Privacy-Utility Equilibrium Protocol for Federated Aggregating Multiparty Genome Data

Hai Liu[1,2,3], Changgen Peng[1,2,3], Youliang Tian[2,3], Feng Tian[4], and Zhenqiang Wu[4]
[1]Guizhou Big Data Academy, Guizhou University, Guiyang 550025, China
[2]College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
[3]State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
[4]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

Cloud server aggregates a large amount of genome data from multi genome donors to facilitate scientific research. However, the untrusted cloud server is prone to violate privacy of aggregating genome data. Thus, each genome donor can randomly perturb her genome data using differential privacy mechanism before aggregating. But this is easy to lead to utility disaster of aggregating genome data due to the different privacy preferences of each genome donor, and privacy leakage of aggregating genome data because of the kinship between genome donors. The key challenge here is to achieve an equilibrium between privacy preserving and data utility of aggregating multiparty genome data. To this end, we proposed federated aggregation protocol of multiparty genome data (MGD-FAP) with privacy-utility equilibrium for guaranteeing desired privacy protection and desired data utility. First, we regarded the privacy budget and the accuracy as the desired privacy-utility metrics of genome data respectively. Second, we constructed the federated aggregation model of multiparty genome data by combining random perturbation method of genome data guaranteeing desired data utility with federated comparing update method of local privacy budget achieving desired privacy preserving. Third, we presented the MGD-FAP maintaining privacy-utility equilibrium under the federated aggregation model of multiparty genome data. Finally, our theoretical and experimental analysis showed that MGD-FAP can maintain privacy-utility equilibrium. The MGD-FAP is practical and feasible to ensure the privacy-utility equilibrium of cloud server aggregating multiparty genome data.

***Index Terms*—Multiparty genome data aggregation, cloud server, federated comparing, strategic game, privacy-utility equilibrium.**

I. Introduction

SINCE the cost of sequencing has decreased significantly, large-scale and high-dimensional genome data have been produced[1]. Genome data have been widely used in scientific research [1]. Because a single institution usually only possesses a limited number of genome data, genome data need to be aggregated to cloud server for supporting meaningful scientific research [2], [3]. Genome data can uniquely identify individuals and keep stability without changing over time, and it is associated with individual information such as heredity, disease, phenotype, and kinship. Thus, the untrusted cloud server is easy to bring about the risk of privacy leakage of aggregating genome data for scientific research.

Cryptographic techniques can achieve privacy preserving of multiparty genome data aggregation (MGDA) [2], [3], but cryptographic techniques have large computational overhead. Differential privacy is a strict privacy preserving framework without considering all the background knowledge except a single record [4]. Therefore, differential privacy has been used to to protect sensitive information of the genome data [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] and dependent genome data [17]. However, since differential privacy has privacy-utility monotonicity [18], these works can only achieve the privacy-utility tradeoff of genome data and dependent genome data. Privacy-utility tradeoff implies that one thing increases and the other inevitably decreases between privacy protection and data utility. Moreover, these works do not consider the differential privacy protection of aggregating multiparty genome data.

But if differential privacy is directly used for aggregating multiparty genome data, it will lead to utility disaster due to the different privacy preferences of each genome donor, and privacy leakage due to the kinship between genome donors. The following example specifically explains these serious results.

Example 1. In aggregating multiparty genome data with differential privacy, some genome donors may use a smaller privacy budget because of their concern for privacy, and other genome donors may use a larger privacy budget because they do not pay attention to privacy. This can lead to utility disaster of aggregating genome data, such that cloud server gets very less utility of aggregating genome data. Because of the kinship between genome donors, this also will bring the risk of privacy leakage.

Thus, the key challenge is to balance the privacy protection and data utility of aggregating genome data according to Example 1. To solve this problem, we proposed MGD-FAP ensuring privacy-utility equilibrium to support meaningful scientific research. Specifically, we gave the definitions of desired privacy preserving metric and desired data utility metric. We provided a federated comparing method that can

Manuscript received November 10, 2021; revised December 30, 2021.
Corresponding author: Changgen Peng (email: peng_stud@163.com).
This research was supported by the National Natural Science Foundation of China (62002081, U1836205, and 61602290), the Funded by China Postdoctoral Science Foundation (2019M663907XB), the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ004), and the Major Scientific and Technological Special Project of Guizhou Province (20183001).

[1]https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

update each genome donor's local privacy budget to obtain the desired privacy budget. We presented the random perturbation method of genome data guaranteeing desired data utility. We constructed a federated aggregation model of multiparty genome data by combining the federated comparing update of local privacy budget achieving desired privacy preserving and the random perturbation method guaranteeing desired data utility. We presented MGD-FAP based on the federated aggregation model of multiparty genome data. Our theoretical and experimental results demonstrated that the MGD-FAP can keep privacy-utility equilibrium. Our protocol can be used for federated aggregation of multiparty genome data to solve the problems of utility disaster and privacy leakage. Our main contributions are as follows.

(1) We stated the desired privacy-utility metrics of genome data. We used the federated comparing update of local privacy budget achieving desired privacy preserving, and presented the random perturbation method of genome data guaranteeing desired data utility.

(2) We constructed the federated aggregation model of multiparty genome data by combining federated comparing update of local privacy budget achieving desired privacy preserving and random perturbation method of genome data guaranteeing desired data utility, and gave the MGD-FAP achieving privacy-utility equilibrium.

(3) We theoretically proved that the MGD-FAP satisfies the desired differential privacy preserving and achieves the desired data utility. We also theoretically proved MGD-FAP ensuring the privacy-utility equilibrium based on strategic game. Moreover, our experimental results confirm the theoretical analysis results.

This paper is organized as follows. Section II introduces the related work. Section III introduces the preliminaries. Section IV presents the aggregation model of multiparty genome data and design goal. Section V presents the federated aggregation model and protocol of multiparty genome data, and takes a theoretical analysis for our protocol. Section VI makes the numerical evaluation on the privacy-utility equilibrium of MGD-FAP. Section VII concludes this paper.

