Notes on Information-Theoretic Privacy*

Shahab Asoodeh, Fady Alajaji, and Tamás Linder
Department of Mathematics and Statistics, Queen's University
{asoodehshahab, fady, linder}@mast.queensu.ca

***Abstract*—We investigate the tradeoff between privacy and utility in a situation where both privacy and utility are measured in terms of mutual information. For the binary case, we fully characterize this tradeoff in case of *perfect privacy* and also give an upper-bound for the case where some privacy leakage is allowed. We then introduce a new quantity which quantifies the amount of private information contained in the observable data and then connect it to the optimal tradeoff between privacy and utility.**

I. Introduction

Suppose Alice has some personal information which is represented by random variable X and she wants to keep this personal information as private as possible. However there exists some correlated information, represented by Y, observable by an advertising company and to be displayed publicly by this company. The company gets paid to send the most information about Y, and at the same time it does not want to violate the privacy of Alice. The question raised in this situation is then how much information about Y can be displayed without breaching privacy? Hence, it is of interest to characterize such competing objectives in the form of a quantitative tradeoff. Such a characterization provides a controllable balance between utility and privacy.

Statistical studies regarding privacy were started by Warner [1] who suggested privacy-preserving methods for survey sampling. More recently, a measure known as differential privacy was introduced by Dwork et al. [2]. In this setting, usually the source is modelled as a *database* $X = (X_1, X_2, \dots, X_n) \in \mathcal{D}^n$ where $\mathcal{D} = \{0,1\}^\ell$. A *mechanism* $\mathcal{M} : \mathcal{D}^n \to \mathcal{S}$, where $\mathcal{S}$ is a set not necessarily equal to $\mathcal{D}^n$, then produces the *sanitized* database based on the tradeoff between accuracy and privacy. The accuracy of differential privacy is defined via a query $q : \mathcal{S} \to \mathcal{R}$ where $\mathcal{R}$ is some abstract set. The query can be viewed as a question about the original database X that one might ask. Each query is then answered using the sanitized database Z. On the one hand, the data provider wants to have an accurate answer to each query, and on the other hand, the provider needs to satisfy a certain level of privacy.

In an information-theoretic context, $\mathcal{M}$ is simply a Markov kernel (i.e., channel) $Q_n(\cdot|X)$ with output $Z := \mathcal{M}(X)$ which takes values in $\mathcal{S}$. The privacy is then measured by the upper-bound of the likelihood ratio of x and x' with Hamming distance 1, that is, the mechanism is called ϵ-differentially private if $\frac{Q_n(B|x)}{Q_n(B|x')} \leq \exp(\epsilon)$ for all measurable $B \subset \mathcal{S}$ and all $x, x' \in \mathcal{D}^n$ such that $d_H(x, x') = 1$, where d_H is the Hamming distance. Note that this definition does not involve the prior distribution of x and x'. Another measure of privacy was recently proposed under the name of *a posteriori differential privacy* which incorporates the prior distributions by Wang et al. [3].

The locality requirement of $d_H(x, x') = 1$ in the definition makes it hard to connect differential privacy to information theory. To overcome this problem, Duchi et al. [4] removed the condition $d_H(x, x') = 1$. This generalized definition yields the upper bound $I(X, Z) \leq \epsilon$ on the mutual information, which gives an information-theoretic interpretation of differential privacy. Hence generalized ϵ-differentially private mechanism leaks at most ϵ private information.

Despite its frequent use in computer science, differential privacy does not characterize the optimal balance between privacy versus accuracy. For example, if we want only 1% privacy leakage, it is not clear what the best achievable accuracy is. Furthermore, it is not clear how to define differential privacy when instead of the database X, another database Y, correlated with X, is observable.

The problem treated in this paper can also be contrasted with the more well-studied concept of *secrecy*. While in secrecy problems, e.g., in cryptography, wiretap channel problems, etc., the aim is to keep information secret only from wiretappers, the problem treated in privacy further aims to keep the correlated source private from the intended receiver.

Although there has been no universal way of measuring privacy in the literature, in this work we follow Yamamoto [5] who proposed a private source coding model. He introduced the equivocation as the conditional entropy of the private message given the observation and then defined the privacy in the system as the equivocation involved in the decoding. He then defined the rate-distortion-equivocation function as the tradeoff between utility (i.e., distortion) and privacy (i.e, equivocation). Inspired by this work, we use the mutual information between private information X and the displayed information Z as the measure of privacy and also use the mutual information between the observable data Y and Z as utility and then define the rate-privacy function as the optimal tradeoff between these quantities. Defining utility and privacy using the mutual information gives a more intuitive measure of how much the receiver knows about Y and how much of the private information is leaked to the receiver.

*The first lemma in the version published in Allerton 2014 is incorrect. We could fortunately prove the other results in this version without invoking that lemma.

