# On the Capacity of Leaky Private Information Retrieval

Islam Samy Ravi Tandon Loukas Lazos
Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ.
Email: {*islamsamy, tandonr, llazos*}@arizona.edu

***Abstract*—Private information retrieval (PIR) allows users to retrieve data from databases without revealing the identity of that data. An extensive body of works has investigated efficient schemes to achieve computational and information-theoretic privacy. The latter guarantees that no information is revealed to the databases, irrespective of their computational power. Although information-theoretic PIR (IT-PIR) provides a strong privacy guarantee, it can be too taxing for certain applications. In this paper, we initiate the study of leaky private information retrieval (L-PIR), where a bounded amount of privacy leakage is allowed and measured through a parameter $\epsilon$. The classical IT-PIR formulation is obtained by setting $\epsilon = 0$, and for $\epsilon > 0$, we explore the opportunities offered for reducing the download cost. We derive new upper and lower bounds on the download cost of L-PIR for any arbitrary $\epsilon$, any number of messages $K$, and for $N = 2$ databases.**

## I. Introduction

Private information retrieval (PIR) introduced by Chor *et al.* [1] is a primitive allowing a user to retrieve a message from a set of databases, without revealing any information about the identity of the message. Initial methods achieved privacy assuming that the databases were computationally bounded. More recently, the problem has received significant attention under information-theoretic privacy. In [2], Sun and Jafar characterized the capacity of information-theoretic PIR (IT-PIR), defined as the maximum possible ratio of the desired data to the total amount of downloaded data, for non-colluding replicated databases. Several follow-up works investigated the PIR problem under different setups including coded databases [3], [4], colluding databases [5], and cache-aided PIR [6].

The majority of prior works in this area have adopted the IT-PIR definition, where no information is leaked about the retrieved message index. This perfect privacy requirement does not allow tuning the PIR efficiency and privacy according to the application requirements. In scenarios of frequent message retrieval, lowering the level of privacy could be desirable to improve efficiency. Ideally, one would select a desired privacy level and choose a PIR scheme that guarantees such privacy while maximizing the PIR capacity. Recently, Toledo *et al.* [7] adopted a a game-based differential privacy definition of PIR to increase the PIR capacity at the expense of a bounded loss in privacy. However, their privacy definition only captures the privacy of the submitted queries. The authors propose several schemes that hide the query identity and study their cost. Although, the query privacy can be thought of a functional equivalent to an IT-PIR in some cases, it does not satisfy the IT-PIR definition. Our work differs in that we propose a leaky PIR definition that meets the IT-PIR criteria. Moreover, we study relevant lower and upper bounds.

Specifically, we study the problem of *leaky PIR* (L-PIR) when some bounded information leakage, determined by a non-negative constant $\epsilon$, is allowed. We present an upper bound on the capacity of L-PIR for an arbitrary number of messages, $K$, and arbitrary number of databases, $N$, and for any privacy budget $\epsilon$. Furthermore, we present an L-PIR scheme for the case of $N = 2$ databases, and arbitrary number of messages. Our scheme matches the lower bound for extremal values of epsilon ($\epsilon = 0$, and $\epsilon = \infty$). The core new elements of our L-PIR scheme are as follows: (a) we devise an alternate perfect privacy scheme through a *path-based approach*, where a user query is equivalent to selecting one of several possible paths across databases that achieve decodability. (b) Since such paths could exhibit different download costs, we leverage this alternative approach and introduce leakage through the idea of biasing the path selection probabilities, and these probabilities are chosen to satisfy the privacy budget, measured by $\epsilon$. Notation: Through this work, we use the notation $[X]$ to represent the set of integers from 1 to $X$.

