# Digital Blind Box: Random Symmetric Private Information Retrieval

Zhusheng Wang Sennur Ulukus
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
*zhusheng@umd.edu* *ulukus@umd.edu*

***Abstract*—We introduce the problem of random symmetric private information retrieval (RSPIR). In canonical PIR, a user downloads a message out of $K$ messages from $N$ non-colluding and replicated databases in such a way that no database can know which message the user has downloaded (user privacy). In SPIR, the privacy is symmetric, in that, not only that the databases cannot know which message the user has downloaded, the user itself cannot learn anything further than the particular message it has downloaded (database privacy). In RSPIR, different from SPIR, the user does not have an input to the databases, i.e., the user does not pick a specific message to download, instead is content with any one of the messages. In RSPIR, the databases need to send symbols to the user in such a way that the user is guaranteed to download a message correctly (random reliability), the databases do not know which message the user has received (user privacy), and the user does not learn anything further than the one message it has received (database privacy). This is the digital version of a blind box, also known as gachapon, which implements the above specified setting with physical objects for entertainment. This is also the blind version of $1$-out-of-$K$ oblivious transfer (OT), an important cryptographic primitive. We study the information-theoretic capacity of RSPIR for the case of $N = 2$ databases. We determine its exact capacity for the cases of $K = 2, 3, 4$ messages. While we provide a general achievable scheme that is applicable to any number of messages, the capacity for $K \geq 5$ remains open.**

## I. Introduction

Gachapon is a vending machine-dispensed capsule toy by means of a roulette mechanism, which makes it random and unpredictable for customers [1]. In addition, gachapon is being adapted as a random-type item in online games and 3D printing, and its digital form is catching on quickly in the worldwide market [2], [3]. Due to packaging requirements prior to official distribution, gachapon is also referred to as a *blind box* [1]. A blind box is a type of packaging that keeps its contents hidden. The covers of blind boxes are identical in every way. Nobody including the manufacturer knows what exactly is inside until the customer opens a blind box [4]. Nowadays, not only constrained to the scope of entertainment, blind box has become a commercial phenomenon in certain parts of the world impacting people's daily lives.

Following the concepts of gachapon as well as blind box, we introduce a *digital blind box* between a user and a server in a communication network with the following characteristics: 1) A user will ultimately receive a random box (content) from the server. However, the user does not know anything about what is in the box (what the content is) until it receives a box (content) from the server. 2) For the sake of unpredictability, a user should also know nothing about the current box (content) based on what it has received in the previous transactions. A user should not know anything about what other users have received before communicating with them. In other words, a user should not know anything beyond what it receives from the current box (content). This requirement also protects the content privacy of the server. 3) In order to protect the privacy of the users, the server should learn nothing about what a specific user has received.

Introduced in [5], [6], private information retrieval (PIR) characterizes a fundamental problem, where a user downloads a message out of multiple messages stored in several non-colluding and replicated databases in such a way that no single database can know which message the user has downloaded. This privacy requirement is referred to as *user privacy*. Some important variations of the PIR problem have been investigated in [7]–[39]. Further extended in [40], [41], symmetric PIR (SPIR) requires in addition that the user learns nothing about the remaining messages stored in the databases after downloading its desired message. This privacy requirement is referred to as *database privacy*. Some important variations of SPIR problem have been investigated in [42]–[51].

In this paper, we introduce a new concept called random SPIR (RSPIR). In reference to the conventional SPIR, the only difference is that, in RSPIR there is no input at the user side. That is, the user does not send any queries to the databases, and ultimately receives a random message from the databases. This requirement is referred to as *random reliability*. Interestingly, the three requirements of RSPIR, namely, random reliability, database privacy and user privacy, strictly correspond to the three characteristics of the digital blind box described above. Thus, the digital blind box is equivalent to the RSPIR.

Oblivious transfer (OT), first introduced in [52] and then developed in [53], is an essential building block in modern cryptography. A 1-out-of-$K$ OT protocol consists of two parties, a sender with $K$ input messages and a receiver with a choice $k \in [K]$. The objective of the protocol is that the receiver will receive the $k$th message without the sender learning the index $k$, while the sender can guarantee that the receiver only received one of the $K$ messages. Note that SPIR is a distributed (multi-database) version of 1-out-of-$K$ OT. An important variant of 1-out-of-$K$ OT is that the receiver has no input. Thus, the receiver will receive each potential message

with equal probability without gaining any partial knowledge about the remaining messages, while the sender is ignorant of which message has been received by the receiver. For example, this variant can be used as a subroutine in contract signing and certified mail protocols [53]. Likewise, RSPIR can be viewed as a distributed version of this variant of 1-out-of-$K$ OT.

