Multiparty Private Set Intersection Cardinality and Its Applications

Jiahui Gao
Arizona State University
jgao76@asu.edu

Ni Trieu
Arizona State University
nitrieu@asu.edu

Avishay Yanai
VMware Research
ay.yanay@gmail.com

ABSTRACT

We describe a new paradigm for multi-party private set intersection cardinality (PSI-CA) that allows n parties to compute the intersection size of their datasets without revealing any additional information. We explore a variety of instantiations of this paradigm. By operating under the assumption that a particular subset of parties refrains from collusion, our protocols avoid computationally expensive public-key operations and are secure in the presence of a semi-honest adversary.

We demonstrate the practicality of our PSI-CA with an implementation. For $n = 16$ parties with data-sets of 2^{20} items each, our server-aided variant takes 71 seconds. Interestingly, in the server-less setting, the same task takes only 7 seconds. To the best of our knowledge, this is the first 'special purpose' implementation of a multi-party PSI-CA from symmetric-key techniques (i.e. an implementation that does not rely on a generic underlying MPC).

We study two interesting applications – heatmap computation and associated rule learning (ARL) – that can be computed securely using a dot-product as a building block. We analyse the performance of securely computing heatmap and ARL using our protocol and compare that to the state-of-the-art.

KEYWORDS

multiparty, private set intersection cardinality, secure dot product, heatmap computation, associated rule learning

1 INTRODUCTION

Secure multi-party computation (MPC) allows a set of parties to jointly invoke a distributed computation while ensuring correctness, privacy, and more. in this work, we study Private Set Intersection Cardinality (PSI-CA), a special case of MPC, that allows multiple parties to compute the intersection size of their private sets without revealing additional information. PSI itself has been motivated by many real-world applications such as contact discovery [26]. Over the last several years PSI has become truly practical with extremely fast cryptographically secure implementations [11, 22, 35, 38]. In the setting of two parties, PSI with post-processing (a.k.a circuit-based PSI), especially PSI-CA, has recently drawn more attention with several applications, such as measuring the effectiveness of online advertising [25], limiting the spread of Child Sexual Abuse Material (CSAM) [7], and private contact tracing related to COVID-19 [6, 15, 18]. However, the state-of-the-art PSI-CA is only efficient in the two-party setting [15, 18, 25]. This work considers a natural generalization to the multi-party setting, which opens the opportunity for richer applications, like the two we showcase below. The state-of-the-art protocol for PSI-CA in the multi-party setting [10] relies on secret-shared computation [12], which might not scale well for a large number of parties. In this work we present a scalable protocol for PSI-CA in the multi-party setting *with an assumption that a particular subset of parties refrains from collusion.* This is an reasonable assumption for real-life applications especially when performance is critical so that a weaker security guarantee is applied as trade off. For example, in the Covid-19 heatmap computation with multiple mobile network operators, a large cloud computing company (e.g. Amazon AWS) can play the role of the server, and the health service running by the government can play the role of P_1. Statutes and regulations imposed upon large companies and governments reduce the likelihood of collusion among these parties. In reality, such participants can be chosen to play the role of server or leaders (i.e. P_1, P_2, or P_n depending on the protocol) when invoking the multi-party protocols we proposed.

Moreover, we present a new protocol, called DotProd, where n parties may compute a sum of element-wise products of their binary vectors without revealing any additional information. Mathematically, suppose party P_i holds the m-element vector x_i, then the parties obtain $\sum_{j=1}^{m} \prod_{i=1}^{n} x_i[j]$, where $x_i[j]$ is the j^{th} element of the vector x_i. Note that in the two-party case, the computation is exactly of the dot product $x_1 \cdot x_2$. We demonstrate the efficiency of our protocols through two real-world applications: a COVID-19 heatmap computation based on PSI-CA and an associated rule learning (ARL) based on DotProd.

In the rest of this section, we will present the related work of PSI-CA and its applications. Additionally, we will delve into the technical overview and outcomes of our proposed protocols. To establish a foundation, Section 2 presents the necessary preliminaries. Furthermore, we will introduce two novel cryptographic gadgets, namely Server-Aided Shuffled OPRF and Server-Aided OPPRF, in Section 3. We will discuss our PSI-CA protocols in Section 4 and explore practical applications in Section 5. Lastly, in Section 6, we will evaluate the performance of our PSI-CA protocols and provide a comparison with existing approaches.

