Protecting Data Against Unwanted Inferences

Supriyo Chakraborty, Nicolas Bitouzé, Mani Srivastava, Lara Dolecek
Electrical Engineering Department
University of California, Los Angeles
Los Angeles, USA
{supriyo,bitouze,mbs}@ucla.edu, dolecek@ee.ucla.edu

***Abstract*—We study the competing goals of utility and privacy as they arise when a provider delegates the processing of its personal information to a recipient who is better able to handle this data. We formulate our goals in terms of the inferences which can be drawn using the shared data. A whitelist describes the inferences that are desirable, i.e., providing utility. A blacklist describes the unwanted inferences which the provider wants to keep private. We formally define utility and privacy parameters using elementary information-theoretic notions and derive a bound on the region spanned by these parameters. We provide constructive schemes for achieving certain boundary points of this region. Finally, we improve the region by sharing data over aggregated time slots.**

I. Introduction

In a wide variety of applications, data is often shared from one entity to another: an information provider delegates the processing of its personal information to a recipient who is better able to handle this data. While the provider generally benefits from the information processing performed by the recipient, the shared data may also be used to draw inferences that are unwanted from the stance of the provider for privacy reasons. For example, the provider can be a smartphone with access to the current activity and whereabouts of its user. The recipient can then be an untrusted app which (based on the current user context, i.e., activity and whereabouts), for any incoming call decides whether the phone should ring, vibrate or remain silent. Specifically, as shown in Fig. 1 (notation explained in detail later), for given data D we consider that the provider specifies a *whitelist* of utility-providing inferences, denoted by X, and a *blacklist* of unwanted inferences Y that she would like to keep private. In the above example, X would be the suitable ringing behavior, and Y the user's current location. We focus on strategies to derive message M under specified utility and privacy constraints (how much information M gives about X and Y).

Research on privacy has recently attracted a lot of attention. Mechanisms following differential privacy have been used to protect membership information within a database by adequately perturbing the aggregate query responses [1]. Differentially private aggregation of data from multiple distributed sources is presented in [2]. In [3], the authors introduce the notion of partial information hiding and increase the uncertainty of individual values by perturbing them. The effect of added noise can often be eliminated by averaging and other sophisticated filtering techniques [2], [4]. A prototype implementation of a privacy-aware framework on an Android smartphone is provided in [5]. Inspired by the work in [5], we consider an information-theoretic analysis of the unwanted inference problem. Elementary tools from information theory can be used to effectively capture how much the recipient can infer from the shared data. Our investigation follows the approaches developed in [6], [7] for databases and [8] for time-series data. We differ from the above works in several key components. Our scheme does not depend on the existence of a multi-user database, the identity of the provider is typically known, and what the provider wishes to keep private is a function of the data and not the raw data itself.

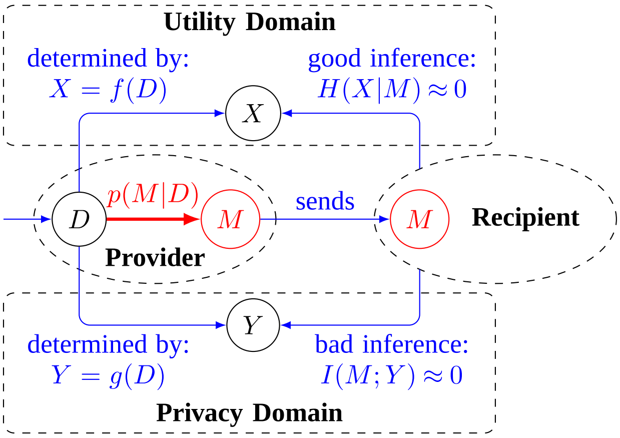


Fig. 1: The provider senses an RV D and wants to send a message M to the recipient, so that recipient can estimate $X = f(D)$ from M without being able to estimate $Y = g(D)$. If f, g and the distribution of D are fixed, the recipient can only choose the conditional distribution $p(M|D)$.

Specifically, our contributions are as follows. In Section II, we present our system model, introduce formal definitions of our utility and privacy parameters, and derive a theoretical bound on the region described by the achievable values of these parameters. In Section III, we outline our strategy for achieving maximum privacy under perfect utility, followed by Section IV in which we derive conditions for achieving maximum utility under perfect privacy. In Section V, we consider sharing data over multiple time slots to better approach the theoretical bound. We conclude in Section VI.