The paper is organized as follows. In section II, we formulate the problem in terms of the rate-privacy function and also study the binary case. We show that if zero privacy leakage is required, then in the binary case, no information from Y can be transmitted. In section III we give a multi-letter version of the rate-privacy function in a special case and show that even if n different copies of Y are observed, non-zero information can be transmitted about Y when vanishing privacy leakage is required. In section IV, we define a new quantity related with privacy and pose an intuitive question connecting the new quantity with the rate-privacy function for the case of zero privacy leakage.

II. Problem Formulation and the Rate-Privacy Function

Consider two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with $|\mathcal{X}|, |\mathcal{Y}| < \infty$ and fixed joint distribution P_{XY}. X is the *private data* and Y is the *observable data* correlated with X. Suppose there exists a channel $P_{Z|Y}$ such that Z, the *displayed data*, has limited information about X. This channel is called the *privacy filter*. The objective is then to find the most informative privacy filter, i.e., a channel which preserves most of the information contained in Y. This setup is shown in Fig. 1. In particular, we are interested in

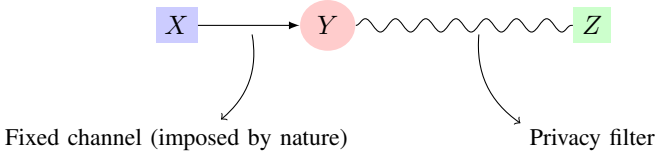

Fig. 1. Information-theoretic privacy.

characterizing the quantity,

$$g_\epsilon(X;Y) := \max_{P_{Z|Y}: I(X;Z)\leq\epsilon} I(Y;Z), \tag{1}$$

which we call, the *rate-privacy function*. The dual representation of $g_\epsilon(X;Y)$ is given in [6] and called the *privacy funnel*. Basically, in this model, the privacy and utility are both measured using mutual information. Note that since $I(Y;Z)$ is a convex function of $P_{Z|Y}$ and furthermore the constraint set $\mathcal{D}_\epsilon := \{P_{Z|Y} : \ I(X;Z) \leq \epsilon\}$ is convex and compact, the maximum in (1) occurs at the extreme points, namely for a $P_{Z|Y} \in \mathcal{D}_\epsilon$ such that $I(X;Z) = \epsilon$. If we restrict $P_{Z|Y}$ to be a deterministic function f, we get the simplified quantity

$$\tilde{g}_\epsilon(X;Y) := \sup_{f: I(f(Y);X)=\epsilon} H(f(Y)). \tag{2}$$

Using the Carathéodory-Fenchel theorem, one can readily show that it suffices that the random variable Z is supported on an alphabet $\mathcal{Z}$ with cardinality $|\mathcal{Z}| \leq |\mathcal{Y}| + 1$.

In the study of $g_\epsilon(X;Y)$ for general P_{XY}, the most interesting case is when $\epsilon = 0$ (the so-called *prefect privacy*), i.e., no privacy leakage is allowed. The following theorem shows that for binary X and Y and an arbitrary channel between X and Y the requirement of perfect privacy allows no information transfer from Y.

Theorem 1. *For any pair of dependent binary random variables X and Y, we have*

$$g_0(X;Y) = 0.$$

Proof. In the perfect privacy regime the constraint set reduces to $\mathcal{D}_0 = \{P_{Z|Y} : \ Z \perp\!\!\!\perp X\}$. Since X, Y and Z form the Markov chain $X \to Y \to Z$, we can write

$$\begin{aligned} P_{Z|Y}(\cdot|0)P_{Y|X}(0|1) + P_{Z|Y}(\cdot|1)P_{Y|X}(1|1) &= P_{Z|X}(\cdot|1) \\ P_{Z|Y}(\cdot|0)P_{Y|X}(0|0) + P_{Z|Y}(\cdot|1)P_{Y|X}(1|0) &= P_{Z|X}(\cdot|0). \end{aligned}$$

The condition $Z \perp\!\!\!\perp X$ implies that $P_{Z|X}(\cdot|1) = P_{Z|X}(\cdot|0) = P_Z(\cdot)$ and hence from the above,

$$\begin{aligned} P_{Z|Y}(\cdot|0)P_{Y|X}(0|1) + P_{Z|Y}(\cdot|1)P_{Y|X}(1|1) &= P_Z(\cdot) \\ P_{Z|Y}(\cdot|0)P_{Y|X}(0|0) + P_{Z|Y}(\cdot|1)P_{Y|X}(1|0) &= P_Z(\cdot). \end{aligned}$$

From the assumption that X and Y are dependent, it follows that the above system of equations has a unique solution. The unique solution turns out to satisfy $P_{Z|Y}(\cdot|0) = P_{Z|Y}(\cdot|1) = P_Z(\cdot)$, which implies that $I(Y;Z) = 0$. □

Note that the theorem does not necessarily hold for non-binary X and Y. In fact, it is easy to construct an example for ternary Y and binary X in which $g_0(X;Y) > 0$ (for instance, see Example 1). Berger and Yeung [7, Appendix II], gave a necessary condition for $g_0(X;Y) > 0$, in a different context.