1.1 State-of-the-Art for PSI Cardinality

Private Set Intersection Cardinality (PSI-CA) is a variant of PSI in which the parties learn the intersection size and nothing else. In this work, we also focus on *server-aided* PSI-CA constructions. By "server-aided", we refer to cases where the parties perform PSI-CA computation with the help of a semi-honest cloud server(s). To the best of our knowledge, this work proposes the first special-purpose

Proceedings on Privacy Enhancing Technologies 2024(2), 73–90
© 2024 Copyright held by the owner/author(s).
https://doi.org/10.56553/popets-2024-0041

PSI-CA protocols from symmetric-key techniques that work in the multi-party setting.

We start with discussing PSI-CA works in the two-party setting. Clearly, one can use circuit-based PSI [36] to implement PSI-CA. However, this generic solution is expensive due to the secure computation inside the circuit. For the special-purpose two-party PSI-CA constructions, the work [13, 25] extends the classic DH-based PSI protocol [32] to support two-party PSI-CA by having a sender shuffle the PRFs of their items before returning to the receiver. Epione [43] also proposed a protocol that is suitable to the unbalanced, client-server setting, in which the server has a large database of m_1 items and the client has a small database of m_2 items. The protocol, however, requires $O(m_1 + m_2)$ expensive public-key operations (group exponentiation). Delegated PSI-CA [18] improves the efficiency of the two-party PSI-CA protocol on the client's device, Catalic [18] proposes a delegated system in which the client (i.e. PSI-CA receiver) can shift most of its PSI-CA computation to multiple untrusted servers while preserving privacy. However, Catalic system requires at least two non-colluding cloud servers with a heavy computation/communication cost. Based on oblivious switching network (OSN), [21] proposes a two-party PSI-CA (so-called OSN-based PSI-CA) which is better than circuit-based PSI-CA protocol [36] in terms of communication cost and running time in the WAN setting. However, it has both communication and computation complexity $O(m \log(m))$ for a set size m due to the expensive OSN construction. Dittmer et al. [15] introduces a variant of two-party PSI-CA (so-called weighted PSI-CA) in which each token of the client has an associated secret weight. The weighted PSI-CA is based on cheap Function Secret Sharing (FSS) constructions [8, 9], thus it is efficient on both client's and server's sides. However, their construction assumes that there exist two non-colluding servers, each holding an identical input set. In Section 4, we show that the more straightforward version of the server-aided PSI [27] yields a fastest PSI-CA protocol in the two-party setting.

A multi-party PSI-CA protocol was first proposed by Kissner and Song [28]. The protocol of [28] is based on oblivious polynomial evaluation which is implemented using additively homomorphic encryption. The basic idea is to represent a dataset as a polynomial whose roots are its elements, and send the homomorphic encryptions of the coefficients to other parties so that they can evaluate the encrypted polynomial on their inputs. The protocol of [28] has a quadratic computation and communication complexity in both the size of dataset and the number of parties.

Mohassel et al. [34] proposed a PSI-CA protocol, but on secret shared data in the honest-majority three-party setting, which is different than the setting in this paper, as we consider a setting with any number of parties, in which the input does not have to be in a secret-sharing form. However, one can extend the protocol of [34] to support the multi-party PSI-CA where all the parties secret-share their input to the three parties of [34] which then jointly compute the final output. We discuss the extension and compare the performance of our protocol and [34]'s in Section 6.2.

Chandran et al. [10] proposed an efficient PSI (not PSI-CA), which can be extended to circuit-based PSI. Hence, one could combine their extended protocol with a circuit that computes the size of the intersection to obtain a protocol for PSI-CA. At the technical core, [10] is built on n-party secret-sharing functionalities introduced by [12]. Their use of generic secure computation protocol for a specific problem (of PSI-CA) makes their extended protocol less attractive. In addition, [10] requires 10 interaction rounds while our server-less multiparty PSI-CA protocol needs 4 rounds. We compare the performance of our protocols and [10] in Section 6.2.

