Privacy). Given a random algorithm M and an adjacent data set D and D', P_M indicates the range of values of M, S_M is any subset of P_M, indicating that M outputs an arbitrary result on the adjacent data sets D and D'. If the inequality (1) is satisfied, the random algorithm M is said to satisfy the ε-differential privacy.

$$\frac{\Pr[M(D) \in S_M]}{\Pr[M(D') \in S_M]} \leq \exp(\varepsilon) \tag{1}$$

where $\Pr[M(D) \in S_M]$ is the probability that the data D output is S_M under the algorithm M, and is also the risk of privacy disclosure. The privacy parameter ε (also called the privacy budget) specifies the degree of privacy offered. Intuitively, a lower value of ε implies stronger privacy guarantee and a larger perturbation noise, and a higher value of ε means higher accuracy and lower privacy guarantees.

Definition 2-2 (Sequence Combination Theorem). With algorithm $M_1, M_2, ..., M_n$, the privacy budgets are respectively $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, then for the same data set D, the combined algorithm composed of these algorithms $M(M_1(D), M_2(D), ...M_n(D))$ provides $(\sum_{i=1}^{n} \varepsilon_i)-$differential privacy.

Definition 2-3 (Post-Processing Property) Given $A_1(\bullet)$ that satisfies ε-differential privacy, then for any (possibly randomized) algorithm A_2, the composition of A_1 and A_2, i.e., $A_2(A_1(\bullet))$ satisfies ε-differential privacy.

2.2 Laplace mechanism

Laplace mechanism is the first generalized noise addition mechanism proposed by Dwork [Dwork (2011)]. This mechanism is suitable for handling privacy protection of numerical results. By injecting independent noise obeying the Laplacian distribution into the numerical data or the statistical result, the noise disturbance query result satisfies the constraint condition shown in (1).

Definition 2-3 (Laplace Mechanism). Given a data set D, with a function $f: D \to R^d$ and a sensitivity of Δf, the random algorithm $M(D) = f(D) + Y$ provides ε- differential privacy, where $Y \sim Lap(b)$ is random noise that obeys the Laplace distribution with the scale parameter $b = \Delta f / \varepsilon$.

2.3 Kalman filter algorithm

For linear filtering and prediction problems, Kalman filter algorithm is an algorithm that can optimally estimate the state of the system from a series of data with measurement noise when the measurement variance is known. As it is convenient for programming and the collected data can be processed in real time, Kalman filtering algorithm is widely applied in communication, navigation, sensing data fusion and other fields.

The system's state equation and measurement equation are given as follows:

$$X(k) = AX(k-1) + BU(k) + \omega(k) \tag{2}$$

$$Z(k) = HX(k) + v(k) \tag{3}$$

where $X(k)$, $Z(k)$ and $U(k)$ respectively represent the system state, measured value, and system control amount at time *k*. A and B represent system parameters and H is the measurement system parameter, $\omega(k)$ and $v(k)$ are the process noise and measurement noise of the system respectively.

The Kalman filter algorithm is described by the following five equations.

$$X(k \mid k-1) = AX(k-1 \mid k-1) + Bu(k) \tag{4}$$

The covariance prediction equation in this state at time *k* is defined in (5).

$$P(k \mid k-1) = AP(k-1 \mid k-1)A' + Q \tag{5}$$

System filter estimation equation is given in (6).

$$X(k \mid k) = X(k \mid k-1) + Kg(k)(Z(k) - HX(k \mid k-1)) \tag{6}$$

Kalman filter gain equation is defined in (7).

$$Kg(k) = \frac{P(k|k-1)H'}{HP(k|k-1)H' + R} \tag{7}$$

Filter covariance update equation is given in (8).

$$P(k \mid k) = (I - Kg(k)H)P(k \mid k-1) \tag{8}$$

2.4 Related works

At present, with the emergence of privacy protection algorithms for location data and sensing data in the MCS environment, how to ensure the accuracy of the results to meet the needs of relevant applications while satisfying privacy requirements, and how to

improve the algorithm performance while maintaining acceptable computational overhead and resisting various attacks have become research hotspots. In order to solve the above problems, researchers have proposed many methods for privacy protection. According to the different protection strategies and ideas adopted, the algorithms can be roughly divided into three categories: data anonymity technology, data encryption technology and data perturbation technology.

The privacy protection algorithms based on data anonymity technology achieve the purpose of privacy protection by transplanting and generalizing data. K-anonymity [Sweeney (2002)] and l-diversity [Machanavajjhala, Gehrke, Kifer et al. (2006)] are two well-known data anonymization models. Based on the expansion of Merkle tree, Li et al. proposed a privacy protection mechanism that can control the other leaf nodes of the Merkle tree, authenticate participants anonymously without the trusted third-party [Li and Cao (2014)]. The proposed mechanism effectively realizes incentive schemes under the anonymous mechanism and avoids the single-point failure of the trusted third-party, the revealed data content of participants and the malicious attacks. Wang et al. [Wang, Cheng, Mohapatra et al. (2014)] proposed a privacy protection policy based on trust management and designed the trust classification levels of participants and privacy information sensitivity levels, where the platform cannot deduce the node identity when evaluating node behaviors. Eventually, the anonymity of the participant credibility assessment, privacy protection and data trust management can be achieved. Chen et al. [Chen, Wu, Li et al. (2014)] introduced a privacy protection technology based on k-anonymity that generalizes users. It is impossible for the server to distinguish which of the k users has completed the crowd sensing task, thus protecting the user's privacy.

