Privacy-Utility Tradeoff and Privacy Funnel

Salman Salamatian, Flavio P. Calmon, Nadia Fawaz, Ali Makhdoumi, Muriel Médard *

January 16, 2020

Abstract

We consider a privacy-utility trade-off encountered by users who wish to disclose some information to an analyst, that is correlated with their private data, in the hope of receiving some utility. We propose a general framework under which data is transformed according to a probabilistic privacy-preserving mapping before it is disclosed. We show that applying this general framework to the setting where the adversary uses the log-loss cost function naturally leads to a non-asymptotic information-theoretic formulation for characterizing the best achievable privacy subject to utility constraints. This formulation can be cast as a modified rate-distortion problem. We justify the relevance and generality of the privacy metric under the log-loss by proving that the inference threat under any bounded cost function can be upper bounded by an explicit function of the mutual information between private data and disclosed data. We then study connections between our framework and differential privacy. In addition, we show that when the log-loss is used in this framework in both the privacy metric and the utility metric, the average information leakage and the utility constraint can be reduced to the mutual information between private data and disclosed data, and between non-private data and disclosed data, respectively. We then show that the privacy-utility tradeoff under the log-loss can be cast as a non-convex optimization termed Privacy Funnel. Finally, we study the problem of finding optimal privacy-preserving mapping.

Contents

*S. Salamatian, and M. Médard are with the Massachusetts Institute of Technology. Email: salmansa@mit.edu, and medard@mit.edu. F. P. Calmon is with Harvard University Email: flavio@seas.harvard.edu. N. Fawaz is with Pinterest. Email: Nadida.Fawaz@gmail.com. Ali Makhdoumi is with the Fuqua School of Business at Duke University. Email: ali.makhdoumi@duke.edu. This paper was presented in part at the 2012 Allerton Conference on Communication, Control, and Computing , the 2013 Allerton Conference on Communication, Control, and Computing, and the 2014 IEEE Information Theory Workshop.

1 Introduction

1.1 Motivation

Increasing volumes of user data are being collected over wired and wireless networks, by a large number of companies who mine this data to provide personalized services or targeted advertising to users. As a consequence, privacy is gaining ground as a major topic in the social, legal, and business realms. This trend has spurred recent research in the area of theoretical models for privacy, and their application to the design of privacy-preserving services. Most privacy-preserving techniques, such as anonymization, k-anonymity Sweeney (2002) and differential privacy Dwork et al. (2006a), are based on some form of perturbation of the data, either before or after the data is used in some computation. These perturbation techniques provide privacy guarantees at the expense of a loss of accuracy in the computation result (i.e., loss of utility), which leads to a privacy-utility trade-off.

In this paper, we consider the general setting where a user wishes to release a set of measurements to an analyst who provides a service (e.g. a recommendation system), while keeping data that are correlated with these measurements private. On one hand, the analyst is a legitimate receiver for these measurements, from which he expects to derive some utility. On the other hand, the correlation of these measurements with the user's private data gives the analyst the ability to illegitimately infer private information. The tension between the privacy requirements of the user and the utility expectations of the analyst gives rise to the problems of privacy-utility trade-off modeling, and the design of release schemes minimizing the privacy risks incurred by the user, while satisfying the utility constraints of the analyst.

1.2 Contributions

We propose a general statistical inference framework to capture the privacy threat incurred by a user who releases information given certain utility constraints. The privacy risk is modeled as an inference cost gain by a curious adversary upon observing the information released by the user. In broad terms, this cost gain represents the "amount of knowledge" learned by an adversary about the private data after observing the user's released data. The design problem of finding the optimal mapping from the user's information to a privacy-preserving output is formulated as an optimization problem where the cost gain of the adversary is minimized for a given set of utility constraints. This formulation is general and given in terms of minimizing the average, being applicable to different cost functions.

We apply this general framework to the case when the adversary uses the self-information cost function (also commonly referred as log-loss cost). We show how this naturally leads to a non-asymptotic information-theoretic framework to characterize the information leakage subject to utility constraints. Based on these results we introduce a privacy metric termed *average information leakage*. We also demonstrate that the problem of designing a privacy preserving mechanism that achieves the optimal privacy-utility tradeoff can be cast as modified rate-distortion problems. We justify the relevance and generality of the privacy metric under the log-loss by proving that the inference threat under any bounded cost function can be upper bounded by an explicit function of the mutual information between private data and disclosed data.

We compare the average information leakage with differential privacy. We show that differential privacy does not provide in general *any* privacy guarantees in terms of average information leakage. Furthermore, we show that local differential privacy provides a bound on the average information leakage. We then show that, when the log-loss is introduced in this framework in both the privacy metric and the distortion metric, the privacy leakage reduces to the mutual information between private data and disclosed data, while the utility requirement is modeled by the mutual information between non-private data and disclosed data. We then

show that the privacy-utility tradeoff under the log-loss can be cast as the Privacy Funnel optimization, and study its connection to the Information Bottleneck Tishby et al. (2000).

Finally, we focus on the design of privacy-preserving mapping. We first show that if the distribution of the private data is not available, then the privacy-utility trade-off reduces to that of rate-distortion. We then consider the case in which the distribution of the private data is available and show that the general modified rate-distortion problem can be expressed as a convex program. As a consequence, the privacy preserving mapping that achieves the optimal privacy-utility tradeoff can be efficiently found using convex minimization algorithms or widely available convex solvers. On the other hand, for general distributions, the privacy funnel optimization is a non-convex problem, so provide a greedy algorithm for the Privacy Funnel that is locally optimal by leveraging connections to the Information Bottleneck method Slonim and Tishby (1999); Tishby et al. (1999), and evaluate its performance on real-world data.

1.3 Related Works

Privacy-utility tradeoffs have been studied under either a local privacy setting, or a centralized privacy setting. In the local privacy setting, users do not trust the entity aggregating data. Thus, each user holds her data locally, and processes it according to a privacy-preserving mechanism before releasing it to the aggregator. Local privacy dates back to randomized response in surveys Warner (1965), and has been considered in privacy for data mining and statistics Agrawal and Srikant (2000); Mishra and Sandler (2006); Evfimievski et al. (2003); Rebollo-Monedero et al. (2010); Kasiviswanathan et al. (2011); Banerjee et al. (2012); Duchi et al. (2013). The setup we consider falls under the local privacy setting, since the analyst is assumed to be untrusted, and users wish to protect against statistical inference of private information from data they release to the analyst. In contrast, the framework we study models non-asymptotic privacy guarantees in terms of the inference cost gain that an adversary achieves by observing the released output. Local privacy has also been considered in the differential privacy Dwork et al. (2006a); Dwork (2006b) corpus, e.g. for learning concept classes Kasiviswanathan et al. (2011), training clustering algorithms Banerjee et al. (2012), and statistical parameter estimation Duchi et al. (2013), from data distorted locally by users. These works are concerned with the problem of learning aggregate statistical properties from the data of several users. In contrast, we focus on providing utility to an individual user while maintaining the privacy of this individual user's attributes.

In the centralized privacy setting, a trusted entity aggregates data from users in a database, while an untrusted analyst asks queries on the database. The trusted aggregator jointly processes data from multiple users according to a centralized privacy-preserving mechanism to produce a privatized answer to the query, that is released to the analyst. The centralized privacy setting is less stringent than the local privacy setting. Information theoretic frameworks have been used to analyze privacy-utility tradeoffs in the centralized database setting. One line of work Reed (1973); Yamamoto (1983b); Sankar et al. (2013) focuses mainly on collective privacy for all or subsets of the entries of a data base, and provide fundamental and asymptotic results on the rate-distortion-equivocation region as the number of data samples grows arbitrarily large. Traditionally, many differential privacy works assumed a centralized setting with a trusted database owner, and focused on making the output of an application running on the database differentially private, e.g. data mining Friedman and Schuster (2010), social recommendations Machanavajjhala et al. (2011), recommender systems McSherry and Mironov (2009), as well as algorithms for statistical estimators Smith (2011); Dwork and Lei (2009), classifiers Chaudhuri et al. (2011); Rubinstein et al. (2009), principal component analysis Chaudhuri et al. (2012), etc. More specifically, McSherry and Mironov (2009) considers the case of a trusted recommender system who has access to ratings from privacy-conscious users, and addresses the challenge of training a differentially-private recommendation algorithm based on these original ratings. In contrast, we study a local privacy setup where the analyst is not trusted by privacy-conscious users, who wish to protect against statistical inference of private information from data they release to the analyst.

Our paper relates to a vast literature on the study of differential privacy introduced in Dwork et al. (2006a); Dwork (2006b). Differential privacy is studied in many contexts including mechanism design McSherry and Talwar (2007); Ghosh and Roth (2015), learning theory Dwork et al. (2010); Blum et al. (2013); Duchi et al. (2013), and data mining Banerjee et al. (2012); Dwork et al. (2015b,a) (see Dwork et al. (2014) for a survey of results). Moreover, Agrawal and Srikant (2000); McGregor et al. (2010) study the class of adding distortion to the public data to protect privacy and Sweeney (2002); Wang et al. (2004) study the use of $k-$anonymity

to mask private information in classification. We compare our framework to that of differential privacy in Section 3.

Several approaches rely on information-theoretic tools to model privacy-utility trade-offs, such as Reed (1973); Yamamoto (1983a); Evfimievski et al. (2003); Sankar et al. (2013). Indeed, information theory, and more specifically rate-distortion theory, appear as natural frameworks to analyze the privacy-utility trade-off resulting from the distortion of correlated data. Although the approach we introduce in this paper involves information theoretic metrics, it is fundamentally different from previous information theoretic privacy models. Indeed, traditional information theoretic privacy models, such as Yamamoto (1983a); Sankar et al. (2013, 2010), focus on collective privacy for all or subsets of the entries of a database, and provide asymptotic guarantees on the average remaining uncertainty per database entry – or equivocation per input variable – after the output release. More precisely, the average equivocation per entry is modeled as the conditional entropy of the input variables given the released output, normalized by the number of input variables. In contrast, the general framework introduced in this paper provides privacy guarantees in terms of bounds on the inference cost gain that an adversary achieves by observing the released output. The use of a self-information cost yields a non-asymptotic information theoretic framework modeling the privacy risk in terms of information leakage. This framework, in turn, can be used to design practical privacy preserving mappings.