One potential application of RSPIR is in the field of symmetric key sharing. Consider the following situation: Two non-colluding databases share multiple symmetric keys. Each database individually selects a random answer and then broadcasts this answer to a set of users denoted by $\mathcal{U}_1$. Thus, one symmetric key is now shared among users in $\mathcal{U}_1$. For another set of users denoted by $\mathcal{U}_2$, the first database selects the same answer while the second database selects another new random answer, and they broadcast these answers to users in $\mathcal{U}_2$. Thus, another distinct symmetric key is now shared among $\mathcal{U}_2$. As a consequence, a message can be circulated within $\mathcal{U}_1$ securely against the eavesdropping attack launched by any single database or any user in $\mathcal{U}_2$. This property also applies to the secure information circulation within $\mathcal{U}_2$. Another instance of RSPIR can be observed in the problem formulation in [44] which considers the SPIR problem with user-side common randomness. The problem formulation in [44] allows the user to fetch a random subset of the common randomness available at the databases to form user-side side-information unknown to the databases (unknown also to the user before it receives them). The purpose of this action is to increase the SPIR rate; in fact, such an action increases the SPIR rate to the level of PIR rate. The common randomness fetching phase of [44] is an instance of RSPIR problem.

In this paper, we formulate $N = 2$ database RSPIR and investigate its capacity. We determine its capacity as well as the minimal amount of required common randomness in the cases of $K = 2, 3, 4$ messages. This determines the capacity of digital blind box. While we give a general achievable scheme for any number of messages, the exact capacity of RSPIR for $K \geq 5$ remains an open problem.

## II. RSPIR: Problem Formulation

In this paper, we consider $N = 2$ non-colluding databases each storing the same set of $K \geq 2$ i.i.d. messages. Each message consists of $L$ i.i.d. uniformly chosen symbols from a sufficiently large finite field $\mathbb{F}_q$, i.e.,

$$H(W_k) = L, \quad k \in [K] \tag{1}$$

$$H(W_{1:K}) = H(W_1) + \cdots + H(W_K) = KL \tag{2}$$

The two databases jointly share a necessary common randomness random variable $\mathcal{S}$, which is generated independent of the message set $W_{1:K}$. Thus,

$$H(W_{1:K}, \mathcal{S}) = H(W_{1:K}) + H(\mathcal{S}) \tag{3}$$

Before the RSPIR process starts, an answer set $\mathcal{A}$ with cardinality $M_1$ is assigned to database 1 while another answer set $\mathcal{B}$ with cardinality $M_2$ is assigned to database 2. Since there is no input at the user side in the RSPIR process, the databases will never receive a query from the user. Therefore, as a simple approach, each database individually selects a random answer under a uniform distribution from its corresponding answer set and then transmits it to the user. The indices of the answers for two databases are denoted by $X$ and $Y$, respectively, i.e., database 1 will select $A_X \in \mathcal{A}$ and database 2 will select $B_Y \in \mathcal{B}$. Moreover, we use $x$ and $y$ to denote the realizations of the random variables $X$ and $Y$, respectively. We note that every answer from any answer set is generated based on the message set and the common randomness, hence, for all $X \in [M_1]$ and $Y \in [M_2]$, we have,

$$\text{[deterministic answer]} \quad H(A_X, B_Y | X, Y, W_{1:K}, \mathcal{S}) = 0 \tag{4}$$

After collecting two arbitrary answers from the databases, the user should always be able to decode a random message reliably. Thus, for all $X \in [M_1]$ and $Y \in [M_2]$, we can always find an index $\theta_{X,Y} \in [K]$ (the mapping here is not deterministic) such that

$$\text{[random reliability]} \quad H(W_{\theta_{X,Y}} | X, Y, A_X, B_Y) = 0 \tag{5}$$

Because of the database privacy constraint, the user is supposed to learn nothing about $W_{\bar{\theta}_{X,Y}}$ which is the complement of the randomly retrieved message $W_{\theta_{X,Y}}$, i.e., $W_{\bar{\theta}_{X,Y}} = \{W_1, \ldots, W_{\theta_{X,Y}-1}, W_{\theta_{X,Y}+1}, \ldots, W_K\}$,

$$\text{[database privacy]} \quad I(W_{\bar{\theta}_{X,Y}}; X, Y, A_X, B_Y) = 0 \tag{6}$$

Because of the user privacy constraint, i.e., the protection of this randomly retrieved message's index in the user, from the perspective of each individual database, this index must be indistinguishable for each randomly selected answer under a uniform distribution. In other words, even though an answer from one database is deterministic, the user can still decode every potential message in the message set with equal probability through the variation of the answer from the other database. Thus, for the first database, given any realization $x \in [M_1]$, we always have the following probability distribution of the random variable $\theta_{x,Y}$ with respect to the random variable $Y$,

$$P(\theta_{x,Y} = k) = \frac{1}{K}, \quad \forall k \in [K] \tag{7}$$

which is equivalent to

$$\text{[user privacy]} \quad I(x, A_x, W_{1:K}, \mathcal{S}; \theta_{x,Y}) = 0 \tag{8}$$