II. Related Work

This section introduces the related work from the following four aspects, and analyses the gap between existing work and contribution of this paper.

A. Correlated Data with Differential Privacy

Kifer and Machanavajjhala [19] had shown that correlated data with differential privacy is easy to lead to weak privacy preserving. Therefore, the current work has carried out extensive research on differential privacy preserving of correlated data. Chen et al. [20] proposed edge differential privacy by introducing correlation coefficient. Considering the background knowledge of correlated data, Kifer and Machanavajjhala [21] proposed the privacy preserving model of Pufferfish. Yang et al. [22] proposed Bayesian differential privacy of correlated data under Pufferfish model. Song et al. [23] proposed Pufferfish privacy mechanism, Wasserstein mechanism, for correlated data. Wang and Wang [24] proposed the correlated tuple differential privacy in correlated tuple data release. However, these work only achieves the privacy-utility tradeoff of correlated data, but do not reach the privacy-utility equilibrium.

B. Genome Data Research with Differential Privacy

Homer et al. [25] demonstrated that statistical test of a complex genome mixture can determine whether a specific individual is in the control group or the case group, and obtain individual's genome data. Therefore, it is urgent to protect the privacy of participants and the confidentiality of genome data research. At present, the existing work has carried out extensive research on genomic privacy preserving based on differential privacy. Differential privacy can prevent an attacker with prior knowledge from leaking the individual privacy in genome-wide association research [5], [6], [7], [8]. Because genome data is highly sensitive, differential privacy has used to privacy preserving of genome data sharing [9], [10], [11], [12]. In genome healthcare, genome donors worry about privacy leakage, while healthcare providers worry about disclosure of trade secrets. Thus, differential privacy also used to achieve privacy preserving of genome healthcare [13], [14], [15]. However, genome data research with differential privacy only achieves the privacy-utility tradeoff. Therefore, Liu et al. [16] proposed adaptive differential privacy of categorical data to achieve desired privacy preserving and desired data utility of genome data sharing. However, adaptive differential privacy of categorical data can not be directly used for aggregating multiparty genome data to achieve privacy-utility equilibrium.

C. Dependent Genome Data with Differential Privacy

Because genome donors have kinship, individual genome data may also disclose sensitive information about her family members' genome data [26], [27]. Thus, Almadhoun et al. [17] introduced the differential privacy for genome data with the probabilistic dependence relationship between dependent tuples and achieves rigorous privacy preserving. However, dependent genome data using differential privacy can only achieve the privacy-utility tradeoff.

D. Federated Genome Data with Differential Privacy

In order to get high quality statistical patterns and relationships between genetic variants and diseases, cloud server usually aggregates a large amount of genome data from multi institutions. But the major problem of aggregating genome data cross-institution is privacy concerns. Thus, existing work uses cryptography methods achieving privacy preserving of federated genome data [2], [3], [28], [29], such as secure multiparty computation, homomorphic encryption, and Intel SGX. Blockchain-based time-stamping scheme [30] can also be used to guarantee privacy of federated genome data in cloud storage. However, cryptography methods require large computational overhead for huge scale genome data aggregation. Moreover, the existing work does not consider federated aggregation of multiparty genome data using differential privacy.

To sum up, the existing work does not consider privacy-utility equilibrium of aggregating genome data with differential privacy. In this study, we achieve this goal by constructing a federated aggregation model and proposing interactive protocol for aggregating multiparty genome data.

III. Preliminaries

This section introduces the preliminaries to genome [8], differential privacy [4], local differential privacy [31], and strategic game [32].

A. Genome

Individual's genome data are a sequence of the diploid genotype, and each diploid genotype takes values in $\{0, 1, 2\}$. For each gene locus, two alleles are observed at the same position, which a major allele is observed at a higher frequency and a minor allele is observed at a lower frequency. B presents the major allele and b presents the minor allele, where $\mathrm{B}, \mathrm{b} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. BB is encoded as 0, Bb as 1, and bb as 2.

B. Differential Privacy

A dataset x is a collection of records, in which x_i represents the i-th element or subset in the dataset x. A natural measure of distance between two databases x and y will be Hamming distance $d(x, y)$.

Definition 1 (Differential Privacy). A random mechanism $\mathcal{M}$ is (ε, δ)-differential privacy if for all $S \subseteq Range(\mathcal{M})$ and for all x, y such that $d(x, y) \leq 1$, then

$$P(\mathcal{M}(x) \in S) \leq e^{\varepsilon} P(\mathcal{M}(y) \in S) + \delta \tag{1}$$

When $\delta = 0$, $\mathcal{M}$ is ε-differential privacy.

Differential privacy guarantees that the probability distribution of response to any query is the same independent of any individual opting presence or absence in the database. To all databases x and y with $d(x, y) \leq 1$, when $\mathcal{M}$ is (ε, δ)-differential privacy, the mechanism $\mathcal{M}$ is ε-differential privacy with probability at least $1 - \delta$.

For any query function $f : x \to \mathbb{R}^k$, Laplace mechanism (LM) $Lap(\frac{\Delta_1 f}{\varepsilon})$ is ε-differential privacy, where $\mathbb{R}$ denotes the set of all real numbers, and $\Delta_1 f = \max_{d(x,y)=1} ||f(x) - f(y)||_1$ is ℓ_1-sensitivity. Unless otherwise stated, the differential privacy mechanism mentioned in the follow-up of this paper refers to Laplace mechanism.

Definition 2 (Local Differential Privacy). A random mechanism $\mathcal{M}$ is (ε, δ)-local differential privacy if for any input b_1 and b_2 and for any possible output b, then

$$P(\mathcal{M}(b_1) = b) \leq e^{\varepsilon} P(\mathcal{M}(b_2) = b) + \delta \tag{2}$$

When $\delta = 0$, $\mathcal{M}$ is ε-local differential privacy.