Definition 1 ([7]). *The random variable X is said to be* weakly independent *of Y if the rows of the transition matrix $P_{X|Y}$, i.e., the set of vectors $\{P_{X|Y}(\cdot|y),\ y \in \mathcal{Y}\}$, are linearly dependent.*

In [7], it is proved that if X is weakly independent of Y then there exists a binary random variable Z such that $Z \perp\!\!\!\perp X$ which is correlated with Y, and hence $g_0(X;Y) > 0$. This condition is met, for example, if $|\mathcal{Y}| > |\mathcal{X}|$. It is also straightforward to show that this condition is indeed a necessary and sufficient for $g_0(X;Y) > 0$. It is straightforward to see that if Y is binary then X is weakly independent of Y if and only if X and Y are independent. This together with the fact that weak independence is a necessary and sufficient condition for $g_0(X;Y) > 0$, imply the following lemma which generalizes Theorem 1.

Lemma 1. *Let Y be a binary random variable. Then $g_0(X;Y)$ is equal to either $H(Y)$ or zero.*

III. A Multi-letter Version of $\tilde{g}_\epsilon(X;Y)$

We next consider the simplified version of the rate-privacy function $\tilde{g}_\epsilon(X;Y)$ defined in (2), in the limit when $\epsilon \to 0$. Suppose for any $x \in \mathcal{X}$, inducing the distribution $P_{Y|X}(\cdot|x)$ over $\mathcal{Y}$, one takes n independent copies of Y with distribution $P_{Y^n|X}(y^n|x) = \prod_{i=1}^n P_{Y|X}(y_i|x)$. The privacy constraint is that; $I(f(Y^n);X) = \epsilon$ for every n and every deterministic function $f : \mathcal{Y}^n \to \mathcal{Z}$ where $|\mathcal{Z}| \leq |\mathcal{Y}|$. Let $\tilde{g}_{n,\epsilon}(X;Y)$ denote $\frac{1}{n}\tilde{g}_\epsilon(X;Y^n)$ when the distribution $P_{Y^n|X}$ is specified as above, so that

$$\tilde{g}_{n,\epsilon}(X;Y) := \frac{1}{n} \sup_{f:\ I(f(Y^n);X)=\epsilon} H(f(Y^n)). \tag{3}$$

The following theorem gives an asymptotic lower bound on $\tilde{g}_{n,\epsilon}(X;Y)$.

Theorem 2. *For any pair of random variables (X,Y) with fixed joint distribution P_{XY}, we have*

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\tilde{g}_{n,\epsilon}(X;Y)\geq H_\infty^*(Y|X),$$

where the min-entropy is defined as

$$H_\infty^*(Y|X):=\min_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}}\left(-\log P_{Y|X}(y|x)\right). \tag{4}$$

Proof. Suppose $|\mathcal{X}|=m$ and each $x_j\in\mathcal{X}$, $j=1,\ldots,m$, induces the product distribution $P_j^n(y^n):=P_{Y^n|X}(y^n|x_j)=\prod_{k=1}^n P_{Y|X}(y_k|x_j)$ over $\mathcal{Y}$. Given these m distributions P_j^n for $j=1,2,\ldots,m$, we construct nearly equiprobable bins $K_j^n(i)\subset\mathcal{Y}^n$ for $i=1,2,\ldots,2^r$, (with r to be determined later), such that $P_j^n(K_j^n(i)):=P_j^n(Y^n\in K_j^n(i))$ is close to 2^{-r} for each $j=1,2,\ldots,m$ and $i=1,2,\ldots,2^r$. Let U^r denote the uniform distribution over $\{0,1\}^r$ and $V(P,Q)$ denote the total variation distance between distributions P and Q.

Note that each bin $K_j^n(i)$ is an agglomeration of some mass points of $P_j^n(y^n)$ for each $j=1,2,\ldots,m$ and therefore the probability of each bin is equal to the sum of the probabilities of points y^n it contains. Recalling the definition of $H_\infty^*(Y|X)$ in (4), we can write

$$P_j^n(y^n)\leq 2^{-nH_\infty^*(Y|X)},\qquad j=1,2,\ldots,m. \tag{5}$$

We start the construction of the bins $K_j^n(1),K_j^n(2),\ldots,K_j^n(J_j)$ for each $j=1,2,\ldots,m$ where $J_j\leq 2^r-1$ is the number of bins for each j. The first bin is constructed as follows. We agglomerate the minimal number of mass points of P_j^n into $K_j^n(1)$ as needed to make sure

$$P_j^n(K_n(1))\geq 2^{-r}-2^{-s}, \tag{6}$$

for some $s<nH_\infty^*(Y|X)$. This together with (5) shows that

$$P_1^n(K_n(1))<2^{-r}-2^{-s}+2^{-nH_\infty^*(Y|X)}, \tag{7}$$

which can be simplified as

$$P_1^n(K_n(1))<2^{-r}, \tag{8}$$

because $s<nH_\infty^*(Y|X)$.

