Privacy-Utility Tradeoff in a Guessing Framework Inspired by Index Coding

Yucheng Liu†, Ni Ding*, Parastoo Sadeghi†, and Thierry Rakotoarivelo*
†Research School of Electrical, Energy and Materials Engineering, Australian National University, Australia
*Data61, Commonwealth Scientific and Industrial Research Organisation, Australia
Emails: †{yucheng.liu, parastoo.sadeghi}@anu.edu.au, *{ni.ding, thierry.rakotoarivelo}@data61.csiro.au

Abstract—This paper studies the tradeoff in privacy and utility in a single-trial multi-terminal guessing (estimation) framework using a system model that is inspired by index coding. There are $n$ independent discrete sources at a data curator. There are $m$ legitimate users and one adversary, each with some side information about the sources. The data curator broadcasts a distorted function of sources to legitimate users, which is also overheard by the adversary. In terms of utility, each legitimate user wishes to perfectly reconstruct some of the unknown sources and attain a certain gain in the estimation correctness for the remaining unknown sources. In terms of privacy, the data curator wishes to minimize the maximal leakage: the worst-case guessing gain of the adversary in estimating any target function of its unknown sources after receiving the broadcast data. Given the system settings, we derive fundamental performance lower bounds on the maximal leakage to the adversary, which are inspired by the notion of confusion graph and performance bounds for the index coding problem. We also detail a greedy privacy enhancing mechanism, which is inspired by the agglomerative clustering algorithms in the information bottleneck and privacy funnel problems.

I. INTRODUCTION

In this paper, we consider an information-theoretic multi-terminal guessing framework with side information. Our model is inspired by that of the index coding problem [1], [2], but with a significant twist to place emphasis on privacy. Index coding is a communication problem where a sender attempts to efficiently broadcast to multiple users through a noiseless channel. See [3] and references therein. Instead of trying to minimize the broadcast rate, in our framework the sender’s goal is to balance the privacy and utility performance in the broadcast, where both are measured based on the success rate of correctly estimating a certain parameter of interest about the sources through a single guess. More specifically, we consider multiple independent sources available at a data curator and assume that there are multiple legitimate users, as well as one adversary in the system. Each party (a user or the adversary) knows some sources a priori as side information. The data curator broadcasts (discloses) a distorted function of the sources to the users, which is also overheard by the adversary.

For the adversary, we adopt the maximal leakage introduced in [4] as the privacy metric. It measures the worst-case information leakage in terms of the gain of the adversary in maximum a posteriori estimation of any target function of the unknown sources after and before observing the disclosed data. For legitimate users, we define our utility metric such that it also reflects the improvement in users’ guessing ability. Quite often in practice different sources are of different levels of priority to a user. We capture this by dividing the unknown sources of each user into two subsets. Some essential source are required to be perfectly reconstructed by the user. That is, the correct guessing probability of such sources after observing the disclosed data is non-negotiable and must be 1. The remaining sources are less critical, for which the user requires the guessing gain about each such source to be larger than a certain negotiable threshold. The privacy-utility tradeoff is cast as a constrained optimization problem, where the objective is to minimize the privacy leakage to the adversary, conditioned that the requirements on the utility of the unknown sources for the legitimate users are satisfied.

This paper contributes mainly in two aspects. First, we derive two lower bounds on the privacy leakage given the source distribution and utility constraints (cf. Theorems 1-2) from different perspectives. These lower bounds serve as converse results (fundamental performance limits) for the privacy leakage. Second, we propose a greedy algorithm as the privacy mechanism (cf. Algorithm 1) inspired by the agglomerative clustering method used in the information bottleneck [5] and privacy funnel problems [6]. We leverage the connection between our data disclosure problem and the index coding problem when investigating both the converse results and the greedy algorithm.