The randomized response (RR) is a major random perturbation mechanism of local differential privacy [33]. For any input $b \in \{0, 1\}$ of the randomized response, the probability of correct response is $\frac{e^{\varepsilon}}{1+e^{\varepsilon}}$ and the probability of wrong response is $\frac{1}{1+e^{\varepsilon}}$.

Moreover, differential privacy has properties of post-processing [4] and parallel composition [34].

Theorem 1 (Post-Processing). A random mechanism $\mathcal{M} : x \to \mathbb{R}$ on dataset x is (ϵ, δ)-differential privacy. Let $f : \mathbb{R} \to \mathbb{R}'$ be a random mapping, then $f \circ \mathcal{M} : x \to \mathbb{R}'$ is (ϵ, δ)-differential privacy.

Theorem 2 (Parallel Composition). Each random mechanism $\mathcal{M}_i$ is ϵ_i-differential privacy. x_i is arbitrary disjoint subsets of the input dataset x. The parallel composition of $\mathcal{M}_i$ is $max\{\epsilon_i\}$-differential privacy.

C. Strategic Game

Definition 3 (Strategic Game). Strategic game is a triplet $G = (N, (S_i)_{i \in N}, (u_i)_{i \in N})$

(1) Players set $N = \{1, \cdots, n\}$.

(2) Strategies set S_i of player $i \in N$.

(3) Utility function $u_i : S_1 \times \cdots \times S_n \to \mathbb{R}$ of player $i \in N$.

Definition 4 (Nash Equilibrium). A vector $s = (s_1, \cdots, s_n)$ of strategies is Nash equilibrium in strategic game if $u_i(s_i, s_{-i}) \geq u_i(s_i^{'}, s_{-i})$ for any strategy $s_i^{'}$ of each player $i \in N$, where s_{-i} means the strategies vector of other players except for strategy $s_i \in S_i$ of player $i \in N$.

IV. Aggregation Model and Design Goal

In this section, we present the aggregation model of multiparty genome data, give privacy threat model, define desired privacy-utility metrics, and identify design goal.

A. Aggregation Model of Multiparty Genome Data

The aggregation model of multiparty genome data is shown in Fig. 1. This model consists of multi genome donors, cloud server, and multi genome users. Genome donors store their genome data on cloud server to save the cost of long-term storage and management. Genome donor can be hospital or human individual. Cloud server aggregates genome data of multi genome donors. Cloud server analyzes aggregating multiparty genome data and shares genome data or analysis results. The cloud server can be a data storage and processing center of a hospital, a third party, or a human individual. Genome users request the cloud server to query genome data or analysis results. Cloud server answers to the corresponding genome data or analysis results according to the query of genome users. Genome users can be biomedical researchers in medical center or genome research center, or healthcare provider.

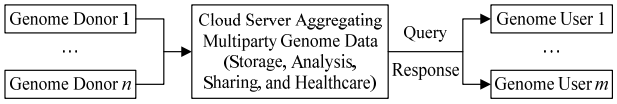

Fig. 1. Aggregation model of multiparty genome data.

B. Privacy Threat Model

In reality, cloud server and genome users are not trusted. Therefore, honest genome donors are unwilling to send genome data to the cloud server because of privacy concerns. Considering self-interest of genome donors, genome data donors are willing to send genome data to the cloud server under the maximal privacy preserving. This study assumes that cloud server is semi-honest. Thus, the cloud server honestly performs the storage, analysis, sharing, and healthcare of aggregating multiparty genome data. However, cloud server is curious about the sensitive information on genome data and which genome donor presence in the aggregating multiparty genome data. Similarly, we assume that genome users are also semi-honest. Genome users honestly execute the query request to the cloud server. Because cloud server honestly shares genome data or analysis results, each genome user can obtain correct response to the corresponding query. But, genome users also want to obtain sensitive information about genome data and which genome donor presence in aggregating multiparty genome data. Moreover, considering genome users is also self-interest. Genome users want to obtain the genome data and analysis results under the maximal data utility. Therefore, the cloud server wants to obtain the desired data utility of aggregating multiparty genome data according to the maximal data utility of genome users.

C. Desired Privacy-Utility Metrics

Because privacy budget quantifies privacy loss of differential privacy, we use privacy budget as the desired privacy metric. Each genome donor selects the desired privacy budget according to the normalized expectation estimation error. The normalized expected estimation error is $\frac{\sum_{j=1}^{m} P(x_{ij}^{'})|x_{ij}^{'}-x_{ij}|}{|x_i|}$, where $x_{ij} \in x_i$, $x_{ij} \in \{0,1,2\}$, and $x_i = (x_{i1}, \cdots, x_{im})$ is the m-dimensional diploid genotype vector of genome donor i. $x_{ij}^{'} \in x_i^{'}$, $x_{ij}^{'} \in \{0,1,2\}$, and $x_i^{'} = (x_{i1}^{'}, \cdots, x_{im}^{'})$ denotes the m-dimensional diploid genotype vector of genome donor i after random perturbation. $P(x_{ij}^{'})$ is probability of randomly perturbing x_{ij} to $x_{ij}^{'}$. $|x_i|$ is the number of all diploid genotypes in x_i.

Because the genome data are categorical data of 0, 1, and 2, it should also be categorical data of 0, 1, and 2 after using differential privacy mechanism, otherwise the genome data are completely unavailable because of random perturbation. Therefore, we use accuracy [6] as the desired data utility U to measure the proportion of correct diploid genotypes after random perturbation. The accuracy is $U = \frac{|x_i \bigcap x_i^{'}|}{|x_i|}$, and $|x_i|U = |x_i \bigcap x_i^{'}|$ represents the number of the same diploid genotypes between vector x_i and $x_i^{'}$. We use $x = (x_1, \cdots, x_n)^{\top}$ representing m-dimensional diploid genotype matrix of n genome donors, $|x|$ is the number of all diploid genotypes in x, and $|x_1| + \cdots + |x_n| = |x|$. Because each genome donor i provides the desired data utility being $U = \frac{|x_i \bigcap x_i^{'}|}{|x_i|}$, this can ensure that the desired data utility of the aggregating genome data of n genome donors is also

$$U = \frac{|x_1|U + \cdots + |x_n|U}{|x|} = \frac{(|x_1| + \cdots + |x_n|)U}{|x|} \quad (3)$$

Therefore, we use accuracy as the desired data utility metric in this study, which not only guarantees the desired data utility of each genome donor, but also guarantees that cloud server gets the desired data utility of aggregating multiparty genome data.