The privacy protection algorithms based on data encryption technology commonly use homomorphic encryption to aggregate the ciphertext directly and the intermediate node can aggregate the ciphertext directly without decrypting the data, thereby achieving privacy preservation. IPHCDA [Ozdemir and Xiao (2011)] uses homomorphic encryption model which is based on elliptic curve and relevant information authentication codes to deal with external attacks and internal attacks. TTP-free method [Li and Cao (2013)] uses pseudonym and bling signature to protect user privacy, but its encryption operations may bring a burden of cost. Besides, trusted third-party node were used to verify the information uploaded by the participants. Through the establishment of various private and shared keys among the participants, the global public key can remove the association between participants and the server, and the MAC address conversion and network coding are considered to prevent IP address attacks [Shin, Cornelius, Peebles et al. (2011)]. Lv et al. [Lv, Mu and Li (2014)] puts forward a kind of non-interactive public key exchange mechanism utilizing the time-evolving topology model as well as two-channel cryptography: the time-evolving topology model is used to simulate the predicable periodic motion of nodes in Interstellar network, thus the nodes can predict when it conduct the public key exchange with whom; two-channel cryptography algorithm is used for non-interactive public key exchange and guarantees its reliability. Although, this mechanism can realize effective spreading of node public key and guarantee the confidentiality of data transmission in MCS, it requires that the nodes have relatively fixed orbit like stars in Interstellar networks which limits its application scope. Basudan et al. [Basudan, Lin and Sankaranarayanan (2017)] propose a lightweight certificateless scheme of signcryption based on which a privacy-preserving protocol is

designed for enhancing security in data transmission. With respect to privacy preserving, all the sensing data are encrypted. By introducing the certificateless scheme of signcryption, malicious roadside units (can be considered as the edge nodes) are prohibited from modifying the sensing data provided by the participants. Nevertheless, this work does not take into consideration the case when some participants become malicious. Fan et al. [Fan, Li and Cao (2015)] proposed a privacy-aware and trustworthy data aggregation protocol to preserve the privacy of the participants and restrict the behavior of malicious participants. Similar to Basudan et al. [Basudan, Lin and Sankaranarayanan (2017)], the goal of preserving privacy is fulfilled by encrypting the sensing data. To deal with malicious participants, Fan et al. [Fan, Li and Cao (2015)] propose the novel concept of a data value vector from which the participants pick one value to be their sensing result. Then a privacy-aware data validation technique is derived to validate whether each participant has submitted valid data from the data value vector. Via this operation, the influence of malicious participants will be limited. However, the accuracy of the sensing result can be degraded.

The privacy protection algorithms based on data perturbation technology are to add specific random noise in the original data or to perturb the real data through data slice recombination to achieve the purpose of privacy protection. SMAPT [He, Liu, Nguyen et al. (2007)] is a classic representative of data slice recombination technology. The basic idea is that the sensor node randomly splits the perceived data into multiple data slices, and the split data slice itself does not carry any valid information. The data slice can be reorganized to obtain the original sensory data, and the data slices are randomly exchanged between the nodes. The received slice is summed to eliminate the distinguishability of data slices, and then sent the sum results to the base station node for further fusion operation. Tang et al. [Tang and Ren (2015)] exploited the time-domain data transmission delay to achieve the data privacy protection. Firstly, the data is fragmented and the forwarding nodes are randomly selected to transmit the data slices to the sensing platform, thus the data processing server and the source nodes are separated to prevent the server from deducting the trueness and identity of participants. In the data transmission process, the participants are assigned with pseudonyms dynamically. The proposed algorithm can balance the relationship among security, delay rate and delivery rate. However, cutting off the connection between the source node and destination node completely makes the node evaluation impossible, resulting in the system performance degradation. Chen et al. [Chen, Ma and Zhao (2017)] designed a data privacy protection method especially for the untrusted server. The data are divided into multiple slices and forwarded to neighbor participants. The carrying participants send the fragment information directly to the server when the hop count reaches the threshold. Xiao et al. [Xiao, Li and Yuan (2010)] and Cormode et al. [Cormode, Procopiuc, Srivastava et al. (2012)] adopted the method of quadtree and kd-tree to evenly divide the space of user data distribution. Adding noise that satisfies the Laplace distribution to the statistical results, and releasing the statistical result after perturbation, so as to achieve the purpose of protecting the user's location privacy information. From the perspective of the generative antagonistic neural network and combining with the differential privacy protection mechanism, Hitaj et al. [Hitaj, Ateniese and Perez-Cruz (2017)] generated internal attack data and challenged the security of collaborative deep learning. Based on random response and Bloom-Filter, Fanti et al. [Fanti, Pihur and Erlingsson (2015)] and Erlingsson et al. [Erlingsson and Korolova (2014)]

