Changchang Liu*, Xi He, Thee Chanyaswad, Shiqiang Wang, and Prateek Mittal

Investigating Statistical Privacy Frameworks from the Perspective of Hypothesis Testing

Abstract: Over the last decade, differential privacy (DP) has emerged as the gold standard of a rigorous and provable privacy framework. However, there are very few practical guidelines on how to apply differential privacy in practice, and a key challenge is how to set an appropriate value for the privacy parameter ϵ. In this work, we employ a statistical tool called *hypothesis testing* for discovering useful and interpretable guidelines for the state-of-the-art privacy-preserving frameworks. We formalize and implement hypothesis testing in terms of an adversary's capability to infer mutually exclusive sensitive information about the input data (such as whether an individual has participated or not) from the output of the privacy-preserving mechanism. We quantify the success of the hypothesis testing using the *precision-recall-relation*, which provides an interpretable and natural guideline for practitioners and researchers on selecting ϵ. Our key results include a quantitative analysis of how hypothesis testing can guide the choice of the privacy parameter ϵ in an interpretable manner for a differentially private mechanism and its variants. Importantly, our findings show that an adversary's auxiliary information – in the form of prior distribution of the database and correlation across records and time – indeed influences the proper choice of ϵ. Finally, we also show how the perspective of hypothesis testing can provide useful insights on the relationships among a broad range of privacy frameworks including differential privacy, Pufferfish privacy, Blowfish privacy, dependent differential privacy, inferential privacy, membership privacy and mutual-information based differential privacy.

Keywords: Differential Privacy, Hypothesis Testing, Auxiliary Information

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

***Corresponding Author: Changchang Liu:** Princeton University, E-mail: cl12@princeton.edu (Currently at IBM Research US, E-mail: changchang.liu33@ibm.com)
Xi He: University of Waterloo, E-mail: xihe@uwaterloo.ca
Thee Chanyaswad: KBTG Machine Learning Team, Thailand, E-mail: theerachai.c@kbtg.tech
Shiqiang Wang: IBM Research US, E-mail: wangshiq@us.ibm.com
Prateek Mittal: Princeton University, E-mail: pmittal@princeton.edu

1 Introduction

An important thread of research in the security community has investigated approaches for protecting the privacy of sensitive user data while enabling data analytics [1–7]. Among these approaches, differential privacy (DP) [1, 8–18] has emerged as the gold standard for providing rigorous and provable privacy protection for individuals. A differentially-private mechanism guarantees that the participation of any individual in the database does not significantly *change* the output of the mechanism, where the degree of change is quantified by a tunable privacy parameter ϵ.

While the concept of differential privacy has received considerable attention in the last decade, including industry and government adoption (e.g., Google, Apple, and US Census), *there are very few guidelines on how to apply it in practice* [19, 20]. As illustrated by the recent controversy surrounding Apple's implementation of differential privacy [19], a key challenge facing system designers and researchers is *how to set an appropriate value of* ϵ. Dwork and Smith have also identified this as an open research direction [12]. Specifically, they considered the choice of ϵ as essentially a social question. However, existing tools provide only a limited support for understanding this social question. In addition, it has been observed in [2–4, 21–23] that the appropriate choice of ϵ may also be affected by the existence of auxiliary information. To address these challenges, we aim to provide a rigorous and quantitative procedure to investigate the choice of an appropriate value of ϵ, from the perspective of adversaries' hypothesis testing. In our work, we also consider adversaries that have access to arbitrary auxiliary information, especially focusing on their influence on the choice of ϵ.

Contributions. In order to convincingly determine an appropriate value of ϵ and analyze the effect of auxiliary information on this choice, we need an *interpretable* notion of how much information is leaked about individuals from the mechanism outputs. In other words, we need a tool that can relate the value of ϵ to a more semantically meaningful and, crucially, measurable quantity. Only a limited number of previous works [24–26]

have investigated the question of how to select a proper value of ϵ, but these approaches either require complicated economic models or lack the analysis of adversaries with arbitrary auxiliary information (see Section 2.3 for more details). Our work is inspired by the interpretation of differential privacy via hypothesis testing, initially introduced by Wasserman and Zhou [27–29]. However, this interpretation has not been systematically investigated before in the context of our research objective, i.e., reasoning about the choice of the privacy parameter ϵ (see Section 2.4 for more details).

We consider hypothesis testing [30–32] as the tool used by the adversary to infer sensitive information of an individual record (e.g., the presence or absence of a record in the database for unbounded DP) from the outputs of privacy mechanisms. In particular, we employ the precision-recall (PR) relation from the perspective of hypothesis testing by the adversary as the measurable quantity of interest. The PR relation considers the trade-off between the *precision* (the fraction of examples classified as input records that are truly existing in the input data for unbounded DP) and the *recall* (the fraction of truly existing input records that are correctly detected for unbounded DP) from the adversary's perspective.

With this context of hypothesis testing, we consider three research questions in this work: *how do we set the value of* ϵ, *how does auxiliary information affect the choice of* ϵ, and *can hypothesis testing be used to systematically compare across heterogeneous privacy frameworks?* We introduce our concrete approach to address these questions below.

Investigating Differential Privacy. To explore the choice of an appropriate value of ϵ, we consider an adversary who tries to infer the existence of a record d_i in the database D from the output of a differentially private mechanism $\mathcal{A}(D)$. Our threat model is an adversary who uses hypothesis testing with the Neyman-Pearson criterion [33], which is one of the most powerful criteria in hypothesis testing, on the noisy query results obtained by DP mechanisms. We focus on using the Neyman-Pearson criterion for the Laplace perturbation mechanism [8] in order to perform a concrete analysis. We also show how to generalize our approach to other mechanisms such as the *Gaussian* perturbation mechanism. Particularly, we leverage the PR-relation and the corresponding $F_{\beta score}$ (the weighted harmonic average of precision and recall [34]) as effective metrics to quantify the performance of adversaries' hypothesis testing, which can provide a natural and interpretable guideline for selecting proper privacy parameters by system designers and researchers. Furthermore, we extend our analysis on unbounded DP to bounded DP and the approximate (ϵ, δ)-DP.

Impact of Auxiliary Information. The conjecture that auxiliary information can influence the design of DP mechanisms has been made in prior work [2–4, 21–23]. We therefore investigate the adversary's capability based on hypothesis testing under three types of auxiliary information: *the prior distribution of the input record*, *the correlation across records*, and *the correlation across time*. Our analysis demonstrates that the auxiliary information indeed influences the appropriate selection of ϵ. The results suggest that, when possible and available, the practitioners of DP should explicitly incorporate adversary's auxiliary information into the parameter design of their privacy frameworks. Hence, our results provide a rigorous and systematic answer to the important question posted by Dwork and Smith [12].

Comparison of Statistical Privacy Frameworks. In addition to the two primary questions regarding differential privacy, we also extend our hypothesis testing analysis to a comparative study of a range of state-of-the-art privacy-preserving frameworks [2–7]. Some of these frameworks [2, 3, 21–23] have considered adversaries with auxiliary knowledge in their definitions, but no prior work has applied a common technique to compare and understand their relationship among each other and with differential privacy.

Overall, our work makes the following contributions.

- We investigate differential privacy from the perspective of hypothesis testing by the adversary who observes the differentially private outputs. We comprehensively analyze (i) the unbounded and (ii) bounded scenarios of DP, and (iii) (ϵ, δ)-DP.
- We theoretically derive the *PR-relation* and the corresponding $F_{\beta score}$ as criteria for selecting the value of ϵ that would limit (to the desired extent) the adversary's success in identifying a particular record, in an interpretable and quantitative manner.
- We analyze the effect of three types of auxiliary information, namely, the prior distribution of the input record, the correlation across records, and the correlation across time, on the appropriate choice of ϵ via the hypothesis testing by the adversary.
- Furthermore, we systematically compare the state-of-the-art statistical privacy notions from the perspective of the adversary's hypothesis testing, including Pufferfish privacy [2], Blowfish privacy [3], dependent differential privacy [4], membership

privacy [5], inferential privacy [6] and mutual-information based differential privacy [7].

2 Background and Related Work

In this section, we briefly discuss the frameworks of differential privacy and hypothesis testing, as well as related works regarding these two topics.

2.1 Differential Privacy

Differential privacy is a rigorous mathematical framework aimed at protecting the privacy of the user's record in a statistical database [1, 8, 11–13]. The goal of DP is to randomize the query results to ensure that the risk to the user's privacy does not increase substantially (bounded by a function of ϵ) as a result of participating in the statistical database. The notion of ϵ-differential privacy is formally defined as follows.

Definition 1. (ϵ-differential privacy [1]) *A randomized algorithm $\mathcal{A}$ provides ϵ-differential privacy if for any two neighboring databases D and D' such that D and D' differ by adding/removing a record, and for any subset of outputs $S \subseteq \mathbb{S}$, $\max_{D,D'} \frac{P(\mathcal{A}(D)\in S)}{P(\mathcal{A}(D')\in S)} \leq \exp(\epsilon)$, where $\mathcal{A}(D)$ (resp. $\mathcal{A}(D')$) is the output of $\mathcal{A}$ on input D (resp. D') and ϵ is the privacy budget.*

It is worth noting that the smaller the privacy budget ϵ, the higher the privacy level. This privacy definition is also known as unbounded DP as the database size is unknown. When the database size is known, D and D' are neighbors if D can be obtained from D' by replacing one record in D' with another record. Definition 1 based on this notion of neighbors is known as bounded DP [21]. Approximate DP is another variant of DP, also named (ϵ, δ)-DP [9], and is defined as $P(\mathcal{A}(D) \in S) \leq P(\mathcal{A}(D') \in S)\exp(\epsilon) + \delta$, which relaxes DP by ignoring noisy outputs with a certain probability controlled by the parameter δ. In Section 4, we will analyze mechanisms that satisfy these DP guarantees from the adversary's perspective of hypothesis testing.

The Laplace perturbation mechanism (LPM) [8] is a classic and popular mechanism that achieves ϵ-DP, which makes use of the concept of global sensitivity.

Definition 2. (Global sensitivity) [8] *The global sensitivity of a query $Q : \mathcal{D} \to \mathbb{R}^q$ is the maximum difference between the values of the function when one input changes, i.e. $\Delta Q = \max_{D,D'} \|Q(D) - Q(D')\|_1$.*

Theorem 1. (Laplace Perturbation Mechanism) *[8] For any query $Q : \mathcal{D} \to \mathbb{R}^q$, the Laplace perturbation mechanism, denoted by $\mathcal{A}$, and any database $D \in \mathcal{D}$, $\mathcal{A}(D) = Q(D) + (\eta_1, \ldots, \eta_q)$ achieves ϵ-differential privacy, where η_i are i.i.d random variables drawn from the Laplace distribution with a parameter $\zeta = \Delta Q/\epsilon$, denoted by $Lap(\zeta)$, that is $\Pr[\eta_i = z] \propto \frac{\epsilon}{2\Delta Q} \exp\left(-\frac{\epsilon|z|}{\Delta Q}\right)$.*

In order to perform a concrete analysis, our work mainly focuses on the Laplace perturbation mechanism. However, our approach also generalizes to other mechanisms such as the *Gaussian* perturbation mechanism (as discussed in Section 4.3).

In the literature, several statistical privacy frameworks have also been proposed as important generalization of DP, such as Pufferfish privacy [2], Blowfish privacy [3], dependent differential privacy [4], membership privacy [5], inferential privacy [6] and mutual-information based differential privacy [7]. These privacy frameworks are important statistical privacy frameworks for releasing aggregate information of databases while ensuring provable guarantees, similar to DP. We will systematically compare them with DP from the adversary's perspective of hypothesis testing in Section 6.

2.2 Hypothesis Testing

Hypothesis testing [30–32] is the use of statistics on the observed data to determine the probability that a given hypothesis is true. The common process of hypothesis testing consists of four steps: 1) state the hypotheses; 2) set the criterion for a decision; 3) compute the test statistic; 4) make a decision. The binary hypothesis testing problem[1] decides between a null hypothesis $H = h_0$ and an alternative hypothesis $H = h_1$ based on observation of a random variable O [30, 31]. Under hypothesis h_0, O follows the probability distribution P_0, while under h_1, O follows distribution P_1. A decision rule $\hat{H}$ is a criterion that maps every possible observation $O = o$ to either h_0 or h_1.

The most popularly used criteria for decision are *maximum likelihood* [35], *maximum posterior probability* [36], *minimum cost* [37] and the *Neyman-Pearson criterion* [33].

1 We consider the binary hypothesis testing problem since the adversary aims to distinguish two neighboring databases in DP.

Definition 3. *The Neyman-Pearson criterion aims to maximize the true detection rate (the probability of correctly accepting the alternative hypothesis) subject to a maximum false alarm rate (the probability of mistakenly accepting the alternative hypothesis) [32, 33], i.e.,*

$$\max P_{TD} \quad s.t., \; P_{FA} \leq \alpha \tag{1}$$

where P_{TD} and P_{FA} denote the true detection rate and the false alarm rate, respectively.