Notation: For non-negative integers $a$ and $b$, $[a]$ denotes the set $\{1, 2, \ldots, a\}$ and $[a : b]$ denotes the set $\{a, a+1, \ldots, b\}$. For a set $S$, $|S|$ denotes its cardinality. For any discrete random variable $Z$ with probability distribution $P_Z$, we denote its alphabet by $Z$ with realizations $z \in Z$. We denote an estimator of $Z$ by $\hat{Z}$, whose alphabet is also $Z$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Assume that a data curator observes $n$ independent discrete random sources $X_1, X_2, \ldots, X_n$. We assume a general distribution $P_{X_i}$ for each source. Without loss of generality, we assume every source has full support. For brevity, when we say source $i$, we mean source $X_i$. For any $S \subseteq [n]$, set $S^c = [n] \setminus S$, $X_S = (X_i : i \in S)$, $x_S = (x_i : i \in S)$, and $\mathcal{X}_S = \prod_{i \in S} \mathcal{X}_i$. The data curator broadcasts a distorted version of $X_{[n]}$, denoted by $Y$, generated according to the privacy mechanism $P_{Y|X_{[n]}}$ to a number of legitimate users, which is also overheard by a single adversary.
We consider a guessing framework where both the legitimate users and the adversary attempt to estimate a certain parameter \( V \) of interest about sources in a single trial. Both the privacy and utility are defined in terms of a guessing gain \( r \) as defined below. Consider any party (an user or the adversary) that wishes to guess \( V \) with the side information \( Z \). Note that \( V \perp Z \) due to source independence. The party aims to maximize the correct guessing probability of \( V \) upon observing \( Y \) (i.e., the party employs the maximum \textit{a posteriori} estimator). For each \((y,z) \in Y \times Z\), we define the ratio between such maximized guessing probability after and before observing a \( y \in Y \) given \( z \in Z \) as

\[
\frac{\max_y \mathbb{E}[\mathbb{P}_{Y|Z=z}(V|y,z)]}{\max_v \mathbb{E}[\mathbb{P}_{Y|Z=z}(V|z)]} \quad (1)
\]

\[
= \frac{\max_y \mathbb{P}_{Y,Z}(v|y,z)}{\max_v \mathbb{P}_{Y,Z}(v|z)} \quad (2)
\]

\textbf{Adversary:} We assume the adversary has side information \( X_P \) for some \( P \subset [n] \), and is interested in a (possibly randomized) discrete function \( U \) of the sources \( X_Q \). It does not know, where \( Q \equiv P^c \). Quite often in practice, this function is chosen by the adversary, and is unknown to the data curator. Therefore, we consider a worst-case privacy leakage measure, a conditional version of maximal leakage (MaxL) from [4], as our privacy metric.

\textit{Definition 1 (Maximal Leakage, [4]):} Given a finite discrete joint distribution \( P_{X[n]:Y} \), the maximal leakage from \( X_Q \) to \( Y \) given \( X_P \) is defined as

\[
\mathcal{L}_{\text{max}}(X_Q \rightarrow Y|X_P) \triangleq \sup_{U \in \mathcal{X}^c} \mathbb{L}(U \rightarrow Y|X_P), \quad (3)
\]

where

\[
\mathbb{L}(U \rightarrow Y|X_P) \triangleq \log \mathbb{E}\left[\mathbb{P}_{Y,X_P}(r(U \rightarrow Y|X_P))\right] \quad (4)
\]

\textit{Remark 1:} Note that the MaxL in Definition 1 assumes a different Markov chain model from [4]; Section II-E for our problem, the Markov chain model studied in [4] always reduces to \( U \rightarrow X[n] \rightarrow Y \) regardless of \( Q \)

For the rest of the paper, we refer to \( \mathcal{L}_{\text{max}}(X_Q \rightarrow Y|X_P) \) as \( \mathcal{L}_{\text{max}} \) when there is no ambiguity. A computable expression of the MaxL in Definition 1 is presented as

\[
\mathcal{L}_{\text{max}} = \log \sum_{y:z} \max_{x|y} \mathbb{P}_{Y,X_P|X_Q}(y,x|z), \quad (5)
\]

which can be obtained following a similar approach to [4]. We omit the derivation details due to limited space.

\textbf{Legitimate Users:} There are \( m \) legitimate users. User \( i \in [m] \) knows some sources \( X_{A_i} \), a priori as side information for some \( A_i \subset [n] \), and is interested in all the remaining sources \( X_{A^c} \). More specifically, for each user \( i \), the sources \( X_{A^c} \) are divided into two groups of different levels of priority:

- Source \( X_{W_i} \) for some \( W_i \subset A_i \) are indispensable to the user, and thus they must be correctly guessed by user \( i \) with probability of 1 (i.e., perfect decoding).