achieved the collection of statistical information of user strings and long-term privacy protection of multiple data collection. Avent et al. [Avent, Korolova, Zeber et al. (2017)] proposed a hybrid model BLENDER with high availability and privacy protection by combining local privacy protection and centralized data mode. Nguyên et al. [Nguyên, Xiao, Yang et al. (2016)] addressed the privacy data collection problem of Samsung's smart mobile terminals, the accurate and efficient Harmony system is built by using the local differential privacy protection mechanism, which realizes the statistical analysis and machine learning functions supporting LDP. Chen et al. [Chen, Yu and Chirkova (2015)] perturbed relevant parameters in the wavelet transformation process to solve the privacy leakage problem in the process of wavelet clustering algorithm, protecting users' location privacy by distorting the number of data points in different equal-width networks through Laplace mechanism and exponential mechanism. To et al. [To, Ghinita and Shahabi (2014)] proposed a personalized localized differential privacy technology to solve the problem of location privacy protection. The concept of security zone is proposed according to the requirements of different privacy protection requirements of each user. Each user specifies a security zone that they can tolerate, then the localized differential privacy technology is used to perturb the user's security zone, so that the attacker can recognize that the concept of a user's security zone is less than a certain threshold. Chen et al. [Chen, Li, Qin et al. (2016)] proposed an architecture for location data acquisition using LDP technology. The user sends the data to a trusted atomic service provider, which is responsible for collecting and updating the location data with privacy parameters ε according to meet the difference of Private Spatial Division (PSD) way. Then the PSD information is stored on the server side and the user responds to requests from the requester.

In summary, each type of technology has its own advantages and disadvantages and performance for different application requirements. There is no way to solve all the problems of privacy protection. The choice of specific methods depends on the application scenarios and the participants' personalized privacy requirements. Usually the tasks published by the MCS platform are sensitive to time and location. Therefore, facing the location sensitive and personalized privacy requirements of participants, how to improve the availability of sensing data on the basis of preserving participants' privacy is an urgent problem to be solved.

3 A data aggregation method based on worker partition and location obfuscation

3.1 System model

Consider an MCS system consisting of a centralized platform, a set of participating workers {1, …, N}, N is the number of workers, as illustrated in Fig. 1. The task requires workers to report to the platform their local sensing data of a specific object or phenomenon (e.g., spectrum sensing and environmental monitoring). To enhance the reliability of the result, the platform will aggregate the sensing data, as the reliability of each worker's sensing data may be different due to different sensor qualities.

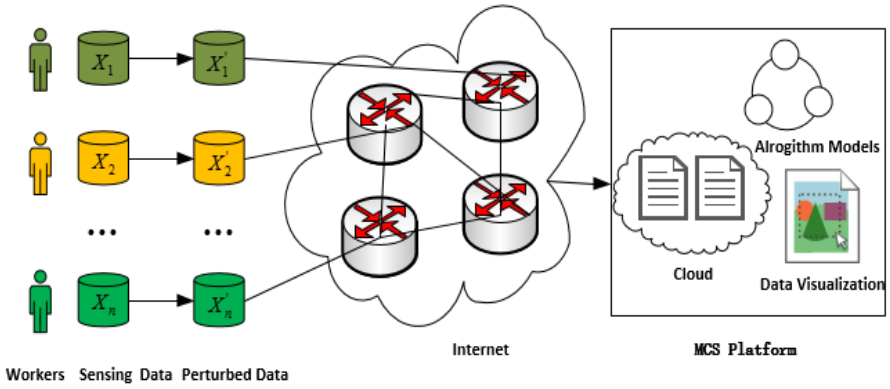

Figure 1: System model of MCS

In this paper, we assume that the MCS platform is untrusted and may be curious to workers' sensing data. We also assume that all workers involved in responding to sensing tasks have been selected under a certain incentive mechanism. In other words, workers are the winners of all participants according to a certain incentive mechanism. And that the sensor has a certain systematic error. Besides, due to various privacy concerns, each worker may have his/her own privacy protection requirements and is not willing to send their precise location and sensing data to the MCS platform directly. Moreover, in this paper, we assume that the number of tasks is smaller than that of participants on the MCS platform, so no selected worker needs to perform more than one task in one snapshot of the task allocation. This assumption is reasonable as today's milestone MCS applications have already attracted millions of users (e.g., Waze [Waze (2016)]), and limiting the number of tasks for each user can benefit both the quality of task performing and user fairness. Last but not the least, further process and analysis of the aggregated data on the MCS platform will be not considered in this paper.

3.2 The framework of the data aggregation method

DP-DAWL method is based on worker partition and location obfuscation. The overall idea of the proposed method is to provide aggregated sensing data to the untrusted MCS platform, while preventing workers' location and original sensing data from acquiring by the platform. Considering that the number of workers in different regions can be various and workers may have his/her privacy requirements, the DP-DAWL method firstly uses improved K-means algorithm to divide workers into several groups according to worker density, then adopts differential privacy as the privacy preserving model and assign different privacy budget to the groups, the assigned privacy budget depend on group density (the number of workers in a group). To protect workers' privacy, each worker's location was obfuscated and sensing data is perturbed by adding Laplace noise. The obfuscated location and perturbed sensing data then are uploaded to the platform, and data aggregation is performed by the platform. Since aggregated data is perturbed and data utility has been decreased, the platform uses improved Kalman filter algorithm to filter out the added noise in best effort without losing workers' privacy. After data filtering, the utility of aggregated data can be improved for further process and analysis.