Once condition (6) is met, the construction for the first bin is completed and we move on to the second bin. This procedure can go on until either we run out of mass points or the restriction $J_j\leq 2^r-1$ is violated. In the latter case, we set $J_j=2^r-1$ and then collect all mass points left into the bin $K_j^n(J_j+1)$. The former happens if the total probability of the left-over is strictly less than $2^{-r}-2^{-s}$ so that we can not meet the requirement (6), in other words,

$$P_j^n\left(\bigcup_{i=1}^{J_j}K_j^n(i)\right)>1-2^{-r}+2^{-s}. \tag{9}$$

On the other hand, we know from (8) that $P_j^n\left(\bigcup_{i=1}^{J_j}K_j^n(i)\right)<J_j2^{-r}$ which, together with (9), implies

$$1-2^{-r}+2^{-s}<P_j^n\left(\bigcup_{i=1}^{J_j}K_j^n(i)\right)<J_j2^{-r}, \tag{10}$$

leading to a lower bound for the number of bins in this case

$$J_j>2^r+2^{r-s}-1, \tag{11}$$

which is greater than the allowable upper-bound 2^r-1. We can hence conclude that with s that satisfies $s<nH_\infty^*$, the procedure stops only when the restriction $J_j\leq 2^r-1$ is violated, and therefore, we assume $J_j=2^r-1$ in what follows.

As specified earlier, we construct the last bin $K(J_j+1)$ by including all the leftover mass there. We therefore have

$$K_j^n(J_j+1)=\text{supp}\{P_j^n\}-\bigcup_{i=1}^{J_j}K_j^n(i), \tag{12}$$

where $\text{supp}\{P_j^n\}$ denotes the support of P_j^n. Since each bin has probability lower-bounded by (6), it follows from (12) that

$$P_j^n(K(J_j+1))=1-\sum_{i=1}^{J_j}P_j^n(K_j^n(i))\leq 1-J_j\left(2^{-r}-2^{-s}\right), \tag{13}$$

which, after substituting $J_j=2^r-1$, is simplified as

$$P_j^n(K(J_j+1))\leq 2^{r-s}+2^{-r}-2^{-s}. \tag{14}$$

So far we have constructed $m\times 2^r$ bins, namely 2^r bins for each P_j^n, $j=1,2,\ldots,m$. Consider now the deterministic mapping $g_n:\mathcal{Y}^n\times\mathcal{X}\to\{0,1\}^r$ defined as follows:

$$g_n(y^n,x_j)=i\qquad\text{if}\qquad y^n\in K_j^n(i).$$

This mapping requires x_j because for each $j\in\{1,2,\ldots,m\}$ the corresponding bins are disjoint. However, we know that by using a proper channel encoding and decoding, ϕ_n and ψ_n, respectively, one can decode Y^n to obtain $\psi_n(Y^n)$ such that $P(X\neq\psi_n(Y^n))$ decays exponentially. So, we can have a deterministic function which acts only on Y^n from which x_j is obtained with probability exponentially close to one. Hence our sequence of deterministic mappings is:

$$f_n(y^n):=g_n(y^n,\psi_n(y^n))=i\qquad\text{if}\qquad y^n\in K_j^n(i).$$

where j is the index of the decoded symbol, that is the j such that $\psi_n(y^n)=x_j$.

Now let us look at the total variation distance between $\tilde{P}_j^n:=f_n\circ P_j^n$ and U^r which is the uniform distribution over the set $\{0,1\}^r$.

$$\begin{aligned}V\left(\tilde{P}_j^n,U^r\right)&=\sum_{i=1}^{2^r}|2^{-r}-P_j^n(K_j^n(i))|\\&=\sum_{i=1}^{J_j}\left(2^{-r}-P_j^n(K_j^n(i))\right)\\&\quad+|2^{-r}-P_j^n(K_j^n(J_j+1))|\end{aligned} \tag{15}$$

$$\leq \sum_{i=1}^{J_j} 2^{-s} + 2^{-r} + P_j^n(K_n(J_j+1)) \quad (16)$$
$$\leq J_j 2^{-s} + 2^{-r} + 2^{r-s} + 2^{-r} - 2^{-s} \quad (17)$$
$$= 2\left(2^{r-s} + 2^{-r} - 2^{-s}\right) < 2\left(2^{r-s} + 2^{-r}\right).$$

where in (15) we use (8), in (16) we use the triangle inequality and (6) and the inequality in (17) follows from (14). To make sure that $V(\tilde{P}_j^n, U^r)$ goes to zero as $n \to \infty$, we set $r = nH_\infty^*(Y|X) - n\delta$ and $s = nH_\infty^*(Y|X) - n\frac{\delta}{2}$ for some $0 < \delta \leq \frac{2}{3}H_\infty^*(Y|X)$. Hence we can make sure that $\tilde{P}_j^n$ and $\tilde{P}_k^n$ for $j \neq k$ are at most 2ϵ-distant in the total variation sense. This is because, for large n

$$V(\tilde{P}_j^n, \tilde{P}_k^n) \leq V(\tilde{P}_j^n, U^r) + V(\tilde{P}_k^n, U^r) \leq 2\epsilon.$$