- The rest of the sources, \( X_{G_i} \), where \( G_i \equiv A_i^c \setminus W_i \), are less essential yet still useful/interesting. Thus, user \( i \) requires the guessing ability gain upon observing \( Y \) to be larger than a certain threshold \( d_i \).

These result in the following two kinds of utility constraints. For any \( i \in [m] \), we have

\[
H(X_{W_i}|Y,X_{A_i}) = 0, \quad \forall i \in [m], \quad (6)
\]

\[
D(X_{G_i} \rightarrow Y|X_{A_i}) \geq d_i, \quad \forall i \in [m], \quad (7)
\]

where \( D(X_{G_i} \rightarrow Y|X_{A_i}) \) is defined as

\[
D(X_{G_i} \rightarrow Y|X_{A_i}) \equiv \mathbb{E}_{Y,X_{A_i}}[\log r(X_{G_i} \rightarrow Y|X_{A_i})]. \quad (8)
\]

Note that for constraint (7), each legitimate user \( i \) is interested in obtaining the source \( X_{G_i} \), rather than a function/feature of \( X_{G_i} \). We simplify the notation \( D(X_{G_i} \rightarrow Y|X_{A_i}) \) to \( D_i \) when there is no ambiguity.

\textit{Remark 2:} Note the subtle difference between \( D_i \) and \( \mathcal{L}(X_{G_i} \rightarrow Y|X_{A_i}) = \log \mathbb{E}_{Y,X_{A_i}}[r(X_{G_i} \rightarrow Y|X_{A_i})] \), \( \mathcal{L} \) defined in (4). The latter is lower bounded by the former due to Jensen’s inequality. From the data curator’s viewpoint, requesting \( D_i \) to be above a certain threshold is more stringent than requesting \( \mathcal{L}(X_{G_i} \rightarrow Y|X_{A_i}) \) to be above that threshold. We use \( D_i \) rather than \( \mathcal{L}(X_{G_i} \rightarrow Y|X_{A_i}) \) as our utility measure as it leads to a simple closed-form result characterizing \( \mathcal{L}_{\text{max}} \) in terms of \( D_i \) (cf. Lemma 2).

\textit{Remark 3:} The two types of utility constraints in (6) and (7) can be unified to be represented in terms of the same function \( D \): The perfect decoding constraint (6) is equivalent to requesting that

\[
D(X_{W_i} \rightarrow Y|X_{A_i}) = \mathbb{E}_{Y,X_{A_i}}[\log \frac{1}{\max_x \mathbb{P}_{X_{W_i}|X_{W_i}}(x|y)}] = H_{\infty}(X_{W_i}), \quad (9)
\]

where \( H_{\infty}(X_{W_i}) \) denotes the min-entropy (i.e., R´enyi entropy of order \( \infty \)) of \( X_{W_i} \). One can show that for any \( i \in [m] \), \( D_i \leq H_{\infty}(X_{G_i}) \). Consequently, to avoid an invalid system model we always require that \( 0 \leq d_i \leq H_{\infty}(X_{G_i}) \).

\textbf{Privacy-Utility Tradeoff:} We denote such system by the 5-tuple \((P_{X[n]}, A, W, d, P)\), where \( A \equiv \{A_i, i \in [m]\}, W \equiv \{W_i, i \in [m]\}, \) and \( d \equiv \{d_i, i \in [m]\}\). Note that \( G_i \) is determined by \( W_i \) and \( A_i \), and \( Q \) is determined by \( P \).

To design the privacy mechanism \( P_{Y|X[n]} \), we need to consider the fundamental tradeoff between the privacy and utility. Any data distortion that reduces the information leakage to the adversary can decrease the utility obtained by the users. Such tradeoff is formulated by the following constrained optimization problem.

\[
\inf_{P_{Y|X[n]} \in \mathcal{P}_{Y|X[n]}(P_{X[n]}; A, W, d, P)} \mathcal{L}_{\text{max}}(X_Q \rightarrow Y|X_P), \quad (10)
\]

where \( \mathcal{P}_{Y|X[n]}(P_{X[n]}; A, W, d, P) \) denotes the collection of randomized mappings \( P_{Y|X[n]} \) that satisfy (6) and (7) for the problem \((P_{X[n]}; A, W, d, P)\).