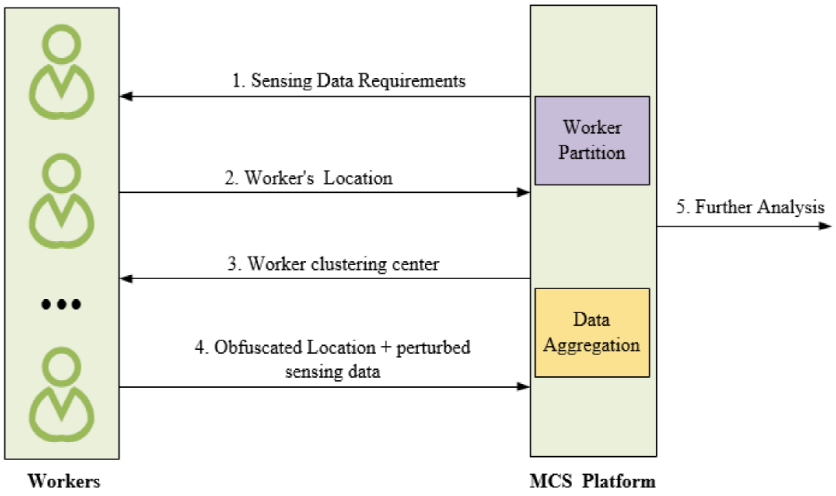

Figure 2: Data flow of DP-DAWL method

The data flow diagram of the proposed data aggregation method in this paper is shown in Fig. 2 and can be described as follows. The MCS platform firstly publish application scenario and sensing data requirements. All workers first perform a simple linear transformation (here relative position shift) of their original location data. And then all workers upload their transformed location data to the platform (here we assume workers have been successfully selected from candidate participants according a certain incentive mechanism). The platform divides the worker into different groups (clusters) according to their location information by using an improved K-means clustering algorithm, detailed clustering process description is presented in the next section. Based on the clustering result and combined with differential privacy model, different privacy budgets are allocated for different groups. Workers in the same group have the same privacy budget, and sensing data of workers will be perturbed before uploading to the platform, the added noise of sensing data depends on the assigned privacy budget. At the same time, according to the privacy budget allocated by different clustering, works in the different group will have different radius and workers in the same group will have same radius. Then take the original location of each worker as the center, and randomly generate four positions within this radius ([Labrador, Wightman and Perez (2011)] recommend N=4 in order to maintain acceptable variability in the final noise, while increasing the average distance from the original point compared with using just one sample). Finally, the farthest location from the center is selected to obfuscate the original location data of the worker. After finishing these two things, workers upload both the obfuscated location data and the perturbed sensing data to the platform. The platform uses am improved Kalman filter algorithm to filter out the added noise in best effort, the added noise includes both added noise of sensing data and the system noise in the sensing process. Through filtering step, the platform can finally gain better utility of aggregated data while preserving workers' privacy.

As mentioned before, we assume that the MCS platform is untrusted. During the whole process of the method, the platform only gets the knowledge of the simple linear

transformed location of workers to perform group partition and cluster center of each group (cluster). Since workers move during the task execution phase and each worker's real location was obfuscated and sensing data is perturbed, worker's privacy can be preserved, strict privacy analysis is presented in Section 3.6.

In summary, the process flow of DP-DAWL method is described as follows:

The DP-DAWL Algorithm	
Input:	worker set W and each worker's initial location l_i, the initial number of cluster $iNum$, privacy budget $\mathcal{E}$, max iteration time T
Output:	filtered aggregated sensing data D', root mean square error
Step 1.	Divide workers into different groups by using the improved clustering algorithm based on Euclidean distance, each group is assigned a privacy budget according to group size. // the improved clustering algorithm is given in Section 3.3
Step 2.	Add Laplace noise into each worker's sensing data and obfuscate each worker's location.
Step 3.	Upload the obfuscated location and the perturbed sensing data to the MCS platform.
Step 4.	Filtering the noisy aggregated data by using the improved Kalman filter algorithm. // the improved Kalman filter algorithm is given in Section 3.5
Step 5.	Calculate the optimal estimation of filtered data and root mean square error

3.3 Worker partition using improved K-means algorithm

Since the MCS platform is untrusted and workers may personalize privacy requirements, we firstly partition the worker into different groups before they begin to execute the task. We do this step for the following reason. It is natural that different regions have various population density. Workers in the population intensive region commonly require more strict privacy preservation while workers in the population sparse region require less. According to the distribution of workers, workers can be classified into different groups, workers in the intensive group will be assigned more privacy budget and workers in the sparse group will be assigned less privacy budget. Workers in the same group will have same privacy budget.

We use an improved K-means algorithm to classify workers into different groups, each group is a cluster of K-means algorithm. K-means classification algorithm is widely used because of its simple implementation and good clustering effect for unsupervised learning. However, there are some disadvantages of K-means such as being sensitive to the selection of initial class cluster center and easily falling into local optimum. To overcome these disadvantages, we improve the K-means algorithm by using contour coefficient to adaptively select of optimal cluster number, thus the clustering result can be optimized.

In stage of data pre-processing, workers' location is normalized to the range of [0, 1]. In the stage of data clustering, the contour coefficient mechanism is introduced to adaptively

select the number of clusters. Firstly, the contour coefficients of all the workers' locations are found and have to be averaged. The average value of contour coefficients has a range of [-1, 1], the larger value of average contour coefficient means better clustering result. The final step is to select the number of clusters with the largest average contour coefficient as the optimal value through multiple iterations.