Very recently, Fenske et al. [19] proposed an efficient and malicious multi-party PSI-CA procotol in the outsourcing setting. Their approach makes use of K servers, with the assumption that at least one of the servers is not colluding with other participants. When $K = 1$, their protocol is comparable to our server-aided one, as both require a non-colluding server. However, our protocol outperforms theirs in this scenario, as we employ symmetric-key operations while [19] heavily relies on the additively homomorphic encryption. For instance, in the case of $n = 8$ and $m = 2^{16}$, our semi-honest protocol can compute PSI-CA within 3 seconds. In contrast, the malicious protocol [19] requires approximately 2 hours for $n = 5$ and $m = 30000$, as indicated in their Figure 10. This demonstrates the efficiency and superiority of our approach in terms of computational time. Note that the protocol [19] only works in the server-aided setting, whereas in this work we also propose a way to work without such an entity, which we call the "server-less" setting.

1.2 Secure Dot Product and Its Applications

Dot-product plays a key role in machine learning and data analysis tasks. Its implementation in a privacy-preserving setting remains expensive as it requires either generating Beaver triples [5] or using fully homomorphic encryption (FHE). There is a long list of results for secure computation of dot product or linear algebra in general [1, 4, 14, 24, 42, 46, 47]. For the applications that we consider in this paper, namely, Covid-heatmap and ARL, dot-product of *sparse vectors* would be sufficient. Many algorithms for linear algebra operations, like matrix multiplication, leverage an apriori knowledge of the operands being sparse, and sometimes these algorithms can even be computed securely, without degrading their asymptotic complexity. None of the above works, however, address the problem of dot product in a setting where the vectors are sparse. The most relevant works to ours are [4, 16, 41, 44, 45].

To the best of our knowledge, Vaidya and Clifton [44] were the first to study secure computation of scalar product of two m-element vectors in the two-party setting and its application to privacy-preserving association rule learning (ARL). Their dot product protocol heavily relies on public-key operations, and requires four communication rounds, communication complexity of $O(m)$ and computation complexity of $O(m^2)$.

Their follow-up work [45] is based on PSI, which makes the complexity dependent only of t, where t is the upper-bound on the Hamming weight of the vectors. They also propose a protocol for the multi-party setting, which requires a commutative one-way hash function so that the input from each party can be encrypted by a common set of keys. The resulting ciphertexts are the same if the original values are the same. Although efficient, their protocol introduces an undesirable leakage; specifically, it leaks the items in the intersection (rather only their sum). Moreover, their protocol is insecure when the input domain is relatively small (e.g. of size 2^{30}) as one party could easily perform a brute force attack [37]. To handle the latter security issue, [16] studied a two-party ARL and proposed a solution via PSI that is built on the Goldwasser-Micali Encryption [23] and Oblivious Bloom Intersection [17]. Their protocol still leaks the items in the intersection, and became much more expensive than the protocol we present in this paper. In addition, they did not consider an extension to the multi-party case.

Recently, Bampoulidis et al. [4] studies COVID-19 heatmap computation and proposes secure dot product based on homomorphic encryption with several optimizations. However, the number of required HE operations is $O(m)$ (regardless of the Hamming weight of the vectors), which makes their protocol expensive. Schoppmann et al. [41] presents efficient two-party protocols for several common sparse linear algebra operations including sparse matrix-vector multiplication. The main building block of their protocols is a new functionality – Read-Only Oblivious Map (ROOM). Using ROOM, the cost of the secure matrix-vector multiplication is dependent only on the number of non-zero entries, instead of the operands' size. However, in all three ROOM constructions the parties invoke generic secure computation in order to obtain a secret-shared output. We compare the performance of our protocol to a ROOM-based dot-product in Section 6.1.

1.3 Our Results and Techniques

1.3.1 Our PSI-CA Approach: We present a new multi-party PSI-CA protocol paradigm with an assumption that a subset of particular parties does not collude. We offer two variants of our protocol. The first protocol relies on a non-colluding semi-honest server that has no input. It is optimized for the number of communication rounds between parties; that is, the protocol leverages a star network topology, where parties mostly communicate with the server. The second protocol removes the need of a server by reducing the problem of n-party PSI-CA to the problem of server-aided $(n-1)$-party PSI-CA with use of a semi-honest party P_n *who may have an input.* The base case with $n = 2$ can be instantiated efficiently by two-party server-aided PSI protocol of Kamara et al. [27]. However, [27] is only for PSI itself (not PSI-CA)[1]. We simplify their PSI protocol and present the server-aided two-party PSI-CA in Section 4.1.