Finally, mutual information as a measure of privacy has been used in the literature (see, e.g., Chatzikokolakis et al. (2010); Zhu and Bettati (2005); Chatzikokolakis et al. (2008)), mostly under the context of quantitative information flow and anonymity systems. The connections between different privacy notions have been studied recently, e.g., Alvim et al. (2011); Mir (2012); Wang et al. (2016). Several works have studied a rate-distortion approach to privacy including Sarwate and Sankar (2014); Asoodeh et al. (2014); Moraffah and Sankar (2015); Basciftci et al. (2016); Asoodeh et al. (2016); Bonomi et al. (2016). More recently, generalizations to the privacy-utility trade-offs have been considered, e.g. Rassouli and Gunduz (2019) measures the privacy leakage in terms of total variation; Asoodeh et al. (2017); Osia et al. (2019) consider privacy against guessing attacks; Liao et al. (2019, 2018) study privacy guarantees under α-maximum leakage; Liao et al. (2017); Li et al. (2018); Sreekumar et al. (2018) are concerned with privacy against an adversary performing a hypothesis test; the estimation formulations of the privacy utility trade-offs have also been extensively considered in Asoodeh et al. (2018); Wang and Calmon (2017); Wang et al. (2019).

1.4 Outline

In Section 2, we introduce the threat model and formulate the general privacy-utility trade-off, and specialize it to the *log-loss* case. The privacy metric with log-loss cost function is termed average information leakage. We then show that privacy guarantees with log-loss cost function (i.e., small average information leakage) provides a performance guarantee under any bounded cost function. In addition, we show that if the average information leakage is small, then the probability of error in inferring the private data becomes large. In Section 3 we compare differential privacy with average information leakage. In Section 4, we consider the case where the distortion is also computed in terms of log-loss, which leads to the so-called *Privacy Funnel* optimization problem. Finally, in Section 5, we consider the problem of finding the optimal privacy-preserving mapping and show that it is convex. We then show that the privacy funnel problem is non-convex and provide a greedy-algorithm for it inspired by algorithms that solve the information bottleneck method. We conclude the paper in Section 6.

Previous Publications: Parts of this manuscript have been published in Calmon and Fawaz (2012); Makhdoumi and Fawaz (2013); Makhdoumi et al. (2014). This publication distinguishes itself in three main ways: (i) it contains a substantial set of examples and simulations throughout the paper which are missing in the above submissions, (ii) contains all the complete proofs as well as some additional technical lemmas such as Lemma 1, Proposition 1, Lemma 4, and (iii) gives a much more complete view on the privacy against inference problem, and thus serves as a main reference for researchers and practitioners interested in designing privacy systems.

1.5 Notation

Matrices are denoted by upper-case bold letters (e.g. $\mathbf{A}$) and column vectors by lower-case bold letters We denote by $\mathbf{1}$ the vector with all entries equal to 1, and the dimension of $\mathbf{1}$ will be clear from the context. Sets are denoted by calligraphic letters (e.g. $\mathcal{A}$), with the exception $[k] \triangleq \{1, \dots, k\}$ for a positive integer k, and the n-dimensional simplex

$$\Delta_n \triangleq \left\{ \mathbf{x} \in \mathbb{R}^n \middle| \sum_{i=1}^{n} x_i = 1,\ x_i \geq 0 \right\}.$$

The convex hull of a set $\mathcal{A}$ is denoted by $\mathsf{conv}\mathcal{A}$, and the boundary of a set is denoted by $\partial\mathcal{A}$.

Let X and Y be two (discrete) random variable, with joint distribution p_{XY} and support $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that $|\mathcal{X}|, |\mathcal{Y}| < \infty$. We assume without loss of generality that $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$. We denote by $\mathbf{p} \in \Delta_m$ the column vector with entries $\mathbf{p} = [P_X(1), \cdots, P_X(m)]$ (equivalently $\mathbf{q} \in \Delta_n$ for P_Y). In addition, we denote by $\mathbf{T} \in \mathbb{R}^{n \times m}$ the column-stochastic matrix whose entries are the channel transformation $P_{Y|X}$, i.e. $[\mathbf{T}]_{i,j} = P_{Y|X}(i|j)$. Observe that $\mathbf{q} = \mathbf{T}\mathbf{p}$. We denote $X \sim P_X$ and $X \sim \mathbf{p}$ interchangeably.

For $\mathbf{p} \in \Delta_m$, we denote by h_m the entropy function, i.e., $h_m : \Delta_m \to \mathbb{R}$ with $h_m(\mathbf{p}) = -\sum_{i \in [m]} p_i \log p_i$ with the usual convention that $0 \log 0 = 0$. For $a \in [0, 1]$, we let $\bar{a} \triangleq 1 - a$. The binary entropy function is given by $h_b(x) \triangleq h_2([x, \bar{x}])$. When $p_{X,Y,Z}(x, y, z) = p_X(x) p_{Y|X}(y|x) p_{Z|Y}(z|y)$ (i.e. X is conditionally independent of Z given Y), we write $X \to Y \to Z$.

For a random variable X with discrete support and $X \sim p_X$, the entropy of X is given by

$$H(X) \triangleq -\mathbb{E}\left[\log\left(p_X(X)\right)\right].$$

For $X, Y \sim p_{XY}$, the mutual information between X and Y is

$$I(X;Y) \triangleq \mathbb{E}\left[\log\left(\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right)\right].$$

For any two pmfs over $\mathcal{X}$ such as p_X and q_X the KL-Divergence is defined as

$$D(p||q) \triangleq \mathbb{E}_{X \sim p_X}\left[\log\left(\frac{p_X(X)}{q_X(X)}\right)\right].$$

The basis of the logarithm will be clear from the context. For any real-valued random variable X, we denote the L_p-norm of X as

$$\|X\|_p \triangleq \left(\mathbb{E}|X|^p\right)^{1/p}.$$

2 General Setup for Privacy-utility trade-offs

In this section we outline the general setup considered in this paper, and some threats models it addresses. Our formulation is general and encompasses many existing privacy-utility scenarios, which are designed under a specific setting. Despite its generality, many of these formulations share similar properties, which will become apparent from the viewpoint we propose.

2.1 General setup

We assume that there are two parties that communicate over a noiseless channel, namely Alice and Bob. Alice has access to a set of measurement points, represented by the variable $X \in \mathcal{X}$, that she wishes to transmit to Bob. At the same time, Alice requires that a set of variables $S \in \mathcal{S}$ should remain private, where S is jointly distributed with X according to the distribution $(X, S) \sim p_{X,S}(x, s)$, $(x, s) \in \mathcal{X} \times \mathcal{S}$. Depending on the considered setting, the variable S can be either directly accessible to Alice or inferred from X. If no privacy mechanism was in place, Alice would simply transmit X to Bob.

Bob has a utility requirement for the information sent by Alice. Furthermore, Bob will try to learn S from Alice's transmission. Alice's goal is to find and transmit a distorted version of X, denoted by $Y \in \mathcal{Y}$, such that Y satisfies a target utility constraint, but "protects" (in a sense made more precise later) the private

variable S. We assume that Bob is passive but computationally unbounded, and will try to infer S based on Y.

We consider, without loss of generality, that $S \to X \to Y$. Note that this model can capture the case where S is directly accessible by Alice by appropriately adjusting the alphabet $\mathcal{X}$. For example, this can be done by representing $S \to Y$ as an injective mapping or allowing $\mathcal{S} \subset \mathcal{X}$. In other words, even though the privacy mechanism is designed as a mapping from $\mathcal{X}$ to $\mathcal{Y}$, it is not limited to an output perturbation, and it encompasses input perturbation settings.

Definition 1. A privacy-preserving mapping is a transition probability $p_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}$. A distortion, or utility measure, is a function $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. We say a privacy mapping $p_{Y|X}$ has D-distortion for some $D \geq 0$, if $\mathbb{E}[d(X,Y)] \leq \delta$ when $(X,Y) \sim p_X p_{Y|X}$.

Throughout the paper we make the following assumptions:

1. Alice and Bob know the prior distribution of $p_{X,S}(\cdot)$. This represents the side information that an adversary has. In Section 5.2, we relax this assumption to the case where only p_X is known.
2. Bob has complete knowledge of the privacy preserving mapping, i.e., g and $p_{Y|X}(\cdot)$ are known.

Note that this represents the *worst-case* statistical side information that an adversary can have about the input.

2.2 Threat model

We assume that Bob selects a revised distribution $q \in \mathcal{P}_S$, where $\mathcal{P}_S$ is the set of all probability distributions over $\mathcal{S}$, in order to minimize an expected cost $C(S,q)$. The cost $C : \mathcal{S} \times \mathcal{P}_S \to \mathbb{R}^+$ models the statistical risk or cost, of picking an estimator q to estimate the random variable S. In other words, the adversary chooses q as the solution of the minimization

$$c_0^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_S[C(S,q)]$$

prior to observing Y, and

$$c_y^* = \min_{q \in \mathcal{P}_S} \mathbb{E}_{S|Y}[C(S,q)|Y=y]$$

after observing the output Y. Note that this restriction on Bob models a very broad class of adversaries that perform statistical inference, capturing how an adversary acts in order to infer a revised belief distribution over the private variables S when observing Y. After choosing this distribution, the adversary can perform an estimate of the input distribution (e.g. using a MAP estimator). However, the quality of the inference is inherently tied to the revised distribution q.

The average cost gain by an adversary after observing the output is

$$\Delta C = c_0^* - \mathbb{E}_Y[c_y^*]. \tag{2.1}$$

We also mention that one can represent similarly the maximum cost gain by an adversary in terms of the most informative output (i.e. the output that give the largest gain in cost), which gives:

$$\Delta C^* = c_0^* - \min_{y \in \mathcal{Y}} c_y^*. \tag{2.2}$$

In the next section we present a formulation for the privacy-utility tradeoff based on this general setting.

2.3 A general formulation for the privacy-utility tradeoff

Our goal is to design privacy preserving mappings that minimize ΔC for a given distortion level D, characterizing the fundamental privacy-utility tradeoff. More precisely, our focus is to solve optimization problems over $p_{Y|X} \in \mathcal{P}_{Y|X}$ of the form

$$\begin{aligned} &\min\ \ \Delta C \\ &\text{s.t.}\ \ \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta\ , \end{aligned} \tag{2.3}$$

where $\mathcal{P}_{Y|X}$ is the set of all conditional probability distributions of Y given X.

Remark 1. In the remainder of the paper we consider only one distortion constraint. However, it is straightforward to generalize the formulation and the subsequent optimization problems to multiple distinct distortion constraints $\mathbb{E}_{X,Y}[d_1(X,Y)] \leq \delta_1, \ldots, \mathbb{E}_{X,Y}[d_n(X,Y)] \leq \delta_n$. This can be done by simply adding an additional linear constraint to the optimization problem.

