Answering Count Queries for Genomic Data With Perfect Privacy

Bo Jiang, Mohamed Seif, Ravi Tandon, *Senior Member, IEEE*, and Ming Li, *Senior Member, IEEE*

***Abstract*— In this paper, we consider the problem of answering count queries for genomic data subject to perfect privacy constraints. Count queries are often used in applications that collect aggregate (population-wide) information from biomedical Databases (DBs) for analysis, such as Genome-wide association studies. Our goal is to design mechanisms for answering count queries of the following form: *How many users in the database have a specific set of genotypes at certain locations in their genome?* At the same time, we aim to achieve perfect privacy (zero information leakage) of the sensitive genotypes at a pre-specified set of secret locations. The sensitive genotypes could indicate rare diseases and/or other health traits one may want to keep private. We present both local and central count-query mechanisms for the above problem that achieves perfect information-theoretic privacy for sensitive genotypes while minimizing the expected absolute error (or per-user error probability, depending on the setting) of the query answer. We also derived a lower bound of the per-user probability of error for an arbitrary query-answering mechanism that satisfies perfect privacy. We show that our mechanisms achieve error close to the lower bound, and match the lower bound for some special cases. We numerically show that the performance of each mechanism depends on the data prior distribution, the intersection between the queried and sensitive genotypes, and the strength of the correlation in the genomic data sequence.**

***Index Terms*— Genomic data protection, perfect information privacy, counting query, information privacy.**

I. Introduction

GENETIC research is experiencing an unprecedented growth in terms of the amount of personal data collected in large repositories [1]. The human genome is the complete set of genetic information, which is composed of four different bases (A, C, G, T). Genetic information is encoded inside chromosomes, and each chromosome contains genes responsible for various functions controlling the human body all together [2]. They provide medical researchers invaluable information that can play a role in different interesting applications such as sequencing [3], genome sequence assembly [4], [5] and Genome-Wide Association Studies (GWAS) [6]. The main objective of these applications is to enable the investigation of variants in human genomes with different biological or health-related traits through interactive queries with biomedical databases (DBs) or directly from each record. For example, medical researchers might want to query how many users in a dataset have specific genotypes at certain locations in their genome. This can help to reveal relationships that exist in certain genotypes and biological traits or discover particular diseases like cancer [7], diabetes, and even more complex ones [8]. Such systems can also be used as a powerful tool for finding new drug targets [9].

Despite the impact of these applications on health services, privacy remains a fundamental concern that needs to be addressed. Indeed, existing query-answering systems such as STRIDE [10] and i2b2 [11] can leak sensitive genotypes about individuals [12], [13], [14]. Therefore, a significant research effort has been made to enable privacy-preserving access to genomic data. Recent interesting works in the literature addressed this issue in genomics under different settings, including sequential genomic data release [15], query-answering in biomedical DBs [16], [17], [18]. Depending on whether the mechanism relies on a trusted third party, these privacy protection models can be classified into two categories: centralized models assume there is a trusted server that processes the dataset and releases privatized answers to specific types of queries. For example, Cho et al. [18] proposed differentially private (DP) mechanisms for count queries using geometric additive noise mechanisms. Local models, on the other hand, enables each data holder (user) to perturb his/her genomic data sequence locally before publishing it to the server, which can be used to answer any query afterward, such as [15].

In the past two decades, DP has been widely accepted as the de facto standard in the privacy research community. However, DP only leads to a few real-world adoptions. One of the main reasons is that DP is a stringent (worst-case) privacy notion, i.e., there is no assumption on the underlying prior distribution of the data. While privacy guarantees hold for the worst-case realizations of the data, this could lead to lower data utility. Hence, it is not ideal for scenarios when the prior distribution of genomic data is available, e.g., through public datasets and/or via existing statistical models of genomic data sequence [19], [20]. In contrast, context-aware privacy notions, such as mutual information privacy, can lead to mechanism designs with a better privacy-utility tradeoff by leveraging the prior information. Note that despite

Manuscript received 15 November 2022; revised 14 June 2023; accepted 16 June 2023. Date of publication 22 June 2023; date of current version 30 June 2023. The work of Ravi Tandon was supported in part by NSF under Grant CNS 2317192, Grant CCF 2100013, Grant CNS 2209951, and Grant CCF 1651492. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Theodorakopoulos. *(Corresponding author: Bo Jiang.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721 USA (e-mail: bjiang@email.arizona.edu; mseif@email.arizona.edu; tandonr@email.arizona.edu; lim@email.arizona.edu).

Digital Object Identifier 10.1109/TIFS.2023.3288812

the high diversity, most of the DNA Sequence is common across the whole human population. Only around 0.5% of each person's DNA is different from the reference genome, owing to genetic variations [21]. This makes it suitable to adopt context-aware privacy notions for genomic sequence. Besides, context-aware privacy notions model each genomic data sequence as a random variable, which considers the correlation inside the data sequence.

In this paper, we consider K individuals participating in a genome variants survey, and each person's N-length genome sequence is drawn from a known distribution. Specifically, our goal is to design mechanisms for answering count queries of the following form: *How many users have a specific set of genotypes (denoted by a set $v_{\mathcal{L}}$) at certain locations (denoted by $\mathcal{L}$) in their genome?* At the same time, we aim to achieve perfect secrecy (zero information leakage) (perfect information-theoretic privacy) about sensitive genotypes at a pre-specified set $\mathcal{S}$ of secret locations. Our main contributions are summarized as follows:

Main Contributions:

1) We present two count-query mechanisms for genomic data sequences in the central and local settings, respectively. The mechanisms achieve perfect privacy for sensitive genotypes at the locations belonging to a predefined set $\mathcal{S}$ while optimizing a utility metric (i.e., minimizing the probability of error for local setting and expected absolute distance for central setting). The proposed mechanisms incur less computation complexity and achieve significantly higher utility than a previous perfect secret mechanism [15] and achieve a better utility-privacy tradeoff than DP-based mechanisms [18].

2) We theoretically analyze the performance of each mechanism under different cases regarding the strength of correlations within the genomic sequence and the length of the overlapped genotypes between the queried locations and sensitive locations. The proposed mechanisms only depend on the statistics of the subsequences derived from the intersected set of locations. Thus, the complexity of our mechanisms only grows with $|\mathcal{L} \cup \mathcal{S}|$, irrespective of N. We also derive a lower bound on the error probability (P_e) for an arbitrary mechanism subject to perfect privacy constraints. We show the optimality of our proposed mechanisms for several special cases.

3) We present comprehensive numerical simulations to show the utility-privacy tradeoff of the proposed mechanisms and compare them with existing DP-based mechanisms. Results show that our mechanisms achieve perfect privacy and incur smaller errors than those based on DP under certain large privacy budgets ϵ (larger ϵ indicates larger privacy leakage). We also show that the performance of the mechanisms is largely dependent on the data prior as well as the data dependence: the performance of each mechanism can be improved when the queried locations are less correlated with the sensitive locations.

II. Related Works

A fundamental need in designing genomic privacy-preserving mechanisms is to selectively limit the leakage of information about biological or health-related traits of an individual that can be inferred from the shared genetic data. Under this context, in [22], Thenen et al. have shown that beacons, previously deemed secure against re-identification attacks, were vulnerable despite their stringent policy, due to the correlation and inter-dependence inside each genomic sequence. To this end, [17], [18] study differential privacy mechanisms for sharing aggregate genomic data. However, DP-based mechanisms provide worst-case privacy guarantees that incur too much noise and lead to decreased utility in terms of accuracy. Furthermore, traditional DP-based mechanisms do not leverage correlations that exist in the genomic sequence, which may degrade the privacy guarantees offered by DP. In [23], Nour et al. first demonstrate this drawback of DP by proposing an attack and then propose a mechanism for privacy-preserving sharing of statistics from genomic datasets to attain privacy guarantees while taking into consideration the data dependence. From there, we assert that information theoretical privacy measurements are more favorable for genomic data processing for the following reasons. 1. Different individuals' genomic sequences possess extremely similar patterns at certain locations, which provides a solid foundation for estimating the underlying data distribution. 2. Dependence and inter-correlations that exist in the genomic sequence can be explicitly measured and leveraged in the mechanism design. We next discuss some of the related works that have considered information-theoretic privacy within this context.

A privacy-preserving genomic data sequence release problem was studied in [15], where the genotypes at some specific sensitive locations and some non-sensitive locations but are correlated to sensitive locations are hidden. This model also provides a baseline approach for query answer release (perturb then query). However, the query-answering mechanism incurs much less noise than sequence release mechanisms (either local or centralized ones). Other than that, the sequence release model was built for the local settings only. In this work, we study query-answering models in both local and central settings.

There is also a line of work on the privacy-preserving data release problem that is similar to our settings [24], [25], [26]. The variables form a Markov chain $S \to X \to Y$, where X and Y represent the input and output data of the mechanism, respectively, and S denotes a private latent variable that correlates to X. Based on this model, [24] proposes principal inertia components, which provide a fine-grained decomposition of the dependence between two random variables. It also proves that the smallest PIC of $P_{S,X}$ plays a central role in achieving perfect privacy (i.e. $I(S;Y) = 0$): If $|X| \leq |S|$, then perfect privacy is achievable with $I(X;Y) > 0$ if and only if the smallest PIC of $p_{S,X}$ is 0. However, the mechanisms provided in [24] are challenging to be implemented in genomic data, as it requires an exhausting search of functions in the defined region.

In [25], the system allows post-processing after taking observation on Y, then data utility is measured by the mean square error (MSE) between X and the estimation $E[X|Y]$. Finally, the utility and privacy tradeoff can be formulated by minimizing the MSE while subject to certain privacy constraints. However, this work focuses on improving the utility-privacy tradeoff and do not provide closed-form

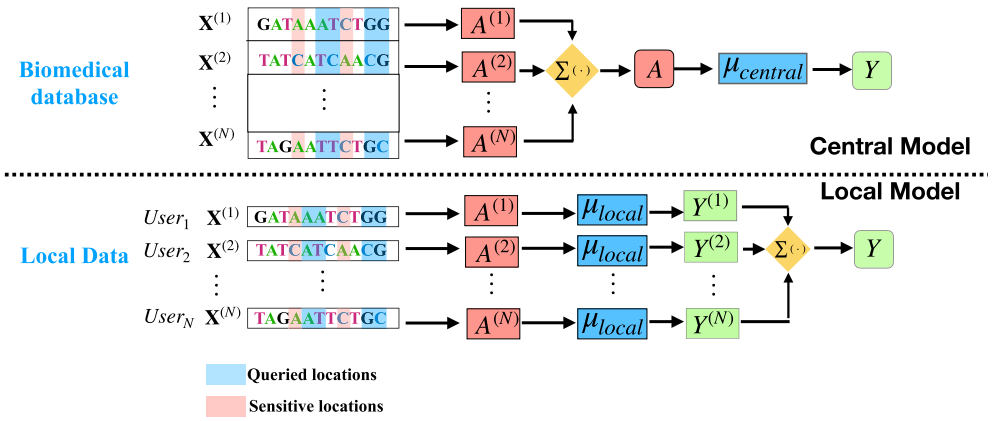


Fig. 1. The system model. The queried sub-sequence $X_{\mathcal{L}}^{(m)}$ for the m-th user and the query $v_{\mathcal{L}}$ altogether pass through a deterministic equality test function with a true answer $A^{(m)}$, followed by a private release mechanism. The perturbed answer Y guarantees perfect privacy for sensitive genotypes.

privatization mechanisms. Also, this paper studies general privacy guarantees, not perfect privacy.

