An Information-Theoretic Scheme for Multi-Party Private Set Intersection

Zhusheng Wang[1], Karim Banawan[2], and Sennur Ulukus[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
[2]Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

***Abstract*—We investigate the problem of multi-party private set intersection (MP-PSI). In MP-PSI, there are M parties, each storing a data set $\mathcal{P}_i$ over N_i replicated and non-colluding databases, and we want to calculate the intersection of the data sets $\cap_{i=1}^{M}\mathcal{P}_i$ without leaking any information beyond the set intersection to any of the parties. For a specific communication protocol, we propose an information-theoretic scheme for MP-PSI based on the connection between the PSI problem and the multi-message symmetric private information retrieval (MM-SPIR) problem. Our scheme is a non-trivial generalization of the 2-party PSI scheme as it needs an intricate design of the shared common randomness. Interestingly, our scheme does not incur any penalty due to the more stringent privacy constraints in the MP-PSI problem compared to the 2-party PSI problem.**

I. Introduction

The two-party private set intersection (PSI) problem refers to a classical privacy problem, which is introduced in [1]. In its classical setting, two parties, each possessing a data set, need to calculate common elements that lie in both data sets. This calculation is performed in such a way that neither party reveals anything to the counterparty except for the elements in the intersection. Ubiquitous schemes have been investigated to tackle the PSI problem using cryptographic techniques; see for example [2]. Reference [3] formulates the 2-party PSI problem from an information-theoretic perspective. Interestingly, [3] explores an intriguing connection between the PSI problem and the private information retrieval (PIR) problem [4]. Specifically, [3] investigates the PSI determination using the multi-message symmetric PIR (MM-SPIR) procedure. Surprisingly, under some technical conditions, MM-SPIR proves to be the most-efficient 2-party PSI protocol under absolute privacy guarantees. The efficiency is measured by the total download cost, which is the number of bits needed to be downloaded to calculate the set intersection at one of the parties. The optimality proof builds on the rich literature of characterizing the fundamental limits of PIR and related problems, starting with the seminal work of Sun-Jafar [5]. Further fundamental limits of many variations of the PIR problem have been investigated; see [6]–[48] for example.

The MM-SPIR framework to solve the PSI problem in [3], however, works only for 2-party PSI. This is because the original PIR problem (and the SPIR problem) involves two parties, the user and the server(s). Unlike PIR, the PSI problem may involve more than two parties. For instance, consider a company which sells a certain product (e.g., shoes), a company which makes ads and posts them at various web-hosts, and another company which is a web-host that hosts ads. All of these parties have their individual lists of clicks that they wish to keep private, but may want to compute the intersection, i.e., actual customers who bought the product from the company after seeing an ad produced by the ad company hosted at the particular web-host company, to determine the effectiveness of the ad company and the web-host company. Note that pairwise intersections leak additional information beyond the three-way intersection. Hence, the MP-PSI problem is a non-trivial extension of the 2-party PSI as it cannot be implemented via multiple pair-wise 2-party PSI. In the computational privacy literature, the first MP-PSI achievable scheme was proposed by Freedman et al. [1]. Though considerable progress has been made in the construction of various 2-party PSI schemes, only few works exist for MP-PSI schemes [49]–[51].

In this paper, we investigate the MP-PSI problem from an information-theoretic perspective. In MP-PSI, there are M independent parties. The ith party is denoted by P_i, for $i = 1, \cdots, M$. Each party possesses a data set $\mathcal{P}_i$. The elements of all data sets are picked from a finite set $\mathbb{S}_K$ with cardinality $|\mathbb{S}_K| = K$. The data set $\mathcal{P}_i$ is stored in N_i replicated and non-colluding databases. We aim at privately calculating the intersection $\mathcal{P} = \cap_{i=1}^{M}\mathcal{P}_i$ in such a way that no party can learn any information beyond the intersection $\mathcal{P}$. Inspired by the classical achievable scheme in [1], [52], we focus on a specific communication strategy between the parties in this work. In particular, we assume that the parties agree on choosing one of them as a *leader* party, while the remaining parties act as *client* parties. The leader party initiates the MP-PSI determination protocol by generating and submitting queries to the client parties in one round. At the clients' side and before MP-PSI, the clients are allowed to generate and share common randomness. Communication between any two client parties is not allowed during the protocol. The client parties respond truthfully to the leader's queries without leaking information about the elements outside $\mathcal{P}$ with the aid of the assigned common randomness.