Combined with the above description, the execution flow of the improved K-means algorithm is described as follows:

The improved K-means Algorithm	
Input.	Sensing data D, Initial cluster number INum, Iterations T
Output.	Optimal cluster number K and its cluster center, the number of workers contained in each cluster
Step 1.	Randomly select a location data from the sensing data D as the initial cluster center G_1
Step 2.	Normalize the location data, calculate the shortest distance Dis(x) between each location data and the current existing cluster center, and select the location data with the largest distance as the new cluster center by contour coefficients.
Step 3.	Repeat Step 2 to determine the optimal number of clusters $K = \{G_1, G_2, \cdots G_k\}$
Step 4.	Calculate the distance of each location data to the K cluster centers and cluster them into the class corresponding to the cluster center with the smallest distance.
Step 5.	For each category G_i, recalculate the centroid of all location data for that class to update the cluster center

3.4 The noise mechanism

The NRand algorithm consists of generating N uniformly distributed random points within a circular domain that is centred on the original coordinates, in order to select the farthest one from the original point; this point will be reported to the service provider as the user's location. Labrador et al. [Labrador, Wightman and Perez (2011)] recommend N=4 in order to maintain acceptable variability in the final noise, while increasing the average distance from the original point compared with using just one sample. In this paper, the NRand algorithm is used to obfuscate the original sensing position of workers based on the results of worker clustering. According to the privacy budget allocated by different clustering, works in the different group will have different radius and workers in the same group will have same radius. Then take the original location of each workers as the center in group, and randomly generate four positions within this radius. Finally, the four positions are located furthest from the center of the circle to obfuscate the original sensing position and upload it to the platform.

As mentioned above, workers located in different areas may need diverse levels of privacy preservation. After we use the improved K-means algorithm to classify workers

into different groups (clusters), each group will be assigned a privacy budget, and according to this assigned privacy budget, the noise that are added to the sensing data of workers in the same group can also be calculated. The main idea of assignment is that workers in the population intensive group will be assigned more privacy budget and workers in the population sparse group will be assigned less privacy budget. Workers in the same group will have same privacy budget. We adopt Differential Privacy model as our privacy preservation method since Differential Privacy gives a strict, quantitative representation and proof of the privacy risk.

We use Laplace mechanism to add different noise perturbations to workers in different groups (clusters). The specific implementation ideas are as follows:

Assume that m is the number of clusters and Ɛ is the total privacy budget, ε_i is the privacy budget of i-th group (cluster) Gi, Ɛ and ε_i satisfy that $\varepsilon = \sum_{C_i=1}^{m} \varepsilon_i$, and $\varepsilon_i = \frac{n_i}{N}$, where n_i is the number of workers in the group Gi, and N is the total number of workers participating the sensing task.

Then the Laplace noise added to workers' sensing data in different groups can be expressed in the following equation:

$$f'(X_{G_{ij}}) = f(X_{G_{ij}}) + Lap(\frac{\Delta f}{\varepsilon_i}) \tag{9}$$

where $f(X_{G_{ij}})$ represents the original sensing data of worker wj in group Gi, and $Lap(\frac{\Delta f}{\varepsilon_i})$ represents the added noise. Δf is global sensitivity and $f'(X_{G_{ij}})$ is the perturbed sensing data.

Only after the noise was added to workers' sensing data, the perturbed sensing data can be uploaded to the MCS platform for further process and analysis.

3.5 Data aggregation using kalman filter

Sensing data of each worker is added Laplace noise according to the assigned privacy budget and each worker's location is obfuscated. Then both perturbed sensing data and obfuscated location are uploaded to the untrusted MCS platform. However, noisy aggregated sensing data may greatly reduce the utility of sensing data. In order to improve the data utility, DP-DAWL method adopts the improved Kalman filter algorithm to filter out noise without compromising workers' privacy.

Since the workers participating in the sensing task are traveling around certain region rather than fixing at one point during phase of task execution, when sensors of a worker collect sensing data, the process noise is also not fixed but changes with time. For example, in remote areas, the sensor signal is weak, the interference signal is small, and the process noise is small. In the bustling downtown area, there are many interference signals. Based on this, improve process noise of system model based on system process noise affected by worker density. The greater the worker density, the greater the process noise of the sensing system. Specifically, larger process noise is set when worker is in a worker intensive area, and smaller process is set when worker is in a worker sparse area. The improve Kalman filter algorithm firstly establishes the sensor system model with noisy sensing data as state quantity. Then the true value of the current moment can be estimated using the state of the

previous moment. That is, if the sensor system is at time k, the state of time k can be predicted and updated by the state of k-1 according to Eq. (4). After updating the system state information, the covariance of the system state can be predictive updated by Eq. (5), where $P(k|k-1)$ is the covariance corresponding to system state $X(k|k-1)$, and Q is the covariance of system process state noise $\omega(k)$. From the previous steps, the prediction result of the state at the current time k can be calculated according to the system state of the previous time k-1. Then, according to the measured value of state k and combined with the predicted value, the state k at the current moment can be estimated according to Eq. (6), where Kg is the Kalman Gain corresponding to the Kalman filter algorithm, which is updated by Eq. (7) itself. And R is the covariance of the system's measured state noise $v(k)$. At this point, the prediction and update of the current state k of the untrusted MCS platform can be realized. Since Kalman filter algorithm is an iteratively updated algorithm, it is necessary to further estimate the state covariance of $X(k|k)$ at the current time k according to Eq. (8), so as to perform the covariance update at time k+1. Thus, when the state covariance $P(k|k)$ is obtained at the current moment and after the system enters the time k+1, $P(k|k)$ is $P(k-1|k-1)$ in Eq. (6). In this way, the algorithm can perform autoregressive and iterative operations to obtain filtered sensing data.