Note that, letting $E_X[\cdot]$ denote the expectation with respect to X, we have in general

$$\begin{aligned} V\left(P_{ZX}, P_Z P_X\right) &= E_X\left[V\left(P_{Z|X}(\cdot|X), P_Z\right)\right], \\ &= E_X\left[V\left(P_{Z|X}(\cdot|X), E_X[P_{Z|X}(\cdot|X)]\right)\right], \end{aligned}$$

and hence by Jensen's inequality

$$\begin{aligned} &V(P_{ZX}, P_Z P_X) \\ &\leq \sum_x \sum_{x'} P_X(x) P_X(x') V\left(P_{Z|X}(\cdot|x), P_{Z|X}(\cdot|x')\right) \end{aligned}$$

We can therefore conclude that $V(\tilde{P}_j^n, \tilde{P}_k^n) \leq 2\epsilon$ for all $j \neq k$ results in the following

$$V\left(P_{Z_n X}, P_{Z_n} P_X\right) \leq 2\epsilon,$$

where $Z_n = f_n(Y^n)$. In other words, Z_n and X are "2ϵ-independent" for sufficiently large n in sense of total variation distance. Invoking [8, Lemma 2.7], the theorem follows. □

This theorem implies that, unlike in the binary case studied in Theorem 1, one can have information transfer at a positive rate while allowing perfect privacy only in the limit instead of requiring absolutely zero privacy leakage.

IV. Non-Private Information vs. the Rate-Privacy Function

Conceptually, $g_\epsilon(X;Y)$ quantifies the "largest" part of Y which carries ϵ amount of information about X. Witsenhausen [9] defined the *private information* of a pair of random variables (X,Y) as

$$M(X;Y) := \max_{W:\ X \to W \to Y} H(X,Y|W). \quad (18)$$

Wyner [10] defined the *common information* of X and Y as

$$C_W(X;Y) := \min_{W:\ X \to W \to Y} I(X,Y;W). \quad (19)$$

Clearly, the definition of private information in (18) implies $C_W(X;Y) = H(X,Y) - M(X;Y)$. Operationally, $M(X;Y)$ is the rate of information that one needs to transmit over two "non-common" channels when $C_W(X;Y)$ is transmitted over the common channel in order to be able to decode X and Y with arbitrarily small error probability. This definition is not immediately useful in our setting, as it is symmetric in X and Y. We seek an asymmetric definition for the private information that Y contains, i.e., the rate of information contained in Y which correlates with X. Inspired by Wyner's common information, $C_W(X;Y)$, and Gács-Körner's common information [11], denoted by $C_{GK}(X;Y)$, we define the *private information about X carried by Y* as follows

$$C_X(Y) := \min_{\substack{W:X \to W \to Y \\ H(W|Y)=0}} H(W), \quad (20)$$

and similar to the connection between $C_W(X;Y)$ and $M(X;Y)$, we define $D_X(Y) := H(Y) - C_X(Y)$ and call it the *non-private information about X carried by Y*. The quantity $C_X(Y)$ as defined above is similar to the so-called *necessary conditional entropy*, $H(Y \dagger X)$, defined by Cuff et al. [12] as $\min H(W|X)$ where the minimum is taken over W that satisfies the same conditions as in (20). Conceptually, we decompose the information contained in Y into two parts, namely, one part which correlates with X, denoted by $C_X(Y)$, and another part which has no correlation with X, denoted by $D_X(Y)$. Using the assumption $H(W|Y) = 0$ in (20), we can obtain the following variational representation for $D_X(Y)$:

$$D_X(Y) = \max_{\substack{W:X \to W \to Y \\ H(W|Y)=0}} H(Y|W). \quad (21)$$

Remark 1. Since $H(W|Y) = 0$ implies that W is a function of Y, one can show that the constraint in the above maximization, i.e., the conditions $X \to W \to Y$ and $H(W|Y) = 0$, is equivalent to the "double Markov relations" $X \to W \to Y$ and $X \to Y \to W$.

The so called *exact common information* has been introduced in [13] and shown to be related to the problem of *exact* generation of a joint distribution P_{XY}. The exact common information is defined as the minimum rate R^* at which an external randomness must be supplied to physically separated agents, each responsible for one of the marginals via the private randomness, so that they are able to *exactly* reproduce joint distribution P_{XY}, in an asymptotic formulation. As illustrated in Fig. 2, the exact common information is the minimum rate of generating W such that two independent processors construct $\hat{X}$ and $\hat{Y}$, using W as an input of separate stochastic decoders, such that $P_{\hat{X}\hat{Y}} = P_{XY}$.