In this paper, we formulate the MP-PSI problem from an information-theoretic perspective. We show that MP-PSI can also be recast as a MM-SPIR problem extending [3]. This can be done by mapping the data sets at each party into an *inci-*

This work was supported by NSF Grants CCF 17-13977 and ECCS 18-07348.

dence vector to facilitate the MM-SPIR of the elements that belong to the leader's data set. We propose a novel achievable scheme for MP-PSI. The structure of the queries is the same as the SPIR queries in [10]. Despite the similarity of the queries, the answering strings in MP-PSI are fundamentally different. This is due to the fact that the leader party cannot perform $M-1$ pair-wise PSI operation to calculate $\mathcal{P}=\cap_{i=1}^{M}\mathcal{P}_i$ without leaking extra information about the individual intersections $\mathcal{P}_M\cap\mathcal{P}_i$, $i=1,\cdots,M-1$. We design an intricate protocol of generating and sharing the common randomness among the databases of the parties. By correlating some of the components of the common randomness in a specific way, we show that the leader party can reliably identify the elements in $\mathcal{P}$, but nothing beyond it. The download cost of our scheme is $\min_{t\in\{1,\cdots,M\}}\sum_{i\in\{1,\cdots M\}\setminus t}\left\lceil\frac{|\mathcal{P}_t|N_i}{N_i-1}\right\rceil$. This implies that there is no penalty due to strengthening the clients' privacy constraint in MP-PSI compared to the PSI problem. Our achievable download cost scales linearly with the cardinality of the leader set, which outperforms the best-known MP-PSI scheme, which scales with the sum of the cardinalities of the data sets [49] with added advantage of simpler implementation and providing absolute privacy guarantees. We only provide sketches of the proofs here due to space limitations; proof details, examples and figures can be found in [53].

II. Problem Formulation

There are M independent parties, denoted by P_i, $i=1,2,\cdots,M$. The ith party possesses a data set $\mathcal{P}_i$ for $i\in[M]$, where $[M]$ denotes integers from 1 to M. The data set $\mathcal{P}_i$ is stored within N_i replicated and non-colluding databases. The elements in each data set $\mathcal{P}_i$ are picked independently from a finite set $\mathbb{S}_K$ of cardinality K with an arbitrary statistical distribution. We assume that the cardinality of data set $|\mathcal{P}_i|$ is public knowledge.

The ith party maps $\mathcal{P}_i$ into a searchable list to facilitate PIR. The party P_i constructs an incidence vector X_i, such that

$$X_{i,j}=\begin{cases}1, & j\in\mathcal{P}_i\\ 0, & j\notin\mathcal{P}_i\end{cases}\tag{1}$$

where $X_{i,j}$ is the jth element of X_i for all $j\in\mathbb{S}_K$. The MP-PSI determination is performed over X_i instead of $\mathcal{P}_i$.

We consider the leader-to-clients communication model. Specifically, the parties agree on a *leader* party, which sends queries to the remaining parties and eventually calculates $\cap_{i=1}^{M}\mathcal{P}_i$. The remaining parties are called *client* parties. The communication between any two client parties is not allowed in our protocol. The leader party P_M (without loss of generality) sends the query $Q_{i,j}^{[\mathcal{P}_M]}$ to the jth database in the client party P_i for all $i\in[M-1]$ and $j\in[N_i]$. Since P_M has no prior information about $\mathcal{P}_i$, the queries are independent of $\mathcal{P}_i$,

$$I(Q_{i,j}^{[\mathcal{P}_M]};\mathcal{P}_i)=0,\quad\forall i\in[M-1],\forall j\in[N_i]\tag{2}$$

The jth database associated with the client party P_i responds truthfully with an answer $A_{i,j}^{[\mathcal{P}_M]}$ for all $i\in[M-1]$, and $j\in[N_i]$. The answer is a deterministic function of the query $Q_{i,j}^{[\mathcal{P}_M]}$, the data set $\mathcal{P}_i$, and some common randomness $\mathcal{R}_{i,j}$ that is available to the jth database of P_i. Thus,

$$H(A_{i,j}^{[\mathcal{P}_M]}|Q_{i,j}^{[\mathcal{P}_M]},\mathcal{P}_i,\mathcal{R}_{i,j})=0,\ \forall i\in[M-1],\forall j\in[N_i]\tag{3}$$

Denote all the queries generated by P_M as $Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}$ and all the answers collected by P_M as $A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}$, i.e.,

$$Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}=\left\{Q_{i,j}^{[\mathcal{P}_M]}:i\in[M-1],j\in[N_i]\right\}\tag{4}$$

$$A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}=\left\{A_{i,j}^{[\mathcal{P}_M]}:i\in[M-1],j\in[N_i]\right\}\tag{5}$$

Three formal requirements are needed to be satisfied for the MP-PSI problem: First, the leader party P_M should be able to reliably determine the intersection $\mathcal{P}=\cap_{i=1}^{M}\mathcal{P}_i$ based on $Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}$, $A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}$ and the knowledge of $\mathcal{P}_M$ without knowing $|\mathcal{P}|$ in advance. The reliability constraint is given by:

$$H(\mathcal{P}|Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]},A_{1:M-1,1:N_i}^{[\mathcal{P}_M]},\mathcal{P}_M)=0\tag{6}$$

Second, the queries sent by P_M should not leak any information about $\mathcal{P}_M$. Thus, $\mathcal{P}_M$ should be independent of all the information available in the jth database of P_i for all $i\in[M-1]$ and $j\in[N_i]$. Thus, the leader's privacy constraint,

$$I(\mathcal{P}_M;Q_{i,j}^{[\mathcal{P}_M]},A_{i,j}^{[\mathcal{P}_M]},\mathcal{P}_i,\mathcal{R}_{i,j})=0,\ i\in[M-1],j\in[N_i]\tag{7}$$