Due to the non-convexity of (10), instead of providing an explicit solution, we derive lower bounds on \( \mathcal{L}_{\text{max}} \) by
taking inspiration from index coding. These bounds serve as fundamental performance limits that cannot be surpassed by any mechanism because they are enforced by the system \((P_{X_{[n]}}, A, W, d, P)\). In Section IV, we design an achievable mechanism based on the idea of agglomerative clustering.

III. LOWER BOUNDS ON THE PRIVACY LEAKAGE

We derive two information-theoretic lower bounds on the privacy leakage. One is based on the utility constraint (6) only, while the other is obtained based on both (6) and (7).

A. Lower Bound Based on the Confusion Graph

The utility constraint (6) indicates that for user \(i\), any different realizations \(x_{W_i} \neq x_{W_i} \in X_{W_i}\) must be distinguishable based on the released \(y \in Y\), as well as the user’s side information \(x_{A_i} \in X_{A_i}\). To describe such distinguishability, we recall the notion of confusion graph for index coding [7].

**Definition 2** (Confusion graph [7]): Any two realizations of the \(n\) sources \(x_{[n]}^{1}, x_{[n]}^{2} \in X_{[n]}\) are confusable if there exists some user \(i \in [n]\) such that \(x_{i}^{1} \neq x_{i}^{2}\) and \(x_{A_i}^{1} = x_{A_i}^{2}\). A confusion graph \(\Gamma\) is an undirected graph with \(|X_{[n]}|\) vertices such that every vertex corresponds to a realization \(x_{[n]}\) and an edge connects two vertices if and only if their corresponding realizations are confusable.

Therefore, to ensure (6), a group of realizations of \(X_{[n]}\) can be mapped to the same \(y\) with nonzero probability only if they are pairwise not confusable. More rigorously, for any \(S \subseteq [n]\), define

\[
\mathcal{Y}(x_S) = \{ y \in Y : P_{Y|x_S}(y|x_S) > 0 \}, \quad \forall x_S \in X_S, \\
\mathcal{X}_S(y) = \{ x_S \in X_S : P_{Y|x_S}(y|x_S) > 0 \}, \quad \forall y \in Y.
\]

Then, we have the following lemma. We omit the proof as it can be simply done by contradiction.

**Lemma 1**: Given a \(P_{Y|x_S}\) satisfying (6), for any two confusable \(x_{[n]}^{1}, x_{[n]}^{2} \in X_{[n]}\), we have \(\mathcal{Y}(x_{[n]}^{1}) \cap \mathcal{Y}(x_{[n]}^{2}) = \emptyset\).

Given a set \(S \subseteq [n]\) and a specific realization \(x_S \in X_S\), we define \(\Gamma(x_S)\) as the subgraph of \(\Gamma\) induced by all the vertices \(x_{[n]}\) such that \(x_{[n]} = (x_S, x_{S^c})\) for some \(x_{S^c} \in X_{S^c}\). Notice that for any \(x_{[n]} \neq x_{[n]}^k \in X_S\) and \(x_{[n]} \neq x_{[n]}^k \in X_{S^c}\), \(x_{[n]}^{k}\) is confusable and if only if \((x_{[n]}^{k}, x_{[n]}^{k})\) and \((x_{[n]}^{k}, x_{[n]}^{k})\) are confusable. Hence, given \(S \subseteq [n]\), the subgraphs \(\Gamma(x_S), x_S \in X_S\) are identical to each other, and thus we simply denote any such subgraph by \(\Gamma(S)\).

We present our main result of this subsection as follows.

**Theorem 1**: For the problem (10) with confusion graph \(\Gamma\),

\[
\mathcal{L}_{\text{max}} \geq \log \omega(\Gamma(P)),
\]

where \(\omega(\cdot)\) is the clique number (size of the largest clique) of a graph.