The main building blocks of our multiparty PSI-CA protocols are oblivious key-value store (OKVS) data structure [22], and/or Oblivious Programmable PRF (OPPRF) [30]. To this end, we propose a very simple and efficient protocol for server-aided OPPRF, which we believe to be of independent interest. Our server-aided OPPRF is based on a two-party server-aided shuffled OPRF, a functionality we formally define in Section 3.1.

We provide an implementation of server-aided and server-less variants of our PSI-CA approach for $n > 2$. To the best of our knowledge, this is the first 'special-purpose' implementation of multi-party PSI-CA from symmetric-key techniques that does not rely on generic secure computation. We find that multi-party PSI-CA is practical, by evaluating our protocols over settings with million items sets and 16 parties. The main reason for the efficiency of our protocol is its reliance on fast symmetric-key primitives. This is in contrast with prior multi-party PSI-CA protocols, which require expensive public-key operations for each item [28] or computation on secret-shared data [10].

[1]Note that [27] has a protocol for multiparty PSI, but it reveals intersection items of each pair-wise parties sets to the server and is non-trivial to support PSI-CA.

Interestingly, the server-less PSI-CA variant is about 10× faster than the server-aided one. We consider colluding model in the semi-honest setting which is introduced in detail in Section 2.1. The two variants, however, offer different security guarantees. Specifically, the former is secure in the presence of an adversary who may passively corrupt any subset from $\{P_3, \ldots, P_n\}$ or one of P_1,P_2 or P_n (i.e. P_1, P_2 and P_n are non-colluding). The latter (server-aided PSI-CA) is secure in the presence of an adversary who may passively corrupt any strict subset of $\{P_1, P_3, \ldots, P_n\}$ or $\{P_2, P_3, \ldots, P_n\}$ (i.e. P_1 and P_2 do not collude) or passively corrupt the cloud server C. In some sense, one may look at the server-less variant as a multi-server-aided PSI-CA but the servers have their private input. Hence, we can use our efficient server-aided OPPRF (instead of the two-party OPPRF [30]) in the server-less PSI-CA protocol, which may explain why it is possible to get a better performance in this case. In the server-less variant, we assign the non-colluding party P_1 the role of a server in the server-aided OPPRF protocol.

The security model employed in this work deviates from the commonly known concept of "threshold security". Rather, we adopt a specific but sufficiently general access structure, in which a designated subset of parties does not collude. Although this approach differs from the conventional notion of threshold security, we do believe our approach can be used as a stepping stone toward achieving security in the 'standard' threshold access structure.

Note that in practice, a server-aided model can be reasonable. Performance is critical and often it makes sense given that the alternative has a weaker security guarantee. For example, in the federated learning setting, there is a server and many clients where the server helps training a machine learning model for the benefit of the clients. In this work, we motivate our protocols with two real-world applications in which using a non-colluding, but semi-honest server, makes complete sense. For example, in the Covid-19 heatmap computation, an established company (e.g. Google or Apple) can play the role of the server.

1.3.2 Our Multi-party Dot-Product of Binary Vectors (DotProd): We propose a new protocol for computing the sum of element-wise products of n sparse binary vectors (so-called multiple dot product, DotProd). Let us begin with the simpler case, where $n = 2$, known as secure dot product. One would expect a solution for a dot product of m-elements vectors to incur communication overhead of at least $O(m)$, for the very fact that the parties need to first input those elements (which usually involves some sort of encryption or secret sharing on each element). In this work, we show that the communication and computation complexity is independent of m and can be reduced to $O(t)$, where t is the upper bound on the Hamming weight of the vectors. This improvement is significant when the vectors are sparse (i.e. $t = o(m)$).

For an m-element binary vector x we define $\mathbf{idx}(x) = \{i \in [m] \mid x[i] = 1\}$ to be the set of non-zero indices in x. Suppose the receiver P_0 and the sender P_1 hold an m-element binary sparse vector x_0 and x_1, respectively. The vectors are sparse and have the number of non-empty elements bounded by $t = o(m)$. As a very simple warm-up, we consider a non-secure dot product computation with the communication complexity cost of $O(t)$. Given the input vector x_0, the receiver computes $A_0 = \mathbf{idx}(x_0)$ and the sender computes $A_1 = \mathbf{idx}(x_1)$. The sender then sends A_1 to the receiver, who is able to compute the dot product $x \cdot y$ by computing the intersection $A_0 \cap A_1$ and outputting its cardinality $|A_0 \cap A_1|$.