The Neyman-Pearson criterion has the highest statistical power [38] since it maximizes the true detection rate under a given requirement of the false alarm rate (as defined above). According to the Neyman-Pearson Lemma [39], an efficient way to solve Eq. 1 is to implement the likelihood ratio test [40, 41]. In practice, the likelihood ratio, or equivalently its logarithm, can be used directly to construct test statistics to compare the goodness of fit of the two hypotheses. Other criteria such as maximum likelihood [35], maximum posterior probability [36] and minimum cost [37] cannot guarantee the highest statistical power in general, and they are based on a fixed threshold on the likelihood ratio that cannot incorporate α. In contrast, the Neyman-Pearson criterion treats the testing performance as a function of the threshold for the likelihood ratio controlled by α ([42], pp. 67). Therefore, the Neyman-Pearson criterion can provide more flexibility to optimize a range of evaluation metrics by setting different values of α [38].

2.3 Setting the Privacy Budget ϵ

Setting the privacy budget ϵ in DP is a challenging task. Prior work [24–26, 43] attempted to address this problem, but has several limitations. Hsu et al. [24] proposed an economic method to express the balance between the accuracy of a DP release and the strength of privacy guarantee in terms of a cost function when bad events happen. However, this work involves complicated economic models consisting of *bad* events and their corresponding *cost* functions. It is difficult to quantify the cost of a bad event for general applications. Krehbiel [25] takes each data owner's privacy preference into consideration and aims to select a proper privacy parameter for achieving a good tradeoff between utility and privacy. This mechanism focuses more on an economic perspective for collecting and distributing payments under a chosen level of privacy parameter and their privacy definition is not the standard DP.

Other related works by Lee and Clifton [26, 43] determine ϵ for LPM based on the posterior probability that the adversary can infer the value of a record. The Neyman-Pearson criterion adopted in our approach has an advantage over the maximum posterior probability analysis [36]. As stated in Section 2.2, the Neyman-Pearson criterion in our work can be used to optimize a range of evaluation metrics by selecting a proper value of the false alarm rate while maximizing the true detection rate. In addition, their analysis [26] only assumes uniform distribution of the input data (equivalent to the scenarios without any prior distribution) in both their experiments and theoretical derivations.

Finally, these previous works are noticeably different from our approach as they do not utilize the statistical tool of hypothesis testing (especially the Neyman-Pearson criterion) by the adversary. Furthermore, our work is not limited to the selection of the ϵ value, but is also extended to the analysis on the impact of the auxiliary information possessed by the adversary and the comparison across the state-of-the-art statistical privacy frameworks, which has not been studied before.

2.4 Hypothesis Testing in DP

Previous work in [44–46] has investigated how to accurately compute the test statistics in hypothesis testing while using DP to protect data. Ding et al. designed an algorithm to detect privacy violations of DP mechanisms from a hypothesis testing perspective [47]. Wasserman and Zhou [27], Hall et al. [28], and Kairouz et al. [29] are our inspiration for using hypothesis testing on DP. These works propose a view of DP from the perspective of hypothesis testing by the adversary who observes the differentially private outputs. Specifically, the observation is first made by Wasserman and Zhou [27] for bounding the probability for the adversary to correctly reject a false hypothesis of the input database. Hall et al. [28] later extend such analysis to (ϵ, δ)-DP. Kairouz et al. [29] then apply this concept in their proof of composition of DP. However, these prior works have not applied hypothesis testing to our objectives of determining the appropriate value of ϵ.

In contrast, our work extensively analyzes the adversary's capability of implementing hypothesis testing using Neyman-Pearson's criterion [33] on outputs of DP mechanisms, such as LPM and the *Gaussian* perturbation mechanism. We apply it to the problem of determining ϵ, the analysis of the effect of auxiliary information on the choice of ϵ, and the comparative analysis

of the privacy guarantee provided by different privacy frameworks. To our knowledge, we are the first to comprehensively investigate how hypothesis testing can be used as a viable tool to guide the selection of the privacy parameter in the design of privacy preservation frameworks. Finally, we note that our approach is not limited to the aforementioned settings and can be generalized to other hypothesis testing techniques and other perturbation mechanisms as well.

3 Methodology Overview

In this section, we describe our quantitative procedure to investigate the choice of an appropriate value of ϵ from the perspective of the adversary's hypothesis testing (shown in Figure 1). Specifically, we discuss the threat model, the quantification method we utilize and the generality of our approach as follows.

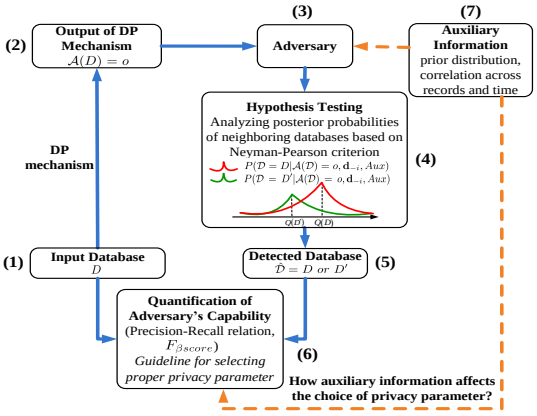


Fig. 1. Overview of our quantitative procedure to investigate the choice of an appropriate value of ϵ from the perspective of the adversary's hypothesis testing.

Threat Model: In this work, we consider the standard adversary in DP that aims to infer the existence of a particular record (for unbounded scenario) or the true value of a record (for bounded scenario) from the noisy output of DP mechanisms. The adversary also has access to the values of all the other records of the input database and/or other auxiliary information such as the prior distribution of the input database and the correlation across records and time. We assume that the adversary exploits the Neyman-Pearson criterion in hypothesis testing to infer sensitive information of a particular record. Note that our analysis considers an information-theoretic/unbounded adversary (and not probabilistic polynomial time (PPT) adversary [48]).

Next, let us briefly describe how we apply the hypothesis testing in our setting. Assume that the adversary has access to the noisy result of DP mechanisms $\mathcal{A}(D) = o$ and tries to infer the existence of a record d_i in the database $D = [d_0, d_1, \cdots]$, where the neighboring database D' assumes the non-existence of the record d_i (for unbounded DP). For the bounded scenario, one database D' can be obtained from its neighbor D by replacing one record with a different value. We use a random variable $\mathcal{D}$ to represent the true database which is unknown to the adversary. The adversary would assume the following two hypotheses:

$$H = \begin{cases} h_0 : \mathcal{D} = D' \\ h_1 : \mathcal{D} = D \end{cases} \tag{2}$$

When the adversary observes the noisy result $\mathcal{A}(D) = o$, he/she tries to distinguish the two events $\mathcal{D} = D$ and $\mathcal{D} = D'$ through analyzing the two posterior probabilities of $P(\mathcal{D} = D|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}, Aux)$ and $P(\mathcal{D} = D'|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}, Aux)$, where $\mathbf{d}_{-i}$ represent the values of all the other records in the input database and Aux represents the auxiliary information that may be accessible to the adversary (which will be discussed in details in Section 5). Using the Neyman-Pearson criterion (instead of directly measuring the statistical difference between the probabilities of neighboring databases), the adversary determines whether to accept D or D' (more details will be described in Sections 4, 5, 6) and we denote this detection result as $\hat{\mathcal{D}}$.

Leveraging PR-relation as Quantification Metric: We utilize the precision-recall (PR)-relation as an effective metric to quantify the adversary's capability of hypothesis testing, which can also serve as a practical guideline for selecting proper values of the privacy parameter ϵ. In statistics, *precision* denotes out of those predicted positive how many of them are actually positive, and *recall* denotes the fraction of the true positives that are labeled as positive. We leverage the PR-relation to quantify the adversary's hypothesis testing (which has not been explored in previous works [27–29]) since it is more useful in practice for problems where the *risk* for the two hypotheses are different. In our setting, the more critical class corresponds to the truly existing records of the input database since the adversary's detection of these records is more serious than that of non-existing records for unbounded DP. Specifically for our problem, *precision* and *recall* are defined as

$$\begin{aligned} precision &= P(h_1|\hat{\mathcal{D}} = D) = P(\mathcal{D} = D|\hat{\mathcal{D}} = D) \\ recall &= P(\hat{\mathcal{D}} = D|h_1) = P(\hat{\mathcal{D}} = D|\mathcal{D} = D) \end{aligned} \tag{3}$$

From Eq. 3, we know that *precision* is the probability that hypothesis h_1, which the adversary's hypothesis testing says is true, is indeed true; and *recall* is the probability that hypothesis h_1, which is indeed true, is detected as true by the adversary's hypothesis testing. Note that the randomness in the process of computing *precision* and *recall* comes from the statistical noise added in DP mechanisms. Therefore, our analysis aims to quantify the capability of the adversary that leverages the hypothesis testing technique in identifying any specific record from DP outputs.

PR-relation quantifies the actual detection accuracy of the adversary, which has a one-to-one correspondence with the false alarm rate P_{FA} and the missed detection rate $1-P_{TD}$ [29]. Different from Kairouz et al. [29] that quantifies the relative relationship between P_{FA} and $1-P_{TD}$, we obtain explicit expression of the precision and recall of the adversary that exploits the Neyman-Pearson criterion (in Sections 4.1.3, 5.1, 5.2, 5.3). It is also interesting to note that there exists a one-to-one correspondence between PR-relation and the receiver operating characteristic (ROC) [49]. Therefore, our analysis in the domain of PR-relation can be directly transferred to the domain of ROC.

Furthermore, we theoretically prove that the adversary that implements Neyman-Pearson criterion achieves the optimal PR-relation (in Theorem 2) and this optimality is generally applicable for correlated records (in Corollary 1). The corresponding proofs are deferred to the appendix.

Theorem 2. *The Neyman-Pearson criterion characterizes the optimal adversary that can achieve the best PR-relation.*

Corollary 1. *The optimality of Neyman-Pearson criterion given in Theorem 2 holds for correlated records in the database.*

In addition, we leverage $F_{\beta score}$ (weighted harmonic average of precision and recall), to further quantify the relationship between the adversary's hypothesis testing and the privacy parameter ϵ, which also provides an interpretable guideline to practitioners and researchers for selecting ϵ. $F_{\beta score} = \frac{1}{\frac{1}{(1+\beta^2)precision} + \frac{\beta^2}{(1+\beta^2)recall}}$ (for any real number $\beta > 0$) is an example for quantifying the PR-relation in order to understand the adversary's power in a more convenient manner, since it combines the two metrics, precision and recall, into a single metric. However, this combination has the potential of information loss of the PR-relation which broadly covers the adversary's hypothesis testing in the entire space (more discussions in Section 4.1.4). Note that every step of our analysis in Figure 1 is accurate in quantifying the adversary's hypothesis testing under the Neyman-Pearson criterion using the metrics of PR-relation and $F_{\beta score}$.

Generalizing to Other Privacy Mechanisms and Practical Adversaries: We further generalize our analysis of the conventional ϵ-DP to its variants such as (ϵ, δ)-DP (adopting the *Gaussian* perturbation mechanism), more advanced privacy notions, and also adversaries with auxiliary information. Our analysis shows that the adversary's auxiliary information in the form of prior distribution of the input database, the correlation across records and time can affect the relationship between the two posterior probabilities (corresponding to the two hypotheses made by the adversary), thus impacting the proper value of privacy parameter ϵ.

4 Quantification of DP from the Adversary's Hypothesis Testing

In this section, we theoretically analyze the capability of the adversary's hypothesis testing for inferring sensitive information of a particular record from DP outputs. Specifically, we implement our analysis on the unbounded and bounded scenarios of DP and (ϵ, δ)-DP.

4.1 Quantification of Unbounded DP

4.1.1 Hypothesis Testing Problem

Recall that the Neyman-Pearson criterion [33] aims to maximize the true detection rate of the hypothesis test given a constrained false alarm rate (Definition 3). Following our threat model and methodology in Section 3, the adversary would assume the following two hypotheses, corresponding to the presence or the absence of a record d_i:

$$H = \begin{cases} h_0 : d_i \text{ does not exists in } \mathcal{D}, \text{ i.e., } \mathcal{D} = D' \\ h_1 : d_i \text{ exists in } \mathcal{D}, \text{ i.e., } \mathcal{D} = D \end{cases} \quad (4)$$

This is clearly the unbounded DP setting (we will analyze the bounded DP setting and other DP variations in the next subsections.). After observing the noisy query result $\mathcal{A}(D) = o$, the adversary tries to distinguish the two events $\mathcal{D} = D$ and $\mathcal{D} = D'$ by analyzing the corresponding posterior probabilities of $P(\mathcal{D} = D|\mathcal{A}(\mathcal{D}) =$

$o, \mathbf{d}_{-i})$ and $P(\mathcal{D} = D'|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})$. Following the *Bayes'* rule of

$$P(\mathcal{D} = D|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}) = \frac{P(\mathcal{A}(\mathcal{D}) = o|\mathbf{d}_{-i}, \mathcal{D} = D)P(\mathcal{D} = D)}{P(\mathcal{A}(\mathcal{D}) = o|\mathbf{d}_{-i})} = \frac{P(\mathcal{A}(D) = o)P(\mathcal{D} = D)}{P(\mathcal{A}(\mathcal{D}) = o|\mathbf{d}_{-i})}, \tag{5}$$

we know that distinguishing the two posterior probabilities is equivalent to differentiating the two conditional probabilities of $P(\mathcal{A}(D) = o)$ and $P(\mathcal{A}(D') = o)$,[2] for adversaries that have no access to the prior distribution of the input database and thus assume a uniform prior, i.e., $P(\mathcal{D} = D) = P(\mathcal{D} = D')$.