## II. Problem Formulation

We consider $K$ independent messages $W_1, W_2, \cdots, W_K$, of size $L$ bits. Each message is stored in $N \geq 2$ non-colluding databases (DBs). A user interested in privately retrieving $W_i$, $i \in [K]$, sends $N$ separate queries $Q_1^i, \cdots, Q_N^i$ to each of the $N$ DBs, where $Q_n^i$ denotes the query sent to the $n^{th}$ database when retrieving message $W_i$. The queries are independent of the messages such that their mutual information is zero.

$$I(W_1, \cdots, W_K; Q_1^i, \cdots, Q_N^i) = 0. \tag{1}$$

After $Q_n^i$ is received by the $n^{th}$ database, it generates the corresponding answer $A_n^i$ as a deterministic function of both $Q_n^i$ and the stored messages. Therefore,

$$H(A_n^i | Q_n^i, W_1, \cdots, W_K) = 0. \tag{2}$$

The user must be able to decode the desired message $W_i$ from all answers received from the $N$ databases. Moreover, the queries and corresponding answers must leak a bounded amount of information about the identity (index $i$) of the

This work was supported in part by NSF grants CNS 1813401, CAREER 1651491, and CNS 1715947.

desired message. Formally, the L-PIR scheme must satisfy the following correctness and privacy definitions, which are constructed under an asymptotic scenario of an arbitrarily long message length ($L$ approaching infinity).

**Correctness:** Given the generated queries, the user must be able to recover the desired message $W_i$ correctly by collecting all answers $A_1^i, \cdots, A_N^i$ from the $N$ DBs,

$$H(W_i|Q_1^i, \cdots, Q_N^i, A_1^i, \cdots, A_N^i) = 0. \tag{3}$$

**$\epsilon$-Privacy:** Given any message subset $\mathbf{W}_\Omega$, the following likelihood ratios must be bounded as follows:

$$\frac{Pr\{Q_n^i|\mathbf{W}_\Omega\}}{Pr\{Q_n^j|\mathbf{W}_\Omega\}} \leq e^\epsilon, \quad \forall i,j \in [K], \quad \forall n \in [N]. \tag{4}$$

$$\frac{Pr\{A_n^i|\mathbf{W}_\Omega\}}{Pr\{A_n^j|\mathbf{W}_\Omega\}} \leq e^\epsilon, \quad \forall i,j \in [K], \quad \forall n \in [N]. \tag{5}$$

$$\frac{Pr\{Q_n^i, A_n^i|\mathbf{W}_\Omega\}}{Pr\{Q_n^j, A_n^j|\mathbf{W}_\Omega\}} \leq e^\epsilon, \quad \forall i,j \in [K], \quad \forall n \in [N]. \tag{6}$$

where $\epsilon$ is a non-negative constant.

Intuitively, the reduction in privacy comes from biasing queries and answers towards certain messages compared with any other message. Note that by setting $\epsilon = 0$, the $\epsilon$-privacy definition becomes equivalent to perfect privacy. The relaxed privacy definition shows us that any pair of answer and query can be biased towards a desired message.

To evaluate the capacity of L-PIR, we consider the total amount of communication between the user and the DBs for retrieving the desired message. Similar to prior works [2]–[5], we adopt the Shannon theoretic formulation where the message size is assumed to be arbitrarily long and therefore, the upload cost is negligible compared with the download cost [2]. In this case, the L-PIR rate is the reciprocal of the download cost $D(\epsilon)$, which characterizes how much total information per bit the user has to download to retrieve a message with $\epsilon$ privacy.

$$D(\epsilon) = \frac{D}{L} = \frac{\sum_i^N H(A_n^i)}{H(W_i)}, \tag{7}$$

where $D$ is the average download cost over all messages. We say that the pair $(L, D)$ is achievable if there exist an L-PIR scheme that satisfies the correctness and $\epsilon$-privacy constraints. Our goal is to find the optimal download cost $D^*(\epsilon)$, such that $D^*(\epsilon) = \min\{D(\epsilon) : (L, D) \text{ is achievable}\}$.

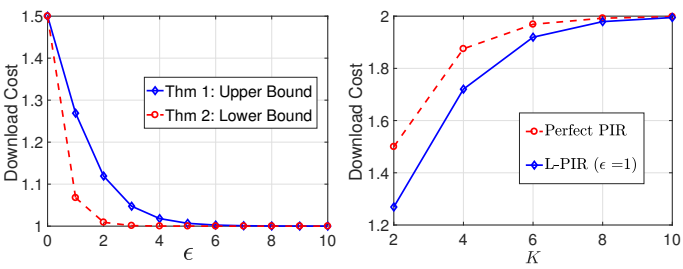
## III. Main Results and Discussion

**Theorem 1.** *For $N = 2$, the optimal download cost of the L-PIR is upper-bounded by,*

$$D^*(\epsilon) \leq 1 + \frac{2^{K-1}}{e^\epsilon + 2^{K-1} - 1}\left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{K-1}}\right). \tag{8}$$