II. Preliminaries

In our model (see Fig.1), a data provider senses a discrete RV D. The provider cooperates with the recipient by sharing information about the whitelist, as specified by the RV $X = f(D)$ (e.g., X can be mute, ring, or vibrate). The provider also wants to keep the blacklist, specified by $Y = g(D)$ (e.g., Y can be work, home, theater, etc.) private. We consider X and Y to be deterministic functions of D. In this work, we focus on provider strategies for generating message M from D, represented by the distribution $p(M|D)$, such that the desired tradeoff between utility (M gives a lot of information about X) and privacy (M provides no information about Y) can be achieved.

Formally, for a given choice of distribution $p(M|D)$ of message M knowing data D, we define a utility parameter $\delta_U(M)$ and a privacy parameter $\delta_P(M)$ as:

$$\delta_U(M) \triangleq \frac{H(X|M)}{H(X)}, \quad \delta_P(M) \triangleq \frac{I(M;Y)}{H(Y)}. \tag{1}$$

When there is no ambiguity, we omit the argument M of δ_U and δ_P. Depending on the application, various other utility metrics could be useful, for instance Hamming or Euclidean distortions [6]. Most of our analysis directly translates under any distortion metric for utility: for reasons of simplicity we choose to present it here under the equivocation metric δ_U.

These parameters expressed in terms of elementary information-theoretic notions conveniently capture the tradeoff between utility and privacy. The smaller $\delta_U(M)$ is, the more useful M is in determining X, and the smaller $\delta_P(M)$ is, the more private M is about Y. We refer to $\delta_U = 0$ as the *perfect utility* case (in which $H(X|M) = 0$ and therefore X can be perfectly inferred from M) and to $\delta_P = 0$ as the *perfect privacy* case (in which $I(M;Y) = 0$ and therefore Y is independent from M). In general, it is not possible to achieve perfect utility and perfect privacy at the same time. In the following, we provide a lower bound on the achievable (δ_U, δ_P) pairs:

Theorem 1. *Let D, X and Y be fixed. For any choice of conditional distribution $p(M|D)$, the following lower bound holds:*

$$\delta_U(M)H(X) + \delta_P(M)H(Y) \geq I(X;Y). \tag{2}$$

For a given choice of M, equality in (2) holds if and only if $H(X|M,Y) = I(M;Y|X) = 0$.

Proof: We prove that for any three RVs M, X and Y, $H(X|M) + I(M;Y) \geq I(X;Y)$, where $\delta_U(M)H(X) = H(X|M)$ and $\delta_P(M)H(Y) = I(M;Y)$:

$$\begin{aligned}
&I(M;Y) + H(X|M) - I(X;Y)\\
&= H(M) + H(Y) - H(M,Y) + H(X|M) - I(X;Y)\\
&= H(M,X) + H(X,Y) - H(M,Y) - H(X)\\
&= H(M) + H(X|M) + H(X) + H(Y|X)\\
&\qquad\qquad - H(M) + H(Y|M) - H(X)\\
&= H(X|M) + H(Y|X) - H(Y|M)\\
&= H(X|M) + I(M;Y|X) + H(Y|M,X) - H(Y|M)\\
&= H(X|M,Y) + I(M;Y|X) \geq 0.
\end{aligned} \tag{3}$$

Because $H(X|M,Y)$ and $I(M;Y|X)$ are both non-negative, equality holds if and only if they are both zero. □

From Theorem 1, we deduce that it is possible to achieve perfect utility and perfect privacy at the same time only if $I(X;Y) = 0$ (in fact, one can observe that reciprocally, if $I(X;Y) = 0$ then we can always achieve perfect utility and perfect privacy simultaneously, e.g. by transmitting $M = X$).

A more intriguing question is, when $I(X;Y) > 0$, what are the achievable pairs (δ_U, δ_P)? In this case, we consider that D, X and Y are fixed, and we want to find RVs M that achieve good tradeoffs between δ_U and δ_P. We denote the respective alphabets of D, X, Y and M by $\mathcal{D}$, $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{M}$. While the former three alphabets are fixed by the problem setup, we have the choice of the alphabet $\mathcal{M}$ of M.