4.1.2 Decision Rule

Adversary's Hypothesis Testing for Scalar Query: We first consider the situation where the query output is a scalar and then generalize our analysis to vector output. It is worth noting that DP is defined in terms of *probability measures* but likelihood is defined in terms of *probability densities*. However, we use them interchangeably in this paper (we can change the probability densities to the probability measures through quantizing over the query output for instance). Without loss of generality, we assume $Q(D) \geq Q(D')$. Then, we have the following theorem.

Theorem 3. *Applying Neyman-Pearson criterion of maximizing the true detection rate under a given requirement of false alarm rate α in Definition 3 is equivalent to the following hypothesis testing which is of a simpler formulation: setting a threshold*

$$\theta = \begin{cases} -\frac{\Delta Q \log 2\alpha}{\epsilon} + Q(D'), & \alpha \in [0, 0.5] \\ \frac{\Delta Q \log 2(1-\alpha)}{\epsilon} + Q(D'), & \alpha \in (0.5, 1] \end{cases} \tag{6}$$

for the output of LPM-based DP mechanisms $\mathcal{A}(\mathcal{D}) = o$, then the decision rule of the adversary's hypothesis testing is $o \underset{h_0}{\overset{h_1}{\gtrless}} \theta$.[3]

Proof. Following the Neyman-Pearson Lemma [39], we utilize the likelihood ratio test [40, 41] to realize the Neyman-Pearson criterion. Therefore, we first compute the likelihood ratio $\Lambda(o)$ of the two hypotheses and set a threshold λ on this ratio for the adversary's decision. Under a given requirement of the false alarm rate α, we can uniquely determine the threshold θ of the noisy output. Then, we can compute λ from θ. Next, let us discuss each step of this proof in detail.

Construct Likelihood Ratio Test: Given the noisy scalar output $o = \mathcal{A}(\mathcal{D}) = Q(\mathcal{D}) + Lap(\zeta) = Q(\mathcal{D}) + Lap(\Delta Q/\epsilon)$ from the LPM, we can compute the likelihood ratio $\Lambda(o)$ corresponding to the two hypotheses defined in Eq. 4 as

$$\Lambda(o) = \frac{P(\mathcal{A}(D) = o)}{P(\mathcal{A}(D') = o)} = \frac{\frac{\epsilon}{2\Delta Q}\exp\left(-\frac{\epsilon|o - Q(D)|}{\Delta Q}\right)}{\frac{\epsilon}{2\Delta Q}\exp\left(-\frac{\epsilon|o - Q(D')|}{\Delta Q}\right)} = \begin{cases} \exp(\epsilon) & \text{if } o > Q(D) \\ \exp\left(\frac{2o - Q(D) - Q(D')}{\Delta Q}\epsilon\right) & \text{if } o \in [Q(D'), Q(D)] \\ \exp(-\epsilon) & \text{if } o < Q(D') \end{cases} \tag{7}$$

Assume the decision threshold for the likelihood ratio is λ, then the corresponding decision rule in Neyman-Pearson criterion is $\Lambda(o) \underset{h_0}{\overset{h_1}{\gtrless}} \lambda$.

Uniquely Determine θ from α: Under a threshold θ on the noisy output o, P_{FA} can be computed as $1 - \int_\theta^\infty P(A(D') = o)do = 1 - \int_\theta^\infty \frac{\epsilon}{2\Delta Q} e^{-\frac{\epsilon|x - Q(D')|}{\Delta Q}} dx$, which is $1 - \frac{1}{2}e^{-\frac{(\theta - Q(D'))\epsilon}{\Delta Q}}$ if $\theta < Q(D')$, or $\frac{1}{2}e^{\frac{(\theta - Q(D'))\epsilon}{\Delta Q}}$ if $\theta \geq Q(D')$. Therefore, for a given requirement of the false alarm rate α, we can obtain Eq. 6.

Compute λ from θ: Based on Eq. 7, we know that $\exp(-\epsilon) \leq \Lambda(o) = \frac{P(\mathcal{A}(D)=o)}{P(\mathcal{A}(D')=o)} \leq \exp(\epsilon)$. Therefore, it is sufficient to choose λ such that $\exp(-\epsilon) \leq \lambda \leq \exp(\epsilon)$. Then, the decision rule becomes $e^{\frac{2o - Q(D) - Q(D')}{\Delta Q}\epsilon} \underset{h_0}{\overset{h_1}{\gtrless}} \lambda \implies o \underset{h_0}{\overset{h_1}{\gtrless}} \frac{\log\lambda\Delta Q}{2\epsilon} + \frac{Q(D)+Q(D')}{2}$.[4] Therefore, the threshold λ for the likelihood ratio $\Lambda(o)$ can be computed from the threshold θ for the noisy output o as

$$\lambda = e^{\frac{2\theta - Q(D) - Q(D')}{\Delta Q}\epsilon} \tag{8}$$

Based on the analysis above, we know that there exists a one-to-one correspondence between the false alarm rate α and the threshold of the noisy output θ. This uniquely determined θ satisfies the *existence* and *sufficient* conditions for achieving Neyman-Pearson Lemma (see Theorem 3.2.1 in [38]). Therefore, the Neyman-Pearson criterion in our setting is equivalent to making a decision rule of $o \underset{h_0}{\overset{h_1}{\gtrless}} \theta$ by setting a threshold θ to the noisy output o, which is of a simpler formulation. □

2 $P(\mathcal{A}(D) = o)$ represents the same conditional probability as $P(\mathcal{A}(\mathcal{D}) = o|\mathcal{D} = D)$.

3 The decision rule is $o \underset{h_1}{\overset{h_0}{\gtrless}} \theta$ if $Q(D) \leq Q(D')$.

4 We consider the natural base for logarithm in this paper.

Note that Theorem 3 holds for any adversary that aims to distinguish the neighboring databases, including those who have access to other auxiliary information such as the prior distribution of the input data and correlation across records and time (see Section 5).

Adversary's Hypothesis Testing for Vector Query: Our analysis for the scalar query can be readily generalized to the vector query output according to the following theorem.

Theorem 4. *The best performance of hypothesis testing that the adversary can achieve on the output of ϵ-DP mechanisms $Q : \mathcal{D} \to \mathbb{R}^q$ is the same as that on the output of $q\epsilon$-DP scalar mechanisms.*

Proof. Compared to the scalar query, the privacy property of the vector query $Q : \mathcal{D} \to \mathbb{R}^q$ would decrease by a factor of q based on the sequential composition theorem of DP [50]. Therefore, the adversary's capability for hypothesis testing is increased to that under the scalar query with a privacy parameter of $q\epsilon$. □

Based on Theorem 4, we focus our analyses on the scalar query output in the subsequent discussion.

4.1.3 Evaluating Hypothesis Testing Performance

As stated in Section 3, we interpret the DP constraint (Definition 1) in the context of hypothesis testing in terms of the precision-recall (PR)-relation. Based on Theorem 3, we show the detailed process of hypothesis testing in Figure 2. The red curve and green curve corresponds to the conditional probabilities of the two hypotheses $h_1 : \mathcal{D} = D$ and $h_0 : \mathcal{D} = D'$, respectively. The decision rule for the adversary's hypothesis testing is $o \underset{h_0}{\overset{h_1}{\gtrless}} \theta$. Next, we define two probabilities of $P(\hat{\mathcal{D}} = D|h_1) = \int_\theta^{+\infty} P(\mathcal{A}(D) = o)do$ and $P(\hat{\mathcal{D}} = D|h_0) = \int_\theta^{+\infty} P(\mathcal{A}(D') = o)do$, and these two probabilities are highlighted by RSR (*red shaded region*), GSR (*green shaded region*) in Figure 2. For the LPM, we can compute RSR and GSR as follows.

$$RSR = \begin{cases} 0.5e^{-\frac{(\theta - Q(D))\epsilon}{\Delta Q}}, & \theta \in [Q(D), +\infty) \\ 1 - 0.5e^{\frac{(\theta - Q(D))\epsilon}{\Delta Q}}, & \theta \in (-\infty, Q(D)] \end{cases} \tag{9}$$

$$GSR = \begin{cases} 0.5e^{-\frac{(\theta - Q(D'))\epsilon}{\Delta Q}}, & \theta \in [Q(D'), +\infty) \\ 1 - 0.5e^{\frac{(\theta - Q(D'))\epsilon}{\Delta Q}}, & \theta \in (-\infty, Q(D')] \end{cases} \tag{10}$$

Based on Eq. 3 and Figure 2, we can compute the precision and recall as

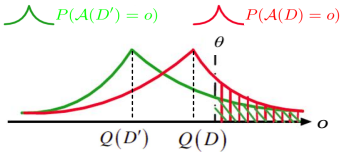


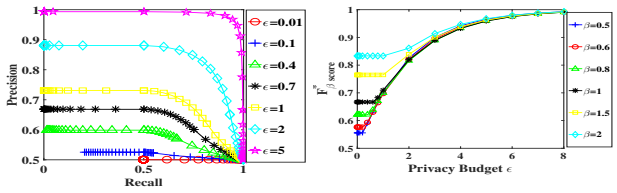
Fig. 2. Adversary's hypothesis testing of detecting a particular record under unbounded DP. By setting a threshold θ, the adversary's decision rule is $o \underset{h_0}{\overset{h_1}{\gtrless}} \theta$.

$$precision = \frac{RSR}{RSR + GSR}, \quad recall \quad = RSR \tag{11}$$

Next, we plot the PR-relation for the adversary's hypothesis testing under different values of ϵ in Figure 3(a). Under each ϵ, the PR-relation is generated using precision and recall values at different θ (thus different values of α in Neyman-Pearson criterion). From Figure 3(a), we have the following observations: 1) we find that the adversary's capability to infer the existence of a particular record decreases as more noise is added (corresponding to a smaller value of ϵ); 2) when the privacy parameter ϵ is very small (e.g., $\epsilon = 0.01$), we have $precision \approx 0.5$, close to the worst hypothesis testing of the adversary (random guessing); 3) when the privacy parameter ϵ is very large (e.g., $\epsilon = 5$), we have very high *precision* and *recall*, close to the best hypothesis testing of the adversary (nearly exact inference). Next, we prove that this analysis of PR-relation is applicable for any query function Q in the following theorem.

Theorem 5. *The PR-relation of the adversary's hypothesis testing on the outputs of LPM is independent of the query function Q.*

Proof. For any query function Q, we define $\psi = \frac{\theta - Q(D')}{\Delta Q}$. The PR-relation shown in Figure 3(a) is directly generated by varying $\theta \in (-\infty, +\infty)$ which is independent of Q. Because θ can be any real number, ψ can also be viewed as a free variable that can take any real number. Then, we have $\frac{\theta - Q(D)}{\Delta Q} = \frac{\theta - Q(D') - \Delta Q}{\Delta Q} = \psi - 1$. Substituting into Eqs. 9, 10, we know that $RSR = 0.5e^{-(\psi-1)\epsilon}$ if $\psi \in [1, +\infty)$, or $1 - 0.5e^{(\psi-1)\epsilon}$ if $\psi \in (-\infty, 1)$, and $GSR = 0.5e^{-\psi\epsilon}$ if $\psi \in [0, +\infty)$, or $1 - 0.5e^{\psi\epsilon}$ if $\psi \in (-\infty, 0)$. Given a RSR, ψ can be computed and then a fixed GSR can be computed accordingly, which is independent of Q. Therefore, the PR-relation generated by precision and recall values (Eq. 11) at different values of $\psi \in (-\infty, +\infty)$ is independent of the query function Q and is only determined by ϵ. □



(a) (b)

Fig. 3. (a) PR-relation and (b) the highest $F_{\beta score}$ of detecting a particular record under unbounded DP.

Table 1. The Maximal ϵ under a Required Bound of $F_{\beta score}$.

$F_{\beta score}$	0.55	0.58	0.62	0.67	0.76	0.83	0.9	0.95
$\beta = 0.5$	0.22	0.34	0.55	0.82	1.42	2.04	3	4.29
$\beta = 0.6$	—	0.33	0.54	0.83	1.45	2.11	3.11	4.43
$\beta = 0.8$	—	—	0.49	0.8	1.46	2.16	3.21	4.58
$\beta = 1$	—	—	—	0.71	1.4	2.12	3.2	4.6
$\beta = 1.5$	—	—	—	—	1.17	1.88	2.99	4.41
$\beta = 2$	—	—	—	—	—	1.61	2.69	4.12

4.1.4 Choosing Proper ϵ

The privacy parameter ϵ can relatively measure the privacy guarantees of DP mechanisms. However, choosing appropriate values for ϵ is non-trivial since its impact on the privacy risks of the input data in practice are not well understood. Our method of choosing ϵ considers the capability of the adversary's hypothesis testing to identify any particular record as being in the database. Specifically, we provide guidelines for selecting proper values of ϵ using PR-relation and $F_{\beta score}$, respectively.