In [26], privacy and utility are defined as the probabilities of correctly guessing S and X given Y, respectively. This paper also mentioned mechanism design. However, firstly, the perfect privacy measured by correct guessing cannot imply independence of the random variable (what we defined as perfect privacy in this paper), as $P_C(S|Y) = P_C(S)$ does not imply $P(S|Y) = P(S)$. Also, the modeling of the correlation between the latent variable and the input data is too simplified: the correlated latent variable considered in [26] is also binary, and the correlation can be represented by two parameters. Finally, the mechanism is derived as a function of the correlation parameters. But for genomic data processing, the data sequence is a vector with each value having a cardinality of 4. Thus the mechanism in [26] does not apply to our problem. On the other hand, the model we considered is in the form of $S \to f(X) \to Y$, where S and X are vectors, $f(X)$ and Y are binary, which is similar to, but different from related works in [24], [25], and [26].

III. Model Setup & Problem Statement

Consider a set of K sequences $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}\}$, where each sequence $\mathbf{X}^{(k)} = \{X_1^{(k)}, X_2^{(k)}, \ldots, X_N^{(k)}\}$, for $k = 1, 2, \ldots, K$ is of length N. The entries $X_n^{(k)}$ are drawn from a finite alphabet $\mathcal{X}$ (e.g. for genomic data, $\mathcal{X} = \{A, T, G, C\}$). We assume that the sequences are independent and each sequence is drawn from a known distribution $P_{\mathbf{X}}(\bar{x}) = P_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N)$ ($P_{\mathbf{X}}(\bar{x})$ can be different from user to user, we simplify by removing users' index). In users' genomic sequence, some locations are private and need to be protected while others are non-sensitive. Denote $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ as the set of indices of the sensitive locations of the genomic data sequence (each $X_n^{(k)}$ is sensitive $\forall k \in \{1, 2, \ldots, K\}, \forall n \in \mathcal{S}$), and let $\mathbf{X}_{\mathcal{S}}^{(k)} = (X_{s_1}^{(k)}, X_{s_2}^{(k)}, \ldots, X_{s_{|\mathcal{S}|}}^{(k)})$ denote the sensitive genomic subsequence of the k-th user.

A count query can be defined as follows: the number of users in the dataset with a specific sequence at some particular locations. Mathematically, we denote the query as $\mathcal{Q} = \{\mathcal{L}, v_{\mathcal{L}}\}$, where $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$ denotes the indices of the queried locations and $\mathbf{X}_{\mathcal{L}}^{(k)} = (X_{l_1}^{(k)}, X_{l_2}^{(k)}, \ldots, X_{l_{|\mathcal{L}|}}^{(k)})$ denotes the corresponding queried sequence for the k-th user; $v_{\mathcal{L}}$ denotes the reference sequence and v_j denotes the reference value at location j. Then the local query of the k-th user is $Q^{(k)} = \{\mathbf{X}_{\mathcal{L}}^{(k)} = v_{\mathcal{L}}\}$. Note that, the sensitive locations $\mathcal{S}$ and the queried locations $\mathcal{L}$ may or may not have common indices. Denote $\bar{\mathcal{L}} = \mathcal{L} \backslash (\mathcal{L} \cap \mathcal{S})$ as the non-sensitive query locations, and $\bar{\mathcal{S}} = \mathcal{S} \backslash (\mathcal{L} \cap \mathcal{S})$.

Let $A^{(k)}$ be the true query answer calculated from the k-th user. Then $A^{(k)} = \mathbb{1}_{\{\mathbf{X}_{\mathcal{L}}^{(k)} = v_{\mathcal{L}}\}} = \prod_{j=1}^{|\mathcal{L}|} \mathbb{1}_{\{\mathbf{X}_{l_j}^{(k)} = v_j\}}$ and the final count can be expressed as $A = \sum_{k=1}^{K} A^{(k)}$, i.e., the aggregated answer is decomposable. To avoid leaking sensitive information about each $\mathbf{X}_{\mathcal{S}}^{(k)}$, we design private mechanisms and release a perturbed version while achieving high utility. Depending on whether there exists a trusted server, there are two different settings for the data protection mechanisms: for the local setting, each user privatizes $A^{(k)}$ independently and publishes a perturbed version $Y^{(k)}$, then $Y = \sum_{k=1}^{K} Y^{(k)}$ is the aggregated result; For the central setting, the data server first aggregates A from each $A^{(k)}$, then perturbs A and releases Y as a privatized version. Each of the local and central mechanisms has its pros and cons. Generally speaking, the central model guarantees a comparable level of privacy protection as the local model with a smaller amount of noise. The local model, on the other hand, provides a customizable privacy budget for each individual without relying on any third party. The system models considered are depicted in Fig. 1.

Data Privacy: In this paper, we aim to satisfy perfect privacy of sensitive genotypes, which is defined as:

$$P_{Y^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(y|\mathbf{x}_{\mathcal{S}}) = P_{Y^{(k)}}(y), \forall k \in \{1, 2, \ldots, K\}, \quad (1)$$

for local model, and

$$P_{Y|\mathbf{X}_{\mathcal{S}}^{(k)}}(y|\mathbf{x}_{\mathcal{S}}) = P_Y(y), \forall k \in \{1, 2, \ldots, K\} \quad (2)$$

for central model. The privacy-preserving mechanism must guarantee that the output of the privacy protection mechanism (Y for central model, $Y^{(k)}$ for the local model) is independent of the genomic data $\mathbf{X}_{\mathcal{S}}^{(k)}$ at sensitive locations $\mathcal{S}$ for every

TABLE I
LIST OF SYMBOLS

$\mathbf{X}$	Genomic data sequence
N	Length of the genomic sequence
k	individual index
$\mathcal{X}$	Finite set {A,T,G,C}
$P(\cdot)$	Probability distribution
$\mathcal{S}$	Set of indices of the sensitive locations
$\mathcal{Q}$	Count query
$\mathcal{L}$	Set of indices of the query locations
$v_{\mathcal{L}}$	Reference sequence
$\mathcal{M}$	Privacy protection mechanism
Y	Release of the privacy protection mechanism
P_e	Probability of error
R	Dependence measurement

user k. In the following, we refer to the conditions in (1) as perfect privacy constraints.

Performance (Utility): The performance of the mechanism is measured by the Expected Average Error (EAE) between A and Y. Mathematically,

$$\text{EAE} = E[|A - Y|]. \tag{3}$$

In this paper, we aim to maximize the utility (minimize the EAE) subject to perfect privacy. We summarized symbols used throughout this paper in Table I.

IV. MAIN RESULTS

In this paper, we aim to minimize the EAE subject to perfect local privacy. In this Section, we investigate privacy-preserving query answer mechanisms for local and central settings, respectively.

A. Local Mechanism Design

Next, we present two mechanisms based on different data processing settings which satisfy perfect local privacy for sensitive genotypes. To measure the performance of the proposed mechanisms for aggregated query, we define the utility for the aggregated query as the Expected Absolute Error (EAE):

$$E[|Y - A|] = E\left[\left|\sum_{k=1}^{K}\left(Y^{(k)} - A^{(k)}\right)\right|\right]. \tag{4}$$

Note that the EAE can be further upper-bounded as follows:

$$E[|Y - A|] \leq \sum_{k=1}^{K} E\left[|Y^{(k)} - A^{(k)}|\right] \triangleq \sum_{k=1}^{K} P_e^{(k)},$$

where $P_e^{(k)}$ denotes the per-user error probability. Given a query, we will design binary perturbation mechanisms for each sequence to minimize the local query error while satisfying per-user perfect privacy constraints. Formally, the problem can be described as

$$\min P_e^{(k)}, \text{ s.t. (1)}, \ \forall k = 1, 2, \ldots, K. \tag{5}$$

A local privacy-protection mechanism $\mathcal{M}^{(k)}$ with parameter $\mu(\mathbf{X}^{(k)}, \mathcal{Q})$ can be described as follows: $\mathcal{M}^{(k)}$ takes as input the local genomic sequence $\mathbf{X}^{(k)}$ of a user and the query $\mathcal{Q} = \{\mathcal{L}, v_{\mathcal{L}}\}$. Then, it extracts the sensitive data sequence $\mathbf{X}_{\mathcal{S}}^{(k)}$ and the queried data sequence $\mathbf{X}_{\mathcal{L}}^{(k)}$ respectively. From there, $\mathcal{M}^{(k)}$ releases $Y^{(k)} = 0$ or $Y^{(k)} = 1$ according to the data prior distribution as well as the correlation between $\mathbf{X}_{\mathcal{S}}^{(k)}$ and $\mathbf{X}_{\mathcal{L}}^{(k)}$. The corresponding release probabilities are defined as:

$$Y^{(k)} = \begin{cases} 1, & \text{w.p.} \quad \mu(\mathbf{X}^{(k)}, \mathcal{Q}), \\ 0, & \text{w.p.} \quad 1 - \mu(\mathbf{X}^{(k)}, \mathcal{Q}). \end{cases}$$

We present two mechanisms, $\mathcal{M}_1^{(k)}$, and $\mathcal{M}_2^{(k)}$, with parameters $\mu_1(\mathbf{X}^{(k)}, \mathcal{Q})$ and $\mu_2(\mathbf{X}^{(k)}, \mathcal{Q})$ given as follows:

$$\mu_1(\mathbf{X}^{(k)}, \mathcal{Q}) \triangleq P_{Y^{(k)}|\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}, \mathbf{X}_{\mathcal{S}}^{(k)}}(1|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) =$$
$$\begin{cases} \dfrac{\min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|w)}{P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}})}, & \text{if } x_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}, \mathcal{E} \leq 1/2, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

$$\mu_2(\mathbf{X}^{(k)}, \mathcal{Q}) \triangleq P_{Y^{(k)}|\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}, X_{\mathcal{S}}^{(k)}}(1|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) =$$
$$\begin{cases} 1, & \text{if } x_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}, \mathcal{E} \leq 1/2, \\ 1 - \dfrac{\min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|w)}{P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}})}, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{E}$ is defined as $\mathcal{E} \triangleq \Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}^{(k)} \neq v_{\mathcal{L}\cap\mathcal{S}})$. Both of the proposed mechanisms use the following probability ratio:

$$R(\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}, \mathbf{X}_{\mathcal{S}}^{(k)}) = \frac{\min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|w)}{P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}})}, \tag{8}$$

which can be (informally) used to measure the statistical dependence between $\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}$ (genotypes in $\mathcal{L}$ which are not in the sensitive locations) and $\mathbf{X}_{\mathcal{S}}^{(k)}$ (genotypes in sensitive locations). In the following, we use R for simplicity. Specifically, a value of $R = 1$ clearly implies independence of $\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}$ and $\mathbf{X}_{\mathcal{S}}^{(k)}$, whereas $R < 1$ implies that $\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}$ and $\mathbf{X}_{\mathcal{S}}^{(k)}$ are dependent, and the dependence grows as R becomes small.

Given a query $Q = \{\mathcal{L}, v_{\mathcal{L}}\}$, mechanism $\mathcal{M}_1^{(k)}$ first looks at the sequence $x_{\bar{\mathcal{L}}}$ that corresponds to the non-sensitive genotypes in the set $\bar{\mathcal{L}}$. If $x_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}$ and $\mathcal{E} < 1/2$, the mechanism releases 1 with probability R, otherwise, it always releases 0. Mechanism $\mathcal{M}_2^{(k)}$ works similarly as follows. It first looks at the sequence $x_{\bar{\mathcal{L}}}$, if $x_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}$ and $\mathcal{E} < 1/2$, then $\mathcal{M}_2^{(k)}$ always releases 1. Otherwise, it releases 1 with probability $1 - R$. Our first main result is stated in the following Theorem.

Theorem 1: Mechanisms $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2^{(k)}$ satisfy perfect privacy.