By symmetry, for database 2, given any realization $y \in [M_2]$, we also have the following probability distribution of the random variable $\theta_{X,y}$ with respect to the random variable $X$,

$$P(\theta_{X,y} = k) = \frac{1}{K}, \quad \forall k \in [K] \tag{9}$$

which is equivalent to

$$\text{[user privacy]} \quad I(y, B_y, W_{1:K}, \mathcal{S}; \theta_{X,y}) = 0 \tag{10}$$

As a consequence, we obtain the following theorem regarding the cardinality of the answer sets, which can be proved by contradiction using the user privacy constraint.

**Theorem 1** *The total possible number of answers in the answer set for each database must be a multiple of $K$, i.e.,*

$$M_1 = t_1 K,\ M_2 = t_2 K, \quad t_1, t_2 \in N^+ \tag{11}$$

Moreover, we also have the following theorem concerning the common randomness distribution in the databases.

**Theorem 2** *As in multi-database SPIR [40], [41], in RSPIR, the databases must share some necessary common randomness that is unknown to the user before the retrieval process starts. Otherwise, RSPIR is not feasible.*

**Proof:** Without any common randomness in the databases, for any $X \in [M_1]$ and $Y \in [M_2]$, the random reliability constraint and the database privacy constraint collectively lead to,

$$0 = I(W_{\bar{\theta}_{X,Y}}; X, Y, A_X, B_Y) \tag{12}$$
$$= I(W_{\bar{\theta}_{X,Y}}; W_{\theta_{X,Y}}, X, Y, A_X, B_Y) \tag{13}$$
$$= H(W_{\bar{\theta}_{X,Y}}) - H(W_{\bar{\theta}_{X,Y}} | W_{\theta_{X,Y}}, X, Y, A_X, B_Y) \tag{14}$$

Then, we consider the following expression

$$I(X, Y, A_X, B_Y; W_{\bar{\theta}_{X,Y}} | W_{\theta_{X,Y}})$$
$$= H(W_{\bar{\theta}_{X,Y}} | W_{\theta_{X,Y}}) - H(W_{\bar{\theta}_{X,Y}} | W_{\theta_{X,Y}}, X, Y, A_X, B_Y) \tag{15}$$
$$= H(W_{\bar{\theta}_{X,Y}}) - H(W_{\bar{\theta}_{X,Y}}) \tag{16}$$
$$= 0 \tag{17}$$

where (16) follows from (14). For any realization $x$,

$$0 = I(A_x; W_{\bar{\theta}_{x,Y}} | x, W_{\theta_{x,Y}}) \tag{18}$$
$$= H(A_x | x, W_{\theta_{x,Y}}) - H(A_x | x, W_{1:K}) \tag{19}$$
$$= H(A_x | W_{\theta_{x,Y}}) \tag{20}$$

where (18) follows from (17), and (20) follows from the deterministic answer constraint $H(A_x|x, W_{1:K}) = 0$ without common randomness. Taking into consideration the fact that (20) is true for any realization $y \in [M_2]$ as well as the user privacy constraint (7), we have $H(A_x|W_1) = \cdots = H(A_x|W_K) = 0$. Since messages are all mutually independent, it is easy to derive that $H(A_x) = 0$, which forms a contradiction. ■

A valid two-database RSPIR achievable scheme is a scheme that satisfies the user privacy constraint (8), (10), the database privacy constraint (6) and the random reliability constraint (5).

The efficiency of a scheme is measured in terms of the total number of downloaded bits by the user from the two databases, named as the download cost. According to the formulation above, the download cost consists of the answer indices $X, Y$ and the answers themself $A_X, B_Y$. Compared with the answer cost, the answer index cost can be neglected as it does not scale with the message length if we reuse them to decode each symbol in the randomly retrieved message. Thus, we use $D_{RSPIR}$ to denote the expected number of bits contained in the answers $A_X, B_Y$ over the indices $X, Y$. Then the retrieval rate of RSPIR is given by,

$$R_{RSPIR} = \frac{L}{D_{RSPIR}} \tag{21}$$

The capacity of RSPIR, $C_{RSPIR}$, is the supremum of the retrieval rates $R_{RSPIR}$ over all valid achievable schemes.

## III. Main Results

**Theorem 3** *In the two-database RSPIR problem, in the case of $K = 2$, the capacity is $\frac{1}{2}$ with minimal amount of required common randomness being $L$. In the case of $K = 3, 4$, the capacity is $\frac{1}{3}$ with minimal amount of required common randomness being $2L$.*

The converse proof of Theorem 3 is given in Section IV, and the achievability proof of Theorem 3 is presented in Section V. The capacity and its minimal amount of required common randomness in the case of $K \geq 5$ is an open problem.