The formulation introduced above is general and can be applied to different cost functions in principle. Throughout the paper, we specialize this formulation to the log-loss, or self-information cost. We will show subsequently how the log-loss can be used to bound any other loss function. In addition to its generality, the log-loss has additional convenient advantages. Namely, it is a local, proper and differentiable loss, which will, as we will see, lead to a convex optimization formulation for privacy-utility trade-offs. For an overview of the central role of the self-information cost function in prediction, we refer the reader to Merhav and Feder (1998). Nevertheless, it is important to emphasize that many of the results presented in this paper hold for more general loss functions, at the expense of additional notation.

The *self information* (or *log-loss*) cost function is given by

$$C(S,q) = -\log q(S).$$

It is straightforward to show that for the log-loss function $c_0^* = H(S)$ and, consequently, $c_y^* = H(S|Y=y)$, and, therefore

$$\Delta C = I(S;Y) = \mathbb{E}_Y[D(p_{S|Y}||p_S)],$$

From this definition, the optimal privacy-preserving mapping (the one with privacy $G_d(\delta, p_{S,X})$) is the solution of the minimization

$$\begin{aligned} \min_{p_{Y|X}} \quad & I(S;Y) \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.4}$$

In extreme cases, we say a privacy-mapping has full privacy if $I(S;Y) = 0$ (which implies the released random variable, Y, is independent from the private random variable, S), and no privacy if $I(S;Y) = H(S)$ (implies that S is fully recoverable from Y).

Observe that finding the mapping $p_{Y|X}(y|x)$ that provides the minimum information leakage is a modified rate-distortion problem. Alternatively, we can rewrite this optimization as

$$\begin{aligned} \min_{p_{Y|X}} \quad & \mathbb{E}_Y[D(p_{S|Y}||p_S)] \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.5}$$

The minimization (2.7) has an interesting and intuitive interpretation. If we consider KL-divergence as a metric for the distance between two distributions, (2.7) states that the revised distribution after observing Y should be as close as possible to the a priori distribution.

It is straightforward to show that for the log-loss function $c_0^* = H(S)$ and, consequently, $c_y^* = H(S|Y=y)$, and, therefore

$$\Delta C = I(S;Y) = \mathbb{E}_Y[D(p_{S|Y}||p_S)],$$

From this definition, the optimal privacy-preserving mapping (the one with privacy $G_d(\delta, p_{S,X})$) is the solution of the minimization

$$\begin{aligned} \min_{p_{Y|X}} \quad & I(S;Y) \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.6}$$

In extreme cases, we say a privacy-mapping has full privacy if $I(S;Y) = 0$ (which implies the released random variable, Y, is independent from the private random variable, S), and no privacy if $I(S;Y) = H(S)$ (implies that S is fully recoverable from Y).

Observe that finding the mapping $p_{Y|X}(y|x)$ that provides the minimum information leakage is a modified rate-distortion problem. Alternatively, we can rewrite this optimization as

$$\begin{aligned} \min_{p_{Y|X}} \ & \mathbb{E}_Y[D(p_{S|Y}||p_S)] \\ \text{s.t.} \ & \mathbb{E}_{X,Y}[d(X,Y)] \leq \delta \ . \end{aligned} \tag{2.7}$$

The minimization (2.7) has an interesting and intuitive interpretation. If we consider KL-divergence as a metric for the distance between two distributions, (2.7) states that the revised distribution after observing Y should be as close as possible to the a priori distribution.

We are now ready to define the privacy-utility region.

Definition 2. For $D \geq 0$, distortion measure $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, and a joint distribution $p_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the optimal *privacy-utility function* $G_d(D, p_{S,X})$ as

$$G_d(D, p_{S,X}) \triangleq \inf \{I(S;Y) : \ \mathbb{E}[d(X,Y)] \leq D, S \to X \to Y\}, \tag{2.8}$$

where the infimum is over all mappings $p_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $p_{S,X}$ and $D \geq 0$, the set of pairs $\{(D, G_d(D, p_{S,X}))\}$ is called the *privacy-utility region* of $p_{S,X}$.

We next characterize a property of the optimal privacy-preserving mapping which will be useful in Section 5 to construct solutions to the optimization problem 2.6. In particular, the next lemma suggests that the size of the output alphabets $|\mathcal{Y}|$ one need to consider is bounded by $|\mathcal{X}| + 1$. This lemma will be used in Section 5 when we find to design algorithms to find the optimum privacy-preserving mapping.

Lemma 1. We have

$$G_d(D, p_{S,X}) = \min_{p_{Y|X}} \{I(S;Y) \ : \mathbb{E}[d(X,Y)] \leq D, S \to X \to Y, \ |\mathcal{Y}| \leq |\mathcal{X}| + 1\}.$$

Proof. Let $p_{S,X}$ and $p_{Y|X}$ be given, with $S \to X \to Y$. Denote by $\mathbf{w}_i$ the vector in the $|\mathcal{X}|$-simplex with entries $p_{X|Y}(\cdot|i)$. Furthermore, let $a_i \triangleq \mathbb{E}[d(X,Y)|Y=i]$, and $b_i \triangleq H(S) - H(S|Y=i)$. Therefore

$$\sum_{i=1}^{|\mathcal{Y}|} p_Y(i) \, [\mathbf{w}_i, a_i, b_i] = [p_X, \mathbb{E}[d(X,Y)], I(S;Y)] \, . \tag{2.9}$$

Since $\mathbf{w}_i$ belongs to the $|\mathcal{X}|$-simplex, the vector $[\mathbf{w}_i, a_i, b_i]$ is taken from a connected, compact $|\mathcal{X}| + 1$ dimensional space. Then, from Fenchel-Eggleston strengthening of Carathéodory's theorem (Eggleston, 2009, Theorem 18, pg. 35), the point $[\mathbf{p}_X, \mathbb{E}[d(X,Y)], \Delta C]$ can also be achieved by at most $|\mathcal{X}| + 1$ non-zero values of $p_Y(i)$. It follows directly that it is sufficient to consider $|\mathcal{Y}| \leq |\mathcal{X}| + 1$ for the infimum (4.2). The set of all mappings $p_{Y|X}$ for $|\mathcal{Y}| \leq |\mathcal{X}| + 1$ is compact, and both $p_{Y|X} \to I(S;Y)$ and $p_{Y|X} \to \mathbb{E}[d(X,Y)]$ are continuous and bounded when S, X and Y have finite support. Consequently, the infimum in (4.2) is attainable. □

Next, we give an example of the optimization given in (2.7) and its solution.

Example 1. Let S be a Bernoulli($\frac{1}{2}$) distribution and X be the result of S passing through a BSC(p) channel where $p \leq \frac{1}{2}$. Suppose the distortion measure is hamming distortion, i.e. $\mathbb{E}[d(X,Y)] = \mathbb{P}[X \neq Y]$, and consider the log-loss. We claim that in this setting for a given $\delta \in (0,1)$, we have

$$G_d(\delta, p_{S,X}) = 1 - h_b(p * \delta),$$

where $p * \delta = p(1-\delta) + (1-p)\delta$. First, note that using the privacy-preserving mapping, $p_{Y|X}$, given by $Y = X \oplus Z$, where Z has a Bern(δ) distribution, we have $\mathbb{E}[d(X,Y)] \leq \delta$ and $I(S;Y) = 1 - h(p * \delta)$. This shows that $G_d(\delta, p_{S,X}) \leq 1 - h_b(p * \delta)$. Next, we show that $G_d(\delta, p_{S,X}) \geq 1 - h_b(p * \delta)$. We have $I(S;Y) = H(S) - H(S|Y) = 1 - H(S \oplus Y|Y) \geq 1 - H(S \oplus Y)$. Using Markov property, it follows that $\mathbb{P}[S \oplus Y = 1] \leq p * \delta$, which completes the proof of the claim. Now suppose we want to have full privacy. Given $G_d(D, p_{S,X}) = 1 - h_b(p * D)$, full privacy is possible only in the following two cases:

1. $p = \frac{1}{2}$, implying X is independent from S. In this case, there is no privacy problem to begin with.
2. $\delta = \frac{1}{2}$, implying Y is independent from X. In this case, full privacy implies no utility is preserved in the released data.

2.3.1 Generality of log-loss as a privacy metric

In this section, we focus on the threat model under the log-loss cost function and show its generality. In particular, we establish that for any bounded cost function $C(S,q)$, the associated inference cost gain ΔC can be upperbounded by an explicit constant factor of $\sqrt{I(S;Y)}$. Thus, controlling the cost gain under the log-loss, so that it does not exceed a target privacy level, is sufficient to ensure that the privacy threat under a different bounded cost function would also be controlled. Therefore, the design of the privacy mapping can be focused on minimizing the privacy leakage as measured by $I(S;Y)$.

Theorem 1. Let $L = \sup_{s\in\mathcal{S}, q\in\mathcal{P}_S} |C(s,q)| < \infty$. We have $\Delta C = c_0^* - \mathbb{E}_{p_Y}[c_Y^*] \leq 2\sqrt{2}L\sqrt{I(S;Y)}$.

The proof of Theorem 1 requires the following lemma.

Lemma 2. Let $C(s,q)$ be a bounded cost function such that $L = \sup_{s\in\mathcal{S}, q\in\mathcal{P}_S} |C(s,q)| < \infty$. For any given $y \in \mathcal{Y}$,

$$\mathbb{E}_{p_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y] \leq 2\sqrt{2}L\sqrt{D(p_{S|Y=y}||p_S)},$$

where q_0^* and q_y^* are the maximizing distributions for c_0^* and c_y^* as defined in Section 2.2, respectively.