Proof: We expand $P_{Y^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(1|x_{\mathcal{S}})$ in terms of $x_{\bar{\mathcal{L}}}$ using total probability Theorem. This yields the following set of steps:

$$\begin{aligned} P_{Y^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(1|x_{\mathcal{S}}) &= \sum_{x_{\bar{\mathcal{L}}}} P_{Y^{(k)}|\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}, X_{\mathcal{S}}^{(k)}}(1|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|X_{\mathcal{S}}^{(k)}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) \\ &= P_{Y^{(k)}|\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}, \mathbf{X}_{\mathcal{S}}^{(k)}}(1|v_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{\mathbf{X}_{\bar{\mathcal{L}}}^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)}}(v_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) \\ &\quad + \sum_{x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} \end{aligned}$$

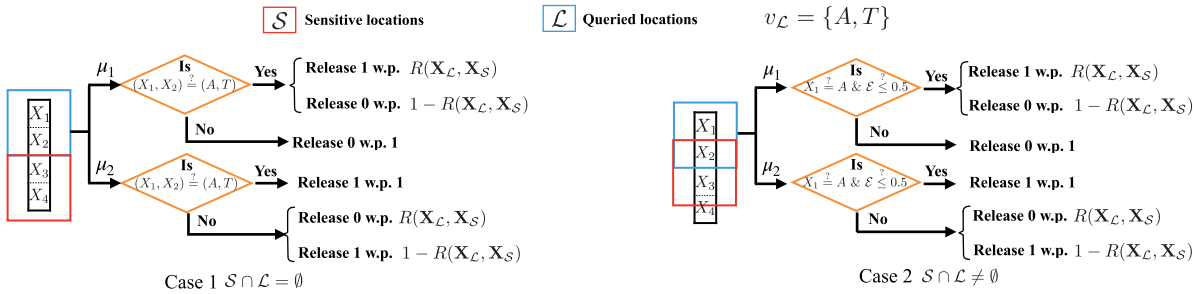


Fig. 2. Description for the illustrative example for the two mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$.

$$
\begin{aligned}
&P_{Y^{(k)}|\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}\mathbf{X}^{(k)}_{\mathcal{S}}}(1|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|\mathbf{X}^{(k)}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) \\
&= \mu_1^{(k)}(v_{\bar{\mathcal{L}}}, Q, S) P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|X^{(k)}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) \\
&\quad + \sum_{x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} \mu_1^{(k)}(x_{\bar{\mathcal{L}}}, Q, S) P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|X^{(k)}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) \\
&\stackrel{(a)}{=} \min_w P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|X^{(k)}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}|x_{\mathcal{S}} = w),
\end{aligned}
$$

where step (a) follows from the proposed mechanism in (6). Observe that the conditional probability does not depend on the realizations $x_{\mathcal{S}}$. Then, it is straightforward to show that

$$P_{Y^{(k)}|\mathbf{X}^{(k)}_{\mathcal{S}}}(1|x_{\mathcal{S}}) = P_Y^{(k)}(1).$$

The calculation for $\mathcal{M}_2^{(k)}$ follows on similar lines, and hence both mechanisms satisfy perfect privacy for sensitive genomes. □

We next present our second result which characterizes the error probability of the proposed mechanisms.

Theorem 2: Define $P_{e,1}^{(k)}$ and $P_{e,2}^{(k)}$ as the per-user error probability for $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2^{(k)}$, respectively, then:
- Case 1: $\mathcal{E} \leq 1/2$:

$$
\begin{aligned}
P_{e,1}^{(k)} &= P_{\mathbf{X}^{(k)}_{\mathcal{L}}}(v_{\mathcal{L}}) + (2\mathcal{E} - 1) \min_w P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|\mathbf{X}^{(k)}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}|w), \\
P_{e,2}^{(k)} &= 1 - P_{\mathbf{X}^{(k)}_{\mathcal{L}}}(v^{(k)}_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} \min_w P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|\mathbf{X}^{(k)}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|w). \qquad (9)
\end{aligned}
$$

- Case 2: $\mathcal{E} > 1/2$:

$$
\begin{aligned}
P_{e,1}^{(k)} &= P_{\mathbf{X}^{(k)}_{\mathcal{L}}}(v_{\mathcal{L}}), \\
P_{e,2}^{(k)} &= 1 - P_{\mathbf{X}^{(k)}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} \min_w P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|\mathbf{X}^{(k)}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|w) \\
&\quad - (2\mathcal{E} - 1) \min_w P_{\mathbf{X}^{(k)}_{\bar{\mathcal{L}}}|\mathbf{X}^{(k)}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}|w).
\end{aligned}
$$

The proof is presented in the Appendix.

1) Illustrative Example: We next explain our mechanisms through an illustrative example. Consider a user's genome sequence $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, and a target query $\mathcal{Q} = \{\mathcal{L} = \{1, 2\}, v_{\mathcal{L}} = \{A, T\}\}$, i.e., the query is of the form "*Is* $(X_1, X_2) = (A, T)$*?*". In this example, we consider two scenarios for the set $\mathcal{S}$ of sensitive locations. Specifically, we can either have $\mathcal{S} \cap \mathcal{L} = \emptyset$ or $\mathcal{S} \cap \mathcal{L} \neq \emptyset$.

Case 1: $\mathcal{S} \cap \mathcal{L} = \emptyset$ When $\mathcal{S} = \{3, 4\}$, if $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}}) = 0$, mechanism $\mathcal{M}_1$ always releases 0 and $\mathcal{M}_2$ always releases 1. When $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}}) = 1$, both mechanisms release the true answer, i.e., $Y = A$. For the case when $0 < R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}}) < 1$, each mechanism perturbs the true answer as follows:

$\mathcal{M}_1$ first examines whether $\{X_1, X_2\} = \{A, T\}$, if no, then it releases the correct answer $Y = 0$; if yes, then it releases $Y = 1$ with probability $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$, and $Y = 0$ with the remaining probability $1 - R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$. On the other hand, for mechanism $\mathcal{M}_2$, if $\{X_1, X_2\} = \{A, T\}$, then it releases $Y = 1$; if not, then it releases $Y = 0$ with probability $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$ (and $Y = 1$ with the remaining probability 1- $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$). The intuition behind the mechanisms is to limit the leakage which arises due to the dependence between the genotypes in non-sensitive query locations $\bar{\mathcal{L}}$ and the sensitive locations $\mathcal{S}$.

Case 2: $\mathcal{S} \cap \mathcal{L} \neq \emptyset$ We next consider the case when $\mathcal{S} = \{2, 3\}$, i.e., $\mathcal{L} \cap \mathcal{S} = \{2\}$. Similar to Case 1, when $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}}) = 0$, mechanism $\mathcal{M}_1$ always releases 0 and $\mathcal{M}_2$ always releases 1 (which clearly satisfies perfect privacy). When $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}}) > 0$, however, since the intersection $\mathcal{L} \cap \mathcal{S} = \{2\}$ is non-empty, both mechanisms first check if $\mathcal{E} = P(X_2 \neq v_2) = P(X_2 \neq T) < 1/2$, and if yes, then proceed in a similar manner as before. Specifically, each mechanism first examines whether the non-sensitive query sequence $X_{\bar{\mathcal{L}}}$ matches $v_{\bar{\mathcal{L}}}$ and $\mathcal{E} < 1/2$, if yes, then $\mathcal{M}_1$ releases $Y = 1$ with probability $R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$, and $Y = 0$ with the remaining probability $1 - R(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{S}})$; if no, it releases the correct answer $Y = 0$. Both cases are illustrated in Fig. 2.

Remark 1: As an alternative baseline scheme, one can perturb the whole genome sequence first by (for instance, by applying the scheme in [15]) and then answer the count. However, perturbing the whole sequence is unnecessary, and the complexity of such a scheme will grow exponentially with the genome sequence length N. In contrast, the complexity of our schemes grows exponentially with $|\mathcal{L} \cup \mathcal{S}|$ which can be substantially smaller than N.

Corollary 1: We can readily specialize the general result of Theorem 2 for the case when genomic data sequences are i.i.d, and each genotype is uniformly distributed over a finite alphabet $\mathcal{X}$. Define $\lambda = \left(\frac{1}{|\mathcal{X}|}\right)^{|\mathcal{L}|}$, then for $\mathcal{M}_1^{(k)}$ and $\mathcal{M}_2^{(k)}$, if $\mathcal{E} \leq 0.5$, $\min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\} = 0$; if $\mathcal{E} > 0.5$, $\min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\} = \lambda$, where $\mathcal{E} = 1 - P_{\mathbf{X}^{(k)}_{\mathcal{L}\cap\mathcal{S}}}(v_{\mathcal{L}\cap\mathcal{S}}) = 1 - \left(\frac{1}{|\mathcal{X}|}\right)^{|\mathcal{L}\cap\mathcal{S}|}$.

2) Lower Bound on P_e: We next derive an information-theoretic lower bound on P_e for any local mechanism that satisfies perfect privacy. The goal is to examine how far our mechanisms are from optimality.

Theorem 3: For any privacy-preserving local data release mechanisms that provide perfect privacy for genomic data at sensitive locations $\mathcal{S}$, the P_e is lower bounded as:

$$P_e^{(k)} \geq h^{-1}\left\{h(A^{(k)}) - \min\{h(A_{\tilde{\mathcal{L}}}^{(k)}|A_{\mathcal{L}\cap\mathcal{S}}^{(k)}), h(A^{(k)}|\mathbf{X}_{\mathcal{S}}^{(k)})\}\right\}, \tag{10}$$

where $h(\cdot)$ denotes the binary entropy function, $A_{\tilde{\mathcal{L}}}^{(k)} \triangleq \mathbb{1}_{\{\mathbf{X}_{\tilde{\mathcal{L}}}^{(k)}=v_{\tilde{\mathcal{L}}}\}}$ and $A_{\mathcal{L}\cap\mathcal{S}}^{(k)} \triangleq \mathbb{1}_{\{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}^{(k)}=v_{\mathcal{L}\cap\mathcal{S}}\}}$.

This bound follows by using Fano's inequality and the privacy constraint in (1). Detailed proof is provided in the Appendix. Denote $LB(P_e^{(k)})$ as the lower bound of $P_e^{(k)}$ calculated from Theorem 3, we next consider some special cases and compare the lower bound with the error probability of the proposed mechanisms.

Remark 2: Under the following scenarios, $LB(P_e^{(k)})$ from Theorem 3 matches $\min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\}$ from Theorem 2:

(1) When $\mathcal{L}\cap\mathcal{S} = \emptyset$: we consider two scenarios regarding the data correlation between $\mathbf{X}_{\mathcal{L}}^{(k)}$ and $\mathbf{X}_{\mathcal{S}}^{(k)}$.

a) If $I(A^{(k)}; \mathbf{X}_{\mathcal{S}}^{(k)}) = h(A^{(k)})$, $LB(P_e^{(k)}) = \min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\} = \min\{P_A^{(k)}(0), P_A^{(k)}(1)\}$, which means when $\mathbf{X}_{\mathcal{S}}^{(k)}$ and $\mathbf{X}_{\mathcal{L}}^{(k)}$ are fully dependent, any mechanism cannot outperform directly sampling result from the prior distribution.

b) If $I(A^{(k)}; \mathbf{X}_{\mathcal{S}}^{(k)}) = 0$, $LB(P_e^{(k)}) = \min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\} = 0$, which implies when $\mathbf{X}_{\mathcal{L}}^{(k)}$ does not depend on $\mathbf{X}_{\mathcal{S}}^{(k)}$, our mechanisms achieve $P_e^{(k)} = 0$, and matches the lower bound.

(2) When $\mathcal{L} \in \mathcal{S}$: $LB(P_e^{(k)}) = \min\{P_{e,1}^{(k)}, P_{e,2}^{(k)}\} = 0$, which is a special case of $I(A^{(k)}; \mathbf{X}_{\mathcal{S}}^{(k)}) = h(A^{(k)})$.

In Section IV, we compare the lower bound with the error of the proposed mechanisms via simulations.

3) Utility of Private Aggregated Query: Next, we show that even though the proposed mechanisms are designed to minimize P_e, under the i.i.d setting (users are independent), using the summation of P_e as a measure of EAE does not lose any optimality.

Proposition 1: For the proposed mechanisms, if the sequences across users are i.i.d, then the EAE is given as follows:

$$E[|Y - A|] = \sum_{k=1}^{K} \min\left[P_{e,1}^{(k)}, P_{e,2}^{(k)}\right].$$

The proof is provided in the Appendix.