In summary, the execution flow of the improved Kalman filtering algorithm is given as follows:

The improved Kalman Algorithm	
Input:	Perturbed sensing data D', Sensing data record count N
Output:	Filtered and aggregated of sensing data and root mean square error
Step 1.	The sensor system model is established and the parameters are set by using the sensing data after the perturbation as the state quantity by the Eqs. (2) and (3).
Step 2.	Improve process noise of system model based on system process noise affected by population density. The greater the population density, the greater the process noise of the sensing system.
Step 3.	Predicting state $X(k|k-1)$ at time k according to the state at time k-1 by Eq. (4).
Step 4.	Estimating the system prediction error $P(k|k-1)$ at time k according to the system prediction error at time k-1 by Eq. (5).
Step 5.	Calculating the Kalman gain Kg by Eq. (7).
Step 6.	Calculating the optimal estimate $X(k|k)$ of the system by Eq. (6).
Step 7.	Calculating the system prediction error $P(k|k)$ of the current moment of the system by Eq. (8).
Step 8.	After reaching N times, the algorithm ends, otherwise it will return to Step 3 to continue execution.
Step 9.	Output filtered and aggregated of sensing data and root mean square error.

3.6 Privacy analysis

In this section, we will prove DP-DAWL method satisfy ε -differential privacy according to the sequence combination theorem and post-processing property.

The total privacy budget is ε , and according to the execution flow of DP-DAWL method, workers are firstly divided into groups and the number of groups is N, the privacy budget assigned to the group Gi is ε_i , and satisfied $\sum_{i=1}^{N} \varepsilon_i = \varepsilon$. Since different groups are all come from the same worker set, according to the sequence combination theorem of differential privacy, if we want prove that the DP-DAWL method satisfies ε -differential privacy under any adjacent datasets D and D', we firstly need to prove that DP-DAWL method satisfies ε_i -differential privacy for adjacent datasets Gi and G_i'.

The proof process is as follows. For adjacent datasets G_i and $G_i^{'}$, the number of different workers between G_i and $G_i^{'}$ can be expressed by global sensitivity. The global sensitivity is defined as $\Delta f = |Count(G_i) - Count(G_i^{'})|$.

Probability ratio of output o on adjacent sets G_i and $G_i^{'}$ is as follows.

$$\begin{aligned}
\frac{\Pr[b[G_i] = o]}{\Pr[b[G_i^{'}] = o]} &= \frac{\frac{\varepsilon_i}{2\Delta f} e^{-\frac{\varepsilon_i}{\Delta f}|Count(G_i)|}}{\frac{\varepsilon_i}{2\Delta f} e^{-\frac{\varepsilon_i}{\Delta f}|Count(G_i^{'})|}} \\
&= \frac{e^{-\frac{\varepsilon_i}{\Delta f}|Count(G_i)|}}{e^{-\frac{\varepsilon_i}{\Delta f}|Count(G_i^{'})|}} \\
&= e^{\frac{\varepsilon_i}{\Delta f}(|Count(G_i^{'}) - Count(G_i)|)} \\
&\le e^{\varepsilon_i}
\end{aligned} \tag{10}$$

From the above proof, it can be seen that DP-DAWL method satisfies ε_i -differential privacy for adjacent datasets G_i and $G_i^{'}$. According to the sequence combination theorem, DP-DAWL method satisfies ε -differential privacy before the stage of Kalman filter.

As we know, the input of the improved Kalman filter algorithm is output of the previous algorithm which satisfies ε -differential privacy as above analysis. According to the post-processing property of differential privacy, we can prove that DP-DAWL method satisfies ε -differential privacy.

3.7 Computational complexity analysis

The main computation overhead of DP-DAWL method are clustering process and data aggregation process using Kalman filter. Firstly, the normalized preprocessing is added to

the data of the traditional K-means algorithm to eliminate the influence of different orders of magnitude data. Secondly, according to the contour coefficient, the optimal number of clusters is selected adaptively in the algorithm, which makes the selection method and number constraint of cluster center more reasonable. The complexity of the algorithm is $O(k*k*m*n)$. Where k is the number of clusters, n is the number of workers in the dataset, and m is the spatial dimension of the dataset. The data aggregation process filters the added noise of the sensing data and the error noise of the sensing system according to the process noise of the improved Kalman filter algorithm to improve the utility of sensing data. The complexity of data aggregation process is $O(m^{2.376}+n^2)$. The total complexity of the algorithm in this paper is $O(k*k*m*n)+O(m^{2.376}+n^2)$. Since n is very large in the sensing data, k and m are much smaller than n. So the final complexity of the algorithm is $O(n^2)$.

4 Experiment analysis

The experimental hardware environment is Intel (R) Xeon (R) CPU E5-2650 v4 @ 2.20 GHz 2.20 GHz (4 processor), 32 G memory, 930 G hard disk storage space. The software environment is Windows 10 operating system, Eclipse2017, related algorithms are realized in Java language. The experimental data is randomly generated by a pseudo-random number generator to generate six sets of synthetic datasets with different number of workers. Each dataset contains a considerable number of workers, each worker's initial location, and sensing data for a task, where the worker ID range is [0,N-1] (N is the number of workers) and the worker's initial location is a two-dimensional array with data range of [-500,500], and the sensing data range of a task is [0,1000]. Results of the improved K-means algorithm are analyzed by clustering compactness (CP) and clustering separation (SP). Perturbed sensing data is aggregated and filtered by the improved Kalman filter algorithm, and the utility of the filtered data is analyzed by using root mean square error and mean absolute error.