**Remark 1** *It is well known [40] that the capacity of multi-database SPIR is $1 - \frac{1}{N}$, where $N$ is the number of replicated and non-colluding databases. As a corollary, the capacity of two-database SPIR is $\frac{1}{2}$, which does not depend on the number of messages $K$ stored in the databases. By contrast, the capacity of RSPIR does depend on the value of $K$. Even though the capacity of RSPIR achieves the same limit as SPIR in the case of $K = 2$, the capacity of RSPIR decreases to $\frac{1}{3}$ when the value of $K$ increases to 3.*

**Remark 2** *Because of the equivalence between RSPIR and the digital blind box, in a digital blind box setting where two non-colluding databases share $K$ messages and some necessary common randomness, perfect digital blind box delivery can be achieved with a linear download cost $KL$. The proof is a direct consequence of the second general achievable scheme given in Section V.*

**Remark 3** *In the problem formulation part of our previous work [44], we assume that the user is able to obtain a random subset of the shared common randomness that is unknown to any individual database before the SPIR retrieval process starts. Although we mention the idea of fetching common randomness like side-information in advance, we do not specify in [44] a corresponding practical implementation. Now, it is clear that the achievability provided here for RSPIR can be used as a practical approach for this problem if common randomness is treated as another independent message system.*

## IV. Converse Proof

**Theorem 4** *In the two-database RSPIR problem, the capacity is realized in the case where $M_1$ and $M_2$ are both exactly $K$.*

**Proof:** We provide a sketch of proof here. The idea of the proof is that once we multiply the value of $M_1$ by an integer $t \geq 2$, it is straightforward to see that additional constraints will be added to each pair $A_X, B_Y$ for all $X \in [tM_1]$ and $Y \in [M_2]$ after considering the index permutation, which will either increase or maintain the minimal value of $H(A_X) + H(B_Y)$. This analysis also applies to the increase of $M_2$. ■

In the case of $K = 2$, motivated by Theorem 4, we consider the simplest case where $M_1 = 2$ and $M_2 = 2$. Then, we only

need to investigate the following constraints since all the other potential system of constraints have the same structure as this one and will lead to the same conclusions,

$$H(W_1|A_1,B_1)=0, \quad H(W_1|A_2,B_2)=0 \tag{22}$$
$$H(W_2|A_1,B_2)=0, \quad H(W_2|A_2,B_1)=0 \tag{23}$$

These constraints exactly reflect the random reliability constraint (5) and user privacy constraint (7), (9) involved in this problem. First, we prove a lower bound for $H(A_1)+H(B_1)$,

$$\begin{aligned}
&H(A_1)+H(B_1)\\
&\geq H(A_1|A_2,B_1)+H(B_1|A_2,B_2) && (24)\\
&= H(A_1,A_2,B_1)+H(A_2,B_1,B_2)\\
&\quad - H(A_2,B_1)-H(A_2,B_2) && (25)\\
&= H(W_1,A_1,A_2,B_1)+H(W_1,A_2,B_1,B_2)\\
&\quad - H(A_2,B_1)-H(A_2,B_2) && (26)\\
&\geq H(W_1,A_2,B_1)+H(W_1,A_1,A_2,B_1,B_2)\\
&\quad - H(A_2,B_1)-H(A_2,B_2) && (27)\\
&= H(A_1,A_2,B_1,B_2)-H(A_2,B_2)+H(W_1) && (28)\\
&\geq H(W_2,A_2,B_2)-H(A_2,B_2)+H(W_1) && (29)\\
&= H(W_2)+H(W_1) && (30)\\
&= 2L && (31)
\end{aligned}$$

where (28) and (30) follow from the database privacy constraint. Likewise, we can always obtain $H(A_X)+H(B_Y)\geq 2L$ for any other answer pair $A_X,B_Y,X,Y\in[2]$. As a result, we reach a converse result for the capacity when $K=2$,

$$R=\frac{L}{D}\leq\frac{L}{H(A_X)+H(B_Y)}\leq\frac{L}{2L}=\frac{1}{2} \tag{32}$$

Next, we prove the minimal required amount of common randomness shared in the two databases.