B. Centralized Mechanism

In the following, we present the privacy-preserving mechanism in the centralized setting. In the centralized setting, the mechanism takes the whole dataset as input and then aggregates the true answer to some query $\mathcal{Q} = \{\mathcal{L}, v_{\mathcal{L}}\}$. The aggregated result is denoted by A. Finally, the mechanism perturbs A and releases a privatized version Y as the output. We denote $\mathbf{X}_{\mathcal{S}} = \{\mathbf{X}_{\mathcal{S}}^{(k)}\}_{k=1}^{(K)}$ as the sensitive data matrix, which includes each user's genomic sequence at sensitive locations.

Problem formulation: The utility measurement in the centralized model is identical to that of (3). For the privacy constraints of the central model, the mechanism should provide the answer Y, which is independent of $\mathbf{X}_{\mathcal{S}}$. It can be easily shown that, Y is independent of each $\mathbf{X}_{\mathcal{S}}^{(k)}$ for $k \in \{1, .., K\}$. As such, the problem in the central setting can be represented as:

$$\min E[|A - Y|] \quad \text{s.t. } P_{Y|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}}) = P_Y(y). \tag{11}$$

The privacy-preserving mechanism in the central setting, μ_3 can be specified as follows:

$$\mu_3(D, \mathcal{Q}) \triangleq P_{Y|A,\mathbf{X}_{\mathcal{S}}}(y|a, \mathbf{x}_{\mathcal{S}}) = \begin{cases} P_A(y) + (1 - P_A(y))\dfrac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})}, & \text{if } y = a, \\ P_A(y)\left[1 - \dfrac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})}\right], & \text{if } y \neq a. \end{cases} \tag{12}$$

Theorem 4: The mechanism provided in (12) achieves perfect privacy.

Detailed proof is provided in Appendix. More related discussion to perfect privacy can be found in [27]. It is worth noting that in the central model, the term

$$R = \frac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})} \tag{13}$$

can be used implicitly to measure the dependence in the dataset. i.e., when the dependence among data is strong, there could be some w, such that $\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w) = 0$, and thus $R = 0$. On the other hand, when data are independent of each other $\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w) = P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}}) = P_A(a)$, and hence $R = 1$. Therefore, $R \in [0, 1]$ implicitly measures the dependence among the dataset. Consider the following three different cases regarding the mechanism in (12):

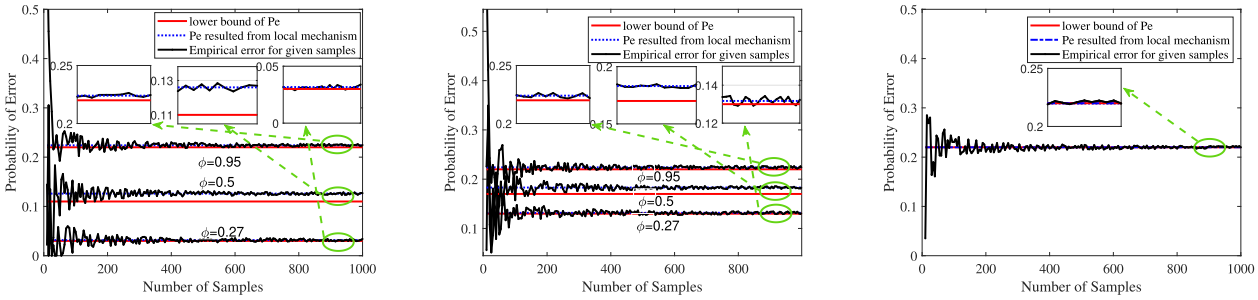
- When $R = 1$, data is independent of each other, the mechanism releases $y = a$ with a probability of 1. i.e., it will always release the true aggregate.
- When $R = 0$, data is highly dependent on each other, the mechanism releases $y = a$ with a probability of $P_A(a)$, and the probability of releasing any other $y \neq a$ is also $P_A(a)$, which means, the mechanism release an answer by sampling from the distribution of A.
- When $R \in (0, 1)$, the mechanism releases $y = a$ with probability that is larger than the prior of $P_A(y)$. Note that the larger the probability is, the higher utility the mechanism achieves, and this accuracy depends on the data dependence.

Theorem 5: The mechanism in (12) incurs the following expected error.

$$\sum_{a=0}^{N}\sum_{y\neq a} |y - a| P_A(y)\left[P_A(a) - \min_w P_{A|X_{\mathcal{S}}}(a|w)\right]. \tag{14}$$

Detailed proof is provided in the Appendix.

Illustrative Example Consider a dataset holding 3 users' genomic data with each data uniformly distributed. Users' data are independent of each other. $\mathcal{L} = \{2\}$, $\mathcal{S} = \{1\}$ and $v = T$. Next, we show how the μ_{central} works.



(a) Length of intersected data is 0 (b) Length of intersected data is 4 (c) Length of intersected data is 8

Fig. 3. Comparison between P_e and lower bound under different lengths of intersected data (correlation is measured by ϕ).

$P_{X_{\mathcal{L}}}(v_{\mathcal{L}}) = P_{X_2}(T) = 1/4$. Then $P_A(a)$ satisfies binomial distribution where

$$P_A(a) = \binom{3}{a}(1/4)^a(3/4)^{(n-a)}. \tag{15}$$

The correlation term can be expressed as:

$$\begin{aligned}
&P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}}) \\
&= P_{A|\{\mathbf{X}_1^{(k)}\}_{k=1}^3}(a|\{\mathbf{x}_1^{(k)}\}_{k=1}^3) \\
&= \sum_{\{x_2^{(k)}\}_{k=1}^3} P_{A|\{X_2^k\}_{k=1}^3}(a|\{x_2^{(k)}\}_{k=1}^3) \\
&\cdot P_{\{X_2^k\}_{k=1}^3|\{X_1^{(k)}\}_{k=1}^3}(\{x_2^k\}_{k=1}^3|\{x_1^k\}_{k=1}^3) \\
&= \sum_{\{x_2^{(k)}\}_{k=1}^3} P_{A|\{X_2^k\}_{k=1}^3}(a|\{x_2^{(k)}\}_{k=1}^3)\prod_{k=1}^3 P_{X_2^k|X_1^k}(x_2^k|x_1^k),
\end{aligned} \tag{16}$$

where

$$P_{A|\{X_2^{(k)}\}_{k=1}^3}(a|\{x_2^{(k)}\}_{k=1}^3) = \begin{cases} 1, & \text{if } \sum_{k=1}^3 \mathbb{1}_{\{x_2^{(k)}=T\}} = a, \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

Then (16) can be writen as:

$$\sum_{\{x_2^{(k)}\}_{k=1}^3 \in \mathcal{B}} \prod_{k=1}^3 P_{X_2^{(k)}|X_1^{(k)}}(x_2^{(k)}|x_1^{(k)}), \tag{18}$$

where $\mathcal{B} = \{\{x_2^{(k)}\}_{k=1}^3 : \sum_{k=1}^3 \mathbb{1}_{\{x_2^{(k)}=T\}} = a\}$. Under the i.i.d model, the EAE of μ_{central} is shown in the following corollary.

Corollary 2: Under the setting where each user's genomic data sequence is i.i.d, and the distribution of each genomic data is uniform, Let $\lambda = \left(\frac{1}{C_x}\right)^{|\mathcal{L}|}$, *then, the EAE for the central mechanism can be expressed as: When* $\mathcal{E} \leq 0.5$, $EAE = 0$, *when* $\mathcal{E} > 0.5$, $EAE = 2\{\sum_{a=0}^N a P_A(a) F_A(a) - N\lambda^2\}$.

V. Numerical Evaluation

We next numerically evaluate the proposed mechanisms for two cases: for the first case, we consider a first-order Markov property in each of the genomic sequences; for the second case, we generate synthetic data with a hidden Markov model (described later in sub-section V-E).

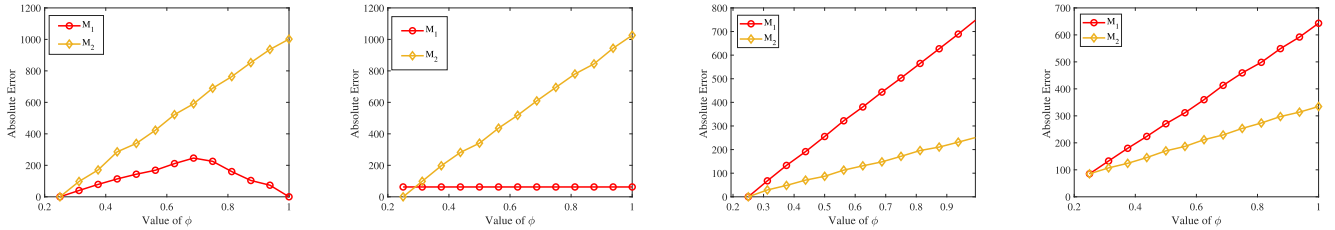
We next summarize the Markov setting: It is assumed that each user's genomic data has a first-order Markov property [15], i.e., $P_{X_{j+1}|X_j,\mathbf{X}_1^{j-1}}(x_{j+1}|x_j,\mathbf{x}_1^{j-1}) = P_{X_{j+1}|X_j}(x_{j+1}|x_j)$. Denote $\mathcal{T}$ as the transition matrix from time j to $j-1$, for all $j \in [1, |X|]$, and can be specified as:

$$\begin{aligned}
&\Pr(X_{j+1} = x|X_j = x) = \phi \\
&\Pr(X_{j+1} = x'|X_j = x) = (1-\phi)/3
\end{aligned} \tag{19}$$

where $x, x' \in \{A, T, G, C\}$ and $x \neq x'$. For the following evaluations, we randomly generate the user's data according to the prior and the transition matrix, and we run Monte-Carlo simulations to get an average error. We consider $K = 1000$ users in the system. Each user possesses a genomic sequence with a length of 10.

A. Comparison of $LB(P_e)$ With P_e From the Local Mechanism

We next compare the P_e resulting from the local mechanism with the lower bound in Theorem 3. We consider three different cases: (i) when $\mathcal{S} \cap \mathcal{L} = \emptyset$, (ii) when $\mathcal{S} \cap \mathcal{L} = 4$ (partially overlapped) and (iii) when $\mathcal{S} \cap \mathcal{L} = \mathcal{S}$ (fully overlapped). Under each case, we consider three types of correlation: low correlation $\phi = 0.27$ (slightly skewed than uniform), moderate correlation $\phi = 0.5$, and high correlation $\phi = 0.95$ (slightly weaker than fully dependent). The comparisons are shown in Fig. 3, where we plotted the lower bound of P_e from Theorem 3, P_e calculated from Theorem 2 as well as the empirical P_e averaged from N samples, and we vary N from 10 to 1000 to illustrate the convergence. Observe that the gap between P_e from Theorem 2 and the lower bound of P_e is different under different scenarios. Generally, extremely high or low correlation leads to smaller gaps. It is worth noting that the lower bound derived in this paper is based on Fano's inequality. There could be some P_e that are not achievable, thus the gap between the P_e from Theorem 2 and the optimal one (assume there is a mechanism that incurs a minimum P_e and satisfies perfect privacy) can be even smaller.



(a) $\mathcal{S} \cap \mathcal{L} = \emptyset$, P_{X_1} is uniform (b) When $\mathcal{S} \cap \mathcal{L} \neq \emptyset$, each X_1 is uniformly generated (c) $\mathcal{S} \cap \mathcal{L} = \emptyset$, $P_{X_1}(A) = 0.8$ (d) $\mathcal{S} \cap \mathcal{L} \neq \emptyset$, $P_{X_1}(A) = 0.8$

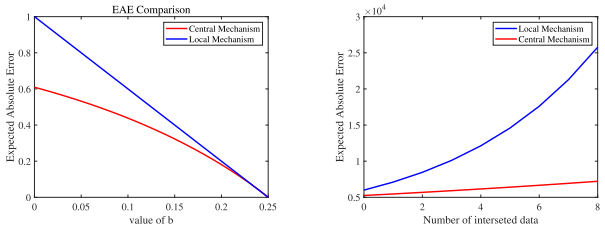
Fig. 4. Numerical results of the Markov model for different cases regarding the intersected length of $\mathcal{L} \cap \mathcal{S}$ and the distribution of X_1 for the local case.