**Proof**: Consider any \(x_P \in X_P\). There exists some realizations \(x_{Q}^{1}, x_{Q}^{2}, \ldots, x_{Q}^{n(\Gamma(x_P))} \in X_Q\) whose corresponding vertices in the subgraph \(\Gamma(x_P)\) form a clique, which indicates that the realizations \((x_P, x_{Q}^{1}), (x_P, x_{Q}^{2}), \ldots, (x_P, x_{Q}^{n(\Gamma(x_P))})\) also form a clique in \(\Gamma\). That is, the realizations \((x_P, x_{Q}^{k})\) are pairwise confusable. Then by Lemma 1, for any \(k \neq k' \in [n(\Gamma(x_P))]\), we have

\[
\mathcal{Y}((x_P, x_{Q}^{k})) \cap \mathcal{Y}((x_P, x_{Q}^{k'})) = \emptyset.
\]

Therefore, we have

\[
\sum_{y} \max_{x_Q} P_{Y|x_Q}(y|x_P, x_Q) \\
\geq \sum_{k \in [\omega(\Gamma(x_P))]} \sum_{y \in \mathcal{Y}(x_P, x_{Q}^{k})} \max_{x_Q} P_{Y|x_Q}(y|x_P, x_Q) \\
\geq \sum_{k \in [\omega(\Gamma(x_P))]} \sum_{y \in \mathcal{Y}(x_P, x_{Q}^{k})} P_{Y|x_Q}(y|x_P, x_{Q}^{k}) \\
= \sum_{k \in [\omega(\Gamma(x_P))]} 1 = \omega(\Gamma(x_P)) = \omega(\Gamma(P)),
\]

where the first inequality follows from (12). Hence, we have

\[
\mathcal{L}_{\text{max}} = \log \sum_{x_P} P_{X_P}(x_P) \sum_{y \in Y} \max_{x_Q} P_{Y|x_Q}(y|x_P, x_Q) \\
\geq \log \sum_{x_P} P_{X_P}(x_P) \cdot \omega(\Gamma(P)) = \log \omega(\Gamma(P)),
\]

where the first equality is due to source independence, and the inequality follows from (13).

B. Lower Bound Based on Guessing Gain and Polymatroidal Functions

We introduce a key lemma that serves as the baseline in the lower bound to be developed in this subsection.

**Lemma 2**: For the problem (10), we have

\[
\mathcal{L}_{\text{max}} \geq \max_{i \in [n]; G_i \subseteq Q} \{ I(X_Q; Y|X_P), \max_{i \in [n]; G_i \subseteq Q} \Delta_i \}.
\]

where \(\Delta_i = D_i + H(X_{W_i \cap Q}) + I(X_{A_i \cap Q}; Y|X_P)\).

The proof is presented in Appendix A.

For a given system \((P_{X_{[n]}}, A, W, d, P)\), \(H(X_{W_i \cap Q})\) has a fixed value and \(D_i\) is lower bounded by \(d_i\) according to (7). Hence, the only terms in (14) that still depend on \(P_{Y|x_S}\) are the mutual information \(I(X_P; Y|X_Q)\) and \(I(X_{A_i \cap Q}; Y|X_P)\). We further bound these mutual information terms below.

We draw inspiration from the polymatroidal bound [8], [9] for index coding. The bound is based on the polymatroidal axioms, which capture Shannon-type inequalities on the entropy function and play a central role in computing converse results in network information theory [10].

**Lemma 3**: Consider the system \((P_{X_{[n]}}, A, W, d, P)\).

For any disjoint \(V, Z \subseteq [n]\), we have

\[
I(X_V; Y|X_Z) = g(Z^c) - g(Z^c \cap V^c),
\]

for some polymatroidal set function \(g(S)\), \(S \subseteq [n]\) such that for any \(i \in [n]\), \(W \subseteq W_i, G \subseteq V^c \cap A_i^c\),

\[
H(X_W) = g(G \cup W) - g(G),
\]

and

\[
g(\emptyset) = 0, \quad g(S') \geq g(S), \quad \text{if } S \subseteq S', \quad g(S' \cup S) \geq g(S' \cap S).
\]

**Proof**: Define \(g(S) = H(Y|X_{S'}) - H(Y|X_{[n]})\), \forall S \subseteq [n]\). We have \(I(X_V; Y|X_Z) = g(Z^c) - g(Z^c \cap V^c)\). It remains to show that this \(g(S)\) satisfies (16)-(19).
For (16), consider any $i \in [n]$, $W \subseteq W_i$, $G \subseteq W^c \cap A_i^c$. Set $A = [n] \setminus W \setminus G$, and one can verify that $A_i \subseteq A$. Hence,

$$H(X_W) = H(X_W | X_A) - H(X_W | Y, X_A)$$

(20)

$$= H(Y | X_A) - H(Y | X_W, X_A) = g(W, G) - g(G),$$

where (20) is due to source independence, (6), and $A_i \subseteq A$.