Third, client's privacy requires that the leader party does not learn any information other than the intersection $\mathcal{P}$ from the collected answer strings. Let $X_{i,\bar{\mathcal{P}}}$ be the set of elements in X_i that do not belong to $\mathcal{P}$, i.e., $X_{i,\bar{\mathcal{P}}}=\{X_{i,k}:k\in\bar{\mathcal{P}}\}$. Hence, the set $\tilde{X}=\left\{X_{1,\bar{\mathcal{P}}},\cdots,X_{M-1,\bar{\mathcal{P}}}\right\}=\left\{X_{1,k},\cdots,X_{M-1,k},k\in\bar{\mathcal{P}}\right\}$ should be independent of all the information available in P_M. Note that if an element in $\mathcal{P}_M$ is not in the intersection $\mathcal{P}$, the leader party is supposed to conclude that not all the client parties contain this element simultaneously. On the basis of this fact, we define a new set $X_{\bar{\mathcal{P}}}=\left\{\tilde{X}:X_{1,k}+\cdots+X_{M-1,k}<M-1,\forall k\in\mathcal{P}_M\cap\bar{\mathcal{P}}\right\}\Big\}$, we have the following client's privacy constraint,

$$I(X_{\bar{\mathcal{P}}};Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]},A_{1:M-1,1:N_i}^{[\mathcal{P}_M]},\mathcal{P}_M)=0\tag{8}$$

For a given K, M, N_i, an achievable MP-PSI scheme satisfies the reliability constraint (6), the leader's privacy constraint (7) and the client's privacy constraint (8). The efficiency of an achievable MP-PSI scheme is measured by its download cost which is the number of downloaded bits (denoted by D) in order to compute the intersection $\mathcal{P}$. The optimal download cost is $D^*=\inf D$ over all MP-PSI achievability schemes.

III. Main Result

Theorem 1 *In the MP-PSI problem with M parties with data sets $\mathcal{P}_i$, assuming that the parties follow a leader-to-clients communication policy, if the data sets are stored within N_i replicated and non-colluding databases for $i=1,\cdots,M$, then the optimal download cost, D^*, is upper bounded by*

$$D^*\leq\min_{t\in\{1,\cdots,M\}}\sum_{i\in\{1,\cdots M\}\setminus t}\left\lceil\frac{|\mathcal{P}_t|N_i}{N_i-1}\right\rceil\tag{9}$$

Remark 1 *The download cost of our scheme is equal to the sum of the download costs of $M-1$ pair-wise PSI schemes. Hence, there is no penalty incurred due to the stringent clients' privacy constraint compared with 2-party PSI.*

Remark 2 *Our scheme is private in the information-theoretic (absolute) sense and is fairly simple to implement. A drawback of our approach is that it needs multiple replicated non-colluding databases as in the 2-party PSI problem in [3]; otherwise, our scheme is infeasible if $N_i = 1$ for any party other than the leader party.*

Remark 3 *Comparing our result with the most closely related information-theoretic MP-PSI schemes [50], our scheme outperforms theirs in terms of the communication cost as our download cost is linear in both the number of parties M and the size of the sets p, assuming that $|\mathcal{P}_i| = p$ for all $i = 1, \cdots, M$ in contrast of $O(M^4p^2)$ in [50]. We note, however, that the work [50] allows for potential distrust between the parties, which is not considered in this work.*

IV. Representative Example: 3 Parties with 3 Databases Each ($M = 3$ with $N_1 = N_2 = N_3 = 3$)

We illustrate our scheme by presenting the following example with $M = 3$ parties, each possessing $N_i = 3$ databases. Assume that each party stores an independently generated set $\mathcal{P}_i \subseteq \mathbb{S}_K$, where $\mathbb{S}_K = \{1, 2, 3, 4\}$. Specifically, we assume that $\mathcal{P}_1 = \{1, 2\}$, $\mathcal{P}_2 = \{1, 3\}$, and $\mathcal{P}_3 = \{1, 4\}$. We aim at reliably calculating the intersection $\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3 = \{1\}$ without leaking any further information to any of the parties. Without loss of generality, we pick P_3 to be the leader party. The remaining parties P_1, P_2 are referred to as clients.

We map the sets into the corresponding incidence vectors,

$$\mathcal{P}_1 = \{1, 2\} \quad \Rightarrow \quad X_1 = [1\ 1\ 0\ 0]^T \tag{10}$$

$$\mathcal{P}_2 = \{1, 3\} \quad \Rightarrow \quad X_2 = [1\ 0\ 1\ 0]^T \tag{11}$$

$$\mathcal{P}_3 = \{1, 4\} \quad \Rightarrow \quad X_3 = [1\ 0\ 0\ 1]^T \tag{12}$$

The parties agree on a finite field $\mathbb{F}_L$, where L is a prime number such that $L \geq M$. We pick $L = 3$ in our case.

