

Improving Achievability of Cache-Aided Private Variable-Length Coding with Zero Leakage

Amirreza Zamani, Mikael Skoglund

Division of Information Science and Engineering, KTH Royal Institute of Technology
Email: amizam@kth.se, skoglund@kth.se

Abstract—A statistical cache-aided compression problem with a privacy constraint is studied, where a server has access to a database of N files, $(Y_1, \dots, Y_N)$, each of size F bits and is linked through a shared channel to K users, where each has access to a local cache memory of size MF bits. During the placement phase, the server fills the users' caches without prior knowledge of their demands, while the delivery phase takes place after the users send their demands to the server. We assume that each file in database Y_i is arbitrarily correlated with a private attribute X , and an adversary is assumed to have access to the shared channel. The users and the server have access to a shared key W . The goal is to design the cache contents and the delivered message $\mathcal{C}$ such that the average length of $\mathcal{C}$ is minimized, while satisfying: i. The response $\mathcal{C}$ does not reveal any information about X , i.e., $I(X; \mathcal{C}) = 0$; ii. User i can decode its demand, Y_{d_i} , by using the shared key W , $\mathcal{C}$, and its local cache Z_i . In a previous work, we have proposed a variable-length coding scheme that combines privacy-aware compression with coded caching techniques. In this paper, we propose a new achievability scheme using minimum entropy coupling concept and a greedy entropy-based algorithm. We show that the proposed scheme improves the previous results. Moreover, considering two special cases we improve the obtained bounds using the common information concept.

Index Terms—Cache-aided networks, private variable-length coding, minimum entropy functional representation.

I. INTRODUCTION

We consider the same scenario as in [1] illustrated in Fig. 1, in which a server has access to a database consisting of N files $Y_1, \dots, Y_N$. Each file, of size F bits, is sampled from the joint distribution $P_{XY_1 \dots Y_N}$, where X denotes the private attribute. We assume that the server knows the realization of the private variable X as well. The server is linked to K users over a shared channel, where user i has access to a local cache memory of size MF bits. Furthermore, we assume that the server and the users have access to a shared key denoted by W , of size T . Similar to [2], the system works in two phases: the placement and delivery phases. In the placement phase, the server fills the local caches using the database. After the placement phase, let Z_k denote the content of the local cache of user k , $k \in [K] \triangleq \{1, \dots, K\}$. In the delivery phase, first the users send their demands to the server, where $d_k \in [N]$ denotes the demand of user k . The server sends a response, denoted by $\mathcal{C}$, over the shared channel to satisfy all the demands, simultaneously. We assume that an adversary has access to the shared link as well, and uses $\mathcal{C}$ to extract information about X . However, the adversary does not have access to the local cache contents or the shared secret key. As argued in [1], since

the files in the database are all correlated with the private latent variable X , the coded caching and delivery techniques introduced in [2] do not satisfy the privacy requirement. Similar to [1], the goal of the cache-aided private delivery problem is to find a response $\mathcal{C}$ with minimum possible average length that satisfies a certain privacy constraint and the zero-error decodability constraint of users. Similar to [1], we consider the worst case demand combinations $d = (d_1, \dots, d_K)$ to construct $\mathcal{C}$, and the expectation is taken over the randomness in the database. We consider a perfect privacy constraint, i.e., we require $I(\mathcal{C}; X) = 0$. Let $\hat{Y}_{d_k}$ denote the decoded message of user k using W , $\mathcal{C}$, and Z_k . User k should be able to decode Y_{d_k} reliably, i.e., $\mathbb{P}\{\hat{Y}_{d_k} \neq Y_{d_k}\} = 0, \forall k \in [K]$. In [1], we have utilized techniques used in privacy mechanisms, data compression, and cache design and coded delivery problems, and combine them to build such a code. In particular, we have used data compression techniques employed in [3] and caching design techniques in [2]. In this work, to build $\mathcal{C}$, we use the minimum entropy coupling concept and a greedy entropy-based algorithm that are studied in [4]–[6]. We compare the new proposed scheme with the existing one in [1] and show that the proposed achievable scheme can significantly improve the previous result.

The privacy mechanism, caching and compression design problems have received increased attention in recent years [1]–[22]. Specifically, in [2], a cache-aided network consisting of a single server connected to several users equipped with local caches over a shared error-free link, is considered, and the rate-memory trade-off has been characterized within a constant gap. This bound has been improved for different scenarios in [7], [8]. In particular, the exact rate-memory trade-off for uncoded placement has been characterized in [9]. A cache-aided coded content delivery problem is studied in [10] where users have different distortion requirements.

Considering the compression problem with privacy constraints, a notion of perfect secrecy is introduced in [11] by Shannon, where the private and public data are statistically independent. In the Shannon cipher system, one of M messages is sent over a channel wiretapped by an eavesdropper, and it is shown that perfect secrecy is achievable if and only if the shared secret key length is at least M [11].

In [12], the problem of privacy-utility trade-off considering mutual information both as measures of utility and privacy is studied. It is shown that under the perfect privacy assumption, the optimal privacy mechanism problem can be obtained as the solution of a linear program. In [13],

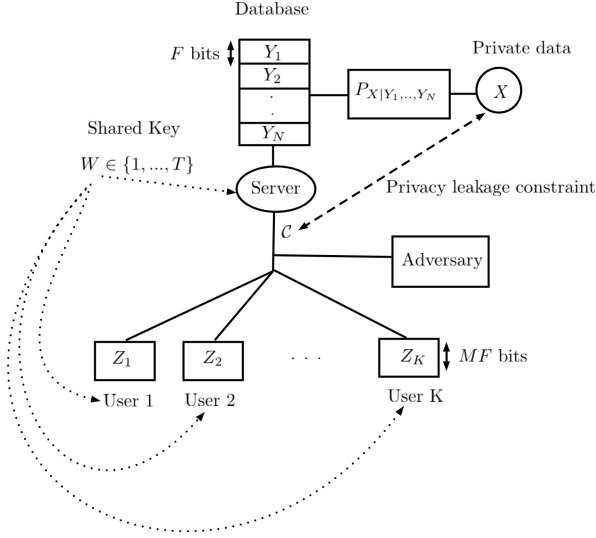


Fig. 1. In this work a server wants to send a response over a shared link to satisfy users' demands, where the database is correlated with the private data. In the delivery phase, we hide the information about X using one-time-pad coding and send the rest of response using a greedy entropy-based algorithm proposed in [4].

the work [12] is generalized by relaxing the perfect privacy assumption allowing some small bounded leakage. This result is generalized to a non-invertible leakage matrix in [14].