The main advantage of the above solution is to reduce dependency on the length of the vectors, especially when the input vectors are sparse. To compute $x_0 \cdot x_1$ securely, the parties run a private set intersection cardinality protocol (PSI-CA) where P_0 inputs A_0 and P_1 inputs A_1. This idea, however, has received little attention due to the large overhead required to compute PSI-CA. We then extend DotProd to the multi-party case. Given an input vector x_i, party P_i computes $A_i = \mathbf{idx}(x_i)$. It is easy to see that the sum of element-wise products of the vectors is equal to the size of their intersection, namely, $\sum_{j=1}^{m} \prod_{i=1}^{n} x_i[j] = |\bigcap_{i=1}^{n} A_i|$. We implement the multi-party DotProd using our multi-party PSI-CA.

1.3.3 Application to PSI-CA and DotProd: We show that our PSI-CA and DotProd techniques can be used to implement and improve the performance of several privacy-preserving applications. More specifically, we consider two running examples: COVID-19 heatmap computation and associated rule learning (ARL).

In the COVID-19 heatmap problem, we consider a scenario where the Department of Health and Human Services (HHS) wants to learn areas with a higher chance of getting infected with the disease without knowing the travel route of infected individuals. The heatmap can be implemented by computing the vector-matrix multiplication as $x^\top Y$, where x and Y are as follows: x is a binary vector of size N, held by the HHS, such that $x[i] = 1$ if the ith user has tested positive to COVID-19 and $x[i] = 0$ otherwise; and $Y = (y_1, \ldots, y_m) \in \mathbb{Z}_2^{N \times m}$ is a user-location matrix, held by a network operator, such that the ith element of the column vector y_j indicates whether the ith user has recently visited the jth location. In that case $y_j[i] = 1$ and otherwise $y_j[i] = 0$. Clearly, $z = x^\top Y$ is an m-element vector where the ith element is equal to the number of users who have tested positive and recently visited the ith location. [4] proposes different optimizations on HE to implement a secure dot product, which still requires $O(Nm)$ independent multiplications (regardless of the Hamming weight of the vectors). In the heatmap example above, we observe that the vector x is sparse because the proportion of diagnosed individuals per day among all N subscribed individuals is small (e.g, 0.01-1% would be a large percentage [2]). Similarly, the matrix Y is also sparse due to people's localized travel habits. In Section 5.2, we apply our DotProd protocol to compute COVID-19 heatmap. In addition, [4] only supports a two-party computation between the HHS and a network provider. In real-world scenarios, there are many network providers. We modify the two-party PSI-CA protocol [27] to support heatmap computation between the HHS and multiple network providers without revealing additional information.

Second, we study associated rule learning (ARL) as an application of DotProd. ARL is a rule-based machine learning method that is used to discover rules/relations of the type $(X \Rightarrow Y)$ between variables X, Y in databases. As a typical example in the sales database of a supermarket, a rule/relation {onions, potatoes $\Rightarrow$ burger} indicates that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. In market design, such information can be used as the basis for decisions about product placements, promotional pricing, and more. However, the ARL training process requires a large transaction database, which may be collected from different sources. Thus, it is highly desirable to maintain the privacy of each source. We study a common ARL training algorithm, called Apriori [3, 40], and adapt it to the privacy-preserving setting. Most steps in Apriori can be computed locally except a step in which the parties want to compute a confidence score of how many transactions across a joint database that contains all attributes/items in both X and Y. This step can be implemented by computing a sum of bit-wise products of multiple binary vectors. We first apply multi-party DotProd for ARL and make its learning process in a privacy-preserving manner.

2 PRELIMINARIES

Computational and statistical security parameters are denoted by κ, λ, respectively. We use $[x]$ to denote the set $\{1, 2, \ldots, x\}$ and $[x, y]$ to denote the set $\{x, x+1, \ldots, y\}$. A set is a collection of distinct elements. We denote the concatenation of two bit strings x and y by $x||y$. For a pseudorandom function (PRF) F, a key k and a set A, we define $F(k, A) = \{F(k, a) \mid a \in A\}$. For an m-element binary vector x, we define $\mathbf{idx}(x) = \{i \in [m] \mid x[i] = 1\}$.