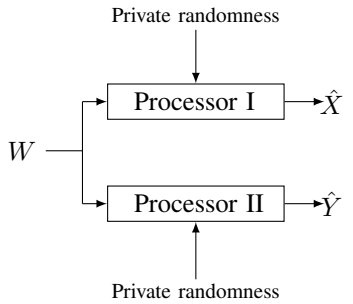

Fig. 2. Exact distribution generation.

A new quantity is then introduced in [13], so called

common entropy defined by

$$G(X;Y) := \min_{W:X\to W\to Y} H(W), \tag{22}$$

and shown that $R^* = \lim_{n\to\infty} \frac{1}{n} G(X^n;Y^n)$. Operationally, $C_X(Y)$ is the exact common information for a setting similar to Fig. 2, except that the common input to each processor is assumed to be a deterministic function of Y as depicted in Fig. 3.

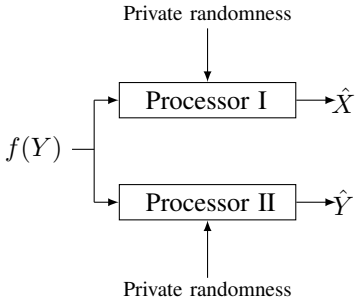

Fig. 3. Exact asymmetric distribution generation.

A. Properties of $C_X(Y)$

1) For any (X,Y) with joint distribution P_{XY}, we have

$$I(X;Y) \le C_W(X;Y) \le G(X;Y) \le C_X(Y) \le H(Y). \tag{23}$$

Proof. The first and the second inequalities are shown respectively in [10] and [13]. The third one becomes clear once we examine the definitions of $G(X;Y)$ and $C_X(Y)$. Indeed, the objective functions in the minimization are equal, however, the constraint set for $C_X(Y)$ is a subset of the constraint set for $G(X;Y)$. The last inequality follows from the fact that Y belongs to the constraint set as well. □

Note that $C_X(Y) = I(X;Y)$ implies that $C_W(X;Y) = I(X;Y) = C_X(Y)$. It is a well-known fact that $C_W(X;Y) = I(X;Y)$ is equivalent to $C_{GK}(X;Y) = I(X;Y)$. Thus, $C_X(Y) = I(X;Y)$ implies that $C_{GK}(X;Y) = C_W(X;Y)$. As Wyner [10, p. 166] pointed out, these two notions of common information are equal if and only if it is possible to write $X = (X',V)$ and $Y = (Y',V)$ such that X' and Y' are conditionally independent given V. Hence $C_X(Y) = I(X;Y)$ implies this decomposition. For the converse, suppose that we have the decomposition $X = (X',V)$ and $Y = (Y',V)$ such that $X' \to V \to Y'$. It is easy to show that for any random variable W that satisfies $X \to W \to Y$ and $H(W|Y) = 0$, there exits a deterministic function f such that $V = f(W)$ with probability one. Hence, on the one hand,

$$\max_{\substack{W:X\to W\to Y\\ H(W|Y)=0}} H(Y|W) \le H(Y|V),$$

and on the other hand, since V also satisfies both conditions of W, we have

$$\max_{\substack{W:X\to W\to Y\\ H(W|Y)=0}} H(Y|W) \ge H(Y|V),$$

and therefore, $D_X(Y) = H(Y|V) = H(Y|X)$ and consequently $C_X(Y) = I(X;Y)$.

2) $C_X(Y) = 0$ if and only if $X \perp\!\!\!\perp Y$.

Proof. Suppose $C_X(Y) = 0$. By the first inequality in (23), we have $I(X;Y) = 0$ which implies $X \perp\!\!\!\perp Y$. Conversely, if $X \perp\!\!\!\perp Y$, then we have the following trivial Markov chain $X \to c \to Y$ for any constant c. This implies $C_X(Y) = 0$. □

3) (*Data-processing inequality*) For any U such that $U \to X \to Y$, we have $C_U(Y) \le C_X(Y)$.

Proof. Let W^* attain the $C_X(Y)$. Hence we have the Markov chain $U \to X \to W^* \to Y$ and also $H(W^*|Y) = 0$. It then follows by the definition that $C_U(Y) \le H(W^*) = C_X(Y)$. □

B. Calculation of $D_X(Y)$

In this section we solve the maximization in the definition of $D_X(Y)$. To do this, we need a definition which also appears in [12], [14] and [15].

Definition 2. *Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, let $T^{\mathcal{X}} : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ be defined by $y \to P_{X|Y}(\cdot|y)$ where $\mathcal{P}(\mathcal{X})$ is the simplex of probability distribution on $\mathcal{X}$.*

To solve the maximization in the definition of $D_X(Y)$, we need the following two lemmas from [16].

