Multi-Party Private Set Intersection: An Information-Theoretic Approach

Zhusheng Wang, *Student Member, IEEE*, Karim Banawan [ID] , *Member, IEEE*, and Sennur Ulukus [ID] , *Fellow, IEEE*

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. We consider a specific communication protocol where one of the parties, called the leader party, initiates the MP-PSI protocol by sending queries to the remaining parties which are called client parties. The client parties are not allowed to communicate with each other. We propose an information-theoretic scheme that privately calculates the intersection $\cap_{i=1}^{M} \mathcal{P}_i$ with a download cost of $D = \min_{t \in \{1,...,M\}} \sum_{i \in \{1,...,M\} \setminus t} \lceil \frac{|\mathcal{P}_t| N_i}{N_i - 1} \rceil$. Similar to the 2-party PSI problem, our scheme builds 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, in terms of the download cost, 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.

Index Terms—Secure multi-party computation, private set intersection, symmetric private information retrieval.

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]–[4]. Many practical applications are tied to PSI. To see this, consider the following scenario: suppose that the national security agency (NSA) and the customs and border protection (CBP) need to check whether a specific group of suspected criminals has entered the country. The NSA has a list of suspected

criminals, while the CBP has a complete list of individuals who entered the country. Both agencies want to find the intersection between these lists. However, the NSA does not want to share its complete list of suspects, and the CBP cannot reveal the entire catalog of records either. This is a natural application for the 2-party PSI problem. Reference [5] formulates the 2-party PSI problem from an information-theoretic perspective. Interestingly, [5] explores an intriguing connection between the PSI problem and the private information retrieval (PIR) problem [6]. Specifically, [5] 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 and Jafar [7]. Further fundamental limits of many variations of the PIR problem have been investigated; see [8]–[62] for example.

The MM-SPIR framework to solve the PSI problem in [5], 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. Returning to the example involving the NSA and CBP above, suppose now that the NSA needs to narrow down the search to check whether the suspects have entered the country via a specific airline. The airline company has a list of all passengers that took its flights all over the world. The company needs to protect the privacy of its passengers as well. The problem of finding the set of suspects who entered the country via this specific airline becomes a 3-party PSI. Unfortunately, the NSA cannot just apply a 2-party PSI scheme with the airline company and the CBP, as the NSA will learn extra information than the intersection of the three lists, for example, the NSA will learn about some of its suspects who boarded a flight with this airline company but never landed in this country. Another example of 3-party PSI is related to ad clicks. 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
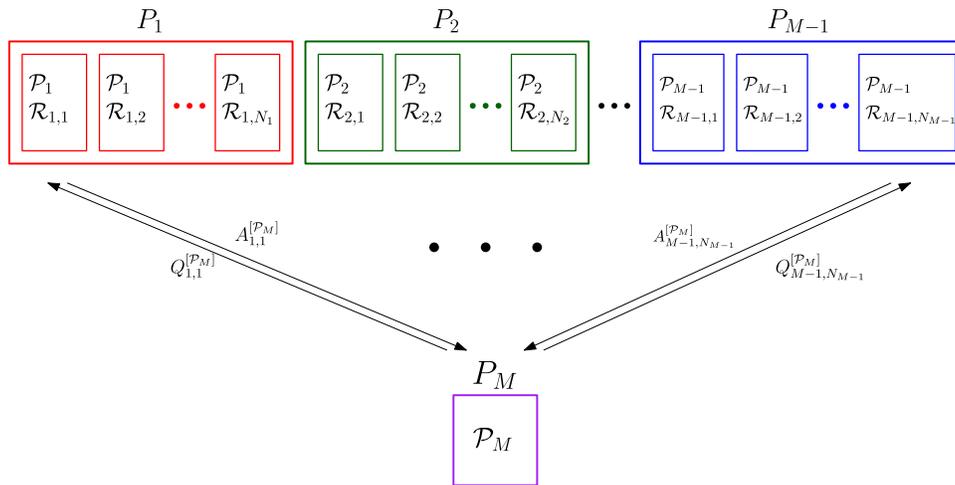
Fig. 1. Multi-party private set intersection (MP-PSI) system model.

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 again that pairwise intersections leak additional information beyond the three-way intersection. These examples motivate the multi-party PSI (MP-PSI) problem. They also illustrate that the MP-PSI is a non-trivial extension of the 2-party PSI as it cannot be implemented via multiple 2-party PSI. To make this point even stronger, consider three parties with sets $\mathcal{P}_1 = \{1, 2, 3\}$, $\mathcal{P}_2 = \{1, 2\}$ and $\mathcal{P}_3 = \{1, 3\}$. The intersection of these three sets is $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3 = \{1\}$. When any one these parties is chosen as the leader party and applies a 3-party PSI protocol, the leader party should learn only this three-way intersection. However, if the leader party applies a 2-party PSI with the two client parties, it will learn information more than the three-way intersection. For instance, if the leader party is the first party, and if it applies a 2-party PSI with the second and third parties, it will learn $\mathcal{P}_1 \cap \mathcal{P}_2 = \{1, 2\}$ and $\mathcal{P}_1 \cap \mathcal{P}_3 = \{1, 3\}$. Even though the leader party can obtain the three-way intersection by taking the intersection of these two two-way intersections, i.e., $\{1, 2\} \cap \{1, 3\} = \{1\}$, this sequential use of 2-party PSI for the 3-party PSI problem leaks further information to the leader party. For instance, the leader party learns that the second party has $\{2\}$ and the third party has $\{3\}$ further than the overall intersection $\{1\}$. Thus, 3-party PSI cannot be implemented by two 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 [63].[1]

[1]An anonymous reviewer brought to our attention the works of [64] and [65] during the review process. The work of [64] is a classical computation paper which is concerned about presenting a minimal, yet powerful, model for secure computation of a function. In [64, Appendix B], the authors present an interesting 'AND' function protocol as an extension for their original 2-input formulation. Their scheme shares some resemblance to our proposed scheme as both use additive and multiplicative common randomness. This idea is generalized in [65]. Nevertheless, our work is significantly different in terms of the problem formulation as we use multiple servers per party in our problem in contrast to an external helper in [64]. Our work is significantly different also in terms of the privacy constraints (i.e., symmetric "user" and "database" privacy constraints we have here as in SPIR) and how our scheme satisfies these constraints.

In this article, 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 $i = 1, \ldots, M$. Each party possesses a data set $\mathcal{P}_i$, where $i \in \{1, \ldots, M\}$. The elements of all data sets are picked from a finite set $\mathbb{S}_K$ with cardinality $|\mathbb{S}_K| = K$ for sufficiently large K.[2] The data set $\mathcal{P}_i$ is stored in N_i replicated and non-colluding databases. We aim at privately determining the intersection of all the M data sets, i.e., we aim at calculating $\mathcal{P} = \cap_{i=1}^M \mathcal{P}_i$ in such a way that no party can learn any information beyond the intersection $\mathcal{P}$. Inspired by the classical achievable scheme in [1], [66], we focus on a specific communication strategy between the parties in this work; see Fig. 1. 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. Without loss of generality, we pick P_M as a leader party, and then the remaining parties $P_1, \ldots, P_{M-1}$ are all client parties. The leader party P_M initiates the MP-PSI determination protocol by generating and submitting queries to the client parties. At the clients' side and before MP-PSI, the clients are allowed to generate and share common randomness (common randomness residing in the jth database of party P_i is shown by $\mathcal{R}_{i,j}$ in Fig. 1). This is motivated by the results of [5], [12], [67], which assert that using common randomness is strictly necessary to enable symmetrically private communication. Furthermore, we assume that the leader party P_M can communicate with each client party in only one round, and 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 article, we first formulate the MP-PSI problem from an information-theoretic perspective. We show that MP-PSI can also be recast as a MM-SPIR problem, which extends the formulation of the 2-party PSI problem [5]. This can be done by mapping the data sets at each party into an *incidence*

[2]Without loss of generality, one can assume that $\mathbb{S}_K = \{1, \ldots, K\}$.

vector[3] to facilitate the MM-SPIR of the elements that belong to $\mathcal{P}_M$ (the leader's data set). Next, we propose a novel achievable scheme for MP-PSI determination. The structure of the queries that the leader party submits in our scheme is the same as the SPIR queries in [12] (which can be traced back to the original work of Chor *et al.* in [6]). 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 $\mathcal{P}_M \cap \mathcal{P}_i$, $i = 1, \ldots, M - 1$, as discussed above. Note that, in general the joint intersection set $\mathcal{P} \neq \cup_{i=1}^{M-1}(\mathcal{P}_M \cap \mathcal{P}_i)$. To alleviate this problem, we design an intricate protocol of generating and sharing the common randomness among the databases of the parties. By properly incorporating the common randomness to the answer strings, we prevent the leader party from learning about the elements that lie in $\cup_{i=1}^{M-1}(\mathcal{P}_M \cap \mathcal{P}_i)$ but not in $\cap_{i=1}^{M} \mathcal{P}_i$. This constraint is referred to as the clients' privacy constraint. 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 $\min_{t \in \{1, \ldots, M\}} \sum_{i \in \{1, \ldots, M\} \setminus t} \lceil \frac{|\mathcal{P}_t| N_i}{N_i - 1} \rceil$. Note that the optimal download cost with one-round communication is $D^* = \min\{\lceil \frac{P_1 N_2}{N_2 - 1} \rceil, \lceil \frac{P_2 N_1}{N_1 - 1} \rceil\}$ in the 2-party PSI problem [5]. This means that although in MP-PSI, the clients' privacy constraint is more stringent than that in the PSI problem, we incur no penalty for it. In addition, 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 [63]. Furthermore, our scheme has an advantage of simpler implementation in addition to providing absolute (information-theoretic) privacy guarantees compared to the computationally private techniques in the literature.

II. PROBLEM FORMULATION

Consider a setting where there are M independent parties,[4] denoted by P_i, $i = 1, 2, \ldots, M$. The ith party possesses a data set $\mathcal{P}_i$ for $i \in [1 : M]$. The data set $\mathcal{P}_i$ is stored within N_i replicated and non-colluding databases.[5] Given that K is large

enough, 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.[6] More specifically, before the data sets generation, the data sets $\mathcal{P}_i$, $i \in [1 : M]$ are all random variables and they are mutually independent. We assume that the cardinality of data set $|\mathcal{P}_i|$ is public knowledge.