To answer the posed question, we first define the *maximum privacy under perfect utility* $\max_{\text{perfU}} \text{P}$ and *maximum utility under perfect privacy* $\max_{\text{perfP}} \text{U}$ points by

$$\begin{aligned}
&\max_{\text{perfU}} \text{P} = (0, \delta_P^*) \text{ where } \delta_P^* \triangleq \min_{\substack{M:\\ \delta_U(M)=0}} \delta_P(M),\\
&\max_{\text{perfP}} \text{U} = (\delta_U^*, 0) \text{ where } \delta_U^* \triangleq \min_{\substack{M:\\ \delta_P(M)=0}} \delta_U(M).
\end{aligned} \tag{4}$$

In the next sections, we consider strategies to achieve the points $\max_{\text{perfU}} \text{P}$ and $\max_{\text{perfP}} \text{U}$.

III. Maximizing Privacy under Perfect Utility

Using the fact that X is a deterministic function of D, we provide a simple strategy to achieve $\max_{\text{perfU}} \text{P}$, and show that the strategy reaches the bound from Theorem 1.

Lemma 1. *For fixed D, X and Y, sharing $M = X$ achieves $\max_{\text{perfU}} \text{P}$ and $\delta_P^* = \frac{I(X;Y)}{H(Y)}$.*

Proof: Follows directly from the definition of $\delta_U(M)$ and $\delta_P(M)$ in (1) and the bound in Theorem 1. □

Sharing $M = X$ is not useful when the computation of X by the recipient is the only reason the provider agrees to release data. However, it is a valid strategy when the provider has additional incentive to share data, or when X is only an intermediate variable which requires further processing by the recipient.

IV. Maximizing Utility under Perfect Privacy

As shown in Section III, achieving $\max_{\text{perfU}} \text{P}$ is straightforward as it does not depend on Y. However, achieving $\max_{\text{perfP}} \text{U}$ is considerably more involved as it depends on both X and Y. This section aims at providing necessary conditions on $p(D,M)$ for M to achieve point $\max_{\text{perfP}} \text{U}$. Combining these conditions greatly restricts the space over which distributions $p(D,M)$ may achieve $\max_{\text{perfP}} \text{U}$; in fact, it allows us to express the maximization of utility under perfect privacy as a linear programming problem.

A. Necessary Conditions on $p(D,M)$ for Maximum Utility under Perfect Privacy

We provide three conditions. The first one is a necessary and sufficient condition for perfect privacy (with no constraint on utility). It is a simple consequence of our definition of perfect privacy, but we state it formally as it is convenient for our analysis. The remaining two are necessary conditions for maximum utility.

Condition 1. *For all $m \in \mathcal{M}$ and $y \in \mathcal{Y}$,*

$$\sum_{d\in g^{-1}(y)} p(d,m) = p(m)p(y). \tag{5}$$

Lemma 2. *RV M achieves perfect privacy if and only if M meets Condition 1.*

Proof: By definition of perfect privacy, $I(M;Y) = 0$ and for all $m \in \mathcal{M}$ and $y \in \mathcal{Y}$ we have $p(m,y) = p(m)p(y)$. The statement follows by noting that

$$p(m,y) = \sum_{d\in g^{-1}(y)} p(d,m). \tag{6}$$

□

We now try to maximize utility when Condition 1 is satisfied. Suppose that for every $m \in \mathcal{M}$ there is a unique $x \in \mathcal{X}$ such that $p(m,x) > 0$. Then, M achieves perfect utility. This observation leads to the intuition that utility is increased when each value $m \in \mathcal{M}$ jointly occurs only with a limited number of different values of $x \in \mathcal{X}$. Therefore, starting from a given RV M that achieves perfect privacy, we might be able to improve utility by performing local rearrangements of the joint distribution $p(D,M)$. The goal is to reduce the number of different values of X that can jointly occur with each value of M while preserving perfect privacy. We identify local patterns in the joint distribution $p(D,M)$ that can be manipulated for a guaranteed increase in utility at no cost in terms of privacy: if for a given M, $p(D,M)$ shows one of these patterns, then M does not achieve $\max_{\text{perfP}} \text{U}$.