$$\begin{aligned}
0 &= I(W_2;A_1,B_1) && (33)\\
&= I(W_2;A_1,B_1|W_1) && (34)\\
&= H(A_1,B_1|W_1)-H(A_1,B_1|W_1,W_2)\\
&\quad + H(A_1,B_1|W_1,W_2,\mathcal{S}) && (35)\\
&= H(A_1,B_1|W_1)-I(A_1,B_1;\mathcal{S}|W_1,W_2) && (36)\\
&= H(A_1,B_1|W_1)-H(\mathcal{S}|W_1,W_2)\\
&\quad + H(\mathcal{S}|W_1,W_2,A_1,B_1) && (37)\\
&\geq H(A_1,B_1|W_1)-H(\mathcal{S}) && (38)
\end{aligned}$$

where (35) follows from the deterministic answer constraint (4) and (38) follows from the independence between message set and the common randomness (3). Therefore, we turn to find a lower bound for the expression $H(A_1,B_1|W_1)$,

$$\begin{aligned}
&H(A_1,B_1|W_1)\\
&= H(A_1|W_1,B_1)+H(B_1|W_1) && (39)\\
&\geq H(A_1|W_1,A_2,B_1)+H(B_1|W_1,A_2,B_2) && (40)\\
&= H(A_1,A_2,B_1)+H(A_2,B_1,B_2)\\
&\quad - H(W_1,A_2,B_1)-H(A_2,B_2) && (41)\\
&= H(A_1,A_2,B_1)+H(A_2,B_1,B_2)-H(A_2,B_1)\\
&\quad - H(A_2,B_2)-H(W_1) && (42)\\
&\geq H(W_2) && (43)\\
&= L && (44)
\end{aligned}$$

where (42) follows from the database privacy constraint and (43) exactly follows from the steps between (25)-(30). As a consequence, we reach a converse result for the minimal amount of required common randomness,

$$H(\mathcal{S})\geq L \tag{45}$$

In the case of $K=3$, $M_1$ and $M_2$ both take the value 3, after converting the random reliability constraint and user privacy constraint into pairwise constraints as in (22)-(23), we can proceed with the converse steps. As in the converse proof in the case of $K=2$ above, the concrete process is to utilize the converse proof of [46, Theorem 2] once more after eliminating the influence of retrieval strategy randomness and its generated queries. Thus, we have the same conclusions as the one in [46, Theorem 2] in the case of $K=3$,

$$R\leq\frac{1}{3}, \quad H(\mathcal{S})\geq 2L \tag{46}$$

In the case of $K=4$, it is easy to verify that each answer pair $A_X,B_Y,X,Y\in[4]$ has more constraints than the one when $K=3$. Thus, a converse proof for the capacity and the minimal amount of required common randomness in the case of $K=4$ can be inherited from the case of $K=3$, i.e.,

$$R\leq\frac{1}{3}, \quad H(\mathcal{S})\geq 2L \tag{47}$$

A tight converse proof for the capacity and the minimal amount of required common randomness remains to be found in the case of $K\geq 5$.

## V. Achievability

The work in [45] provides a scheme that can be readily converted into an achievable scheme (albeit suboptimal) for the two-database RSPIR problem. For clarity, we restate the result from the new perspective of RSPIR here. Assuming that $L=1$ for the time being, two databases share $K$ common randomness symbols $S_1,\ldots,S_K$, which are all uniformly selected from $\mathbb{F}_q$. For database 1, the answer set $\mathcal{A}$ is composed of $K$ elements in the following form,

$$\begin{aligned}
A_1 &= (W_1+S_1, W_2+S_2, \ldots, W_K+S_K) && (48)\\
A_2 &= (W_1+S_2, W_2+S_3, \ldots, W_K+S_1) && (49)\\
&\vdots\\
A_K &= (W_1+S_K, W_2+S_1, \ldots, W_K+S_{K-1}) && (50)
\end{aligned}$$

Basically, we only rotate common randomness symbols by one place in the sequence of answers. A homomorphic variation of $\mathcal{A}$ is to rotate message symbols by one place without imposing any influence on the answer set $\mathcal{B}$ and it is shown as follows,

$$A_1 = (W_1+S_1, W_2+S_2, \ldots, W_K+S_K) \tag{51}$$

$$A_2 = (W_2+S_1, W_3+S_2, \ldots, W_1+S_K) \quad (52)$$
$$\vdots$$
$$A_K = (W_K+S_1, W_1+S_2, \ldots, W_{K-1}+S_K) \quad (53)$$

For database 2, the answer set $\mathcal{B}$ also including $K$ elements is shown as follows,

$$B_1 = S_1, \quad B_2 = S_2, \quad \cdots \quad B_K = S_K \quad (54)$$

The answer set construction in these two databases is public knowledge to the user. Afterwards, database 1 selects a random answer under a uniform distribution from $\mathcal{A}$, and then sends the values of symbols as well as the index belonging to this answer to the user. Likewise, database 2 performs the same selection and transmission. The reason for sending the answer indices is that the user does not know how to use the values of symbols in the answers to decode a random message without the help of the answer indices. After receiving two answers, the user is always able to decode one random message reliably. Moreover, since each database is doing the uniform selection, this random message is equally likely to be any message in the message set. Therefore, it is impossible for each individual database to learn the index of this randomly retrieved message at the user side. Meanwhile, the user cannot learn any information about the remaining messages because of the existence of unknown common randomness symbols. When each message includes multiple symbols, we can simply perform this scheme repeatedly for each symbol while there is no need to do the new selection nor send the answer index for each database after first execution. Thus, the download cost of answer index can be ignored as illustrated in the problem formulation when $L$ is large enough. Obviously, the download cost is $D = (K+1)L$ in this scheme but it is not optimal.