2.1 Security Model

Secure computation allows mutually untrusted parties to jointly compute a function on their private inputs without revealing any additional information. There are two classical security models: colluding model is modeled by considering a single monolithic adversary that captures the possibility of collusion between the dishonest participants; and non-colluding model is modeled by considering independent adversaries, each captures the view of each independent dishonest party. There are also two adversarial models, which are usually considered. In the semi-honest (passive) model, the adversary is assumed to follow the protocol, but may try to learn information from the protocol transcript. In the malicious (active) model, the adversary follows an arbitrary polynomial-time strategy to learn additional information. This paper introduces two variations of PSI-CA in the semi-honest model, each providing distinct security guarantees. Firstly, the server-aided variant of PSI-CA ensures security in the presence of an adversary who may passively corrupt any subset of $\{P_1, P_3, \ldots, P_n\}$ or $\{P_2, P_3, \ldots, P_n\}$ (i.e. P_1 and P_2 do not collude) or passively corrupt the server C. The "server-less" protocol guarantees security in the presence of an adversary who may passively corrupt any subset from $\{P_3, \ldots, P_n\}$ or passively corrupt one of P_1, P_2, or P_n.

2.2 Oblivious Key-Value Store (OKVS)

A Key Value Store (KVS) consists of two algorithms: i) Encode takes as input a set of (k_i, v_i) key-value pairs from the key-value domain, $\mathcal{K} \times \mathcal{V}$, and outputs an object S (or, with negligible probability, an error indicator $\perp$); ii) Decode takes as input an object S, a key x and outputs a value y.

```
EXPERIMENT 2.2.1. ( $\mathsf{Exp}^{\mathcal{A}}(\mathcal{K} = (k_1, \ldots, k_m))$ )
    (1) for $i \in [m]$: choose uniform $v_i \leftarrow \mathcal{V}$
    (2) return $\mathcal{A}(\mathsf{Encode}(\{(k_1, v_1), \ldots (k_m, v_m)\}))$
```

A KVS is correct if, for all $A \subseteq \mathcal{K} \times \mathcal{V}$ with distinct keys: i) $Pr[\mathsf{Encode}(A) = \perp]$ is negligible, and ii) if $\mathsf{Encode}(A) = S \neq \perp$ and $(k, v) \in A$ then $\mathsf{Decode}(S, k) = v$.

We say that a KVS is oblivious if for all $\mathcal{K}_1, \mathcal{K}_2$ of size m and all PPT adversaries $\mathcal{A}$: $\left|\Pr[\mathsf{Exp}^{\mathcal{A}}(\mathcal{K}_1) = 1] - \Pr[\mathsf{Exp}^{\mathcal{A}}(\mathcal{K}_2) = 1]\right| = \frac{1}{2} + \varepsilon$ where $\varepsilon \le \mathsf{negl}(\kappa)$. In other words, if the values v_i are chosen uniformly then the output of Encode hides the choice of the keys k_i. Oblivious Key-Value Store (OKVS)[22] is given in Experiment 2.2.1, where $\mathcal{A}$ is an arbitrary PPT algorithm.

2.3 Oblivious PRF (OPRF) and Programmable PRF (OPPRF)

An oblivious PRF (OPRF) [20] is a 2-party protocol in which the sender learns a PRF key k and the receiver learns the PRF values $F(k, q_1), \dots, F(k, q_m)$. Here, F is a PRF and $(q_1, \dots, q_m)$ are inputs chosen by the receiver. Functionality 1 presents a variant of OPRF where the receiver obtains outputs of multiple statically chosen queries.

FUNCTIONALITY 1. (*Oblivious PRF* - $\mathcal{F}^{m}_{\text{oprf}}$)
Parameters: A PRF F, and a bound m on the number of queries.
Behavior: Wait for distinct queries $(q_1, \dots, q_m)$ from the receiver where $q_i \in \{0,1\}^{\kappa}$. Sample a random PRF key k and give it to the sender. Give $\{F(k, q_1), \dots, F(k, q_m)\}$ to the receiver.

An oblivious programmable PRF (OPPRF) functionality is introduced by [30]. It is similar to the plain OPRF functionality except that it allows the sender to initially provide a set of points $\mathcal{P}$ which will be programmed into the PRF. Functionality 2 presents a simple version of OPPRF defined in [30]. For a comprehensive and in-depth understanding of OPPRF, we refer the reader [30, 35].