Proof. we have

$$\begin{aligned}
&\mathbb{E}_{p_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y] = \sum_s p(s|y)[C(s,q_0^*) - C(s,q_y^*)] \\
&= \sum_s (p_{S|Y}(s|y) - p_S(s) + p_S(s))[C(s,c_0^*) - C(s,q_y^*)] = \sum_s (p_{S|Y}(s|y) - p_S(s))[C(s,q_0^*) - C(s,q_y^*)] \\
&+ \sum_s p(s)[C(s,q_0^*) - C(s,q_y^*)] \leq 2L\sum_s |p(s|y) - p(s)| + (\mathbb{E}_{p_S}[C(S,q_0^*)] - \mathbb{E}_{p_S}[C(S,q_y^*)]), \\
&\leq 2L\sum_s |p_{S|Y}(s|y) - p_S(s)| = 4L||p_{S|Y=y} - p_S||_{TV} \leq 4L\sqrt{\frac{1}{2}D(p_{S|Y=y}||p_S)},
\end{aligned}$$

where we used that $C(s,q_0^*) - C(s,q_y^*) \leq 2L$ and $\mathbb{E}_{p_S}[C(S,q_0^*)] - \mathbb{E}_{p_S}[C(S,q_y^*)] \leq 0$. And the last inequality follows from using Pinsker's inequality (Csiszár and Körner, 2011, Problem 3.18) (where the log in the definition of divergence is natural log). □

We now prove Theorem 1.

proof of Theorem 1. We have

$$\begin{aligned}
\Delta C &= \mathbb{E}_{p_S}[C(S,q_0^*)] - \mathbb{E}_{p_Y}\left[\mathbb{E}_{p_{S|Y}}[C(S,q_y^*)|Y=y]\right] \\
&= \mathbb{E}_{p_Y}\left[\mathbb{E}_{p_{S|Y}}[C(S,q_0^*) - C(S,q_y^*)|Y=y]\right] \\
&\leq 2\sqrt{2}L\mathbb{E}_{p_Y}\left[D(p_{S|Y=y}||p_S)\right] \leq 2\sqrt{2}L\sqrt{I(S;Y)},
\end{aligned}$$

where the last step follows from concavity of square root function and the one before that follows from Lemma 2. □

2.3.2 Inference Defeat through Privacy

One natural and related question is whether a privacy mapping which is designed to minimize average information leakage, privacy, by solving problem (2.6), also provides guarantees on the probability of correctly inferring S from the observation of Y, using any inference algorithm. Next, we show a lower bound on the error probability in inferring S from Y, based on a bound on privacy, using Fano's inequality.

Proposition 1. Assume $|\mathcal{S}| > 2$ and $I(S;Y) \leq \epsilon H(S)$, for some $\epsilon \in [0,1]$. Let $\hat{S}$ be an estimator of S based on the observation Y (possibly randomized). We have

$$p_e \triangleq \mathbb{P}[\hat{S}(Y) \neq S] \geq \frac{(1-\epsilon)H(S) - 1}{\log(|\mathcal{S}| - 1)}.$$

For $|\mathcal{S}| = 2$, we have $h(p_e) \geq (1-\epsilon)H(S)$.

Proof. Denote $p_e = \mathbb{P}[\hat{S}(Y) \neq S]$. From Fano's inequality Cover and Thomas (2012), Theorem 2.10.1, we have

$$p_e \left(\log(|\mathcal{S}| - 1)\right) \geq H(S|Y) - h(p_e).$$

Since $I(Y;S) = H(S) - H(S|Y) \leq \epsilon H(S)$, we have $H(S|Y) \geq (1-\epsilon)H(S)$. Therefore,

$$p_e \geq \frac{(1-\epsilon)H(S) - h(p_e)}{\log(|\mathcal{S}| - 1)} \geq \frac{(1-\epsilon)H(S) - 1}{\log(|\mathcal{S}| - 1)}.$$

The proof when $|\mathcal{S}| = 2$ is similar. □

Note that one can obtain tighter bounds than the one in Proposition 1 by considering β-conditional entropies as the privacy metric, as shown in Sason and Verdú (2017). In particular, as β goes to ∞, the bound becomes tight as the loss considered becomes the 0-1 loss (see Example **??**).

2.4 Application examples

We illustrate next how the proposed model can be cast in terms of privacy preserving queries and hiding features within data sets.

2.4.1 Privacy-preserving queries to a database

The framework described above can be applied to database privacy problems, such as those considered in differential privacy. In this case we denote the private variable as a vector $\mathbf{S} = S_1, \ldots, S_n$, where $S_j \in \mathcal{S}$, $1 \leq j \leq n$ and $S_1, \ldots, S_n$ are discrete entries of a database that represent, for example, the entries of n users. A (not necessarily deterministic) function $f : \mathcal{S}^n \to \mathcal{X}$ is calculated over the database with output X such that $X = f(S_1, \ldots, S_n)$. The goal of the privacy preserving mapping is to present a query output Y such that the individual entries $S_1, \ldots, S_n$ are "hidden", i.e. the estimation cost gain of an adversary is minimized according to the previous discussion, while still preserving the utility of the query in terms of the target distortion constraint. We illustrate this case with the counting query, which will be a recurring example throughout the rest of this paper.

Example 2 (Counting query)**.** Let $S = (S_1, \ldots, S_n)$, where S_i's are the entries in a database, and define:

$$X = f(S_1, \ldots, S_n) = \sum_{i=1}^{n} \mathbf{1}_A(S_i), \tag{2.10}$$

where

$$\mathbf{1}_A(z) = \begin{cases} 1 & \text{if } z \text{ has property } A, \\ 0 & \text{otherwise.} \end{cases}$$

In this case there are two possible approaches: (i) output perturbation, where X is distorted directly to produce Y, and (ii) input perturbation, where each individual entry S_i is distorted directly, resulting in a new query output Y. In particular, if each database input S_i, $1 \leq i \leq n$ satisfies $\mathbb{P}\left[\mathbf{1}_A(S_i) = 1\right] = p$ and are independent and identically distributed. Then X is a binomial random variable with parameter (n, p). It follows that $H(S|X = x) = \log \binom{n}{x}$. Consequently, the optimal privacy preserving mapping will be the one that results in a posterior probability $p_{X|Y}(x|y)$ that is proportional to the size of the pre-image of x, i.e. $p_{X|Y}(x|y) \propto |f^{-1}(x)| = \binom{n}{x}$.

2.4.2 Hiding dataset features

Another important particularization of the proposed framework is the obfuscation of a set of features S by distorting the entries of a data set X. In this case $|\mathcal{S}| \ll |\mathcal{X}|$, and S represents a set of features that might be inferred from the data X, such as age group or salary. The distortion can be defined according to the the utility of a given statistical learning algorithm (e.g. a recommendation system) used by Bob.

3 Comparison of Privacy Metrics

In this section, we compare the existing privacy measures in the literature with each other and also with our measures of privacy. In particular, we compare average information leakage privacy (or privacy in short) and differential privacy and show that, while privacy guarantees a small probability of inferring private random variable based on the released data (Proposition 1), differential privacy does not necessarily guarantee that.

Definition 3 (Differential Privacy).

- *Differential privacy*(see Dwork et al. (2006b)): For a given ϵ, $p_{Y|S}$ is $\epsilon-$ differentially private if

$$\sup_{y,s,s':\ s\sim s'} \frac{p(y\in A|s)}{p(y\in A|s')} \leq e^{\epsilon},$$

  for any measurable set A, where $s \sim s'$ denotes that s and s' are neighbors. The notion of neighboring can have multiple definitions, e.g. Hamming distance 1 (differ in a single coordinate) and ℓ_p ball.

- *local differential privacy* (see Kasiviswanathan et al. (2011)): For a given ϵ, $p_{Y|S}$ is $\epsilon-$locally differential private if

$$\sup_{y,s,s'} \frac{p(y\in A|s)}{p(y\in A|s')} \leq e^{\epsilon},$$

  for any measurable set A and s and s'. Note that this definition is stronger than differential privacy, because we relaxed the neighboring assumption.

- *Information privacy* : For a given ϵ, $P_{Y|S}$ is $\epsilon-$information private if

$$e^{-\epsilon} \leq \frac{p(s\in B|y\in A)}{p(s\in B)} \leq e^{\epsilon},$$

  for any measurable sets A and B.

In the next theorem we first show that local differential privacy implies average information leakage privacy. We then show that differential privacy *does not guarantee* average information leakage privacy *in general*. More specifically, guaranteeing that a mechanism is ϵ-differentially private *does not* provide *any* guarantee on the average information leakage. We also compare other measures of privacy.

Theorem 2. (a) If $p_{Y|S}$ is $\epsilon-$ locally differential private then it is $\epsilon-$average information private. Moreover, for every $\epsilon > 0$ and $\delta \geq 0$, there exists a (S, Y) such that $p_{Y|S}$ is ϵ-differentially private but it is not δ-average information private.

(b) $\epsilon-$ local differential privacy implies $\epsilon-$ information privacy and $\epsilon-$ information privacy implies $2\epsilon-$ local differential privacy.

Proof. **Part (a):** The first part follows because $\epsilon-$ local differential privacy implies

$$\frac{p_{S|Y}(s|y)}{p_S(s)} = \frac{p_{Y|S}(y|s)}{p_Y(y)} = \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s')p_S(s')} \leq \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s)p_S(s')e^{-\epsilon}} \leq e^{\epsilon}, \quad \forall s\in\mathcal{S}, y\in\mathcal{Y},$$

which in turn leads to

$$I(S;Y) = \mathbb{E}_{p_{SY}}\left[\log\frac{p_{SY}}{p_S p_Y}\right] = \mathbb{E}_{p_{SY}}\left[\log\frac{p_{Y|S}}{p_Y}\right] \leq \epsilon.$$

We prove the second part by explicitly constructing an example that is ϵ-differentially private, but an arbitrarily large amount of information can leak on average from the system. For this, we return to the counting query discussed in Example 2. We also use Hamming distance being 1 as the definition of neighboring. In particular, we let $X = \sum_{i=1}^{n} \mathbf{1}_A(S_i)$ and $\mathcal{Y} = \mathcal{X}$. We do not assume independence of the inputs. For the counting query and for any given prior, adding Laplacian noise to the output provides ϵ-differential privacy

Dwork (2006a). More precisely, for the output of the query given in (2.10), denoted as $X \sim p_X(x), 0 \leq x \leq n$, the mapping

$$Y = X + N, \quad N \sim \text{Lap}(1/\epsilon), \tag{3.1}$$

where the pdf of the additive noise N given by

$$p_N(r;\epsilon) = \frac{\epsilon}{2}\exp(-|r|\epsilon),$$

is ϵ-differentially private. Now assume that ϵ is given, and denote $S = (S_1, \ldots, S_n)$. Set k and n such that $n \mod k = 0$, and let $p_S(\cdot)$ be such that

$$p_X(x) = \begin{cases} \frac{1}{1+n/k} & \text{if } x \mod k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

With the goal of lower-bounding the information leakage, assume that Bob, after observing Y, maps it to the nearest value of x such that $p_X(x) > 0$, i.e. does a maximum a posteriori estimation of X. The probability that Bob makes a correct estimation (and neglecting edge effects), denoted by $\alpha_{k,n}(\epsilon)$, is given by:

$$\alpha_{k,n}(\epsilon) = \int_{\frac{-k}{2}}^{\frac{k}{2}} \frac{\epsilon}{2}\exp(-|x|\epsilon)dx = 1 - \exp\left(-\frac{k\epsilon}{2}\right).$$

Let E be a binary random variable that indicates the event that Bobs makes a wrong estimation of X given Y. Then

$$\begin{aligned} I(X;Y) &\geq I(E,X;Y) - 1 \geq I(X;Y|E) - 1 \\ &\geq \mathbb{P}\left[E=0\right] I(X;Y|E=0) - 1 = \left(1 - e^{-\frac{k\epsilon}{2}}\right)\log\left(1+\frac{n}{k}\right) - 1, \end{aligned}$$

which can be made arbitrarily larger than δ by appropriately choosing the values of n and k. Since X is a deterministic function of S, $I(X;Y) = I(S;Y)$ and the result follows.