Guidelines under PR-relation: Our analysis by leveraging PR-relation generally covers the adversary's trade-offs between precision and recall in the entire space. Figure 3(a) demonstrates that, with sufficiently low privacy budget ϵ, an adversary's ability to identify an individual record in the database is indeed limited. Under a given requirement of trade-offs between precision and recall of the adversary's hypothesis testing, the privacy mechanism designer can refer to the PR-relation obtained in Eqs. 9, 10, 11 and Figure 3(a) to choose an appropriate privacy budget ϵ.

Guidelines under $F_{\beta score}$: Besides the PR-relation, we also leverage the $F_{\beta score}$, which is a weighted harmonic average of precision and recall, as another appropriate metric for quantifying the effectiveness of the adversary's hypothesis testing. Furthermore, we can theoretically derive the highest $F_{\beta score}$ (by selecting a proper threshold θ) that the adversary can achieve for arbitrary real number $\beta > 0$ as

$$F^*_{\beta score} = \max_{\theta} \frac{1}{\frac{1}{(1+\beta^2)precision} + \frac{\beta^2}{(1+\beta^2)recall}} = \begin{cases} \frac{1+\beta^2}{2+\beta^2}, & \epsilon < \log(1+\beta^2) \\ \frac{(1+\beta^2)(\sqrt{1+4\beta^2 e^{\epsilon}}-1)}{(1+\beta^2)\sqrt{1+4\beta^2 e^{\epsilon}}-1+\beta^2}, & \epsilon \geq \log(1+\beta^2) \end{cases} \quad (12)$$

and the detailed proof is deferred to the Appendix.

Next, we show $F^*_{\beta score}$ with varying privacy parameter ϵ in Figure 3(b). Our results in Eq. 12 and Figure 3(b) are accurate quantification of the adversary's hypothesis testing from the perspective of $F_{\beta score}$, from which we observe that the adversary's capability of inferring the existence of an individual record is generally enhanced with an increasing value of ϵ.

$F_{\beta score}$ can be interpreted as a summary statistic for the PR-relation, which provides a more convenient way of quantifying the relationship between the adversary' hypothesis testing and the privacy parameter ϵ. Under a desired bound of $F^*_{\beta score}$ that the adversary's hypothesis testing can achieve, the privacy mechanism practitioners can refer to Eq. 12 and Figure 3(b) to choose an appropriate value of ϵ. Furthermore, we provide numerical bounds of ϵ under different requirements of $F_{\beta score}$ for commonly-used weights of $\beta \in [0.5, 2]$ in Table 1, as an easier way to look up for privacy practitioners.

When the summary statistics of the precision and recall are used (as opposed to using the full PR-relation information), such as the use of the $F_{\beta score}$, there is potential for information loss, especially in the regime corresponding to smaller values of ϵ (Eq. 12 and Figure 3(b)). It is interesting to note that $F^*_{\beta score}$ keeps the same for $\epsilon < \log(1+\beta^2)$, and then monotonically increases with ϵ for $\epsilon \geq \log(1+\beta^2)$. This turning point $\epsilon = \log(1+\beta^2)$ approaches 0 for smaller values of β, making $F^*_{\beta score}$ closer to be monotonically increasing with ϵ thus capturing the privacy benefits of smaller values of ϵ (as shown Figure 3(b)).

Finally, we emphasize that the alternative approach of using the entire PR-relation to guide the selection of ϵ does not suffer from the limitations discussed above, and also shows privacy benefits of using smaller ϵ (smaller *precision* for a given *recall* as shown in Figure 3(a)).

4.1.5 Plausible Deniability Property

There are multiple ways to interpret semantics of DP guarantees such as hypothesis testing [27–29] and plausible deniability (Page 9 in Dwork [1], Page 2 in Dwork and Smith [12], Definition 1 in Dwork, McSherry, Nis-

sim and Smith [8], Section 2 in Kasiviswanathan and Smith [15], Section 4 in Li et al. [5]). The potential of randomness providing plausible deniability was first recognized by Warner [51]. Bindschaedler et al. provide a formal definition of plausible deniability for data synthesis, compared to which DP is a stronger privacy guarantee [52].

Definition 4. (Plausible Deniability) [52] *For any database D with $|D| > k$ ($|D|$ is the number of records in D), and any record y generated by a perturbation mechanism $\mathcal{M}$ such that $y = \mathcal{M}(d_1)$ for $d_1 \in D$, we state that y is releasable with (k, γ)-plausible deniability, if there exist at least $k-1$ distinct records $d_2, \cdots, d_k \in D \setminus d_1$ such that $\gamma^{-1} \leq \frac{P(\mathcal{M}(d_i)=y)}{P(\mathcal{M}(d_j)=y)} \leq \gamma$ for any $i, j \in \{1, 2, \cdots, k\}$.*

We interpret DP as hypothesis testing — how well an adversary in DP can infer the existence of an individual record (unbounded DP) or the exact value of a record (bounded DP) in binary hypothesis testing problem involving two neighboring databases (Dwork [1], Dwork, McSherry, Nissim and Smith [8], Kifer and Machanavajjhala [21]). Theorem 3 demonstrates that the adversary implements the likelihood ratio test $\Lambda(o) = \frac{P(\mathcal{A}(D)=o)}{P(\mathcal{A}(D')=o)}$ to satisfy the Neyman-Pearson criterion and the decision rule in Eq. 6 is equivalent to $\Lambda(o) \underset{h_0}{\overset{h_1}{\gtrless}} \lambda$. Combining Eq. 6 and Eq. 8, we know that $\lambda = \frac{e^{-\epsilon}}{4\alpha^2}$ if $\alpha \in [0, 0.5]$, or $\frac{e^{-\epsilon}}{4(1-\alpha)^2}$ if $\alpha \in (0.5, 1]$. According to Definition 4, the plausible deniability also quantifies the likelihood ratio between two data (although it only considers the scenario of privacy-preserving data synthesis [52]). Therefore, our analysis of using hypothesis testing to guide selection of proper privacy parameters in DP has implicitly incorporated the plausible deniability of any individual records in the database (controlled by the maximum false alarm rate α in the Neyman-Person criterion). Furthermore, α determines a trade-off between the false alarm rate P_{FA} and the true detection rate P_{TD} (Definition 3). Therefore, our analysis of using PR-relation (which has a one-to-one correspondence with P_{FA}, P_{TD}) and $F_{\beta score}$ (summary statistics of precision and recall) generated by varying α quantifies the randomness and the plausible deniability [51] [52] of any individual records in the database.

4.2 Quantification of Bounded DP

Bounded DP corresponds to the setting where the neighboring databases differ in one (record's) value and their size is the same. For simplicity, we first discuss a scenario where the i-th record of the input data d_i can only take binary values d_{i1}, d_{i2}. Note that the hypothesis testing by the adversary in the bounded scenario is different from the unbounded scenario in that the adversary is no longer aiming to distinguish the absence/presence of a record, but to estimate the true value of a record. Thus, the adversary's hypothesis testing now becomes:

$$H = \begin{cases} h_0: \ d_i = d_{i1} \\ h_1: \ d_i = d_{i2} \end{cases} \tag{13}$$

Comparing Eq. 4 and Eq. 13, we know that the two hypotheses for unbounded and bounded cases are different. However, according to Theorem 5, their corresponding PR-relation (and $F_{\beta score}$) are the same. This means, the hypothesis testing implemented by the adversary for bounded DP with binary records is the same as that for unbounded DP.

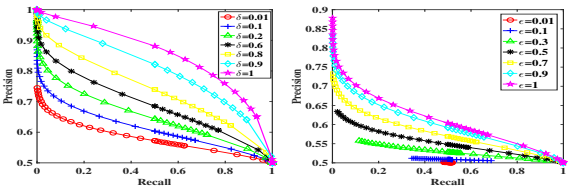
Next, we consider a more general scenario where d_i takes multiple values $d_{i1}, d_{i2}, \cdots, d_{ik}$. Without loss of generality, we assume $Q(d_{i1}) \leq Q(d_{i2}) \leq \cdots \leq Q(d_{ik})$. Therefore, the distance between any two query results computed over two different values of d_i is smaller than the sensitivity of the query $\Delta Q = \max \|Q(d_{ik}) - Q(d_{i1})\|$. Since the inserted noise for satisfying DP is calculated based on ΔQ (i.e., $Lap(\frac{\Delta Q}{\epsilon})$), we know that the hypothesis testing achieved by the adversary in distinguishing any two values of d_i is not worse than distinguishing d_{i1} and d_{ik}. We thus conclude that the best hypothesis testing of the adversary for the bounded scenario is the same as that for the unbounded scenario.

4.3 Quantification of (ϵ, δ)-DP

Approximate DP, also named (ϵ, δ)-DP [9] is defined as $P(\mathcal{A}(D) \in S) \leq \exp(\epsilon) P(\mathcal{A}(D') \in S) + \delta$ for any neighboring databases D, D'. One of the most popular mechanisms to achieve (ϵ, δ)-DP is the *Gaussian* perturbation mechanism, where a *Gaussian* noise with zero mean and standard variant $\sigma = \sqrt{2 \log(1.25/\delta) \Delta Q / \epsilon}$ is added to the query output [9, 53].

Similar to Section 4.1, we first derive the mechanism for the adversary's hypothesis testing that satisfies the Neyman-Pearson criterion based on the following theorem (detailed proof is deferred to the Appendix).

Theorem 6. *Applying Neyman-Pearson criterion in Definition 3 is equivalent to the following hypothesis testing which is of a simpler formulation: setting a threshold $\theta = \Phi^{-1}(1-\alpha)\sigma + Q(D')$ (where α is the max-*



(a) (b)

Fig. 4. PR-relation of detecting a particular record from (ϵ, δ)-DP results with varying (a) ϵ and (b) δ, respectively.

imum P_{FA} and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution) for the output of the Gaussian perturbation mechanism o, the decision for the adversary's hypothesis testing is $o \underset{h_0}{\overset{h_1}{\gtrless}} \theta$.

According to Theorem 6 and Eq. 11, we can theoretically derive *precision* and *recall* of the adversary's hypothesis testing for the *Gaussian* mechanism as

$$precision = \frac{1 - \Phi\left(\frac{\theta - Q(D)}{\sigma}\right)}{2 - \Phi\left(\frac{\theta - Q(D)}{\sigma}\right) - \Phi\left(\frac{\theta - Q(D')}{\sigma}\right)}$$
$$recall = 1 - \Phi\left(\frac{\theta - Q(D)}{\sigma}\right) \quad (14)$$

We further show the corresponding PR-relation of the adversary's hypothesis testing in Figure 4(a) and Figure 4(b) with varying ϵ and δ, respectively. We observe that the adversary's performance of hypothesis testing is enhanced with an increasing value of ϵ and δ. Comparing Figure 3(a) and Figure 4, we know that different mechanisms vary in their power of defending against adversaries' hypothesis testing since their output distributions are different. Note that our approach can be generally applied to any DP mechanism (Figure 1), although we focus on LPM and GPM.

Next, the highest $F_{\beta score}$ of the adversary's hypothesis testing under different values of privacy parameters ϵ, δ can be directly derived from Eq. 14 as

$$F^*_{\beta score} = \max_{\theta} \frac{(1+\beta^2)\left(1 - \Phi\left(\frac{\theta - Q(D)}{\sigma}\right)\right)}{2 + \beta^2 - \Phi\left(\frac{\theta - Q(D)}{\sigma}\right) - \Phi\left(\frac{\theta - Q(D')}{\sigma}\right)} \quad (15)$$

Balle and Wang [54] developed the analytic Gaussian mechanism whose variance is calibrated using the Gaussian cumulative density function instead of a tail bound approximation. Other recent works also proposed tight lower bounds for the variance of the added Gaussian noise while satisfying (ϵ, δ)-DP [55, 56]. How to adapt our analysis for the classical Gaussian mechanism in Theorem 6 and Eqs. 14, 15 to these improved Gaussian mechanisms [54–56] is an interesting future direction.

5 Quantification of DP under Auxiliary Information from the Adversary's Hypothesis Testing

We now demonstrate how to control the adversary's success rate in identifying of a particular record with several important variations of the adversary's belief including the input data's prior distribution, record correlation and temporal correlation.

5.1 Quantification of DP under Prior Distribution

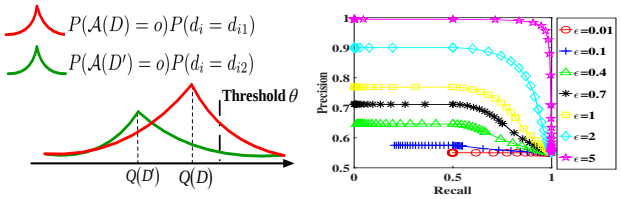
Let us first consider an adversary with known prior distribution of the input data. Although such an adversary is not explicitly considered in conventional DP frameworks[5], we still analyze this adversary's inference for sensitive information in a particular record as an interesting and practical case study. In some scenarios, the adversary's prior is non-uniform, which will result in a different decision rule. Similar to our analysis in Section 4.2, we still consider a binary hypothesis testing problem where the adversary aims to distinguish the two neighboring databases in Eq. 13.