In [3], the *secrecy by design* concept has been introduced and is applied to privacy mechanism and lossless compression design problems. For the privacy problem, bounds on privacy-utility trade-off are derived by using the Functional Representation Lemma. These results are derived under the perfect privacy assumption. The privacy problems considered in [3] are generalized in [17] by relaxing the perfect privacy constraint. Moreover, the problems of fixed length and variable length compression with a certain privacy constraint have been studied in [3] and upper and lower bounds on the average length of encoded message have been derived. A similar approach has been used in [15], where in a lossless compression problem the relations between shared key, secrecy, and compression considering perfect secrecy, maximal leakage, secrecy by design, mutual information leakage and local differential privacy have been studied. In [1], we have studied a cache-aided coded content delivery problem with a certain privacy constraint which is closely related to [2] and [3]. We have provided upper and lower bounds on the average length of the server's response $\mathcal{C}$. Specifically, we have generalized the problem considered in [2] by considering correlation between the database and the private attribute, and we have used variable-length lossless compression techniques as in [3] to build the response $\mathcal{C}$ in the presence of an adversary. As argued in [1], for the achievability scheme we use a two-part code construction, which is based on the Functional Representation Lemma (FRL) and one-time pad coding to hide the information about X and reconstruct the demanded files at user side.

The main contribution of this work is to improve the achievability scheme obtained in [1]. Here, the achievability scheme corresponds to the upper bound on the average

length of the response $\mathcal{C}$. To do so we use a two-part code construction, which is based on the greedy entropy-based algorithm proposed in [4] and one-time pad coding to hide the information about X and reconstruct the demanded files at user side. Finally, considering two special cases, we improve the obtained bounds by using the common information concept. We show that when the size of the private data is large, the obtained bounds can be significantly improved using less shared key size.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this work, the i -th file in the database is denoted by Y_i , which is of size F bits, i.e., $\mathcal{Y}_i \in \{1, \dots, 2^F\}$ and $|\mathcal{Y}_i| = 2^F$. Similar to [1] we assume that $N \geq K$; however, the results can be generalized to other cases as well. The discrete random variable (RV) X defined on the finite alphabet $\mathcal{X}$ describes the private attribute and is arbitrarily correlated with the files in the database $Y = (Y_1, \dots, Y_N)$ where $|\mathcal{Y}| = |\mathcal{Y}_1| \times \dots \times |\mathcal{Y}_N| = (2^F)^N$ and $\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_N$. The joint distribution of the private data and the database is shown by $P_{XY_1 \dots Y_N}$ and marginal distributions of X and Y_i by vectors P_X and P_{Y_i} defined on $\mathbb{R}^{|\mathcal{X}|}$ and $\mathbb{R}^{2^F}$ given by the row and column sums of $P_{XY_1 \dots Y_N}$. The relation between X and Y is given by the matrix $P_{Y_1 \dots Y_N | X}$ defined on $\mathbb{R}^{(2^F)^N \times |\mathcal{X}|}$. Each user is equipped with a local cache of size MF bits. The shared secret key is denoted by the discrete RV W defined on $\{1, \dots, T\}$, and is assumed to be known by the server and the users, but not the adversary. Furthermore, we assume that W is uniformly distributed and is independent of (X, Y) . Similar to [2], we have K caching functions to be used during the placement phase:

$$\theta_k : [|\mathcal{Y}|] \rightarrow [2^{\lfloor FM \rfloor}], \forall k \in [K], \quad (1)$$

such that

$$Z_k = \theta_k(Y_1, \dots, Y_N), \forall k \in [K]. \quad (2)$$

Let the vector $(Y_{d_1}, \dots, Y_{d_K})$ denote the demands sent by the users at the beginning of the delivery phase, where $(d_1, \dots, d_K) \in [N]^K$. A variable-length prefix-free code with a shared secret key of size T consists of mappings:

$$(\text{encoder}) \mathcal{C} : [|\mathcal{Y}|] \times [T] \times [N]^K \rightarrow \{0, 1\}^*$$

$$(\text{decoder}) \mathcal{D}_k : \{0, 1\}^* \times [T] \times [2^{\lfloor FM \rfloor}] \times [N]^K \rightarrow 2^F, \forall k \in [K].$$

The output of the encoder $\mathcal{C}(Y, W, d_1, \dots, d_K)$ is the codeword the server sends over the shared link in order to satisfy the demands of the users $(Y_{d_1}, \dots, Y_{d_K})$. At the user side, user k employs the decoding function $\mathcal{D}_k$ to recover its demand Y_{d_k} , i.e., $\hat{Y}_{d_k} = \mathcal{D}_k(Z_k, W, \mathcal{C}(Y, W, d_1, \dots, d_K), d_1, \dots, d_K)$. The variable-length code $(\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K)$ is lossless if for all $k \in [K]$ we have

$$\mathbb{P}(\mathcal{D}_k(\mathcal{C}(Y, W, d_1, \dots, d_K), W, Z_k, d_1, \dots, d_K) = Y_{d_k}) = 1. \quad (3)$$