Part (b): We have

$$\frac{p_{S|Y}(s|y)}{p_S(s)} = \frac{p_{Y|S}(y|s)}{p_Y(y)} = \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s')p_S(s')} \leq \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s)p_S(s')e^{-\epsilon}} \leq e^{\epsilon},$$

and

$$\frac{p_{S|Y}(s|y)}{p_S(s)} = \frac{p_{Y|S}(y|s)}{p_Y(y)} = \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s')p_S(s')} \geq \frac{p_{Y|S}(y|s)}{\sum_{s'} p_{Y|S}(y|s)p_S(s')e^{\epsilon}} \geq e^{-\epsilon}.$$

On the other hand, we have

$$\frac{p(y|s)}{p(y|s')} = \frac{p(s|y)}{p(s)}\frac{p(s')}{p(s'|y)} \leq e^{2\epsilon},$$

which completes the proof.

□

The counterexample used in the proof of the previous theorem can be extended to allow the adversary to recover *exactly* the inputs generated the output Y. This can be done by assuming that the inputs are ordered and correlated in such a way that $X = x$ if and only if $S_1 = 1, \ldots, S_x = 1$. In this case, for n and k sufficiently large, the adversary can exploit the input correlation to correctly learn the values of $S_1, \ldots, S_n$ with arbitrarily high probability.

Differential privacy does not necessarily guarantee low leakage of information – in fact, an arbitrarily large amount of information can be leaking from a differentially private system, as shown in Theorem 2. In addition, it follows as a simple extension of (McGregor et al., 2011, Prop. 4.3) that $I(S;Y) \leq O(\epsilon n)$, corroborating that differential privacy does not bound above the average information leakage when n is sufficiently large.

4 Log-loss Distortion and Privacy Funnel

The log-loss distortion is defined as $d(x,y) = -\log p_{X|Y}(x|y)$. Note that this distortion (unlike the one in Definition 1) is a function of x and y as well as $p_{Y|X}$. Using log-loss, the average distortion becomes $\mathbb{E}[d(X,Y)] = \mathbb{E}_{p_{X,Y}}[-\log p_{X|Y}] = H(X|Y)$. Therefore, for a given $D \geq 0$, the distortion bound $H(X|Y) \leq D$ is equivalent to $I(X;Y) \geq t$, where $t = H(X) - D$. It should be noted that the average distortion under the log-loss is not linear in $p_{Y|X}$ (unlike the one in Definition 1).

4.1 Privacy-Utility Trade-off under Log-loss

Using log-loss distortion the tradeoff between between utility and privacy becomes minimizing $I(S;Y)$ while $I(X;Y) \geq t$ for some $t \geq 0$. Therefore, the trade-off between utility and privacy in the design of the privacy-preserving mapping is represented by the following optimization, that we refer to as the *Privacy Funnel*:

$$\begin{aligned} &\min\ I(S;Y) \\ &p_{Y|X}:\ I(X;Y) \geq t. \end{aligned} \tag{4.1}$$

For a given utility level t, among all feasible privacy mappings $p_{Y|X}$ satisfying $I(X;Y) \geq t$, the privacy funnel selects the one that minimizes $I(S;Y)$.

Similar to Definition 2 We define next the privacy funnel function, which captures the smallest amount of disclosed private information for a given threshold on the amount of disclosed useful information. We then characterize properties of the privacy funnel function in the rest of this section.

Definition 4. For $0 \leq t \leq H(X)$ and a joint distribution $p_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the *privacy funnel function* $G_I(t, p_{S,X})$ as

$$G_I(t, p_{S,X}) \triangleq \inf \{I(S;Y):\ I(X;Y) \geq t, S \to X \to Y\}, \tag{4.2}$$

where the infimum is over all mappings $p_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $p_{S,X}$ and $t \geq 0$, the set of pairs $\{(t, G_I(t, p_{S,X}))\}$ is called the *privacy funnel region* of $p_{S,X}$.

Before we proceed to the rest of the discussion, we can prove the counterpart of Lemma 1 for this setting, which gives a bound on the size of the alphabet $\mathcal{Y}$ one needs to consider.

Lemma 3. We have

$$G_I(t, p_{S,X}) = \min_{p_{Y|X}} \{I(S;Y)\ : I(X;Y) \leq t, S \to X \to Y,\ |\mathcal{Y}| \leq |\mathcal{X}| + 1\}. \tag{4.3}$$

Proof. In the Proof of Lemma 1, we let $a_i \triangleq H(X) - H(X|Y=i)$ and the rest of the proof is identical to that of Lemma 1. □

We now prove a few useful properties of $G_I(t, p_{S,X})$ and the privacy region.

Lemma 4. For $0 \leq t \leq H(X)$, we have

$$\max\{t - H(X|S), 0\} \leq G_I(t, p_{S,X}) \leq \frac{tI(X;S)}{H(X)}. \tag{4.4}$$

Proof. Observe that $G_I(H(X), p_{S,X}) = I(X;S)$, since $I(X;Y) = H(X)$ implies that $p_{Y|X}$ is a one-to-one mapping of X. The upper bound then follows from Lemma 3 as follows. For $0 < t \leq H(X)$ and $p_{S,X}$ fixed, let $G_I(t, p_{S,X}) = \alpha$. From the discussion above, there exists $p_{Y|X}$ that achieves $I(S;Y) = \alpha$ for $I(X;Y) \geq t$. Now consider $p_{\widetilde{Y}|X}$ where $\widetilde{\mathcal{Y}} = [|\mathcal{Y}| + 1]$ and, for $0 < \lambda \leq 1$,

$$p_{\widetilde{Y}|X}(y|x) = (1-\lambda)\mathbf{1}_{\{y=|\mathcal{Y}|+1\}} + \lambda \mathbf{1}_{\{y \neq |\mathcal{Y}|+1\}} p_{Y|X}(y|x).$$

Intuitively, $\widetilde{Y}$ is an "erased" version of Y, with the erasure symbol being $|\mathcal{Y}| + 1$. It follows directly that $I(S;\widetilde{Y}) = \lambda I(S;Y) = \lambda\alpha$, $I(X;\widetilde{Y}) = \lambda I(X;Y) \geq \lambda t$, and

$$\frac{G_I(\lambda t, p_{S,X})}{\lambda t} \leq \frac{\lambda I(S;Y)}{\lambda t} = \frac{G_I(t, p_{S,X})}{t}.$$

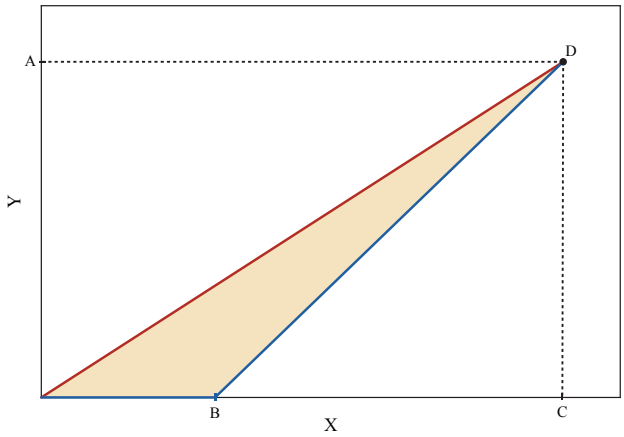

Figure 1: For a fixed $p_{S,X}$, the privacy region is contained within the shaded area. The red and the blue lines correspond, respectively, to the upper and lower bounds presented in Lemma 4.

Since this holds for any $0 < \lambda \leq 1$, then $\frac{G_I(t,p_{S,X})}{t}$ is non-decreasing in t. Finally, for a fixed $p_{S,X}$, the set of points $(\mathbf{w}_i, a_i, b_i) \in \mathbb{R}^{|\mathcal{X}|+2}$ that satisfies (2.9) is convex, and thus, for a fixed p_X, it's lower-boundary, which corresponds to the graph of $(t, G_I(t, p_{S,X}))$, is convex. Clearly $G_I(t, p_{S,X}) \geq 0$. In addition, for any $p_{Y|X}$,

$$I(S;Y) = I(X;Y) - I(X;Y|S) \geq I(X;Y) - H(X|S) \geq t - H(X|S),$$

proving the lower bound. □

Figure 1 illustrates the bounds from Lemma 4. The privacy region is contained withing the shaded area. The next two examples illustrate that both the upper bound (red line) and the lower bound (blue line) of the privacy region can be achieved for particular instances of $p_{S,X}$.

Example 3.

- Let $X = (S, W)$, where $W \perp\!\!\!\perp S$. Then by setting $Y = W$, we have $I(S;Y) = 0$ and $I(X;Y) = H(W) = H(X|S)$. Consequently, from Lemmas 3 and 4, $G_I(t, p_{S,X}) = 0$ for $t \in [0, H(X|S)]$. By letting $Y = W$ with probability λ and $Y = (S, W)$ with probability $1 - \lambda$ for $\lambda \in [0, 1]$, the lower-bound $G_I(t, p_{S,X}) = t - H(X|S)$ can be achieved for $H(X|S) = H(W) \leq t \leq H(X)$. Consequently, the lower bound in (4.4) is sharp.
- Now let $X = f(S)$. Then $I(X;S) = H(X)$ and

  $$I(S;Y) = I(X;Y) - I(X;Y|S) = I(X;Y).$$

  Consequently, $G_I(t, p_{S,X}) = t$, and the upper bound in (4.4) is sharp.

4.2 Connections to the Information Bottleneck Method

The information bottleneck method, introduced in Tishby et al. (1999), considers the setting where a variable X is to be compressed, while maintaining the information it bears about another correlated variable S. The information bottleneck method is a technique generalizing rate-distortion, as it seeks to optimize the tradeoff between the compression length of X and the accuracy of the information preserved about S in the compressed output Y. The information bottleneck optimization Tishby et al. (1999) is

$$\begin{aligned} &\min\ I(X;Y) \\ &\quad p_{Y|X}:\ I(S;Y) \geq C \end{aligned} \tag{4.5}$$

for some constant C. In the information bottleneck, the compression mapping $p_{Y|X}$ is designed to make X and Y as far as possible from each other (minimizes $I(X;Y)$) while guaranteeing that S and Y are close to

each other. In other words, in the information botteleneck the mapping $p_{Y|S}$ is designed to make $I(S;Y)$ large and $I(X;Y)$ small. The information bottleneck optimization (4.5) bears some resemblance to the privacy funnel (4.1), but is actually the opposite optimization. Indeed, in the privacy funnel, the privacy mapping is designed to make $I(S;Y)$ small and $I(X;Y)$ large.