Next, we quantify the hypothesis testing of the adversary to distinguish the two posterior distributions of $P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}), P(d_i = d_{i2}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})$. According to *Bayes*' rule, we have $P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}) = \frac{P(\mathcal{A}(\mathcal{D})=o|\mathbf{d}_{-i}, d_i=d_{i1})P(d_i=d_{i1})}{P(\mathcal{A}(\mathcal{D})=o|\mathbf{d}_{-i})} = \frac{P(\mathcal{A}(D)=o)P(d_i=d_{i1})}{P(\mathcal{A}(\mathcal{D})=o|\mathbf{d}_{-i})}$. Then, we get

$$\frac{P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})}{P(d = d_{i2}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})} = \frac{P(\mathcal{A}(D) = o)P(d_i = d_{i1})}{P(\mathcal{A}(D') = o)P(d_i = d_{i2})} \quad (16)$$

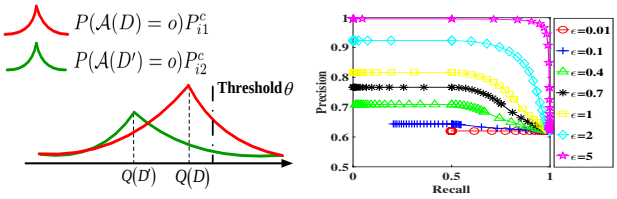
Based on Eq. 16, we know that the adversary's hypothesis testing under prior distribution is equivalent to distinguishing the two probabilities of $P(\mathcal{A}(D) = o)P(d_i = d_{i1})$ and $P(\mathcal{A}(D') = o)P(d_i = d_{i2})$. Figure 5(a) shows the hypothesis testing procedure of the adversary, where θ is the decision threshold on the noisy query outputs and the adversary's decision rule is $o \underset{d_{i2}}{\overset{d_{i1}}{\gtrless}} \theta$. We further define the *coefficient of prior distribution* ρ_p as $\rho_p = 1 - \min_{d_{i1}, d_{i2}} \frac{P(d_i=d_{i1})}{P(d_i=d_{i2})}$, where $\rho_p \in [0, 1]$. $\rho_p = 0$ corresponds to the scenario where the adversary has no knowledge about the prior distribution and thus makes the assumption of uniform distribution (the same as in Section 4.1). Combining Eq. 11 with Eq. 16, we can de-

5 DP guarantees are not influenced by the prior distribution of the input data.



(a) (b)

Fig. 5. (a) Hypothesis testing and (b) PR-relation of detecting a particular record for adversaries with access to the prior distribution of the input data ($\rho_p = 0.2$).



(a) (b)

Fig. 6. (a) Hypothesis testing and (b) PR-relation of detecting a particular record for adversaries with access to record correlation ($\rho_p = 0.2, \rho_c = 0.1$).

rive *precision, recall* as

$$\begin{aligned} precision &= \frac{P(d_i = d_{i1})RSR}{P(d_i = d_{i1})RSR + P(d_i = d_{i2})GSR} \\ &= \frac{1}{1 + (1-\rho_p)\frac{GSR}{RSR}}, \\ recall &= RSR \end{aligned} \tag{17}$$

From Eq. 17, we know that *precision* is increased while *recall* is kept unchanged for adversaries knowing prior distribution of the input data. We further show this enhanced PR-relation in Figure 5(b) by setting $\rho_p = 0.2$ for instance, as a comparison to Figure 3(a) (corresponding to $\rho_p = 0$).

Furthermore, we theoretically derive the highest $F_{\beta score}$ of the adversary's hypothesis testing under different values of ϵ and ρ_p as

$$F^*_{\beta score} = \begin{cases} \dfrac{1+\beta^2}{2+\beta^2-\rho_p}, & \epsilon < \log(1 + \dfrac{\beta^2}{1-\rho_p}) \\ \dfrac{(1+\beta^2)(\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p}} - 1)}{(1+\beta^2)\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p}} - 1 + \beta^2}, & \epsilon \geq \log(1 + \dfrac{\beta^2}{1-\rho_p}) \end{cases} \tag{18}$$

The detailed proof is deferred to the Appendix. Comparing Eq. 18 and Eq. 12, we know that 1) the adversary achieves an improved hypothesis testing by possessing auxiliary information of prior distribution and 2) a larger value of ρ_p results in a higher confidence for the adversary to select the correct value of d_i (with higher $F_{\beta score}$). Therefore, we conclude that the prior distribution of the input data should be considered when trying to select a proper ϵ in practice.

5.2 Quantification of DP under Record Correlation

Records in real world data often exhibit inherent dependencies or correlations. Handling correlated records is thus a significant problem, which has been demonstrated in previous work [21–23, 57–60]. Tschantz, Sen and Datta in [60] investigate an interesting causal view of DP as limiting the effect of a single data point without independence assumptions, in order to resolve the confusion in prior work about DP under correlated data. Here, we will show the enhanced hypothesis testing of the adversary who has access to the correlation relationship across records. Note that we now consider the bounded case since the presence/absence of a record in the unbounded case has no relation with the correlation among records. This is the same for correlation across time which will be discussed in Sections 5.3. Let us consider a general setting where the adversary aims to infer the value of d_i while having access to values of all the other records $\mathbf{d}_{-i}$ and the relationship between d_i and its correlated records $d_{c1}, d_{c2}, \cdots$. Therefore, the adversary's hypothesis testing tries to distinguish the two posterior probabilities of $P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})$ and $P(d_i = d_{i2}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})$. Define $P^c_{i1} = P(d_i = d_{i1}|d_{c1}, d_{c2}, \cdots)$ and $P^c_{i2} = P(d_i = d_{i2}|d_{c1}, d_{c2}, \cdots)$. According to *Bayes*' rule, we can derive $P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i}) = \frac{P(\mathcal{A}(\mathcal{D})=o|d_i=d_{i1},\mathbf{d}_{-i})P^c_{i1}}{P(\mathcal{A}(\mathcal{D})=o|\mathbf{d}_{-i})} = \frac{P(\mathcal{A}(D)=o)P^c_{i1}}{P(\mathcal{A}(\mathcal{D})=o|\mathbf{d}_{-i})}$. Then, we know

$$\frac{P(d_i = d_{i1}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})}{P(d_i = d_{i2}|\mathcal{A}(\mathcal{D}) = o, \mathbf{d}_{-i})} = \frac{P(\mathcal{A}(D) = o)P^c_{i1}}{P(\mathcal{A}(D') = o)P^c_{i2}} \tag{19}$$

Based on Eq. 19, we know that the adversary's hypothesis testing under record correlation is equivalent to distinguishing the probabilities of $P(\mathcal{A}(D) = o)P^c_{i1}$ and $P(\mathcal{A}(D') = o)P^c_{i2}$ as shown in Figure 6(a). Let us further define the *coefficient of record correlation* ρ_c as $\rho_c = 1 - \min_{d_{i1}, d_{c1}, d_{c2}, \cdots} \frac{P(d_i = d_{i1})}{P(d_i = d_{i1}|d_{c1}, d_{c2}, \cdots)}$, where $\rho_c \in [0, 1]$, and $\rho_c = 0$ corresponds to the scenario of independent records. Combining Eq. 11 and Eq. 19, we can compute the *precision* and *recall* for this adversary's

hypothesis testing as

$$\begin{aligned} precision &= \frac{P^c_{i1} RSR}{P^c_{i1} RSR + P^c_{i2} GSR} \\ &= \frac{1}{1 + (1 - \rho_p - \rho_c(2-\rho_p))\frac{GSR}{RSR}}, \\ recall &= RSR. \end{aligned} \tag{20}$$

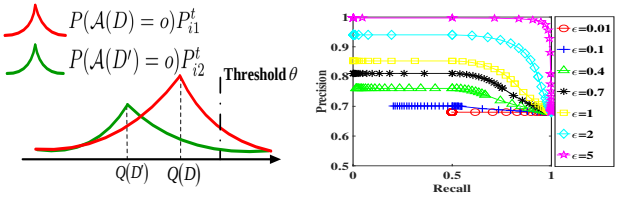
Eq. 20 holds based on the fact that $\max \frac{P^c_{i1}}{P^c_{i2}} = \max \frac{P^c_{i1}}{1-P^c_{i1}} = \frac{\frac{1}{1-\rho_p}}{(1+\frac{1}{1-\rho_p})(1-\rho_c)} / (1 - \frac{\frac{1}{1-\rho_p}}{(1+\frac{1}{1-\rho_p})(1-\rho_c)}) = \frac{1}{1-\rho_p-\rho_c(2-\rho_p)}$. From Eq. 20, we know that *precision* is increased under the same level of *recall* for adversaries that possess record correlation of the input data. We show the enhanced PR-relation in Figure 6(b) by setting $\rho_p = 0.2, \rho_c = 0.1$ for example. Furthermore, we theoretically derive the relationship between the highest $F_{\beta score}$ and privacy budget ϵ, coefficient of prior distribution ρ_p, coefficient of record correlation ρ_c as

$$F^*_{\beta score} = \begin{cases} \frac{1+\beta^2}{2+\beta^2-\rho_p-\rho_c(2-\rho_p)}, & \epsilon < \epsilon(\rho_p, \rho_c) \\ \frac{(1+\beta^2)(\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p-\rho_c(2-\rho_p)}}-1)}{(1+\beta^2)\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p-\rho_c(2-\rho_p)}}-1+\beta^2}, & \epsilon \geq \epsilon(\rho_p, \rho_c) \end{cases} \tag{21}$$

where $\epsilon(\rho_p, \rho_c) = \log(1 + \frac{\beta^2}{1-\rho_p-\rho_c(2-\rho_p)})$ and the corresponding proof is deferred to the Appendix. From Eq. 21, we know that $F^*_{\beta score}$ is further improved by record correlation accessible to the adversary, meaning that the adversary has more confidence to detect the true value of d_i. Therefore, the correlation across records should also be taken into consideration when selecting appropriate values of ϵ in practice.

5.3 Quantification of DP under Temporal Correlation

Temporal dynamics exist naturally in information networks, e.g., social network, mobility data, health records, etc. For instance, users' location data which is provided to location-based services or applications are usually temporally correlated. Location-based social networks allow users to share locations with friends, to find friends, and to provide recommendations about points of interest based on their locations. Yet, individual privacy has been a major obstacle to data sharing. Many privacy frameworks including DP schemes do not explicitly incorporate such dynamics. Most of current perturbation mechanisms only consider static scenarios or perturb the location at single timestamps without considering temporal correlations of a moving user's data. Even advanced variants of DP frameworks such as dependent differential privacy [4] only considers correlation among records in a single static database. Xiao et al. [61] and Cao et al. [62] consider temporal correlation across a single user's data instead of under multiple users' database. In practice, a time-series of users' data may need to be published to enable real-world applications while satisfying rigorous privacy guarantees.



(a) (b)

Fig. 7. (a) Hypothesis testing and (b) F1-score of detecting a particular record for adversaries with access to temporal correlation ($\rho_p = 0.2, \rho_c = 0.1, \rho_t = 0.1$).

Let us consider a general temporal setting where the adversary aims to infer the value of d^t_i at timestamp t while having access to the values of all the other records $\mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i$ and the relationship between d^t_i and its correlated records (across time and within this timestamp) $\mathbf{d}^{[t-1]}_i, \mathbf{d}^{[t]}_{c1}, \mathbf{d}^{[t]}_{c2}, \cdots$. [6] Similar to Sections 5.1, 5.2, the adversary aims to distinguish two posterior probabilities of $P(d^t_i = d_{i1} | \mathcal{A}(\mathcal{D}) = o, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)$ and $P(D^t_i = d_{i2} | \mathcal{A}(\mathcal{D}) = o, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)$. Defining $P^t_{i1} = P(d^t_i = d_{i1} | \mathbf{d}^{[t-1]}_i, \mathbf{d}^{[t]}_{c1}, \mathbf{d}^{[t]}_{c2}, \cdots)$ and $P^t_{i2} = P(d^t_i = d_{i2} | \mathbf{d}^{[t-1]}_i, \mathbf{d}^{[t]}_{c1}, \mathbf{d}^{[t]}_{c2}, \cdots)$. According to *Bayes*' rule, we have $P(d^t_i = d_{i1} | \mathcal{A}(\mathcal{D}) = o, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i) = \frac{P(\mathcal{A}(\mathcal{D})=o | d^t_i = d_{i1}, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i) P^t_{i1}}{P(\mathcal{A}(\mathcal{D})=o | \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)} = \frac{P(\mathcal{A}(D)=o) P^t_{i1}}{P(\mathcal{A}(\mathcal{D})=o | \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)}$. Therefore, we can derive

$$\frac{P(d^t_i = d_{i1} | \mathcal{A}(\mathcal{D}) = o, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)}{P(d^t_i = d_{i2} | \mathcal{A}(\mathcal{D}) = o, \mathbf{d}^{[t]}_{-i}, \mathbf{d}^{[t-1]}_i)} = \frac{P(\mathcal{A}(D) = o) P^t_{i1}}{P(\mathcal{A}(D') = o) P^t_{i2}} \tag{22}$$

From Eq. 22, we know that the adversary's hypothesis testing under temporal dynamics is equivalent to distinguishing the two probabilities of $P(\mathcal{A}(D) = o)P^t_{i1}$ and $P(\mathcal{A}(D') = o)P^t_{i2}$ as shown in Figure 7(a). Let us define the *coefficient of temporal correlation* as $\rho_t = 1 - \min_{d_{i1}, \mathbf{d}^{[t-1]}_i, \mathbf{d}^{[t]}_{c1}, \mathbf{d}^{[t]}_{c2}, \cdots} \frac{P(d_i = d_{i1} | d_{c1}, d_{c2}, \cdots)}{P(d^t_i = d_{i1} | \mathbf{d}^{[t-1]}_i, \mathbf{d}^{[t]}_{c1}, \mathbf{d}^{[t]}_{c2}, \cdots)}$, where $\rho_t \in [0, 1]$ and $\rho_t = 0$ corresponds to the static scenario. Therefore, we can compute the precision and

6 Here, $[t]$ represents timestamps from 1 to t.

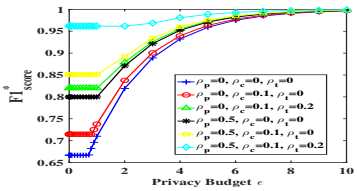


Fig. 8. The highest $F_{\beta score}$ achieved by the adversary under different auxiliary information (setting $\beta = 1$, i.e., $F1_{score}$).

recall for this adversary's hypothesis testing as

$$\begin{aligned} precision &= \frac{P^t_{i1} RSR}{P^t_{i1} RSR + P^t_{i2} GSR} \\ &= \frac{1}{1 + (1 - \rho_p - (2 - \rho_p)(\rho_c + \rho_t(1 - \rho_c)))\frac{GSR}{RSR}}, \\ recall &= RSR. \end{aligned} \tag{23}$$

Eq. 23 holds since $\max \frac{P^t_{i1}}{P^t_{i2}} = \max \frac{P^t_{i1}}{1-P^t_{i1}} = \frac{\frac{1}{1-\rho_p}(1+\frac{1}{1-\rho_p})(1-\rho_c)(1-\rho_t)}{1-\frac{1}{1-\rho_p}(1+\frac{1}{1-\rho_p})(1-\rho_c)(1-\rho_t)} = \frac{1}{(2-\rho_p)(1-\rho_c)(1-\rho_t)-1} = \frac{1}{1-\rho_p-(2-\rho_p)(\rho_c+\rho_t(1-\rho_c))}$. From Eq. 23, we know that *precision* is increased under the same level of *recall* when the adversary has access to the temporal correlation of the input data, resulting in a better PR-relation compared to the static scenario. We show the enhanced PR-relation in Figure 7(b) by setting $\rho_p = 0.2, \rho_c = 0.1, \rho_t = 0.1$ for example. Furthermore, we theoretically compute the highest $F_{\beta score}$ under given values of privacy budget ϵ, coefficient of prior distribution ρ_p, coefficient of record correlation ρ_c and coefficient of temporal correlation ρ_t as

$$F^*_{\beta score} = \begin{cases} \frac{1+\beta^2}{2+\beta^2-\rho_p-(2-\rho_p)(\rho_c+\rho_t(1-\rho_c))}, & \epsilon < \epsilon(\rho_p,\rho_c,\rho_t) \\ \frac{(1+\beta^2)(\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p-(2-\rho_p)(\rho_c+\rho_t(1-\rho_c))}}-1)}{(1+\beta^2)\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p-(2-\rho_p)(\rho_c+\rho_t(1-\rho_c))}}-1+\beta^2}, & \epsilon \geq \epsilon(\rho_p,\rho_c,\rho_t) \end{cases} \tag{24}$$

where $\epsilon(\rho_p, \rho_c, \rho_t) = \log(1 + \frac{\beta^2}{1-\rho_p-(2-\rho_p)(\rho_c+\rho_t(1-\rho_c))})$ and the corresponding proof is deferred to the Appendix. From Eq. 24, we know that the temporal correlation can benefit the adversary's hypothesis testing to achieve an enhanced $F^*_{\beta score}$. Therefore, the correlation of the input data across time should be considered in selecting appropriate privacy parameters of privacy preserving mechanisms.

Summary for the Quantification of DP under Auxiliary Information: Figure 8 shows the highest $F_{\beta score}$ (setting $\beta = 1$) varying with ϵ under different auxiliary information. We can observe that the adversary can infer more information of the input data with more auxiliary information (higher values of ρ_p, ρ_c, ρ_t). This property can also be explained by using *conditioning always reduces entropy (uncertainty)* in information theory [63]. Therefore, choosing a proper privacy budget ϵ needs more careful consideration when designing privacy-preserving mechanisms against adversaries who have access to these auxiliary information.

5.3.1 Relating PR-relation to DP Guarantees

Since the computation of PR-relation involves a specific adversary model, the connection between the PR-relation and privacy guarantees (DP) depend on what assumptions we make for the adversaries.

Optimal Adversary: considering an optimal adversary that implements the Neyman-Pearson criterion and with full access to the possible auxiliary information of the data, under a specific mechanism, the PR-relation achieved by the adversary would be fixed (since every step of our analysis in Figure 1 is exact) as shown in Eqs. 9, 10, 11 for Laplacian mechanism, Eq. 14 for Gaussian mechanism, and Eqs. 17, 20, 23 under auxiliary information. For a given PR-relation that follows this fixed pattern (named as $PR_{optimal}$), we thus can infer the values of privacy parameters in DP guarantees by computing them directly from the corresponding equations (Eqs. 9, 10, 11, 14, 17, 20, 23) of PR-relation or the corresponding Figures 3(a), 4, 5(b), 6(b), 7(b).

Realistic Adversary: considering a realistic adversary that may not have access to the full auxiliary information, the PR-relation (named as $PR_{realistic}$) may be different. Nevertheless, we can still obtain a lower bound for the corresponding privacy parameters, by finding a lower bound of $PR_{realistic}$, denoted as $PR_{optimalLow}$, within all possible $PR_{optimal}$ relations (corresponding to different privacy parameters). Since the best PR-relation achieved by the optimal adversary under this mechanism should be better than $PR_{realistic}$, we know that the values of privacy parameters corresponding to the given $PR_{realistic}$ should be larger than the privacy parameters corresponding to the lower-bound optimal PR-relation $PR_{optimalLow}$.

6 Quantification of Other Privacy Notions from the Adversary's Hypothesis Testing

In this section, we systematically compare several existing statistical privacy frameworks from the perspective

of the adversary's hypothesis testing, including Pufferfish privacy [2], Blowfish privacy [3], dependent differential privacy [4], membership privacy [5], inferential privacy [6], and mutual-information based differential privacy [7] (detailed definitions are deferred to the Appendix). Our analysis can deepen the understanding of these privacy notions as well as guide the design of their deployment in real world applications.

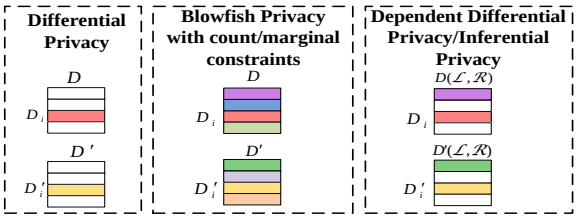


Fig. 9. Comparison of neighboring databases in the state-of-the-art statistical privacy frameworks.

6.1 Qualitative Comparison of Different Privacy Metrics

From the perspective of adversary's hypothesis testing, the adversary aims to distinguish between two neighboring databases from the noisy outputs satisfying different privacy metrics (recall Eqs. 2, 4, 13). We thus compare the definitions of neighboring databases in various privacy notions as shown in Figure 9. The neighboring databases in DP [1] considers the change of only one record in the database. Pufferfish privacy [2] aims to protect any potential secret of the database and Blowfish privacy [3] is a special class of privacy notions in the Pufferfish framework. The neighboring databases in Blowfish privacy with count query constraint and marginal constraint [3] can generally consider all the possible records' differences in the databases. Dependent differential privacy [4] and inferential privacy [6] aim to protect a particular record while taking its correlation with other records into consideration, therefore their neighboring databases are generated by the direct change of one record followed by possible changes of its correlated records. Furthermore, we analyze the relationship between these two notions. According to *Bayes*' analysis in Eq. 5, we know inferential privacy that requires $\max_{D,D'} \frac{P(\mathcal{D}=D|\mathcal{A}(\mathcal{D})=o)}{P(\mathcal{D}=D'|\mathcal{A}(\mathcal{D})=o)} \leq e^{\epsilon} \frac{P(\mathcal{D}=D)}{P(\mathcal{D}=D')}$ is equivalent to dependent differential privacy that requires $\max_{D,D'} \frac{P(\mathcal{A}(D)=o)}{P(\mathcal{A}(D')=o)} \leq e^{\epsilon}$.

Note that membership privacy [5] and mutual-information differential privacy [7] are different from the above privacy metrics since their frameworks have not explicitly defined neighboring databases. We will show the relationship between them and all the other privacy metrics in the next subsection from the perspective of adversary's hypothesis testing. Under the same privacy parameter, a privacy notion that places less limitations to the neighboring databases can better restrict an adversary's capability of performing hypothesis testing to infer sensitive information of an individual record. Therefore, using the same privacy budget, Blowfish privacy with count/marginal constraints can provide stronger defenses against the adversary's hypothesis testing than dependent differential privacy and inferential privacy. Furthermore, these advanced variants of DP have explicitly taken the correlation among records into consideration, which are powerful in defending against adversaries who aim to utilize auxiliary knowledge to infer sensitive information (Section 5.2).

6.2 Quantitative Comparison of Different Privacy Metrics

For quantitative analysis, we obtain the main results for comparison of these privacy notions above in Theorem 7 below, and two propositions thereafter considering two special database scenarios (detailed proofs are deferred to the Appendix). Note that there is no general perturbation mechanism to achieve these advanced privacy notions. Therefore, it is difficult to numerically analyze the adversary' hypothesis testing over these privacy notions as we did in the DP setting (recall Sections 4, 5).

Theorem 7. Privacy Comparison Main Result: *Under the same performance of adversary's hypothesis testing (denoted as ht which can be PR-relation for instance), we have the following relationship for the privacy parameter ϵ used in different privacy notions.*

$$\begin{aligned} &\epsilon_{BP}(ht) \geq \epsilon_{DDP}(ht) = \epsilon_{IP}(ht) \geq \epsilon_{DP}(ht) \\ &\epsilon_{DDP}(ht) \leq 2\epsilon_{MP}(ht) \\ &\epsilon_{MIDP}(ht) \leq \epsilon_{MP}(ht) \end{aligned} \tag{25}$$

where the subscripts *BP, DDP, IP, DP, MP, MIDP* represent Blowfish privacy with count/marginal constraints, dependent differential privacy, inferential privacy, differential privacy, membership privacy, mutual-information differential privacy, respectively. Note that Blowfish privacy is a special subclass of the general Pufferfish framework that handles a set of deterministic constraints such as count/marginal constraints.

Theorem 7 states the relationship among values of ϵ in different privacy notions under the same level of

PR-relation that can be achieved by the adversary. A privacy notion with a smaller ϵ in Theorem 7 is weaker in restricting an adversary's capability of performing hypothesis testing to infer an individual record. Combining the qualitative analysis in Section 6.1 and the quantitative comparison shown in Theorem 7, we know that under the same level of PR-relation, a larger value of ϵ can be selected for a privacy notion with less restrictions in the definition of neighboring databases. Based on Theorem 7, we further compare these privacy notions under two special data distributions: 1) independent records and 2) independent and uniform records, as shown in the following propositions.

Proposition 1. Privacy Comparison under Independent Records: *If the individual records in the database are independent of each other, we have*

$$\epsilon_{BP}(ht) = \epsilon_{DDP}(ht) = \epsilon_{IP}(ht) = \epsilon_{DP}(ht)$$
$$\epsilon_{DDP}(ht) \leq 2\epsilon_{MP}(ht)$$
$$\epsilon_{MIDP}(ht) \leq \epsilon_{MP}(ht)$$

Proposition 2. Privacy Comparison under Independent and Uniform Records: *If the individual records in the database are independent of each other, and each record is uniformly distributed, we have*

$$\epsilon_{BP}(ht) = \epsilon_{DDP}(ht) = \epsilon_{IP}(ht) = \epsilon_{DP}(ht)$$
$$\epsilon_{DDP}(ht) \leq 2\epsilon_{MP}(ht)$$
$$\epsilon_{MIDP}(ht) \leq \epsilon_{MP}(ht)$$

7 Discussions, Limitations and Future Works

Differential privacy provides a stability condition to the perturbation mechanism towards changes to the input, and there are ways to interpret its semantic privacy guarantee, such as hypothesis testing [27–29] and plausible deniability [51][52]. In our work, we focus on leveraging hypothesis testing to provide a privacy interpretation for DP mechanisms, which has implicitly taken the plausible deniability of any individual records in the database into consideration (recall Section 3).

Our analysis focuses on the popular LPM-based DP mechanisms, based on which we illustrate how hypothesis testing can be used for the selection of privacy parameters. We have shown the generality of our approach by applying it to the *Gaussian* perturbation mechanism in Section 4.3. Investigating how to generalize our analysis to a broader range of privacy mechanisms and metrics such as the exponential mechanism, randomized response, local DP and geo-indistinguishability [64] could be interesting future directions.

In our work, we consider the adversary who aims to infer the presence/absence of any particular record (for unbounded DP) or the true value of a record (for bounded DP), which is the standard adversary considered in DP framework. In practice, the adversary may be more interested in some aggregate statistics of the record, for instance, whether the value of the record d_i is higher than a given value γ. Under this scenario, the two hypotheses of the adversary can be constructed as $h_0 : d_i > \gamma, \quad h_1 : d_i \leq \gamma$ and then similar analysis in Sections 4, 5 can be conducted for implementing hypothesis testing. We will study the hypothesis testing of these adversaries in the future.

Our analysis considers adversaries with accurate auxiliary information of the prior distribution and correlation across records/time of the input database. In practice, it can be challenging for defenders to have an accurate estimate of the adversary's auxiliary information. Therefore, investigating the capability of adversary's hypothesis testing with approximate auxiliary information could be another interesting future work.

Motivated by composition properties of DP [29, 50, 65], it is interesting to investigate the composability of our analysis across different privacy mechanisms and explore tighter composition properties under specific mechanisms similar to [29] in the future.

8 Conclusion

In this paper, we investigate the state-of-the-art statistical privacy frameworks (focusing on DP) from the perspective of hypothesis testing of the adversary. We rigorously analyze the capability of an adversary for inferring a particular record of the input data using hypothesis testing. Our analysis provides a useful and interpretable guideline for how to select the privacy parameter ϵ in DP, which is an important question for practitioners and researchers in the community. Our findings show that an adversary's auxiliary information – in the form of prior distribution of the database, and correlation across records and time – indeed influences the proper choice of ϵ. Finally, our work systematically compares several state-of-the-art privacy notions from the perspective of adversary's hypothesis testing and showcases their relationship with each other and with DP.

9 Acknowledgement

The authors would like to thank Esfandiar Mohammadi for shepherding the paper, the anonymous reviewers for their valuable feedback. This work is supported in part by the National Science Foundation (NSF) under the grant CNS-1553437 and CCF-1617286, an Army Research Office YIP Award, and faculty research awards from Google, Cisco, Intel, and IBM. This work is also partly sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] C. Dwork, "Differential privacy," in *Automata, languages and programming*, 2006.

[2] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*. ACM, 2012, pp. 77–88.

[3] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 1447–1458.

[4] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *The Network and Distributed System Security Symposium (NDSS)*, 2016.

[5] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: a unifying framework for privacy definitions," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013, pp. 889–900.

[6] A. Ghosh and R. Kleinberg, "Inferential privacy guarantees for differentially private mechanisms," *arXiv preprint arXiv:1603.01508*, 2016.

[7] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 43–54.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Springer Theory of cryptography*, 2006.

[9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[10] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *arXiv preprint arXiv:1603.01887*, 2016.

[11] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, 2008.

[12] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," *Journal of Privacy and Confidentiality*, 2010.

[13] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, 2011.

[14] A. Machanavajjhala, X. He, and M. Hay, "Differential privacy in the wild: A tutorial on current practices & open challenges," in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 1727–1730.

[15] S. P. Kasiviswanathan and A. Smith, "On the'semantics' of differential privacy: A bayesian formulation," *Journal of Privacy and Confidentiality*, vol. 6, no. 1, 2014.

[16] I. Mironov, "Renyi differential privacy," in *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*. IEEE, 2017, pp. 263–275.

[17] T. Chanyaswad, A. Dytso, H. V. Poor, and P. Mittal, "Mvg mechanism: Differential privacy under matrix-valued query," in *Proceedings of the 25nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018.

[18] S. Meiser and E. Mohammadi, "Tight on budget?: Tight bounds for r-fold approximate differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 247–264.

[19] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on macos 10.12," *arXiv preprint arXiv:1709.02753*, 2017.

[20] A. Haeberlen, B. C. Pierce, and A. Narayan, "Differential privacy under fire." in *USENIX Security Symposium*, 2011.

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

[22] R. Chen, B. C. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," vol. 23, no. 4. Springer-Verlag New York, Inc., 2014, pp. 653–676.

[23] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-iid dataset," *Information Forensics and Security, IEEE Transactions on*, 2013.

[24] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *Computer Security Foundations Symposium (CSF), 2014 IEEE 27th*. IEEE, 2014, pp. 398–410.

[25] S. Krehbiel, "Markets for database privacy," 2014.

[26] J. Lee and C. Clifton, "How much is enough? choosing ε for differential privacy," in *International Conference on Information Security*. Springer, 2011, pp. 325–340.

[27] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.

[28] R. Hall, A. Rinaldo, and L. Wasserman, "Differential privacy for functions and functional data," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 703–727, 2013.

[29] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037–4049, 2017.

[30] D. R. Anderson, K. P. Burnham, and W. L. Thompson, "Null hypothesis testing: problems, prevalence, and an alternative," *The journal of wildlife management*, pp. 912–923, 2000.

[31] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics*, vol. 4, pp. 2111–2245, 1994.

[32] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[33] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part i," *Biometrika*, pp. 175–240, 1928.

[34] C. J. van Rijsbergen, "Information retrieval," in *Butterworth-Heinemann Newton, MA, USA*, 1979.

[35] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE transactions on medical imaging*, vol. 1, no. 2, pp. 113–122, 1982.

[36] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 271–279, 1989.

[37] D. K. Merchant and G. L. Nemhauser, "Optimality conditions for a dynamic traffic assignment model," *Transportation Science*, vol. 12, no. 3, pp. 200–207, 1978.

[38] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[39] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Soc. Lond*, pp. 289–337, 1933.

[40] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, pp. 307–333, 1989.

[41] A. Satorra and W. E. Saris, "Power of the likelihood ratio test in covariance structure analysis," *Psychometrika*, vol. 50, no. 1, pp. 83–90, 1985.

[42] "Detection, decision, and hypothesis testing," http://web.mit.edu/gallager/www/papers/chap3.pdf.

[43] J. Lee and C. Clifton, "Differential identifiability," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1041–1049.

[44] R. Rogers, A. Roth, A. Smith, and O. Thakkar, "Max-information, differential privacy, and post-selection hypothesis testing," *arXiv preprint arXiv:1604.03924*, 2016.

[45] M. Gaboardi, H.-W. Lim, R. M. Rogers, and S. P. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," in *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR, 2016.

[46] Y. Wang, J. Lee, and D. Kifer, "Differentially private hypothesis testing, revisited," *ArXiv e-prints*, 2015.

[47] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 475–489.

[48] Y. Tsiounis and M. Yung, "On the security of elgamal based encryption," in *International Workshop on Public Key Cryptography*. Springer, 1998, pp. 117–134.

[49] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[50] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 265–273.

[51] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[52] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 481–492, 2017.

[53] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[54] B. Balle and Y.-X. Wang, "Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *International Conference on Machine Learning (ICML)*, 2018.

[55] D. Sommer, S. Meiser, and E. Mohammadi, "Privacy loss classes: The central limit theorem in differential privacy," *Proceedings on privacy enhancing technologies*, 2019.

[56] Q. Geng, W. Ding, R. Guo, and S. Kumar, "Optimal Noise-Adding Mechanism in Additive Differential Privacy," in *Proceedings of the 22th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[57] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*. ACM, 2015, pp. 747–762.

[58] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 1291–1306.

[59] X. Wu, T. Wu, M. Khan, Q. Ni, and W. Dou, "Game theory based correlated privacy preserving analysis in big data," *IEEE Transactions on Big Data*, 2017.

[60] M. C. Tschantz, S. Sen, and A. Datta, "Differential privacy as a causal property," *arXiv preprint arXiv:1710.05899*, 2017.

[61] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1298–1309.

[62] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy under temporal correlations," in *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*. IEEE, 2017, pp. 821–832.

[63] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[64] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013*

ACM SIGSAC conference on Computer & communications security. ACM, 2013, pp. 901–914.

[65] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

10 Appendix

10.1 Proof for Theorem 2

Proof. **Achieve the minimal P_{FA} under a given level of P_{TD}:** For a given false alarm rate P_{FA}, the maximal true detection rate P_{TD} can be achieved according to the Neyman-Pearson criterion (Definition 3). Furthermore, note that the maximal P_{TD} is not decreasing with the increasing of P_{FA}. Thus, under a given level of the true detection rate $P_{TD} = P(\hat{\mathcal{D}} = D|\mathcal{D} = D) = P(\hat{\mathcal{D}} = D, \mathcal{D} = D)/P(\mathcal{D} = D)$, the adversary implementing the Neyman-Pearson criterion can achieve the minimal false alarm rate $P_{FA} = P(\hat{\mathcal{D}} = D|\mathcal{D} = D') = P(\hat{\mathcal{D}} = D, \mathcal{D} = D')/P(\mathcal{D} = D')$. As a direct result of this, we obtain a minimal $P(\hat{\mathcal{D}} = D, \mathcal{D} = D')$ under a fixed $P(\hat{\mathcal{D}} = D, \mathcal{D} = D)$.

Achieve the maximal precision under a given level of recall: Since both P_{TD} and recall correspond to $P(\hat{\mathcal{D}} = D|\mathcal{D} = D)$, we know that a given level of recall is equivalent to the same level of P_{TD}. Therefore, under a given level of recall (i.e., P_{TD}), the precision can be computed as $P(\mathcal{D} = D|\hat{\mathcal{D}} = D) = P(\hat{\mathcal{D}} = D, \mathcal{D} = D)/\left(P(\hat{\mathcal{D}} = D, \mathcal{D} = D) + P(\hat{\mathcal{D}} = D, \mathcal{D} = D')\right)$, which is maximized under the Neyman-Pearson criterion (with a minimal $P(\hat{\mathcal{D}} = D, \mathcal{D} = D')$ under a fixed $P(\hat{\mathcal{D}} = D, \mathcal{D} = D)$).

Therefore, the Neyman-Pearson criterion can achieve the maximal precision under any given level of recall, thus characterizing the optimal adversary that can achieve the best PR-relation. □

10.2 Proof for Corollary 1

Proof. This optimality is generally applicable for adversaries implementing the Neyman-Pearson criterion, under any distribution of the input data. This is because the proof in Theorem 2 demonstrates the inherent relationship among these quantification metrics (i.e., the maximization of P_{TD} under a given level of P_{FA} is equivalent to maximizing *precision* under a given level of *recall*), regardless of the distribution of the data. Therefore, the optimality of Neyman-Pearson criterion (maximizing P_{TD} under a given level of P_{FA}) to achieve the best PR-relation holds for correlated records. where the correlation relationships are naturally incorporated in the likelihood ratio detection process of the Neyman-Pearson criterion. □

10.3 $F^*_{\beta score}$ for Unbounded DP

According to the definition of $F_{\beta score}$, we know that maximizing $F_{\beta score}$ is equivalent to minimizing $\frac{GSR+\beta^2}{RSR}$ since $F_{\beta score} = \frac{1}{\frac{1}{(1+\beta^2)precision} + \frac{\beta^2}{(1+\beta^2)recall}} = \frac{1+\beta^2}{1+\frac{GSR+\beta^2}{RSR}}$. Then, we define $f = \frac{GSR+\beta^2}{RSR}$ and analyze f by considering three different intervals of θ in $(-\infty, Q(D')], (Q(D'), Q(D)), [Q(D), +\infty)$, respectively.

1) For the first interval of $(-\infty, Q(D')]$, we have $f = \frac{GSR+\beta^2}{RSR} = \frac{1+\beta^2-0.5e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}}{1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}}$. Then, we take the derivative of f with respect to θ as

$$\begin{aligned}\frac{\partial f}{\partial \theta} &= \frac{-\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}(1+\beta^2-0.5e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}((1+\beta^2)e^{-\epsilon}-1)}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{26}$$

Therefore, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D')} > 0$, i.e., the function f increases monotonically with $\theta \in (-\infty, Q(D')]$, if $\epsilon < \log(1+\beta^2)$. Otherwise, it decreases monotonically.

2) For the second interval of $(Q(D'), Q(D))$, we have $f = \frac{GSR+\beta^2}{RSR} = \frac{0.5e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2}{1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}}$. We then compute the derivative of f with respect to θ as

$$\begin{aligned}\frac{\partial f}{\partial \theta} &= \frac{-\frac{0.5\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}(0.5e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2)}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5\epsilon}{\Delta Q}(e^{-\epsilon}-e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2 e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{27}$$

When $\theta = Q(D)$, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D)} = \frac{0.5\epsilon}{\Delta Q}(e^{-\epsilon} - e^{-\epsilon} + \beta^2) > 0$. When $\theta = Q(D')$, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D')} = \frac{0.5\epsilon}{\Delta Q}((1+\beta^2)e^{-\epsilon}-1)$. Therefore, we know that $\frac{\partial f}{\partial \theta}_{\theta=Q(D')} > 0$. i.e., the function f increases monotonically with $\epsilon \in [Q(D'), Q(D)]$, if $\epsilon < \log(1+\beta^2)$. Otherwise, it decreases and then increases thus there is

a mininum within this interval. To solve for this minimum, we set the derivative to 0, i.e., $e^{-\epsilon} - e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon} + \beta^2 e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon} = 0$ to obtain $\theta = \frac{\Delta Q}{\epsilon}\log\frac{-1+\sqrt{1+4\beta^2e^{\epsilon}}}{2\beta^2} + Q(D')$. The corresponding minimum value for $F_{\beta score}$ can be computed as $\frac{(1+\beta^2)(1+4\beta^2e^{\epsilon}-\sqrt{1+4\beta^2e^{\epsilon}})}{(1+\beta^2)(1+4\beta^2e^{\epsilon})-(1-\beta^2)\sqrt{1+4\beta^2e^{\epsilon}}}$.
3) For the third interval $[Q(D), +\infty)$, we have $f = \frac{GSR+\beta^2}{RSR} = \frac{0.5e^{\frac{-(\theta-Q(D')\epsilon)}{\Delta Q}}+\beta^2}{0.5e^{\frac{-(\theta-Q(D)\epsilon)}{\Delta Q}}}$. We then take the derivative of f with respect to θ as

$$\begin{aligned}\frac{\partial f}{\partial\theta} &= \frac{-\frac{0.25\epsilon^2}{\Delta Q}e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}(0.5e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2)}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5\beta^2\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{28}$$

Therefore, we know that f increases monotonically with $\theta \in [Q(D), +\infty)$. Combining all the three intervals 1)-3), we obtain the highest $F_{\beta score}$ as in Eq. 12.