D. Design Goal

Because of each genome donor selecting different privacy budget and existing the kinship between genome donors, aggregating multiparty genome data satisfying differential privacy will lead to the extreme results of utility disaster and privacy leakage.

Combining the privacy threat model and the desired privacy-utility metrics, it is preferable to study the federated aggregation model and protocol of multiparty genome data ensuring privacy-utility equilibrium.

V. Our Model and Protocol

This section introduces our model and protocol, and theoretically proves our protocol keeping privacy-utility equilibrium.

A. Federated Aggregation Model of Multiparty Genome Data

Combining the aggregation model of multiparty genome data in Fig. 1, we construct a federated aggregation model of multiparty genome data in Fig. 2. As shown in Fig. 2, the cloud server sends the desired data utility to genome donors according to the maximal data utility of genome users. In this way, the federated aggregation model can achieve the desired data utility of multiparty genome data. Each genome donor encodes genome data to diploid genotype according to the allele frequencies. Each genome donor randomly perturbs diploid genotype data using differential privacy mechanism under desired data utility. However, genome donors have different privacy preferences when randomly perturb their diploid genotype data using differential privacy mechanism. Therefore, cloud server obtains the global privacy budget by federated comparing the aggregating local privacy budget of multi genome donors. In order to achieve the desired privacy protection of each genome donor, each genome donor updates the global privacy budget and regards it as desired privacy budget. The specific process of constructing a federated aggregation model of multiparty genome data is as follows.

(1) Encoding of genome data

This study encodes genome data to diploid genotype data to achieve privacy preserving using differential privacy mechanism. For m-dimensional genome data of i-th genome donor, genome donor encodes it to m-dimensional vector x_i by combining her genome data with corresponding allele frequency dataset.

(2) Federated comparing update of local privacy budget achieving desired privacy preserving

Because each genome donor i only uses random noise to perturb genome data to achieve differential privacy preserving, each genome donor i publishes privacy budget will not violate privacy. In addition, the privacy budget is smaller, the privacy

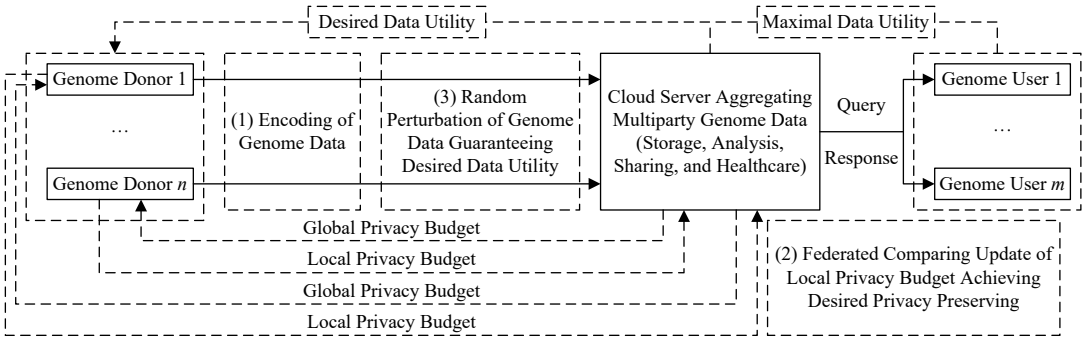

Fig. 2. Federated aggregation model of multiparty genome data.

preserving level is better. Therefore, cloud server uses the federated comparing to update the local privacy budget, so that all genome donors can achieve the desired privacy preserving. The specific process of doing federated comparing update of local privacy budget achieving desired privacy preserving is as follows.

- Each genome donor i selects the initial privacy budget ε_i according to her own normalized expected estimation error. Each genome donor i sends local privacy budget ε_i to cloud server.
- Cloud server makes a federated comparing to local privacy budget of all genome donors and obtains the global privacy budget $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$. Cloud server sends the global privacy budget ε to all genome donors.
- Considering the desired privacy budget for each genome donor i, genome donor i updates her own local privacy budget $\varepsilon_i = \varepsilon$.

The global privacy budget ε is equal to the minimum value of the initial privacy budget of all genome donors. That is to say, $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$. In this way, each genome donor i gets the desired privacy budget $\varepsilon_i = \varepsilon$. On this basis, each genome donor i can achieve the desired privacy preserving.

Example 2. There are three genome donors representing 1, 2, and 3 respectively. The initial privacy budget of the three genome donors are $\varepsilon_1 = 0.3$, $\varepsilon_2 = 1$, and $\varepsilon_3 = 0.5$ respectively. Cloud server can get the global privacy budget $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\} = 0.3$ by federated comparing. Thus, each genome donor obtains the desired privacy budget $\varepsilon_i = \varepsilon = 0.3 (i = 1, 2, 3)$.

(3) Random perturbation of genome data guaranteeing desired data utility

If the differential privacy mechanism is directly used for diploid genotype data, the diploid genotype data are completely unavailable. Therefore, each genome donor uses the differential privacy mechanism to randomly perturb diploid genotypes to categorical data $\{0, 1, 2\}$ under desired data utility. The specific process of random perturbation of genome data for guaranteeing desired data utility is as follows.