B. Comparison Between M_1 and M_2

Next, we examine the impact from (a) different prior distributions; (b) whether $\mathcal{S} \cap \mathcal{L}$ is an empty set or not. We fix $\mathcal{S} = \{3, 4\}$ and evaluate with parameters according to the following cases: for (a), we assume X_1 is uniformly distributed or $P_{X_1} = 0.8$, and then generate the whole sequence based on the value of X_1 and $\mathcal{T}$. (b), we consider either $\mathcal{L} = \{4, 5\}$ or $\mathcal{L} = \{5, 6\}$. We compare the absolute error resulted from each mechanism under each case.

The results are shown in Fig. 4(a). Observe from Fig. 4(a), all mechanisms achieve zero-EAE when $\phi = 0.25$. When $\phi \neq 0.25$, $\mathcal{M}_1$ outperforms $\mathcal{M}_2$, the reason is that, to guarantee perfect privacy, $\mathcal{M}_1$ tends to release most of the local answers as 0 while $\mathcal{M}_2$ releases most answers as 1. Since $P_{X_{\mathcal{L}}^{(k)}}(v_{\mathcal{L}})$ happens with probability less than 1/2, releasing more 0s is more accurate than releasing more 1s. Also, observe that the EAE of $\mathcal{M}_1$ first increases then decreases to 0 with ϕ. Intuitively, the value of $R(\mathbf{X}, \mathcal{Q})$ decreases with ϕ, and larger $R(\mathbf{X}, \mathcal{Q})$ implies larger probability to directly release the answer. That is why EAE is smaller when ϕ is small. For large ϕ, the dependence of data increases, and each user's genomic sequence can hardly match $v_{\mathcal{L}}$. As a result, the real local answer for each user is 0. Thus, $\mathcal{M}_1$ which releases more 0s achieves 0-EAE when $\phi = 1$. For case 2, from Fig. 4(b), the EAE for $\mathcal{M}_1$ is a constant, this is because when data is uniformly distributed, $\mathcal{E} = \frac{1}{4}$, which is smaller than 0.5. As a result, $\mathcal{M}_1$ releases 0 all the time, hence incurring a constant error for different ϕ. It is worth noting that each case in Fig. 4(a) provides lower P_e than those in Fig. 4(b). The reason is that the prior of case 2 is more skewed and requires less perturbation.

Then, we examine how data prior affects the performance of $\mathcal{M}_1$ and $\mathcal{M}_2$. We let $P_{X_1}(A) = 0.8$, and let $v_{\mathcal{L}} = \{A, A\}$, which means, each local answer $A^{(k)}$ is more likely to be 1 than 0, and as ϕ increases the probability of $P_{A^{(k)}}(1)$ increases. The comparison of the two mechanisms' performance is shown in Fig. 4(c) and Fig. 4(d). We can observe that under case 1 and case 2, both mechanisms' EAEs increase as ϕ increases since larger ϕ implies smaller $R(\mathbf{X}, \mathcal{Q})$. On the other hand, $\mathcal{M}_2$ performs better than $\mathcal{M}_1$, because each true local answer is more likely to be 1, and $\mathcal{M}_2$ tends to release 1 while $\mathcal{M}_1$ tends to release 0 under high data correlations. It is worth noting that when $\phi = 0.25$, the data is i.i.d. For case 1, each mechanism releases the true answer and achieves 0-EAE; for case 2, $\mathcal{M}_1$ and $\mathcal{M}_2$ release 1 when $X_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}$, since $\mathcal{E} \leq 0.5$.



(a) (b)

Fig. 5. EAE comparison between the central and local mechanisms (a) with different correlation parameters (b) with different length of intersected data.

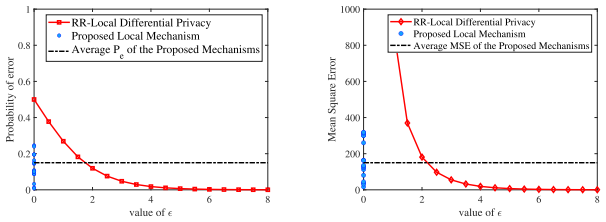
As a result, an error is incurred when $X_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}$ (each local $A^{(k)} = 1$ when $X_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}$ and $X_{\mathcal{L}\cap\mathcal{S}} = v_{\mathcal{L}\cap\mathcal{S}}$).

C. Comparison Between Centralized and Local Mechanism

We next compare the EAEs incurred by the local and centralized mechanisms, consider K independent users who hold genomic data, denote $\mathbf{X}^{(k)}$ as the i-th user's genomic data sequence with a length of N, i.e., $\mathbf{X}^{(k)} = \{X_1^{(k)}, X_2^{(k)}, \ldots, X_N^{(k)}\}$. It is assumed that the prior of each $X_1^{(k)}$ is uniformly distributed, and the correlation of between the data from one to another follows Markov property, which can be summarized as follows: $P_{X_{k+1}^{(k)}|X_k^{(k)}}(\delta|\gamma) = b$, and $P_{X_{k+1}^{(k)}|X_k^{(k)}}(\gamma|\gamma) = a$, where k denotes the index of different genomic data, $a \geq b$ and δ, γ are possible realizations from the support of the genomic data: $\{A, T, G, C\}$. Therefore: $a + 3b = 1$. Under the above setting, we first assume $\mathcal{S} = 1$, 2 and $\mathcal{L} = 3, 4$, $v = \{A, T\}$. That is, $A^{(k)} = 1$ iff $X_{3,4}^{(k)} = \{A, T\}$. We next show how different mechanisms perform under this setting.

We vary the value of b from 0 (strong dependence) to 0.25 (independent), and observe the EAEs resulted from different mechanisms. Observe that when $b = 0.25$, which means data within the genomic sequence are independent, both mechanisms achieve zero expected absolute error; when $b = 0$, referring to the strongest dependence scenario, the central mechanism always results in smaller EAE than the local mechanism.

We then fix the correlation to be $b = 0.5$, and then vary the number of intersected data length from 0 to 8, i.e., $\mathcal{L} = \{3, 4, 5, 6, 7, 8, 9, 10\}$, and increase the length of $\mathcal{S}$, from $\{1, 2\}$ to $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We then compare the resulted expected absolute error, and the result is shown in Fig. 5. Observe that, with the increase of the number of



(a) ϵ $v.s.$ P_e (per-user case) (b) ϵ $v.s.$ MSE (aggregated case)

Fig. 6. Comparison between the privacy guarantee by $\mathcal{M}_1$ or $\mathcal{M}_2$ and DP.

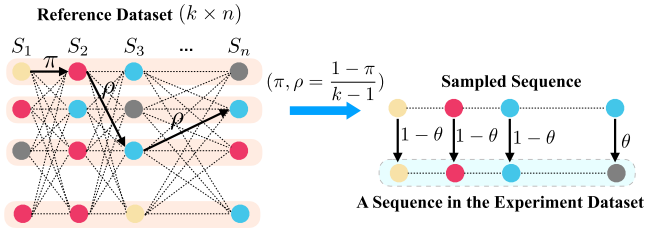


Fig. 7. Generating experiment dataset from reference dataset with Hidden Markov Model (hmm).

intersected data lengths, both mechanisms result in a larger expected error. However, the error of the local model increases much faster than the central model.

D. Comparison With Differential Privacy-Based Mechanisms

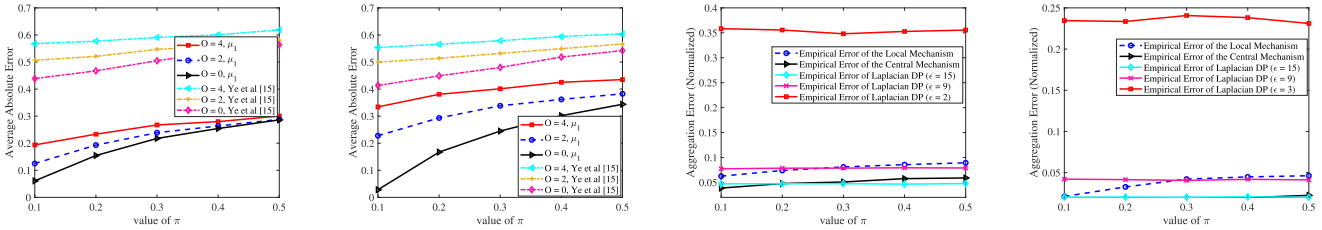
In this part, we compare the performance of the proposed mechanism with (Local) Differential Privacy [28]. Since our mechanisms guarantee perfect privacy, equivalently, under the privacy notion of DP, $\epsilon = 0$. Then, in the following, we show that under different P_e or EAE, the minimum value of ϵ (a smaller ϵ indicates a stronger privacy guarantee) provided by different mechanisms. Under the local setting, we use a binary randomized response perturbation mechanism that satisfies LDP. In [28], an optimal mechanism is derived under LDP constraints: $\Pr(Y = A) = \frac{e^\epsilon}{1+e^\epsilon}$, and $\Pr(Y \neq A) = \frac{1}{e^\epsilon+1}$, wherein the binary case, the latter stands for the P_e. On the other hand, in [29], the minimum MSE resulted by the LDP-based mechanism for aggregated count query is given by: $N\frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2}$. We next present the comparison under two cases: the comparison of the P_e and the comparison of the MSE. The results are shown in Fig. 6. Observe that under a larger P_e or MSE, Differential Privacy can provide a better privacy guarantee with a small ϵ. However, our proposed mechanisms always provide perfect privacy.

E. Evaluation With Hidden Markov Model

Since the real-world genomic dataset contains private information and is infeasible to access. We next evaluate with the Hidden Markov Model (HMM), which is widely adopted in genetics for a wide range of tasks that require a probabilistic model of the genome [15], [19], [30]. In the HMM model, the experimental data (set) is generated from a reference dataset with certain parameters (π, θ). In this experiment, we consider the reference dataset contains $K = 100$ users genomic sequence, each with a length of $N = 20$. We treat each genotype at a certain location as a state. When generating the experiment data sequence, we randomly pick one data sequence from the reference dataset to begin with. For each of the following genotypes, we make it identical to the state of the same sequence with a probability of π, and switch to another state from another sequence with a probability of $(1-\pi)/(N-1)$. It is worth noting that the switching probability of π can be used to measure the dependence in the genomic sequence. A larger value of π indicates that genomic sequences in the experimental dataset would preserve similar patterns to the reference dataset. Thus the dependence is strong. On the other hand, for a small π, each genotype in the experiment dataset contains large randomness, and hence the dependence is weak. Besides the switching probability, we add randomness to the experimental dataset with an error probability θ. That said, each time HMM model copies a genotype from the reference dataset, there is a probability of θ to substitute by randomly sampling one from $\{A, T, G, C\}$. Finally, we compare it with the mechanism proposed in [15], which is the most relative mechanism in the literature. Note that the mechanism in [15] hides genotypes at certain locations, which makes the query aggregation not directly applicable. To this end, we consider two types of post-processing strategies: one is to sample genotypes at hidden locations from a uniform distribution, and the other is to sample from the prior distribution calculated from the frequency. Figure 7 depicts the process that an HMM model generates an experimental dataset from the reference dataset.