For (17), we have $g(0) = H(Y | X_{[i]}) - H(Y | X_{[i]}) = 0$. For (18), for any $S \subseteq S' \subseteq [n]$, $S'' \subseteq S'$, and thus

$$g(S') = H(Y | X_{S''}) \geq H(Y | X_{S'}) = g(S).$$

For (19), consider any $S, S' \subseteq [n]$. Set $S'' \cap S'' = S_0$, $S' \cap S_0 = S_1$, and $S'' \cap S_0 = S_2$. We have

$$g(S') + g(S) = \Delta(Y, Z) \geq \max \{g(Z') - g(Z' \cap V')\}.$$

Combining Lemmas 2 and 3 gives the following result.

**Theorem 2:** For the problem (10), we have

$$\mathcal{L}_{\max} \geq \max\{\Delta(Q, P), \max_{i \in [n]: G_i \subseteq Q} \Gamma^{LP}_i\},$$

where $\Gamma^{LP}_i \triangleq \Delta_i + H(X_{W_i} | Q) + \Delta(A_i \cap Q, P)$.

**Remark 4:** In the index coding problem, we usually assume uniformly distributed independent sources and a deterministic mapping $P_{Y|x_{[i]}}$. In contrast, Theorems 1 and 2 hold for any discrete independent source distribution and make no assumption on the privacy mechanism.

In general, Theorems 1 and 2 can outperform each other.

**IV. Privacy Mechanism Design**

We develop a greedy algorithm to provide a solution for the problem (10). The algorithm is based on the agglomerative clustering method, which has been used in the information bottleneck [5] and the privacy funnel problem [6].

Consider a given system $(P_{X_{[i]}}, A, W, d, P)$. To design the privacy mechanism $P_{Y|x_{[i]}}$, we start from the one-to-one deterministic mapping with $\hat{Y} = X_{[i]}$ and $P_{Y|x_{[i]}}(y|x_{[i]}) = 1$ iff $y = x_{[i]}$ and then iteratively merge some elements of $\hat{Y}$ to make the privacy leakage smaller (in other words, we “blur” the revealed information), while still ensuring the utility for the users at an acceptable level, i.e., satisfying (6) and (7). In particular, to ensure (6), we again utilize the notion of confusion graph [7] introduced in Section III-A.

Based on the merging idea discussed above, we propose an agglomerative clustering algorithm in Algorithm 1. Let $Y_{y_1,y_2}$ be the resulting Y from merging any $y_1, y_2 \in Y$. Let $\Theta$ denote the collection of $\{y_1, y_2\}$ such that merging them does not violate (6) and (7) and strictly reduces the privacy leakage to the adversary:

$$\Theta \leftarrow \{\{y_1, y_2\} \in Y \times Y: y_1 \neq y_2, \text{any two} \ x_{[i]}, x_{[j]} \in X_{[i]}(y_1) \cup X_{[i]}(y_2) \text{are not confusable,} \right.$$

$$\left. D(X_{G_i} \rightarrow Y_{y_1,y_2}|X_{A_i}) \geq d_i, \forall i \in [m], \right.$$ (21)

$$\mathcal{L}_{\max}(X_Q \rightarrow Y_{y_1,y_2}|X_P) < \mathcal{L}_{\max}(X_Q \rightarrow Y|X_P).$$

The algorithm terminates when $\Theta$ becomes an empty set.