- Each genome donor i uses differential privacy mechanism $\mathcal{M}$ generating random noise X under desired privacy budget ε.
- Each genome donor i uniformly selects the random noise $X_{ij} \in X$, gets noise vector $X_i = (X_{i1}, \cdots, X_{im})$, and takes the rounding noise vector $round(X_i) = (round(X_{i1}), \cdots, round(X_{im}))$ while guaranteeing the desired data utility $U = \frac{|round(X_i) \mathrm{mod} 3 \equiv 0|}{|X_i|}$ via modular operation. Note that when each genome donor uniformly selects random noise, it is necessary to ensure that the ratio of the rounding noise divisible 3 is the desired data utility.
- Each genome donor i uses the rounding noise vector $round(X_i)$ to randomly perturb the diploid genotype data x_i, and then gets $x_i^{'} = x_i + round(X_i) \mathrm{mod} 3$ by carrying out modular operation to achieve the desired data utility $U = \frac{|x_i^{'} \cap x_i|}{|x_i|}$ of diploid genotype data.

All genome donors perform the same process as above, and finally get desired data utility $U = \frac{|x^{'} \cap x|}{|x|}$ of aggregating multiparty genome data. In this study, $x^{'} = (x_1^{'}, \cdots, x_n^{'})^{\top}$ represents the m-dimensional diploid genotype matrix of n genome donors after random perturbation using differential privacy mechanism under desired data utility.

B. Federated Aggregation Protocol of Multiparty Genome Data

We give the MGD-FAP under federated aggregation model of multiparty genome data, and its interactive implementation process is as follows.

Step 1: Cloud server sends the desired data utility U of aggregating multiparty genome data to all genome donors.

Step 2: Each genome donor i encodes her own genome data into a diploid genotype vector x_i. Each genome donor i selects local privacy budget ε_i according to her own normalized expected estimation error and sends local privacy budget ε_i to cloud server.

Step 3: Cloud server makes a federated comparing to the local privacy budget of all genome donors, and sends the global privacy budget $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$ to each genome donor i.

Step 4: Each genome donor i updates her own local privacy budget and getting the desired privacy budget $\varepsilon_i = \varepsilon$. Each genome donor i generates random noise X using differential privacy mechanism $\mathcal{M}$ under desired privacy budget $\varepsilon_i = \varepsilon$.

Each genome donor i selects random noise $X_{ij} \in X$ by uniform selection, and gets random noise vector X_i and rounding noise vector $round(X_i)$. Each genome donor i must ensure that the rounding noise vector satisfies equation $U = \frac{|round(X_i)\text{mod}3\equiv 0|}{|X_i|}$. Each genome donor i randomly perturbs diploid genotype vector x_i using rounding noise vector $round(X_i)$, and gets random diploid genotype vector $x_i^{'} = x_i + round(X_i)\text{mod}3$. Each genome donor i sends the random diploid genotype vector $x_i^{'}$ to cloud server.

Step 5: Cloud server aggregates random diploid genotype vector $x_i^{'}$ of each gnome donor i, and gets random diploid genotype matrix $x^{'}$ of all genome donors.

Thus, each genome donor i can achieve desired privacy preserving using desired privacy budget $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$ in MGD-FAP. Since the ratio of uniformly selected rounding noise vector divisible 3 is the desired data utility U for each genome donor, the MGD-FAP can achieve desired data utility $U = \frac{|x^{'} \cap x|}{|x|}$ of aggregating multiparty genome data. In this study, we use Laplace mechanism achieving desired privacy preserving of aggregating multiparty genome data. We do not consider the discrete Laplace mechanism and Gaussian mechanism. Because the discrete Laplace mechanism only produces integer random noise, it takes a long time to produce the amount of random noise that meets the desired data utility. Because the Gaussian mechanism also depends on the probability value δ, it is required to select an appropriate parameter δ.

C. Theoretical Analysis of Our Protocol

In this Section, we theoretically analyze our protocol from three theorems.

Theorem 3. The MGD-FAP achieves the desired privacy preserving.

Proof. The Laplace mechanism generates random noise X under the desired privacy budget $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$. Considering desired data utility $U = \frac{|round(X_i)\text{mod}3\equiv 0|}{|X_i|}$, each genome donor i uniformly selects random noise $X_{ij} \in X$ and gets random noise vector X_i. Each genome donor i gets the rounding noise vector $round(X_i)$ by rounding operation of X_i. Each genome donor i uses $round(X_i)$ to randomly perturb her diploid genotype vector x_i and gets $x_i + round(X_i)$. Each genome donor i performs modular operation on $x_i + round(X_i)$, and gets random diploid genotype vector $x_i^{'} = x_i + round(X_i)\text{mod}3$. Thus, each gnome donor i can achieve ε-differentia privacy of her genome data according to Theorem 1. By Theorem 2, the random perturbation $x + round(X)\text{mod}3$ of x is ε-differential privacy in the MGD-FAP. Thus, the MGD-FAP can achieve the desired privacy preserving. □

Theorem 4. The MGD-FAP achieves the desired data utility.

Proof. Each genome donor i uniformly selects random noise $X_{ij} \in X$ and gets random noise vector X_i under desired data utility U, and each genome donor i requires $U = \frac{|round(X_i)\text{mod}3\equiv 0|}{|X_i|}$. Thus, random perturbation $x_i^{'} = x_i + round(X_i)\text{mod}3$ of diploid genotype vector x_i achieves the desired data utility $U = \frac{|x_i^{'} \cap x_i|}{|x_i|}$. The random perturbation $x^{'} = x + round(X)\text{mod}3$ of aggregating multiparty genome data x also achieves the desired data utility $U = \frac{|x^{'} \cap x|}{|x|}$ according to the Eq. (3). □

Theorem 5. The MGD-FAP reaches the privacy-utility equilibrium.

Proof. We consider the strategic game between each genome donor i and cloud server. Each genome donor i updates their local privacy budgets ε_i by federated comparing, and obtains the desired privacy budget $\varepsilon_i = \varepsilon$ under the condition of $\varepsilon = \min\{\varepsilon_i\}_{i=1}^{n}$. The cloud server can get the desired data utility U by Theorem 4. Thus, MGD-FAP can reach the equilibrium between desired privacy preserving and desired data utility according to the Definition 4. □

According to the above theorems, MGD-FAP can achieve privacy-utility equilibrium using Laplace mechanism. In the following example, we show that MGD-FAP can solve the problems of Example 1.