Here, we provide a new general scheme that achieves the download cost of $D = KL$ when $L$ goes to infinity. Assuming that $L = 1$ temporarily, let two databases share $K-1$ common randomness symbols $S_1, \ldots, S_{K-1}$. For database 1, the answer set $\mathcal{A}$ contains $K$ elements in the following form,

$$A_1 = (S_1, S_2, \ldots, S_{K-1}) \quad (55)$$
$$A_2 = (W_1+W_2+S_1, W_2+W_3+S_2, \ldots, W_{K-1}+W_K+S_{K-1}) \quad (56)$$
$$A_3 = (W_1+W_3+S_1, W_2+W_4+S_2, \ldots, W_{K-1}+W_1+S_{K-1}) \quad (57)$$
$$\vdots$$
$$A_K = (W_1+W_K+S_1, W_2+W_1+S_2, \ldots, W_{K-1}+W_{K-2}+S_{K-1}) \quad (58)$$

Basically, except for the first answer, we only rotate the second message symbol by one place in the sequence of answers while keeping the first message symbol. For database 2, the answer set $\mathcal{B}$ consists of $K$ elements in the following form,

$$B_1 = W_1+S_1, \quad \cdots \quad B_{K-1} = W_{K-1}+S_{K-1},$$
$$B_K = W_K+S_1+S_2+\ldots S_{K-1} \quad (59)$$

Now, note that, since this scheme achieves a download cost of $D = KL$, it achieves a rate of $R = \frac{L}{D} = \frac{L}{KL} = \frac{1}{K}$. For $K = 2$ and $K = 3$, this scheme achieves rates $\frac{1}{2}$ and $\frac{1}{3}$ meeting the converse bounds in (32) and (46), respectively. Specifically, when $K = 2$, we have the following answer sets,

$$A_1 = S_1 \quad B_1 = W_1+S_1 \quad (60)$$
$$A_2 = W_1+W_2+S_2 \quad B_2 = W_2+S_2 \quad (61)$$

When $K = 3$, we have the following answer sets,

$$A_1 = (S_1, S_2), \; B_1 = W_1+S_1 \quad (62)$$
$$A_2 = (W_1+W_2+S_1, W_2+W_3+S_2), \; B_2 = W_2+S_2 \quad (63)$$
$$A_3 = (W_1+W_3+S_1, W_2+W_1+S_2), \; B_3 = W_3+S_1+S_2 \quad (64)$$

When $K = 4$, this achievable scheme achieves a rate $R = \frac{1}{K} = \frac{1}{4}$ whereas the converse in (47) gives a bound of $\frac{1}{3}$. Now, we provide a better scheme that achieves the converse in the case of $K = 4$. The message length $L$ is assumed to be 2 such that $W_1 = \{a_1, a_2\}, W_2 = \{b_1, b_2\}, W_3 = \{c_1, c_2\}$ and $W_4 = \{d_1, d_2\}$. Moreover, two databases share 4 common randomness symbols $S_1, S_2, S_3, S_4$. For database 1, the answer set $\mathcal{A}$ containing 4 elements is in the following form,

$$A_1 = (S_1, S_2, S_3) \quad (65)$$
$$A_2 = (a_1+c_1+c_2+S_1, b_2+d_1+S_1+S_3, c_2+S_4) \quad (66)$$
$$A_3 = (a_1+d_2+S_1+S_4, a_2+d_1+d_2+S_2, b_1+c_2+S_2+S_3) \quad (67)$$
$$A_4 = (b_1+S_4, a_1+a_2+b_1+b_2+S_1+S_2, c_1+d_2+S_1+S_2+S_3) \quad (68)$$

For database 2, the answer set $\mathcal{B}$ with 4 elements is as follows,

$$B_1 = (a_1+S_1, a_2+S_2, S_4) \quad (69)$$
$$B_2 = (b_1+b_2+S_1+S_2, b_1+S_2+S_3, a_1+c_1+d_2+S_1+S_4) \quad (70)$$
$$B_3 = (d_1+d_2+S_2, b_1+c_2+S_4, d_1+S_1+S_3) \quad (71)$$
$$B_4 = (c_2+S_2+S_3, c_1+c_2+S_1, a_1+a_2+b_2+c_1+d_1+S_3+S_4) \quad (72)$$

Here, the download cost is $D = 6$ and the rate is $R = \frac{L}{D} = \frac{1}{3}$. The remaining steps and verification of this specific achievable scheme are the same as the last two general ones. Specifically, regarding verification, we can use the bipartite graph in Fig. 1. In this bipartite graph, by using colors red, yellow, green and blue for messages $W_1, W_2, W_3$ and $W_4$, respectively, the color of links indicates which message should be decoded while keeping all the other messages private. Moreover, each node is always connected to 4 links with different colors.