In the following, we define perfectly private codes. The code $(\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K)$ is *perfectly private* if

$$I(\mathcal{C}(Y, W, d_1, \dots, d_K); X) = 0. \quad (4)$$

Let ξ be the support of $\mathcal{C}(Y, W, d_1, \dots, d_K)$, where $\xi \subseteq \{0, 1\}^*$. For any $c \in \xi$, let $\mathbb{L}(c)$ be the length of the codeword. The lossless code $(\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K)$ is $(\alpha, T, d_1, \dots, d_K)$ -variable-length if

$$\mathbb{E}(\mathbb{L}(\mathcal{C}(Y, w, d_1, \dots, d_K))) \leq \alpha, \quad \forall w \in [T] \text{ and } \forall d_1, \dots, d_K, \quad (5)$$

and $(\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K)$ satisfies (3). Finally, let us define the set $\mathcal{H}(\alpha, T, d_1, \dots, d_K)$ as follows:
 $\mathcal{H}(\alpha, T, d_1, \dots, d_K) \triangleq \{(\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K) : (\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K) \text{ is perfectly-private and } (\alpha, T, d_1, \dots, d_K)\text{-variable-length}\}$. The cache-aided private compression design problems can be stated as follows

$$\mathbb{L}(P_{XY_1 \cdot Y_N}, T) = \inf_{d_1, \dots, d_K, (\mathcal{C}, \mathcal{D}_1, \dots, \mathcal{D}_K) \in \mathcal{H}(\alpha, T, d_1, \dots, d_K)} \alpha. \quad (6)$$

Remark 1. By letting $M = 0$, $N = 1$, and $K = 1$, (6) leads to the privacy-compression rate trade-off studied in [3] and [17], where upper and lower bounds have been derived.

Remark 2. Similar to [1], to design a code, we consider the worst case demand combinations. This follows since (5) must hold for all possible combinations of the demands.

III. MAIN RESULTS

In this section, we derive upper bounds on $\mathbb{L}(P_{XY_1 \cdot Y_N}, T)$ defined in (6). Providing new converse bounds is challenging and due to the limited space is left as a future work. Similar to [1], we employ the two-part code construction, which has been used in [3]. We first encode the private data X using a one-time pad [15, Lemma 1], then encode the RV found by the achievable scheme in [2, Theorem 1] by using the greedy entropy based algorithm in [4]. To do this, let us first recall the important results regarding upper and lower bounds on the minimum entropy coupling as obtained in [4]–[6]. Similar to [6], for a given joint distribution P_{XY} let the minimum entropy of functional representation of (X, Y) be defined as

$$H^*(P_{XY}) = \inf_{H(Y|X, U)=0, I(X;U)=0} H(U). \quad (7)$$

Remark 3. By letting $\alpha = 1$ in [6, Definition 1], it leads to the same problem in (7).

Remark 4. As shown in [6, Lemma 1], the minimum entropy functional representation and the minimum entropy coupling are related functions. More specifically, $H^*(P_{XY})$ equals to the minimum entropy coupling of the set of PMFs $\{P_{Y|X=x_1}, \dots, P_{Y|X=x_n}\}$, where $\mathcal{X} = \{x_1, \dots, x_n\}$.

Let $\mathcal{G}_S$ be the output of the greedy entropy-based algorithm which is proposed in [4, Section 3], i.e., $H^*(P_{XY}) \leq H(\mathcal{G}_S)$. More specifically, the corresponding algorithm aims to solve (7) but does not achieve the optimal solution in general. Next, we recall a result obtained in [5] which shows that $\mathcal{G}_S$ is optimal within $\frac{\log e}{e} \approx 0.53$ bits for $|\mathcal{X}| = 2$ and is optimal within $\frac{1+\log e}{2} \approx 1.22$ bits for $|\mathcal{X}| > 2$. Let U^* achieve the optimal solution of (7), i.e., $H(U^*) = H^*(P_{XY})$.

Theorem 1. [5, Th. 3.4, Th. 4.1, Th. 4.2] Let $(X, Y) \sim P_{XY}$ and have finite alphabets. When X is binary, we have

$$H(\text{Profile}) \leq H(U^*) \leq H(\mathcal{G}_S) \leq H(\text{Profile}) + \frac{\log e}{e} \approx H(\text{Profile}) + 0.53. \quad (8)$$

Moreover, for $|\mathcal{X}| > 2$ we have

$$\begin{aligned} H(\text{Profile}) &\leq H(U^*) \leq H(\mathcal{G}_S) \\ &\leq H(\text{Profile}) + \frac{1 + \log e}{2} \\ &\approx H(\text{Profile}) + 1.22. \end{aligned} \quad (9)$$

Here, *Profile* corresponds to the profile method proposed in [5, Section 3].

Next, we present results on lower bounds on $H^*(P_{XY})$ obtained in a parallel work [6]. The lower bounds are obtained by using information spectrum and majorization concepts.

Theorem 2. [6, Corollary 2, Th. 2] Let $(X, Y) \sim P_{XY}$ and have finite alphabets. By letting $\alpha = 1$ in [6, Corollary 2, Th. 2], we have

$$H(\wedge_{x \in \mathcal{X}} P_{Y|x}) \leq H(Q^*) \leq H(U^*). \quad (10)$$

where $\wedge$ corresponds to the greatest lower bound with respect to majorization and Q^* is defined in [6, Lemma 3].

Remark 5. In contrast with [5], the lower bounds in [6] are obtained considering Rényi entropy in (7). In this paper, we consider Shannon entropy which is a special case of Rényi entropy.

Remark 6. As argued in [6, Remark 1], for $\alpha = 1$ the (largest) lower bounds obtained in [5] and [6] match. Thus, using Theorem 1, for binary X we have

$$H(Q^*) \leq H(U^*) \leq H(\mathcal{G}_S) \leq H(Q^*) + \frac{\log e}{e}, \quad (11)$$

and for $|\mathcal{X}| > 2$,

$$H(Q^*) \leq H(U^*) \leq H(\mathcal{G}_S) \leq H(Q^*) + \frac{1 + \log e}{2}. \quad (12)$$

Moreover, in some cases the lower bound $H(Q^*)$ is tight, e.g., see [6, Example 2].