Condition 2. *Given $d_1 \neq d_2$ in $\mathcal{D}$ and $m_1 \neq m_2$ in $\mathcal{M}$ such that $f(d_1) \neq f(d_2)$ and $g(d_1) = g(d_2)$, there exists a pair $(i,j) \in \{1,2\}^2$ such that $p(d_i, m_j) = 0$.*

Lemma 3. *If M achieves $\max_{\text{perfP}} \text{U}$, then M meets Condition 2.*

Proof: Suppose that M achieves $\max_{\text{perfP}} \text{U}$ but does not meet Condition 2. Then, there exist d_1, d_2, m_1 and m_2 so that all of the $p(d_i, m_j)$ terms, $(i,j) \in \{1,2\}^2$, are non-zero. We show that this leads to a contradiction by building an RV M_α which also achieves perfect privacy, but with better utility than M.

For a real number α in a well-chosen range, we define the RV M_α by its joint distribution $p_\alpha(D,M)$:

$$p_\alpha(d,m) = \begin{cases} p(d,m) + \alpha & \text{if } (d = d_1 \wedge m = m_1) \\ & \text{or } (d = d_2 \wedge m = m_2), \\ p(d,m) - \alpha & \text{if } (d = d_1 \wedge m = m_2) \\ & \text{or } (d = d_2 \wedge m = m_1), \\ p(d,m) & \text{otherwise.} \end{cases} \tag{7}$$

Notice that the joint distribution $p_\alpha(D, M_\alpha)$ is a locally perturbed version of $p(D,M)$. We now consider how $H(X|M_\alpha)$ varies with α. Because $p_\alpha(d,m)$ must remain between 0 and 1, α can only vary in some line segment $[\alpha_{\min}, \alpha_{\max}]$. Note that the constraints that the probabilities must remain non-negative are sufficient to find the bounds: taking for instance $p(d_1,m_1) + \alpha > 1$ would imply that $p(d_1,m_2) - \alpha < 0$. Therefore, $\alpha_{\min} = -\min(p(d_1,m_1), p(d_2,m_2))$ and $\alpha_{\max} = \min(p(d_2,m_1), p(d_1,m_2))$.

We now prove that $H(X|M_\alpha)$ is a concave function of α and therefore reaches its minimum for $\alpha = \alpha_{\min}$ or $\alpha = \alpha_{\max}$ (so that one of the $p_\alpha(d_i,m_j) = 0$). We use the following notation shortcuts

$$p^{ij} \triangleq p(m_j, x_i) = \sum_{d\in f^{-1}(x_i)} p(d,m_j), \tag{8}$$

$$p^{ij}_\alpha \triangleq p_\alpha(m_j, x_i) = \sum_{d\in f^{-1}(x_i)} p_\alpha(d,m_j), \tag{9}$$

where $x_i = f(d_i)$, and decompose $H(X|M_\alpha)$ into a term that depends on α and a constant. We use the fact that for all m and all x, $p_\alpha(m) = p(m)$ and $p_\alpha(x) = p(x)$. After some lengthy derivations we obtain

$$H(X|M_\alpha) = \\ -p^{11}_\alpha \log p^{11}_\alpha - p^{12}_\alpha \log p^{12}_\alpha - p^{21}_\alpha \log p^{21}_\alpha - p^{22}_\alpha \log p^{22}_\alpha + C. \tag{10}$$

Differentiating twice, we get

$$\frac{\partial^2}{\partial\alpha^2} H(X|M_\alpha) = -\frac{1}{p^{11}_\alpha} - \frac{1}{p^{12}_\alpha} - \frac{1}{p^{21}_\alpha} - \frac{1}{p^{22}_\alpha} < 0. \tag{11}$$

Thus, there exists an M_α such that $\delta_U(M_\alpha) < \delta_U(M)$.

Also, $p_\alpha(m,y) = \sum_{d\in g^{-1}(y)} p_\alpha(d,m)$, thus for $m \notin \{m_1, m_2\}$ or for $y \neq g(d_1) = g(d_2)$, we have $p_\alpha(m,y) = p(m,y)$ because for these values of d and m, $p_\alpha(d,m) = p(d,m)$. For $m \in \{m_1,m_2\}$ and $y = g(d_1) = g(d_2)$, we have

$$p_\alpha(m,y) = \sum_{d\in g^{-1}(y)} p_\alpha(d,m) = \sum_{d\in g^{-1}(y)} p(d,m) + \alpha - \alpha = p(m,y). \tag{12}$$

Therefore, $\delta_P(M_\alpha) = \delta_P(M) = 0$, which together with $\delta_U(M_\alpha) < \delta_U(M)$ contradicts the fact that M achieves $\max_{\text{perfP}} \text{U}$. □