**Theorem 2.** *The download cost of the L-PIR is lower-bounded by,*

$$D^*(\epsilon) \geq 1 + \frac{1}{Ne^{2\epsilon}} + \frac{1}{(Ne^{2\epsilon})^2} + \cdots + \frac{1}{(Ne^{2\epsilon})^{K-1}}. \tag{9}$$




(a) (b)

Fig. 1: (a) Lower and upper bounds of L-PIR for $N = 2$, and $K = 2$. The download cost of perfect privacy is obtained when $\epsilon = 0$, and (b) Upper bound of L-PIR and the download cost of perfect privacy as $K$ grows.

We make the following remarks:

- From the upper bound in (8), we can observe that L-PIR reduces the download cost by at least $^{e^\epsilon-1}/_{(e^\epsilon+2^{K-1}-1)}(1/2 + \cdots + 1/2^{K-1})$ compared to the perfect PIR. The latter has an optimal download cost of is $1 + 1/2 + \cdots + 1/2^{K-1}$ when $N = 2$, [2].
- If we ignore the privacy requirement (set $\epsilon \to \infty$ in (8) and (9)), the download cost approaches 1, which means the user needs to only download the required message.
- As $K$ grows, $^{2^{K-1}}/_{(e^\epsilon+2^{K-1}-1)}$ approaches 1, and the upper bound approaches the optimal download cost for perfect privacy. That is, the benefits of relaxing privacy by $\epsilon$ diminish with the database size (as can be seen in Fig. 1(b) for $\epsilon = 1$, and $K = 10$ messages).
- For $\epsilon = 0$, the upper and lower bounds for L-PIR match the optimal download cost with perfect privacy [2].

Figure 1(a) shows upper and lower bounds on the capacity of L-PIR for $K = 2$. One can observe the decrease in the download cost (increase in capacity) with $\epsilon$, quickly approaching the minimum possible value of 1. Fig. 1(b) compares the download cost of perfect PIR ($\epsilon = 0$) with L-PIR when $\epsilon = 1$.

### A. A Leaky PIR Example

Consider the simplest PIR setting where $N = K = 2$ and each message, $W_1 = \{a_1, \ldots, a_4\}$ and $W_2 = \{b_1, \ldots, b_4\}$, is $L = 4$ bits long. To motivate the construction of a leaky PIR, we first recall the perfect PIR scheme proposed in [2]. Figure 2 shows a retrieval structure for $W_1$. The main idea is that one can use coding and leverage side information from the other database to reduce the download cost to $3/2$. We highlight that the shown bit indices represent one possible permutation of the real indices. Thus, $W_1$ retrieval can be obtained through multiple bit structures that are selected uniformly and have equal download cost of $\frac{3}{2}$. Figure 3 shows an alternative PIR scheme in which the requested message can be downloaded via sequences of structures that give unequal download cost. In particular, when the user wants to retrieve message $W_1$, it picks one of the four possible queries/paths:

- Path $\mathcal{P}_1$:$(\emptyset, W_1)$: Send no request to DB 1 and request $W_1$ from DB 2. This is what we term as a *path*, and this path/query has a download cost of $L$ bits.

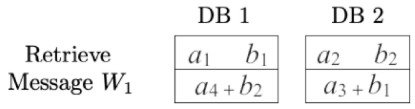



Fig. 2: A PIR scheme for $N = 2, K = 2$, and $L = 4$.

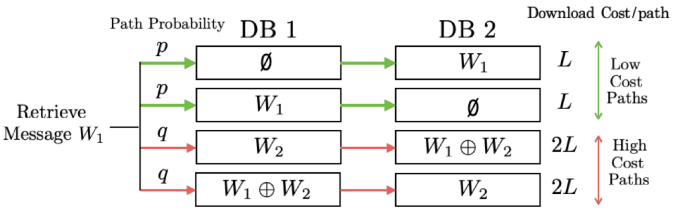



Fig. 3: L-PIR scheme for $N = 2, K = 2$, and $L = 4$.