Several techniques were developed to solve the information bottleneck problem such as alternating iteration Tishby et al. (1999) and agglomerative information bottleneck Slonim and Tishby (1999). This connection is harvested in Section 5.3 to design algorithms for the privacy funnel optimization inspired by the existing litterature on the information bottleneck optimization.

5 Privacy-Preserving Mappings Design

In this section, we consider the problem of finding optimal privacy mapping by solving (2.6). We first consider the case with the knowledge of p_{SX} and then consider the case without the knowledge on p_{SX} by only knowing p_X.

5.1 Known p_{SX}

Consider the optimization given in (2.6). The following theorem shows that the problem can be expressed as a convex optimization problem. We note that this optimization is solved in terms of the unknowns $p_{Y|X}(\cdot|\cdot)$ and $p_{Y|S}(\cdot|\cdot)$, which are coupled together through a linear equality constraint.

Proposition 2. Given $p_{S,X}(\cdot,\cdot)$, a distortion function $d(\cdot,\cdot)$ and a distortion constraint D, the mapping $p_{Y|X}(\cdot|\cdot)$ that minimizes the average information leakage can be found by solving the following convex optimization (assuming the usual simplex constraints on the probability distributions):

$$\min_{p_{Y|X},\ p_{Y|S},\ |\mathcal{Y}|\leq|\mathcal{X}|+1} \sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} p_{Y|S}(y|s)p_S(s)\log\left(\frac{p_{Y|S}(y|s)}{p_Y(y)}\right) \tag{5.1}$$

$$\text{s.t.} \quad \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{Y|X}(y|x)p_X(x)d(y,x) \leq D, \tag{5.2}$$

$$\sum_{x\in\mathcal{X}} p_{X|S}(x|s)p_{Y|X}(y|x) = p_{Y|S}(y|s)\ \forall y,s, \tag{5.3}$$

$$\sum_{s\in\mathcal{S}} p_{Y|S}(y|s)p_S(s) = p_Y(y)\ \forall y. \tag{5.4}$$

Proof. Clearly the previous optimization is the same as (2.6). To prove the convexity of the objective function, note that $h(x,a) = ax\log x$ is convex for a fixed $a\geq 0$ and $x\geq 0$, and, therefore, the perspective of $g_1(x,z,a) = ax\log(x/z)$ is also convex in x and z for $z>0, a\geq 0$ (Boyd and Vandenberghe (2004)). Since the objective function (5.1) can be written as

$$\sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} g(p_{Y|S}(y|s), p_Y(y), p_S(s)),$$

it follows the optimization is convex. In addition, since $p(y)\to 0 \Leftrightarrow p(y|s)\to 0\ \forall y$, the minimization is well defined over the probability simplex. Finally, the constraint $|\mathcal{Y}|\leq|\mathcal{X}|+1$ follows from Lemma 1. □

Corollary 1. If X is a deterministic function of S and $S\to X\to Y$ then the minimization in (2.6) can be simplified to a rate-distortion problem:

$$\min_{p_{Y|X}} \ I(X;Y)$$
$$\text{s.t.} \ \ \mathbb{E}_{X,U}[d(X,Y)] \leq D\ .$$

Furthermore, by restricting $Y = X+Z$ and $d(X,Y) = d(X-Y)$, the optimization reduces to

$$\max_{p_Z} \ H(Z)$$
$$\text{s.t.} \ \ \mathbb{E}_Z[d(Z)] \leq D\ .$$

Proof. Since X s a deterministic function of S and $S \to X \to Y$, then

$$\begin{aligned} I(S;Y) &= I(S,X;Y) - I(X;Y|S) \\ &= I(X;Y) + I(S;Y|X) - I(X;Y|S) \\ &= I(X;Y), \end{aligned} \tag{5.5}$$

where (5.5) follows from the fact that X is a deterministic function of S ($I(X;Y|S) = 0$) and $S \to X \to Y$ ($I(S;Y|X) = 0$). For the additive noise case, the result follows by observing that $H(X|Y) = H(Z)$. □

5.1.1 Maximum information leakage

The minimum over all possible maximum cost gains of an adversary that uses a log-loss function in (2.2) is given by

$$C^* = \max_{y \in \mathcal{Y}} H(S) - H(S|Y = y).$$

The previous expression motivates the definition of *maximum information leakage*, presented below.

Definition 5. The *maximum information leakage* of a set of features S is defined as the maximum cost gain, given in terms of the log-loss function, that an adversary obtains by observing a single output, and is given by $\max_{y \in \mathcal{Y}} H(S) - H(S|Y = y)$. A privacy-preserving mapping $p_{Y|X}(\cdot)$ is said to achieve the *minmax information leakage* for a distortion constraint D if it is a solution of the minimization

$$\min_{p_{Y|X}} \max_{y \in \mathcal{Y}} \; H(S) - H(S|Y = y) \tag{5.6}$$

$$\text{s. t. } \mathbb{E}[d(Y,X)] \leq D \tag{5.7}$$

The following theorem demonstrates how the mapping that achieves the minmax information leakage can be determined as the solution of a related convex program that finds the minimum distortion given a constraint on the maximum information leakage.

Proposition 3. Given $p_{S,X}(\cdot,\cdot)$, a distortion function $d(\cdot,\cdot)$ and a constraint ϵ on the maximum information leakage, the minimum achievable distortion and the mapping that achieves the minmax information leakage can be found by solving the following convex optimization (assuming the implicit simplex constraints on the probability distributions):

$$\begin{aligned} \min_{p_{Y|X}, p_{Y|S}} \quad & \sum_{y \in \mathcal{Y}} \sum_{s \in \mathcal{S}} p_{Y|x}(y|x) p_X(x) d(y,x) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} p_{X|S}(x|s) p_{Y|X}(y|x) = p_{Y|S}(y|s) \;\; \forall y, s, \\ & \sum_{s \in \mathcal{S}} p_{Y|S}(y|s) p_S(s) = p_Y(y) \;\; \forall y, \end{aligned} \tag{5.8}$$

$$D p_Y(y) + \sum_{s \in \mathcal{S}} p_{Y,S}(y,s) \log \frac{p_{Y,S}(y,s)}{p_Y(y)} \leq 0 \;\; \forall y, \tag{5.9}$$

where $D = H(S) - \epsilon$. Therefore, for a given value of D, the optimization problem in (5.6) can be efficiently solved with arbitrarily large precision by performing a line-search over $\epsilon \in [0, H(S)]$ and solving the previous convex program at each step of the search.

Proof. The convex program in (5.6) can be reformulated to return the minimum distortion for a given constraint ϵ on the minmax information leakage as

$$\begin{aligned} \min_{p_{Y|X}} \quad & \mathbb{E}[d(Y,X)] \\ \text{s.t. } & H(S|Y = y) \geq D \; . \end{aligned} \tag{5.10}$$

It is straightforward to verify that constraint (5.9) can be written as (5.10). Following the same steps as the proof of Theorem 2 and noting that the function $g_2(x,z,a) = ax\log(ax/z)$ is convex for $a, x \geq 0$, $z > 0$, it follows that (5.10) and, consequently, (5.9), is a convex constraint. Finally, since the optimal distortion value in the previous program is a decreasing function of ϵ, it follows that the solution of (5.6) can be found through a line-search in ϵ. □

Corollary 2. For $X = f(S)$, where $f : \mathcal{S} \to \mathcal{Y}$ is a deterministic function, $S \to X \to Y$ and a fixed prior $p_{X,S}(\cdot,\cdot)$, the privacy preserving mapping that minimizes the maximum information leakage is given by

$$p^*_{Y|X} = \arg\min_{p_{Y|X}} \max_{y\in\mathcal{Y}} D(p_{X|Y}||\zeta)$$
$$\text{s.t. } \mathbb{E}[d(Y,X)] \leq D,$$

where $\zeta(x) = \frac{2^{H(S|X=x)}}{\sum_{x'\in\mathcal{X}} 2^{H(S|X=x')}}$.

Proof. Under the assumptions of the corollary, note that for a given $y \in \mathcal{Y}$ (and assuming that the logarithms are in base 2)

$$\begin{aligned}
H(S|Y=y) &= -\sum_{s\in\mathcal{S}} p_{S|Y}(s|y)\log p_{S|Y}(s|y)\\
&= -\sum_{s\in\mathcal{S}}\left(\sum_{x\in\mathcal{X}} p_{S|X}(s|x)p_{X|Y}(x|y)\right)\left(\log\sum_{x'\in\mathcal{X}} p_{S|X}(s|x')p_{X|Y}(x'|y)\right)\\
&= -\sum_{s\in\mathcal{S}} p_{S|X}(s|f(s))p_{X|Y}(f(s)|y)\log p_{S|X}(s|f(s))p_{X|Y}(f(s)|y) \qquad (5.11)\\
&= -\sum_{s\in\mathcal{S},x\in\mathcal{X}} p_{S|X}(s|x)p_{X|Y}(x|y)\log p_{S|X}(s|x)p_{X|Y}(x|y) \qquad (5.12)\\
&= H(X|Y=y) + \sum_{x\in\mathcal{X}} p_{X|Y}(x|y)H(S|X=x)\\
&= \sum_{x\in\mathcal{X}} p_{X|Y}(x|y)\log\frac{2^{H(S|X=x)}}{p_{X|Y}(x|y)} = -D(p_{X|Y}||\zeta) + \log\left(\sum_{x\in\mathcal{X}} 2^{H(S|X=x)}\right), \qquad (5.13)
\end{aligned}$$

where (5.11) and (5.12) follows by noting that $p_{S|X}(s|x) = 0$ if $x \neq f(s)$. The result follows directly by substituting (5.13) in (5.6). □

For X a deterministic function of S, the optimal privacy preserving mechanism is the one that approximates (in terms of KL-divergence) the posterior distribution of X given Y to $\zeta(\cdot)$. Note that the distribution $\zeta(\cdot)$ captures the inherent uncertainty that exists in the function f for different outputs $x \in \mathcal{X}$. The purpose of the privacy preserving mapping is then to augment this uncertainty, while still satisfying the distortion constraint. In particular, the larger the uncertainty $H(S|X=x)$, the larger the probability of $p_{X|Y}(x|y)$ for all y. Consequently, the optimal privacy mapping (exponentially) reinforces the posterior probability of the values of x for which there is a large uncertainty regarding the features S. This fact is illustrated in the next example, where we revisit the counting query presented in Example 2.