Lemma 2 ([16])**.** *The random variable $T^{\mathcal{X}}(Y)$ satisfies the Markov chain $X \to T^{\mathcal{X}}(Y) \to Y$.*

This lemma shows that the random variable $T^{\mathcal{X}}(Y)$ is basically a sufficient statistics of Y with respect to X.

Lemma 3 ([16])**.** *Let X, Y and V form a Markov chain, $X \to V \to Y$ and also $H(V|Y) = 0$. Then there exists a deterministic function g, such that $T^{\mathcal{X}}(Y) = g(V)$ with probability one.*

This lemma together with Lemma 2 implies that $T^{\mathcal{X}}(Y)$ is the *minimal* sufficient statistics of Y with respect to X, i.e., all other sufficient statistics of Y are a function of $T^{\mathcal{X}}(Y)$.

The following theorem shows that $T^{\mathcal{X}}(Y)$ solves the minimization in the definition of $C_X(Y)$.

Theorem 3. *For any pair of random variables (X,Y) with joint distribution P_{XY}, we have*

$$D_X(Y) = H(Y) - H(T^{\mathcal{X}}(Y)).$$

Proof. Since $H(Y|W) = H(Y) - H(W)$ for all W that satisfies the condition $H(W|Y) = 0$, we will show that $T^{\mathcal{X}}(Y)$ minimizes $H(W)$ over all W that satisfies Markov chain $X \to W \to Y$ and $H(W|Y) = 0$. Note that Lemma 2 shows that $T^{\mathcal{X}}(Y)$ belongs to the constraint set

of the maximization in the theorem and Lemma 3 shows that $T^{\mathcal{X}}(Y)$ has the smallest entropy among all the random variables in the constraint set. These two lemmas therefore together imply that $H(Y|W)$ attains its maximum value at $W = T^{\mathcal{X}}(Y)$. □

As mentioned before, $C_X(Y) \leq H(Y)$. From the previous theorem we can now give the condition under which $C_X(Y) = H(Y)$. Assume that $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$.

Lemma 4. *$C_X(Y) = H(Y)$ if and only if there exists no $y_1, y_2 \in \mathcal{Y}$ such that $P_{X|Y}(\cdot|y_1) = P_{X|Y}(\cdot|y_2)$.*

Proof. From Theorem 3, it is easy to see that if for all $y \in \mathcal{Y}$, $P_{X|Y}(\cdot|y)$ are different, then $C_X(Y) = H(Y)$. Conversely, suppose that W^* attains $C_X(Y)$ and also suppose $H(W^*) = H(Y)$. Assume that there exist y_1 and y_2 such that $P_{X|Y}(\cdot|y_1) = P_{X|Y}(\cdot|y_2)$. Then define a new random variable $\tilde{Y}$ which takes on values on set $\{y', y_3, \ldots, y_m\}$ with probabilities $(P_Y(y_1) + P_Y(y_2), P_Y(y_3), \ldots, P_Y(y_m))$. This random variable satisfies the conditions $X \to \tilde{Y} \to Y$ and $H(\tilde{Y}|Y) = 0$. However, $H(\tilde{Y}) < H(Y) = H(W^*)$ which contradicts the minimality of W^*. □

C. Connecting $D_X(Y)$ with $g_0(X;Y)$

Considering the definition of $C_X(Y)$, one can loosely say that all the information contained in Y which is correlated with X is concentrated on $T^{\mathcal{X}}(Y)$, and therefore $D_X(Y)$ represents the amount of information contained in Y and not correlated with X. This suggests that $D_X(Y)$ is equal to $g_0(X;Y)$. In what follows, we study two different cases where $D_X(Y) = g_0(X;Y)$. First we look at the case when $C_X(Y) = I(X;Y)$. We previously showed that this happens if and only if there exists the decomposition $X = (X', V)$ and $Y = (Y', V)$ such that Y' is conditionally independent of X' given V. In this case, $D_X(Y) = H(Y|X)$. It is straightforward to show that in this case we have $g_0(X;Y) \leq H(Y'|V)$. To see this, assume otherwise, that is, suppose that there exists a random variable, say, Z such that $Z \perp\!\!\!\perp X$ and also $I(Y;Z) > H(Y'|V)$. Since

$$I(Y;Z) = I(Y', V; Z) = I(V;Z) + I(Y';Z|V),$$

the assumption $I(Y;Z) > H(Y'|V)$ implies

$$I(V;Z) > H(Y'|V,Z).$$

This contradicts our assumption that $Z \perp\!\!\!\perp X = (X', V)$. Hence we conclude that $g_0(X;Y) \leq H(Y'|V)$. One special case of this decomposition is the case studied by Wyner [10] where X', Y' and V are mutually independent. Consider now the following deterministic function f acting on $Y = (Y', V)$, defined by $f(y) = (y', 0)$. Then we set $Z = f(Y)$ and hence the privacy filter is $P_{Y'|Y}$. By construction we have $Z \perp\!\!\!\perp (V, X')$ and hence $Z \perp\!\!\!\perp X$. Note that $I(Y;Z) = H(Z) = H(Y')$. Since we showed above that $g_0(X;Y) \leq H(Y'|V)$ and since in this special case $H(Y'|V) = H(Y')$, one can conclude that $g_0(X;Y) = H(Y')$. Therefore, in this case the equality $D_X(Y) = g_0(X;Y)$ holds.