- Path $\mathcal{P}_2$:$(W_1, \emptyset)$: Request $W_1$ from DB 1 and send no request to DB 2. This path has a download cost of $L$ bits.
- Path $\mathcal{P}_3$:$(W_2, W_1 \oplus W_2)$: Request $W_2$ from DB 1 and $W_1 \oplus W_2$ from DB 2. This path has a download cost of $2L$ bits.
- Path $\mathcal{P}_4$:$(W_1 \oplus W_2, W_2)$: Request $W_1 \oplus W_2$ from DB 1 and $W_2$ from DB 2. This path has a download cost of $2L$ bits.

Paths $\mathcal{P}_1$ and $\mathcal{P}_2$ are selected with probability $p$, whereas $q$ is the selection probability for both $\mathcal{P}_3$ and $\mathcal{P}_4$. The answer of a DB $n$ can take four different structures, $\pi_{n,1}, \ldots, \pi_{n,4}$. These structures represent the element addition of all possible subsets of $\{W_1, W_2\}$. Note that the probability of any structure $\pi_{n,j}$, $j \in [4]$, to be selected equals the selection probability of all paths coming through that structure. Also, there is one path per message that comes through each structure $\pi_{n,j}$. For example, $\pi_{1,2}$ is paired with $\pi_{2,2}$ to retrieve $W_1$, or it can be paired with $\pi_{2,3}$ for $W_2$ retrieval. Let the path selection probabilities be uniform, i.e., $p = q = \frac{1}{4}$. Thus, each structure is selected with probability $\frac{1}{4}$, irrespective of the requested message index. It is straightforward to show that this probability assignment satisfies the perfect privacy constraints. Moreover, although the cost varies per path, the uniform path selection yields an optimal average download cost of $\frac{3}{2}$. Therefore, this path-based PIR scheme is also optimal and matches the result of Sun and Jafar [2] for perfect privacy.

**Improving the download cost via path biasing.** The leaky privacy definition in (4)-(6), together with the alternative path-based scheme described above, leads us to consider schemes that bias the path selection process for retrieving desired messages. We next show that this helps reduce the average download cost for any non-zero $\epsilon$. Intuitively, if the biased paths have lower download cost (for example $L$), an overall lower cost can be achieved at the expense of some bounded loss of privacy due to the biasing. The question we pose is whether there are values $p \neq q$ that yield an average download cost less than $\frac{3}{2}$ and simultaneously satisfy the $\epsilon$-privacy definitions in (4)-(6). To meet these definitions, the possible structures to each query must satisfy:

$$e^{-\epsilon} \leq \frac{\Pr(\pi_{n,j}|i=1)}{\Pr(\pi_{n,j}|i=2)} < e^{\epsilon}, \quad \forall n,\ j. \tag{10}$$

where $P(\pi_{n,j}|i=k)$ is the probability of retrieving structure $\pi_{n,j}$ when the desired message is $k$. This satisfies (4) directly, whereas it satisfies (5) and (6) as each possible answer of the selected structure is obtained with probability $c_{n,j} \cdot P(\pi_{n,j}|i=k)$, where $c_{n,j}$ is a constant independent of the requested message. For example, we have $c_{n,j} = 2^{-L}$ for the structures $W_1$, $W_2$, and $W_1 \oplus W_2$. Based on the scheme in Fig. 3, there are two cases for each structure $\pi_{n,j}$: (i) $\pi_{n,j}$ is used to recover $W_1$ and $W_2$ with the same probability either $p$ or $q$, then $^{P(\pi_{n,j}|i=1)}/_{P(\pi_{n,j}|i=2)} = 1$, which intuitively satisfies (4)-(6). (ii) $\pi_{n,j}$ is selected with different probability $p$ and $q$ to retrieve $W_1$ and $W_2$, respectively, and vice versa. Then, $p$ and $q$ must satisfy

$$\frac{\Pr(\pi_{n,j}|i=1)}{\Pr(\pi_{n,j}|i=2)} = \frac{p}{q} \leq e^{\epsilon}. \tag{11}$$

Invoking the fact that the sum of path probabilities must equal one ($2p + 2q = 1$) and substituting $q = 0.5 - p$, we can rewrite (11) as

$$p \leq \frac{e^{\epsilon}}{2(1+e^{\epsilon})}. \tag{12}$$

We thus pick $p$ that satisfies (12) with equality, and then select $q$ as $q = 0.5 - p$, as a valid choice of path selection probabilities which satisfy the $\epsilon$-privacy constraints.