Example 4 (Counting query continued)**.** Assume that each database input S_i, $1 \leq i \leq n$ satisfies $P(\mathbf{1}_A(S_i) = 1) = p$ and are independent and identically distributed. Then Y is a binomial random variable with parameter (n,p). It follows that $H(\mathbf{S}|X=x) = \log\binom{n}{x}$. Consequently, the optimal privacy preserving mapping will be the one that results in a posterior probability $p_{X|Y}(x|y)$ that is proportional to the size of the pre-image of y, i.e. $p_{X|Y}(x|y) \propto |f^{-1}(x)| = \binom{n}{x}$.

In all the problems we have considered so far, the complete knowledge of the prior distribution p_{SX} is assumed. Next, we relax this assumption, and consider a worst-case optimization with partial knowledge of the distribution, namely when p_X is known, but $p_{S|X}$ is unknown.

5.2 Known p_X, unknown $p_{S|X}$

In practice, we may not have access to the probability of the underlying variable S and only know the probability of random variable X. Consequently, finding the exact solution of problem (2.6) is not possible. This raises the question of the design of privacy-preserving mappings under this partial knowledge on the priors. In particular, suppose p_X is known and $p_{S|X}$ is unknown. We consider the privacy-preserving mapping which minimizes the worst-case privacy over all possible $p_{S|X}$ while satisfying the utility constraint. Therefore, the optimal privacy-preserving mapping under this partial knowledge is solution of

$$\begin{aligned} &\min_{p_{Y|X}} \ \max_{p_{S|X}} I(S;Y), \\ &\text{s.t.} \ \ \mathbb{E}_{X,Y}[d(X,Y)] \leq D \ . \end{aligned} \tag{5.14}$$

Proposition 4. The problem in (5.14) is equivalent to the following rate distortion problem.

$$\begin{aligned} \min_{p_{Y|X}:\ |\mathcal{Y}|\leq|\mathcal{X}|+1} \quad & I(X;Y), \\ \text{s.t.} \quad & \mathbb{E}_{X,Y}[d(X,Y)] \leq D \end{aligned} \tag{5.15}$$

Proof. First note that by letting $S = X$, we obtain $I(X;Y) \leq \max_{p_{S|X}} I(S;Y)$ which then result in

$$\min_{p_{Y|X}} \ I(X;Y) \leq \min_{p_{Y|X}} \ \max_{p_{S|X}} I(S;Y).$$

The other direction follows from the Markov chain property, i.e. $S \to X \to Y$. In particular, for any $p_{S|X}$ we have $I(X;Y) \geq I(S;Y)$ which results in $I(X;Y) \geq \max_{p_{S|X}} I(S;Y)$ for any $p_{Y|X}$. Therefore, we have

$$\min_{p_{Y|X}} \ I(X;Y) \geq \min_{p_{Y|X}} \ \max_{p_{S|X}} I(S;Y).$$

Also, note that the constraint $|\mathcal{Y}| \leq |\mathcal{X}| + 1$ follows from the same argument as in Lemma 1, which completes the proof. □

Proposition 4 shows that optimization (5.15) can be solved by using any convex solver. Also, note that the optimization (5.15) can be solved using an Expectation-Minimization (EM) algorithm such as Arimoto-Blahut algorithm Cover and Thomas (2012). The problem (5.15) becomes intractable for large alphabet sizes $|\mathcal{X}|$ since the objective function is non-linear. In particular, in our experiments, classical interior-point methods could hardly overcome alphabet sizes of a few hundred. One line of work consists in finding good algorithms tailored to this optimization problem, as in Salamatian et al. (2014), where it is noticed that the solution of the optimization is often sparse. This observation is then used to construct an algorithm similar to Dantzig-Wolf methods in Linear-Programming where only the relevant optimization variables are generated in a greedy way. This allows to efficiently solve the optimization over much larger alphabets.

5.3 Algorithm for Privacy Funnel

The alternating iteration algorithm Tishby et al. (1999) finds a stationary point of the Lagrangian of information bottleneck optimization (4.5) defined as $\mathcal{L} = I(X;Y) - \beta I(S;Y)$ for some β. The stationary point can be a local minimum, which addresses the information bottleneck, or a local maximum in which case it addresses the privacy funnel. However, there is no guarantee on the convergence of this alternating algorithm to either a local minimum or a local maximum.

$$\begin{aligned} &\min \ I(S;Y) \\ &\quad p_{Y|X} : \ I(X;Y) \geq t. \end{aligned}$$

For a given utility level t, among all feasible privacy mappings $p_{Y|X}$ satisfying $I(X;Y) \geq t$, the privacy funnel selects the one that minimizes $I(S;Y)$.

Note that $I(X;Y)$ is convex in $p_{Y|X}$ and since $p_{Y|S}$ is linear in $p_{Y|X}$ and $I(S;Y)$ is convex in $p_{Y|S}$, the objective function $I(S;Y)$ is convex in $p_{Y|X}$. However, because of the constraint $I(X;Y) \geq t$, the Privacy

Algorithm 1 Greedy algorithm-privacy funnel

Input: t, $p_{S,X}$
Initialization: $\mathcal{Y}=\mathcal{X}$, $p_{Y|X}(y|x)=\mathbf{1}\{y=x\}$.
while there exists $i',j'\in\mathcal{Y}$ such that $I(X;Y^{(i',j')})\geq t$ **do**
among those i',j', let
$\{y_i,y_j\}=\operatorname{argmax}_{y_{i'},y_{j'}\in\mathcal{Y}} I(S;Y)-I(S;Y^{(i',j')})$
merge: $\{y_i,y_j\}\to y_{ij}$
update: $\mathcal{Y}=\{\mathcal{Y}\setminus\{y_i,y_j\}\}\cup\{y_{ij}\}$ and $p_{Y|X}$
end while
Output: $p_{Y|X}$

Funnel (4.1) is not a convex optimization (Boyd and Vandenberghe, 2004, Chap. 4). As mentioned previously, the Privacy Funnel (4.1) is not a convex optimization. In this section, we provide a greedy algorithm and an alternating iteration algorithm to solve optimization (4.1), and we evaluate them on simulated data.

We use a greedy algorithm to find a privacy mapping as described next. Assume $I(X;Y)\geq t$ is given and we are looking for $p_{Y|X}$ that minimizes $I(S;Y)$. Note that for $\mathcal{Y}=\mathcal{X}$ and $p_{Y|X}(y|x)=\mathbf{1}\{x=y\}$ (where $\mathbf{1}\{x=y\}=1$ if and only if $x=y$), the condition $I(X;Y)\geq t$ is satisfied, but, $I(S;Y)$ might be too large. The idea is to merge two elements of $\mathcal{Y}$ to make $I(S;Y)$ smaller, while satisfying $I(X;Y)\geq t$. This method is motivated by agglomerative information method introduced in Slonim and Tishby (1999). We merge y_i and y_j and denote the merged element by y_{ij}. We then update $p_{Y|X}$ as $p(y_{ij}|x)=p(y_i|x)+p(y_j|x)$, for all $x\in\mathcal{X}$. After merging, we also have $p(y_{ij})=p(y_i)+p(y_j)$. Let $Y^{(i,j)}$ be the resulting Y from merging i and j. Algorithm 1 is a greedy algorithm to solve optimization (4.1). Proposition 5 shows that, there is an efficient way to calculate $I(S;Y)-I(S;Y^{(i,j)})$ and $I(X;Y)-I(X;Y^{(i,j)})$ at each iteration of algorithm 1.

Proposition 5. For a given joint distribution $p_{S,X,Y}=p_{S,X}p_{Y|X}$, we have $I(S;Y)-I(S;Y^{(i,j)})=$

$$p(y_{ij})H\left(\frac{p(y_i)p_{S|Y=y_j}+p(y_j)p_{S|Y=y_j}}{p(y_{ij})}\right)-\left(p(y_i)H(p_{S|Y=y_i})+p(y_j)H(p_{S|Y=y_j})\right).$$

We also have $I(X;Y)-I(X;Y^{(i,j)})=$

$$p(y_{ij})H\left(\frac{p(y_i)p_{X|Y=y_j}+p(y_j)p_{X|Y=y_j}}{p(y_{ij})}\right)-\left(p(y_i)H(p_{X|Y=y_i})+p(y_j)H(p_{X|Y=y_j})\right).$$

Proof. After merging y_i and y_j, we have

$$p(s|y_{ij})=\frac{p(y_i)}{p(y_{ij})}p(s|y_i)+\frac{p(y_j)}{p(y_{ij})}p(s|y_j),\ \text{for all } s\in\mathcal{S},$$
$$p(x|y_{ij})=\frac{p(y_i)}{p(y_{ij})}p(x|y_i)+\frac{p(y_j)}{p(y_{ij})}p(x|y_j),\ \text{for all } x\in\mathcal{X}.$$

The proof follows from writing $I(S;Y)-I(S;Y^{(i,j)})=H(S|Y^{(i,j)})-H(S|Y)$ and $I(X;Y)-I(X;Y^{(i,j)})=H(X|Y^{(i,j)})-H(X|Y)$. □

Note that the greedy algorithm is locally optimal at every step since we minimize $I(S;Y)$. However, there is no guarantee that such a greedy algorithm induces a global optimal privacy mapping.

Note 1. The minimum of $I(S;Y)$ in (4.1) is a decreasing function of $I(X;Y)$ and is achieved for a mapping $P_{Y|X}$ that satisfies $I(X;Y)=t$ (if possible due to discrete alphabets). For a given mutual information, t, there are many conditional probability distributions, $P_{Y|X}$, achieving $I(X;Y)=t$. Among which there is one that gives the minimum $I(S;Y)$ and one that gives the maximum $I(S;Y)$. We can modify the greedy algorithm so that it converges to a local maximum of $I(S;Y)$ for a given $I(X;Y)=t$. The algorithm which we call *greedy algorithm-information bottleneck* is given in Algorithm (2). Algorithm (1) and Algorithm (2)

Algorithm 2 Greedy algorithm-information bottleneck

Input: Δ, $p_{S,X}$
Initialization: $\mathcal{Y} = \mathcal{X}$, $p_{Y|X}(y|x) = \mathbf{1}\{y = x\}$
while there exists $i', j' \in \mathcal{Y}$ such that $I(S;Y^{(i',j')}) \geq \Delta$ **do**
  among those i', j', let
  $\{y_i, y_j\} = \text{argmax}_{y_{i'}, y_{j'} \in \mathcal{Y}} I(X;Y) - I(X;Y^{(i',j')})$
  merge: $\{y_i, y_j\} \to y_{ij}$
  update: $\mathcal{Y} = \{\mathcal{Y} \setminus \{y_i, y_j\}\} \cup \{y_{ij}\}$ and $p_{Y|X}$
end while
Output: $p_{Y|X}$

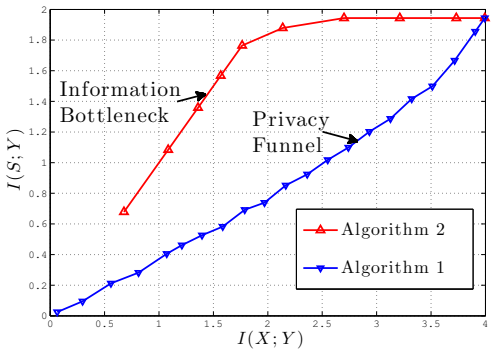

Figure 2: Maximum and minimum of $I(S;Y)$ for a given $I(X;Y)$: using greedy algorithms.

allow us to approximately characterize the range of values $I(S;Y)$ can take for a given value of $I(X;Y)$ as being those between the local minimum and the local maximum. Interestingly, by observing the gap between the local maximum and the local minimum, we have a relative idea on the effectiveness of the Greedy algorithm, i.e., if the difference is significant it means a negligent mapping may lie anywhere between those values, possibly leading to a much higher privacy threat.