Motivated by the relation between 2-party PSI and MM-SPIR in [5], the ith party maps its data set $\mathcal{P}_i$ into a searchable list to facilitate PIR. To that end, the party P_i constructs an incidence vector X_i, which is a binary vector of size K associated with the data set $\mathcal{P}_i$ for all $i \in [1 : M]$, such that

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

where $X_{i,j}$ is the jth element of X_i for all $j \in \mathbb{S}_K$. Note that X_i is a sufficient statistic for $\mathcal{P}_i$ for a given K. Hence, the MP-PSI determination is performed over X_i instead of $\mathcal{P}_i$.

We consider a specific communication protocol in this work. The parties agree on a *leader* party, which sends queries to the remaining parties and eventually calculates the desired intersection $\cap_{i=1}^{M} \mathcal{P}_i$. The remaining parties are called *client* parties. Without loss of generality, assume that the leader party is P_M. The leader party P_M sends the query $Q_{i,j}^{[\mathcal{P}_M]}$ to the jth database in the client party P_i for all $i \in [1 : M - 1]$ and $j \in [1 : N_i]$. Since P_M has no information about data set $\mathcal{P}_i$ before the communication, the generated queries $Q_{i,j}^{[\mathcal{P}_M]}$ are independent from $\mathcal{P}_i$. Hence,

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

The jth database associated with the client party P_i responds truthfully with an answer $A_{i,j}^{[\mathcal{P}_M]}$ for all $i \in [1 : M - 1]$, and $j \in [1 : N_i]$. 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[7] $\mathcal{R}_{i,j}$ that is available to the jth database of P_i. Thus,

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

Let us denote all the queries generated by the leader party 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.,

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

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

---

[3]As investigated in [5], in MM-SPIR problem, a user needs to retrieve P messages from N replicated servers containing K messages. The 2-party PSI problem can be recast as MM-SPIR by considering that the messages correspond to *incidences* of elements in its data set with respect to the finite set of all elements. Specifically, [5] transforms each data set into a library of K binary messages of one-bit length. Finally, in [5], party P_1 (or P_2) performs MM-SPIR of the messages corresponding to its data set $\mathcal{P}_1$ (or $\mathcal{P}_2$) within the databases of the other entity.

[4]In this work, we only consider semi-honest (honest but curious) parties in the sense that parties exactly follow the prescribed protocol but curious to learn more about the others. MP-PSI under malicious/adversarial attacks and in the presence of dishonest parties is an interesting future direction that is outside the scope of this work.

[5]We note that the multi-server assumption exists in almost all information-theoretic PIR literature. In practice, the data content may be distributed to the databases by a central content generator who does not communicate directly with other parties, i.e., does not have access to the exchanged queries. The databases do not have any direct communication links among each other and they update their content by downloading the data from the content generator. Hence, in this setting, the databases are replicated but not colluding.

[6]The presented achievability scheme works for any data set generation model and even for distribution-free data sets. The specific data set generation model in the 2-party PSI problem in [5] was introduced only for settling the converse.

[7]We note that the common randomness (key) exchange is an interesting stand-alone problem that is outside the scope of this article. One practical solution to this problem in our setting is to have an external *helper*, who generates and shares the common randomness prior to the MP-PSI determination process. The external helper is not involved in the MP-PSI process itself, i.e., it does not observe the queries or the answers. In this case, the client parties do not need to communicate with each other to exchange the common randomness and there is no leakage from their queries/answers to the external helper. We note that the SPIR problem [12] (and by extension our scheme) is infeasible if no common randomness exists.

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. This is captured by the following MP-PSI reliability constraint,

$$\left[\text{MP-PSI reliability}\right]$$
$$H\left(\mathcal{P}|Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) = 0 \quad (6)$$

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

$$\left[\text{Leader's privacy}\right] \quad \forall i \in [1 : M-1], \forall j \in [1 : N_i]$$
$$I\left(\mathcal{P}_M; Q_{i,j}^{[\mathcal{P}_M]}, A_{i,j}^{[\mathcal{P}_M]}, \mathcal{P}_i, \mathcal{R}_{i,j}\right) = 0 \quad (7)$$

Note that the communication between any two client parties is not allowed in our protocol. This implies that the party P_i is not able to get any information about the remaining $M-2$ client parties. Thus, the mutual independence required by the problem formulation is thereby satisfied from the perspective of the party P_i.

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 $\{X_{1,\bar{\mathcal{P}}}, \ldots, X_{M-1,\bar{\mathcal{P}}}\} = \{X_{1,k}, \ldots, X_{M-1,k}, k \in \bar{\mathcal{P}}\}$ 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 $X_{\bar{\mathcal{P}}} = \{\{X_{1,\bar{\mathcal{P}}}, \ldots, X_{M-1,\bar{\mathcal{P}}}\} : X_{1,k} + \cdots + X_{M-1,k} < M-1, \forall k \in \mathcal{P}_M \cap \bar{\mathcal{P}}\}$, we have the following client's privacy constraint,

$$\left[\text{Client's privacy}\right]$$
$$I\left(X_{\bar{\mathcal{P}}}; Q_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, A_{1:M-1,1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) = 0 \quad (8)$$

For a given field size K and individual parties with associated databases, an MP-PSI achievability scheme is a scheme that satisfies the MP-PSI 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[8] which is the number of downloaded bits

---

[8] We note that although a more natural performance metric is to consider the combined upload and download cost, we argue that the upload cost may not scale with the number of MP-PSI determination rounds if the MP-PSI is regularly repeated [5, Footnote 8]. Since the core of [5] (and the current paper also) relies on SPIR, we give a detailed discussion of how to reduce the upload cost of the SPIR scheme without sacrificing the download cost in [5, Sec. 7.2]. The optimal download cost of the SPIR problem is characterized in [12] with keeping the upload cost unconstrained. In addition, the optimal upload cost of the SPIR problem is characterized in [68] with keeping the download cost unconstrained. The optimal combined download and upload cost for the canonical SPIR problem is still an open problem.

(denoted by D) by one of the parties 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

In this section, we state our main result concerning the performance of our MP-PSI scheme in terms of the download cost. This is summarized in the following theorem, whose proof is given in Section V.

Theorem 1: In the MP-PSI problem with M independent 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 $i = 1, \ldots, M$, then the optimal download cost, D^*, is upper bounded by

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

Remark 1: In the special case of having an arbitrary party P_i where $|\mathcal{P}_i| = K$, we discard this party P_i before we perform the MP-PSI determination process, and thereby, the M-party MP-PSI problem reduces to an $M-1$-party MP-PSI problem. In the extreme case, where all parties have $|\mathcal{P}_i| = |\mathbb{S}_K| = K$, the download cost becomes zero, i.e., no party needs to exchange any information with any other, as the intersection is immediate.

Remark 2: The minimization problem in (9) in Theorem 1 corresponds to the fact that the parties can agree on the party with the minimum $\sum_{i \in \{1, \ldots, M\} \setminus t} \lceil \frac{|\mathcal{P}_t|N_i}{N_i - 1} \rceil$ to be the leader party. We note that the leader party may not be the party with the least $|\mathcal{P}_i|$, as the download cost also depends on the number of the databases at all parties.

Remark 3: The download cost of our achievability scheme is equal to the sum of the download costs of $M-1$ pair-wise PSI schemes. This implies that there is no penalty incurred due to adopting a stringent clients' privacy constraint over the E_2 privacy constraint. Note that the E_2 privacy constraint is a relaxed version of client's privacy (8) when $M = 2$ [5]. More specifically, the E_2 privacy constraint asserts that the leakage from elements outside the set $\mathcal{P}_1$ in the answers returned by E_2 is zero, i.e., $I(\bar{\mathcal{P}}_1; A_{1:N_2}^{[\mathcal{P}_1]}) = 0$.

Remark 4: Our achievability 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 [5]; otherwise, our scheme is infeasible if $N_i = 1$ for all i.

Remark 5: Comparing our result with the most closely related information-theoretic MP-PSI schemes [69], we argue that 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 $i = 1, \ldots, M$ in contrast of $O(M^4 p^2)$ in [69]. We note, however, that the work [69] allows for potential distrust between the parties in the sense that an active adversary may corrupt up to $M/3$ parties. The issue of parties' misbehavior is an

interesting future direction for our work, which is outside the scope of this article.

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

In this section, we motivate our scheme by presenting the following example. In this example, we have $M = 3$ parties, each possessing $N_i = 3$ replicated and non-colluding 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 according to the defined communication policy. 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 as in [5], i.e., we construct a vector X_i, such that $X_{i,k} = 1$ if $k \in \mathcal{P}_i$, hence,

$$\text{Party } P_1: \quad \mathcal{P}_1 = \{1, 2\}$$
$$\Rightarrow \quad X_1 = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & X_{1,4} \end{bmatrix}^T = [1 \ 1 \ 0 \ 0]^T \quad (10)$$
$$\text{Party } P_2: \quad \mathcal{P}_2 = \{1, 3\}$$
$$\Rightarrow \quad X_2 = \begin{bmatrix} X_{2,1} & X_{2,2} & X_{2,3} & X_{2,4} \end{bmatrix}^T = [1 \ 0 \ 1 \ 0]^T \quad (11)$$
$$\text{Party } P_3: \quad \mathcal{P}_3 = \{1, 4\}$$
$$\Rightarrow \quad X_3 = \begin{bmatrix} X_{3,1} & X_{3,2} & X_{3,3} & X_{3,4} \end{bmatrix}^T = [1 \ 0 \ 0 \ 1]^T \quad (12)$$

To carry out the MP-PSI calculations, the parties agree on a finite field $\mathbb{F}_L$, where L is a prime number such that $L \geq M$. Therefore, we pick $L = 3$ in our case, i.e., all summations are performed as modulo-3 arithmetic.