**Computing the download cost $D(\epsilon)$:** Since our scheme is symmetric, the same download cost is obtained for the retrieval of message $W_1$ or message $W_2$. Then, the average download cost can be written as

$$D(\epsilon) = \frac{\sum_{j=1}^{4} \Pr\{\mathcal{P} = \mathcal{P}_j\} \cdot D_{\mathcal{P}_j}}{L}, \tag{13}$$

where $\Pr\{\mathcal{P} = \mathcal{P}_j\} \in \{p, q\}$ is the probability that path $\mathcal{P}_j$ is chosen and $D_{\mathcal{P}_j}$ is the cost of path $\mathcal{P}_j$. From Fig. 3, we know that $D_{\mathcal{P}_1} = D_{\mathcal{P}_2} = L$, and $D_{\mathcal{P}_3} = D_{\mathcal{P}_4} = 2L$. Hence, $D(\epsilon)$ equals

$$D(\epsilon) = \frac{2 \times p \times L + 2 \times q \times (2L)}{L} = 2 - 2p \geq 2 - \frac{e^{\epsilon}}{(1+e^{\epsilon})}, \tag{14}$$

where the last inequality comes from (12). The achievable download cost of our L-PIR scheme can be rewritten as

$$D(\epsilon) = \frac{3}{2} - \frac{e^{\epsilon} - 1}{2(e^{\epsilon} + 1)}, \tag{15}$$

which is lower than $\frac{3}{2}$, the optimal download cost under perfect privacy, for any $\epsilon$.

## IV. Proof of Theorem 1

In this section, we construct an L-PIR scheme for general $K$ and $N = 2$. The download cost of this scheme gives the upper bound in Theorem 1. Assume there are $K \geq 2$ messages, $W_1, \ldots, W_K$. For each DB $n \in \{1, 2\}$, there are $2^K$ different structures, $\pi_{n,1}, \ldots, \pi_{n,2^K}$ that represent the addition of mes-

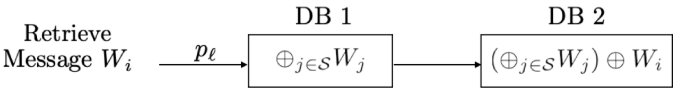



Fig. 4: One path of the L-PIR scheme that retrieves $W_i$.

sages belonging to any possible subset $S$ of the $K$ messages. To decode the desired message $W_i$, the two structures that form each path are selected such that their $\oplus$ addition gives the requested message. For instance, $W_1$ can be recovered from paths $(W_1, \emptyset)$, $(W_2, W_1 \oplus W_2)$, $(W_3, W_1 \oplus W_3)$, etc. A general form for a path selected with probability $p_\ell$ is shown in Fig. 4. This construction yields: (i) two paths with a cost of $L$ bits which include the $\emptyset$ structure. (ii) $(2^K - 2)$ other paths with a cost of $2L$ bits. Without loss of generality, we assign probabilities $p$ and $q$ to paths with cost $L$ and $2L$, respectively, such that $2p+(2^K-2)q = 1$. Due to symmetry, these paths are assigned the same probability. Assigning different probabilities does not improve the download cost or the privacy. Figure 5 shows the L-PIR scheme for $K = 3$.

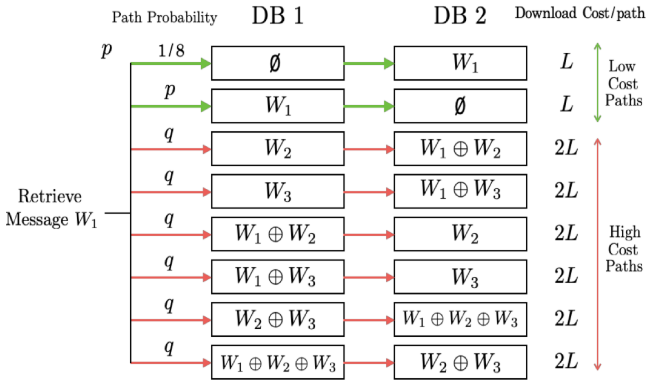



Fig. 5: An L-PIR scheme for $N = 2$, and $K = 3$.