Condition 3. *For all $m_0 \in \mathcal{M}$ and $d_1, d_2 \in \mathcal{D}$ such that $g(d_1) = g(d_2)$,*

$$d_1 \neq d_2 \Rightarrow p(d_1, m_0) = 0 \vee p(d_2, m_0) = 0. \tag{13}$$

Lemma 4. *If M achieves $\max_{\text{perfP}} \text{U}$, then M meets Condition 3.*

Proof: Suppose that M achieves $\max_{\text{perfP}} \text{U}$ but does not meet Condition 3. Then, there exist m_0, d_1 and d_2 such that $g(d_1) = g(d_2)$ and both $p(d_1, m_0) > 0$ and $p(d_2, m_0) > 0$. Now consider the RV M' on alphabet $\mathcal{M}' = \mathcal{M} \cup \{m'_0\}$ (where $m'_0 \notin \mathcal{M}$ is a new symbol), obtained by splitting symbol m_0 into m_0 and m'_0. Formally,

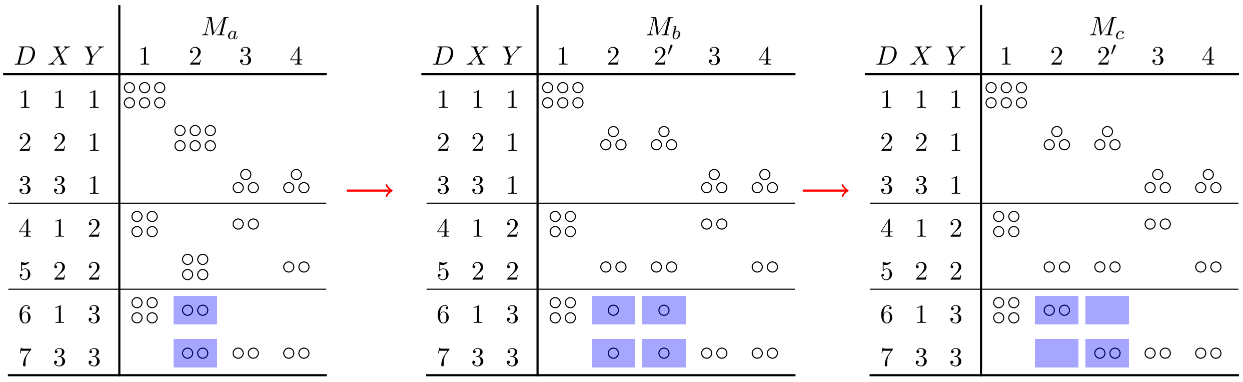
$$p'(d,m) = \begin{cases} p(d,m) & \text{if } m \notin \{m_0, m'_0\}, \\ p(d,m)/2 & \text{if } m \in \{m_0, m'_0\}. \end{cases} \tag{14}$$

Let us first notice that $\delta_U(M') = \delta_U(M)$:

$$H(X)\delta_U(M') = \sum_{m\in\mathcal{M}'} p'(m) H(X|M'=m) \\ = \sum_{m\in\mathcal{M}\setminus\{m_0\}} p(m)H(X|M=m) + 2\frac{p(m_0)}{2} H(X|M=m_0) \\ = H(X)\delta_U(M). \tag{15}$$

Similarly, $\delta_P(M') = \delta_P(M) = 0$. As M' does not satisfy Condition 2 (for m_0, m'_0, d_1 and d_2), by Lemma 3, M' does not achieve $\max_{\text{perfP}} \text{U}$, and so neither does M. □

Under the assumption that Condition 1 is satisfied, Fig. 2 illustrates for a simple synthetic example of D, X and Y how better utility can be obtained once Conditions 2 and 3 are met.



(a) While Condition 2 is met, Condition 3 is not: we split column $m = 2$ into 2 and $2'$, and obtain M_b that violates Condition 2.

(b) While Condition 3 is now met, Condition 2 is not. We rearrange the highlighted probabilities and gain utility without losing privacy.

(c) Lastly, Conditions 1–3 are met, the table is a candidate for maximum utility under perfect privacy (and does in fact achieve it).