The leader party P_3 initiates the MP-PSI protocol by sending queries $Q_{i,j}^{[\mathcal{P}_3]}$ for $i \in [2]$ and $j \in [3]$. The queries aim at *privately retrieving* the messages $X_{1,1}$, $X_{1,4}$ and $X_{2,1}$, $X_{2,4}$ using the SPIR retrieval scheme in [10]. More specifically, let h_k, where $k = 1, \cdots, 4$, be a random variable picked uniformly and independently from $\mathbb{F}_3$, then, for client party P_1, the queries sent from the leader party P_3 are,

$$Q_{1,1}^{[\mathcal{P}_3]} = [h_1\ \ h_2\ \ h_3\ \ h_4]^T \tag{13}$$

$$Q_{1,2}^{[\mathcal{P}_3]} = [h_1 + 1\ \ h_2\ \ h_3\ \ h_4]^T \tag{14}$$

$$Q_{1,3}^{[\mathcal{P}_3]} = [h_1\ \ h_2\ \ h_3\ \ h_4 + 1]^T \tag{15}$$

For client party P_2, the leader party submits the same set of queries, i.e., $Q_{2,j}^{[\mathcal{P}_3]} = Q_{1,j}^{[\mathcal{P}_3]}$ for all $j \in [3]$.

Originally in 2-party PSI, the client databases obtain the inner product of X_i and $Q_{i,j}^{[\mathcal{P}_3]}$ and add a common randomness. In MP-PSI, however, we note that applying the answering strategy of [3], [10] compromises the clients' privacy constraint (8). This is due to the fact that the leader, in this case, can decode that $X_{1,4} = 0$ and $X_{2,4} = 0$ and not only the intersection $\cap_{i=1,2,3}\ \mathcal{P}_i$. Consequently, the clients' databases need to share intricate common randomness prior to the retrieval phase to prevent that. To that end, the client parties generate and/or share the following randomness:

1) *Local randomness:* The local randomness $s_i \sim \text{uniform}\{0, 1, 2\}$ is generated independently from all sets and all other randomness sources. s_i is shared among all the databases belonging to the ith client party and not shared with other parties. s_i acts as the common randomness needed for SPIR [10].
2) *Individual correlated randomness:* This is possessed by each client's database, and is denoted by the random variables $t_{i,j}$. This prevents the leader party from decoding $X_{1,4}$, and $X_{2,4}$. The random variables $t_{i,j}$ need to be correlated such that their effect can be removed if $X_{i,j}$ belongs to the intersection. We choose $t_{1,1} = t_{2,1} = 0$. Furthermore, the jth database of P_1 generates $t_{1,j}$ randomly and sends it to the jth database of P_2 to calculate $t_{2,j}$ as,

$$t_{1,j} \sim \text{uniform}\{0, 1, 2\}, \quad j = 2, 3 \tag{16}$$

$$t_{1,j} + t_{2,j} = 1, \quad j = 2, 3 \tag{17}$$

This randomness is added to each response as well.
3) *Global randomness:* The global randomness $c \sim \text{uniform}(\mathbb{F}_3 \setminus \{0\})$ is generated randomly and independently of all other randomness sources. The global randomness is shared among all databases of all client parties P_1 and P_2. The global randomness is used as a multiplier to the responses.

The jth database of the ith party responds to $Q_{i,j}^{[\mathcal{P}_3]}$ as follows,

$$A_{i,j}^{[\mathcal{P}_3]} = c(X_i^T Q_{i,j}^{[\mathcal{P}_3]} + s_i + t_{i,j}), \quad i = 1, 2,\ j = 1, 2, 3 \tag{18}$$

Hence, the answer strings from P_i can be written as,

$$A_{i,1}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{i,k} + s_i\right) \tag{19}$$

$$A_{i,2}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{i,k} + X_{i,1} + s_i + t_{i,2}\right) \tag{20}$$

$$A_{i,3}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{i,k} + X_{i,4} + s_i + t_{i,3}\right) \tag{21}$$

1) Reliability: The leader party subtracts $A_{i,1}^{[\mathcal{P}_3]}$ for $i = 1, 2$ from the remaining answer strings. Denote the result of subtraction related to the jth element at P_i by $Z_{i,j}$. Thus,

$$Z_{i,1} = c(X_{i,1} + t_{i,2}) = A_{i,2}^{[\mathcal{P}_3]} - A_{i,1}^{[\mathcal{P}_3]} \tag{22}$$

$$Z_{i,4} = c(X_{i,4} + t_{i,3}) = A_{i,3}^{[\mathcal{P}_3]} - A_{i,1}^{[\mathcal{P}_3]} \tag{23}$$

Now, let E_j be an indicator of having the jth element in $\mathbb{S}_K$ in the intersection $\cap_{i=1,2,3}\ \mathcal{P}_i$, such that $E_j = 0$ if and only

if $j \in \cap_{i=1,2,3} \mathcal{P}_i$. To that end, define E_j as the modulo-L sum of $Z_{i,j}$ along all clients, i.e., $E_j = \sum_{i=1}^{M-1} Z_{i,j}$. Looking deeper at E_1, we note that,

$$E_1 = c(X_{1,1}+X_{2,1}+t_{1,2}+t_{2,2}) = c(X_{1,1}+X_{2,1}+1) \quad (24)$$

where $t_{1,2} + t_{2,2} = 1$ by the construction of the individual correlated randomness. Therefore, $E_1 = 0$ if and only if $X_{1,1} = 1$ and $X_{2,1} = 1$ simultaneously irrespective of the value of c and the leader party verifies that $\{1\} \subseteq \cap_{i=1,2,3}\mathcal{P}_i$.