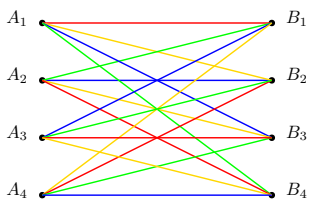


Fig. 1. A two-database RSPIR bipartite graph for $K = 4$ messages.

## REFERENCES


[1] https://en.wikipedia.org/wiki/Gashapon#cite_note-2.

[2] M. Fujihara and A. Shibuya. How is the Gacha system reported on in Japan? In *Digital Games Research Association (DiGRA) International Conference: Play Everywhere*, June 2020.

[3] K. Charnsil, K. Choochuen, et al. 3D-gARt—A new Gachapon 3D-printed toy played with augmented reality and story narration. In *IEEE Global Conf. on Life Sciences and Technologies (LifeTech)*, March 2022.

[4] https://www.kidrobot.com/pages/wtf.

[5] H. Sun and S. A. Jafar. The capacity of private information retrieval. *IEEE Trans. on Info. Theory*, 63(7):4075–4088, July 2017.

[6] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–981, 1998.

[7] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*, 64(10):6842–6862, October 2018.

[8] N. Shah, K. Rashmi, and K. Ramchandran. One extra bit of download ensures perfectly private information retrieval. In *IEEE ISIT*, June 2014.

[9] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*, 64(3):1945–1956, March 2018.

[10] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat. Achieving maximum distance separable private information retrieval capacity with linear codes. *IEEE Trans. on Info. Theory*, 65(7):4243–4273, July 2019.

[11] R. Zhou, C. Tian, H. Sun, and T. Liu. Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size. *IEEE Trans. on Info. Theory*, 66(8):4904–4916, August 2020.

[12] K. Banawan and S. Ulukus. Private information retrieval through wiretap channel II: Privacy meets security. *IEEE Trans. on Info. Theory*, 66(7):4129–4149, July 2020.

[13] S. Kumar, A. Graell i Amat, E. Rosnes, and L. Senigagliesi. Private information retrieval from a cellular network with caching at the edge. *IEEE Trans. on Commun.*, 67(7):4900–4912, July 2019.

[14] T. Guo, R. Zhou, and C. Tian. On the information leakage in private information retrieval systems. *IEEE Trans. on Info. Forensics and Security*, 15:2999–3012, 2020.

[15] H. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi. Multi-server weakly-private information retrieval. *IEEE Trans. on Info. Theory*, 68(2):1197–1219, February 2022.

[16] C. Tian, H. Sun, and J. Chen. Capacity-achieving private information retrieval codes with optimal message size and upload cost. *IEEE Trans. on Info. Theory*, 65(11):7613–7627, November 2019.

[17] H. Yang, W. Shin, and J. Lee. Private information retrieval for secure distributed storage systems. *IEEE Trans. on Info. Forensics and Security*, 13(12):2953–2964, December 2018.

[18] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*, 65(2):1206–1219, February 2019.

[19] Z. Jia, H. Sun, and S. A. Jafar. Cross subspace alignment and the asymptotic capacity of $X$-secure $T$-private information retrieval. *IEEE Trans. on Info. Theory*, 65(9):5783–5798, September 2019.

[20] Z. Jia and S. A. Jafar. $X$-secure $T$-private information retrieval from MDS coded storage with Byzantine and unresponsive servers. *IEEE Tran. on Info. Theory*, 66(12):7427–7438, December 2020.

[21] M. A. Attia, D. Kumar, and R. Tandon. The capacity of private information retrieval from uncoded storage constrained databases. *IEEE Trans. on Info. Theory*, 66(11):6617–6634, November 2020.

[22] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus. The capacity of private information retrieval from heterogeneous uncoded caching databases. *IEEE Trans. on Info. Theory*, 66(6):3407–3416, June 2020.

[23] N. Raviv, I. Tamo, and E. Yaakobi. Private information retrieval in graph-based replication systems. *IEEE Trans. on Info. Theory*, 66(6):3590–3602, June 2020.

[24] K. Banawan and S. Ulukus. Asymmetry hurts: Private information retrieval under asymmetric traffic constraints. *IEEE Trans. on Info. Theory*, 65(11):7628–7645, November 2019.

[25] H. Sun and S. A. Jafar. The capacity of private computation. *IEEE Trans. on Info. Theory*, 65(6):3880–3897, June 2019.