Example 3. Considering genome donors are kinship, genome donors can get desired privacy budget guaranteeing the same level of privacy protection by federated comparing update of local privacy budget. Since each genome donor randomly perturb her genome data under desired data utility, cloud server can get the desired data utility of aggregating multiparty genome data. Thus, our MGD-FAP can solve the problems of utility disaster and privacy leakage of Example 1.

Moreover, since differential privacy guarantees better indistinguishability between adjacent databases with smaller privacy budget, our MGD-FAP can achieve any desired data utility to obtain any meaningful results of scientific research under desired privacy budget.

VI. Experimental Evaluation

In the experimental evaluation, we use the publicly available genome data and allele frequencies of chromosome 22 in the 1000 Genome Project (Phase 3)[2]. We encode the genome data to diploid genotype according to the allele frequencies of the Haplotypemap (HapMap) of the human genome. It contains the genome data of chromosome 22 of 165 CEU populations. In all experiments, we aggregate 100, 300, and 500 gene loci of 165 individuals from the diploid genotype dataset respectively, and we get the average experimental results of repeating the experiment 10 times. Because we consider the adjacent diploid genotype datasets with Hamming distance being 1, the ℓ_1-sensitivity is $\Delta_1 f = 2$.

A. Desired Data Utility

We use accuracy as both data utility and desired data utility metrics. Genome donors directly use Laplace mechanism, which makes the aggregating multiparty genome data completely unavailable. Therefore, we use the rounding value of the random noise of Laplace mechanism to randomly perturb the diploid genotype, and then do modular operation to obtain the random diploid genotype data in this experimental analysis. As shown in Fig. 3, the data utility of MGDA using Laplace mechanism increases with the increasing of the desired privacy budget, because the ratio of rounding noise

[2]ftp://ftp.ncbi.nlm.nih.gov/hapmap/

divisible 3 increases with the increasing of desired privacy budget. Because the ratio of different amount of rounding noise divisible 3 is identical under the same desired privacy budget, the data utility of MGDA using Laplace mechanism is the same for aggregating diploid genotype data of different dimensions under the same desired privacy budget. The data utility of MGDA using randomized response almost maintains the data utility $U = 0.5$, because the randomized response uses either true or false as the answer for equally likely. Thus, the data utility of MGDA using randomized response is the same for aggregating diploid genotype data of different dimensions under the different desired privacy budget.

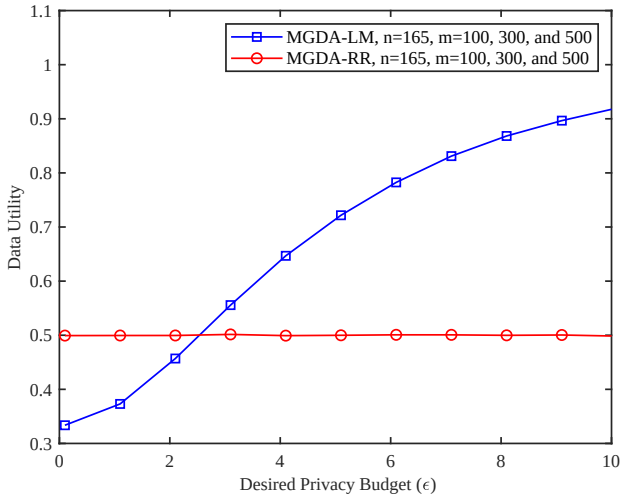

Fig. 3. Data utility of MGDA using LM or RR.

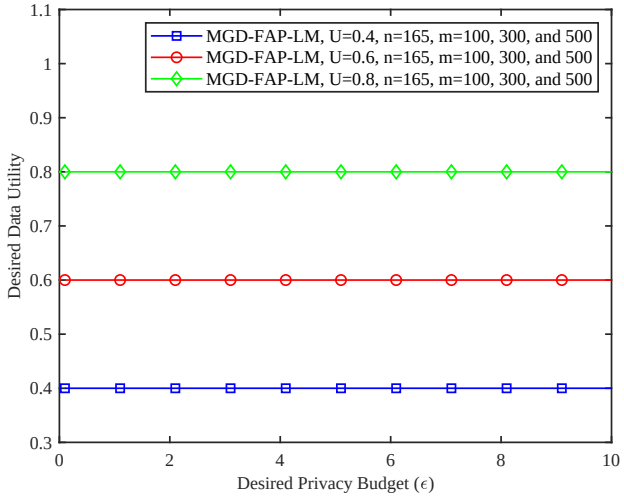

Fig. 4. Desired data utility of MGD-FAP using LM.

According to the desired data utility of cloud server, the MGD-FAP using Laplace mechanism can achieve desired data utility for aggregating diploid genotype data of different dimensions under different desired privacy budget in Fig. 4. Because the ratio of different amount of rounding noise divisible 3 is the desired data utility under any desired privacy budget, the desired data utility of aggregating genome data using Laplace mechanism for diploid genotype of different dimensions are identical in Fig. 4. Thus, these experimental results verify Theorem 4.

B. Desired Privacy Preserving

In Fig. 5, the normalized expected estimation error of MGDA using Laplace mechanism or randomized response decreases with the increasing of desired privacy budget, because the ratio of rounding noise divisible 3 increases with increasing of the desired privacy budget. Because the ratio of different amount of rounding noise divisible 3 is identical under the same desired privacy budget, the MGDA using Laplace mechanism has the same normalized expected estimation error for aggregating diploid genotype data of different dimensions under the same desired privacy budget. Since the randomized response uses either true or false as the answer for equally likely, the MGDA using randomized response has the same normalized expected estimation error for aggregating diploid genotype data of different dimensions under the same desired privacy budget. By Fig. 3, we can conclude that MGDA using Laplace mechanism can achieve privacy-utility tradeoff. Although the MGDA using randomized response has a small normalized expected estimation error under the same desired privacy budget, the MGDA using randomized response maintains the data utility being 0.5 in Fig. 3.