On the other hand, when P_3 calculates E_4,

$$E_4 = Z_{1,4} + Z_{2,4} = c(X_{1,4} + X_{2,4} + 1) \neq 0 \quad (25)$$

Thus, $\cap_{i=1,2,3} \mathcal{P}_i = \{1\}$ and does not include 4.

2) Leader's Privacy: Similar to [10], each element in the queries is uniformly distributed over the finite field $\mathbb{F}_3$. Hence, no information about $\mathcal{P}_3$ is leaked from the queries.

3) Client's Privacy: No information is leaked about $\overline{\mathcal{P}_1 \cap \mathcal{P}_3}$ or $\overline{\mathcal{P}_2 \cap \mathcal{P}_3}$ due to s_1 and s_2, respectively. Moreover, if $E_4 = 1$, $\mathbb{P}(X_{1,4} + X_{2,4} = 0) = \mathbb{P}(X_{1,4} + X_{2,4} = 1) = \frac{1}{2}$ because c is uniformly distributed over 1 and 2 and the sum $t_{1,3} + t_{2,3} = 1$ by construction. The conclusion is exactly the same when E_4 equals 2. Thus, the only information that P_3 can obtain for the element 4 is that client parties P_1 and P_2 do not contain it at the same time. Hence, c is used such that the leader party P_3 does not know the value of $X_{1,4} + X_{2,4}$.

4) Download Cost: In our example, the leader party P_3 downloads $N_i = |\mathcal{P}_M| + 1$ symbols from each client party. Hence, the total download cost is $D = (M-1)(|\mathcal{P}_M|+1) = 6$.

V. Achievability Proof

We describe our general achievable scheme for arbitrary M, $|\mathcal{P}_i|$, N_i, for $i \in [M]$. The leader's querying policy is based on the SPIR scheme presented in [10]. The novel aspect of the scheme is the intricate design of generating and sharing common randomness among the clients' databases in such a way that the leader party cannot learn anything but $\cap_{i=1}^{M} \mathcal{P}_i$.

A. General Achievability Scheme

1) *Initialization:* The parties agree on a retrieval finite field $\mathbb{F}_L$ such that, $L = \min\{L \geq M : L \text{ is a prime}\}$. The parties agree on a leader P_{t^*} such that:

$$t^* = \arg \min_{t \in \{1,\cdots,M\}} \sum_{i \neq t} \left\lceil \frac{|\mathcal{P}_t| N_i}{N_i - 1} \right\rceil \quad (26)$$

We assume that $t^* = M$ and $\mathcal{P}_{t^*} = \mathcal{P}_M = \{Y_1, Y_2, \cdots, Y_R\}$ with cardinality $|\mathcal{P}_M| = R$.

2) *Query generation:* The leader party P_M independently and uniformly generates κ random vectors $\{\mathbf{h}_\ell\}_{\ell=1}^{\kappa}$, where $\kappa = \max_{i \in \{1,\cdots,M-1\}} \left\lceil \frac{|\mathcal{P}_M|}{N_i - 1} \right\rceil$. The vector $\mathbf{h}_\ell$ is picked uniformly from $\mathbf{F}_L^K$ such that,

$$\mathbf{h}_\ell = [h_\ell(1) \quad h_\ell(2) \quad \cdots \quad h_\ell(K)] \quad (27)$$

Denote $\eta_i = \left\lceil \frac{|\mathcal{P}_M|}{N_i - 1} \right\rceil$, and let $\mathcal{P}_M^{\ell_i} = \{Y_1^{\ell_i}, Y_2^{\ell_i}, \cdots Y_{N_i-1}^{\ell_i}\}$. The leader party P_M submits η_i random vectors from $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_\kappa\}$ to the first database of the ith client party as queries. Each submitted random vector can be reused in the remaining $N_i - 1$ databases to retrieve $N_i - 1$ symbols. This can be done by adding 1 to the positions corresponding to the desired symbols. More specifically, assume that $\mathcal{P}_M = \cup_{\ell_i=1}^{\eta_i} \mathcal{P}_M^{\ell_i}$, where $\mathcal{P}_M^{\ell_i} \subseteq \mathcal{P}_M$ are disjoint partitions of $\mathcal{P}_M$ such that $|\mathcal{P}_M^{\ell_i}| = N_i - 1$, then for $i = 1, 2, \cdots, M-1$, the query structure is given by:

$$Q_{i,1}^{[\mathcal{P}_M^{\ell_1}]} = [h_1(1) \quad h_1(2) \quad \cdots \quad h_1(K)] \quad (28)$$

$$Q_{i,2}^{[\mathcal{P}_M^{\ell_1}]} = [h_1(1) \ \cdots \ h_1(Y_1^{\ell_1}) + 1 \ \cdots \ h_1(K)] \quad (29)$$

$$\vdots$$

$$Q_{i,N_i}^{[\mathcal{P}_M^{\ell_1}]} = [h_1(1) \ \cdots \ h_1(Y_{N_i-1}^{\ell_1}) + 1 \cdots h_1(K)] \quad (30)$$

$$\vdots$$

$$Q_{i,1}^{[\mathcal{P}_M^{\eta_i}]} = [h_{\eta_i}(1) \quad h_{\eta_i}(2) \quad \cdots \quad h_{\eta_i}(K)] \quad (31)$$

$$\vdots$$

$$Q_{i,N_i}^{[\mathcal{P}_M^{\eta_i}]} = [h_{\eta_i}(1) \cdots h_{\eta_i}(Y_{N_i-1}^{\eta_i}) + 1 \cdots h_{\eta_i}(K)] \quad (32)$$

3) *Common randomness generation:* The clients need to generate and share common randomness. Specifically,

- *Local randomness:* This is denoted by $\mathbf{s}_i = [s_i(1)s_i(2)s_i(\eta_i)]$, where $s_i(\ell) \sim \text{uniform}(\mathbb{F}_L)$. The local randomness $\mathbf{s}_i$ is shared between the databases associated with P_i. Note that each database uses a different element from $\mathbf{s}_i$ for each submitted query.
- *Individual correlated randomness:* The jth database associated with the ith client possesses an individual randomness $\mathbf{t}_{i,j} = [t_{i,j}(1) \ t_{i,j}(2) \ t_{i,j}(\eta_i)]$ for $i \in [M-1]$, and $j \in [N_i]$. $t_{i,1} = 0$ for all i. For $i \in [M-2]$, the vector $\mathbf{t}_{i,j}$ is independently and uniformly picked from $\mathbf{F}_L^{\eta_i}$. The client P_{M-1} generates a correlated randomness. For simplicity, let us (re)denote the individual randomness components by $\tilde{t}_{i,k}$, where i is the index of the client party and $k = 1, 2, \cdots, R$ is just a monotonically increasing index of the randomness component used within the databases 2 to N_i of the ith client. Thus, $\tilde{t}_{i,1} = t_{i,2}(1)$, $\tilde{t}_{i,1} = t_{i,2}(2), \cdots, \tilde{t}_{i,R} = t_{i,N_i}(\eta_i)$. With this re-definition, the client P_{M-1} calculates its individual randomness as, for $j = 1, 2, \cdots, R$

$$\tilde{t}_{M-1,j} = L - (M-1) - \sum_{i=1}^{M-2} \tilde{t}_{i,j} \quad (33)$$

This ensures that the individual randomness are correlated such that $\sum_{i=1}^{M-1} \tilde{t}_{i,j} = L - (M-1)$. The individual randomness is added to the responses.
- *Global randomness:* This is denoted by $c \sim \text{uniform}(\mathbb{F}_L \setminus \{0\})$ and is shared among all the databases at all clients.

4) *Response generation:* The answer string of the jth database associated with the ith client to retrieve one of the elements of the partition $\mathcal{P}_M^{\ell_i}$, $A_{i,j}^{[\mathcal{P}_M^{\ell_i}]}$, is given by,

$$A_{i,j}^{[\mathcal{P}_M^{\ell_i}]} = c\left(X_i^T Q_{i,j}^{[\mathcal{P}_M^{\ell_i}]} + s_i(\ell_i) + t_{i,j}(\ell_i)\right) \quad (34)$$

B. Download Cost, Reliability, Leader's and Clients' Privacy

1) Download cost: By observing the queries associated with the MP-PSI scheme in the previous section, one can note that the desired symbols are divided into $\eta_i = \left\lceil \frac{|\mathcal{P}_M|}{N_i - 1} \right\rceil$ subsets. Each subset consists of $N_i - 1$ desired symbols. The leader needs to download 1 bit from all N_i databases to query the entire subset, as the leader downloads useless random linear combination of the contents from the first database. Hence,

$$D = \sum_{i=1}^{M-1} N_i \eta_i = \sum_{i=1}^{M-1} \left\lceil \frac{|\mathcal{P}_M| N_i}{N_i - 1} \right\rceil \quad (35)$$

2) Reliability: We note that the answer string that is returned from database 1 is a random linear combination of the contents of the database besides the common randomness,

$$A_{i,1}^{[\mathcal{P}_M^{\ell_i}]} = c\left(\sum_{k=1}^{K} h_{\ell_i}(k) X_{i,k} + s_i(\ell_i)\right), \quad i \in [M-1] \quad (36)$$

Note that $t_{i,1} = 0$ by construction. The leader subtracts this response from each response that belongs to the same partition. Denote the subtraction result at the ith client that contains the element $X_{i,k}$ by $Z_{i,k}$, hence,

$$Z_{i,k} = c(X_{i,k} + \tilde{t}_{i,k}) = A_{i,j^*}^{[\mathcal{P}_M^{\ell_i}]} - A_{i,1}^{[\mathcal{P}_M^{\ell_i}]}, \quad k \in \mathcal{P}_M^{\ell_i} \quad (37)$$

for some unique j^* that $A_{i,j^*}^{[\mathcal{P}_M^{\ell_i}]}$ is a response of the query that adds 1 to the kth position of the query vector. In particular, for the special case of $N_i = |\mathcal{P}_i| + 1$ for all $i = 1, \cdots, M-1$, we have $j^* = k+1$ and $\mathcal{P}_M^{\ell_i} = \mathcal{P}_M$ (one partition). Note that we used the alternative notation $\tilde{t}_{i,k}$ as it is counted in sequence.

Next, the leader constructs the intersection indicator variable E_k, where E_k is given by,

$$E_k = \sum_{i=1}^{M-1} Z_{i,k} = c\left(\sum_{i=1}^{M-1} X_{i,k} + L - (M-1)\right) \quad (38)$$

where (38) follows from the construction of the individual randomness. Now, the element $E_k = 0$ if and only if $\sum_{i=1}^{M-1} X_{i,k} = M-1$, which implies that $X_{i,k} = 1$ for all $i = 1, 2, \cdots, M-1$. Consequently, $Y_k \in \cap_{i=1}^{M} \mathcal{P}_i$ if and only if $E_k = 0$. This proves the reliability of the scheme.

3) Leader's privacy: The leader's privacy follows from the fact that the random vectors $\{\mathbf{h}_1, \cdots, \mathbf{h}_\kappa\}$ are uniformly generated over $\mathbb{F}_L^K$. Adding 1 to these vectors does not change the statistical distribution of the vector. Since the leader submits independent vectors each time it queries a database, all queries are equally likely and the leader's privacy is preserved.

4) Clients' privacy: Without loss of generality, we derive the proof of the client's privacy for the homogeneous number of databases, i.e., $N_i = R+1, \forall i \in [1 : M-1]$. In the following proof, we adopt the notation that for a random variable $\zeta_{i,j}$ indexed by two indices (i,j),

$$\zeta_{i_1:i_M, j_1:j_R} = \{\zeta_{i,j} : i \in \{i_1, \cdots, i_M\},\ j \in \{j_1, \cdots, j_R\}\} \quad (39)$$

For the proof, we need the following lemmas, whose proofs can be found in [53]. Lemma 1 shows that the effect of the local randomness is to make the response of the first database at all parties independent of $X_{\bar{\mathcal{P}}}$.

Lemma 1 *For the presented achievable scheme, we have,*

$$I(X_{\bar{\mathcal{P}}}; A_{1:M-1,1}^{[\mathcal{P}_M]} | Z_{1:M-1,Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) = 0 \quad (40)$$

Lemma 2 asserts that for $i \in [1 : M-2]$, $j \in [1 : R]$ the effect of individual randomness $t_{i,j+1}$ is to force the random variables Z_{i,Y_j} to be independent of $X_{\bar{\mathcal{P}}}$.

Lemma 2 *For the presented scheme, we have,*

$$I(X_{\bar{\mathcal{P}}}; Z_{1:M-2,Y_1:Y_R} | E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) = 0 \quad (41)$$

Lemma 3 states that indicator functions E_{Y_j} for all j do not leak any information about $X_{\bar{\mathcal{P}}}$.

Lemma 3 *For the presented scheme, we have,*

$$I(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) = 0 \quad (42)$$

Now, we are ready to show that our achievability satisfies the client's privacy constraint,

$$\begin{aligned}
&I(X_{\bar{\mathcal{P}}}; Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) \\
&= I(X_{\bar{\mathcal{P}}}; A_{1:M-1,1}^{[\mathcal{P}_M]}, Z_{1:M-1,Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) && (43)\\
&= I(X_{\bar{\mathcal{P}}}; Z_{1:M-1,Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) \\
&\quad + I(X_{\bar{\mathcal{P}}}; A_{1:M-1,1}^{[\mathcal{P}_M]} | Z_{1:M-1,Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) && (44)\\
&= I(X_{\bar{\mathcal{P}}}; Z_{1:M-1,Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) && (45)\\
&= I(X_{\bar{\mathcal{P}}}; Z_{1:M-2,Y_1:Y_R}, E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) && (46)\\
&= I(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) \\
&\quad + I(X_{\bar{\mathcal{P}}}; Z_{1:M-2,Y_1:Y_R} | E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) && (47)\\
&= I(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M) = 0 && (48)
\end{aligned}$$

VI. Conclusion

We formulated the problem of MP-PSI from an information-theoretic point of view. We investigated a specific mode of communication, namely, single round communication between the leader and clients. We proposed a novel achievable scheme for the MP-PSI problem. Our scheme hinges on a careful design and sharing of randomness between client parties prior to commencing the MP-PSI operation. Our scheme is not a straightforward extension to the 2-party PSI scheme, as applying the 2-party PSI scheme $M-1$ times leaks information beyond the intersection $\cap_{i=1}^{M} \mathcal{P}_i$. The download cost of our scheme matches the sum of download cost of pair-wise PSI despite the stringent privacy constraint in the case of MP-PSI.

REFERENCES

[1] M. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 1–19. Springer, 2004.

[2] E. De Cristofaro and G. Tsudik. Practical private set intersection protocols with linear complexity. In *International Conference on Financial Cryptography and Data Security*, pages 143–159, 2010.

[3] Z. Wang, K. Banawan, and S. Ulukus. Private set intersection: A multi-message symmetric private information retrieval perspective. Available at arXiv: 1912.13501.

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

[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] H. Sun and S. A. Jafar. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. on Info. Theory*, 64(4):2361–2370, April 2018.

[7] R. Tajeddine and S. El Rouayheb. Robust private information retrieval on coded data. In *IEEE ISIT*, June 2017.

[8] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. In *IEEE ISIT*, June 2017.

[9] R. Bitar and S. El Rouayheb. Staircase-PIR: Universally robust private information retrieval. In *IEEE ITW*, pages 1–5, November 2018.

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

[11] Q. Wang, H. Sun, and M. Skoglund. Symmetric private information retrieval with mismatched coded messages and randomness. In *IEEE ISIT*, pages 365–369, July 2019.

[12] 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.

[13] 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.

[14] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. Karpuk. Private information retrieval from coded databases with colluding servers. *SIAM Journal on Applied Algebra and Geometry*, 1(1):647–664, 2017.

[15] 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.

[16] 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.

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

[18] 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.

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

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

[21] 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.

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

[23] 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.

[24] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali. Multi-message private information retrieval with private side information. In *IEEE ITW*, pages 1–5, November 2018.

[25] S. Li and M. Gastpar. Converse for multi-server single-message PIR with side information. In *IEEE CISS*, March 2020.

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

[27] M. Mirmohseni and M. A. Maddah-Ali. Private function retrieval. In *IWCIT*, pages 1–6, April 2018.

[28] Z. Chen, Z. Wang, and S. Jafar. The asymptotic capacity of private search. In *IEEE ISIT*, June 2018.

[29] M. A. Attia, D. Kumar, and R. Tandon. The capacity of private information retrieval from uncoded storage constrained databases. Available at arXiv:1805.04104v2.

[30] 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, Nov 2019.

[31] 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.

[32] K. Banawan, B. Arasli, and S. Ulukus. Improved storage for efficient private information retrieval. In *IEEE ITW*, August 2019.

[33] C. Tian. On the storage cost of private information retrieval. Available at arXiv:1910.11973.

[34] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus. The capacity of private information retrieval from decentralized uncoded caching databases. *Information*, 10, December 2019.

[35] 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.

[36] 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.

[37] K. Banawan and S. Ulukus. Private information retrieval from non-replicated databases. In *IEEE ISIT*, pages 1272–1276, July 2019.

[38] 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.

[39] H. Sun and S. A. Jafar. Optimal download cost of private information retrieval for arbitrary message length. *IEEE Trans. on Info. Forensics and Security*, 12(12):2920–2932, December 2017.

[40] Q. Wang, H. Sun, and M. Skoglund. The capacity of private information retrieval with eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3198–3214, May 2019.

[41] 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.

[42] Z. Jia, H. Sun, and S. 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.

[43] 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.

[44] 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.

[45] 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.

[46] R. G. L. D'Oliveira and S. El Rouayheb. One-shot PIR: Refinement and lifting. *IEEE Trans. on Info. Theory*, 66(4):2443–2455, April 2020.

[47] R. Tajeddine, A. Wachter-Zeh, and C. Hollanti. Private information retrieval over random linear networks. Available at arXiv:1810.08941.

[48] S. Vithana, K. Banawan, and S. Ulukus. Semantic private information retrieval. Available at arXiv: 2003.13667.

[49] C. Hazay and M. Venkitasubramaniam. Scalable multi-party private set-intersection. In *Public-Key Cryptography – PKC 2017*, pages 175–203. Springer Berlin Heidelberg, 2017.

[50] R. Li and C. Wu. An unconditionally secure protocol for multi-party set intersection. In *Applied Cryptography and Network Security*, pages 226–236, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[51] V. Kolesnikov, N. Matania, B. Pinkas, M. Rosulek, and N. Trieu. Practical multi-party private set intersection from symmetric-key techniques. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1257–1272. Association for Computing Machinery, 2017.

[52] L. Kissner and D. Song. Privacy-preserving set operations. In *Advances in Cryptology – CRYPTO 2005*, pages 241–257. Springer Berlin Heidelberg, 2005.

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