Fig. 2: Illustration of the impact of Conditions 2 and 3 on the joint distributions $p(D, M)$, when M already meets Condition 1. Each $\circ$ represents a measure of $\frac{1}{49}$. Starting from M_a (left panel) which does not meet Condition 3, we reach M_b (center panel) with $\delta_U(M_b) = \delta_U(M_a)$ and $\delta_P(M_b) = \delta_P(M_a) = 0$, where M_b does not respect Condition 2. From M_b, we reach M_c (right panel) which still has $\delta_P(M_c) = 0$, and has $\delta_U(M_c) < \delta_U(M_b)$.

D	X	Y	M 1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	a_1	a_2	a_3	a_4	0	0	0	0	0	0	0	0
2	2	1	0	0	0	0	a_5	a_6	a_7	a_8	0	0	0	0
3	3	1	0	0	0	0	0	0	0	0	a_9	a_{10}	a_{11}	a_{12}
4	1	2	b_1	b_2	0	0	b_5	b_6	0	0	b_9	b_{10}	0	0
5	2	2	0	0	b_3	b_4	0	0	b_7	b_8	0	0	b_{11}	b_{12}
6	1	3	c_1	0	c_3	0	c_5	0	c_7	0	c_9	0	c_{11}	0
7	3	3	0	c_2	0	c_4	0	c_6	0	c_8	0	c_{10}	0	c_{12}

Fig. 3: General structure of the joint distribution for M satisfying Conditions 1–3 (for an example choice of X and Y). For each $m \in \{1, \dots, 12\}$, $\frac{a_m}{3} = \frac{b_m}{2} = \frac{c_m}{2}$. The table is thus entirely determined by the choice of the values $p(m) = a_m + b_m + c_m$ for each $m \in \{1, \dots, 12\}$.

B. Linear Programming Approach for Maximum Utility under Perfect Privacy

We now use Conditions 1–3 to formulate the maximization of utility under perfect privacy as a LP problem to compute $p(D, M)$ so that M achieves $\max_{\text{perfP}} \mathrm{U}$.

We start by defining the *support of D given $M = m$* as the set $S(m) = \{d \in \mathcal{D} : p(d, m) > 0\}$. Condition 3 can be written in terms of S as follows:

$$\forall m \in \mathcal{M}, \forall y \in \mathcal{Y}, |S(m) \cap g^{-1}(y)| = 1. \qquad (16)$$

Without loss of generality, we only consider those M's for which no distinct m_1 and m_2 have the same support. Otherwise, m_1 and m_2 can be merged into a single symbol with no effect on δ_U and δ_P (for the same reason for which we could split m_0 into two symbols in the proof of Lemma 4).

We denote by $\mathcal{S}$ the set of all possible supports S so that (16) is met. For the same instance of D, X and Y as in Fig. 2, we show in Fig. 3 the structure of the candidate joint distributions for $\max_{\text{perfP}} \mathrm{U}$. Notice that for any given m, Condition 1 requires that $\frac{a_m}{3} = \frac{b_m}{2} = \frac{c_m}{2}$, which in turn determines the conditional distribution of Y given $M = m$. Also, $|\mathcal{S}| = \prod_{y \in \mathcal{Y}} |g^{-1}(y)| = 3 \times 2 \times 2 = 12$. For M achieving Conditions 1–3, we can therefore characterize the joint distribution $p(D, M)$ with only $|\mathcal{S}|$ values (the vector $(p(m))_{m \in \mathcal{M}}$ with $\mathcal{M} = \{1, \dots, |\mathcal{S}|\}$). This greatly reduces the dimensionality of the space of the candidate joint distributions for $\max_{\text{perfP}} \mathrm{U}$.

We therefore formulate achieving $\max_{\text{perfP}} \mathrm{U}$ as the problem of finding $(p(m))_{m \in \mathcal{M}} \in [0, 1]^{|\mathcal{M}|}$ which minimizes

$$H(X|M) = \sum_{m \in \mathcal{M}} p(m) H(X|M = m), \qquad (17)$$

(where $\mathcal{M}$ is chosen so that $\{S(m) : m \in \mathcal{M}\} = \mathcal{S}$ with each support represented exactly once), under the constraints that for each $d \in \mathcal{D}$,

$$\sum_{m : d \in S(m)} p(d, m) = p(d), \qquad (18)$$

which can be written in terms of the $p(m)$:

$$\sum_{m : d \in S(m)} p(m) \Pr(Y = g(d)) = p(d). \qquad (19)$$

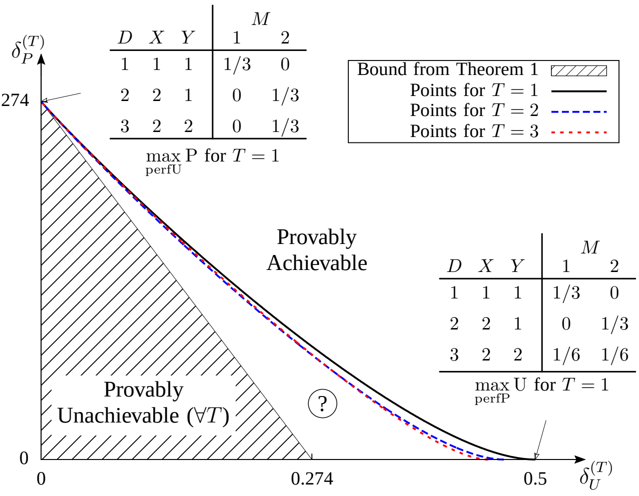
The above LP has an average-case complexity which is polynomial in $|\mathcal{S}|$, where $|\mathcal{S}|$ is upper bounded by $\left(\frac{|\mathcal{D}|}{|\mathcal{Y}|}\right)^{|\mathcal{Y}|}$.

V. Obtaining Better Parameters by Using Time

While $\max_{\text{perfU}} \mathrm{P}$ always achieves the bound from Theorem 1 as discussed in Section III, $\max_{\text{perfP}} \mathrm{U}$ in general does not. We try to reduce the gap between $\max_{\text{perfP}} \mathrm{U}$ and the theoretical bound by simultaneously using multiple time-slots. Rather than considering three RVs D, X and Y, we consider that they are part of three i.i.d. random processes (D_t), (X_t) and (Y_t). If we send a message M_t independently at every time slot t, the analysis remains the same as before. Hoping to reach better privacy and utility, we decompose the processes into groups of T time slots and treat each of these groups as a whole.

For an RV $M^{(T)}$, we define the utility and privacy parameters corresponding to a group of T time slots:

$$\delta_U^{(T)}(M^{(T)}) \triangleq \frac{H(X_1^T|M^{(T)})}{H(X_1^T)}, \delta_P^{(T)}(M^{(T)}) \triangleq \frac{I(M^{(T)};Y_1^T)}{H(Y_1^T)}. \qquad (20)$$



			M	
D	X	Y	1	2
1	1	1	1/3	0
2	2	1	0	1/3
3	2	2	0	1/3

$\max_{\text{perfU}} \text{P}$ for $T = 1$

			M	
D	X	Y	1	2
1	1	1	1/3	0
2	2	1	0	1/3
3	2	2	1/6	1/6

$\max_{\text{perfP}} \text{U}$ for $T = 1$

Fig. 4: Points $(\delta_U^{(T)}, \delta_P^{(T)})$ for $\max_{\text{perfU}} \text{P}$, $\max_{\text{perfP}} \text{U}$ and tradeoffs between the two, for T from 1 to 3, on an example with $|\mathcal{D}| = 3$. While $\max_{\text{perfU}} \text{P}$ does not change with an increase in T, $\max_{\text{perfP}} \text{U}$ is shifted leftwards, and better tradeoff points are achieved for $T = 2, 3$ than for $T = 1$.

While $M^{(T)}$ plays the same role as M did before, we include a superscript (T), to indicate that M spans T time slots. We may omit this superscript when $T = 1$. The following lemma provides a way of obtaining RVs $M^{(T)}$ using an RV M designed for a single time slot.

Lemma 5. *Let f and g be two (deterministic) functions, and (D_t), (X_t) and (Y_t) be three i.i.d. random processes so that for each t, $X_t = f(D_t)$ and $Y_t = g(D_t)$. Then, for any RV M given by the joint distribution $p(D, M)$, consider the i.i.d. random vector M_1^T over $\mathcal{M}^T$ given by its joint distribution with D_1^T,*

$$p^{(T)}(d_1^T, m_1^T) = \prod_{t=1}^{T} p(d_t, m_t). \qquad (21)$$

Here, M_1^T conserves the utility and privacy parameters of M:

$$\delta_U^{(T)}(M_1^T) = \delta_U(M), \ \delta_P^{(T)}(M_1^T) = \delta_P(M). \qquad (22)$$