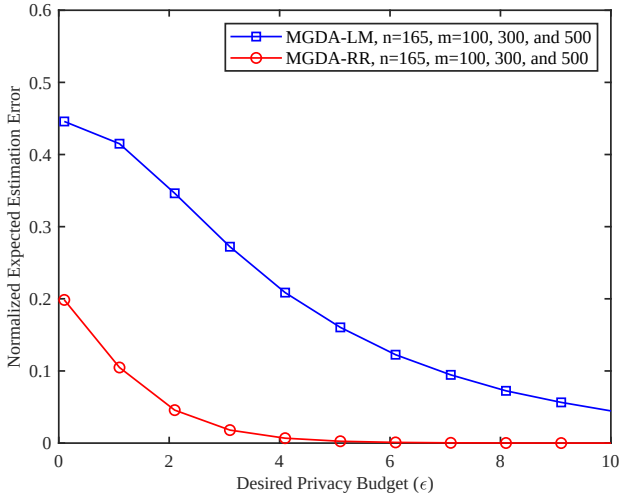

Fig. 5. Normalized expected estimation error of MGDA using LM or RR.

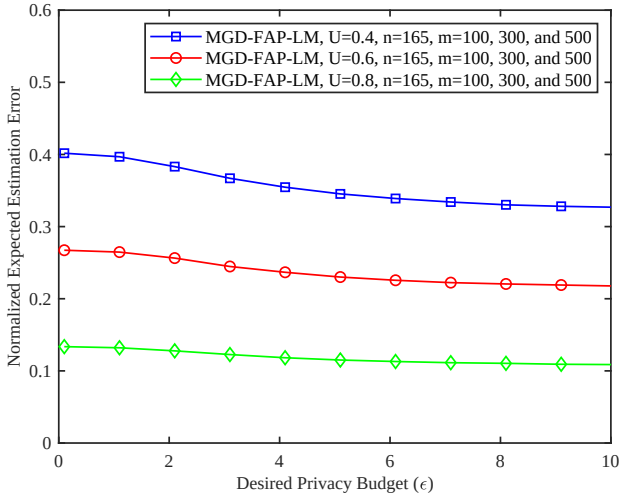

Fig. 6. Normalized expected estimation error of MGD-FAP using LM.

TABLE I
COMPARISON OF MGDA USING LM OR RR, AND MGD-FAP USING LM.

	Differential privacy mechanisms	Desired privacy budget	Desired data utility	Privacy-utility equilibrium
MGDA	LM RR	Yes	No	No
MGD-FAP	LM	Yes	Yes	Yes

As shown in Fig. 6, the normalized expected estimation error of MGD-FAP using Laplace mechanism decreases with the increasing of desired privacy budget, because the effect of privacy protection decreases with the increasing of the desired privacy budget. Moreover, when the desired privacy budget is identical, the normalized expected estimation error increases as the decrease of the desired data utility. The reason for this result is the ratio of retaining the correct genotype by random perturbation decreases with the decreasing of desired data utility. The MGD-FAP using Laplace mechanism has the same normalized expected estimation error for aggregating diploid genotype data of different dimensions under the same desired privacy budget, because the ratio of retaining the correct genotype is the same. In Fig. 6, we observed that the normalized expected estimation error of MGD-FAP using Laplace mechanism are the same for aggregating diploid genotype data of different dimensions under the same desired data utility, because the ratio of different amount of rounding noise divisible 3 is identical under the same desired privacy budget. This facilitates each genome donor to select the initial privacy budget based on the normalized expected estimation error in MGD-FAP. Therefore, we can use Laplace mechanism achieving the desired privacy preserving of MGD-FAP under the desired data utility.

In Table I, we make a comparative analysis of MGDA using Laplace mechanism or randomized response, and MGD-FAP using Laplace mechanism. Each genome donor obtains the desired privacy budget using the federated comparing update of local privacy budget in MGD-FAP. Considering the Laplace mechanism with desired privacy budget, each genome donor randomly perturbs her own diploid genotype data according to the desired data utility of the cloud server in MGD-FAP. Therefore, our experimental and theoretical results show that MGD-FAP can ensure an equilibrium between desired privacy preserving and desired data utility.

VII. CONCLUSION

This paper proposed a federated aggregation model of multiparty genome data and presented the MGD-FAP ensuring privacy-utility equilibrium. We used the desired privacy budget as the desired privacy protection measurement, and used accuracy as the desired data utility measurement. In MGD-FAP, each genome donor can get the desired privacy budget by federated comparing update of local privacy budget under desired data utility of cloud server. Our theoretical and experimental results show that MGD-FAP using Laplace mechanism can ensure privacy-utility equilibrium. This work is conducive to the federated aggregation of multiparty genome data by cloud server, and can be used to solve the problem of privacy-utility contradiction of aggregating multiparty genome data with differential privacy.

REFERENCES

[1] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X.F. Wang, "Privacy in the genomic era," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 6:1-6:44, 2015.

[2] S. D. Constable, Y. Tang, S. Wang, X. Jiang, and S. Chapin, "Privacy-preserving GWAS analysis on federated genomic datasets," *BMC Med. Inform. Decis.*, vol. 15, no. Suppl 5, pp. S2:1-S2:9, 2015.

[3] M. N. Sadat, M. M. A. Aziz, N. Mohammed, F. Chen, X. Jiang, and S. Wang, "SAFETY: Secure GWAS in federated environment through a hybrid solution," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 1, pp. 93-102, 2019.

[4] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends®Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211-407, 2014.

[5] M. M. A. Aziz, S. Kamali, N. Mohammed, and X. Jiang, "Online algorithm for differentially private genome-wide association studies," *ACM Trans. Comput. Heal.*, vol. 2, no. 2, pp. 13:1-13:27, 2021.

[6] S. Simmons and B. Berger, "Realizing privacy preserving genome-wide association studies," *Bioinform.*, vol. 32, no. 9, pp. 1293-1300, 2016.