The second setting that we examine is the binary case. From Theorem 1 we know that $g_0(X;Y) = 0$ for any binary correlated X and Y.

Suppose $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $Y \sim \mathsf{Bernoulli}(p)$, $P_{X|Y}(\cdot|0) = \mathsf{Bernoulli}(\alpha)$ and $P_{X|Y}(\cdot|1) = \mathsf{Bernoulli}(\beta)$. The condition $H(W|Y) = 0$ implies that there exists a deterministic function $f : \mathcal{Y} \to \mathcal{W}$ with $|\mathcal{W}| \leq |\mathcal{Y}|$ such that $W = f(Y)$ and therefore, $P_{W|Y}(w|y) = 1_{\{w=f(y)\}}$. The only possible cases for $P_{W|Y}$ are

$$P_{W|Y} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and

$$P_{W|Y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where each column corresponds to a value of $Y \in \{0,1\}$. Thus $W \sim \mathsf{Bernoulli}(p)$ or $W \sim \mathsf{Bernoulli}(1-p)$. In either case, $H(W) = H(Y)$. Hence, $C_X(Y) = H(Y)$, i.e., $D_X(Y) = 0$ and thus $g_0(X;Y) = D_X(Y) = 0$ which is what we wanted to show. Note that this argument does not depend on the cardinality of $\mathcal{X}$. In other words, it is impossible to send any information at non-zero rate with zero privacy leakage when $|\mathcal{Y}| = 2$ which is a restatement of Lemma 1.

Although the relation $D_X(Y) = g_0(X;Y)$ holds for the two cases described above, in the following example we have $g_0(X;Y) > D_X(Y)$.

Example 1. Consider X distributed according to $\mathsf{Bernoulli}(p)$ and the binary erasure channel $P_{Y|X}$ with erasure probability δ. The output alphabet is therefore ternary $\{0, e, 1\}$ where e denotes the erasure. Letting $Z = f(Y)$ where f maps $Y = 1$ and $Y = 0$ to 1 and e to 0, we conclude that $g_0(X;Y) \geq h(\delta)$. On the other hand, $H(Y|X) = h(\delta)$ which implies that $g_0(X;Y) = h(\delta)$. Furthermore, Lemma 4 implies that $C_X(Y) = H(Y)$ and thus $D_X(Y) = 0$. Therefore, although $D_X(Y) = 0$, we can extract independent information of X from Y with positive rate.

In general, one can ask under what condition on P_{XY} does the relation $D_X(Y) = g_0(X;Y)$ hold?

V. Conclusion

In this paper we defined a new privacy-utility tradeoff where both privacy and utility are measured in terms of mutual information. The resulting rate-privacy function characterizes the best utility when the privacy leakage is required to be less than ϵ. For the case when $\epsilon = 0$ (perfect privacy) we calculated the rate-privacy function for the binary case. We also introduced a new quantity which quantifies the private information contained in the observable data and examined the connection between this quantity and the rate-privacy function.

References

[1] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 39, pp. 63–69, March 1965.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14

[3] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy and mutual-information privacy," *arxiv:1402.3757*, 2014.

[4] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *arxiv:1210.2085*, 2013.

[5] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 918–923, Nov 1983.

[6] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Medard, "From the information bottleneck to the privacy funnel," *arxiv/1402.1774v4*, 2014.

[7] T. Berger and R. Yeung, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 237–244, Mar 1989.

[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[9] H. S. Witsenhausen, "Values and bounds for the common information of two discrete random variables," *SIAM Journal on Applied Mathematics*, vol. 31, no. 2, pp. 313–333, 1976.

[10] A. Wyner, "The common information of two dependent random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.

[11] P. Gács and J. Körner, "Common information is far less than mutual information," *Probl. Inform. Control*, vol. 2, no. 2, pp. 149–162, 1973. [Online]. Available: http://citeseer.ifi.unizh.ch/context/562456/0

[12] P. Cuff, H. Permuter, and T. Cover, "Coordination capacity," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4181–4206, Sept. 2010.

[13] G. Kumar, C. T. Li, and A. El Gamal, "Exact common information," *arxiv:1402.0062v1*, 2014.

[14] S. Kamath and V. Anantharam, "A new dual to the Gács-Körner common information defined via the Gray-Wyner system," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, Sept 2010, pp. 1340–1346.

[15] H. Tyagi, "Common information and secret key capacity," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5627–5640, Sept. 2013.

[16] S. Wolf and J. Wultschleger, "Zero-error information and applications in cryptography," in *Information Theory Workshop, 2004. IEEE*, Oct. 2004, pp. 1–6.