Proof: The proof follows from elementary information-theoretic properties of i.i.d. processes. □

Lemma 5 ensures that the optimal $(\delta_U^{(T)}, \delta_P^{(T)})$ pairs are no worse than the optimal (δ_U, δ_P) pairs. It is in fact possible to build RVs for T time slots that achieve results strictly better than the best that can be obtained for a single time slot. For instance, in Fig. 4, we consider a small example with $\mathcal{D} = \{1, 2, 3\}$, $f(1) = f(2) = g(1) = 1$, $f(3) = g(2) = g(3) = 2$. We plot the bound from Theorem 1, which delimits a provably unachievable region. For each $T \in \{1, 2, 3\}$, we also plot the points $\max_{\text{perfU}} \text{P}$ and $\max_{\text{perfP}} \text{U}$ obtained using respectively Sections III and IV, and use a heuristic algorithm to achieve and plot tradeoffs between utility and privacy by combining the joint distributions of $\max_{\text{perfU}} \text{P}$ and $\max_{\text{perfP}} \text{U}$. Combining these two distributions requires computing "compatibility scores" between elements of the alphabets for the extreme points and using these scores to carefully construct a common alphabet. The detailed description of this procedure is left for an extended version of the paper. For each T, these tradeoffs form a curve that delimits a region of provably achievable pairs (δ_U, δ_P). The maximum utility that can be reached under perfect privacy using a single time slot is $\delta_U^{(1)*} = 0.5$. However, with two time slots, it can be reduced to $\delta_U^{(2)*} = 0.468$, and to $\delta_U^{(3)*} = 0.452$ for three time slots. It appears challenging to establish whether the points in the area in between the theoretical bound and the heuristically obtained tradeoffs are achievable or not (except for the points $(\delta_U^{(T)}, 0)$ with $\delta_U^{(T)} < \delta_U^{(T)*}$, which are unachievable by definition).

The bound from Theorem 1 remains the same regardless of T because of the assumption that the processes are i.i.d.: for any T, $\delta_U^{(T)*} \geq \frac{I(X;Y)}{H(X)} = 0.274$. An interesting question is to determine whether $\delta_U^{(T)*}$ approaches this bound when T goes to infinity. The time complexity of the method from Section IV-B prohibits the computation of $\max_{\text{perfP}} \text{U}$ for $T > 3$, even for simple examples like the one shown in Fig. 4.

VI. Conclusion

In this work, we took the first step towards creating a framework for protecting data against unwanted inferences. We define utility-privacy parameters based on information-theoretic notions, which allow us to effectively capture how much the recipient can infer from the shared data. We identify bounds on these parameters, and provide constructive mechanisms for achieving these bounds. There are multiple facets to the problem which we intend to study, e.g., the effect of side channel information and correlation between the shared data samples over time. Tools from coding theory may also prove valuable in designing schemes in order to achieve better tradeoff points, especially in the case of multiple time slots.

Acknowledgement

This research is based upon work supported in part by the U.S. ARL, U.K. Ministry of defense under Agreement Number W911NF-06-3-0001, by the NSF under awards CNS-0910706 and CNS-1213140, by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via US Navy (USN) SPAWAR Systems Center Pacific (SSCPac). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, ODNI, IARPA, USN, SSCPac, or represent the official policies of the U.S. ARL, the U.S. Government, the U.K. Ministry of Defense or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, TAMC, 2008.
[2] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," SIGMOD, 2010.
[3] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time series compressibility and privacy," VLDB, 2007.
[4] Y.-S. Moon, H.-S. Kim, S.-P. Kim, and E. Bertino, "Publishing time-series data under preservation of privacy and distance orders," DEXA, 2010.
[5] S. Chakraborty, K. R. Raghavan, M. P. Johnson, and M. B. Srivastava, "A framework for context-aware privacy of sensor data on mobile systems," HotMobile, 2013.
[6] L. Sankar, S. Rajagopalan, and H. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 6, pp. 838–852, 2013.
[7] M. Bezzi, "An information theoretic approach for privacy metrics," *Transactions on Data Privacy*, vol. 3, pp. 199–215, Dec. 2010.
[8] L. Sankar, S. Rajagopalan, S. Mohajer, and H. Poor, "Smart meter privacy: A theoretical framework," *Smart Grid, IEEE Transactions on*, vol. 4, no. 2, pp. 837–846, 2013.