Figure 4 shows that for any structure, there is a path that passes through it to recover any desired message. This is crucial to satisfy $\epsilon$-privacy because accessing a structure does not eliminate any of message possibilities. In total, there are $K$ paths passing through each structure. The following lemma generalizes the condition in (12) for satisfying $\epsilon$-privacy to any $K$. It provides an upper bound on the path biasing probability that does not violate the $\epsilon$-privacy.

**Lemma 1.** *To preserve $\epsilon$-privacy*

$$p \leq \frac{e^\epsilon}{2(e^\epsilon + 2^{K-1} - 1)}. \tag{16}$$

*Proof.* Similar to the privacy analysis in (10) and (11) for the L-PIR scheme in Fig. 2, we only need to guarantee that $p \leq q\ e^\epsilon$. Substituting the inequality in $2p + (2^K - 2)q = 1$, we get

$$2p(1 + e^{-\epsilon}(2^{K-1} - 1)) \leq 2p + (2^K - 2)q = 1. \tag{17}$$

This gives

$$p \leq \frac{1}{2(1 + e^{-\epsilon}(2^{K-1} - 1))} = \frac{e^\epsilon}{2(e^\epsilon + 2^{K-1} - 1)}. \tag{18}$$

This proves the lemma. □

Given that all messages are requested equiprobably, the download cost can be written as,

$$\begin{aligned} D(\epsilon) &= \frac{\sum_{s=1}^{2^K} Pr\{\mathcal{P} = \mathcal{P}_s\} \cdot D_{\mathcal{P}_s}}{L} \\ &= \frac{2 \times p \times L + (2^K - 2) \times q \times 2L}{L} = 2p + 2(2^K - 2)q \\ &\overset{(a)}{=} 2p + 2(1 - 2p) = 2 - 2p \\ &\overset{(b)}{\geq} 2 - \frac{e^\epsilon}{e^\epsilon + 2^{K-1} - 1} = 1 + \frac{2^{K-1} - 1}{e^\epsilon + 2^K - 1} \\ &= 1 + \frac{2^{K-1}}{e^\epsilon + 2^K - 1} \cdot \frac{2^{K-1} - 1}{2^{K-1}} \\ &= 1 + \frac{2^{K-1}}{e^\epsilon + 2^{K-1} - 1}(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^{K-1}}), \end{aligned} \tag{19}$$

where $(a)$ comes from the equality $2p + (2^K - 2)q = 1$, and $(b)$ is due to Lemma 1. This gives the upper bound in (8) on the download cost of L-PIR for arbitrary $K$, and $N = 2$, and completes the proof of Theorem 1.

## V. Proof of Theorem 2

In this Section, we prove the lower bound stated in Theorem 2. For $N \geq 2$ and without loss of generality, assume the requested message is $W_1$, then $D$ can be lower bounded as

$$\begin{aligned} D &= H(A_1^1) + \cdots + H(A_N^1) \geq H(A_{[1:N]}^1) \\ &\geq H(A_{[1:N]}^1 | Q_{[1:N]}^1) \\ &= H(W_1, A_{[1:N]}^1 | Q_{[1:N]}^1) - H(W_1 | A_{[1:N]}^1, Q_{[1:N]}^1) \\ &\overset{(a)}{=} H(W_1, A_{[1:N]}^1 | Q_{[1:N]}^1) = L + H(A_{[1:N]}^1 | Q_{[1:N]}^1, W_1) \\ &\overset{(b)}{=} L + I(W_{[2:K]}; A_{[1:N]}^1 | Q_{[1:N]}^1, W_1) \\ &\overset{(c)}{=} L + I(W_{[2:K]}; A_{[1:N]}^1, Q_{[1:N]}^1 | W_1) \\ &\geq L + I(W_{[2:K]}; A_1^1, Q_1^1 | W_1) \\ &\overset{(d)}{=} L + H(A_1^1 | W_1, Q_1^1), \end{aligned}$$

where $(a)$ follows from the correctness constraint in (3), $(b)$ follows from (2), $(c)$ follows from (1), and $(d)$ follows again from (1) and (2).