4.1 Clustering result analysis

This experiment uses datasets with various workers to verify the clustering results of the improved K-means algorithm. Because the datasets used in the experiment is randomly generated by the pseudo-random number generator, there is no cluster evaluation tag. Therefore, we use clustering compactness (CP) and clustering separation (SP) as internal evaluation indexes of clustering to evaluate the utility of clustering results. Lower value of CP means the closer distance of intra-cluster, higher value of SP means the further distance of inter-cluster. Different datasets with size of 2000, 4000, 6000, 8000, and 10000 respectively, are used to verify the improved K-means algorithm and calculate CP & SP. The experimental result is given in Tab. 1.

Table 1: Clustering result analysis

Evaluation Index	Dataset Size (workers)				
	2000	4000	6000	8000	10000
CP	0.043	0.052	0.082	0.082	0.023
SP	0.757	0.734	0.779	0.707	0.706

It can be seen from the above Tab. 1 that the clustering result is more accurate by optimizing the data normalization preprocessing and the contour system adaptively selecting the optimal number of clusters. Among the clustering results of different datasets, the clustering distance in the same group (cluster) are very close, and the optimal value is 0.023. The clustering distance between the different groups (cluster) is very far, and the optimal value is 0.779.

4.2 Data aggregation result analysis

The experiment uses two sets of datasets with 5000 and 10000 workers respectively, to verify the utility of aggregated sensing data after filtering in the untrusted MCS platform. In the process of aggregation filtering, the sensor system model is established by using perturbation sensor data as a state variable. Based on the clustering results, the utility of data aggregation is compared by assigning different privacy budgets. Considering that the process noise of the system increases with the increase of workers, with a total privacy budget of 0.1, 0.3, 0.5, 0.7, and 0.9, the process noise of the system is set to 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. Finally, using the improved Kalman filter algorithm, the utility of the filtered data is verified by the root mean square error and the mean absolute error under different privacy budgets and process noises. Root mean square error is shown in Fig. 3 and mean absolute error is shown in Fig. 4.

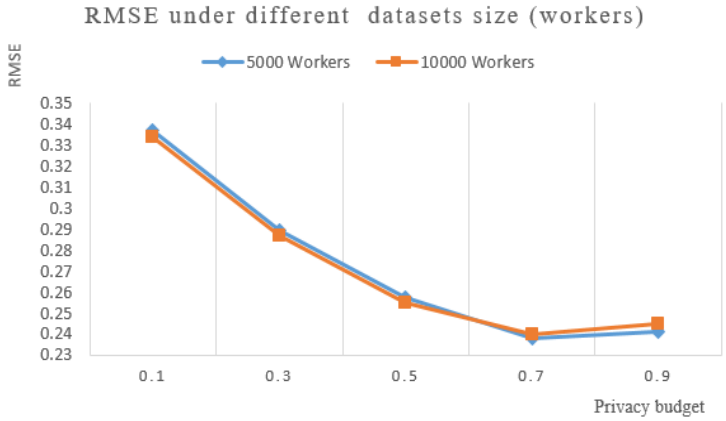

Figure 3: Root mean square error of data aggregation under different datasets

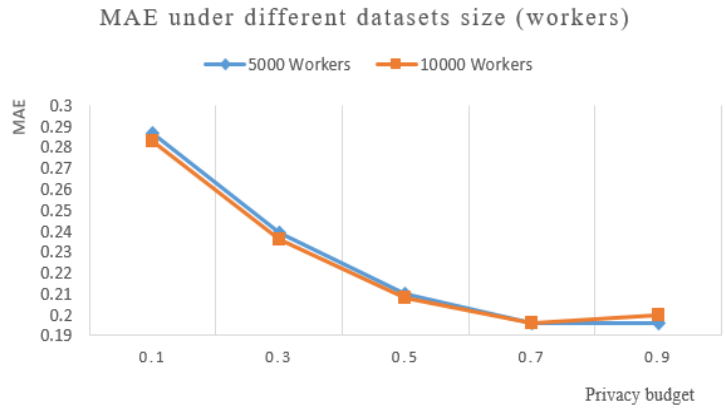

Figure 4: Mean absolute error of data aggregation under different datasets

Observing the above experimental results, it can be found that with the increase of the privacy budget, root mean square error and average absolute error of the datasets are getting smaller and smaller. That is, the data aggregation accuracy is getting higher and higher. With a total privacy budget of 0.1, root mean square error reached 0.334 and average absolute error reached 0.287. With a total privacy budget of 0.9, the root means square error reached 0.241 and the average absolute error reached 0.2. These results demonstrate that DP-DAWL method achieves good data utility without compromising workers' privacy.

5 Conclusions

Trying to preserve workers' location and sensing data from the untrusted MCS platform, a data aggregation method named DP-DAWL is proposed in the paper. DP-DAWL method adopts an improved K-means algorithm to divide workers into groups and assigns different privacy budget to groups according to the group size. Because each worker's location is obfuscated and his/her sensing data is perturbed by adding Laplace noise before uploading to the platform, DP-DAWL method adopts an improved Kalman filter algorithm to filter out the added noise during the data aggregation phase. Through using optimal estimation of noisy aggregated sensing data, the platform can finally gain better utility while preserving workers' privacy. Synthetic datasets with various size are generated and compared to validate the effectiveness of DP-DAWL method. The results of experiment show that DP-DAWL method achieves good data utility without compromising workers' privacy. In future work, we will mainly focus on how to cluster workers in a more accurate way and provide personalized privacy preservation for workers. Besides, the efficiency of DP-DAWL method can be improved in future research.