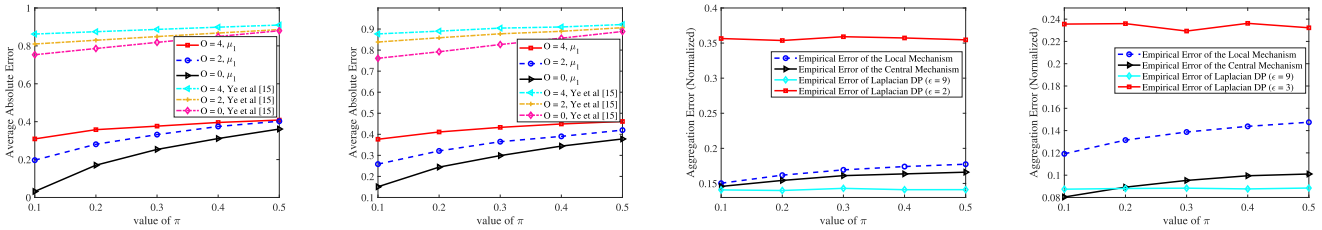
To generate the experimental dataset, we consider two sets of reference dataset, the first reference dataset is sampled from a uniform distribution. The second reference dataset is a sample from Sample GenBank Record [31]. We generate $K = 1000$ users' genomic sequence as the experimental data with π and θ. In the following experiments, we vary the value of π from 0 to 0.5, and fix the value of θ to be 0.01 and 0.05, respectively. Then, we calculate the frequency of the appearance of different combinations of genotypes as the prior distribution and the conditional probabilities. Specifically, to calculate the distribution of A in the central setting, we find the ratio of $\bar{A}/K$, where $\bar{A}$ denotes the true aggregate value, this value converges to $P_{A^{(k)}}(1)$ for each individual (for large K). Under the independent user assumption (each user in the experimental dataset is generated by HMM independently), A is binomially distributed. Hence, its prior distribution can be calculated accordingly by each $P_{A^{(k)}}$. From there, we implement our local and central mechanisms accordingly for $T = 1000$ times and calculate the average values.

We first compare the local models with different length of $|\mathcal{L} \cap \mathcal{S}|$, the results are shown in Fig.8(a) and Fig.8(b) respectively (Fig.9(a) and Fig.9(b) for real-world sample). Observer that, as π increases, the P_e of each case increases accordingly. That is because the dependence increases with π, and P_e increases with the strength of the dependence. Another observation is that P_e also increases with the length of intersected locations. The reason is the dependence between the queried genotypes and the sensitive genotypes increases



(a) AAE comparison when $\theta = 0.01$ (b) AAE comparison when $\theta = 0.05$ (c) EAE comparison when $\theta = 0.01$ (d) EAE comparison when $\theta = 0.05$

Fig. 8. Numerical results of the Hidden Markov Model generated from a synthetic reference dataset, (a), (b) for individual average absolute error (AAE) under different cases of the overlap $\mathcal{L} \cap \mathcal{S}$ (denoted as O), We sample from the prior for the genotypes at the hidden locations for Ye's mechanism in [15]; (c), (d) for aggregate error, which compares empirical error in aggregation when deploying local and central mechanisms.



(a) AAE comparison when $\theta = 0.01$ (b) AAE comparison when $\theta = 0.05$ (c) EAE comparison when $\theta = 0.01$ (d) EAE comparison when $\theta = 0.05$

Fig. 9. Numerical results of the Hidden Markov Model generated from real-world genomic data, (a), (b) for individual average absolute error (AAE) under different cases of $\mathcal{L} \cap \mathcal{S}$(denoted as O), We sample from the uniform distribution for the genotypes at the hidden locations for Ye's mechanism in [15]; (c), (d) for aggregate error, which compares empirical error in aggregation when deploying local and central mechanisms.

with the length of intersected locations. It is worth noting that for larger θ, the randomness in the dataset increases, which results in smaller dependence among the genomic sequence, and hence incurs a smaller P_e, which explains why subcase (b) always incurs smaller errors than subcase (a).

We then compare the local model and the central model according to different θ and π. In addition, we compared our models with those based on DP with Laplace mechanisms. For each model, we modify the value of the privacy budget ϵ, and compute the empirical absolute error by implementing ϵ-DP. Then we find those mechanisms that incur similar aggregation errors to our mechanisms. The corresponding values of ϵ and the comparison results are shown in Fig. 8(c) and Fig. 8(d), respectively (Fig.8(c) and Fig.9(d) for real-world sample). Observe that the central model always provides smaller empirical errors compared to the local models. On the other hand, to achieve comparable performance to our mechanism, the DP-based mechanism causes too much privacy leakage (ϵ is greater than 4). It is worth noting that, the ϵ values yielding comparable accuracy to our local/central mechanisms were found to be 15 and 9 for synthetic and real-world data, respectively. That is to say, DP mechanisms need different amounts of privacy budget to achieve comparable utility (in terms of error probability) to our mechanisms for synthetic and real-world data respectively. The primary reason for this difference is rooted in the inherent characteristics of real-world data sequences, where each local $A^{(k)}$ is more likely to be 0. In view of privacy preservation, our mechanisms may introduce some noise by perturbing certain 0s to 1s. Consequently, this perturbation leads to an increase in the accumulated error. Another observation is when choosing DP with a reasonable budget for privacy, such as 2 or 3, the accuracy gaps with our mechanisms are significant.

VI. Conclusion

The problem of privacy-preserving counting query-answering mechanisms for genotype aggregation is studied in this paper. We propose information theoretical privacy-preserving mechanisms based on both the local and the central settings. The proposed mechanisms guarantee perfect privacy for genome data at sensitive locations. We then derive a lower bound of P_e for an arbitrary mechanism under perfect privacy and show the optimality of our mechanisms under some special cases. Finally, we simulate with both synthetic data and real-world data sample combination with Hidden Markov Model to show the performance of our proposed mechanisms under different scenarios. The results also contain a comparison with Differential Privacy-based mechanisms, which shows that to provide comparable query accuracy, DP-based mechanisms incur much larger privacy leakages.

Appendix A
Proof of Theorem 2

In this Section, we provide some more intuition behind the derivation of the mechanisms and then derive the error probability for the mechanisms. Since the proof for Theorem 2 is for local mechanism only, we remove the superscript (k) for simplicity.

A. Deriving the Mechanisms and Error Probability

We first expand the term $P_{Y|\mathbf{X}_{\mathcal{S}}}(y|x_{\mathcal{S}})$ as follows

$$P_{Y|\mathbf{X}_{\mathcal{S}}}(y|x_{\mathcal{S}}) = \sum_{x_{\bar{\mathcal{L}}}} P_{Y|\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(y|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}). \tag{20}$$

In order to satisfy the per-user perfect privacy condition, we require that $P_{Y|\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(y|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}})\Pr(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}) = f(y, x_{\bar{\mathcal{L}}})$, $\forall x_{\bar{\mathcal{L}}}, y$, i.e., each term in (20) does not depend on $x_{\mathcal{S}}$. To this end, we have

$$P_{Y|\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(y|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) = \frac{f(y, x_{\bar{\mathcal{L}}})}{P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}})}, y \in \{0, 1\}. \tag{21}$$

From the above equation, we make $P_{Y|\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(y|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}})$ a valid probability mass function, by picking $f(y, x_{\bar{\mathcal{L}}})$ as follows

$$\begin{aligned} &0 \le f(y, x_{\bar{\mathcal{L}}}) \le P_{\mathbf{X}_{\bar{\mathcal{L}}}|X_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}}), \forall x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}, \\ &\Rightarrow f(y, x_{\bar{\mathcal{L}}}) \le \min_{w} P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}} = w). \end{aligned} \tag{22}$$

B. Error Probability for Mechanism $\mathcal{M}_1$

Recall that our goal is to minimize the probability of error per user. We first expand the first term as follows:

$$\begin{aligned} P_{Y,\mathbf{X}_{\mathcal{L}}}(0, v_{\mathcal{L}}) &= \sum_{x_{\bar{\mathcal{S}}}} P_{Y|\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(0|v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= \sum_{x_{\bar{\mathcal{S}}}} P_{Y|X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(0|v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= \sum_{x_{\bar{\mathcal{S}}}} (1 - P_{Y|\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(1|x_{\mathcal{L}}, x_{\bar{\mathcal{S}}})) P_{\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{S}}}} P_{Y|\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(1|v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &\overset{(a)}{=} P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{S}}}} P_{Y|\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(1|v_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{\mathbf{X}_{\bar{\mathcal{L}}},\mathbf{X}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) \\ &\overset{(b)}{=} P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{S}}}} f(1, v_{\bar{\mathcal{L}}}) P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}). \\ &= P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - f(1, v_{\bar{\mathcal{L}}}) P_{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}}(v_{\mathcal{L}\cap\mathcal{S}}) \end{aligned} \tag{23}$$

where step (a) follows from the fact that $\mathcal{L}\cap\bar{\mathcal{S}} = \bar{\mathcal{L}}\cap\mathcal{S}$, while step (b) follows from (21).

Similarly, we have the following set of steps for the second term as follows:

$$\begin{aligned} &\sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} P_{Y,X_{\mathcal{L}}}(1, x_{\mathcal{L}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} \sum_{x_{\bar{\mathcal{S}}}} P_{Y|X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(1|x_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(x_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} \sum_{x_{\bar{\mathcal{S}}}} P_{Y|X_{\bar{\mathcal{L}}},X_{\mathcal{S}}}(1|x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{X_{\bar{\mathcal{L}}},X_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) \\ &\overset{(b)}{=} \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} f(1, x_{\bar{\mathcal{L}}}) \sum_{x_{\bar{\mathcal{S}}}} P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}), \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}: x_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}} f(1, v_{\bar{\mathcal{L}}}) \sum_{x_{\bar{\mathcal{S}}}} P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}) \\ &\quad + \sum_{x_{\mathcal{L}}\neq v: x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) \sum_{x_{\bar{\mathcal{S}}}} P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}) \\ &\overset{(c)}{=} f(1, v_{\bar{\mathcal{L}}}) \Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) + \sum_{x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}), \end{aligned} \tag{24}$$

where step (a) follows from (21), while in step (b), the function $f(0, v_{\bar{\mathcal{L}}})$ does not depend on $x_{\mathcal{S}}$. Step (c) follows from the following:

$$\begin{aligned} &\sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}: x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) \sum_{x_{\bar{\mathcal{S}}}} P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}: x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) \sum_{x_{\bar{\mathcal{S}}}} P_{\mathbf{X}_{\bar{\mathcal{S}}},\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}}(x_{\bar{\mathcal{S}}}, x_{\mathcal{L}\cap\mathcal{S}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}: x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) P_{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}}(x_{\mathcal{L}\cap\mathcal{S}}) \\ &= \sum_{x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) \sum_{x_{\mathcal{L}\cap\mathcal{S}}} P_{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}}(x_{\mathcal{L}\cap\mathcal{S}}) = \sum_{x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}). \end{aligned} \tag{25}$$

C. Error Probability for Mechanism $\mathcal{M}_2$

We write the error probabilities in terms of $f(0, x_{\bar{\mathcal{L}}})$ and following similar steps as in $\mathcal{M}_1$. For the first term, we have the following set of steps:

$$\begin{aligned} P_{Y,X_{\mathcal{L}}}(0, v_{\mathcal{L}}) &= \sum_{x_{\bar{\mathcal{S}}}} P_{Y|X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(0|v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{X_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(v_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= \sum_{x_{\bar{\mathcal{S}}}} P_{Y|X_{\bar{\mathcal{L}}},X_{\mathcal{S}}}(0|v_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) P_{X_{\bar{\mathcal{L}}},X_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}, x_{\mathcal{S}}) \\ &= \sum_{x_{\bar{\mathcal{S}}}} f(0, v_{\bar{\mathcal{L}}}) P_{X_{\mathcal{S}}}(x_{\mathcal{S}}) \\ &= f(0, v_{\bar{\mathcal{L}}}) P_{X_{\mathcal{L}\cap\mathcal{S}}}(v_{\mathcal{L}\cap\mathcal{S}}). \end{aligned}$$

For the second term, we have the following set of steps:

$$\begin{aligned} &\sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} P_{Y,\mathbf{X}_{\mathcal{L}}}(1, x_{\mathcal{L}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} \sum_{x_{\bar{\mathcal{S}}}} P_{Y|\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(1|x_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) P_{\mathbf{X}_{\mathcal{L}},X_{\bar{\mathcal{S}}}}(x_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} \sum_{x_{\bar{\mathcal{S}}}} (1 - P_{Y|\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(0|x_{\mathcal{L}}, x_{\bar{\mathcal{S}}})) P_{\mathbf{X}_{\mathcal{L}},\mathbf{X}_{\bar{\mathcal{S}}}}(x_{\mathcal{L}}, x_{\bar{\mathcal{S}}}) \\ &= 1 - P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\mathcal{L}}\neq v_{\mathcal{L}}} \sum_{x_{\bar{\mathcal{S}}}} f(0, x_{\bar{\mathcal{L}}}) P_{\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{S}}) \\ &= 1 - P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) \\ &\quad - f(0, v_{\bar{\mathcal{L}}}) \Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) - \sum_{x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(0, x_{\bar{\mathcal{L}}}). \end{aligned}$$

We optimize the per-user error probability for the two release mechanisms for fixed q, s and $v_{\mathcal{L}}$ as follows.