**Algorithm 1: Agglomerative clustering algorithm for problem (10)**

**Input:** The system $(P_{X_{[i]}}, A, W, d, P)$.

**Output:** Privacy-preserving mechanism $P_{Y|x_{[i]}}$

1. Initialization: $\Theta \leftarrow X_{[i]}$, $P_{Y|x_{[i]}}(y|x_{[i]}) \leftarrow 1$ if $y = x_{[i]}$ and obtain $\Theta$ based on $Y$ by (21);
2. repeat
3. $\{y_1', y_2'\} \leftarrow \arg \min_{\{y_1, y_2\} \in \Theta} \mathcal{L}_{\max}(X_Q \rightarrow Y_{y_1',y_2'}|X_P)$;
4. Merge $y_1'$ and $y_2'$ into $\tilde{y}$: $\tilde{y} \leftarrow \{y_1', y_2'\}$;
5. Obtain the new $Y$ by letting $Y \leftarrow Y \setminus \{y_1', y_2'\} \cup \{\tilde{y}\}$ and $P_{Y|x_{[i]}}'(y|x_{[i]}) \leftarrow P_{Y|x_{[i]}}(y_1'|x_{[i]}) + P_{Y|x_{[i]}}(y_2'|x_{[i]})$ for any $x_{[i]} \in X_{[i]}$ while keeping the rest of $P_{Y|x_{[i]}}$ unchanged;
6. Obtain the new $\Theta$ by (21) based on updated $Y$;
7. until $\Theta = \emptyset$;
8. return $P_{Y|x_{[i]}}$;

**Remark 5:** To compute Algorithm 1 more efficiently, notice that finding $\arg \min_{\{y_1, y_2\} \in \Theta} \mathcal{L}_{\max}(X_Q \rightarrow Y_{y_1,y_2}|X_P)$ is equivalent to finding $\arg \max_{\{y_1, y_2\} \in \Theta} (2\mathcal{L}_{\max}(X_Q \rightarrow Y|X_P) - 2\mathcal{L}_{\max}(X_Q \rightarrow Y_{y_1,y_2}|X_P))$ in step 3, which can be computed as

$$\quad \mathcal{L}_{\max}(X_Q \rightarrow Y|X_P) - 2 \mathcal{L}_{\max}(X_Q \rightarrow Y_{y_1,y_2}|X_P)$$

$$\quad = \sum_{x_P} \max_{x_Q} P_{X_P}(x_P) \cdot P_{Y|x_{[i]}} (y_1 | x_P, x_Q)$$

$$\quad + \sum_{x_P} \max_{x_Q} P_{X_P}(x_P) \cdot P_{Y|x_{[i]}} (y_2 | x_P, x_Q)$$

$$\quad - \sum_{x_P} \max_{x_Q} P_{X_P}(x_P) \cdot P_{Y|x_{[i]}}(\tilde{y}|x_P, x_Q)$$

$$\quad = \sum_{x_P \in X_P(y_1)} P_{X_P}(x_P) \cdot 1 + \sum_{x_P \in X_P(y_2)} P_{X_P}(x_P) \cdot 1$$

$$\quad - \sum_{x_P \in X_P(\tilde{y})} P_{X_P}(x_P) \cdot 1$$

$$\quad = P_{X_P}(X_P(y_1)) + P_{X_P}(X_P(y_2)) - P_{X_P}(X_P(\tilde{y})).$$

**V. Concluding Remarks**

To evaluate the performance of our main results, we consider 200 systems $(P_{X_{[i]}}, A, W, d, P)$ randomly generated according to the following conditions:

1. Note that such one-to-one mapping allows every user to perfectly reconstruct every source and thus definitely satisfies (6) and (7). Nevertheless, it also leads to the largest privacy leakage as the adversary can also perfectly reconstruct $X_Q$ and subsequently any function $U$ it is intersected in.
\[ n = m = 5, W_i = \{i\} \text{ for any user } i \in [5], \text{ and } A \text{ is generated based on a randomly chosen graph } G \text{ from the 9608 nonisomorphic 5-vertex directed graphs [3]} \text{ such that } A_i = \{j \in [5] : (j, i) \in G\}. \]

- For any \( i \in [5], X'_i = \{0, 1\} \) and \( X_i \sim \text{Bern}(p_i) \), where \( p_i \) is uniformly randomly chosen from range \((0, 1)\);
- For any \( i \in [5], d_i = \hat{d}_i \cdot 1/\max(X_G), \) where \( \hat{d}_i \) is uniformly randomly chosen from range \((0, 1)\);
- \( P \subseteq [5] \) is randomly generated assuring that \( |P| \leq 2 \).