As discussed in [6], Q^* can be obtained by a greedy construction. To do so, let $Q^* = (q_1^*, q_2^*, \dots)$ with $q_1^* \geq q_2^* \geq \dots$, where $q_i = P(Q = q_i)$. Let $P_{Y|X}$ be a matrix with columns $P_{Y|X=x}$ where each is a conditional distribution vector and assume that each column has a descending order (re-order each column). Let $q_1^* = \min_{x \in \mathcal{X}} \{\max_{y \in \mathcal{Y}} P_{Y|X}(y|x)\}$. In other words, we choose the smallest number in the first row of the matrix $P_{Y|X}$. Next, we subtract q_1^* from the first row and reorder each column and update the matrix. We then choose the smallest number from the first row of the updated matrix and represent it by q_2^* . We continue this procedure until the summation of q_i^* reaches one. To see an example refer to [6, Example 1].

Next, we present a summary of the achievable scheme proposed in [2, Theorem 1]. We first consider a cache size

$M \in \{\frac{N}{K}, \frac{2N}{K}, \dots, N\}$ and define $p \triangleq \frac{MK}{N}$, which is an integer. In the placement phase, each file, e.g., Y_n , $n \in [N]$, is split into $\binom{K}{p}$ equal size subfiles and labeled as follows

$$Y_n = (Y_{n,\Omega} : \Omega \subset [K], |\Omega| = p). \quad (13)$$

For all n , the server places $Y_{n,\Omega}$ in the cache of user k if $k \in \Omega$. As argued in [2], each user caches total of $N \binom{K-1}{p-1} \frac{F}{\binom{K}{p}} = MF$ bits, which satisfies the memory constraint with equality. In the delivery phase, the server sends the following message over the shared link

$$\mathcal{C}' \triangleq (C_{\gamma_1}, \dots, C_{\gamma_L}), \quad (14)$$

where $L = \binom{K}{p+1}$ and for any $i \in \{1, \dots, L\}$, γ_i is the i -th subset of $[K]$ with cardinality $|\gamma_i| = p + 1$, furthermore,

$$C_{\gamma_i} \triangleq \bigoplus_{j \in \gamma_i} Y_{d_j, \gamma_i \setminus \{j\}}, \quad (15)$$

where $\oplus$ denotes bitwise XOR function. Note that $Y_{d_j, \gamma_i \setminus \{j\}}$ is the subfile that is not cached by user j , but is requested by it. In other words, considering each subset of $[K]$ with cardinality $|\gamma_i| = p + 1$, using the message C_{γ_i} , each user, e.g., user j , is able to decode the subfile $Y_{d_j, \gamma_i \setminus \{j\}}$ that is not cached by it. Considering all the messages in (14) user i can decode file Y_{d_i} completely using $\mathcal{C}'$ and its local cache content Z_i . Note that each subfile C_{γ_i} has size $\frac{F}{\binom{K}{p}}$ bits. As pointed out in [2], for other values of M we can use the memory-sharing technique. For more details see [2, Proof of Theorem 1].

A. New achievable scheme:

In this part, we present our achievable scheme which leads to upper bounds on (6). For simplicity let $|\mathcal{C}'|$ be the cardinality of the codeword defined in (14) where $|\mathcal{C}'| = |\mathcal{C}_{\gamma_1}| \times \dots \times |\mathcal{C}_{\gamma_L}|$. In the next result let Q^* achieve the lower bound in Theorem 2 for the following problem

$$H^*(P_{X\mathcal{C}'}) = \inf_{H(\mathcal{C}'|X, U)=0, I(X;U)=0} H(U), \quad (16)$$

where in (7), Y is substituted by $\mathcal{C}'$, i.e., $\mathcal{C}' \leftarrow Y$, and $\mathcal{C}'$ is as defined in (14). Using Theorem 2 and Remark 6, $H^*(P_{X\mathcal{C}'})$ can be lower bounded by $H(Q^*)$ and upper bounded by $H(Q^*) + 0.53$ when $|X| = 2$ and by $H(Q^*) + 1.22$ when $|X| > 2$. For binary X we have

$$H(Q^*) \leq H^*(P_{X\mathcal{C}'}) \leq H(Q^*) + \frac{\log e}{e}, \quad (17)$$

and for $|X| > 2$,

$$H(Q^*) \leq H^*(P_{X\mathcal{C}'}) \leq H(Q^*) + \frac{1 + \log e}{2}. \quad (18)$$

We emphasize that Q^* that is used in (17) and (18) is constructed using the greedy approach based on the matrix $P_{\mathcal{C}'|X}$. We use the same Q^* for the following result.

Theorem 3. Let RVs $(X, Y) = (X, Y_1, \dots, Y_N)$ be distributed according to $P_{XY_1 \dots Y_N}$ supported on alphabets $\mathcal{X}$ and $\mathcal{Y}$, where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are finite, and let the shared secret key size be $|\mathcal{X}|$, i.e., $T = |\mathcal{X}|$. Furthermore, let $M \in \{\frac{N}{K}, \frac{2N}{K}, \dots, N\}$. Let $|X| = 2$, we have

$$\mathbb{L}(P_{XY}, 2) \leq H(Q^*) + \frac{\log e}{e} + 2, \quad (19)$$

where $\mathcal{C}'$ is as defined in (14). When $|X| > 2$, we have

$$\mathbb{L}(P_{XY}, |\mathcal{X}|) \leq H(Q^*) + \frac{1 + \log e}{2} + 1 + \lceil \log(|\mathcal{X}|) \rceil, \quad (20)$$

Proof: The proof is similar to [1] and the main difference is to use minimum entropy output of (16) instead of FRL that is used in two-part construction coding in [1]. In the placement phase, we use the same scheme as discussed before. In the delivery phase, we use the following strategy. Similar to [1], we use two-part code construction to achieve the upper bounds. As shown in Fig. 2, we first encode the private data X using one-time pad coding [15, Lemma 1], which uses $\lceil \log(|\mathcal{X}|) \rceil$ bits. Next, we produce U based on greedy entropy-based algorithm proposed in [4] which solves the minimum entropy problem in (16), where $\mathcal{C}'$, defined in (14), is the response that the server sends over the shared link to satisfy the users' demands [2]. Thus, we have

$$H(\mathcal{C}'|X, U) = 0, \quad (21)$$

$$I(U; X) = 0, \quad (22)$$

Note that in Remark 6 we substitute $\mathcal{G}_S$ by U and for binary X we have

$$H(U) \leq H(Q^*) + \frac{\log e}{e}, \quad (23)$$

and for $|X| > 2$,

$$H(U) \leq H(Q^*) + \frac{1 + \log e}{2}. \quad (24)$$

Thus, we obtain (17) and (18). Moreover, for the leakage constraint we note that the randomness of one-time-pad coding is independent of X and the output of the greedy entropy-based algorithm U . As shown in Fig. 3, at user side,

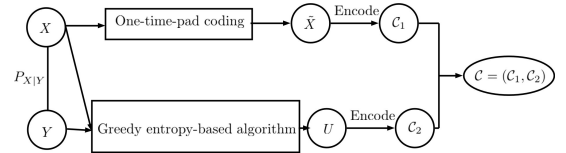


Fig. 2. Encoder design: illustration of the achievability scheme of Theorem 3. Two-part code construction is used to produce the response of the server, $\mathcal{C}$. The server sends $\mathcal{C}$ over the channel, which is independent of X .

each user, e.g., user i , first decodes X using one-time-pad decoder. Then, based on (21) it decodes $\mathcal{C}'$ using U and X . Finally, it decodes Y_{d_i} using local cache Z_i and the response $\mathcal{C}'$. ■

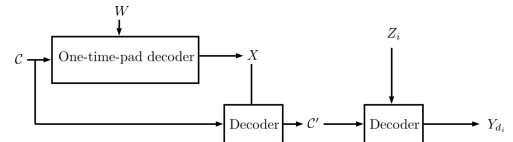


Fig. 3. Illustration of the decoding process for the achievability scheme of Theorem 3.

Remark 7. As we mentioned earlier, the main difference between the present scheme and [1] is to use greedy

entropy-based algorithm which aims to minimize the output of FRL and is optimal within a constant gap.

Remark 8. The complexity of the algorithm to find Q^* in Theorem 3 is linear in $|C'| \times |\mathcal{X}|$. This can be shown by using [6, Lemma 3].

Remark 9. Although in Theorem 3 we assume that $M \in \{\frac{N}{K}, \frac{2N}{K}, \dots, N\}$, the results can be extended for other values of M using the memory sharing technique of [2, Theorem 1].

Remark 10. Similar to [1], we assume that the privacy leakage constraint needs to be fulfilled in the delivery phase. This assumption can be motivated since the placement phase occurs during the off-peak hours and we can assume that the adversary does not listen to the channel during that time. Considering the scenarios in the presence of the adversary during the placement phase, the server can employ the same strategy as used during the delivery phase. The server can fill the caches using the two-part code construction coding.

Next we study a numerical example to better illustrate the achievable scheme in Theorem 3.

Example 1. ([1, Example 1]) Let $F = N = K = 2$ and $M = 1$. Thus, $Y_1 = (Y_1^1, Y_1^2)$ and $Y_2 = (Y_2^1, Y_2^2)$, where $Y_i^j \in \{0, 1\}$ for $i, j \in \{1, 2\}$. We assume that X is the pair of first bits of the database, i.e., $X = (Y_1^1, Y_2^1)$, Y_1 and Y_2 are independent and have the following distributions

$$\begin{aligned} P(Y_1^1 = Y_1^2 = 0) &= P(Y_1^1 = Y_1^2 = 1) = \frac{1}{16}, \\ P(Y_1^1 = 1, Y_1^2 = 0) &= P(Y_1^1 = 0, Y_1^2 = 1) = \frac{7}{16}, \\ P(Y_2^1 = Y_2^2 = 0) &= P(Y_2^1 = Y_2^2 = 1) = \frac{1}{10}, \\ P(Y_2^1 = 1, Y_2^2 = 0) &= P(Y_2^1 = 0, Y_2^2 = 1) = \frac{2}{5}, \end{aligned}$$

In this case, the marginal distributions can be calculated as $P(Y_1^1 = 1) = P(Y_1^2 = 1) = P(Y_2^1 = 1) = P(Y_2^2 = 1) = \frac{1}{2}$. In the placement phase, the server fills the first local cache by the first bits of the database, i.e., $Z_1 = \{Y_1^1, Y_2^1\}$ and the second local cache by the second bits, i.e., $Z_2 = \{Y_1^2, Y_2^2\}$. In the delivery phase, assume that users 1 and 2 request Y_1 and Y_2 , respectively, i.e., $Y_{d_1} = Y_1$ and $Y_{d_2} = Y_2$. In this case, $C' = Y_1^2 \oplus Y_2^1$, where C' is the server's response without considering the privacy constraint. Thus, $|C'| = 2$ and $P_{C'|X} = \begin{bmatrix} \frac{7}{8} & \frac{1}{8} & \frac{1}{8} & \frac{7}{8} \\ \frac{1}{8} & \frac{7}{8} & \frac{7}{8} & \frac{1}{8} \end{bmatrix}$. Moreover, Q^* has the following distribution $P_{Q^*} = [\frac{7}{8}, \frac{1}{8}]$, hence, $H(Q^*) = h(1/8) = 0.5436$. Next, we encode X using W as follows

$$\tilde{X} = X + W \bmod 4,$$

where $W \perp X$ is a RV with uniform distribution over $\mathcal{X}$. To encode $\tilde{X}$ we use 2 bits. We then encode C' using greedy entropy-based algorithm. Let U denote the output of the algorithm which satisfies (24). Let C_1 and C_2 describe the encoded $\tilde{X}$ and U , respectively. The server sends $C = (C_1, C_2)$ over the shared link. For this particular demand vector, using (20), the average codelength is 4.7636 bits. For the request vector $(Y_{d_1}, Y_{d_2}) = (Y_1, Y_2)$, the average

length of the code is 4.7636 bits to satisfy the zero leakage constraint. Thus, for $(Y_{d_1}, Y_{d_2}) = (Y_1, Y_2)$, we have

$$\mathbb{L}(P_{XY}, 4) \leq 4.7636 \text{ bits.}$$

Using [1], for this particular demand vector we need 5 bits. Both users first decode X using $\tilde{X}$ and W , then decode $C' = Y_1^2 \oplus Y_2^1$ by using X and U , since from FRL C' is a deterministic function of U and X . User 1 can decode Y_1^2 using $Y_1^2 \oplus Y_2^1$ and Y_1^1 , which is available in the local cache Z_1 , and user 2 can decode Y_2^2 using $Y_1^2 \oplus Y_2^1$ and Y_2^1 , which is in Z_2 . Moreover, we choose W to be independent of X and U . As a result, X and $(\tilde{X}, U)$ become independent. Thus, $I(C; X) = 0$, which means there is no leakage from X to the adversary. Next, assume that in the delivery phase both users request Y_1 , i.e., $Y_{d_1} = Y_{d_2} = Y_1$. In this case, $C' = Y_1^2 \oplus Y_1^1$ with $|C'| = 2$. Using the same arguments we need 4.7636 bits. Next, let $Y_{d_1} = Y_{d_2} = Y_2$. In this case, $C' = Y_2^2 \oplus Y_2^1$. In this case, $H(Q^*) = h(1/5) = 0.7219$ and we need 4.9419 bits. Finally, let $Y_{d_1} = Y_2$, $Y_{d_2} = Y_1$. In this case, $C' = Y_2^1 \oplus Y_1^1$. Since C' is a function of X it is enough to only send X using on-time pad coding. Thus, for the request vector $Y_{d_1} = Y_2$, $Y_{d_2} = Y_1$, the average length of the code is 2 bits to satisfy the zero leakage constraint. We conclude that in all cases we need less bits to send compared to [1], since by using [1, Example 1] we need 5 bits on average to send over the channel.

B. Special case: improving the bounds using the common information concept

In this section, we improve the bounds obtained in Theorem 3 considering a special case. To do so, let us recall the privacy mechanism design problems considered in [23] with zero leakage as follows

$$g_0(P_{XY}) = \max_{\substack{P_{U|Y}: X-Y-U \\ I(U;X)=0}} I(Y;U), \quad (25)$$

$$h_0(P_{XY}) = \max_{P_{U|Y,X}: I(U;X)=0} I(Y;U). \quad (26)$$

Finally, we define a set of joint distributions $\hat{\mathcal{P}}_{XY}$ as follows

$$\hat{\mathcal{P}}_{XY} \triangleq \{P_{XY} : g_0(P_{XY}) = h_0(P_{XY})\}. \quad (27)$$

As outlined in [24, Lemma 1], a sufficient condition to have $g_0(P_{XY}) = h_0(P_{XY})$ is to have $C(X;Y) = I(X;Y)$, where $C(X;Y)$ denotes the common information between X and Y , where common information corresponds to the Wyner [25] or Gács-Körner [26] notions of common information. Moreover, a sufficient condition for $C(X;Y) = I(X;Y)$ is to let X be a deterministic function of Y or Y be a deterministic function of X . In both cases, $C(X;Y) = I(X;Y)$ and $g_0(P_{XY}) = h_0(P_{XY})$. For more detail see [17, Proposition 6]. Moreover, in [24, Lemma 2], properties of the optimizers for $g_0(P_{XY})$ and $h_0(P_{XY})$ are obtained considering $P_{XY} \in \hat{\mathcal{P}}_{XY}$. It has been shown that the optimizer U^* achieving $g_0(P_{XY}) = h_0(P_{XY})$ satisfies

$$H(Y|U^*, X) = 0, \quad (28)$$

$$I(X; U^*|Y) = 0, \quad (29)$$

$$I(X; U^*) = 0. \quad (30)$$

Next, we recall the definitions of a set $\mathcal{U}^1(P_{XY})$ and a function $\mathcal{K}(P_{XY})$ in [24] as follows.

$$\mathcal{U}^1(P_{XY}) \triangleq \{U : U \text{ satisfies (28), (29), (30)}\} \quad (31)$$

$$\mathcal{K}(P_{XY}) \triangleq \min_{U \in \mathcal{U}^1(P_{XY})} H(U). \quad (32)$$

Noting that the function $\mathcal{K}(P_{XY})$ finds the minimum entropy of all optimizers satisfying $g_0(P_{XY}) = h_0(P_{XY})$. A simple bound on $\mathcal{K}(P_{XY})$ has been obtained in [24, Lemma 4]. Next, we define

$$A_{XY} \triangleq \begin{bmatrix} P_{y_1} - P_{y_1|x_1} & \cdots & P_{y_{|\mathcal{Y}|}} - P_{y_{|\mathcal{Y}|}|x_1} \\ \vdots & \ddots & \vdots \\ P_{y_1} - P_{y_1|x_t} & \cdots & P_{y_q} - P_{y_q|x_t} \end{bmatrix} \in \mathbb{R}^{t \times q}, \quad (33)$$

$$b_{XY} \triangleq \begin{bmatrix} H(Y|x_1) - H(Y|X) \\ \vdots \\ H(Y|x_t) - H(Y|X) \end{bmatrix} \in \mathbb{R}^t, \quad \mathbf{a} \triangleq \begin{bmatrix} a_1 \\ \vdots \\ a_q \end{bmatrix} \in \mathbb{R}^q. \quad (34)$$

where $t = |\mathcal{X}|$ and $q = |\mathcal{Y}|$. Noting that in [24, Theorem 1], bounds on $\mathcal{K}(P_{XY})$ and entropy of any $U \in \mathcal{U}^1$ have been obtained. Specifically, when $\text{rank}(A_{XY}) = |\mathcal{Y}|$, the exact value of $\mathcal{K}(P_{XY})$ is obtained by solving simple linear equations in [24, eq. (26)]. We emphasize that by using [12], $g_0(P_{XY})$ can be obtained by solving a linear program in which the size of the matrix in the system of linear equations is at most $|\mathcal{Y}| \times \binom{|\mathcal{Y}|}{\text{rank}(P_{X|Y})}$ with at most $\binom{|\mathcal{Y}|}{\text{rank}(P_{X|Y})}$ variables. By solving the linear program as proposed in [12] we can find the exact value of $\mathcal{K}(P_{XY})$ and the joint distribution $P_{U|YX}$ that achieves it. The complexity of the linear program in [12] can grow faster than exponential functions with respect to $|\mathcal{Y}|$, however the complexity of the proposed method in [24] grows linearly with $|\mathcal{Y}|$. Thus, our proposed upper bound has less complexity compared to the solution in [12]. The bounds on $\mathcal{K}(P_{XY})$ help us to obtain the next result. Next, we improve the bounds obtained in Theorem 3.

Theorem 4. Let RVs $(X, Y) = (X, Y_1, \dots, Y_N)$ be distributed according to $P_{XY_1 \dots Y_N}$ supported on alphabets $\mathcal{X}$ and $\mathcal{Y}$, where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are finite, and let the shared secret key size be $|\mathcal{X}|$, i.e., $T = |\mathcal{X}|$. Furthermore, let $M \in \{\frac{N}{K}, \frac{2N}{K}, \dots, N\}$. Let $P_{XC'} \in \hat{\mathcal{P}}_{XC'}$ and let $q = |\mathcal{C}'|$ and $\beta = \log(\text{null}(P_{X|C'}) + 1)$, where $\mathcal{C}'$ is defined in (14). Then, we have

$$\mathbb{L}(P_{XC'}, |\mathcal{X}|) \leq \mathcal{K}(P_{XC'}) + 1 + \lceil \log(|\mathcal{X}|) \rceil \quad (35)$$

$$\leq H(\mathcal{C}'|X) + \max_{\substack{a_i: A_{XY}\mathbf{a} = b_{XY}, \mathbf{a} \geq 0, \\ \sum_{i=1}^q P_{c'_i} a_i \leq \beta - H(\mathcal{C}'|X)}} \sum_{i=1}^q P_{c'_i} a_i + 1 + \lceil \log(|\mathcal{X}|) \rceil \quad (36)$$

$$\leq \beta + 1 + \lceil \log(|\mathcal{X}|) \rceil, \quad (37)$$

where c'_i is the i -th element (alphabet) of $\mathcal{C}'$. Moreover, we have

$$\mathbb{L}(P_{XY}, 2) \leq H(Q^*) + \frac{\log e}{e} + 2, \quad (38)$$

$$\mathbb{L}(P_{XY}, |\mathcal{X}|) \leq H(Q^*) + \frac{1 + \log e}{2} + 1 + \lceil \log(|\mathcal{X}|) \rceil, \quad (39)$$

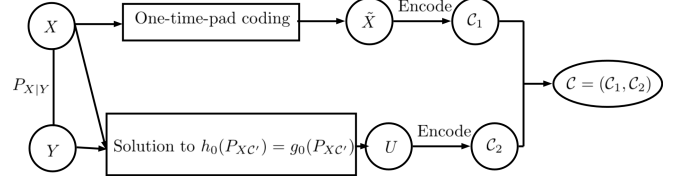


Fig. 4. In this work, we use two-part construction coding strategy to send codewords over the channels. We hide the information of X using one-time-pad coding and we then use the solution of $h_0(P_{XC'}) = g_0(P_{XC'})$ to construct U .

where Q^* is defined in Theorem 3. Finally, for any $P_{XC'}$ (not necessarily $P_{XC'} \in \hat{\mathcal{P}}_{XC'}$) with $|\mathcal{C}'| \leq |\mathcal{X}|$ we have

$$\mathbb{L}(P_{XC'}, |\mathcal{C}'|) \leq \lceil \log |\mathcal{C}'| \rceil. \quad (40)$$

Proof: The proof is based on two-part construction coding and is similar to Theorem 3 and [24, Theorem 2]. To achieve (35), we use the solution to $h_0(P_{XC'}) = g_0(P_{XC'})$ instead of the greedy entropy-based algorithm. Moreover, to achieve (36) and (37), we use two-part construction coding and inequalities obtained in [24, Theorem 1]. Upper bounds (38) and (39) are obtained in Theorem 3. Finally, to achieve (40), let the shared key W be independent of (X, Y) and has uniform distribution $\{1, \dots, T\} = \{1, \dots, |\mathcal{C}'|\}$. We construct $\tilde{C}$ using one-time pad coding. We have

$$\tilde{C} = C' + W \text{ mod } |\mathcal{Y}|,$$

where C' is defined in (14) and clearly we have

$$I(\tilde{C}; X) = 0. \quad (41)$$

Then, $\tilde{C}$ is encoded using any lossless code which uses at most $\lceil \log(|\mathcal{C}|) \rceil$ bits. At decoder side, we first decode C' using the shared key. We then decode each demanded file by using the cache contents and C' . ■

Remark 11. Clearly, the upper bound obtained in (40) improves the bounds in Theorem 3. Since, when $|\mathcal{C}'| \leq |\mathcal{X}|$ we have

$$\lceil \log |\mathcal{C}'| \rceil \leq H(Q^*) + \frac{1 + \log e}{2} + 1 + \lceil \log(|\mathcal{X}|) \rceil. \quad (42)$$

Next, we provide a numerical example that shows (35) improves (19).

Example 2. Let $P_{X|C'} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$ and $P_{C'} = [\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}]$. Clearly, in this case X is a deterministic function of C' . Using the linear program proposed in [12], we obtain a solution as $P_{C'|u_1} = [0.75, 0, 0, 0.25, 0, 0]$, $P_{C'|u_2} = [0, 0.75, 0, 0.25, 0, 0]$, $P_{C'|u_3} = [0, 0, 0.75, 0, 0.25, 0]$, $P_{C'|u_4} = [0, 0, 0.75, 0, 0, 0.25]$ and $P_U = [\frac{1}{6}, \frac{1}{3}, \frac{1}{4}, \frac{1}{4}]$ which results $H(U) = 1.9591$ bits. We have $H(U) = \mathcal{K}(P_{XY}) \leq 1.9591$. Moreover, we have

$$P_{C'|X} = \begin{bmatrix} \frac{1}{6} & 0 \\ \frac{1}{3} & 0 \\ \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \\ 0 & \frac{1}{4} \\ 0 & \frac{1}{4} \end{bmatrix}.$$

Using the greedy search algorithm we have $P_{Q^*} = [\frac{1}{2} \frac{1}{4} \frac{1}{6} \frac{1}{12}]$, hence, $H(Q^*) = 1.7296$. Thus,

$$H(Q^*) + \frac{\log e}{e} = 2.2596 \geq \mathcal{K}(P_{XY}) = 1.9591.$$

IV. CONCLUSION

We have studied a cache-aided compression problem with a perfect privacy constraint, where the information delivered over the shared link during the delivery phase is independent of X that is correlated with the files in the database that can be requested by the users. We have strengthened the previous achievable scheme by using a greedy entropy-based algorithm instead of the FRL. The greedy algorithm aims to solve the minimum entropy functional representation and is optimal within a constant gap. Considering two special cases the obtained bounds are strengthened. Specifically, when the size of the private data is large, we need significantly less bits to send over the channel with less shared key size.

V. ACKNOWLEDGMENT

The authors would like to express their gratitude to Yanina Shkel for suggesting the minimum entropy coupling problem, the greedy entropy-based approach, and providing related references.

REFERENCES

- [1] A. Zamani, T. J. Oechtering, D. Gündüz, and M. Skoglund, "Cache-aided private variable-length coding with zero and non-zero leakage," in *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2023, pp. 247–254.
- [2] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.
- [3] Y. Y. Shkel, R. S. Blum, and H. V. Poor, "Secrecy by design with applications to privacy and compression," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 824–843, 2021.
- [4] M. Kocaoglu, A. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causal inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [5] S. Compton, D. Katz, B. Qi, K. Greenewald, and M. Kocaoglu, "Minimum-entropy coupling approximation guarantees beyond the majorization barrier," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 10 445–10 469.
- [6] Y. Y. Shkel and A. K. Yadav, "Information spectrum converse for minimum entropy couplings and functional representations," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 66–71.
- [7] S. H. Lim, C.-Y. Wang, and M. Gastpar, "Information-theoretic caching: The multi-user case," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7018–7037, 2017.
- [8] C.-Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," in *Information Theory and Applications Workshop, 2016*, 2016, pp. 1–6.
- [9] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Transactions on Information Theory*, 2017.
- [10] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4347–4364, 2018.
- [11] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.
- [12] B. Rassouli and D. Gündüz, "On perfect privacy," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 177–191, 2021.
- [13] A. Zamani, T. J. Oechtering, and M. Skoglund, "A design framework for strongly χ^2 -private data disclosure," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2312–2325, 2021.
- [14] A. Zamani, T. J. Oechtering, and M. Skoglund, "Data disclosure with non-zero leakage and non-invertible leakage matrix," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 165–179, 2022.
- [15] Y. Y. Shkel and H. V. Poor, "A compression perspective on secrecy measures," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 163–176, 2021.
- [16] S. Asodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, p. 15, 2016.
- [17] A. Zamani, T. J. Oechtering, and M. Skoglund, "On the privacy-utility trade-off with and without direct access to the private data," *IEEE Transactions on Information Theory*, vol. 70, no. 3, pp. 2177–2200, 2024.
- [18] T.-Y. Liu and I.-H. Wang, "Privacy-utility tradeoff with nonspecific tasks: Robust privatization and minimum leakage," in *2020 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–5.
- [19] M. A. Zarabian, N. Ding, and P. Sadeghi, "On the lift, related privacy measures, and applications to privacyutility trade-offs," *Entropy*, vol. 25, no. 4, 2023. [Online]. Available: <https://www.mdpi.com/1099-4300/25/4/679>
- [20] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 594–603, 2020.
- [21] S. Asodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, 2016. [Online]. Available: <https://www.mdpi.com/2078-2489/7/1/15>
- [22] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop*, 2014, pp. 501–505.
- [23] A. Zamani, T. J. Oechtering, and M. Skoglund, "Bounds for privacy-utility trade-off with non-zero leakage," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 620–625.
- [24] A. Zamani, T. J. Oechtering, D. Gündüz, and M. Skoglund, "Private variable-length coding with zero leakage," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023, pp. 1–6.
- [25] A. D. Wyner, "The wire-tap channel," *The Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, 1975.
- [26] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 149–162, 1973.