For $\mathcal{M}_1$: from (23) and (24), the per-user error probability is

$$\begin{aligned} P_{e,1} = P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) &+ \sum_{x_{\bar{\mathcal{L}}}\neq v_{\bar{\mathcal{L}}}} f(1, x_{\bar{\mathcal{L}}}) \\ &+ f(1, v_{\bar{\mathcal{L}}})(2\Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) - 1). \end{aligned} \tag{26}$$

We minimize the probability of error as follows. Here, we have two cases depending on the value of $\Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}})$.

If $\Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) \geq 1/2$, we pick $f(1, x_{\bar{\mathcal{L}}}) = 0, \forall x_{\bar{\mathcal{L}}}$. When $\Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) < 1/2$, we pick $f(1, v_{\bar{\mathcal{L}}}) = \min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|x_{\mathcal{S}} = w)$ and $f(1, x_{\bar{\mathcal{L}}}) = 0, \forall x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}$.

For $\mathcal{M}_2$ The per-user error probability for $\mathcal{M}_2$ is

$$P_{e,2} = 1 - P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} f(0, x_{\bar{\mathcal{L}}}) - f(0, v_{\bar{\mathcal{L}}})(2\Pr(\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) - 1), \quad (27)$$

where the error probability is minimized when $f(0, x_{\bar{\mathcal{L}}}) = \min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}|X_{\mathcal{S}}}(x_{\bar{\mathcal{L}}}|w), \forall x_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}$. For $f(0, v_{\bar{\mathcal{L}}})$, we have two special cases: when $\Pr(X_{\mathcal{L}\cap\mathcal{S}} \neq v_{\mathcal{L}\cap\mathcal{S}}) > 1/2$, we pick $f(0, v_{\bar{\mathcal{L}}}) = \min_w P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}}|w)$, otherwise we set $f(0, v_{\bar{\mathcal{L}}}) = 0$.

This concludes the proof for Theorem 2.

Appendix B
Proof of Theorem 3

Proof: According to Fano's inequality [32], the lower bound of the error probability can be expressed as:

$$h(A|Y) \leq h(P_e) + P(e)\log(|A| - 1). \quad (28)$$

As $|A| = 2$, Eq. (28) can be expressed as:

$$h(P_e) \geq h(A|Y) = h(A) - I(A; Y). \quad (29)$$

In Eq.(29), $h(A)$ is a constant given the distribution of A (which is a deterministic function of $\mathbf{X}_{\mathcal{L}}$). Next, we focus on $I(A; Y)$. To derive a lower bound of P_e, we are looking at the upper bound of $I(A; Y)$. Note that random variable A can be represented as:

$$A = \mathbb{1}_{\{\mathbf{X}_{\mathcal{L}} = v_{\mathcal{L}}\}} = \mathbb{1}_{\{\mathbf{X}_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}\}} \times \mathbb{1}_{\{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} = v_{\mathcal{L}\cap\mathcal{S}}\}}. \quad (30)$$

Denote $A_{\bar{\mathcal{L}}} = \mathbb{1}_{\{\mathbf{X}_{\bar{\mathcal{L}}} = v_{\bar{\mathcal{L}}}\}}$ and $A_{\mathcal{L}\cap\mathcal{S}} = \mathbb{1}_{\{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}} = v_{\mathcal{L}\cap\mathcal{S}}\}}$. Therefore, using this notation, we can write $A = A_{\bar{\mathcal{L}}} \times A_{\mathcal{L}\cap\mathcal{S}}$. Then the upper bound of $I(A; Y)$ can be derived as:

$$\begin{aligned} I(A; Y) &= I(A; Y|\mathbf{X}_{\mathcal{S}}) && (31)\\ &\leq I(A; \mathbf{X}|\mathbf{X}_{\mathcal{S}}) \\ &= h(A|\mathbf{X}_{\mathcal{S}}) - h(A|\mathbf{X}, \mathbf{X}_{\mathcal{S}}) \\ &= h(A|\mathbf{X}_{\mathcal{S}}), && (32)\end{aligned}$$

where (22) follows from the privacy constraint in (1), (23) follows by data processing. From another perspective:

$$\begin{aligned} I(A; Y) &= I(A_{\bar{\mathcal{L}}} A_{\mathcal{L}\cap\mathcal{S}}; Y) \\ &\leq I(A_{\bar{\mathcal{L}}}, A_{\mathcal{L}\cap\mathcal{S}}; Y) \\ &= I(A_{\mathcal{L}\cap\mathcal{S}}; Y) + I(A_{\bar{\mathcal{L}}}; Y|A_{\mathcal{L}\cap\mathcal{S}}) \\ &= I(A_{\bar{\mathcal{L}}}; Y|A_{\mathcal{L}\cap\mathcal{S}}) && (33)\\ &\leq I(A_{\bar{\mathcal{L}}}; X|A_{\mathcal{L}\cap\mathcal{S}}) \\ &= h(A_{\bar{\mathcal{L}}}|A_{\mathcal{L}\cap\mathcal{S}}) - h(A_{\bar{\mathcal{L}}}|X, A_{\mathcal{L}\cap\mathcal{S}}) \\ &= h(A_{\bar{\mathcal{L}}}|A_{\mathcal{L}\cap\mathcal{S}}), && (34)\end{aligned}$$

where (24) follows from the privacy constraint in (1), (25) follows from data processing. From (31) and (33), we obtain an upper bound on $I(Y; A)$ as follows: $I(A; Y) \leq \min(h(A_{\bar{\mathcal{L}}}|A_{\mathcal{L}\cap\mathcal{S}}), h(A|X_{\mathcal{S}}))$ Then, substituting the above bound in (29), we arrive at the lower bound on error probability stated in Theorem 3.

$$h(P_e) \geq h(A) - \min\{h(A_{\bar{\mathcal{L}}}|A_{\mathcal{L}\cap\mathcal{S}}), h(A|X_{\mathcal{S}})\}. \quad (35)$$

□

Appendix C
Proof of Claims Made in Remark 2

1) when $\mathcal{L} \cap \mathcal{S} = \emptyset$, from Theorem 3, we have:

$$P_e \geq h^{-1}\big(I(A; \mathbf{X}_{\mathcal{S}})\big). \quad (36)$$

On the other hand, according to Theorem 2, when $\mathcal{L}\cap\mathcal{S} = \emptyset$, $\mathcal{E} = 0$, and

$$\begin{aligned} P_{e,1} &= P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \min_w P_{\mathbf{X}_{\mathcal{L}}|\mathbf{X}_{\mathcal{S}}}(v_{\mathcal{L}}|w), \\ P_{e,2} &= 1 - P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{x_{\mathcal{L}} \neq v_{\mathcal{L}}} \min_w P_{\mathbf{X}_{\mathcal{L}}|\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{L}}|w). \quad (37)\end{aligned}$$

We next consider two sub-cases regarding the relationship between A and $X_{\mathcal{S}}$:

a). When $I(A; \mathbf{X}_{\mathcal{S}}) = h(A)$, from (36):

$$\begin{aligned} P_e &\geq h^{-1}(h(A) - H(A|\mathbf{X}_{\mathcal{S}})) = h^{-1}(h(A)) \\ &= \min\{P_A(0), P_A(1)\}. \quad (38)\end{aligned}$$

Since $h(A|X_{\mathcal{S}}) = 0$, $\min_w P_{\mathbf{X}_{\mathcal{L}}|\mathbf{X}_{\mathcal{S}}}(x_{\mathcal{L}}|w) = 0$ for all $x_{\mathcal{L}}$. From (37), we have $P_{e,1} = P_A(1)$, $P_{e,2} = P_A(0)$, which implies that the minimal P_e resulted by $\mathcal{M}_1$ and $\mathcal{M}_2$ is $\min\{P_A(0), P_A(1)\}$.

b) When $I(A; \mathbf{X}_{\mathcal{S}}) = 0$, from (36), the lower bound implies $P_e \geq 0$. From (37), we have $P_{e,1} = P_{e,2} = 0$, which implies that the minimal P_e resulted by $\mathcal{M}_1$ and $\mathcal{M}_2$ is 0.

The above discussion shows that under the two extreme cases, the proposed mechanism achieves P_e matches the lower bound.

2) When $\mathcal{L} \in \mathcal{S}$, $H(A|\mathbf{X}_{\mathcal{S}}) = H(A_{\bar{\mathcal{L}}}|A_{\mathcal{L}\cap\mathcal{S}}) = 0$. The result follows sub-case a), which matches the lower bound from Theorem 3. The intuition is when all queried locations are sensitive, both mechanisms sample random answers according to the data prior to keep the queried sequence perfectly private.

Appendix D
Proof of Proposition 1

We next show the lower bound of the EAE under the proposed mechanisms, the goal is to show that when users are i.i.d, the bounds are tight (i.e. the inequality can be replaced by equality).

In the following, we let Y denote the aggregated results from each local answer $Y^{(k)}$: $Y = \sum_{k=1}^{K} Y^{(k)}$, and A denote the real answer from each user: $A = \sum_{k=1}^{K} A^{(k)}$. By Jensen's Inequality:

$$\begin{aligned} E|Y - A| &\geq |E(Y - A)| \\ &= \left| E\sum_{k=1}^{K}(Y^{(k)} - A^{(k)}) \right| \\ &= \left| \sum_{k=1}^{K} E(Y^{(k)} - A^{(k)}) \right| \end{aligned}$$

$$= \left| \sum_{k=1}^{K} \left[P_{Y^{(k)},A^{(k)}}(1,0) - P_{Y^{(k)},A^{(k)}}(0,1) \right] \right|$$
$$= \left| \sum_{k=1}^{K} P_{Y^{(k)}|A^{(k)}}(1|0) P_{A^k}(0) - \sum_{k=1}^{K} P_{Y^{(k)}|A^{(k)}}(0|1) P_{A^{(k)}}(1) \right|$$
$$= \Big| \sum_{k,x_{\bar{\mathcal{S}}}} P(Y^{(k)} = 1 | \mathbf{X}_{\mathcal{L}} \neq v_{\mathcal{L}}, \mathbf{X}_{\bar{\mathcal{S}}} = x_{\bar{\mathcal{S}}})$$
$$P(\mathbf{X}_{\mathcal{S}} = x_{\mathcal{S}} | \mathbf{X}_{\mathcal{L}} \neq v_{\mathcal{L}}) P(\mathbf{X}_{\mathcal{L}} \neq v_{\mathcal{L}})$$
$$- \sum_{k,x_{\bar{\mathcal{S}}}} P(Y^{(k)} = 0 | \mathbf{X}_{\mathcal{L}} = V_{\mathcal{L}}, \mathbf{X}_{\bar{\mathcal{S}}} = x_{\bar{\mathcal{S}}})$$
$$P(\mathbf{X}_{\mathcal{S}} = x_{\mathcal{S}} | \mathbf{X}_{\mathcal{L}} = v_{\mathcal{L}}) P(\mathbf{X}_{\mathcal{L}} = v_{\mathcal{L}}) \Big|, \tag{39}$$

where $P(Y^{(k)} = 0 | \mathbf{X}_{\mathcal{L}} = v_{\mathcal{L}}, \mathbf{X}_{\bar{\mathcal{S}}} = x_{\bar{\mathcal{S}}})$ and $P(Y^{(k)} = 1 | \mathbf{X}_{\mathcal{L}} \neq v_{\mathcal{L}}, \mathbf{X}_{\bar{\mathcal{S}}} = x_{\bar{\mathcal{S}}})$ are parameters of the mechanisms. If users are i.i.d, substituting μ_1 we have:

$$|E(Y - A)| = K \left| \left[P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \min_{w} P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(v_{\bar{\mathcal{L}}|w}) P_{\mathbf{X}_{\mathcal{L}\cap\mathcal{S}}}(v_{\mathcal{L}\cap\mathcal{S}}) \right] \right|. \tag{40}$$

Similarly, substituting μ_2, we have:

$$|E(Y - A)| K \left| \left[1 - P_{\mathbf{X}_{\mathcal{L}}}(v_{\mathcal{L}}) - \sum_{u_{\bar{\mathcal{L}}} \neq v_{\bar{\mathcal{L}}}} \min_{w} P_{\mathbf{X}_{\bar{\mathcal{L}}}|\mathbf{X}_{\mathcal{S}}}(u_{\bar{\mathcal{L}}|w}) \right] \right|. \tag{41}$$

As such, when users are i.i.d (same distribution leads to the same mechanism, either $\mathcal{M}_1$ or $\mathcal{M}_2$), the lower bound and the upper bound of the proposed mechanisms match each other, which means under the i.i.d assumption, the bounds we derived are tight.