For each system, we compute the lower bounds \( \mathcal{L}_{201}^{\text{Thm.1}} \) and \( \mathcal{L}_{201}^{\text{Thm.2}} \) by Theorems 1 and 2, respectively. We also compute the MaxL according to the privacy mechanism for the majority of tested problems. A lower ratio \( R = \mathcal{L}_{201}^{\text{Alg.1}} / \max(\mathcal{L}_{201}^{\text{Thm.1}}, \mathcal{L}_{201}^{\text{Thm.2}}) \). A lower ratio \( R \) (close to 1) means that our converse and achievable results perform well and are quite close to the optimal \( \mathcal{L}_{201}^{\text{max}} \), while a higher ratio indicates bad performance. We summarize the values of \( R \) from 200 tests Table I, from which we can see that the proposed techniques achieves a satisfactory level of performance for the majority of tested problems.

<table>
<thead>
<tr>
<th>Ratio, ( R )</th>
<th>(&lt; 1.05)</th>
<th>(&lt; 1.1)</th>
<th>(&lt; 1.2)</th>
<th>(\geq 1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Problems</td>
<td>161</td>
<td>177</td>
<td>187</td>
<td>13</td>
</tr>
</tbody>
</table>

Future directions include improving the privacy mechanism and the converse results for the multi-terminal guessing problem (10), as well as studying the multi-terminal privacy-utility tradeoff using different privacy and utility measures.

**APPENDIX A**

**PROOF OF LEMMA 2**

To show \( \mathcal{L}_{\text{max}} \geq I(X_Q; Y|X_P) \), we have

\[
\mathcal{L}_{\text{max}} = \log \sum_{y,x} P_{Y,X_P}(y,x) \max_{x_Q} \frac{P_{Y,X_P,X_Q}(y,x|x_Q)}{P_{Y,X_P}(y,x)} \\
\geq I(X_Q; Y|X_P) = I(Y; X_A|X_Q),
\]

where the inequality follows from replacing maximum over \( x_Q \) with expectation over \( P_{X_Q|Y,X_P} \) and Jensen’s inequality, and the last equality is due to source independence.

It remains to show \( \mathcal{L}_{\text{max}} \geq \Gamma_i \) for any user \( i \in [m] \) with \( G_i \subseteq Q \). For brevity, we drop the subscript \( i \) remembering that \( W, A, G \) stand for \( W_i, A_i, G_i, \) respectively. Set

\[
W_P = P \cap W, \quad Q_W = Q \cap W, \\
A_P = P \cap A, \quad A_Q = Q \cap A.
\]

Since \( G \subseteq Q \), we have \( P \cap G = \emptyset \), \( Q \cap G = G \), and thus \( P = W_P \cup A_P \), and \( Q = W_Q \cup A_Q \cup G \). We have

\[
\mathcal{L}_{\text{max}}(X_Q \rightarrow Y|X_P) \\
\geq \sum_{y,x_P} P_{Y,X_P}(y,x_P) \log \max_{x_Q} \frac{P_{Y,X_P,X_Q}(y,x_p|x_Q)}{P_{Y,X_P}(y,x_P)} (22) \\
= \sum_{y,x_P} P_{Y,X_P}(y,x_P)
\]

**where (a)** follows from the Markov chain \( X_W \rightarrow (Y, X_A) \rightarrow X_G \) as a result of the utility constraint (6);
- (24) follows from replacing maximum over \( P_{X_Q|Y,X_P} \) and Jensen’s inequality;
- (25) follows from the fact that \( X_{W_Q} \) is a deterministic function of \( Y, X_P, X_A \) according to (6);
- (26) follows from that

\[
\sum_{y,x_A} P_{Y,X_A}(y,x_A) \log \max_{x_G} \frac{P_{Y,X_A,X_G}(y,x_A|x_G)}{P_{Y,X_A}(y,x_A)} \\
\geq \sum_{y,x_A} P_{Y,X_A}(y,x_A) \log \max_{x_G} \frac{P_{X_G}(x_G) \cdot P_{Y,X_A}(y,x_A)}{P_{X_G}(x_G) \cdot P_{Y,X_A}(y,x_A)} \\
= E_{P_{Y,X_A}}[\log r(X_G \rightarrow Y|X_A)] = D_i;
\]
- (27) follows from source independence, as well as (6).

This concludes the proof.
REFERENCES