[26] K. Banawan and S. Ulukus. Noisy private information retrieval: On separability of channel coding and information retrieval. *IEEE Trans. on Info. Theory*, 65(12):8232–8249, December 2019.

[27] Z. Chen, Z. Wang, and S. A. Jafar. The asymptotic capacity of private search. *IEEE Trans. on Info. Theory*, 66(8):4709–4721, August 2020.

[28] S. Vithana, K. Banawan, and S. Ulukus. Semantic private information retrieval. *IEEE Trans. on Info. Theory*, 68(4):2635–2652, April 2022.

[29] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information. *IEEE Trans. on Info. Theory*, 66(4):2032–2043, April 2020.

[30] S. Kadhe, A. Heidarzadeh, et al. Single-server private information retrieval schemes are equivalent to locally recoverable coding schemes. *IEEE Jour. on Selected Areas in Info. Theory*, 2(1):391–402, 2021.

[31] A. Heidarzadeh, B. Garcia, et al. On the capacity of single-server multi-message private information retrieval with side information. In *Allerton Conference*, pages 180–187, October 2018.

[32] S. Li and M. Gastpar. Single-server multi-message private information retrieval with side information. In *Allerton Conference*, October 2018.

[33] R. Tandon. The capacity of cache aided private information retrieval. In *Allerton Conference*, October 2017.

[34] Y.-P. Wei, K. Banawan, and S. Ulukus. Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits. *IEEE Jour. on Selected Areas in Commun.*, 36(6):1126–1139, June 2018.

[35] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. *IEEE Trans. on Info. Theory*, 65(5):3215–3232, May 2019.

[36] Y.-P. Wei, K. Banawan, and S. Ulukus. The capacity of private information retrieval with partially known private side information. *IEEE Trans. on Info. Theory*, 65(12):8222–8231, December 2019.

[37] Y.-P. Wei and S. Ulukus. The capacity of private information retrieval with private side information under storage constraints. *IEEE Trans. on Info. Theory*, 66(4):2023–2031, April 2020.

[38] Z. Chen, Z. Wang, and S. Jafar. The capacity of $T$-private information retrieval with private side information. *IEEE Trans. on Info. Theory*, 66(8):4761–4773, August 2020.

[39] M. J. Siavoshani, S. P. Shariatpanahi, and M. A. Maddah-Ali. Private information retrieval for a multi-message scenario with private side information. *IEEE Trans. on Commun.*, 69(5):3235–3244, May 2021.

[40] H. Sun and S. A. Jafar. The capacity of symmetric private information retrieval. *IEEE Trans. on Info. Theory*, 65(1):322–329, January 2019.

[41] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin. Protecting data privacy in private information retrieval schemes. In *ACM Symposium on Theory of Computing*, May 1998.

[42] Z. Wang, K. Banawan, and S. Ulukus. Private set intersection: A multi-message symmetric private information retrieval perspective. *IEEE Trans. on Info. Theory*, 68(3):2001–2019, March 2022.

[43] Z. Wang, K. Banawan, and S. Ulukus. Multi-party private set intersection: An information-theoretic approach. *IEEE Jour. on Selected Areas in Info. Theory*, 2(1):366–379, March 2021.

[44] Z. Wang and S. Ulukus. Symmetric private information retrieval with user-side common randomness. In *IEEE ISIT*, July 2021.

[45] Y. Zhou, Q. Wang, H. Sun, and S. Fu. The minimum upload cost of symmetric private information retrieval. In *IEEE ISIT*, June 2020.

[46] Z. Wang and S. Ulukus. Communication cost of two-database symmetric private information retrieval: A conditional disclosure of multiple secrets perspective. In *IEEE ISIT*, June 2022.

[47] Q. Wang and M. Skoglund. On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3183–3197, May 2019.

[48] Q. Wang, H. Sun, and M. Skoglund. Symmetric private information retrieval with mismatched coded messages and randomness. In *IEEE ISIT*, July 2019.

[49] Q. Wang and M. Skoglund. Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers. *IEEE Trans. on Info. Theory*, 65(8):5160–5175, August 2019.

[50] J. Zhu, Q. Yan, and X. Tang. Multi-user blind symmetric private information retrieval from coded servers. *IEEE Jour. on Selected Areas in Commun.*, 40(3):815–831, 2022.

[51] J. Cheng, N. Liu, and W. Kang. The capacity of symmetric private information retrieval under arbitrary collusion and eavesdropping patterns. Available at arXiv:2010.08249.

[52] M. O. Rabin. How to exchange secrets with oblivious transfer. In *IACR Eprint archive*, 2005. Originally published as: Technical Report TR-81, Aiken Computation Lab, Harvard University, 1981.

[53] S. Even, O. Goldreich, and A. Lempel. A randomized protocol for signing contracts. *Communications of the ACM*, 28(6):637–647, June 1985.