Appendix E
Proof of Theorem 4

Proof:

$$P_{Y|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}})$$
$$= \sum_{a} P_{Y|A,\mathbf{X}_{\mathcal{S}}}(y|a, \mathbf{x}_{\mathcal{S}}) P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})$$
$$= \sum_{a \neq y} P_{Y|A,\mathbf{X}_{\mathcal{S}}}(y|a, \mathbf{x}_{\mathcal{S}}) P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})$$
$$+ P_{Y|A,\mathbf{X}_{\mathcal{S}}}(y|y, \mathbf{x}_{\mathcal{S}}) P_{A|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}}),$$
$$= \sum_{a \neq y} P_A(y) \left[1 - \frac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})} \right] P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})$$
$$+ \left[P_A(y) + (1 - P_A(y)) \frac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})} \right] P_{A|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}})$$
$$= P_A(y) \sum_{a \neq y} \left[P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}}) - \min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w) \right]$$
$$+ P_A(y) P_{A|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}}) + (1 - P_A(y)) \min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w) \tag{42}$$

which can be further expressed as:

$$1 + P_A(y) - P_A(y) \sum_{a} \min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w).$$

As a result, $P_{Y|\mathbf{X}_{\mathcal{S}}}(y|\mathbf{x}_{\mathcal{S}})$ is independent of $\mathbf{X}_{\mathcal{S}}$. □

Appendix F
Proof of Theorem 5

Proof:

$$E[|Y - A|] = \sum_{y} \sum_{a \neq y} \sum_{\mathbf{x}_{\mathcal{S}}} |y - a| P_{Y,A,\mathbf{X}_{\mathcal{S}}}(y, a, \mathbf{x}_{\mathcal{S}})$$
$$= \sum_{y} \sum_{a \neq y} \sum_{\mathbf{x}_{\mathcal{S}}} |y - a| P_{Y|A,\mathbf{X}_{\mathcal{S}}}(y|a, \mathbf{x}_{\mathcal{S}}) P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}}) P_{\mathbf{X}_{\mathcal{S}}}(\mathbf{x}_{\mathcal{S}})$$
$$= \sum_{y} \sum_{a \neq y} \sum_{\mathbf{x}_{\mathcal{S}}} |y - a| P_{\mathbf{X}_{\mathcal{S}}}(\mathbf{x}_{\mathcal{S}})$$
$$\cdot P_A(y) \left[1 - \frac{\min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w)}{P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})} \right] P_{A|\mathbf{X}_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{S}})$$
$$= \sum_{y} \sum_{a \neq y} \sum_{\mathbf{x}_{\mathcal{S}}} |y - a|$$
$$\cdot P_A(y) \left[P_{A,\mathbf{X}_{\mathcal{S}}}(a, \mathbf{x}_{\mathcal{S}}) - \min_w P_{A|\mathbf{X}_{\mathcal{S}}}(a|w) P_{\mathbf{X}_{\mathcal{S}}}(\mathbf{x}_{\mathcal{S}}) \right]$$
$$= \sum_{a=0}^{N} \sum_{y \neq a} |y - a| P_A(y) \left[P_A(a) - \min_w P_{A|X_{\mathcal{S}}}(a|w) \right].$$

□

References

[1] L. Ohno-Machado et al., "Finding useful data across multiple biomedical data repositories using DataMed," *Nature Genet.*, vol. 49, no. 6, pp. 816–819, Jun. 2017.

[2] M. Akgün, A. O. Bayrak, B. Ozer, and M. Ş. Sağıroğlu, "Privacy preserving processing of genomic data: A survey," *J. Biomed. Inf.*, vol. 56, pp. 103–111, Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415001100

[3] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.

[4] I. Shomorony, S. H. Kim, T. A. Courtade, and D. N. C. Tse, "Information-optimal genome assembly via sparse read-overlap graphs," *Bioinformatics*, vol. 32, no. 17, pp. i494–i502, Sep. 2016.

[5] H. Si, H. Vikalo, and S. Vishwanath, "Information-theoretic analysis of haplotype assembly," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3468–3479, Jun. 2017.

[6] H. Cho, D. J. Wu, and B. Berger, "Secure genome-wide association analysis using multiparty computation," *Nature Biotechnol.*, vol. 36, no. 6, pp. 547–551, Jul. 2018.

[7] D. F. Easton and R. A. Eeles, "Genome-wide association studies in cancer," *Hum. Mol. Genet.*, vol. 17, no. R2, pp. R109–R115, Oct. 2008.

[8] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genet.*, vol. 6, no. 2, pp. 95–108, Feb. 2005.

[9] C. Cao and J. Moult, "GWAS and drug targets," *BMC Genomics*, vol. 15, no. S4, p. S5, May 2014.

[10] H. J. Lowe, T. A. Ferris, P. M. Hernandez, and S. C. Weber, "STRIDE—An integrated standards-based translational research informatics platform," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, Nov. 2009, pp. 391–395.

[11] S. N. Murphy, V. Gainer, M. Mendis, S. Churchill, and I. Kohane, "Strategies for maintaining patient privacy in i2b2," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 1, pp. 103–108, Dec. 2011.

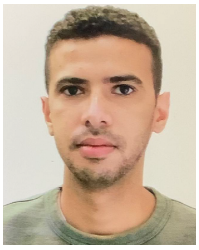
[12] N. Homer et al., "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genet.*, vol. 4, no. 8, Aug. 2008, Art. no. e1000167.

[13] S. S. Shringarpure and C. D. Bustamante, "Privacy risks from genomic data-sharing beacons," *Amer. J. Hum. Genet.*, vol. 97, no. 5, pp. 631–646, Nov. 2015.

[14] I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, and E. Ayday, "An inference attack on genomic data using kinship, complex correlations, and phenotype information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1333–1343, Jul. 2018.

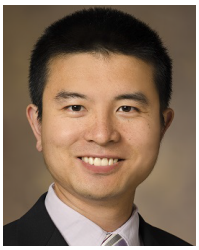
[15] F. Ye, H. Cho, and S. E. Rouayheb, "Mechanisms for hiding sensitive genotypes with information-theoretic privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 902–907.
[16] J. L. Raisaro et al., "Addressing beacon re-identification attacks: Quantification and mitigation of privacy risks," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 4, pp. 799–805, Jul. 2017.
[17] S. Simmons, C. Sahinalp, and B. Berger, "Enabling privacy-preserving GWASs in heterogeneous human populations," *Cell Syst.*, vol. 3, no. 1, pp. 54–61, Jul. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405471216301211
[18] H. Cho, S. Simmons, R. Kim, and B. Berger, "Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs," *Cell Syst.*, vol. 10, no. 5, pp. 408–416, May 2020.
[19] N. Li and M. Stephens, "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data," *Genetics*, vol. 165, no. 4, pp. 2213–2233, Dec. 2003, doi: 10.1093/genetics/165.4.2213.
[20] B. Jiang, M. Seif, R. Tandon, and M. Li, "Context-aware local information privacy," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3694–3708, 2021.
[21] J. C. Venter et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001.
[22] N. von Thenen, E. Ayday, and A. E. Cicek, "Re-identification of individuals in genomic data-sharing beacons via allele inference," *Bioinformatics*, vol. 35, no. 3, pp. 365–371, Jul. 2018, doi: 10.1093/bioinformatics/bty643.
[23] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *Proc. NDSS*, vol. 16, Feb. 2016, pp. 21–24.
[24] F. D. P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5011–5038, Aug. 2017.
[25] H. Wang and F. P. Calmon, "An estimation-theoretic view of privacy," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 886–893.
[26] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1512–1534, Mar. 2019.
[27] T. Berger and R. W. Yeung, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 237–244, Mar. 1989.
[28] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proc. 27th Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2014, pp. 2879–2887.
[29] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proc. 26th USENIX Secur. Symp.*, Aug. 2017, pp. 729–745.
[30] Y. S. Song, "Na Li and Matthew Stephens on modeling linkage disequilibrium," *Genetics*, vol. 203, no. 3, pp. 1005–1006, Jul. 2016, doi: 10.1534/genetics.116.191817.
[31] *Sample GenBank Record*. Accessed: Sep. 30, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/samplerecord/#MoleculeTypeB
[32] J. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2005, ch. 17, pp. 657–687.

Bo Jiang received the bachelor's and first master's degrees from the Harbin Institute of Technology, China, in 2013 and 2015, respectively, and the second master's degree from the Worcester Polytechnic Institute, Worcester, MA, USA, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Arizona. He is a Research Assistant with the Department of Electrical and Computer Engineering, The University of Arizona. His current research interests include security and privacy enhancing technologies, machine learning, radar image processing, and signal processing.



Mohamed Seif received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2014, and the M.Sc. degree in wireless technologies from Nile University, Giza, Egypt, in 2016. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, The University of Arizona. He joined The University of Arizona, as a Graduate Research Assistant, in 2017. His research interests include information theory, machine learning, and wireless communications.

Ravi Tandon (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur (IIT Kanpur), in 2004, and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park (UMCP), in 2010. From 2010 to 2012, he was a Post-Doctoral Research Associate with Princeton University. He is currently the Litton Industries John M. Leonis Distinguished Associate Professor with the Department of ECE, The University of Arizona. Prior to joining The University of Arizona in Fall 2015, he was a Research Assistant Professor with Virginia Tech with positions at the Bradley Department of ECE, Hume Center for National Security and Technology; and the Discovery Analytics Center, Department of Computer Science. His current research interests include information theory and its applications to wireless networks, signal processing, communications, security and privacy, machine learning, and data mining. He was a recipient of the 2018 Keysight Early Career Professor Award, the NSF CAREER Award in 2017, and the Best Paper Award from IEEE GLOBECOM 2011. He serves as an Editor for IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON COMMUNICATIONS.



Ming Li (Senior Member, IEEE) is currently an Associate Professor with the ECE Department, The University of Arizona. He has published more than 130 journals and conference papers, with an H-index of 41. His research interests include wireless networks and security, privacy enhancing technologies, and cyber-physical system security. He is a member of ACM. He received the NSF CAREER Award in 2014, the ONR YIP Award in 2016, and the Best Paper Award from ACM WiSec 2020. He was the TPC Co-Chair of IEEE CNS 2022. He served on the editorial boards for IEEE TRANSACTIONS ON MOBILE COMPUTING and IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING.