Information Theory of Data Privacy

Genqiang Wu[1,2], Xianyao Xia[2], and Yeping He[2]

[1] SIE, Lanzhou University of Finance and Economics, Lanzhou 730020, China
[2] Institute of Software Chinese Academy of Sciences, Beijing 100190, China
genqiang80@gmail.com, {xianyao,yeping}@nfs.iscas.ac.cn

Abstract. By combining Shannon's cryptography model with an assumption to the lower bound of adversaries' uncertainty to the queried dataset, we develop a secure Bayesian inference-based privacy model and then in some extent answer Dwork et al.'s question [1]: "why Bayesian risk factors are the right measure for privacy loss".
This model ensures an adversary can only obtain little information of each individual from the model's output if the adversary's uncertainty to the queried dataset is larger than the lower bound. Importantly, the assumption to the lower bound almost always holds, especially for big datasets. Furthermore, this model is flexible enough to balance privacy and utility: by using four parameters to characterize the assumption, there are many approaches to balance privacy and utility and to discuss the group privacy and the composition privacy properties of this model.

1 Introduction

Data privacy protection [2,3,4] studies how to query dataset while preserving the privacy of individuals whose sensitive information is contained in the dataset. The crux of this field is to find suitable privacy protection model which can provide better tradeoffs between privacy protection and data utility. Differential privacy model [5,6] is currently the most important and popular privacy protection model. Dwork [2] illustrated differential privacy as "differential privacy will ensure that the ability of an adversary to inflict harm (or good, for that matter)of any sort, to any set of peopleshould be essentially the same, independent of whether *any individual* opts in to, or opts out of, the dataset." This illustration can be explained as that differential privacy minimizes the increased risk to an individual's privacy incurred by joining (or leaving) the dataset of *the individual*. This implies that differential privacy seldom cares about the increased risk to the individual's privacy incurred by joining (or leaving) the dataset of other individuals, which is unreasonable since other individuals' data may also be related to the individual's privacy.[3] For example, to a dataset containing individuals'

---

[3] It seems that an individual can't legally claim a breach of his privacy if the breach is due to other individuals' data. However, this doesn't mean that the individual doesn't fear this kind of privacy breach.

genomic data [7,8], the joining of an old man's 100 descendants clearly increases the risk to the man's privacy. In this paper, we will formally analyze the influence of the other individuals' joining the dataset to an individual's privacy and propose our solution.

Our powerful tool to analyze the influence is derived from Shannon's perfect secrecy [9], whose computational complexity relaxation is the famous semantic security [10], one fundamental concept in cryptography. Specifically, the perfect secrecy ensures that outputs (or ciphertexts) of a crypto system contain no information about inputs (or plaintexts), i.e., no information about the inputs can be extracted by any adversary, and the semantic security implies that any information revealed cannot be extracted by the probabilistic polynomial time (PPT) adversaries [10,11][12, p. 476]. To discuss the privacy problems more precisely, let us first review Shannon's theory to cryptography.

In Shannon's theory [9,11,10], a cryptography model/system is defined as a set of (probabilistic) transformations of the plaintext universe into the ciphertext universe. Each particular transformation of the set corresponds to enciphering with a particular key. The transformations are supposed reversible so that unique deciphering is possible when the key is known [10,13]. For a plaintext X and a secret key K, let Y be the corresponding ciphertext. Consider X, K, Y as random variables, where the probability distributions of X, K, Y are the adversary's probabilities for the choices in question, and represent his knowledge of the situation. Then the mutual information $I(X;Y)$ [14] or the max-mutual information $I_\infty(X;Y)$, defined in Definition 3, will be a measure of information about X which the adversary obtains from Y. The perfect secrecy is defined as $I_\infty(X;Y) = 0$ and the semantic security is defined as $I_\infty(X;Y) = O(1/m^t)$ [10,11].

We now borrow the above Shannon's cryptography models to construct data privacy models. In a (data) privacy model/system there are $n \geq 1$ individuals $X_1, \ldots, X_n$. A dataset $x := \{x_1, \ldots, x_n\}$ is a (multi)set of records, where each x_i is an assignment of X_i. For a query $f : \mathcal{D} \to \mathcal{R}$ [15], a privacy model is defined as a set of (probabilistic) transformations of the set $\mathcal{D}$ of possible datasets into the set $\mathcal{R}$ of possible query outputs Y. Each particular transformation of the set is called a (privacy) mechanism. Note that, being different from the cryptography models, a mechanism does not need to be reversible since there is no deciphering step in the privacy models.[4] This implies that the data consumer and the adversary are indistinguishable in their ways to extract information contained in Y. For the query f and the dataset x, the output Y is a probabilistic approximation of $f(x)$, which implies that we can use the expected distortion between $f(x)$ and Y to measure data utility, whose formal definition is deferred until Section 3.2. Note

---

[4] Note that one difference between privacy models and cryptography models is that, for different outputs in cryptography, the secret key is unchanged and the plaintexts are changing, but that, for different outputs in privacy protection, the "plaintext", i.e., the dataset x, is unchanged and the implicit "secret keys" may be changing. Or equivalently, in privacy protection, one can consider the dataset as the "secret key" to encrypt the implicit "plaintexts".

that a differential privacy mechanism is a special kind of privacy mechanisms defined as above, which is formally defined in Definition 2.

Consider X_i, Y as random variables, whose probability distributions are the adversary's probabilities for the choices in question, and represent his knowledge of the situation. Then the max-mutual information $I_\infty(X_i; Y)$ will be the amount of information about the individual X_i which the adversary obtains from Y. Following the semantic security and the perfect secrecy, the setting $I_\infty(X_i; Y) \leq \epsilon$ with $\epsilon > 0$ would be a reasonable choice as a privacy concept. One needs to be mentioned is that the "perfect privacy" [16,17], i.e., the setting $I_\infty(X_i; Y) = 0$, is not practical since this will result in poor data utility even in the assumption of the PPT adversaries by the results in [6]. Due to technical reasons, the formal definition of the privacy concept is deferred until Section 3.

One may find an interesting thing that we seem to pick up the semantic security that Dwork et al. had claimed to be impractical to privacy problems [6,2][18, Section 2.2]. We stress that Dwork [6] mainly proves that the "perfect privacy", i.e., the setting $I_\infty(X_i; Y) = 0$, is impractical due to poor data utility (even in the assumption of the PPT adversary), but seldom claims that $I_\infty(X_i; Y) \leq \epsilon$ is impractical. In this paper, we will continue Dwork's work [6] to discuss whether $I_\infty(X_i; Y) \leq \epsilon$ is suitable to be a privacy concept, and accurately in what extent to be; that is, we will employ Shannon's theory to answer Dwork et al.'s question [1]: "why Bayesian risk factors are the right measure for privacy loss". We will also continue Dwork's work [6] to discuss the tight upper bound of $I_\infty(X_i; Y)$ for the differential privacy output Y, which is obviously important but is neglected by Dwork [6] and the related works [19,20]. In fact, we have the following result.

Corollary 1 (Corollary of Proposition 4). *The mechanism $\mathcal{M}$ satisfies* $\max_{i \in [n]} I_\infty(X_i; Y) \leq n\epsilon$ *for all $X := (X_1, \ldots, X_n)$ if and only if $\mathcal{M}$ satisfies*

$$\max_{x,y \in \mathcal{D}, r \in \mathcal{R}} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(y) = r]} := \max_{x,y \in \mathcal{D}, r \in \mathcal{R}} \frac{\Pr[Y = r | X = x]}{\Pr[Y = r | X = y]} \leq \exp(n\epsilon). \quad (1)$$

Note that (1) is implied when $\mathcal{M}$ satisfies ϵ-differential privacy by the group privacy property of differential privacy in Lemma 1. Therefore, Corollary 1 implies that ϵ-differential privacy mechanism $\mathcal{M}$ allows its output Y such that $I_\infty(X_i; Y) \approx n\epsilon$, which will disclose too much information about the individual X_i so long as the number n of individuals is large enough, and which is our main motivation. For the ϵ-differential privacy mechanism $\mathcal{M}$, one interesting thing in Corollary 1 is that the $n\epsilon$ in (1), which is intended to be the maximal amount of disclosed information, or in other words the privacy budget [1], to the group $X_1, \ldots, X_n$ of individuals by the theory of differential privacy, however, becomes the maximal amount of disclosed information to the individual X_i. We will show in Proposition 4 that this is due to the other individuals' data also contains information of the individual X_i.

One needs to be emphasized is that *it is reasonable to accept $I_\infty(X_i; Y) \leq \epsilon$ as one minimal requirement for any secure privacy mechanism.* The reason is

the same as that $I_\infty(X;Y) \approx 0$ is one mininal requirement for secure cryptography models since large $I_\infty(X;Y)$ must result in information disclosure of the plaintext X, which has been testified for more than 60 years.

Definition 1 (The Knowledge of an Adversary). *Let the random vector $X := (X_1, \ldots, X_n)$ denote the uncertainty of an adversary to the queried dataset. Then X or its probability distribution is called the knowledge of the adversary.*

Note that, before this paper, there have been many Bayesian inference-based privacy models, such as [17,21,19,20,22]. These models share a common feature: they all restrict adversaries' knowledges. Many results, such as those in [19,22] and Proposition 4 of this paper, show that this restriction is inevitable for better utility. Traditionally, it is direct to restrict adversaries to be PPT as in cryptography. However, the current studies in data privacy don't suggest this restriction since most current works in data privacy are not based on it [18,23,24,25]. On the other hand, the current works to restrict adversaries' knowledges are almost no discussion on what are reasonable assumptions [17,21,19,20,22]. Note that the main obstacle to adopt these privacy models is that these models put restrictions to adversaries' knowledges but can't provide the reasonability of these restrictions. In this paper, our restriction to adversaries' knowledges is shown in Assumption 1.

Assumption 1. *Let b be a positive constant. Then, for any one adversary's knowledge X, there must be $H(X) \geq b$, where $H(X)$ is the entropy of X.*

We have the following evidences to support the reasonability of the restriction.

1. The maximal entropy $\max_X H(X)$, in general, is *huger* in privacy models than in cryptography models. For example, to the AES-256 encryption model [13], the adversary only needs to recover the 256 bits secret key in order to recover the information contained in the output Y and therefore it is reasonable to assume that $H(X)$ can be very small or even zero since $H(X)$ is at most 256 bits. However, to the Netflix Prize dataset [26] in data privacy, the adversary, in principle, needs to recover the whole dataset in order to recover the information contained in the output Y[5] and therefore it is reasonable to assume that $H(X)$ is relatively large since the Netflix Prize dataset is large and then $\max_X H(X)$ is at least larger than $100,480,507$ bits, which is huge compared to 256 bits.[6]
2. The long tail phenomenon[7] implies that there are too much "outlier data" in big dataset, which increases the uncertainty $H(X)$.

---

[5] Note that there is no fixed "secret key" in data privacy protection models.

[6] The Netflix Prize dataset contains $100, 480, 507$ movie ratings and each rating has at least two choices [26].

[7] https://en.wikipedia.org/wiki/Long_tail

3. Someone may doubt of the assumption since there are too much background knowledge in data privacy protection compared to in cryptography. For example, to the Netflix Prize dataset [26], it is inevitable that there exists open data, such as the IMDb dataset, as the adversary's background knowledge. Our comment is that, when the dataset is large enough, such as the Netflix dataset, the background knowledge, such as the IMDb dataset, in general, can't have large part, such as over 50%, to be overlapped with the secret dataset. In fact, the Netflix Prize dataset has very small part to be overlapped with the IMDb dataset. Therefore, the entropy $H(X)$ is still large for big dataset even though the diversity of background knowledges.

4. Theoretically, a dataset can be completely recovered by querying the dataset too many times as noted in [27,28][18, Chapter 8]; that is, theoretically, the entropy $H(X)$ can be very small or even zero [9, p. 659]. However, if we restrict the query times[8] and assume the dataset is big enough, we can ensure $H(X)$ to be not too small.

Due to the above evidences, it would be reasonable to adopt Assumption 1 as a reasonable restriction to adversaries' knowledges. Notice that Assumption 1 can achieve the idea of "crowd-blending privacy" (but with a way different from [30,31]), where each individual's privacy is related to other individuals' data; that is, if some other individuals' data is kept private, then Assumption 1 holds, which in turn ensure $I(X_i; Y) \leq \epsilon$ to be holding.

1.1   Contribution and Outline

This paper aims to provide some "mathematical underpinnings of formal privacy notions" [1] and tries to answer "why Bayesian risk factors are the right measure for privacy loss" [1] by employing Shannon's cryptography model and Assumption 1. Our contributions focus on studying how to control $I_\infty(X_i; Y)$ and related quantities based on Assumption 1.

1. We introduce Assumption 1 into privacy models. Compared to the restrictions to adversaries' knowledges in [17,21,19,20,22], the restriction in Assumption 1 is much more reasonable and universally applicable, especially for big datasets.

2. Four parameters are developed to characterize Assumption 1, which makes it easy to control $I_\infty(X_i; Y)$ and to discuss utility. This part is our main contribution; many bounds of $I_\infty(X_i; Y)$ and of utility are obtained.

3. We formalize the group privacy, i.e., the privacy of a group of individuals, and the composition privacy, i.e., the privacy problem when multiple results are output, of the information privacy model. Several results are proved.

---

[8] The differential privacy model also needs to restrict the query times [18, Chapter 8] and a cryptography model also needs to change its secret key after a time of usage. Furthermore, it seems to be computationally impractical to query a much big dataset too many times [29].

The following part of this paper is organized as follows. Section 2 presents some preliminaries. Section 3 introduces the information privacy model and compares it with other privacy models. In Section 4 we discuss the tradeoffs between privacy and utility based on Assumption 1. Section 5 discusses how to preserve the privacy of a group of individuals. Section 6 discusses the privacy problem when multiple results are output. Section 7 gives other related works. Section 8 concludes the results.

2   Preliminaries

The notational conventions of this paper are summarized in Table 1, of which some are borrowed from information theory [14].

2.1   The Setting

This section provides mathematical settings of our model, where most materials contain many mathematical symbols and seem to be boring. However, we emphasize that these symbols are necessary to make the presentation clear and shorter. Therefore, the readers can skip these settings at a first reading and go back to consult them later where necessary.

Let the random variables $X_1, \ldots, X_n$ denote n individuals. Let $\mathcal{X}_i$ denote the record universe of X_i. The probability distribution of X_i denotes an adversary's knowledge about the individual X_i's record. A dataset is a collection (a multiset) of n records $x_1, \ldots, x_n$, where $x_i \in \mathcal{X}_i$ denotes the assignment of X_i. We differentiate a *record sequence* $(x_1, \ldots, x_n)$ from a *dataset* $\{x_1, \ldots, x_n\}$ the record sequence corresponds to: the former has order among the records but the later does not. The universe of record sequences $\mathcal{Z}$ is defined as $\mathcal{Z} = \{(x_1, \ldots, x_n) : x_i \in \mathcal{X}_i, i \in [n]\}$. The universe of datasets $\mathcal{D}$ is defined as $\mathcal{D} = \{\{x_1, \ldots, x_n\} : x_i \in \mathcal{X}_i, i \in [n]\}$. We remark that $\mathcal{D}$ is *not* a multiset, in which the same datasets are merged as one dataset. There may be multiple record sequences which correspond to a same dataset. We call the dataset $\{x_1, \ldots, x_n\}$ as *the dataset of the record sequence* $(x_1, \ldots, x_n)$. For a dataset $y \in \mathcal{D}$, let D^y denote the set of all record sequences corresponding to the same dataset y.

Set $X = (X_1, \ldots, X_n)$. Set $X_{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, $\mathcal{X}_{(i)} = \prod_{j \in [n] \setminus \{i\}} \mathcal{X}_j$ and $x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Let $F(x), x \in \mathcal{Z}$ denote the probability distribution of X. For each $y \in \mathcal{D}$, set

$$F(y) := \sum_{x \in D^y} F(x). \tag{2}$$

In this manner, X can also be considered as a $\mathcal{D}$-valued random variable with the probability distribution $F(y), y \in \mathcal{D}$. Let $\mathbb{P}$ denote the universe of probability distributions over $\mathcal{Z}$ (or over $\mathcal{D}$). Note that, by letting all adversaries' knowledges be derived from a subset Δ of $\mathbb{P}$, we achieve a restriction to adversaries'

knowledges. If the probability distribution of the random variable X is within Δ, we say that X is in Δ, denoted as $X \in \Delta$.

For a query function f, let $\mathcal{R} \supseteq \{f(x) : x \in \mathcal{D}\}$ denote a set including all possible query results. Let $\mathbb{P}(\mathcal{R})$ denote the set of all the probability distributions on $\mathcal{R}$. A *mechanism* $\mathcal{M}$ takes a record sequence $x \in \mathcal{Z}$ as input and outputs a random variable $\mathcal{M}(x)$ valued in $\mathcal{R}$. Let Y be the random variable denoting the adversary's observation about the output. In this manner, for $x \in \mathcal{Z}$ and $r \in \mathcal{R}$, we set

$$\Pr[\mathcal{M}(x) = r] := \Pr[Y = r | X = x]. \tag{3}$$

In this paper, we abuse the notation $\mathcal{M}(x)$ as either denoting a probability distribution in $\mathbb{P}(\mathcal{R})$ or denoting a random variable following the probability distribution. Furthermore, for any $x \in \mathcal{D}$, set $\mathcal{M}(y) \equiv \mathcal{M}(z)$ for any two $y, z \in D^x$. Therefore, for a dataset $x \in \mathcal{D}$, we set $\mathcal{M}(x) := \mathcal{M}(z)$ for $z \in D^x$.

In this paper, we append an empty record, denoted as $\bot$, to each $\mathcal{X}_i$. In this setting, if $x_i = \bot$, it means that the individual X_i does not generate record in the dataset x. Let $\mathcal{X} = \cup_{i=1}^n \mathcal{X}_i \setminus \{\bot\}$. For a dataset $x \in \mathcal{D}$, we use the histogram representation $x \in \bar{\mathbb{N}}^{|\mathcal{X}|}$ to denote the dataset x, where the ith entry of x represents the number of elements in x of type $i \in \mathcal{X}$ [32,18,33]. Two datasets $x, y \in \mathcal{D}$ are said to be *neighbors (or neighboring datasets) of distance* k if $\|x - y\|_1 = k$. If $k = 1$, x, y are said to be *neighbors (or neighboring datasets)*. Two record sequences $x, x' \in \mathcal{Z}$ are said to be *neighbors (or neighboring record sequences)* if their corresponding datasets are neighbors.

For notational simplicity, in the following of this paper, we assume $\mathcal{Z}$ and $\mathcal{D}$ are both *discrete*.

2.2   Differential Privacy

Differential privacy characterizes the changes of outputs when one's record in a dataset is changed. The later changing is captured by the notion of the neighboring datasets.

Definition 2 (ϵ-Differential Privacy [5,6,18]). *Let the notations be as in Section 2.1. A mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{P}(\mathcal{R})$ gives ϵ-differential privacy if*

$$\max_{x,x' \in \mathcal{D}, r \in \mathcal{R} : \|x - x'\|_1 = 1} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]} \leq \exp(\epsilon), \tag{4}$$

where $\Pr[\mathcal{M}(x) = r] := \Pr[Y = r | X = x]$.

Note that Definition 2 is the same as those in [34,35], and is also equivalent to the definition of differential privacy in [5,6,18].

Differential privacy has group privacy property, which ensures that the strength of the privacy guarantee drops linearly with the size of the group of individuals.

Lemma 1 (Group Privacy [18]). *Let $\mathcal{M}$ be an ϵ-differentially private mechanism. Then*

$$\max_{x,y\in\mathcal{D},r\in\mathcal{R}:\|x-y\|_1\le s}\frac{\Pr[\mathcal{M}(x)=r]}{\Pr[\mathcal{M}(y)=r]}\le\exp(s\epsilon). \tag{5}$$

The composition privacy of differential privacy implies that the strength of the privacy guarantee drops in a controllable way when the number of outputs about a dataset raises.

Lemma 2 (Composition Privacy [18]). *Let the mechanism $\mathcal{M}^i$ satisfy ϵ_i-differential privacy on $\mathcal{R}^i$ for $i \in \{1,\ldots,s\}$. Then the composition mechanism $\mathcal{M}$, defined as $\mathcal{M}(x) = (\mathcal{M}^1(x),\ldots,\mathcal{M}^s(x))$, $x \in \mathcal{D}$, satisfies $\sum_{i=1}^s \epsilon_i$-differential privacy on the cartesian set $\prod_{i=1}^s \mathcal{R}^i$.*

2.3   Other Materials

Lemma 3. *Let $g(x) = \frac{a_0+a_1x}{b_0+b_1x}, x \in \mathbb{R}$ and let $a_i \ge 0, b_i > 0$ for $i \in \{0,1\}$. If $\frac{a_0}{b_0} < \frac{a_1}{b_1}$, then $g(x)$ is increasing. Otherwise, if $\frac{a_0}{b_0} \ge \frac{a_1}{b_1}$, then $g(x)$ is decreasing.*

Proof. Note that the derivative of $g(x)$ is $g'(x) = \frac{a_1b_0-a_0b_1}{(b_0x+b_1)^2}$, by which the claims are immediate.   □

Definition 3 (Max-Mutual Information [36]). *The max-mutual information of the random variables X, Y is defined as*

$$I_\infty(X;Y) = \max_{x\in\mathcal{X},r\in\mathcal{R}} \log \frac{\Pr[X=x,Y=r]}{\Pr[X=x]\Pr[Y=r]} = \max_{x\in\mathcal{X},r\in\mathcal{R}} \log \frac{\Pr[X=x|Y=r]}{\Pr[X=x]}.$$

Lemma 4. *There is $I(X;Y) \le I_\infty(X;Y)$.*

Proof. By the definition of $I(X;Y)$ [14], there is

$$I(X;Y) = \sum_{x\in\mathcal{X},r\in\mathcal{R}} \Pr[X=x,Y=r] \log \frac{\Pr[X=x,Y=r]}{\Pr[X=x]\Pr[Y=r]} \tag{6}$$

$$\le \sum_{x\in\mathcal{X},r\in\mathcal{R}} \Pr[X=x,Y=r]I_\infty(X;Y) = I_\infty(X;Y). \tag{7}$$

The claim is proved.   □

3   The Model of Information Privacy

Now it's time to give the formal definition of privacy concept. As discussed in Section 1, our privacy concept is to limit the amount of information of each individual X_i obtained by the adversary from the output Y, i.e., control the value of the max-mutual information $I_\infty(X_i;Y)$ or the mutual information $I(X;Y)$. For mathematical convenience, we only consider how to control the quantity $I_\infty(X_i;Y)$ in this paper. We formalize the discussions in Section 1 as the following definition.

Definition 4 (ϵ-Information Privacy). *Let $\Delta \subseteq \mathbb{P}$. Let $\mathcal{M} : \mathcal{D} \to \mathbb{P}(\mathcal{R})$ be a mechanism and let Y be the output random variable. The mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to Δ if for any $X \in \Delta$ and $i \in [n]$ there is*

$$\max_{x_i \in \mathcal{X}_i, r \in \mathcal{R}} \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']} \leq \exp(\epsilon).$$
(8)

Note that the inequality (8) is equivalent to

$$I_\infty(X_i; Y) \leq \epsilon \tag{9}$$

since

$$\frac{\Pr[X_i = x_i, Y = r]}{\Pr[X_i = x_i] \Pr[Y = r]}$$
$$= \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']}.$$
(10)

The parameter Δ in the above definition is used to model adversaries' knowledges. In this paper, we mainly set Δ to be

$$\Delta_b := \{X \in \mathbb{P} : H(X) \geq b\}, \tag{11}$$

which will be discussed in Section 4.

In information theory, the relative entropy is used to measure the distance between two probability distributions and the mutual information is used to measure the amount of information that one random variable contains about another random variable [14]. The relative entropy of $(X_i | Y = r)$ and X_i, denoted as $D((X_i | Y = r) \| X_i)$, and the mutual information of X_i and Y, i.e., $I(X_i; Y)$, have the following results.

Proposition 1. *Let the mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to X and let Y be its output random variable. We have*

$$\max_{r \in \mathcal{R}} D((X_i | Y = r) \| X_i) \leq \epsilon \quad and \quad I(X_i; Y) \leq \epsilon, \tag{12}$$

for $i \in [n]$.

Proof. The proof is direct and is omitted here. $\square$

Note that, as Definition 4, we can also define the ϵ-*relative entropy privacy*, i.e., $\max_r D((X_i | Y = r) \| X_i) \leq \epsilon$, and the ϵ-*mutual information privacy*, i.e., $I(X_i; Y) \leq \epsilon$. Furthermore, the paper [37] proposes a privacy concept called ϵ-*inferential privacy*, i.e.,

$$\max_{x_i, x_i', r} \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']} \leq \exp(\epsilon). \tag{13}$$

Note also that the inequalities (1) and (2) in [19] are essentially equivalent to the inequality (13). We now discuss the relations among the above three privacy concepts and the ϵ-information privacy. There are the following results.

Proposition 2. *We have the following relation among the privacy concepts: ϵ-inferential privacy $\Rightarrow_a$ ϵ-information privacy $\Rightarrow_b$ ϵ-relative entropy privacy $\Rightarrow_c$ ϵ-mutual information privacy.*

Proof. The claim $\Rightarrow_a$ is due to the inequality

$$\max_{x_i'} \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']} \geq$$

$$\frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']}.$$

The claim $\Rightarrow_b$ is due to Proposition 1. The claim $\Rightarrow_c$ is due to the equation

$$I(X_i; Y) = \sum_{r \in \mathcal{R}} \Pr[Y = r] D((X_i | Y = r) \| X_i).$$

The claims are proved. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Proposition 2 shows that the four privacy concepts, ϵ-inferential privacy, ϵ-information privacy, ϵ-relative entropy privacy and ϵ-mutual information privacy, are in decreasing order in terms of their strength to protect privacy. One can choose any one of the four concepts as the privacy concept, of which the choosing criterion depends on the privacy level of demand.

Proposition 3 (Data-Processing Inequality/Post-Processing). *Assume the mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{P}(\mathcal{R})$ satisfies ϵ-information privacy with respect to Δ and let Y be its output random variable. Let $Z = g(Y)$ and let $\mathcal{R}' = \{g(r) : r \in \mathcal{R}\}$. Then the composed mechanism $g \circ \mathcal{M} : \mathcal{D} \to \mathbb{P}(\mathcal{R}')$ satisfies ϵ-mutual information privacy with respect to Δ, where $g \circ \mathcal{M}(x) := g(\mathcal{M}(x))$ for $x \in \mathcal{D}$.*[9]

Proof. Recall that ϵ-mutual information privacy of $\mathcal{M}$ is implied by its ϵ-information privacy by Proposition 2. Then the claim is a direct corollary of the data-processing inequality in [14, Theorem 2.8.1]. $\qquad\qquad\qquad\square$

It is direct to define the personalized information privacy as the personalized differential privacy [38].

Definition 5 (ϵ-Personalized Information Privacy). *The mechanism $\mathcal{M}$ satisfies ϵ-personalized information privacy with respect to Δ if, for each $X \in \Delta$ and each $i \in [n]$, there is*

$$I_\infty(X_i; Y) \leq \epsilon_i, \tag{14}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.

---

[9] Currently, we can't strengthen the result to be ϵ-information privacy since we can't prove the data-processing inequality to the max-mutual information I_∞.

3.1   Comment on Parameter Setting

In this section we consider how to set the parameter $\mathcal{D}$ (or $\mathcal{Z}$) of the information privacy model. The setting of the parameter Δ is deferred to Section 4.

One needs to be emphasized is that the dataset universe $\mathcal{D}$ (or the record sequence universe $\mathcal{Z}$) should be set carefully since $\mathcal{D}$ itself may leak individuals' privacy and result in tracing attacks [28]. In order to see the above result clearly, we consider the query function $f(x) = x$, $x \in \mathcal{D}$ as an example, which can be considered as the abstraction of data publishing function [39,40,41]. Note that the codomain of f is $\mathcal{R} = \{f(x) : x \in \mathcal{D}\} = \mathcal{D}$. Both of the differential privacy model and the information privacy model employ randomized techniques to protect privacy: When the real dataset is x, in order to preserve privacy, a privacy mechanism first samples a dataset $y \in \mathcal{D}$ (according to a probability distribution) and then outputs $f(y) \in \mathcal{R}$ as the final query result of f. Or equivalently, the privacy mechanism directly samples a value r from the codomain $\mathcal{R}$ of f as the final query result. The major difference of the two models is that the probability distributions used to sample y or r are different. Assume that the individual X_i's record universe $\mathcal{X}_i$ has no overlapped record with all other individuals' record universes. Then, finding a record $r_i \in \mathcal{X}_i$ within an output dataset x would strongly conclude the participation of the individual X_i, which obviously is a successful tracing attack. Therefore, we should set appropriate $\mathcal{D}$ and therefore appropriate $\mathcal{X}_i$ for $i \in [n]$ such that the set $\mathcal{D}$ itself does not leak the participation of an individual. The privacy-oriented (but less utility-oriented) setting is to set $\mathcal{X}_i = \mathcal{X}$ for all $i \in [n]$ as in [18, p. 227].

3.2   Utility Measure

For the query f and the dataset universe $\mathcal{D}$, let the set $\mathcal{R} \supseteq \{f(x) : x \in \mathcal{D}\}$. We equip a metric d over the set $\mathcal{R}$ [15]. That is, the parity $(\mathcal{R}, d)$ is a metric space. Note that the output $\mathcal{M}(x)$ of the mechanism $\mathcal{M}$ is a probabilistic approximation of $f(x)$. Therefore,

for two datasets x, y, if $\|x-y\|_1$ is large, the distance of the outputs
$\mathcal{M}(x), \mathcal{M}(y)$ being small would result in poor data utility.

In most parts of this paper, the above utility measuring method can be used to measure the utility of mechanisms. However, for the completeness of this paper, we will present the formal definition of utility measure in (15). Note that the two utility measure methods are consistent since the former will result in $\mathcal{M}(x), x \in \mathcal{D}$ be more similar with the uniform probability distribution on $\mathcal{R}$, which obviously raises the distortion of $d(f(X), Y)$.

Let $F_o(x), x \in \mathcal{D}$ denote the occurring probability distribution of the individuals X. Then the utility of the mechanism $\mathcal{M}$ is measured by the expected value of the distortion $d(f(X), Y)$, i.e.,

$$
\begin{aligned}
\mathbb{E}[d(f(X), Y)] &= \sum_{x \in \mathcal{D}, r \in \mathcal{R}} \Pr[X = x, Y = r] d(f(x), r) \\
&= \sum_{x \in \mathcal{D}, r \in \mathcal{R}} F_o(x) \Pr[\mathcal{M}(x) = r] d(f(x), r).
\end{aligned}
\tag{15}
$$

We stress that $F_o(x), x \in \mathcal{D}$ is different from the probability distribution $F(x), x \in \mathcal{D}$ defined in (2), where the former is the factual occurring probabilities of datasets but the later denotes the knowledge of the adversary to datasets.

The third quantity to measure the utility is $I(X;Y)$ or $I_\infty(X;Y)$, which is used to measure the information of the individuals X contained in the output Y. Note that large $I(X;Y)$ implies better utility of the mechanism since the output Y contains more information about X that the mining or learning algorithms can mine or learn.

3.3   Some Related Works

One motivation of this paper is to solve the weakness of the differential privacy model [5,6] as shown in Corollary 1, which implies that the differential privacy model allows $I_\infty(X_i;Y)$ to be very large. Corollary 2, which is also appeared in [19,22], shows that the differential privacy model is equivalent to the information privacy model with respect to $\mathbb{P}_1$. Note that the setting $\mathbb{P}_1$ is obviously less reasonable than the setting (11). Therefore, the information privacy model with respect to (11) is more reasonable than the differential privacy model.

As noted in Section 1, the models in [17,21,19,20,22] and the information privacy model all are the Bayesian inference-based models and restrict adversaries' knowledges; that is, they all employ a subset of $\mathbb{P}$, like Δ in this paper, to model adversaries' knowledges. The advantage of these models and the restrictions is clear: powerful both to model privacy problems and to balance privacy and utility. However, the disadvantage is also large: the restrictions seem to be unreasonable since there are many examples, where making such a restriction may quickly lead to a disastrous breach of privacy. We imagine that the first impressions of most readers to these privacy models in [17,21,19,20,22] are similar with ours: Compared to conciseness of the differential privacy model, these privacy models set too many kinds of Δ's but none of these settings seems to be reasonable, which makes it hard to adopt these models. However, the rigorous analysis of the privacy problems by using Shannon's cryptography theory as in Section 1 makes us revisit these models, which results in the introduction of the parameter Δ into the information privacy model. Of course, we also face the problem of how to find a reasonable Δ. Assumption 1 is our solution and the evidences in Section 1 show that it is reasonable, especially for big datasets. In the following sections of this paper we will present our results based on Assumption 1.

Furthermore, the papers [17,19,42,43] discuss the impact of previously released data or query results, called constraints, to the privacy guarantee. The information privacy model treat these constraints by using Assumption 1; this is, these constraints can be summarized as the adversary's knowledge to the queried dataset, and if these constraints can't result in the adversary's knowledge go out of the set Δ in (11), then we can ensure the adversary can only obtain little information of each individual. Note that the above treatment to the constraints is similarly with the semantic security model in cryptography.

The papers [44,45,34] employ either $I(X;Y) \leq \epsilon$ or $I_\infty(X;Y) \leq \epsilon$ to define privacy concepts. We stress that both of the above two inequalities will result in poor data utility. The reason is that $I(X;Y)$ or $I_\infty(X;Y)$ is just the amount of information of X contained in Y that the data consumer needs to mine since the data consumer is also a special kind of adversaries. In contrast, the inequalities $I(X_i;Y) \leq \epsilon$, $I_\infty(X_i;Y) \leq \epsilon$ only restrict the information disclosure of each individual X_i, which, in general, allows the quantities $I(X;Y)$ or $I_\infty(X;Y)$ to be large enough, so long as the number of individuals n is large enough.

4   Privacy-Utility Tradeoff for Big Dataset

In this section, we consider how to set the parameter Δ in Definition 4 in order to give appropriate privacy-utility tradeoffs, where Δ denotes adversaries' knowledges. As noted in Section 3, the setting in (11) is a reasonable restriction to adversaries' knowledges. Before discussing the information privacy model based on this setting, we first discuss why we must restrict adversaries' knowledges. The following results show that the setting $\Delta = \mathbb{P}$ will result in poor utility.

Proposition 4. *The following three conditions are equivalent:*

1. $\max_{i \in [n]} I_\infty(X_i;Y) \leq \epsilon$ *with respect to* $\Delta = \mathbb{P}$.
2. $I_\infty(X;Y) \leq \epsilon$ *with respect to* $\Delta = \mathbb{P}$.
3. $\max_{r \in \mathcal{R}, x, x' \in \mathcal{D}} \frac{\Pr[\mathcal{M}(x)=r]}{\Pr[\mathcal{M}(x')=r]} \leq \exp(\epsilon)$.

Proof. The equivalence between the claim 1 and the claim 3 is due to

$$
\frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)} | X_i = x_i']}
$$
$$
\leq \max_{(x_i', x_{(i)}') \in \mathcal{Z}} \frac{\Pr[\mathcal{M}(x_i, x_{(i)}) = r]}{\Pr[\mathcal{M}(x_i', x_{(i)}') = r]},
$$
(16)

with equality when the record sequence $(x_i', x_{(i)}') \in \mathcal{Z}$ satisfies $\Pr[X_i = x_i'] = 1$ and $\Pr[X_{(i)} = x_{(i)}' | X_i = x_i'] = 1$, where $(x_i', x_{(i)}')$ is just the record sequence satisfying the above maximality.

The equivalence between the claim 2 and the claim 3 is due to

$$
\frac{\Pr[Y = r, X = x]}{\Pr[Y = r] \Pr[X = x]} = \frac{\Pr[\mathcal{M}(x) = r]}{\sum_{x' \in \mathcal{Z}} \Pr[\mathcal{M}(x') = r] \Pr[X = x']} \leq \max_{x' \in \mathcal{Z}} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]},
$$
(17)

with equality when the record sequence $x' \in \mathcal{Z}$ satisfies $\Pr[X = x'] = 1$, where x' is just the record sequence satisfying the above maximality.

The proof is complete. $\square$

The claim 2 of Proposition 4 shows that ϵ-information privacy with respect to $\mathbb{P}$ will result in poor utility since $I_\infty(X;Y) \leq \epsilon$ but $I_\infty(X;Y)$ denotes the information of X contained in Y, which is just the information the utility needs. Note also that the claim 3 of Proposition 4 shows that ϵ-information privacy with respect to $\mathbb{P}$ will result in two datasets even with distance n must have similar outputs, which obviously results in poor utility. Therefore, it is needed to restrict adversaries' knowledges for better utility.

Now we discuss how to control the quantity $I_\infty(X_i;Y)$ with respect to (11). We first formalize the reasons which make Assumption 1 hold. Note that

$$H(X) \leq \sum_{i=1}^{n} H(X_i) \leq \sum_{i=1}^{n} \log|\mathcal{X}_i|, \tag{18}$$

with equality to the first inequality if and only if $X_1, \ldots, X_n$ are independent, and with equality to the second inequality if and only if each X_i has uniform distribution over $\mathcal{X}_i$ [14]. Therefore, there are mainly two reasons which make $H(X) \geq b$:

1. The random variables $X_1, \ldots, X_n$ are not strongly dependent.
2. There exist some X_i's with $H(X_i) > 0$.

Traditionally, we can use the mutual information $I(X_i; X_{(i)})$ and the entropy $H(X_i)$ to characterize the above two reasons, respectively. However, for mathematical convenience, we develop four parameters to characterize them:

1. Use the parameter k to denote the maximal number of dependent random variables in X.
2. Use the parameter δ to denote the maximal dependent extent among the random variables X.
3. Use the parameter ℓ to denote the maximal number of random variables in X with $H(X_i) > 0$.
4. Use the parameter τ to characterize the minimal entropies of the above ℓ random variables.

Subsequently, also for mathematical convenience, we will approximate the set Δ_b in (11) with a set $_\ell^\tau \mathbb{P}_k^\delta$, which is parameterized by the four parameters k, δ, ℓ, τ and will be defined later; that is,

$$\Delta_b = \{X \in \mathbb{P} : H(X) \geq b\} \approx {}_\ell^\tau \mathbb{P}_k^\delta. \tag{19}$$

In the following parts of this section, we will explicitly define k, δ, ℓ, τ and then $_\ell^\tau \mathbb{P}_k^\delta$ and discuss how to control $I_\infty(X_i;Y)$ based on them.

4.1   The Parameter k

Recall that the parameter k denotes the maximal number of dependent random variables in X, which is mainly motivated by the group privacy method in [46] to deal with the dependent problem and by the need to explain differential privacy

using the information privacy model. Let $\mathbb{P}_k$ be the largest subset of $\mathbb{P}$ such that, for any $X \in \mathbb{P}_k$, the maximal number of dependent random variables within X is at most k, where $1 \le k \le n$. Formally, let

$$\mathbb{P}_k = \left\{ X \in \mathbb{P} : \Pr[X = x] = \Pr[X_I = x_I] \prod_{i \in [n]-I} \Pr[X_i = x_i], \forall x \in \mathcal{Z} \right\} \quad (20)$$

where $I \in [n]$ with $|I| \le k$, each $x_I \in \mathcal{X}_I$, each $x_i \in \mathcal{X}_i$ for $i \in [n]-I$. Note that, in this manner, $\mathbb{P}$ equals $\mathbb{P}_n$ and $\mathbb{P}_1$ denotes the universe of probability distributions of the independent random variables $X_1, \ldots, X_n$. We have the following result.

Theorem 1. *The mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}_k$ if and only if $\mathcal{M}$ satisfies*

$$\max_{x,y \in \mathcal{D}, r \in \mathcal{R}: \|x-y\|_1 \le k} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(y) = r]} \le \exp(\epsilon). \quad (21)$$

Proof. Let $X = (X_i, \bar{X}_{(i)}, \tilde{X}_{(i)}) \in \mathbb{P}_k$, where $\bar{X}_{(i)}, \tilde{X}_{(i)}$ denote the random variables in X which are independent to and dependent to X_i, respectively. Let $\bar{x}_{(i)}, \tilde{x}_{(i)}$ and $\bar{\mathcal{X}}_{(i)}, \tilde{\mathcal{X}}_{(i)}$ denote one assignment and the record universe of $\bar{X}_{(i)}, \tilde{X}_{(i)}$, respectively.

"$\Leftarrow$" Assume the inequality (21) holds. For one $\bar{x}_{(i)}$, set $M_{\bar{x}_{(i)}} = \max_{x_i \in \mathcal{X}_i, \tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} \Pr[\mathcal{M}(x_i, \bar{x}_{(i)}, \tilde{x}_{(i)}) = r]$, $m_{\bar{x}_{(i)}} = \min_{x_i \in \mathcal{X}_i, \tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} \Pr[\mathcal{M}(x_i, \bar{x}_{(i)}, \tilde{x}_{(i)}) = r]$. We have

$$\frac{\Pr[X_i = x_i, Y = r]}{\Pr[X_i = x_i]\Pr[Y = r]}$$

$$= \frac{\sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} \sum_{\tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} \Pr[\mathcal{M}(x_i,\bar{x}_{(i)},\tilde{x}_{(i)})=r] \Pr[\bar{X}_{(i)}=\bar{x}_{(i)},\tilde{X}_{(i)}=\tilde{x}_{(i)}|X_i=x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i=x_i'] \sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} \sum_{\tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} \Pr[\mathcal{M}(x_i',\bar{x}_{(i)},\tilde{x}_{(i)})=r] \Pr[\bar{X}_{(i)}=\bar{x}_{(i)},\tilde{X}_{(i)}=\tilde{x}_{(i)}|X_i=x_i']}$$

$$\le \frac{\sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} \sum_{\tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} M_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)}=\bar{x}_{(i)},\tilde{X}_{(i)}=\tilde{x}_{(i)}|X_i=x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i=x_i'] \sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} \sum_{\tilde{x}_{(i)} \in \tilde{\mathcal{X}}_{(i)}} m_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)}=\bar{x}_{(i)},\tilde{X}_{(i)}=\tilde{x}_{(i)}|X_i=x_i']}$$

$$= \frac{\sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} M_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}|X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} m_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}|X_i = x_i']}$$

$$=_a \frac{\sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} M_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} m_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}]}$$

$$\le_b \frac{\sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} M_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{\bar{x}_{(i)} \in \bar{\mathcal{X}}_{(i)}} e^{-\epsilon} M_{\bar{x}_{(i)}} \Pr[\bar{X}_{(i)} = \bar{x}_{(i)}]}$$

$$= e^{\epsilon},$$

$$(22)$$

where $=_a$ is due to the independence between $\bar{X}_{(i)}$ and X_i, and $\le_b$ is due to the inequality (21) and that there are at most k random variables in $(X_i, \tilde{X}_{(i)})$.

"$\Rightarrow$" Assume $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}_k$. Without loss of generality, assume the two datasets $(\bar{x}_1, \bar{x}_{(1)}, \tilde{x}_{(1)}), (\hat{x}_1, \bar{x}_{(1)}, \hat{x}_{(1)}) \in \mathcal{D}$ of distance $\leq k$ and $\bar{r} \in \mathcal{R}$ satisfy

$$\max_{x,y \in \mathcal{D}, r \in \mathcal{R}: \|x-y\|_1 \leq k} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(y) = r]} = \frac{\Pr[\mathcal{M}(\bar{x}_1, \bar{x}_{(1)}, \tilde{x}_{(1)}) = \bar{r}]}{\Pr[\mathcal{M}(\hat{x}_1, \bar{x}_{(1)}, \hat{x}_{(1)}) = \bar{r}]}. \tag{23}$$

We construct the following probability distribution in $\mathbb{P}_k$. Set $\Pr[X_1 = \hat{x}_1] = 1, \Pr[\bar{X}_{(1)} = \bar{x}_{(1)}] = 1, \Pr[\tilde{X}_{(1)} = \tilde{x}_{(1)} | X_1 = \bar{x}_1] = 1, \Pr[\tilde{X}_{(1)} = \hat{x}_{(1)} | X_1 = \hat{x}_1] = 1$. Then

$$\frac{\Pr[\mathcal{M}(\bar{x}_1, \bar{x}_{(1)}, \tilde{x}_{(1)}) = \bar{r}]}{\Pr[\mathcal{M}(\hat{x}_1, \bar{x}_{(1)}, \hat{x}_{(1)}) = \bar{r}]} = \frac{\Pr[X_1 = \bar{x}_1, Y = \bar{r}]}{\Pr[X_1 = \bar{x}_1] \Pr[Y = \bar{r}]} \tag{24}$$

by the first two lines of the equation (22). Furthermore, since $\mathcal{M}$ satisfies ϵ-information privacy with respect $\mathbb{P}_k$, we have

$$\frac{\Pr[\mathcal{M}(\bar{x}_i, \bar{x}_{(i)}, \tilde{x}_{(i)}) = \bar{r}]}{\Pr[\mathcal{M}(\hat{x}_i, \bar{x}_{(i)}, \hat{x}_{(i)}) = \bar{r}]} \leq \exp(\epsilon), \tag{25}$$

which gives (21) by the equation (23).

The proof is complete. □

Note that $\mathbb{P}_n = \mathbb{P}$. There are the following corollaries for $\mathbb{P}_1$ and $\mathbb{P}$.

Corollary 2. *The mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}_1$ if and only if*

$$\max_{x,x' \in \mathcal{D}, r \in \mathcal{R}: \|x-x'\|_1 \leq 1} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]} \leq \exp(\epsilon), \tag{26}$$

and therefore if and only if $\mathcal{M}$ satisfies ϵ-differential privacy.

Corollary 3. *The mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}$ if and only if $\mathcal{M}$ satisfies*

$$\max_{x,y \in \mathcal{D}, r \in \mathcal{R}} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(y) = r]} \leq \exp(\epsilon). \tag{27}$$

Corollary 2, which is also appeared in [19] and is in some extent equivalent to [22, Theorem 4.5, Theorem 4.8], implies that the differential privacy model effectively controls $I_\infty(X_i; Y)$ when the adversary's knowledge X are independent random variables.

Corollary 3 is equivalent to [19, Theorem 3.1], which is also appeared in Proposition 4 and implies Corollary 1. It implies that the differential privacy model can't effectively control $I_\infty(X_i; Y)$ when the adversary's knowledge X are dependent random variables.

Notice that there is a drawback when using Theorem 1 to balance privacy and utility: hard to set the value of k. This is because of that small k will result in

bad privacy since, in general, this may result in the parameter b in Assumption 1 to be large, and then result in Assumption 1 doesn't hold, and that large k will obviously result in poor utility. Therefore, except the parameter k, there should be another one parameter δ to model the dependent extent among X, which is the task of Section 4.2.

4.2   The Parameter δ

The parameter δ denotes the dependent extent among X, which is mainly motivated by the "correlated sensitivity" in [47], the "dependence coefficient" in [48] and the "multiplicative influence matrix" in [37].

   The dependence among the individuals is popular. For example, the spreading of the Black Death in the 14th century[10] and the spreading of the SARS coronavirus in 2002-2003[11] (if without effective controlling) show that people all over the world are dependent. Furthermore, the small world phenomenon [49,50] also shows the dependence among people. However, the dependent extent of these relationships are low and therefore an adversary, in general, will have low dependent relationship knowledge. We now consider how to measure the dependent extent among X.

   Traditionally, it is appropriate to use $I(X_i; X_{(i)})$ to measure the dependent extent between X_i and $X_{(i)}$. However, for mathematical convenience, we develop a new quantity δ to measure it. Roughly speaking, $I(X_i; X_{(i)})$ uses $\log \frac{\Pr[X_{(i)}=x_{(i)}]}{\Pr[X_{(i)}=x_{(i)}|X_i=x_i]}$ but δ uses $\Pr[X_{(i)} = x_{(i)}] - \Pr[X_{(i)} = x_{(i)}|X_i = x_i]$ to measure it. Note that the independence among X ensures that, for each X_i, there are $\Pr[X_{(i)} = x_{(i)}] = \Pr[X_{(i)} = x_{(i)}|X_i = x_i]$ for all $x_i \in \mathcal{X}_i$ and all $x_{(i)} \in \mathcal{X}_{(i)}$. This implies that the weak dependence among X will result in $\Pr[X_{(i)} = x_{(i)}] \approx \Pr[X_{(i)} = x_{(i)}|X_i = x_i]$ for all $x_i \in \mathcal{X}_i$ and all $x_{(i)} \in \mathcal{X}_{(i)}$, and then result in

$$\Pr[X_{(i)} = x_{(i)}] \approx \min\left\{\Pr[X_{(i)} = x_{(i)}|X_i = x_i], \Pr[X_{(i)} = x_{(i)}|X_i = x_i']\right\}$$

for any two records $x_i, x_i' \in \mathcal{X}_i$ and all $x_{(i)} \in \mathcal{X}_{(i)}$. By setting

$$a_{x_{(i)},x_i,x_i'} = \min\left\{\Pr[X_{(i)} = x_{(i)}|X_i = x_i], \Pr[X_{(i)} = x_{(i)}|X_i = x_i']\right\} \quad (28)$$

we can therefore use

$$\max_{x_i,x_i'\in\mathcal{X}_i} \sum_{x_{(i)}\in\mathcal{X}_{(i)}} \left(\Pr[X_{(i)} = x_{(i)}] - a_{x_{(i)},x_i,x_i'}\right) = 1 - \min_{x_i,x_i'\in\mathcal{X}_i} \sum_{x_{(i)}\in\mathcal{X}_{(i)}} a_{x_{(i)},x_i,x_i'}$$
$$(29)$$

to measure the dependence extent between X_i and $X_{(i)}$. In this manner, we can use

$$\sigma := 1 - \min_{i\in[n]} \min_{x_i,x_i'\in\mathcal{X}_i} \sum_{x_{(i)}\in\mathcal{X}_{(i)}} a_{x_{(i)},x_i,x_i'} \quad (30)$$

---

[10] https://en.wikipedia.org/wiki/Black_Death
[11] https://en.wikipedia.org/wiki/SARS_coronavirus

to measure the dependence extent among X; that is, if σ is small (≈ 0), the dependence extent among X would be weak. Note that $0 \le \sigma \le 1$ since

$$0 \le \sum_{x_{(i)} \in \mathcal{X}_{(i)}} 0 \le \sum_{x_{(i)} \in \mathcal{X}_{(i)}} a_{x_{(i)}, x_i, x_i'} \le \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[X_{(i)} = x_{(i)} | X_i = x_i] = 1.$$

In the following part of this section, for notational simplicity, we set $a_{x_{(i)}} := a_{x_{(i)}, x_i, x_i'}$. Let

$$\mathbb{P}^\delta = \{X \in \mathbb{P} : \sigma \le \exp(\delta)\} \tag{31}$$

denote the set of probability distributions satisfying $\sigma \le \exp(\delta)$, where $\delta \in [-\infty, 0]$ and then $\exp(\delta) \in [0, 1]$. Then smaller δ implies that X are low dependent if $X \in \mathbb{P}^\delta$. Therefore, we can use δ to denote the dependence extent among X. We have the following results about $\Delta = \mathbb{P}^\delta$.

Theorem 2. *Assume the mechanism $\mathcal{M}$ satisfy ϵ/n-differential privacy. Then*

$$\max_{i \in [n]} I_\infty(X_i; Y) \le \exp\left(\frac{\epsilon}{n}\right) \times (1 - \exp(\delta)) + \exp(\epsilon) \times \exp(\delta), \tag{32}$$

for all $X \in \mathbb{P}^\delta$.

Proof. Let $X \in \mathbb{P}^\delta$. For any $i \in [n]$ and any $x_{(i)} \in \mathcal{X}_{(i)}$, set $b_{x_{(i)}} = \Pr[X_{(i)} = x_{(i)} | X_i = x_i] - a_{x_{(i)}}$ and $c_{x_{(i)}} = \Pr[X_{(i)} = x_{(i)} | X_i = x_i'] - a_{x_{(i)}}$. Set $M = \max_{x_i' \in \mathcal{X}_i, x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r]$, $m = \min_{x_i' \in \mathcal{X}_i, x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r]$. Set $\alpha = 1 - \sum_{x_{(i)} \in \mathcal{X}_{(i)}} a_{x_{(i)}}$. We have

$$\frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)} | X_i = x_i']}$$

$$= \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times a_{x_{(i)}} + \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times b_{x_{(i)}}}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \left\{ \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times a_{x_{(i)}} + \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times c_{x_{(i)}} \right\}}$$

$$\le \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times a_{x_{(i)}} + M \times \sum_{x_{(i)} \in \mathcal{X}_{(i)}} b_{x_{(i)}}}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \left\{ \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times a_{x_{(i)}} + m \times \sum_{x_{(i)} \in \mathcal{X}_{(i)}} c_{x_{(i)}} \right\}}$$

$$= \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times a_{x_{(i)}} + M \times \alpha}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \left\{ \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times a_{x_{(i)}} + m \times \alpha \right\}}$$

$$\le \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times a_{x_{(i)}} + M \times \alpha}{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} a_{x_{(i)}} \times \min_{x_i' \in \mathcal{X}_i} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] + m \times \alpha}$$

$$\le_a \frac{\exp(\epsilon/n) \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \min_{x_i' \in \mathcal{X}_i} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times a_{x_{(i)}} + M \times \alpha}{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \min_{x_i' \in \mathcal{X}_i} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times a_{x_{(i)}} + m \times \alpha}$$

$$\le_b \frac{\exp(\epsilon/n) \sum_{x_{(i)} \in \mathcal{X}_{(i)}} m \times a_{x_{(i)}} + m \exp(\epsilon) \times \alpha}{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} m \times a_{x_{(i)}} + m \times \alpha}$$

$$= \exp\left(\frac{\epsilon}{n}\right) \times \sum_{x_{(i)} \in \mathcal{X}_{(i)}} a_{x_{(i)}} + \exp(\epsilon) \times \alpha \le_c \exp\left(\frac{\epsilon}{n}\right) (1 - \exp(\delta)) + \exp(\epsilon) \times \exp(\delta)$$

where the inequality $\leq_a$ is due to the fact that $\mathcal{M}$ satisfies ϵ/n-differential privacy, the inequality $\leq_b$ is due to Lemma 3 and the group privacy property of $\mathcal{M}$, and the inequality $\leq_c$ is due to $X \in \mathbb{P}^\delta$.

The claim is proved.   □

Theorem 3. *Assume $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}^\delta$. Then*

$$\max_{\substack{x_{(i)},x'_{(i)},x''_{(i)} \in \mathcal{X}_{(i)} \\ r \in \mathcal{R}, x_i, x'_i \in \mathcal{X}_i}} \frac{\Pr[\mathcal{M}(x_i,x_{(i)})=r] \times \exp(\delta) + \Pr[\mathcal{M}(x_i,x''_{(i)})=r] \times (1-\exp(\delta))}{\Pr[\mathcal{M}(x'_i,x_{(i)})=r] \times \exp(\delta) + \Pr[\mathcal{M}(x'_i,x'_{(i)})=r] \times (1-\exp(\delta))} \leq \exp(\epsilon),$$
(33)

for $i \in [n]$.

Proof. Assume there exist $\bar{\imath} \in [n]$, $\bar{r} \in \mathcal{R}$, $\bar{x}_{\bar{\imath}}, \bar{x}'_{\bar{\imath}} \in \mathcal{X}_{\bar{\imath}}$ and $\bar{x}_{(\bar{\imath})}, \bar{x}'_{(\bar{\imath})}, \bar{x}''_{(\bar{\imath})} \in \mathcal{X}_{(\bar{\imath})}$ such that the left side of (33) equals

$$\frac{\Pr[\mathcal{M}(\bar{x}_{\bar{\imath}}, \bar{x}_{(\bar{\imath})}) = \bar{r}] \times \exp(\delta) + \Pr[\mathcal{M}(\bar{x}_{\bar{\imath}}, \bar{x}''_{(\bar{\imath})}) = \bar{r}] \times (1 - \exp(\delta))}{\Pr[\mathcal{M}(\bar{x}'_{\bar{\imath}}, \bar{x}_{(\bar{\imath})}) = \bar{r}] \times \exp(\delta) + \Pr[\mathcal{M}(\bar{x}'_{\bar{\imath}}, \bar{x}'_{(\bar{\imath})}) = \bar{r}] \times (1 - \exp(\delta))}.$$
(34)

We construct a probability distribution $X \in \mathbb{P}^\delta$ as follows. Set $\Pr[X_{\bar{\imath}} = \bar{x}'_{\bar{\imath}}] = 1$, $\Pr[X_{(\bar{\imath})} = \bar{x}_{(\bar{\imath})} | X_{\bar{\imath}} = \bar{x}'_{\bar{\imath}}] = \Pr[X_{(\bar{\imath})} = \bar{x}_{(\bar{\imath})} | X_{\bar{\imath}} = \bar{x}_{\bar{\imath}}] = \exp(\delta)$ and $\Pr[X_{(\bar{\imath})} = \bar{x}'_{(\bar{\imath})} | X_{\bar{\imath}} = \bar{x}'_{\bar{\imath}}] = \Pr[X_{(\bar{\imath})} = \bar{x}''_{(\bar{\imath})} | X_{\bar{\imath}} = \bar{x}_{\bar{\imath}}] = 1 - \exp(\delta)$. Then, by setting X to be the above probability distribution, we have that (34) equals

$$\frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)} | X_i = x_i]}{\sum_{x'_i \in \mathcal{X}_i} \Pr[X_i = x'_i] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x'_i, x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)} | X_i = x'_i]},$$

which ensures the inequality (33) by combining the ϵ-information privacy of $\mathcal{M}$.   □

Theorem 2 and Theorem 3 have very interesting connections with Corollary 2 and Corollary 3. Corollary 2 and Corollary 3 show the utilities of an ϵ-information privacy mechanism for the cases $\exp(\delta) = 0$ and $\exp(\delta) = 1$, respectively, whereas the left side of (33) is a (mediant-like[12]) linear combination of the left sides of (26) and (27) with the weight $\exp(\delta)$. Furthermore, since

$$\max_{x,x' \in \mathcal{D}, r \in \mathcal{R}} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]} \leq \exp\left(\frac{\epsilon}{n}\right)$$
(35)

is equivalent to $I_\infty(X_i; Y) \leq \epsilon/n$ with respect to $\mathbb{P}$ which is equivalent to $\exp(\delta) = 1$, and (26) is equivalent to $I_\infty(X_i; Y) \leq \epsilon$ with respect to $\mathbb{P}_1$ which is equivalent to $\exp(\delta) = 0$, the bound of $I_\infty(X_i; Y)$ in (36) is just the linear combination of the above two bounds of $I_\infty(X_i; Y)$ with the weight $\exp(\delta)$. Therefore, Theorem 2 and Theorem 3 provide a (in some extent) sufficient and necessary condition, which is a tradeoff between the sufficient and necessary conditions in Corollary 2 and Corollary 3 with the weight $\exp(\delta)$, and which shows how the parameter δ balances privacy and utility.

By combining Theorem 1 and Theorem 2, we have the following result.

---

[12] https://en.wikipedia.org/wiki/Mediant_(mathematics)

Corollary 4. *Assume the mechanism $\mathcal{M}$ satisfy ϵ/k-differential privacy. Then*

$$\max_{i \in [n]} I_\infty(X_i; Y) \leq \exp\left(\frac{\epsilon}{k}\right) \times (1 - \exp(\delta)) + \exp(\epsilon) \times \exp(\delta), \qquad (36)$$

for all $X \in \mathbb{P}_k^\delta$, where $\mathbb{P}_k^\delta = \mathbb{P}_k \cap \mathbb{P}^\delta$.

Proof. The proof of the theorem is the combination of Theorem 2 and the proof techniques of Theorem 1. □

4.3   The Parameters ℓ, τ

The parameters ℓ, τ are motivated partially by the parameters "k, δ" in [21, Definition 2.4] and partially by the need and the works, such as [51,30,19,20,22], to relax the differential privacy model to obtain better utility.

We now discuss how to relax the differential privacy model, from which the parameters ℓ, τ are derived. By Corollary 2, $\mathcal{M}$ satisfies ϵ-differential privacy if and only if $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}_1$. Note that $\mathbb{P}_1$ contains those probability distributions X such that $H(X_i) = 0$ for most or even all $i \in [n]$; that is, the adversary can know most or even every records in the dataset, which is a too strong assumption when the dataset is big enough as discussed in Section 1. Therefore, it is reasonable to assume that there exists a set $I \subset [n]$ of individuals such that $H(X_i) > 0$ for $i \in I$. Formally, set

$$_\ell^\tau \mathbb{P} = \left\{ X \in \mathbb{P} : \exp(-\tau) \leq \frac{\Pr[X_i = x_i]}{1/|\mathcal{X}_i|} \leq \exp(\tau), \forall i \in I \text{ with } |I| \geq \ell \right\}, \quad (37)$$

where $x_i \in \mathcal{X}_i$, $\ell \in [n]$ and $\tau \geq 0$. Note that

$$\exp(-\tau) \leq \frac{\Pr[X_i = x_i]}{1/|\mathcal{X}_i|} \leq \exp(\tau) \qquad (38)$$

ensures $H(X_i) \geq \log|\mathcal{X}_i| - \tau$. Then, by setting Δ in Definition 4 to be

$$_\ell^\tau \mathbb{P}_1 := {}_\ell^\tau \mathbb{P} \cap \mathbb{P}_1, \qquad (39)$$

we can generate a relaxation to the differential privacy model.

Set

$$_\ell^\tau \mathbb{P}_k = {}_\ell^\tau \mathbb{P} \cap \mathbb{P}_k. \qquad (40)$$

We now consider the case where $\ell = n - k$. Specifically, let

$$_{n-k}^{\ \tau} \mathbb{P}_k = \left\{ X \in \mathbb{P} : \Pr[X = x] = \Pr[X_{(I)} = x_{(I)}] \prod_{i \in I} \Pr[X_i = x_i], \forall x \in \mathcal{Z}, \quad (41) \right.$$

$$\left. \text{where } |I| = n - k, \exp(-\tau) \leq \frac{\Pr[X_i = x_i]}{1/|\mathcal{X}_i|} \leq \exp(\tau) \text{ for } i \in I, x_i \in \mathcal{X}_i \right\}. \quad (42)$$

Note that, for each $X \in {}_{n-k}^{\tau}\mathbb{P}_k$, there is

$$H(X) \geq \sum_{i \in I} H(X_i) \geq \sum_{i \in I} \log |\mathcal{X}_i| - |I|\tau. \tag{43}$$

We have the following result.

Theorem 4. *For any $r \in \mathcal{R}$, any $i \in [n]$ and any $I' \subset [n] \setminus \{i\}$ such that $|I'| = n - k - 1$, if*

$$\max_{x_i, x_i' \in \mathcal{X}_i, x_J, x_J' \in \mathcal{X}_J} \frac{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i, x_{I'}, x_J) = r] \Pr[X_{I'} = x_{I'}]}{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i', x_{I'}, x_J') = r] \Pr[X_{I'} = x_{I'}]} \leq \exp(\epsilon), \tag{44}$$

then the mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to ${}_{n-k}^{\tau}\mathbb{P}_k$, where $J = [n] \setminus (\{i\} \cup I')$ and where, for each $i \in I'$, the X_i satisfies (38).

Proof. Let $X \in {}_{n-k}^{\tau}\mathbb{P}_k$. Without loss of generality, let $|I'| = n - k - 1$ such that, for each $i \in I'$, the random variable X_i satisfies (38). Let $J = [n] - I' - \{i\}$, where $i \notin I'$. Then,

$$\max_{x_i \in \mathcal{X}_i, r \in \mathcal{R}} \frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)}]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \Pr[X_{(i)} = x_{(i)}]}$$

$$\leq \max_{x_i, x_i' \in \mathcal{X}_i, r \in \mathcal{R}} \frac{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[X_{I'} = x_{I'}] \sum_{x_J \in \mathcal{X}_J} \Pr[X_J = x_J | X_i = x_i] \Pr[\mathcal{M}(x_i, x_{I'}, x_J) = r]}{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[X_{I'} = x_{I'}] \sum_{x_J \in \mathcal{X}_J} \Pr[X_J = x_J | X_i = x_i'] \Pr[\mathcal{M}(x_i', x_{I'}, x_J) = r]}$$

$$\leq \max_{x_i, x_i' \in \mathcal{X}_i, x_J, x_J' \in \mathcal{X}_J, r \in \mathcal{R}} \frac{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i, x_{I'}, x_J) = r] \Pr[X_{I'} = x_{I'}]}{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i', x_{I'}, x_J') = r] \Pr[X_{I'} = x_{I'}]}$$

$$\leq \exp(\epsilon).$$

$$\tag{45}$$

The claim is proved.   □

By setting $\tau = 0$, we have the following result.

Corollary 5. *For any $r \in \mathcal{R}$, any $i \in [n]$ and any $I' \subset [n] \setminus \{i\}$ such that $|I'| = n - k - 1$, if*

$$\max_{x_i, x_i' \in \mathcal{X}_i, x_J, x_J' \in \mathcal{X}_J} \frac{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i, x_{I'}, x_J) = r]}{\sum_{x_{I'} \in \mathcal{X}_{I'}} \Pr[\mathcal{M}(x_i', x_{I'}, x_J') = r]} \leq \exp(\epsilon), \tag{46}$$

then the mechanism $\mathcal{M}$ satisfies ϵ-information privacy with respect to ${}_{n-k}^{0}\mathbb{P}_k$, where $J = [n] \setminus (\{i\} \cup I')$ and where, for each $i \in I'$, the X_i satisfies (38) where $\tau = 0$.

The inequalities (44) and (46) are two expectation-case relaxations of the worst-case inequality (21). Of course, we must acknowledge that the results in Theorem 4 and Corollary 5 are somewhat weak. Currently, we are unable to further simplify the inequalities (44) and (46) since we face some complicated inequalities which are related to the generalized mediant inequalities[13]. We hope, in future, we can find new approaches to simplify them.

---

[13] `https://en.wikipedia.org/wiki/Mediant_(mathematics)`

4.4   Discussion

The idea of Section 4 is to first discuss the tradeoffs of privacy and utility based on $\mathbb{P}_k, \mathbb{P}^\delta$ and $^\tau_\ell\mathbb{P}$ individually, and then synthesize these results as those based on $^\tau_\ell\mathbb{P}^\delta_k$, where

$$^\tau_\ell\mathbb{P}^\delta_k := \mathbb{P}^\delta_k \cap {}^\tau_\ell\mathbb{P} = \mathbb{P}_k \cap \mathbb{P}^\delta \cap {}^\tau_\ell\mathbb{P}. \tag{47}$$

The results of Section 4 show that the "divide and conquer" approach works. We stress that our final aim is to discuss how to control $I_\infty(X_i; Y)$ based on

$$\Delta_b = \{X \in \mathbb{P} : H(X) \geq b\} \approx {}^\tau_\ell\mathbb{P}^\delta_k. \tag{48}$$

Note that k, δ are the quantities to substitute for the mutual information to measure the dependent extent among X. Currently, we are unable to know the quantitative relation between k, δ and the mutual information $I(X_i; X_{(i)})$, which results in that we are unable to know the quantitative relation between k, δ, ℓ, τ and b. Nevertheless, at least qualitatively, we can find suitable k, δ, ℓ, τ to let (48) hold.

Clearly, setting Δ to be $^\tau_\ell\mathbb{P}^\delta_k$ would be more reasonable and flexible than to be $\mathbb{P}_1$, which implies that, theoretically, ϵ-information privacy with respect to $^\tau_\ell\mathbb{P}^\delta_k$ will achieve more privacy and utility than ϵ-differential privacy by Corollary 2.

Proposition 4 shows *the badness of big datasets* to data privacy; that is, when the number n of individuals increases, the utility of data must be bad in order to satisfy information privacy. Conversely, Assumption 1 shows *the goodness of big datasets* to data privacy; that is, when the number n of individuals increases, the lower bound b of the uncertainty $H(X)$'s of adversaries in the set

$$\Delta_b = \{X \in \mathbb{P} : H(X) \geq b\} \approx {}^\tau_\ell\mathbb{P}^\delta_k, \tag{49}$$

increases accordingly, which provides us opportunities to improve the utility of data. Explicitly, when n increases, the parameter b increases, which results in that the parameters δ, ℓ increase, the parameter τ decreases and the parameter k increase (but slowly than n), which provides us opportunities to improve the utility of data by using the results in Section 4. This in some extent implies that the information privacy model achieves the so called "crowd-blending privacy" [30], but of a flavor different from [30]; that is, an individual's privacy is "blended" with the adversaries' uncertainty to other individuals' data.

Computational Complexity Relaxation Note that the perfect secrecy can be considered as a special case of the information privacy by setting $n = 1, \Delta = \mathbb{P}, \epsilon = 0$. Similarly, the semantic security [10,11] can also be considered as a special case of the information privacy, roughly, by setting $n = 1, \epsilon = O(1/\log^t |\mathcal{X}|)$ and $\Delta = \mathbb{P}_{\mathrm{ppt}}$, where $\mathbb{P}_{\mathrm{ppt}}$ is the subset of $\mathbb{P}$ that the PPT adversaries can evaluate. Also, it is direct to define "computational" information privacy similar as in [11,52], just by setting

$$\Delta = \mathbb{P}_{\mathrm{ppt}} \tag{50}$$

and $\epsilon = \epsilon + O(1/(n\log|\mathcal{X}|)^t)$. Note that the zero-knowledge privacy model [53] is essentially equivalent to the information privacy with respect to $\mathbb{P}_{\mathrm{ppt}}$. One important thing is to discuss the information privacy with respect to

$$\Delta_{b,\mathrm{ppt}} := \Delta_b \cap \mathbb{P}_{\mathrm{ppt}} \approx {}^\tau_\ell \mathbb{P}^\delta_{k,\mathrm{ppt}} := {}^\tau_\ell \mathbb{P}^\delta_k \cap \mathbb{P}_{\mathrm{ppt}}. \tag{51}$$

Noticing that there have been many works on "computational" differential privacy [52,54,55,56,57,58,59,60,61], the "computational" information privacy with respect to ${}^\tau_\ell \mathbb{P}^\delta_{k,\mathrm{ppt}}$ would be one interesting future work.

5 Group Privacy

The group privacy problem is to study how to preserve the privacy of a group of individuals. Let $I = \{i_1, \ldots, i_s\} \subseteq [n]$ and $X_I = (X_{i_1}, \ldots, X_{i_s})$. The group privacy of the group of individuals X_I is to let the mutual information $I(X_I; Y)$ or the max-mutual information $I_\infty(X_I; Y)$ be controllable.

Definition 6 (Group Information Privacy). *For non-empty set $I = \{i_1, \ldots, i_s\}$ $\subseteq [n]$, set $(I) = [n] \setminus I = \{i_{s+1}, \ldots, i_n\}$. Set $X_I = (X_{i_1}, \ldots, X_{i_s})$, $X_{(I)} =$ $(X_{i_{s+1}}, \ldots, X_{i_n})$. Let $\Delta \subseteq \mathbb{P}$. The quantities $\mathcal{X}_I, \mathcal{X}_{(I)}, x_I,$ and $x_{(I)}$ are set accordingly. A mechanism $\mathcal{M}$ satisfies c-group information privacy with respect to Δ if for any $X \in \Delta$ and any non-empty set $I \subseteq [n]$, there is*

$$\max_{x_I \in \mathcal{X}_I, r \in \mathcal{R}} \frac{\sum_{x_{(I)} \in \mathcal{X}_{(I)}} \Pr[\mathcal{M}(x_I, x_{(I)}) = r] \Pr[X_{(I)} = x_{(I)} | X_I = x_I]}{\sum_{x'_I \in \mathcal{X}_I} \Pr[X_I = x'_I] \sum_{x_{(I)} \in \mathcal{X}_{(I)}} \Pr[\mathcal{M}(x'_I, x_{(I)}) = r] \Pr[X_{(I)} = x_{(I)} | X_I = x'_I]} \leq \exp(|I| c\epsilon),$$

where c is a positive constant.

Differential privacy has the good property that a mechanism satisfying ϵ-differential privacy will ensure to satisfy 1-group differential privacy as shown in Lemma 1, which implies that ϵ-information privacy with respect to $\mathbb{P}_1$ implies 1-group information privacy with respect to $\mathbb{P}_1$ by Corollary 2. We now generalize this result to $\mathbb{P}_k$.

Theorem 5. *Assume $\mathcal{M}$ satisfies ϵ-information privacy with respect to $\mathbb{P}_k$. Then $\mathcal{M}$ satisfies 1-group information privacy with respect to $\mathbb{P}_k$, where $k \in [n]$.*

Proof. Let $X \in \mathbb{P}_k$. By using the proving techniques in Theroem 1, we have

$$\frac{\sum_{x_{(I)} \in \mathcal{X}_{(I)}} \Pr[\mathcal{M}(x_I, x_{(I)}) = r] \Pr[X_{(I)} = x_{(I)} | X_I = x_I]}{\sum_{x'_I \in \mathcal{X}_I} \Pr[X_I = x'_I] \sum_{x_{(I)} \in \mathcal{X}_{(I)}} \Pr[\mathcal{M}(x'_I, x_{(I)}) = r] \Pr[X_{(I)} = x_{(I)} | X_I = x'_I]}$$

$$\leq \max_{x,x' : |x-x'|_1 \leq |I|-1+k} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]} \leq_a \exp\left(\left(\left\lceil \frac{|I|-1}{k} \right\rceil + 1\right)\epsilon\right) \leq \exp(|I|\epsilon),$$

where $\leq_a$ is due to (21).

The claim is proved. $\square$

Currently, we lack some techniques to prove the group privacy properties for $\mathbb{P}^\delta, {}^\tau_\ell \mathbb{P}$ and then ${}^\tau_\ell \mathbb{P}^\delta_k$. However, we believe they are true, whose proofs would be one future work.

6   Composition Privacy

The composition privacy problem is to study how to guarantee privacy while multiple datasets or multiple query results are output. There are two kinds of scenarios. First, multiple query results of *one dataset* are output. We call this kind of scenario as *the basic composition privacy problem*. To differential privacy, the privacy problem of this scenario is treated by the *composition privacy property* [39,41,62] as shown in Lemma 2. Second, multiple query results of *multiple datasets* generated by the same group of individuals are output, respectively. We call this knid of scenario as *the general composition privacy problem*. For example, the independent data publications of data of the Netflix and the IMDb [26], the independent data publications of the online and offline data [63], and the independent data publications of the voter registration data and the medical data [31]. For each of the above applications, the *composition attack* [64,26] techniques may employ the relationship between/among different datasets/queries to infer the privacy of individuals whose data is contained in these datasets.

6.1   The Basic Composition Privacy

The basic composition privacy problem, i.e., the privacy problem of multiple queries of a dataset, can be modeled as follows. For the sources $X = (X_1, \ldots, X_n)$ and the s query outputs $Y^1, \ldots, Y^s$, the composition privacy is to let $I(X_i; Y^1, \ldots, Y^s)$ or $I_\infty(X_i; Y^1, \ldots, Y^s)$ be controllable.

Definition 7 (Basic Composition Privacy). *Assume $\mathcal{M}^i$ satisfies ϵ_i-information privacy with respect to Δ and let Y^i be its output random variable, $i \in [s]$. Then the composition mechanism $\mathcal{M}$, which is defined as*

$$\mathcal{M}(x) = (\mathcal{M}^1(x), \ldots, \mathcal{M}^s(x)), x \in \mathcal{D}, \tag{52}$$

is said to satisfy c-basic composition information privacy with respect to Δ if, for each $X \in \Delta$, there are

$$I_\infty(X_i; Y) \leq c \sum_{i=1}^s \epsilon_i, i \in [n], \tag{53}$$

where $Y = (Y^1, \ldots, Y^s)$, c is a positive constant, and

$$\Pr[\mathcal{M}(x) = r] = \prod_{j=1}^s \Pr[\mathcal{M}^j(x) = r^j], \tag{54}$$

for $r = (r^1, \ldots, r^s)$.

Note that, by combining Lemma 2 with Corollary 2, we have that ϵ-information privacy with respect to $\mathbb{P}_1$ implies 1-basic composition information privacy with respect to $\mathbb{P}_1$. We now generalize this result to $\mathbb{P}_k$.

Theorem 6. *Let $\mathcal{M}$ be as shown in Definition 7 and let $\Delta = \mathbb{P}_k$. Then $\mathcal{M}$ satisfies 1-basic composition privacy with respect to $\mathbb{P}_k$.*

Proof. Let $X \in \mathbb{P}_k$. By using the proving techniques in Theroem 1, we have

$$\frac{\sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i, x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)}|X_i = x_i]}{\sum_{x_i' \in \mathcal{X}_i} \Pr[X_i = x_i'] \sum_{x_{(i)} \in \mathcal{X}_{(i)}} \Pr[\mathcal{M}(x_i', x_{(i)}) = r] \times \Pr[X_{(i)} = x_{(i)}|X_i = x_i']}$$

$$\leq \max_{|x-x'|_1 \leq k} \frac{\Pr[\mathcal{M}(x) = r]}{\Pr[\mathcal{M}(x') = r]} \leq_a \prod_{j=1}^{s} \max_{|x-x'|_1 \leq k} \frac{\Pr[\mathcal{M}^j(x) = r]}{\Pr[\mathcal{M}^j(x') = r]} \leq \exp\left(\sum_{j=1}^{s} \epsilon_j\right),$$

where $\leq_a$ is due to (54).

The claim is proved.   □

6.2   The General Composition Privacy

In this section, we discuss the general composition privacy problem. We remark that this problem is different from the basic composition privacy problem in Section 6.1, where the former is to output different privacy-preserving results of *the different datasets* generated by a same group of individuals but the later is to output different privacy-preserving results of *the same dataset*. Except some simple discussions in [19, Section 9.1], this is an almost unexplored problem in privacy protection. In this scenario, an individual X_i should be represented by a stochastic process $X_i := \{X_i^t : t \in T\}$ (but not a random variable as in former sections). The output Y also should be represented by a stochastic process $Y := \{Y^t : t \in T\}$. In this setting, we need to control the value of the mutual information

$$I(X_i; Y) = I\left(\{X_i^t : t \in T\}; \{Y^t : t \in T\}\right). \tag{55}$$

That is, the outputs $\{Y^t : t \in T\}$ should contain little information of each individual $\{X_i^t : t \in T\}$ that the adversary can obtain. For the information privacy, we need to control the quantity

$$I_\infty(X_i; Y) = I_\infty\left(\{X_i^t : t \in T\}; \{Y^t : t \in T\}\right), \tag{56}$$

which is formalized as the following definition.

Definition 8 (General Composition Privacy/Privacy for Stochastic Processes). *Let $X := (X_1, \ldots, X_n)$ be the sources, where each $X_i := \{X_i^t : t \in T\}$ is a stochastic process. Let $\mathbb{P}$ be the universe of probability distributions of X and let let $\Delta \subseteq \mathbb{P}$. Let $\mathbb{P}^{\{t\}}$ be the universe of probability distributions of $X^t := (X_1^t, \ldots, X_n^t)$ and let $\Delta^t \subseteq \mathbb{P}^{\{t\}}$. Set $x = (x_1, \ldots, x_n)$, where $x_i := \{x_i^t : t \in T\}$ with $x_i^t \in \mathcal{X}_i^t$. Set $X_{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Set $\mathcal{X}_i = \prod_{t \in T} \mathcal{X}_i^t$, $\mathcal{X}_{(i)} = \prod_{j \in [n] \setminus \{i\}} \mathcal{X}_j$. Let $\mathcal{M}$ be a mechanism and let $Y := \{Y^t : t \in T\}$ be the output stochastic process valued in a domain $\mathcal{R} := \prod_{t \in T} \mathcal{R}^t$, where, for each $r := \{r^t : t \in T\} \in \mathcal{R}$,*

there are $r^t \in \mathcal{R}^t, t \in T$. For each $x = \{x^t : t \in T\} \in \prod_{i=1}^n \mathcal{X}_i$, a mechanism $\mathcal{M} := \{\mathcal{M}^t : t \in T\}$ is defined as

$$\mathcal{M}(x) = \{\mathcal{M}^t(x^t) : t \in T\}.$$

Assume each $\mathcal{M}^t$ satisfies ϵ_t-information privacy with respect to $\Delta^t, t \in T$. Then the mechanism $\mathcal{M}$ satisfies c-general composition privacy with respect to Δ if, for each $X \in \Delta$, there are

$$I_\infty(X_i; Y) = I_\infty\left(\{X_i^t, t \in T\}; \{Y^t, t \in T\}\right) \leq c \sum_{t \in T} \epsilon_t, i \in [n], \qquad (57)$$

where c is a positive constant, and

$$\Pr[\mathcal{M}(x) = r] = \prod_{t \in T} \Pr[\mathcal{M}^t(x) = r^t]. \qquad (58)$$

We now show that the basic composition privacy problem in Definition 7 is a special case of the general composition privacy problem in Definition 8 where, for each stochastic process $X_i := \{X_i^t : t \in T\}$, the random variables $X_i^t, t \in T$ are all equal.[14] The result is shown in the following proposition.

Proposition 5. *Let the notations be as shown in Definition 8. Assume, for each* $i \in [n]$, *the random variables* $X_i^t, t \in T$ *in the stochastic process* $X_i := \{X_i^t : t \in T\}$ *are all equal. Then, for any* x_i^1 *and any* r, *there is*

$$\frac{\Pr[X_i = \bar{x}_i, Y = r]}{\Pr[X_i = \bar{x}_i]\Pr[Y = r]} = \frac{\Pr[X_i^1 = x_i^1, Y = r]}{\Pr[X_i^1 = x_i^1]\Pr[Y = r]}, \qquad (59)$$

where $\bar{x}_i = \{x_i^t : t \in T\}$ *with* $x_i^t \equiv x_i^1$ *for* $t \in T$, *which implies* $I_\infty(X_i; Y) = I_\infty(X_i^1; Y)$.

Proof. Note that

$$\Pr[X_i^1 = x_i^1, Y = r] = \sum_{x_i^{(1)} \in \mathcal{X}_i^{(1)}} \Pr[X_i^1 = x_i^1, X_i^{(1)} = x_i^{(1)}, Y = r] \qquad (60)$$

$$= \Pr[X_i = \bar{x}_i, Y = r] \qquad (61)$$

since the random variables $X_i^1, \ldots, X_i^{|T|}$ are equal. Similarly, there is $\Pr[X_i = \bar{x}_i] = \Pr[X_i^1 = x_i^1]$. The equation (59) follows by the above two results. The equation $I_\infty(X_i; Y) = I_\infty(X_i^1; Y)$ is an immediate corollary of the equation (59). □

Theorem 7. *Let* $\mathcal{M}$ *be as shown in Definition 8. For each* $t \in T$, *let* $\mathcal{M}^t$ *satisfy* ϵ_t-*information privacy with respect to* $\mathbb{P}_k^{\{t\}}$, *which is similarly defined as in (20). Then* $\mathcal{M}$ *satisfies* 1-*general composition privacy with respect to* $\mathbb{P}_k$.

Proof. The proof is similar with the proof of Theorem 6. □

---

[14] The definition of the equality of random variables please see https://en.wikipedia.org/wiki/Random_variable

Independent Applications Scenario For some applications, the assumption of the independence among different applications are reasonable. For example, for a group of individuals, their shopping data in Amazon would be independent (or less dependent) to their research data in DBLP, or their health data would be independent (or less dependent) to their movie rating data in IMDb. Therefore, for the above applications, it is in some extent reasonable to assume that an adversary's knowledge is only limited to be the independent relationship of different datasets. This setting can be modeled as that the $|T|$ random vectors (X^1, Y^1), ..., $(X^{|T|}, Y^{|T|})$ are independent, where $X^j = (X_1^j, \ldots, X_n^j)$.

Proposition 6. *Let the notations be as shown in Definition 8. Assume the $|T|$ random vectors (X^1, Y^1), ..., $(X^{|T|}, Y^{|T|})$ are mutually independent. Then, for each $i \in [n]$, any $x_i \in \mathcal{X}_i$ and any $r \in \mathcal{R}$, there are*

$$\frac{\Pr[X_i = x_i, Y = r]}{\Pr[X_i = x_i]\Pr[Y = r]} = \prod_{t \in T} \frac{\Pr[X_i^t = x_i^t, Y^t = r^t]}{\Pr[X_i^t = x_i^t]\Pr[Y^t = r^t]}, \tag{62}$$

and $I_\infty(X_i; Y) = \sum_{t \in T} I_\infty(X_i^t; Y^t)$.

Proof. The equation (62) is due to the independence of the $|T|$ random vectors (X^1, Y^1), ..., $(X^{|T|}, Y^{|T|})$, which implies $I_\infty(X_i; Y) = \sum_{t \in T} I_\infty(X_i^t; Y^t)$.   □

Note that, in Proposition 6, for each t, the random variables $X_1^t, \ldots, X_n^t, Y^t$ do not need to be mutually independent.

6.3   Discussion

The information privacy model for stochastic processes in Definition 8 is powerful to model the privacy problems of many complicated application scenarios, such as those applications in the start of Section 6. For each of these applications scenarios, the composition attack [64,26] technique may employ the relation between/among different datasets to infer the privacy of individuals whose data is contained in these datasets. Definition 8 accurately models an adversary's knowledge about relationship among the datasets generated by the individuals and our idea is to set

$$\Delta = \{X \in \mathbb{P} : H(X) \geq b\} \approx {}_\ell^\tau \mathbb{P}_k^\delta. \tag{63}$$

In this manner, the information privacy model would be immune to the composition attack.

Currently, we lack some techniques to prove the composition privacy properties for $\mathbb{P}^\delta, {}_\ell^\tau \mathbb{P}$ and then ${}_\ell^\tau \mathbb{P}_k^\delta$. However, we believe they are true, whose proofs would be one future work.

Furthermore, Definition 8 is also suitable to model the privacy problems of the streaming data [65], the set-valued data [41] and the trajectory data [66] applications. Notice that, when modeling these application scenarios, the sub-mechanisms $\mathcal{M}^1, \ldots, \mathcal{M}^{|T|}$ may be *dependent*; that is, the equation (58) would not hold. (Of course, these application scenarios are more suitable to be modeled by Definition 4 where both each X_i and Y are a stochastic process.)

7   Other Related Works

Data privacy protection has a long history [67,4] and has been developing rapidly for the last decade [3,2,24,25]. We now briefly summarize other related works.

The k-anonymity model [31] is the first privacy model that obtains extensive study. To a dataset, its main idea is to generalize the identifiers and the semi-identifier attributes to ensure at least k records has the same identifiers and the semi-identifiers. However, the k-anonymity model does not change the sensitive attributes in order to preserve data utility. Later, researchers find that the sensitive attributes themself disclose privacy. This urges the development of many variants of the k-anonymity model, such as the ℓ-diversity [68] and t-closeness [69] among others. However, these variants still have many drawbacks. Therefore, a more rigorous privacy model is needed.

There are a lot of works to adapt differential privacy to be resistant to dependent relationship attacks. The paper [70] is believed to be the first to point out that differential privacy is vulnerable to the dependent relationship attack. The paper [46] uses the group privacy property of differential privacy (Lemma 1) to deal with the dependent relationship attack. Explicitly, if there are at most k sources are dependent, one can alleviate the influence of dependent sources to the privacy guarantee of differential privacy by achieving ϵ/k-differential privacy or, equivalently, by multiplying k to the global sensitivity of the query function. This treatment is similar with the result of Theorem 1 and motivates the parameter k in Section 4.1. However, since large k will result in poor utility as discussed in the last part of Section 4.1, the paper [47] and the paper [48] introduce the notions of "correlated sensitivity" and "dependence coefficient", respectively. The two notions can be explained as introducing a dependent coefficient (which is much less than 1 in general) between/among individuals to decrease the raising speed of the global sensitivity of the query function. The two notions motivate the parameter δ in Section 4.2. Although the two notions can add less noise than the group privacy method, the privacy guarantee of the two methods have less theoretical foundation, whereas the result in Theorem 2 achieves similar aim as the above methods but with strong privacy guarantee as shown in Proposition 1.

The paper [71] is an application of the Pufferfish model and designs some mechanisms. The paper [35] relates differential privacy with the conditional mutual information $I(X_i; Y|X_{(i)})$, where $I(X_i; Y|X_{(i)}) \leq \epsilon$ is proved to be weaker than ϵ-differential privacy but stronger than (ϵ, δ)-differential privacy. By combining the above result with Corollary 2, we have that the quantity $I(X_i; Y|X_{(i)}) \leq \epsilon$ can't resist the dependent relation attack. By Proposition 2, the ϵ-inferential privacy model in [37] can be considered as a special case of our model. Furthermore, the method to measure dependence extent in Corollary 4 is more simpler than the corresponding one in [37] since the latter needs to compute complicated matrix operations, such as the matrix inverse.

The paper [72] relates the utility function in Exponential mechanism to the rate distortion function and then discusses the relation between information leakage and privacy. Another kind of work for treating differential privacy via

information theory is to measure the bound of noise complexity of differential privacy output [36,73,54]. Our model uses the relative entropy $D(X_i \| (X_i | Y = r))$ and the mutual information $I(X_i; Y)$ to treat the dependent sources problem of differential privacy. The results in Proposition 1 show that the information privacy model can ensure individual information disclosure to be upper bounded by a small value ϵ.

The outlier privacy model [51] tries to reduce the influence of the outlier records to the model's output. Conversely, the information privacy model tries to "utilize" the outlier records; specifically, in general, the more outlier records in the queried dataset, the more larger of $H(X)$ and then the more larger of b in Assumption 1, which provides more opportunities to improve utility.

8   Conclusion

The main obstacle to adopt Bayesian inference-based privacy models is that these models put restrictions to adversaries' knowledges but can't provide the reasonability of these restrictions. This paper shows that Assumption 1 is a very reasonable restriction to adversaries' knowledges and simultaneously allows flexible approaches to balance privacy and utility.

Of course, we must acknowledge that, even though there are many reasonable evidences, Assumption 1 is really not as stronger as the hardness assumptions in cryptography; the latter are founded on the computational complexity theory [74,11,10] but the former seems can't. This reminds us of Edmonds' remark [75]: "It would be unfortunate for any rigid criterion to inhibit the practical development of algorithms which are either not known or known not to conform nicely to the criterion." Therefore, maybe, we should allow the existence of Assumption 1 due to its usefulness in data privacy even though it doesn't conform nicely to the computational complexity criterion.

Furthermore, this paper leaves many unsolved problems. First, the utility bounds about the parameter ℓ, τ need to be further explored. Second, we only prove the group privacy and the composition privacy properties about the parameter k; the two properties about the parameters δ, ℓ, τ also need to be explored. Third, how to control $I(X_i; Y)$ and $\max_{r \in \mathcal{R}} D(X_i | Y = r \| X_i)$ is another one urgent future work in order to provide more choices to balance privacy and utility. Fourth, the computational information privacy with respect to Δ in (51) is another one interesting future work.

References

1. John M. Abowd, Lorenzo Alvisi, Cynthia Dwork, Sampath Kannan, Ashwin Machanavajjhala, and Jerome P. Reiter. Privacy-preserving data analysis for the federal statistical agencies. *CoRR*, abs/1701.00752, 2017.
2. Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, 2011.

3. Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4), 2010.
4. Charu C. Aggarwal and Philip S. Yu, editors. *Privacy-Preserving Data Mining - Models and Algorithms*, volume 34 of *Advances in Database Systems*. Springer, 2008.
5. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 265–284, 2006.
6. Cynthia Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.
7. Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 319–330, 2016.
8. Cynthia Dwork, Adam D. Smith, Thomas Steinke, Jonathan Ullman, and Salil P. Vadhan. Robust traceability from trace amounts. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 650–669, 2015.
9. C. E. Shannon. Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715, Oct 1949.
10. Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *J. Comput. Syst. Sci.*, 28(2):270–299, 1984.
11. Andrew Chi-Chih Yao. Theory and applications of trapdoor functions (extended abstract). In *23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982*, pages 80–91, 1982.
12. Oded Goldreich. *The Foundations of Cryptography - Volume 2, Basic Applications.* Cambridge University Press, 2004.
13. Joan Daemen and Vincent Rijmen. *The Design of Rijndael: AES - The Advanced Encryption Standard.* Information Security and Cryptography. Springer, 2002.
14. Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Tsinghua University Press, 2003.
15. Genqiang Wu, Xianyao Xia, and Yeping He. Analytic theory to differential privacy. *CoRR*, abs/1702.02721, 2018.
16. Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 5(2–1):429–444, 1977.
17. Gerome Miklau and Dan Suciu. A formal analysis of information disclosure in data exchange. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, pages 575–586, 2004.
18. Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
19. Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1):3, 2014.
20. Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 439–448, 2013.
21. Vibhor Rastogi, Michael Hay, Gerome Miklau, and Dan Suciu. Relationship privacy: output perturbation for queries with joins. In *Proceedings of the Twenty-Eigth*

ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009, June 19 - July 1, 2009, Providence, Rhode Island, USA, pages 107–116, 2009.

22. Ninghui Li, Wahbeh H. Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: a unifying framework for privacy definitions. In *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*, pages 889–900, 2013.

23. Salil Vadhan. The complexity of differential privacy. `http://privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy_1.pdf`, 2016.

24. Anand D. Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Process. Mag.*, 30(5):86–94, 2013.

25. Tianqing Zhu, Gang Li, Wanlei Zhou, and Philip S. Yu. Differentially private data publishing and analysis: A survey. *IEEE Trans. Knowl. Data Eng.*, 29(8):1619–1638, 2017.

26. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125, 2008.

27. Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210, 2003.

28. Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017.

29. Jianzhng Li and Yingshu Li. Research progress in the complexity theory and algorithms of big-data computation (in Chinese). *SCIENTIA SINICA Informationis*, 46(9):1255–1275, 2016.

30. Johannes Gehrke, Michael Hay, Edward Lui, and Rafael Pass. Crowd-blending privacy. In *Advances in Cryptology - CRYPTO 2012 - 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*, pages 479–496, 2012.

31. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

32. Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 351–360, 2013.

33. Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB J.*, 24(6):757–781, 2015.

34. Weina Wang, Lei Ying, and Junshan Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Trans. Information Theory*, 62(9):5018–5029, 2016.

35. Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 43–54, 2016.

36. Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *IEEE*

57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA, pages 487–494, 2016.

37. Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms. In *Proceedings of the 2017 ACM Conference on Innovations in Theoretical Computer Science, Berkeley, USA, January 9-11, 2017*, pages –, 2017.

38. Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1023–1034, 2015.

39. Jun Zhang, Xiaokui Xiao, and Xing Xie. Privtree: A differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 155–170, 2016.

40. Rui Chen, Gergely Ács, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *ACM Conference on Computer and Communications Security*, pages 638–649, 2012.

41. Rui Chen, Noman Mohammed, Benjamin C. M. Fung, Bipin C. Desai, and Li Xiong. Publishing set-valued data via differential privacy. *PVLDB*, 4(11):1087–1098, 2011.

42. Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: tuning privacy-utility trade-offs using policies. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1447–1458, 2014.

43. Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 747–762, 2015.

44. Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012, Allerton Park & Retreat Center, Monticello, IL, USA, October 1-5, 2012*, pages 1401–1408, 2012.

45. Ali Makhdoumi and Nadia Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *51st Annual Allerton Conference on Communication, Control, and Computing, Allerton 2013, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013*, pages 1627–1634, 2013.

46. Rui Chen, Benjamin C. M. Fung, Philip S. Yu, and Bipin C. Desai. Correlated network data publication via differential privacy. *VLDB J.*, 23(4):653–676, 2014.

47. Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Trans. Information Forensics and Security*, 10(2):229–242, 2015.

48. Changchang Liu, Prateek Mittal, and Supriyo Chakraborty. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *23nd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016*, 2016.

49. David A. Easley and Jon M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World.* Cambridge University Press, 2010.

50. Fan Chung and Linyuan Lu. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics).* American Mathematical Society, Boston, MA, USA, 2006.

51. Edward Lui and Rafael Pass. Outlier privacy. In *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part II*, pages 277–305, 2015.

52. Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil P. Vadhan. Computational differential privacy. In *Advances in Cryptology - CRYPTO 2009, 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009. Proceedings*, pages 126–142, 2009.

53. Johannes Gehrke, Edward Lui, and Rafael Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of Cryptography - 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings*, pages 432–449, 2011.

54. Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The limits of two-party differential privacy. In *FOCS*, pages 81–90, 2010.

55. Vipul Goyal, Ilya Mironov, Omkant Pandey, and Amit Sahai. Accuracy-privacy tradeoffs for two-party differentially private protocols. In *Advances in Cryptology - CRYPTO 2013 - 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part I*, pages 298–315, 2013.

56. Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *Advances in Cryptology - CRYPTO 2008, 28th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2008. Proceedings*, pages 451–468, 2008.

57. Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2879–2887, 2014.

58. Vipul Goyal, Dakshita Khurana, Ilya Mironov, Omkant Pandey, and Amit Sahai. Do distributed differentially-private protocols require oblivious transfer? In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 29:1–29:15, 2016.

59. Adam Groce, Jonathan Katz, and Arkady Yerukhimovich. Limits of computational differential privacy in the client/server setting. In *Theory of Cryptography - 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings*, pages 417–431, 2011.

60. Mark Bun, Yi-Hsiu Chen, and Salil P. Vadhan. Separating computational and statistical differential privacy in the client-server model. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, pages 607–634, 2016.

61. Dakshita Khurana, Hemanta K. Maji, and Amit Sahai. Black-box separations for differentially private protocols. In *Advances in Cryptology - ASIACRYPT 2014 - 20th International Conference on the Theory and Application of Cryptology and Information Security, Kaoshiung, Taiwan, R.O.C., December 7-11, 2014, Proceedings, Part II*, pages 386–405, 2014.

62. Noman Mohammed, Rui Chen, Benjamin C. M. Fung, and Philip S. Yu. Differentially private data release for data mining. In *KDD*, pages 493–501, 2011.

63. Ping Luo, Su Yan, Zhiqiang Liu, Zhiyong Shen, Shengwen Yang, and Qing He. From online behaviors to offline retailing. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 175–184, 2016.

64. Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 265–273, 2008.
65. Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *STOC*, pages 715–724, 2010.
66. Rui Chen, Benjamin C. M. Fung, Bipin C. Desai, and Nériah M. Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *KDD*, pages 213–221, 2012.
67. Nabil R. Adam and John C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, 1989.
68. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 24, 2006.
69. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115, 2007.
70. Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*, pages 193–204, 2011.
71. Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1291–1306, 2017.
72. Darakhshan J. Mir. Information-theoretic foundations of differential privacy. In *Foundations and Practice of Security - 5th International Symposium, FPS 2012, Montreal, QC, Canada, October 25-26, 2012, Revised Selected Papers*, pages 374–381, 2012.
73. Gilles Barthe and Boris Köpf. Information-theoretic bounds for differentially private mechanisms. In *Proceedings of the 24th IEEE Computer Security Foundations Symposium, CSF 2011, Cernay-la-Ville, France, 27-29 June, 2011*, pages 191–204, 2011.
74. Whitfield Diffie and Martin E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, 22(6):644–654, 1976.
75. Jack Edmonds. Paths, trees and flowers. *Canadian Journal of Mathematics*, pages 449–467, 1965.

Table 1: Table of notation

Notation	Description
$\|z\|_1$	the ℓ_1-norm of the real vector z
$n, [n]$	the number of individuals, the set $\{1,\ldots, \text{n}\}$, respectively
I, X_I	the subset $\{i_1,\ldots,i_\ell\}$ of $[n]$, the vector $(X_{i_1},\ldots,X_{i_\ell})$, respectively
X_i	the random variable denoting the ith individual
Y	the output random variable
$I(X_i;Y)$	the mutual information of X_i and Y
$I_\infty(X_i;Y)$	the max-mutual information of X_i and Y
$H(X)$	the entropy of X
$X_{(i)}$	the random vector $(X_1,\ldots,X_{i-1}, X_{i+1},\ldots,X_n)$
$x_{(i)}$	the vector $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$
$\mathcal{X}_i$	the record universe of the ith individual X_i
$\mathcal{X}_{(i)}$	the Cartesian set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_n$
$\mathcal{X}$	the set $\cup_{i=1}^n \mathcal{X}_i \setminus \{\bot\}$, where $\bot$ denotes an empty record
X	the sequence of the individuals $(X_1,\ldots,X_n)$
$\mathcal{Z}$	the universe of record sequences $\prod_{i=1}^n \mathcal{X}_i$
$\mathcal{D}$	the universe of datasets
$\mathcal{R}$	a set containing the query function f's codomain $\{f(x) : x \in \mathcal{D}\}$
$\mathbb{P}$	the universe of probability distribution over $\mathcal{Z}$ (or over $\mathcal{D}$)
Δ	a subset of $\mathbb{P}$
$X \in \Delta$	the probability distribution of X is in Δ
k	the maximum number of dependent individuals
δ	the dependent extent among the individuals
τ	the parameter to measure the uncertainty of the adversary to each individual
ℓ	the parameter to measure the number of unknown individuals
$\mathbb{P}^\delta$	the subset of $\mathbb{P}$ with dependent parameters $\leq \delta$
$\mathbb{P}_k$	the subset of $\mathbb{P}$ with dependent parameters $\leq k$
$_\ell^\tau\mathbb{P}$	the subset of $\mathbb{P}$ with parameters τ, ℓ
$_\ell^\tau\mathbb{P}_k^\delta$	the set $\mathbb{P}_k \cap \mathbb{P}^\delta \cap {_\ell^\tau\mathbb{P}}$
PPT	the abbreviation of "probabilistic polynomial time"
$\mathbb{P}_{\text{ppt}}$	the subset of $\mathbb{P}$ that the PPT adversaries can evaluate