Under perfect privacy condition ($\epsilon = 0$), the proof is continued by setting $H(A_1^1 | W_1, Q_1^1) = H(A_1^j | W_1, Q_1^j)$, for $j \in [K]$. However, this does not hold under $\epsilon$-privacy, whenever $\epsilon > 0$. Thus, we obtain a bound on $H(A_1^1 | W_1, Q_1^1)$ by invoking the following lemma.

**Lemma 2.** *Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables, over $[G]$, satisfying the following condition,*

$$e^{-\epsilon} \leq \frac{\Pr(\mathbf{X} = g)}{\Pr(\mathbf{Y} = g)} \leq e^\epsilon, \quad \forall g \in [G], \tag{20}$$

*then we get*

$$H(\mathbf{X}) \geq e^{-\epsilon} H(Y). \tag{21}$$

*Proof.* Let $x_g$ and $y_g$ be the probabilities $\Pr(\mathbf{X} = g)$ and $\Pr(\mathbf{Y} = g)$, respectively. As $|\mathbf{X}| = |\mathbf{Y}| = G$, where $|.|$ represents the set cardinality, we can write $H(\mathbf{X})$, $H(\mathbf{Y})$ as,

$$H(\mathbf{X}) = \log_2(G) - D(\mathbf{X}||\mathbf{U}), \tag{22}$$

$$H(\mathbf{Y}) = \log_2(G) - D(\mathbf{Y}||\mathbf{U}), \tag{23}$$

where $D(\cdot||\cdot)$ is the KL divergence and $\mathbf{U}$ is a random variable, uniformly distributed over $[G]$. This yields

$$H(\mathbf{X}) - H(\mathbf{Y}) = D(\mathbf{Y}||\mathbf{U}) - D(\mathbf{X}||\mathbf{U}). \tag{24}$$

From (20) and $\forall g \in [G]$, we have $e^{-\epsilon}y_g \leq x_g \leq e^{\epsilon}y_g$, then $x_g = e^{-\epsilon}y_g + \delta_g$, where $0 \leq \delta_g \leq e^{\epsilon}y_g - x_g$. From this relation, we get

$$\sum \delta_g = \sum x_g - \sum e^{-\epsilon}y_g = 1 - e^{-\epsilon}. \tag{25}$$

We write $x_g$ equivalently as

$$x_g = e^{-\epsilon}y_g + (1 - e^{-\epsilon})\frac{\delta_g}{1 - e^{-\epsilon}}. \tag{26}$$

Note that both $y_g$ and $\frac{\delta_g}{1-e^{-\epsilon}}$ are valid probability mass functions (p.m.f.s) with sums equal one. Let $\mathbf{T}$ be a random variable over $[G]$ such that $\Pr\{\mathbf{T} = g\} = \frac{\delta_g}{1-e^{-\epsilon}}$. Next, we utilize the convexity of the KL divergence and the relation in (26) to get

$$D(\mathbf{X}||\mathbf{U}) \leq e^{-\epsilon}D(\mathbf{Y}||\mathbf{U}) + (1 - e^{-\epsilon})D(\mathbf{T}||\mathbf{U}). \tag{27}$$

The last inequality and (24) yield

$$\begin{aligned} &H(\mathbf{X}) - H(\mathbf{Y}) \geq (1 - e^{-\epsilon})(D(\mathbf{Y}||\mathbf{U}) - D(\mathbf{T}||\mathbf{U})) \\ &= (1 - e^{-\epsilon})(D(\mathbf{Y}||\mathbf{U}) - \log_2(G) + \log_2(G) - D(\mathbf{T}||\mathbf{U})) \\ &= (1 - e^{-\epsilon})(H(\mathbf{T}) - H(\mathbf{Y})). \end{aligned} \tag{28}$$

Then, we have

$$H(\mathbf{X}) = e^{-\epsilon}H(\mathbf{Y}) + (1 - e^{-\epsilon})H(\mathbf{T}) \geq e^{-\epsilon}H(\mathbf{Y}). \tag{29}$$