10.4 Proof for Theorem 6

Proof. According to the Neyman-Pearson Lemma [39], the likelihood ratio test [40, 41] can be utilized to achieve the Neyman-Pearson criterion. For an adversary with access to the noisy scalar output $o = \mathcal{A}(\mathcal{D}) = Q(\mathcal{D}) + \mathcal{N}(2\log(1.25/\delta)\Delta Q/\epsilon)$, we can compute the likelihood ratio corresponding to the two hypotheses defined in Eq. 4 as

$$\begin{aligned}\Lambda(o) &= \frac{\mathcal{L}(o|h_1)}{\mathcal{L}(o|h_0)} = \frac{\mathbb{P}(A(D)=o)}{\mathbb{P}(A(D')=o)} \\ &= \frac{\frac{1}{\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon}}\exp\left(-\frac{(o-Q(D))^2}{4\log(1.25/\delta)\Delta Q/\epsilon}\right)}{\frac{1}{\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon}}\exp\left(-\frac{(o-Q(D'))^2}{4\log(1.25/\delta)\Delta Q/\epsilon}\right)} \\ &= \exp\left(\frac{\Delta Q(2o-Q(D)-Q(D'))}{4\log(1.25/\delta)\Delta Q/\epsilon}\right)\end{aligned} \tag{29}$$

Then, we can compute the false alarm rate α according to $1 - \int_{\theta}^{\infty}\mathbb{P}(A(D')=o)do = 1 - \int_{\theta}^{\infty}\frac{1}{\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon}}\ 1 - \exp\left(-\frac{(o-Q(D'))^2}{4\log(1.25/\delta)\Delta Q/\epsilon}\right)do$, which is $1 - \Phi(\frac{\theta-Q(D')}{\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon}})$, i.e., $\theta = \Phi^{-1}(1-\alpha)\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon} + Q(D')$, where $\Phi(\cdot)$ is the cumulative distribution probability (CDF) of the standard normal distribution. Then, the threshold λ for the likelihood ratio can be computed as $\exp\left(\frac{\Delta Q(2\theta-Q(D)-Q(D'))}{4\log(1.25/\delta)\Delta Q/\epsilon}\right)$ and the true detection rate can be computed according to $P_{TD} = 1 - \int_{\theta}^{\infty}\mathbb{P}(A(D)=o)do = 1 - \Phi(\frac{\theta-Q(D)}{\sqrt{2\log(1.25/\delta)\Delta Q/\epsilon}})$.

For a given false alarm rate α, we can uniquely determine the threshold of the likelihood ratio λ, the threshold of the output θ and the true detection rate P_{TD}. Since θ can be any possible value of the private query result, we know that the Neyman-Pearson criterion is equivalent to setting a threshold θ for the Gaussian mechanism which is of a simpler formulation. □

10.5 $F^*_{\beta score}$ under Auxiliary Information

10.5.1 $F^*_{\beta score}$ under Prior Distribution

Consider an adversary with access to the prior distribution of the input database. Since $F_{\beta score} = \frac{1}{\frac{1}{(1+\beta^2)precision}+\frac{\beta^2}{(1+\beta^2)recall}} = \frac{1+\beta^2}{1+\frac{(1-\rho_p)GSR+\beta^2}{RSR}}$, we know that maximizing $F_{\beta score}$ is equivalent to minimizing $\frac{(1-\rho_p)GSR+\beta^2}{RSR}$. Next, we define $f = \frac{(1-\rho_p)GSR+\beta^2}{RSR}$ and analyze its property under three intervals.
1) For the first interval of $(-\infty, Q(D')]$, we have $f = \frac{(1-\rho_p)GSR+\beta^2}{RSR} = \frac{1-\rho_p+\beta^2-0.5(1-\rho_p)e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}}{1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}}$. Next, we take the derivative of f as

$$\begin{aligned}\frac{\partial f}{\partial\theta} &= \frac{-\frac{0.5(1-\rho_p)\epsilon}{\Delta Q}e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}(1-\rho_p+\beta^2-0.5(1-\rho_p)e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D')}{\Delta Q}\epsilon}((1-\rho_p+\beta^2)e^{-\epsilon}-(1-\rho_p))}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{30}$$

Therefore, we have $\frac{\partial f}{\partial\theta}_{\theta=Q(D')} > 0$, i.e., the function f increases monotonically for $\theta \in (-\infty, Q(D')]$, if $\epsilon < \log(1+\frac{\beta^2}{1-\rho_p})$. Otherwise, it decreases monotonically.
2) For the second interval of $(Q(D'), Q(D))$, we have $f = \frac{(1-\rho_p)GSR+\beta^2}{RSR} = \frac{0.5(1-\rho_p)e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2}{1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}}$. Next, we take the derivative of f as

$$\begin{aligned}\frac{\partial f}{\partial\theta} &= \frac{-\frac{0.5(1-\rho_p)\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon}(0.5(1-\rho_p)e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2)}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5(1-\rho_p)\epsilon}{\Delta Q}(e^{-\epsilon}-e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\frac{\beta^2}{1-\rho_p}e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})}{(1-0.5e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{31}$$

For $\theta = Q(D)$, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D)} = \frac{0.5(1-\rho_p)\epsilon}{\Delta Q}(e^{-\epsilon} - e^{-\epsilon} + \frac{\beta^2}{1-\rho_p}) > 0$. For $\theta = Q(D')$, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D')} = \frac{0.5(1-\rho_p)\epsilon}{\Delta Q}((1 + \frac{\beta^2}{1-\rho_p})e^{-\epsilon} - 1)$. Therefore, we have $\frac{\partial f}{\partial \theta}_{\theta=Q(D')} > 0$, i.e., the function f increases monotonically for $\epsilon \in (Q(D'), Q(D))$, if $\epsilon < \log(1 + \frac{\beta^2}{1-\rho_p})$. Otherwise, it decreases and then increases thus there is a mininum within this interval. To solve for the mininum, we set the derivative to 0, i.e., $e^{-\epsilon} - e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon} + \frac{\beta^2}{1-\rho_p} \cdot e^{\frac{\theta-Q(D)}{\Delta Q}\epsilon} = 0$ to obtain $\theta = \frac{\Delta Q}{\epsilon}\log(\frac{1-\rho_p}{2\beta^2}(\sqrt{1+\frac{4\beta^2 e^\epsilon}{1-\rho_p}}-1)) + Q(D')$. The corresponding $F_{\beta score} = \frac{(1+\beta^2)(\sqrt{1+\frac{4\beta^2e^\epsilon}{1-\rho_p}}-1)}{(1+\beta^2)\sqrt{1+\frac{4\beta^2e^\epsilon}{1-\rho_p}}-1+\beta^2}$.

3) For the third interval of $[Q(D'), +\infty$, we have $f = \frac{(1-\rho_p)GSR+\beta^2}{RSR} = \frac{0.5(1-\rho_p)e^{\frac{-(\theta-Q(D')\epsilon)}{\Delta Q}}+\beta^2}{0.5e^{\frac{-(\theta-Q(D)\epsilon)}{\Delta Q}}}$. Then, we take the derivative of f as

$$\begin{aligned}\frac{\partial f}{\partial \theta} &= \frac{-\frac{0.25(1-\rho_p)\epsilon^2}{\Delta Q}e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &+ \frac{\frac{0.5\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}(0.5(1-\rho_p)e^{-\frac{\theta-Q(D')}{\Delta Q}\epsilon}+\beta^2)}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2} \\ &= \frac{\frac{0.5\beta^2\epsilon}{\Delta Q}e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon}}{(0.5e^{-\frac{\theta-Q(D)}{\Delta Q}\epsilon})^2}\end{aligned} \tag{32}$$

Therefore, f increases monotonically for $\theta \in (-\infty, Q(D')]$. Combining the analysis for all the three intervals, we can achieve $F^*_{\beta score}$ as in Eq. 12.

10.5.2 $F^*_{\beta score}$ under Record Correlation

The computation of the highest $F_{\beta score}$ for adversaries under record correlation is similar to that of adversaries with prior distribution. By comparing Eq. 17 and Eq. 20, we can also simply replace ρ_p in Eq. 18 with $\rho_p + \rho_c(2-\rho_p)$ to obtain Eq. 21.

10.5.3 $F^*_{\beta score}$ under Temporal Correlation

The computation of the highest $F_{\beta score}$ for adversaries under temporal correlation is similar to that of adversaries with prior distribution and record correlation. By comparing Eq. 17 with Eq. 23, we can also simply replace ρ_p in Eq. 18 with $\rho_p + (2-\rho_p)(\rho_c + \rho_t(1-\rho_c))$ to obtain Eq. 24.

10.6 Proof for Theorem 7

Proof. The Blowfish framework [3], which is a subclass of the Pufferfish framework, allows user to specify adversarial knowledge about the database in the form of deterministic policy constraints. In the presence of general deterministic constraints, pairs of neighboring databases can differ in any number of tuples. The neighboring databases of dependent differential privacy differ in $\mathcal{L}$ tuples caused by one direct modification of one tuple. Therefore, under the same performance achieved by the adversary's hypothesis testing (the same amount of perturbation), we know that $\epsilon_{DDP}(ht) \leq \epsilon_{BP}(ht)$.

From the definition of inferential privacy, we have $\max_{D,D'} \frac{P(\mathcal{D}=D|\mathcal{A}(\mathcal{D})=o)}{P(\mathcal{D}=D'|\mathcal{A}(\mathcal{D})=o)} \leq e^\epsilon \frac{P(\mathcal{D}=D)}{P(\mathcal{D}=D')} \iff \max_{D,D'} \frac{P(\mathcal{D}=D|\mathcal{A}(\mathcal{D})=o)}{P(\mathcal{D}=D)} \leq e^\epsilon \frac{P(\mathcal{D}=D'|\mathcal{A}(\mathcal{D})=o)}{P(\mathcal{D}=D')} \iff \max_{D,D'} \frac{P(\mathcal{D}=D,\mathcal{A}(\mathcal{D})=o)}{\mathcal{A}(\mathcal{D})=o),P(\mathcal{D}=D)} \leq e^\epsilon \frac{P(\mathcal{D}=D',\mathcal{A}(\mathcal{D})=o)}{\mathcal{A}(\mathcal{D})=o),P(\mathcal{D}=D')} \iff \max_{D,D'} P(\mathcal{A}(D)=o) \leq e^\epsilon P(\mathcal{A}(D')=o)$, which is equivalent to dependent differential privacy that requires $\max_{D,D'} \frac{P(\mathcal{A}(D)=o)}{P(\mathcal{A}(D')=o)} \leq e^\epsilon$. Therefore, we have $\epsilon_{DDP}(ht) = \epsilon_{IP}(ht)$ under the same hypothesis testing performance achieved by the adversary.

Since the neighboring databases in DP only differ in one tuple, we know that $\epsilon_{DP}(ht) \leq \epsilon_{DDP}(ht)$ under the same hypothesis testing achieved by the adversary.

Next, we consider membership privacy whose mathematical form is different from other privacy metrics. Membership privacy does not consider two neighboring databases, but only bounds the ratio between the posterior probability and the prior probability for any data record. Based on the definition of membership privacy and inferential privacy, we know that the membership privacy satisfying $\exp(-\epsilon) \leq \frac{P(\mathcal{D}=D|\mathcal{A}(\mathcal{D})=o)}{P(\mathcal{D}=D)} \leq \exp(\epsilon)$ would lead to $\max_{d_{i1},d_{i2}} \frac{P(D_i=d_{i1}|\mathcal{A}(\mathcal{D})=o)}{P(D_i=d_{i2}|\mathcal{A}(\mathcal{D})=o)} \leq \exp(2\epsilon)\frac{P(D_i=d_{i1})}{P(D_i=d_{i2})}$ in inferential privacy (and thus dependent differential privacy). That is to say, ϵ-membership privacy would lead to 2ϵ-dependent differential privacy. Therefore, we have $\epsilon_{DDP}(ht) \leq 2\epsilon_{MP}(ht)$ under the same performance of the adversary's hypothesis testing.

Furthermore, membership privacy considers the worst-case difference between posterior probability and prior probability, and the mutual information based differential privacy considers the average difference between posterior probability and prior probability. Therefore, we have that ϵ-membership privacy would lead to ϵ-mutual information based differential privacy. Under the same performance of the adversary's hypothesis testing, we have $\epsilon_{MIDP}(ht) \leq \epsilon_{MP}(ht)$. □