Example 5 (Numerical Example)**.**
Data Set: The US 1994 Census dataset Asuncion and Newman (2007) is a well-known dataset in the machine learning community, which is a sample of the US population from 1994. For each of the entries, it contains features such age, work-class, education, gender, and native country, as well as an income category. The income level is a binary variable which determines whether the income is above or below USD 50000, gender is a binary variable, education level is a variable with four categories, age is a variable divided into seven categories. For our purposes, we consider the private attributes $S =$ (age, income level) and the attributes to be released as $X =$ (age, gender, education level). The goal of the privacy mapping is to release a modified version of attributes Y which is informative about X but that renders the inference of S based on Y hard.
Numerical illustration In Fig. 2, we plot the minimum and maximum of $I(S;Y)$ for a given $I(X;Y)$. This figure is based on US 1994 census data set described before. The top curve shows the maximum of $I(S;Y)$ versus $I(X;Y)$, using Algorithm (2). The bottom curve shows the minimum of $I(S;Y)$ versus $I(X;Y)$, using Algorithm (1). The area between the two curves shows the possible pairs of $(I(X;Y), I(S;Y))$ as $P_{Y|X}$ varies (a subset of possible pairs, since the algorithms are sub-optimal). Indeed, we will design the mapping to lie on the bottom curve. For a given t, if we design the mapping negligently, we may have $I(S;Y)$ on the top curve instead of the bottom curve.

6 Conclusions

We consider a privacy-utility trade-off encountered by users who wish to disclose some information to an analyst, that is correlated with their private data, in the hope of receiving some utility. We propose a general framework under which data is transformed according to a probabilistic privacy-preserving mapping before it is disclosed. We show that applying this general framework to the setting where the adversary uses the log-loss cost function naturally leads to a non-asymptotic information-theoretic formulation for characterizing the best achievable privacy subject to utility constraints. This formulation can be cast as a modified rate-distortion problem which, in turn, can be formulated as a convex program. We justify the relevance and generality of the privacy metric under the log-loss by proving that the inference threat under any bounded cost function can be upper bounded by an explicit function of the mutual information between private data and disclosed data. We compare our framework with differential privacy. In addition, we show that when the log-loss is used in this framework in both the privacy metric and the distortion metric, the average information leakage and the utility constraint can be reduced to the mutual information between private data and disclosed data, and between non-private data and disclosed data, respectively. We then show that the privacy-utility tradeoff under the log-loss can be cast as the non-convex Privacy Funnel optimization, and we leverage its connection to the Information Bottleneck, to provide a greedy algorithm for solving it.

Acknowledgement

The authors are grateful to Prof. Thomas Courtade, Prof. Kave Salamatian, and Prof. Tsachy Weissman for encouraging us to study the connections between privacy and the information bottleneck.

References

Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *ACM Sigmod Record*, 29(2): 439–450, 2000.

Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer, 2011.

Shahab Asoodeh, Fady Alajaji, and Tamás Linder. Notes on information-theoretic privacy. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 1272–1278. IEEE, 2014.

Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Information extraction under privacy constraints. *Information*, 7(1):15, 2016.

Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Privacy-aware guessing efficiency. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 754–758. IEEE, 2017.

Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamás Linder. Estimation efficiency under privacy constraints. *IEEE Transactions on Information Theory*, 65(3):1512–1534, 2018.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html.

Siddhartha Banerjee, Nidhi Hegde, and Laurent Massoulié. The price of privacy in untrusted recommendation engines. *arXiv preprint arXiv:1207.3269*, 2012.

Yuksel Ozan Basciftci, Ye Wang, and Prakash Ishwar. On privacy-utility tradeoffs for constrained data release mechanisms. 2016.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

Luca Bonomi, Liyue Fan, and Hongxia Jin. An information-theoretic approach to individual sequential data sanitization. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 337–346. ACM, 2016.

Stephen Poythress Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

F. P. Calmon and Nadia Fawaz. Privacy against statistical inference. In *Proc. 50th Ann. Allerton Conf. Commun., Contr., and Comput.*, pages 1401–1408, 2012.

Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Anonymity protocols as noisy channels. *Information and Computation*, 206(2-4):378–401, 2008.

Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical measurement of information leakage. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 390–404. Springer, 2010.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, 2011.

Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *NIPS*, pages 998–1006, 2012.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-interscience, 2012.

Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition, August 2011. ISBN 0521196817.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006a. URL http://www.cs.bgu.ac.il/~{}kobbi/papers/sensitivity-tcc-final.pdf.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*. Springer, 2006a.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, volume 4052, pages 1–12. Springer, 2006b. ISBN 978-3-540-35907-4, 978-3-540-35908-1.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. Springer, 2006b.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 117–126. ACM, 2015b.

H. G. Eggleston. *Convexity*. Cambridge University Press, Cambridge England, 1 edition edition, January 2009.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM Symposium on Principles of Database Systems*, pages 211–222, New York, NY, USA, 2003.

Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502. ACM, 2010.

Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior*, 91:334–346, 2015.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3), 2011.

Zuxing Li, Tobias J Oechtering, and Deniz Gündüz. Privacy against a hypothesis testing adversary. *IEEE Transactions on Information Forensics and Security*, 14(6):1567–1581, 2018.

Jiachun Liao, Lalitha Sankar, Vincent YF Tan, and Flavio du Pin Calmon. Hypothesis testing under mutual information privacy constraints in the high privacy regime. *IEEE Transactions on Information Forensics and Security*, 13(4):1058–1071, 2017.

Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. A tunable measure for information leakage. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 701–705. IEEE, 2018.

Jiachun Liao, Lalitha Sankar, Oliver Kosut, and Flavio P Calmon. Robustness of Maximal α-Leakage to Side Information. *arXiv preprint arXiv:1901.07105*, 2019.

Ashwin Machanavajjhala, Aleksandra Korolova, and Atish Das Sarma. Personalized social recommendations: accurate or private. *Proceedings of the VLDB Endowment*, 4(7):440–450, 2011.

A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Medard. From the information bottleneck to the privacy funnel. In *IEEE Inf. Theory Workshop (ITW)*, pages 501–505, 2014.

Ali Makhdoumi and Nadia Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *Proc. 51th Annual Allerton Conference on Communication, Controrl, and Computation*, pages 1627–1634, 2013.

A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan. The limits of two-party differential privacy. *Electronic Colloquium on Computational Complexity (ECCC)*, 18(106), 2011.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 81–90. IEEE, 2010.

Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636. ACM, 2009.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6): 2124–2147, 1998.

Darakhshan J Mir. Information-theoretic foundations of differential privacy. In *International Symposium on Foundations and Practice of Security*, pages 374–381. Springer, 2012.

Nina Mishra and Mark Sandler. Privacy via pseudorandom sketches. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 143–152. ACM, 2006.

Bahman Moraffah and Lalitha Sankar. Information-theoretic private interactive mechanism. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 911–918. IEEE, 2015.

Seyed Ali Osia, Borzoo Rassouli, Hamed Haddadi, Hamid R Rabiee, and Deniz Gündüz. Privacy against brute-force inference attacks. *arXiv preprint arXiv:1902.00329*, 2019.

Borzoo Rassouli and Deniz Gunduz. Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 2019.

David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11): 1623–1636, 2010.

I. S. Reed. Information Theory and Privacy in Data Banks. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, AFIPS '73, pages 581–587. ACM, 1973.

Irving S Reed. Information theory and privacy in data banks. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*. ACM, 1973.

Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

Salman Salamatian, Nadia Fawaz, Branislav Kveton, and Nina Taft. Sspm : Sparse privacy preserving mappings. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2014.

L. Sankar, S. R. Rajagopalan, and H. V. Poor. Utility-privacy tradeoff in databases: An information-theoretic approach. *IEEE Trans. Inf. Forensics Security*, 2013.

Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility and privacy of data sources: Can shannon help conceal and reveal information? In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–7. IEEE, 2010.

Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

Anand D Sarwate and Lalitha Sankar. A rate-disortion perspective on local differential privacy. In *Allerton*, pages 903–908, 2014.

I. Sason and S. Verdú. Arimoto-rényi conditional entropy and bayesian m-ary hypothesis testing. *IEEE Transactions on Information Theory*, PP(99):1–1, 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2757496.

N. Slonim and N. Tishby. Agglomerative information bottleneck. *Proc. of Neural Information Processing Systems (NIPS-99)*, 1999.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

Sreejith Sreekumar, Asaf Cohen, and Deniz Gündüz. Distributed hypothesis testing with a privacy constraint. *arXiv preprint arXiv:1807.02764*, 2018.

L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proc. 37th Ann. Allerton Conf. Commun., Contr., and Comput.*, pages 368–377, 1999.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv:physics/0004057 [physics.data-an]*, April 2000.

Hao Wang and Flavio P Calmon. An estimation-theoretic view of privacy. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 886–893. IEEE, 2017.

Hao Wang, Lisa Vo, Flavio P Calmon, Muriel Médard, Ken R Duffy, and Mayank Varia. Privacy with estimation guarantees. *IEEE Transactions on Information Theory*, 65(12):8025–8042, 2019.

Ke Wang, Philip S Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004.

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

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver of wiretappers. 29(6), 1983a.

Hirosuke Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.). *Information Theory, IEEE Transactions on*, 29(6), 1983b.

Ye Zhu and Riccardo Bettati. Anonymity vs. information leakage in anonymity systems. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 514–524. IEEE, 2005.