□

To continue the converse proof, we utilize (4) and (6) in bounding the following ratio $\forall i, j \in [K]$,

$$\frac{\Pr\{A_n^i|Q_n^i, \mathbf{W}_\Omega\}}{\Pr\{A_n^j|Q_n^j, \mathbf{W}_\Omega\}} = \frac{\Pr\{Q_n^i, A_n^i|\mathbf{W}_\Omega\} \cdot \Pr\{Q_n^j|\mathbf{W}_\Omega\}}{\Pr\{Q_n^i|\mathbf{W}_\Omega\} \cdot \Pr\{Q_n^j, A_n^j|\mathbf{W}_\Omega\}} \leq e^{2\epsilon}. \tag{30}$$

Replacing $\mathbf{X}$ and $\mathbf{Y}$ in Lemma 2 with the conditional random variables of answers given queries and messages, we get

$$H(A_n^i|Q_n^i, \mathbf{W}_\Omega) \geq e^{-2\epsilon}H(A_n^j|Q_n^j, \mathbf{W}_\Omega), \quad \forall i, j \in [K], \forall n \tag{31}$$

Given the previous inequality, $D$ can be bounded as

$$D \geq L + e^{-2\epsilon}H(A_n^2|W_1, Q_n^2), \quad \forall n. \tag{32}$$

The addition of (32) over all possible $n$'s gives us the following

$$N \times D \geq N \times L + e^{-2\epsilon}\sum_{n=1}^{N} H(A_n^2|W_1, Q_n^2). \tag{33}$$

Dividing by $N$,

$$\begin{aligned} D &\geq L + e^{-2\epsilon} \cdot \frac{1}{N}\sum_{n=1}^{N} H(A_n^2|W_1, Q_n^2) \\ &\geq L + e^{-2\epsilon} \cdot \frac{1}{N}H(A_1^2, \ldots, A_N^2|W_1, Q_1^2, \ldots, Q_N^2) \\ &= L + e^{-2\epsilon} \cdot \frac{1}{N}H(W_2, A_1^2, \ldots, A_N^2|W_1, Q_1^2, \ldots, Q_N^2) \\ &\quad - H(W_2|A_1^2, \ldots, A_N^2, W_1, Q_1^2, \ldots, Q_N^2) \\ &\overset{(a)}{=} L + e^{-2\epsilon} \cdot \frac{1}{N}H(W_2, A_1^2, \ldots, A_N^2|W_1, Q_1^2, \ldots, Q_N^2) \\ &\overset{(b)}{=} L + e^{-2\epsilon} \cdot \frac{1}{N}H(W_2) \\ &\quad + \frac{1}{N}H(A_1^2, \ldots, A_N^2|W_1, W_2, Q_1^2, \ldots, Q_N^2) \\ &= L + e^{-2\epsilon} \cdot \frac{L}{N} + \frac{1}{N}H(A_1^2, \ldots, A_N^2|W_1, W_2, Q_1^2, \ldots, Q_N^2), \end{aligned} \tag{34}$$

where $(a)$ comes from (3), and $(b)$ is due to the independence of the messages and queries. Following the same iterative process used in [2] and applying Lemma 2 yields a lower bound on the download cost equal to:

$$D^*(\epsilon) \geq 1 + \frac{1}{Ne^{2\epsilon}} + \cdots + \frac{1}{(Ne^{2\epsilon})^{K-1}}. \tag{35}$$

This proves the lower bound in Theorem 2.

## VI. CONCLUSION

In this paper, we studied the PIR problem under a relaxed definition of privacy that trades off some bounded amount of information leakage controlled by a non-negative parameter $\epsilon$ for lower download cost. We explored the opportunities offered by this relaxation and proposed an L-PIR scheme that lowers the download cost for any $\epsilon$ and any arbitrary number of messages. We derived new upper and lower bounds on the download cost of L-PIR for $N = 2$ databases. Closing the gap between the two bounds was left as an interesting open problem for further research.

## REFERENCES


[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*. IEEE, 1995, pp. 41–50.

[2] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, 2017.

[3] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, 2018.

[4] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from mds coded data in distributed storage systems," *IEEE Transactions on Information Theory*, 2018.

[5] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2361–2370, 2018.

[6] R. Tandon, "The capacity of cache aided private information retrieval," in *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*. IEEE, 2017, pp. 1078–1082.

[7] R. R. Toledo, G. Danezis, and I. Goldberg, "Lower-cost-private information retrieval," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 4, pp. 184–201, 2016.