Acknowledgement: This research was funded by Key Research and Development Program of Shaanxi Province (No. 2017GY-064) and the National Key R&D Program of China (No. 2017YFB1402102).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Avent, B.; Korolova, A.; Zeber, D.; Hovden, T.; Livshits, B. (2017): {BLENDER}: enabling local search with a hybrid differential privacy model. *26th {USENIX} Security Symposium ({USENIX} Security 17)*, pp. 747-764.

Basudan, S.; Lin, X.; Sankaranarayanan, K. (2017): A privacy-preserving vehicular crowdsensing-based road surface condition monitoring system using fog computing. *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 772-782.

Chen, J.; Ma, H.; Zhao, D. (2017): Private data aggregation with integrity assurance and fault tolerance for mobile crowd-sensing. *Wireless Networks*, vol. 23, no. 1, pp. 131-144.

Chen, L.; Yu, T.; Chirkova, R. (2015): Wavecluster with differential privacy. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1011-1020.

Chen, R.; Li, H.; Qin, A. K.; Kasiviswanathan, S. P.; Jin, H. (2016): Private spatial data aggregation in the local setting. *IEEE 32nd International Conference on Data Engineering*, pp. 289-300.

Chen, X.; Wu, X.; Li, X. Y.; Ji, X.; He, Y. et al. (2015): Privacy-aware high-quality map generation with participatory sensing. *IEEE Transactions on Mobile Computing*, vol. 15, no. 3, pp. 719-732.

Cormode, G.; Procopiuc, C.; Srivastava, D.; Shen, E.; Yu, T. (2012): Differentially private spatial decompositions. *IEEE 28th International Conference on Data Engineering*, pp. 20-31.

Dwork, C. (2011): Differential privacy. *Encyclopedia of Cryptography and Security*, pp. 338-340.

Erlingsson, Ú.; Pihur, V.; Korolova, A. (2014): Rappor: randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054-1067.

Fan, J.; Li, Q.; Cao, G. (2015): Privacy-aware and trustworthy data aggregation in mobile sensing. *IEEE Conference on Communications and Network Security*, pp. 31-39.

Fanti, G.; Pihur, V.; Erlingsson, Ú. (2015): Building a RAPPOR with the unknown: privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41-61.

Guo, B.; Yu, Z.; Zhou, X.; Zhang, D. (2014): From participatory sensing to mobile crowd sensing. *IEEE International Conference on Pervasive Computing and Communication Workshops*, pp. 593-598.

He, W.; Liu, X.; Nguyen, H.; Nahrstedt, K.; Abdelzaher, T. (2007): PDA: privacy-preserving data aggregation in wireless sensor networks. *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pp. 2045-2053.

Hitaj, B.; Ateniese, G.; Pérez-Cruz, F. (2017): Deep models under the GAN: information leakage from collaborative deep learning. *Proceedings of the 2017 ACM*

SIGSAC Conference on Computer and Communications Security, pp. 603-618.

Labrador, M. A.; Perez, A. J.; Wightman, P. M. (2010): *Location-Based Information Systems: Developing Real-Time Tracking Applications*. CRC Press.

Li, Q.; Cao, G. (2013): Providing privacy-aware incentives for mobile sensing. *IEEE International Conference on Pervasive Computing and Communications*, pp. 76-84.

Li, Q.; Cao, G. (2014): Providing efficient privacy-aware incentives for mobile sensing. *IEEE 34th International Conference on Distributed Computing Systems*, pp. 208-217.

Lv, X.; Mu, Y.; Li, H. (2014): Non-interactive key establishment for bundle security protocol of space DTNs. *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 1, pp. 5-13.

Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. (2006): L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering*, pp. 24-24.

Nguyên, T. T.; Xiao, X.; Yang, Y.; Hui, S. C.; Shin, H. et al. (2016): Collecting and analyzing data from smart device users with local differential privacy. arXiv:1606.05053.

Ozdemir, S.; Xiao, Y. (2011): Integrity protecting hierarchical concealed data aggregation for wireless sensor networks. *Computer Networks*, vol. 55, no. 8, pp. 1735-1746.

Shin, M.; Cornelius, C.; Peebles, D.; Kapadia, A.; Kotz, D. et al. (2011): Anonysense: a system for anonymous opportunistic sensing. *Pervasive and Mobile Computing*, vol. 7, no. 1, pp. 16-30.

Sweeney, L. (2002): Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571-588.

Tang, D.; Ren, J. (2015): A novel delay-aware and privacy-preserving data-forwarding scheme for urban sensing network. *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2578-2588.

To, H.; Ghinita, G.; Shahabi, C. (2014): A framework for protecting worker location privacy in spatial crowdsourcing. *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 919-930.

Wang, X. O.; Cheng, W.; Mohapatra, P.; Abdelzaher, T. (2014): Enabling reputation and trust in privacy-preserving mobile sensing. *IEEE Transactions on Mobile Computing*, vol. 13, no. 12, pp. 2777-2790.

Waze (2016): Waze-GPS, Maps, Traffic Alerts & Live Navigation. https://play.google.com/store/apps/details?id=com.waze&hl=en.

Xiao, Y.; Xiong, L.; Yuan, C. (2010): Differentially private data release through multidimensional partitioning. *Workshop on Secure Data Management*, pp. 150-168.