The leader party P_3 initiates the MP-PSI determination protocol by sending queries $Q_{i,j}^{[\mathcal{P}_3]}$ for $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$. 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 [12] (the same query structure was introduced in the original work of [6]). Note that in this example we have $N_i = |\mathcal{P}_3| + 1$, thus, the leader party sends exactly 1 query to each client database. More specifically, let h_k, where $k = 1, \ldots, 4$, 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 generated as follows,

$$Q_{1,1}^{[\mathcal{P}_3]} = [h_1 \ h_2 \ h_3 \ h_4]^T \quad (13)$$
$$Q_{1,2}^{[\mathcal{P}_3]} = [h_1 + 1 \ h_2 \ h_3 \ h_4]^T \quad (14)$$
$$Q_{1,3}^{[\mathcal{P}_3]} = [h_1 \ h_2 \ h_3 \ h_4 + 1]^T \quad (15)$$

i.e., the leader party sends a random vector $\mathbf{h} = [h_1 \ h_2 \ h_3 \ h_4] \in \mathbb{F}_3^4$ to the first database as a query. The queries for the remaining databases add a 1 to the positions corresponding to $\mathcal{P}_3$. For client party P_2, the leader party submits the same set of queries,

$$Q_{2,1}^{[\mathcal{P}_3]} = [h_1 \ h_2 \ h_3 \ h_4]^T \quad (16)$$

$$Q_{2,2}^{[\mathcal{P}_3]} = [h_1 + 1 \ h_2 \ h_3 \ h_4]^T \quad (17)$$
$$Q_{2,3}^{[\mathcal{P}_3]} = [h_1 \ h_2 \ h_3 \ h_4 + 1]^T \quad (18)$$

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 [5], [12] 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 (see Fig. 2).

1) *Local randomness:* This is denoted by the random variable s_i, for $i = 1, 2$. The random variable s_i is picked uniformly from $\mathbb{F}_3$ independent of all data sets and other randomness sources. The local randomness s_i is shared among all the databases belonging to the ith client party and not shared with other parties. This local randomness acts as the common randomness needed for SPIR [12], and is added to the inner product of the incidence vector and the query.

2) *Individual correlated randomness:* This is possessed by each client's database, and is denoted by the random variables $t_{i,j}$ for $i = 1, 2$, and $j = 1, 2, 3$. This is needed to prevent the leader party from decoding $X_{1,4}$, and $X_{2,4}$. However, since we also need the leader party to decode the intersection, 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. To that end, we choose $t_{1,1} = t_{2,1} = 0$. Database 2 of the party P_1 generates uniformly and independently $t_{1,2}$ from $\mathbb{F}_3$ and sends it to database 2 of party P_2. Database 2 of the party P_2 calculates $t_{2,2} = 1 - t_{1,2}$. Similarly, database 3 of the party P_1 generates $t_{1,3}$ uniformly and independently from $\mathbb{F}_3$ and shares it with database 3 of P_2. Hence,

$$t_{1,j} \sim \text{uniform}\{0, 1, 2\}, \quad j = 2, 3 \quad (19)$$
$$t_{1,j} + t_{2,j} = 1, \quad j = 2, 3 \quad (20)$$

This randomness is added to each response as well. Note that client parties do not know each other's data sets while generating/sharing this randomness.

3) *Global randomness:* This is denoted by the random variable c. The random variable c is generated randomly and independently of all data sets and other randomness variables. The global randomness c is picked uniformly from $\mathbb{F}_3 \setminus \{0\} = \{1, 2\}$. 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.

After sharing the common randomness needed to construct the answer strings as shown above, the jth database of the ith client party responds to the query $Q_{i,j}^{[\mathcal{P}_3]}$ as follows,

$$A_{i,j}^{[\mathcal{P}_3]} = c\Big(X_i^T Q_{i,j}^{[\mathcal{P}_3]} + s_i + t_{i,j}\Big), \quad i = 1, 2, \ j = 1, 2, 3 \quad (21)$$
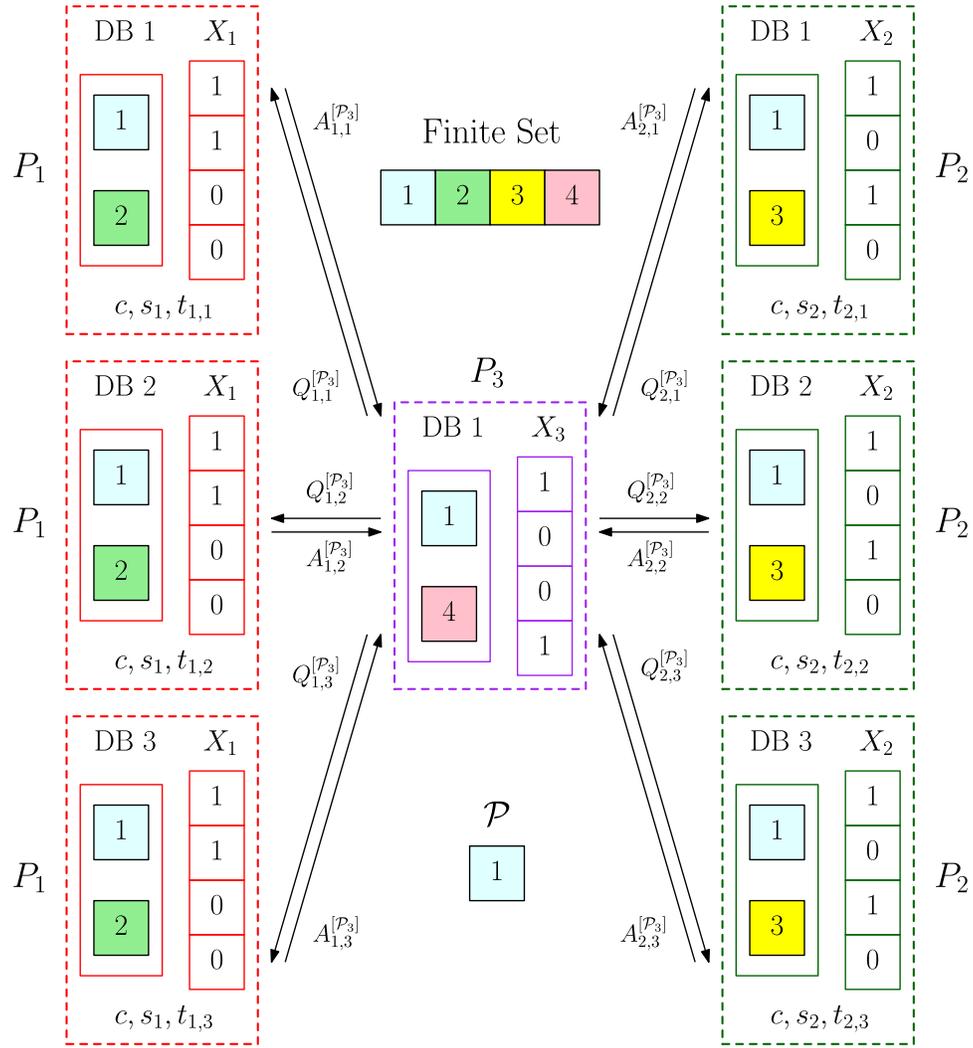
Fig. 2. MP-PSI for the motivating example.

Hence, noting that $t_{1,1} = 0$, the answer strings from P_1 can be explicitly written as,

$$A_{1,1}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{1,k} + s_1\right) \tag{22}$$

$$A_{1,2}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{1,k} + X_{1,1} + s_1 + t_{1,2}\right) \tag{23}$$

$$A_{1,3}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{1,k} + X_{1,4} + s_1 + t_{1,3}\right) \tag{24}$$

Similarly, the answer strings from P_2 are,

$$A_{2,1}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{2,k} + s_2\right) \tag{25}$$

$$A_{2,2}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{2,k} + X_{2,1} + s_2 + t_{2,2}\right) \tag{26}$$

$$A_{2,3}^{[\mathcal{P}_3]} = c\left(\sum_{k=1}^{4} h_k X_{2,k} + X_{2,4} + s_2 + t_{2,3}\right) \tag{27}$$

Note that, by this construction, the local randomness s_i is used to protect the random sum $\sum_{k=1}^{4} h_k X_{i,k}$ as in SPIR, and the individual randomness $t_{i,j}$ is needed to prevent the leader party from directly decoding $X_{i,j+1}$. Note that s_1 and s_2 need to be independent to avoid the information leakage about the relationship between $\sum_{k=1}^{4} h_k X_{1,k}$ and $\sum_{k=1}^{4} h_k X_{2,k}$.

a) *Reliability:* To calculate $\cap_{i=1,2,3} \mathcal{P}_i$ based on the answer strings the leader party has received, the leader party subtracts $A_{1,1}^{[\mathcal{P}_3]}$ and $A_{2,1}^{[\mathcal{P}_3]}$ from the remaining answer strings. Denote the result of subtraction related to the jth element in $\mathbb{S}_K$ at P_i by $Z_{i,j}$. This leads to,

$$Z_{1,1} = c(X_{1,1} + t_{1,2}) = A_{1,2}^{[\mathcal{P}_3]} - A_{1,1}^{[\mathcal{P}_3]} \tag{28}$$

$$Z_{1,4} = c(X_{1,4} + t_{1,3}) = A_{1,3}^{[\mathcal{P}_3]} - A_{1,1}^{[\mathcal{P}_3]} \tag{29}$$

$$Z_{2,1} = c(X_{2,1} + t_{2,2}) = A_{2,2}^{[\mathcal{P}_3]} - A_{2,1}^{[\mathcal{P}_3]} \tag{30}$$

$$Z_{2,4} = c(X_{2,4} + t_{2,3}) = A_{2,3}^{[\mathcal{P}_3]} - A_{2,1}^{[\mathcal{P}_3]} \tag{31}$$

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.,

$$E_j = \sum_{i=1}^{M-1} Z_{i,j} \qquad (32)$$

Looking deeper at E_1, we note that,

$$E_1 = Z_{1,1} + Z_{2,1} \qquad (33)$$
$$= c\big(X_{1,1} + X_{2,1} + t_{1,2} + t_{2,2}\big) \qquad (34)$$
$$= c\big(X_{1,1} + X_{2,1} + 1\big) \qquad (35)$$

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. In this case, $E_1 = 0$ 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,

$$E_4 = Z_{1,4} + Z_{2,4} = c\big(X_{1,1} + X_{2,1} + 1\big) \neq 0 \qquad (36)$$

Consequently, the leader party confirms that $\cap_{i=1,2,3} \mathcal{P}_i = \{1\}$ and does not include 4.

b) *Leader's Privacy:* The leader's privacy constraint follows from the user's privacy constraint of the inherent SPIR scheme [12]. The queries of the leader to any party have the same structure as the queries of the user in the SPIR problem. More specifically, the privacy of leader party is preserved as 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.

c) *Client's Privacy:* To see the client's privacy, we note that 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. Nevertheless, in MP-PSI, we need to verify that the leader does not know which of the two parties possesses the element $\{4\}$, i.e., knowing the fact that $E_4 \neq 0$, we need to show that $\mathbb{P}(X_{1,4} + X_{2,4} = 0) = \mathbb{P}(X_{1,4} + X_{2,4} = 1) = \frac{1}{2}$. Specifically, if E_4 is 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 (this is no further leak, as if they did contain it at the same time, it would have been in the intersection). Hence, c is used such that the leader party P_3 does not know whether the sum $X_{1,4} + X_{2,4}$ is 0 or 1.

d) *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

In this section, we describe our general achievable scheme for MP-PSI for arbitrary number of parties M, arbitrary set sizes $|\mathcal{P}_i|$, and arbitrary number of databases per party N_i, for $i \in \{1, \ldots, M\}$. The leader's querying policy is based on the SPIR scheme presented in [12] (originally introduced in [6]). Our novel ideas in this scheme are concerned with the construction of the answering strings. More specifically, the scheme hinges on 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 the intersection $\cap_{i=1}^{M} \mathcal{P}_i$.

A. General Achievability Scheme

In the following, assume that $\mathcal{P}_i \subseteq \mathbb{S}_K$, where $|\mathbb{S}_K| = K$.

1) *Initialization:* The parties agree on a retrieval finite field $\mathbb{F}_L$ to carry out the calculations needed for MP-PSI determination protocol. L is chosen such that,

$$L = \min \{L \geq M : L \text{ is a prime}\} \qquad (37)$$

The parties agree on a leader P_{t^*} such that:

$$t^* = \arg \min_{t \in \{1, \ldots, M\}} \sum_{i \neq t} \left\lceil \frac{|\mathcal{P}_t| N_i}{N_i - 1} \right\rceil \qquad (38)$$

Without loss of generality, we assume that $t^* = M$ in the sequel. Furthermore, assume that $\mathcal{P}_{t^*} = \mathcal{P}_M = \{Y_1, Y_2, \ldots, Y_R\}$ with cardinality $|\mathcal{P}_M| = R$.

2) *Query generation:* The leader party P_M independently and uniformly generates κ random vectors $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_\kappa\}$, where κ is given by,

$$\kappa = \max_{i \in \{1, \ldots, M-1\}} \left\lceil \frac{|\mathcal{P}_M|}{N_i - 1} \right\rceil \qquad (39)$$

The vector $\mathbf{h}_\ell$, for $\ell = 1, 2, \ldots, \kappa$ is picked uniformly from $\mathbf{F}_L^K$ such that,

$$\mathbf{h}_\ell = [h_\ell(1) \quad h_\ell(2) \quad \cdots \quad h_\ell(K)] \qquad (40)$$

Denote $\eta_i = \lceil \frac{|\mathcal{P}_M|}{N_i - 1} \rceil$, and $\mathcal{P}_M^{\ell_i} = \{Y_1^{\ell_i}, Y_2^{\ell_i}, \ldots, Y_{N_i-1}^{\ell_i}\}$, for $i = 1, \ldots, M-1$. The leader party P_M submits η_i random vectors from $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_\kappa\}$ 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, take ℓ_i to be a running index, i.e., $\ell_i = 1, 2, \ldots, \eta_i$, and 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 $|\mathcal{P}_M^{\ell_i}| = N_i - 1$ (except potentially for the last subset $\mathcal{P}_M^{\eta_i}$), then for $i = 1, 2, \ldots, M-1$, the query structure is given by:

$$Q_{i,1}^{\left[\mathcal{P}_M^{\ell_1}\right]} = [h_1(1) \quad h_1(2) \quad \cdots \quad h_1(K)\} \qquad (41)$$

$$Q_{i,2}^{\left[\mathcal{P}_M^{\ell_1}\right]} = \Big[h_1(1) \quad \cdots \quad h_1\big(Y_1^{\ell_1} - 1\big) \quad h_1\big(Y_1^{\ell_1}\big) + 1$$
$$h_1\big(Y_1^{\ell_1} + 1\big) \quad \cdots \quad h_1(K)\Big] \qquad (42)$$

$$\vdots$$

$$Q_{i,N_i}^{\left[\mathcal{P}_M^{\ell_1}\right]} = \Big[h_1(1) \quad \cdots \quad h_1\big(Y_{N_i-1}^{\ell_1} - 1\big) \quad h_1\big(Y_{N_i-1}^{\ell_1}\big) + 1$$
$$h_1\big(Y_{N_i-1}^{\ell_1} + 1\big) \quad \cdots \quad h_1(K)\Big] \qquad (43)$$

$$\vdots$$

$$Q_{i,1}^{\left[\mathcal{P}_M^{\eta_i}\right]} = [h_{\eta_i}(1) \quad h_{\eta_i}(2) \quad \cdots \quad h_{\eta_i}(K)\} \qquad (44)$$

$$Q_{i,2}^{\left[\mathcal{P}_M^{\eta_i}\right]} = \Big[h_{\eta_i}(1) \quad \cdots \quad h_{\eta_i}\big(Y_1^{\eta_i} - 1\big) \quad h_{\eta_i}\big(Y_1^{\eta_i}\big) + 1$$
$$h_{\eta_i}\big(Y_1^{\eta_i} + 1\big) \quad \cdots \quad h_{\eta_i}(K)\Big] \qquad (45)$$

$$\vdots$$

$$Q_{i,N_i}^{\left[\mathcal{P}_M^{\eta_i}\right]} = \left[h_{\eta_i}(1) \quad \cdots \quad h_{\eta_i}\left(Y_{N_i-1}^{\eta_i} - 1\right) \quad h_{\eta_i}\left(Y_{N_i-1}^{\eta_i}\right) + 1 \right.$$
$$\left. h_{\eta_i}\left(Y_{N_i-1}^{\eta_i} + 1\right) \quad \cdots \quad h_{\eta_i}(K) \right] \tag{46}$$

i.e., P_M simply partitions the set $\mathcal{P}_M$ into subsets of size $N_i - 1$. For each set, P_M uses different $\mathbf{h}_\ell$. P_M submits $\mathbf{h}_\ell$ into the first database. For the remaining databases, it adds 1 for the positions that corresponds to the partition.

3) *Common randomness generation:* In order to respond to the leader party, the clients need to generate and share common randomness. Specifically, there are three types of randomness.

- *Local randomness:* This is denoted by $\mathbf{s}_i = [s_i(1) \; s_i(2) \; s_i(\eta_i)]$. Each element of $\mathbf{s}_i$ is generated independently and uniformly from $\mathbb{F}_L$. The local randomness $\mathbf{s}_i$ is shared between the databases associated with P_i. The local randomness is added to the responses as in SPIR [12]. 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 $i = 1, \ldots, M - 1$, and $j = 1, \ldots, N_i$. The elements $t_{i,1} = 0$ for all i. For $i = 1, \ldots, M - 2$, the vector $\mathbf{t}_{i,j}$ is independently and uniformly picked from $\mathbf{F}_L^{\eta_i}$. All these random vectors are sent to the party P_{M-1}. The client P_{M-1} generates its individual randomness $\mathbf{t}_{M-1,j}$ according to the received individual randomness from the remaining parties. 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 $k = 1, 2, \ldots, R$ is just a monotonically increasing index of the randomness component used within the databases 2 to N_i of the ith client. Thus,

$$\tilde{t}_{i,1} = t_{i,2}(1), \quad \tilde{t}_{i,1} = t_{i,2}(2), \ldots, \tilde{t}_{i,R} = t_{i,N_i}(\eta_i) \tag{47}$$

With this re-definition, the client P_{M-1} calculates its individual randomness as,

$$\tilde{t}_{M-1,j} = L - (M-1) - \sum_{i=1}^{M-2} \tilde{t}_{i,j}, j = 1, 2, \ldots, R \tag{48}$$

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. c is picked uniformly and independently from $\mathbb{F}_L \setminus \{0\}$. c is shared among all the databases at all clients. c is used as a multiplier for the answering string.

4) *Response generation:* The clients respond to the submitted queries by using the queries as a combining vector to their contents, i.e., each database calculates the inner product of the query and its contents. Next, it adds the local and individual randomness. Finally, it multiplies the result by the global randomness. More specifically,

the answer string of the jth database, which is associated with the ith client to retrieve one of the elements of the partition $\mathcal{P}_M^{\ell_i}$, $A_{i,j}^{\left[\mathcal{P}_M^{\ell_i}\right]}$, is given by,

$$A_{i,j}^{\left[\mathcal{P}_M^{\ell_i}\right]} = c\left(X_i^T Q_{i,j}^{\left[\mathcal{P}_M^{\ell_i}\right]} + s_i(\ell_i) + t_{i,j}(\ell_i) \right) \tag{49}$$

From the collected answers the leader party can determine the intersection $\cap_{i=1}^{M} \mathcal{P}_i$ reliably and privately.

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

a) *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 $\eta_i = \lceil \frac{|\mathcal{P}_M|}{N_i-1} \rceil$ 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, the download cost is given by,

$$D = \sum_{i=1}^{M-1} N_i \eta_i \tag{50}$$

$$= \sum_{i=1}^{M-1} \left\lceil \frac{|\mathcal{P}_M| N_i}{N_i - 1} \right\rceil. \tag{51}$$

b) *Reliability:* To verify reliability, we follow the leader's processing of the responses. First, 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, and is given by,

$$A_{i,1}^{\left[\mathcal{P}_M^{\ell_i}\right]} = c\left(\sum_{k=1}^{K} h_{\ell_i}(k) X_{i,k} + s_i(\ell_i) \right), \quad i = 1, \ldots, M-1 \tag{52}$$

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,

$$Z_{i,k} = c\left(X_{i,k} + \tilde{t}_{i,k} \right) = A_{i,j^*}^{\left[\mathcal{P}_M^{\ell_i}\right]} - A_{i,1}^{\left[\mathcal{P}_M^{\ell_i}\right]}, \quad k \in \mathcal{P}_M^{\ell_i} \tag{53}$$

for some unique j^* that $A_{i,j^*}^{\left[\mathcal{P}_M^{\ell_i}\right]}$ 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 $i = 1, \ldots, M-1$, 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,

$$E_k = \sum_{i=1}^{M-1} Z_{i,k} \tag{54}$$

$$= c\left(\sum_{i=1}^{M-1} X_{i,k} + \sum_{i=1}^{M-1} \tilde{t}_{i,k} \right) \tag{55}$$

$$= c\left(\sum_{i=1}^{M-1} X_{i,k} + L - (M-1) \right) \tag{56}$$

where (56) 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 $i = 1, 2, \ldots, M - 1$. 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.

c) Leader's privacy: The leader's privacy follows from the fact that the random vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_\kappa\}$ 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.

d) Clients' privacy: Without loss of generality, we derive the proof of the client's privacy for the homogeneous number of databases, i.e., $N_i = R + 1, \forall i \in [1 : M - 1]$. The general proof in the heterogeneous case follows the same steps and after removing the response of the first databases, we will be left with $Z_{i,k}$ that has the same structure of homogeneous case. Consequently, we present the homogeneous case here for convenience only. In the following proof, we adopt the notation that for a random variable $\zeta_{i,j}$ indexed by two indices (i, j),

$$\zeta_{i_1:i_M, j_1:j_R} = \left\{ \zeta_{i,j} : i \in \{i_1, \ldots, i_M\}, j \in \{j_1, \ldots, j_R\} \right\} \quad (57)$$

For the proof, we need the following lemmas. 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,

$$I\left(X_{\bar{\mathcal{P}}}; A_{1:M-1,1}^{[\mathcal{P}_M]} | Z_{1:M-1, Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) = 0. \quad (58)$$

Proof: Intuitively, the proof follows from the fact that $A_{i,1}^{[\mathcal{P}_M]}, i \in [1 : M - 1]$ is a random variable uniformly distributed over $[0 : L - 1]$ because of the local randomness s_i, and thus, is independent of the data sets, queries and the subtraction results. More specifically,

$$I\left(X_{\bar{\mathcal{P}}}; A_{1:M-1,1}^{[\mathcal{P}_M]} | Z_{1:M-1, Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right)$$
$$= H\left(A_{1:M-1,1}^{[\mathcal{P}_M]} | Z_{1:M-1, Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right)$$
$$\quad - H\left(A_{1:M-1,1}^{[\mathcal{P}_M]} | X_{\bar{\mathcal{P}}}, Z_{1:M-1, Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) \quad (59)$$
$$\leq H\left(A_{1:M-1,1}^{[\mathcal{P}_M]}\right) - H\left(A_{1:M-1,1}^{[\mathcal{P}_M]} | X_{1:M-1}, c, X_{\bar{\mathcal{P}}},\right.$$
$$\left. Z_{1:M-1, Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) \quad (60)$$
$$\leq (M - 1) - H(s_1, \ldots, s_{M-1}) \quad (61)$$
$$= (M - 1) - (M - 1) = 0 \quad (62)$$

This concludes the proof, since,

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

∎

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 $X_{\bar{\mathcal{P}}}$. Note that we do not claim anything about Z_{M-1,Y_j} as the individual randomness are correlated at party $M - 1$.

Lemma 2: For the presented scheme, we have,

$$I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-2, Y_1:Y_R} | E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) = 0. \quad (63)$$

Proof: Intuitively, similar to the proof of Lemma 1, the proof follows from the fact that $Z_{i,Y_j}, i \in [1 : M - 2], j \in [1 : R]$ is a random variable uniformly distributed over $[0 : L - 1]$ because of the individual randomness $t_{i,j+1}$, and thus, is independent of the data sets, queries, and the data sets in the client parties E_{Y_j},

$$I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-2, Y_1:Y_R} | E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right)$$
$$= H\left(Z_{1:M-2, Y_1:Y_R} | E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right)$$
$$\quad - H\left(Z_{1:M-2, Y_1:Y_R} | X_{\bar{\mathcal{P}}}, E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) \quad (64)$$
$$\leq H\left(Z_{1:M-2, Y_1:Y_R}\right)$$
$$\quad - H\left(Z_{1:M-2, Y_1:Y_R} | X_{1:M-1}, c, X_{\bar{\mathcal{P}}}, E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) \quad (65)$$
$$\leq ((M - 2)R) - H(t_{1:M-2, Y_1:Y_R}) \quad (66)$$
$$= ((M - 2)R) - ((M - 2)R) = 0 \quad (67)$$

This concludes the proof as the reverse implication is true by the non-negativity of mutual information. ∎

The following lemma asserts 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,

$$I\left(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) = 0. \quad (68)$$

Proof: Note that if $Y_j \in \mathcal{P}_M$ is in the intersection, $E_{Y_j} = 0$ has nothing to do with $X_{\bar{\mathcal{P}}}$ since $X_{\bar{\mathcal{P}}}$ is defined on the elements not in the intersection. However, if Y_j is not in the intersection, $E_{Y_j} = c(X_{1,Y_j} + \cdots + X_{M-1,Y_j} + L - (M - 1)), Y_j \in \mathcal{P}_M \cap \bar{\mathcal{P}}$ received by the leader party would be a realization within the range of $\mathbb{F}_L \setminus \{0\}$ because of the global randomness c. However, the leader party only knows that the global randomness c is uniformly distributed over $\mathbb{F}_L \setminus \{0\}$ and has no information about the specific value of c in the client parties. As a result, from the perspective of the leader part P_M, $X_{1,Y_j} + \cdots + X_{M-1,Y_j} + L - (M - 1)$ is uniformly distributed over $[1 : L - 1]$ according to the information contained in E_{Y_j}. This comes from the fact that the set $\mathbb{F}_L \setminus \{0\}$ of all $L - 1$ non-zero elements must form a finite cyclic group under multiplication given a finite field $\mathbb{F}_L$. That means that, in the additive table under multiplication operation, each element in $\mathbb{F}_L \setminus \{0\}$ appears precisely once in each row and column of the table. The probability $\mathbb{P}(X_{1,Y_j} + \cdots + X_{M-1,Y_j} + L - (M - 1) = l)$ would always be $\frac{1}{L-1}$ for any $l \in [1 : L - 1]$. Then, $X_{1,Y_j} + \cdots + X_{M-1,Y_j}$ is uniformly distributed over $[M - L : M - 2]$ (i.e., $[0 : M - 2] \cup [M : L - 1]$) and we can further conclude that $X_{1,Y_j} + \cdots + X_{M-1,Y_j}$ is uniformly distributed over $[0 : M - 2]$ because its largest possible value is $M - 2$ if Y_j is not in the intersection. Thus, the only information we can learn from $E_{Y_1}, \ldots, E_{Y_R}$ and the accompanying queries about $X_{\bar{\mathcal{P}}}$ is $X_{1,k} + \cdots + X_{M-1,k} < M - 1, \forall k \in \mathcal{P}_M \cap \bar{\mathcal{P}}$ without knowing the specific value of $X_{1,k} + \cdots + X_{M-1,k}$, which already exists in the definition of $X_{\bar{\mathcal{P}}}$. Thus, we obtain,

$$I\left(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right)$$
$$= I\left(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R} | Q_{1:M-1, 1:N_i}^{[\mathcal{P}_M]}, \mathcal{P}_M\right) \quad (69)$$

$$= H\left(X_{\bar{\mathcal{P}}} | Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right)$$

$$- H\left(X_{\bar{\mathcal{P}}} | E_{Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (70)$$

$$= H(X_{\bar{\mathcal{P}}}) - H(X_{\bar{\mathcal{P}}}) \quad (71)$$

$$= 0 \quad (72)$$

where (69) follows from the fact that queries and $\mathcal{P}_M$ are independent of the data sets in the client parties E_{Y_j} in (2). ■

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

$$I\left(X_{\bar{\mathcal{P}}}; Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, A^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right)$$

$$= I\left(X_{\bar{\mathcal{P}}}; A^{[\mathcal{P}_M]}_{1:M-1,1}, Z_{1:M-1,Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (73)$$

$$= I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-1,Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right)$$

$$+ I\left(X_{\bar{\mathcal{P}}}; A^{[\mathcal{P}_M]}_{1:M-1,1} | Z_{1:M-1,Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (74)$$

$$= I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-1,Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (75)$$

$$= I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-2,Y_1:Y_R}, E_{Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (76)$$

$$= I\left(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right)$$

$$+ I\left(X_{\bar{\mathcal{P}}}; Z_{1:M-2,Y_1:Y_R} | E_{Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (77)$$

$$= I\left(X_{\bar{\mathcal{P}}}; E_{Y_1:Y_R}, Q^{[\mathcal{P}_M]}_{1:M-1,1:N_i}, \mathcal{P}_M\right) \quad (78)$$

$$= 0 \quad (79)$$

where (73) follows from the fact that there is a bijective transformation between $A^{[\mathcal{P}_M]}_{1:M-1,1:N_i}$ and $(A^{[\mathcal{P}_M]}_{1:M-1,1}, Z_{1:M-1,Y_1:Y_R})$, (75) follows from Lemma 1, (76) follows from the fact that there is a bijective transformation between $Z_{1:M-1,Y_1:Y_R}$ and $(Z_{1:M-2,Y_1:Y_R}, E_{Y_1:Y_R})$, (78) follows from Lemma 2, and (79) follows from Lemma 3.

VI. FURTHER EXAMPLES

In this section, we present two examples of our achievable scheme. Unlike the motivating example in Section IV, in these examples, the number of databases per party does not need to be $N_i = |\mathcal{P}_M| + 1$ or even be homogeneous in general.[9]

A. An Example for $N_i < |\mathcal{P}_M| + 1$

In this example, we use the same setting of Section IV with $\mathcal{P}_1 = \{1, 2\}$, $\mathcal{P}_2 = \{1, 3\}$, and $\mathcal{P}_3 = \{1, 4\}$ with P_3 being the leader party and the retrieval field being $\mathbb{F}_3$. The incidence vectors X_i, for $i = 1, 2$ remain the same. However, to illustrate that our scheme works for $N_i < |\mathcal{P}_M| + 1$, we assume that $N_1 = N_2 = 2$. As we will show next, when $N_i < |\mathcal{P}_M| + 1$, we need to send $\kappa = \eta_i = \lceil \frac{|\mathcal{P}_M|}{N_i - 1} \rceil = 2$ queries to the first database of the ith party (in contrast to 1 query only when $N_i \geq |\mathcal{P}_M| + 1$). Moreover, the common randomness components $\mathbf{s}_i$, and $\mathbf{t}_{i,j}$ need to be vectors of size $\lceil \frac{|\mathcal{P}_M|}{N_i - 1} \rceil = 2$. Note that, in this case, the leader's set is divided into 2 subsets $\mathcal{P}^{\ell_1}_M = \{1\}$ and $\mathcal{P}^{\ell_1}_M = \{4\}$ as $|\mathcal{P}^{\ell_i}_M| = N_i - 1 = 1$.

[9]For $N_i > |\mathcal{P}_M| + 1$, we just use any arbitrary $|\mathcal{P}_M| + 1$ databases to execute the MP-PSI determination protocol.

For the queries, since both client parties have the same number of databases, the leader P_3 submits the same query vectors to the databases of both clients. The first databases of each client receives 2 uniformly generated vectors $\mathbf{h}, \bar{\mathbf{h}} \in \mathbb{F}^4_3$, where $\mathbf{h} = [h_1\ h_2\ h_3\ h_4]^T$ and $\bar{\mathbf{h}} = [\bar{h}_1\ \bar{h}_2\ \bar{h}_3\ \bar{h}_4]^T$. P_3 submits the same two vectors to the second databases of P_1 and P_2 with adding 1 to the desired positions. More specifically, let $Q^{[k]}_{i,j}$ be the query to the jth database of P_i to retrieve the element k, then P_3 submits the following queries:

$$Q^{[1]}_{1,1} = Q^{[1]}_{2,1} = [h_1\ h_2\ h_3\ h_4]^T \quad (80)$$

$$Q^{[1]}_{1,2} = Q^{[1]}_{2,2} = [h_1 + 1\ h_2\ h_3\ h_4]^T \quad (81)$$

$$Q^{[4]}_{1,1} = Q^{[4]}_{2,1} = [\bar{h}_1\ \bar{h}_2\ \bar{h}_3\ \bar{h}_4]^T \quad (82)$$

$$Q^{[4]}_{1,2} = Q^{[4]}_{2,2} = [\bar{h}_1\ \bar{h}_2\ \bar{h}_3\ \bar{h}_4 + 1]^T \quad (83)$$

At the clients' side, the clients share a global randomness $c \sim \text{uniform}\{1, 2\}$ among all the databases of both clients. For $i = 1, 2$, the ith client generates and shares a local randomness $\mathbf{s}_i = [s_i(1)\ s_i(2)]^T$, such that $s_i(\ell) \sim \text{uniform}\{0, 1, 2\}$ among the databases that belong to the ith client. Finally, for $i = 1, 2$, the second database of the ith client has an individual correlated randomness $\mathbf{t}_{i,2} = [t_{i,2}(1)\ t_{i,2}(2)]^T$, such that $t_{1,2}(1) \sim t_{1,2}(2) \sim \text{uniform}\{0, 1, 2\}$, $t_{1,2}(1) + t_{2,2}(1) = 1$, and $t_{1,2}(2) + t_{2,2}(2) = 1$. Assume that $t_{1,1} = t_{2,1} = 0$. All randomness components are independently generated of each other and of the data sets.

The answer string $A^{[k]}_{i,j}$, for $i = 1, 2$, $j = 1, 2$, $k = 1, 4$, is given by,

$$A^{[k]}_{i,j} = c\left(X^T_i Q^{[k]}_{i,j} + s_i(\ell(k)) + t_{i,j}(\ell(k))\right) \quad (84)$$

where $\ell(1) = 1$ and $\ell(4) = 2$.

Thus, the leader party receives the following answer strings from P_1,

$$A^{[1]}_{1,1} = c\left(\sum_{k=1}^{4} h_k X_{1,k} + s_1(1)\right) \quad (85)$$

$$A^{[1]}_{1,2} = c\left(\sum_{k=1}^{4} h_k X_{1,k} + X_{1,1} + s_1(1) + t_{1,2}(1)\right) \quad (86)$$

$$A^{[4]}_{1,1} = c\left(\sum_{k=1}^{4} \bar{h}_k X_{1,k} + s_1(2)\right) \quad (87)$$

$$A^{[4]}_{1,2} = c\left(\sum_{k=1}^{4} \bar{h}_k X_{1,k} + X_{1,4} + s_1(2) + t_{1,2}(2)\right) \quad (88)$$

and the following answer strings from P_2,

$$A^{[1]}_{2,1} = c\left(\sum_{k=1}^{4} h_k X_{2,k} + s_2(1)\right) \quad (89)$$

$$A^{[1]}_{2,2} = c\left(\sum_{k=1}^{4} h_k X_{2,k} + X_{1,1} + s_2(1) + t_{2,2}(1)\right) \quad (90)$$

$$A^{[4]}_{2,1} = c\left(\sum_{k=1}^{4} \bar{h}_k X_{2,k} + s_2(2)\right) \quad (91)$$

$$A_{2,2}^{[4]} = c\left(\sum_{k=1}^{4} \bar{h}_k X_{2,k} + X_{1,4} + s_2(2) + t_{2,2}(2)\right) \quad (92)$$

The leader party constructs the subtractions $Z_{i,j}$ as follows,

$$Z_{1,1} = c(X_{1,1} + t_{1,2}(1)) = A_{1,2}^{[1]} - A_{1,1}^{[1]} \quad (93)$$

$$Z_{1,4} = c(X_{1,4} + t_{1,2}(2)) = A_{1,2}^{[4]} - A_{1,1}^{[4]} \quad (94)$$

$$Z_{2,1} = c(X_{1,1} + t_{2,2}(1)) = A_{2,2}^{[1]} - A_{2,1}^{[1]} \quad (95)$$

$$Z_{2,4} = c(X_{1,4} + t_{2,2}(2)) = A_{2,2}^{[4]} - A_{2,1}^{[4]} \quad (96)$$

These are exactly the statistics in (28)–(31). Hence, the reliability and privacy constraints follow exactly as in Section IV. The total download cost in this case is $D = 4+4 = 8$, which is consistent with the download cost for the general case in (9), $D = \lceil\frac{2*2}{2-1}\rceil + \lceil\frac{2*2}{2-1}\rceil = 8$.

B. An Example for Heterogeneous Number of Databases

In this example, we consider a general case, where there are no constraints on the number of databases associated with each party or on the cardinality of the sets. In this example, we have $M = 4$ parties with $N_1 = 2$, $N_2 = 3$, $N_3 = 5$, and $N_4 = 4$ associated databases. The four parties have the following data sets and the corresponding incidence vectors,

Party P_1: $\mathcal{P}_1 = \{1, 2, 3, 4\}$

$\Rightarrow$ $X_1 = [X_{1,1}\ X_{1,2}\ X_{1,3}\ X_{1,4}\ X_{1,5}]^T = [1\ 1\ 1\ 1\ 0]^T$ (97)

Party P_2: $\mathcal{P}_2 = \{1, 2, 4\}$

$\Rightarrow$ $X_2 = [X_{2,1}\ X_{2,2}\ X_{2,3}\ X_{2,4}\ X_{2,5}]^T = [1\ 1\ 0\ 1\ 0]^T$ (98)

Party P_3: $\mathcal{P}_3 = \{1, 3, 4\}$

$\Rightarrow$ $X_3 = [X_{3,1}\ X_{3,2}\ X_{3,3}\ X_{3,4}\ X_{3,5}]^T = [1\ 0\ 1\ 1\ 0]^T$ (99)

Party P_4: $\mathcal{P}_4 = \{1, 4, 5\}$

$\Rightarrow$ $X_4 = [X_{4,1}\ X_{4,2}\ X_{4,3}\ X_{4,4}\ X_{4,5}]^T = [1\ 0\ 0\ 1\ 1]^T$ (100)

First, we choose party P_4 for the role of the leader party, as it results in the minimum download cost $D_t = \sum_{i\neq t} \lceil\frac{|\mathcal{P}_t|N_i}{N_i-1}\rceil$. Since $M = 4$, we choose a retrieval field $\mathbb{F}_L$, such that $L = 5$, as L is the smallest prime number that satisfies $L \geq M$.

Now, $\kappa = \max_i \lceil\frac{|\mathcal{P}_4|}{N_i-1}\rceil = 3$. Hence, for the queries, the leader P_4 generates $\kappa = 3$ random vectors. From which, it submits $\eta_i = \lceil\frac{|\mathcal{P}_4|}{N_i-1}\rceil$ to the first database associated with the ith party, $i = 1, 2, 3$. Each random vector can be reused for retrieving $N_i - 1$ elements from the remaining databases by adding 1 to the query vector in the positions of the desired symbols.

Specifically, party P_1 has only two databases and P_4 is supposed to submits $\eta_1 = \lceil\frac{3}{2-1}\rceil = 3$ random vectors to database 1, denoted by $\mathbf{h}_\ell = [h_\ell(1)\ h_\ell(2) \cdots h_\ell(5)]^T$, where $\ell = 1, 2, 3$. The leader's set is divided as $\mathcal{P}_4^{11} = \{1\}$, $\mathcal{P}_4^{12} = \{4\}$, and $\mathcal{P}_4^{13} = \{5\}$ with $|\mathcal{P}_M^{\ell_1}| = N_1 - 1 = 1$. These random vectors are generated uniformly from $\mathbb{F}_5^5$. Thus, the queries sent from

P_4 to P_1 are generated as follows,

$$Q_{1,1}^{[1]} = [h_1(1)\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (101)$$

$$Q_{1,2}^{[1]} = [h_1(1)+1\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (102)$$

$$Q_{1,1}^{[4]} = [h_2(1)\quad h_2(2)\quad h_2(3)\quad h_2(4)\quad h_2(5)]^T \quad (103)$$

$$Q_{1,2}^{[4]} = [h_2(1)\quad h_2(2)\quad h_2(3)\quad h_2(4)+1\quad h_2(5)]^T \quad (104)$$

$$Q_{1,1}^{[5]} = [h_3(1)\quad h_3(2)\quad h_3(3)\quad h_3(4)\quad h_3(5)]^T \quad (105)$$

$$Q_{1,2}^{[5]} = [h_3(1)\quad h_3(2)\quad h_3(3)\quad h_3(4)\quad h_3(5)+1]^T \quad (106)$$

Party P_2 has three databases and P_4 only needs to send $\eta_2 = \lceil\frac{4}{3-1}\rceil = 2$ random vectors to database 1 of client P_2. Each random vector can be reused at databases 2, 3 to retrieve 2 desired symbols. The leader's set is divided as $\mathcal{P}_4^{21} = \{1, 4\}$, and $\mathcal{P}_4^{22} = \{5\}$. Without loss of generality, P_4 uses $\mathbf{h}_1$ to obtain the information of $X_{2,1}, X_{2,4}$ and $\mathbf{h}_2$ is used to obtain the information of $X_{2,5}$. Note that, in this case no query is needed to be sent to the third database to retrieve $X_{2,5}$. Thus, the queries sent from P_4 to P_2 are generated as follows,

$$Q_{2,1}^{[1,4]} = [h_1(1)\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (107)$$

$$Q_{2,2}^{[1,4]} = [h_1(1)+1\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (108)$$

$$Q_{2,3}^{[1,4]} = [h_1(1)\quad h_1(2)\quad h_1(3)\quad h_1(4)+1\quad h_1(5)]^T \quad (109)$$

$$Q_{2,1}^{[5]} = [h_2(1)\quad h_2(2)\quad h_2(3)\quad h_2(4)\quad h_2(5)]^T \quad (110)$$

$$Q_{2,2}^{[5]} = [h_2(1)\quad h_2(2)\quad h_2(3)\quad h_2(4)\quad h_2(5)+1]^T \quad (111)$$

Party P_3 has five databases and P_4 needs to send $\eta_3 = \lceil\frac{4}{5-1}\rceil = 1$ random vector to database 1 and reuse this vector to retrieve all the desired symbols from databases 2 through 4. Thus, the queries sent from P_4 to P_3 are generated as follows,

$$Q_{3,1}^{[1,4,5]} = [h_1(1)\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (112)$$

$$Q_{3,2}^{[1,4,5]} = [h_1(1)+1\quad h_1(2)\quad h_1(3)\quad h_1(4)\quad h_1(5)]^T \quad (113)$$

$$Q_{3,3}^{[1,4,5]} = [h_1(1)\quad h_1(2)\quad h_1(3)\quad h_1(4)+1\quad h_1(5)]^T \quad (114)$$

$$Q_{3,4}^{[1,4,5]} = [h_2(1)\quad h_2(2)\quad h_2(3)\quad h_2(4)\quad h_2(5)+1]^T \quad (115)$$

The clients share the following common randomness. A global randomness $c \sim$ uniform$\{1, 2, 3, 4\}$ is shared among all databases at all clients. A local randomness $\mathbf{s}_1 = [s_1(1)\ s_1(2)\ s_1(3)]$ is shared among the databases of P_1, and similarly $\mathbf{s}_2 = [s_2(1)\ s_1(2)]$, $\mathbf{s}_3 = [s_3(1)]$ are shared among the databases of P_2 and P_3, respectively. The random variable $s_i(\ell) \sim$ uniform$\{0, 1, 2, 3, 4\}$. Finally, database 2 which is associated with P_1, generates the individual randomness $t_{1,2} = [t_{1,2}(1)\ t_{1,2}(2)\ t_{1,2}(3)]$. Similarly, at P_2, database 2 generates $t_{2,2} = [t_{2,2}(1)\ t_{2,2}(2)]$, and database 3 generates $t_{2,3}$. Each element of the common randomness $\mathbf{t}_{i,j}$ for $i = 1, 2$ and $j = 2, 3$ is generated uniformly and independently from $\mathbb{F}_5$. The variables ($\mathbf{t}_{i,j}$, $i = 1, 2$, $j = 2, 3$) are sent to P_3. The individual correlated randomness $t_{3,j}$ at P_3 is calculated as,

$$t_{3,2} = 2 - t_{1,2}(1) - t_{2,2}(1)$$
$$\Longleftrightarrow t_{1,2}(1) + t_{2,2}(1) + t_{3,2} = 2 \quad (116)$$

$$t_{3,3} = 2 - t_{1,2}(2) - t_{2,3}$$
$$\Longleftrightarrow t_{1,2}(2) + t_{2,3} + t_{3,3} = 2 \quad (117)$$

$$t_{3,4} = 2 - t_{1,2}(3) - t_{2,2}(2)$$
$$\iff t_{1,2}(3) + t_{2,2}(2) + t_{3,4} = 2 \qquad (118)$$

According to this construction, the leader receives the following answer strings from P_1,

$$A_{1,1}^{[1]} = c\left(\sum_{k=1}^{5} h_1(k)X_{1,k} + s_1(1)\right) \qquad (119)$$

$$A_{1,2}^{[1]} = c\left(\sum_{k=1}^{5} h_1(k)X_{1,k} + s_1(1) + X_{1,1} + t_{1,2}(1)\right) \quad (120)$$

$$A_{1,1}^{[4]} = c\left(\sum_{k=1}^{5} h_2(k)X_{1,k} + s_1(2)\right) \qquad (121)$$

$$A_{1,2}^{[4]} = c\left(\sum_{k=1}^{5} h_2(k)X_{1,k} + s_1(2) + X_{1,4} + t_{1,2}(2)\right) \quad (122)$$

$$A_{1,1}^{[5]} = c\left(\sum_{k=1}^{5} h_3(k)X_{1,k} + s_1(3)\right) \qquad (123)$$

$$A_{1,2}^{[5]} = c\left(\sum_{k=1}^{5} h_3(k)X_{1,k} + s_1(3) + X_{1,5} + t_{1,2}(3)\right) \quad (124)$$

Similarly, P_4 receives the following responses from P_2,

$$A_{2,1}^{[1,4]} = c\left(\sum_{k=1}^{5} h_1(k)X_{2,k} + s_2(1)\right) \qquad (125)$$

$$A_{2,2}^{[1,4]} = c\left(\sum_{k=1}^{5} h_1(k)X_{2,k} + s_2(1) + X_{2,1} + t_{2,2}(1)\right) \quad (126)$$

$$A_{2,3}^{[1,4]} = c\left(\sum_{k=1}^{5} h_1(k)X_{2,k} + s_2(1) + X_{2,4} + t_{2,3}\right) \qquad (127)$$

$$A_{2,1}^{[5]} = c\left(\sum_{k=1}^{5} h_2(k)X_{2,k} + s_2(2)\right) \qquad (128)$$

$$A_{2,2}^{[5]} = c\left(\sum_{k=1}^{5} h_2(k)X_{2,k} + s_2(2) + X_{2,5} + t_{2,2}(2)\right) \quad (129)$$

Finally, P_4 receives the following responses from P_3,

$$A_{3,1}^{[1,4,5]} = c\left(\sum_{k=1}^{5} h_1(k)X_{3,k} + s_3(1)\right) \qquad (130)$$

$$A_{3,2}^{[1,4,5]} = c\left(\sum_{k=1}^{5} h_1(k)X_{3,k} + s_3(1) + X_{3,1} + t_{3,2}\right) \quad (131)$$

$$A_{3,3}^{[1,4,5]} = c\left(\sum_{k=1}^{5} h_1(k)X_{3,k} + s_3(1) + X_{3,4} + t_{3,3}\right) \quad (132)$$

$$A_{3,4}^{[1,4,5]} = c\left(\sum_{k=1}^{5} h_1(k)X_{3,k} + s_3(1) + X_{3,5} + t_{3,4}\right) \quad (133)$$

The leader party P_4 proceeds with decoding by removing the random responses created at database 1 of all clients P_1, P_2 and P_3, i.e., it constructs $Z_{i,j}$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$ by subtracting the responses $A_{i,1}$,

$$Z_{1,1} = c(X_{1,1} + t_{1,2}(1)) = A_{1,2}^{[1]} - A_{1,1}^{[1]} \qquad (134)$$

$$Z_{1,4} = c(X_{1,4} + t_{1,2}(2)) = A_{1,2}^{[4]} - A_{1,1}^{[4]} \qquad (135)$$

$$Z_{1,5} = c(X_{1,5} + t_{1,2}(3)) = A_{1,2}^{[5]} - A_{1,1}^{[5]} \qquad (136)$$

$$Z_{2,1} = c(X_{2,1} + t_{2,2}(1)) = A_{2,2}^{[1,4]} - A_{2,1}^{[1,4]} \qquad (137)$$

$$Z_{2,4} = c(X_{2,4} + t_{2,3}) = A_{2,3}^{[1,4]} - A_{2,1}^{[1,4]} \qquad (138)$$

$$Z_{2,5} = c(X_{2,5} + t_{2,2}(2)) = A_{2,2}^{[5]} - A_{2,1}^{[5]} \qquad (139)$$

$$Z_{3,1} = c(X_{3,1} + t_{3,2}) = A_{3,2}^{[1,4,5]} - A_{3,1}^{[1,4,5]} \qquad (140)$$

$$Z_{3,4} = c(X_{3,4} + t_{3,3}) = A_{3,3}^{[1,4,5]} - A_{3,1}^{[1,4,5]} \qquad (141)$$

$$Z_{3,5} = c(X_{3,5} + t_{3,4}) = A_{3,4}^{[1,4,5]} - A_{3,1}^{[1,4,5]} \qquad (142)$$

The MP-PSI determination at P_4 concludes by evaluating the following indicators, E_j, for $j = 1, 4, 5$ as,

$$E_1 = \sum_{i=1}^{3} Z_{i,1}$$
$$= c(X_{1,1} + X_{2,1} + X_{3,1} + t_{1,2}(1) + t_{2,2}(1) + t_{3,2}) \quad (143)$$

$$E_4 = \sum_{i=1}^{3} Z_{i,4}$$
$$= c(X_{1,4} + X_{2,4} + X_{3,4} + t_{1,2}(2) + t_{2,3} + t_{3,3}) \quad (144)$$

$$E_5 = \sum_{i=1}^{3} Z_{i,5}$$
$$= c(X_{1,5} + X_{2,5} + X_{3,5} + t_{1,2}(3) + t_{2,2}(2) + t_{3,4}) \quad (145)$$

By observing that the sum of the correlated randomness in E_j according to (116)–(118) is equal 2, we note that $E_j = 0$ if and only if $\sum_{i=1}^{3} X_{i,j} = 3$, i.e., if $X_{1,j} = X_{2,j} = X_{3,j} = 1$ simultaneously. Consequently, $E_1 = E_4 = 0$ irrespective to c, while $E_5 \neq 0$ and P_4 can reliably calculate $\cap_{i=1,2,3,4} \mathcal{P}_i = \{1, 4\}$. On the other hand, for E_5, $X_{1,5} + X_{2,5} + X_{3,5} + t_{1,4} + t_{2,4} + t_{3,4}$ is equal to 2 and then E_5 must be one value in the set $\{1, 2, 3, 4\}$ depending on the value of c. Now, we calculate the value of the expression $X_{1,5} + X_{2,5} + X_{3,5}$ from the perspective of the leader party P_4. If E_5 is 1, $\mathbb{P}(X_{1,5} + X_{2,5} + X_{3,5} = l) = \frac{1}{4}, \forall l = \{0, 1, 2, 3\}$ because c is uniformly distributed over $\{1, 2, 3, 4\}$. The conclusion is exactly the same when E_5 is equal to 2, 3 or 4. Thus, the only information that P_4 can obtain for the element 5 is that client parties P_1, P_2 and P_3 cannot contain it at the same time. The privacy of leader party is preserved because each element in the queries is uniformly distributed over the finite field $\mathbb{F}_5$. Hence, no information about P_4 is leaked from the queries. The total download cost in this case is $D = 6 + 5 + 4 = 15$, which is consistent with the download cost for the general case in (9), $D = \lceil \frac{3*2}{2-1} \rceil + \lceil \frac{3*3}{3-1} \rceil + \lceil \frac{3*5}{5-1} \rceil = 15$.

VII. CONCLUSION AND FUTURE WORK

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. We note that this work provides only an achievable scheme with no claim of optimality. A converse proof is needed to assess the efficiency of our scheme. Furthermore, several interesting directions can be pursued based on this work. First, one can investigate the MP-PSI in more general communication settings (not necessarily leader-to-clients). Second, one can study the case where the communication between the parties is done over multiple rounds (in contrast to the single round of communication in this work). Third, one can investigate the case of calculating more general set functions (not necessarily the intersection).

REFERENCES

[1] M. Freedman, K. Nissim, and B. Pinkas, "Efficient private matching and set intersection," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.*, 2004, pp. 1–19.

[2] H. Chen, K. Laine, and P. Rindal, "Fast private set intersection from homomorphic encryption," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 1243–1255.

[3] D. Dachman-Soled, T. Malkin, M. Raykova, and M. Yung, "Efficient robust private set intersection," in *Proc. Int. Conf. Appl. Cryptogr. Netw. Security*, 2009, pp. 125–142.

[4] E. De Cristofaro and G. Tsudik, "Practical private set intersection protocols with linear complexity," in *Proc. Int. Conf. Financ. Cryptogr. Data Security*, 2010, pp. 143–159.

[5] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," 2020. [Online]. Available: arXiv: 1912.13501

[6] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

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

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

[9] 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 *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1908–1912.

[10] R. Tajeddine and S. El Rouayheb, "Robust private information retrieval on coded data," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1903–1907.

[11] R. Bitar and S. El Rouayheb, "Staircase-PIR: Universally robust private information retrieval," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[12] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[13] Q. Wang and M. Skoglund, "Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5160–5175, Aug. 2019.

[14] Q. Wang and M. Skoglund, "Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers," in *Proc. IEEE IEEE Inf. Theory Workshop (ITW)*, Nov. 2017, pp. 71–75.

[15] Q. Wang, H. Sun, and M. Skoglund, "Symmetric private information retrieval with mismatched coded messages and randomness," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 365–369.

[16] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2999–3012, 2020.

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

[18] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM J. Appl. Algebra Geometry*, vol. 1, no. 1, pp. 647–664, 2017.

[19] Y. Zhang and G. Ge, "A general private information retrieval scheme for MDS coded databases with colluding servers," *Designs Codes Cryptogr.*, vol. 87, no. 11, pp. 2611–2623, Nov. 2019.

[20] S. Kumar, H.-Y. Lin, E. Rosnes, and A. G. I. Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4243–4273, Jul. 2019.

[21] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij–Hollanti *et al.*" *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[22] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[23] Y. Zhang and G. Ge, "Private information retrieval from MDS coded databases with colluding servers under several variant models," 2017. [Online]. Available: arXiv:1705.03186

[24] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[25] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Private information retrieval from coded storage systems with colluding, Byzantine, and unresponsive servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3898–3906, Jun. 2019.

[26] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. Allerton Conf.*, Oct. 2017, pp. 1078–1082.

[27] M. Kim, H. Yang, and J. Lee, "Cache-aided private information retrieval," in *Proc. Asilomar Conf.*, Oct. 2017, pp. 398–402.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.

[29] Y.-P. Wei, K. Banawan, and S. Ulukus, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1126–1139, Jun. 2018.

[30] 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. Commun.*, vol. 67, no. 7, pp. 4900–4912, Jul. 2019.

[31] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[32] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

[33] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8222–8231, Dec. 2019.

[34] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[35] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, Oct. 2018, pp. 180–187.

[36] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, Oct. 2018, pp. 173–179.

[37] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2020, pp. 1–6.

[38] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2023–2031, Apr. 2020.

[39] H. Sun and S. A. Jafar, "The capacity of private computation," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3880–3897, Jun. 2019.

[40] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," in *Proc. Iran Workshop Commun. Inf. Theory (IWCIT)*, Apr. 2018, pp. 1–6.

[41] Z. Chen, Z. Wang, and S. Jafar, "The asymptotic capacity of private search," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2122–2126.

[42] M. Abdul-Wahid, F. Almoualem, D. Kumar, and R. Tandon, "Private information retrieval from storage constrained databases—Coded caching meets PIR," 2017. [Online]. Available: arXiv:1711.05244

[43] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," 2018. [Online]. Available: arXiv:1805.04104

[44] K. Banawan, B. Arasli, and S. Ulukus, "Improved storage for efficient private information retrieval," in *Proc. IEEE Inf. Theory Workshop*, Aug. 2019, pp. 1–5.

[45] C. Tian, "On the storage cost of private information retrieval," 2019. [Online]. Available: arXiv:1910.11973

[46] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus, "The capacity of private information retrieval from decentralized uncoded caching databases," *Information*, vol. 10, p. 372, Dec. 2019.

[47] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3407–3416, Jun. 2020.

[48] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3590–3602, Jun. 2020.

[49] K. Banawan and S. Ulukus, "Private information retrieval from non-replicated databases," in *Proc. IEEE Int. Symp. Inf. sTheory (ISIT)*, Jul. 2019, pp. 1272–1276.

[50] K. Banawan and S. Ulukus, "Private information retrieval through wire-tap channel II: Privacy meets security," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4129–4149, Jul. 2020.

[51] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

[52] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[53] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2953–2964, 2018.

[54] Z. Jia, H. Sun, and S. Jafar, "Cross subspace alignment and the asymptotic capacity of *X*-secure *T*-private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5783–5798, Sep. 2019.

[55] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, pp. 2920–2932, 2017.

[56] 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. Inf. Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.

[57] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[58] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7628–7645, Nov. 2019.

[59] K. Banawan and S. Ulukus, "Noisy private information retrieval: On separability of channel coding and information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8232–8249, Dec. 2019.

[60] R. G. L. D'Oliveira and S. El Rouayheb, "One-shot PIR: Refinement and lifting," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2443–2455, Apr. 2020.

[61] R. Tajeddine, A. Wachter-Zeh, and C. Hollanti, "Private information retrieval over random linear networks," 2018. [Online]. Available: arXiv:1810.08941

[62] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," 2020. [Online]. Available: arXiv: 2003.13667

[63] C. Hazay and M. Venkitasubramaniam, "Scalable multi-party private set-intersection," in *Public-Key Cryptography (PKC)*. Heidelberg, Germany: Springer, 2017, pp. 175–203.

[64] U. Feige, J. Killian, and M. Naor, "A minimal model for secure computation," in *Proc. 26th Annu. ACM Symp. Theory Comput.*, 1994, pp. 554–563.

[65] Y. Zhao and H. Sun, "Expand-and-randomize: An algebraic approach to secure computation," 2020. [Online]. Available: arXiv: 2001.00539

[66] L. Kissner and D. Song, "Privacy-preserving set operations," in *Advances in Cryptology (CRYPTO)*. Heidelberg, Germany: Springer, 2005, pp. 241–257.

[67] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, May 1998, pp. 151–160.

[68] Y. Zhou, Q. Wang, H. Sun, and S. Fu, "The minimum upload cost of symmetric private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1030–1034.

[69] R. Li and C. Wu, "An unconditionally secure protocol for multi-party set intersection," in *Applied Cryptography and Network Security*. Heidelberg, Germany: Springer, 2007, pp. 226–236.

Zhusheng Wang (Student Member, IEEE) received the B.Sc. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2016, and the M.Sc. degree in electrical and computer engineering from the University of Michigan at Ann Arbor, MI, USA, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. His research interests include information theory, private information retrieval, and machine learning.

Karim Banawan (Member, IEEE) received the B.Sc. and M.Sc. degrees (Highest Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008 and 2012, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2017 and 2018, respectively, with the Ph.D. thesis on private information retrieval and security in networks.

In 2019, he joined the Department of Electrical Engineering, Alexandria University, as an Assistant Professor. His research interests include information theory, wireless communications, physical layer security, private information retrieval, and applications of machine learning in wireless networks. He was a recipient of the Distinguished Dissertation Fellowship from the Department of Electrical and Computer Engineering, University of Maryland College Park, for his Ph.D. thesis work.

Sennur Ulukus (Fellow, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University and the Ph.D. degree in electrical and computer engineering from Wireless Information Network Laboratory (WINLAB), Rutgers University.

She is the Anthony Ephremides Professor in Information Sciences and Systems with the Department of Electrical and Computer Engineering, University of Maryland at College Park, where she also holds a joint appointment with the Institute for Systems Research (ISR). Prior to joining UMD, she was a Senior Technical Staff Member of AT&T Labs-Research. She is a Distinguished Scholar-Teacher with the University of Maryland. Her research interests are in information theory, wireless communications, machine learning, signal processing and networks, with recent focus on private information retrieval, age of information, distributed coded computation, energy harvesting communications, physical layer security, and wireless energy and information transfer. She received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 IEEE Communications Society Best Tutorial Paper Award, the 2005 NSF CAREER Award, from 2010 to 2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She has been an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2019, and a Senior Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING since 2020. She was an Area Editor for the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING from 2016 to 2020, an Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS' Series on Green Communications and Networking from 2015 to 2016, an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2007 to 2010, and an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2003 to 2007. She was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2015 and 2008, JOURNAL OF COMMUNICATIONS AND NETWORKS in 2012, and the IEEE TRANSACTIONS ON INFORMATION THEORY in 2011. She is a TPC Chair/Co-Chair of the 2021 IEEE Globecom, the 2019 IEEE ITW, the 2017 IEEE ISIT, the 2016 IEEE Globecom, the 2014 IEEE PIMRC, and the 2011 IEEE CTW. She was a Distinguished Lecturer of the IEEE Information Theory Society form 2018 to 2019.