[7] S. Simmons, C. S. Sahinalp, and B. Berger, "Enabling privacy-preserving GWAS in heterogeneous human populations," *Cell Syst.*, vol. 3, no. 1, pp. 54-61, 2016.

[8] F. Tramèr, Z. Huang, J.-P. Hubaux, and E. Ayday, "Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies," in *Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1286-1297.

[9] I. Hagestedt, Y. Zhang, M. Humbert, P. Berrang, H. Tang, X.F. Wang, and M. Backes, "MBeacon: Privacy-preserving beacons for DNA methylation data," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, 2019.

[10] S. Simmons, B. Berger, and C. Sahinalp, "Protecting genomic data privacy with probabilistic modeling," in *Proceedings of the Pacific Symposium on Biocomputing*, 2018, pp. 403-414.

[11] E. Yilmaz, E. Ayday, T. Ji, and P. Li, "Preserving genomic privacy via selective sharing," in *Proceedings of the 19th Workshop on Privacy in the Electronic Society*, 2020, pp. 163-179.

[12] A. Yamamoto and T. Shibuya, "More practical differentially private publication of key statistics in GWAS," *Bioinform. Adv.*, vol. 1, no. 1, pp. 1-10, 2021.

[13] M. Fredrikson, E. Lantz, S. Jha, S. M. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proceedings of the 23rd USENIX Security Symposium*, 2014, pp. 17-32.

[14] A. Honkela, M. Das, O. Dikmen, and S. Kaski, "Efficient differentially private learning improves drug sensitivity prediction," *Biol. Direct*, vol. 13, Article no. 1, 2018.

[15] T. T. Le, W. K. Simmons, M. Misaki, J. Bodurka, B. C. White, J. Savitz, and B. A. McKinney, "Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests," *Bioinform.*, vol. 33, no. 18, pp. 2906-2913, 2017.

[16] H. Liu, Z. Wu, C. Peng, X. Lei, F. Tian, and L. Lu, "Adaptive differential privacy of character and its application for genome data sharing," in *Proceedings of the International Conference on Networking and Network Applications*, 2019, pp. 429-436.

[17] N. Almadhoun, E. Ayday, and Özgür Ulusoy, "Differential privacy under dependent tuples - the case of genomic privacy," *Bioinform.*, vol. 36, no. 6, pp. 1696-1703, 2020.

[18] H. Liu, Z. Wu, C. Peng, F. Tian, and L. Lu, "Bounded privacy-utility monotonicity indicating bounded tradeoff of differential privacy mechanisms," *Theor. Comput. Sci.*, vol. 816, pp. 195-220, 2020.

[19] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2011, pp. 193-204.

[20] R. Chen, B. C. M. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," *VLDB J.*, vol. 23, no. 4, pp. 653-676, 2014.
[21] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 3:1-3:36, 2014.
[22] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2015, pp. 747-762.
[23] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2017, pp. 1291-1306.
[24] H. Wang and H. Wang, "Correlated tuple data release via differential privacy," *Inf. Sci.*, vol. 560, pp. 347-369, 2021.
[25] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genet.*, vol. 4, no. 8, pp. e1000167:1-e1000167:9, 2008.
[26] E. Ayday and M. Humbert, "Inference attacks against kin genomic privacy," *IEEE Secur. Priv.*, vol. 15, no. 5, pp. 29-37, 2017.
[27] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the lacks family: Quantification of kin genomic privacy," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2013, pp. 1141-1152.
[28] T. Pascoal, J. Decouchant, A. Boutet, and P. Esteves-Verissimo, "DyPS: Dynamic, private and secure GWAS," *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 2, pp. 1-21, 2021.
[29] X. Wu, H. Zheng, Z. Dou, F. Chen, J. Deng, X. Chen, S. Xu, G. Gao, M. Li, Z. Wang, Y. Xiao, K. Xie, S. Wang, and H. Xu, "A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis," *Briefings Bioinform.*, vol. 22, no. 3, pp. 1-10, 2021.
[30] Y. Zhang, C. Xu, N. Cheng, H. Li, H. Yang, and X. Shen, "Chronos^{+}: An accurate blockchain-based time-stamping scheme for cloud storage," *IEEE Trans. Serv. Comput.*, vol. 13, no .2, pp. 216-229, 2020.
[31] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, 2013, pp. 429-438.
[32] M. Maschler, E. Solan, and S. Zamir, Game Theory. Cambridge, UK: Cambridge University Press, 2013.
[33] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Am. Stat. Assoc.*, vol. 60, no. 309, pp. 63-69, 1965.
[34] F. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2009, pp. 19-30.

Hai Liu received his B.S. and M.S. degrees from Guizhou University, Guiyang, China, in 2012 and 2015, respectively, and the Ph.D. degree from Shaanxi Normal University, Xi'an, China, in 2019. He is currently working toward the Postdoctoral Researcher with Guizhou University, Guiyang, China. His research interests include applied cryptography and information security, and big data security and privacy preserving.

Changgen Peng received his Ph.D. degree from Guizhou University, China, in 2007. He is currently a professor of Guizhou University, China. His research interests include cryptography, information security, big data security and privacy preserving.

Youliang Tian received his B.S. and M.S. degrees from Guizhou University, China, in 2004 and 2009, respectively, and the Ph.D. degree from Xidian University, China, in 2012. He is currently a professor of Guizhou University. His current research interests include algorithmic game theory, cryptography and security protocols, big data security and privacy preserving, blockchain, electronic currency, etc.

Feng Tian received his Ph.D. degree from Xi'an Jiaotong University, China, in 2015. He is currently an associate professor of Shaanxi Normal University, China. His research interests include location privacy preserving and data mining with privacy preserving.

Zhenqiang Wu received his Ph.D. degree from Xidian University, China, in 2007. He is currently a professor of Shaanxi Normal University, China. His research interests include computer communications networks, wireless networks, network security, anonymous communication, and privacy preserving.