FUNCTIONALITY 2. (*Oblivious Programmable PRF* - $\mathcal{F}^{m_1,m_2}_{\text{opprf}}$)
Parameters: A PRF F, an upper bound m_1 on the number of points to be programmed, and a bound m_2 on the number of queries.
Behavior: Wait for points $\mathcal{P} = \{(a_1, t_1), \dots, (a_{m_1}, t_{m_1})\}$, with distinct keys a_i's, from the sender $\mathcal{S}$, and distinct queries $(q_1, \dots, q_{m_2})$ from the receiver $\mathcal{R}$. Run $k \leftarrow \mathsf{KeyGen}(\kappa, \mathcal{P})$. Give k to $\mathcal{S}$ and $(F(k, q_1), \dots, F(k, q_{m_2}))$ to $\mathcal{R}$.

2.4 Unconditional Zero Sharing

Unconditional zero sharing provides the parties with a sharing function $S : \{0,1\}^{\kappa} \times \{0,1\}^{\ell} \to \{0,1\}^{\kappa}$ and a key K_i for party P_i, such that for every $x \in \{0,1\}^{\ell}$, we have that $s_i = S(K_i, x)$ is P_i's random share, and $\bigoplus_{i=1}^{n} s_i = 0$. The functionality and its construction from [30] are given in Functionality 3 and Protocol 17.

FUNCTIONALITY 3. (*Zero-Sharing* - $\mathcal{F}_{\text{ZS}}$)
Parameters: n parties. The dictionary store is initialized to $\emptyset$.
Behavior: P_i obtains a zero-sharing key K_i for a sharing function S. Upon an input x from P_i, if store_x does not exist, generate random values $s_1, \dots, s_n$ where $s_i = S(K_i, x)$ s.t. $\bigoplus_{i=1}^{n} s_i = 0$ and store $\text{store}_{x,i} = s_i$ for $i \in [n]$. Output $K_i, \text{store}_{x,i}$ to P_i.

2.5 Private Set Intersection Cardinality

Private set intersection cardinality (PSI-CA) allows n parties, each holding a set of m items, to learn the intersection size of their private sets without revealing anything else. In the server-aided PSI-CA, we assume there is a distrusted server that has no input and does not collude with the parties. The server is involved in the PSI-CA protocol while learning nothing. PSI-CA and server-aided PSI-CA are formally presented in Functionality 4. The highlighted text is required for the server-aided case.

FUNCTIONALITY 4. (*PSI Cardinality* - $\mathcal{F}_{\text{PSI-CA}}$)
Parameters: n parties $P_1, \dots, P_n$; an untrusted server C; the set size m.
Behavior:
- Wait for input set X_i of m distinct items from P_i.
- Give the server C nothing.
- Give P_1 an intersection set size $|\bigcap_{i=1}^{n} X_i|$.

2.6 Secure Dot Product of Binary Vectors

Secure dot product functionality allows n parties, each holding an m-element binary vector, to learn the dot product of their private vectors without revealing any additional information. In this work, we consider the problem of the secure dot product of n binary vectors, in a server-aided setting, in which we make use of a non-colluding distrusted server. Our protocols are extremely efficient when the upper bound on the Hamming weight of the vectors, denoted t, is in $o(m)$. The dot product of n vectors $x_1, \dots, x_n$, each with m elements, is defined by $\sum_{j=1}^{m} \prod_{i=1}^{n} x_i[j]$ and is called DotProd. DotProd is presented in Functionality 5. The highlighted text is required for the server-aided case.

FUNCTIONALITY 5. (*Secure Dot Product* - $\mathcal{F}_{\text{DotProduct}}$)
Parameters: n parties: $P_{i \in [n]}$; an untrusted server C; an upper-bound t.
Behavior:
- Wait for input m-element binary vector x_i from P_i.
- Give the server C nothing.
- Give to $\sum_{j=1}^{m} \prod_{i=1}^{n} x_i[j]$ the party P_1.

3 SERVER-AIDED OPRF AND OPPRF

In this section, we introduce new OPRF and OPPRF constructions which make use of a semi-honest non-colluding cloud server.

3.1 Server-Aided Shuffled OPRF

The server-aided OPRF functionality involves a sender $\mathcal{S}$, a receiver $\mathcal{R}$ and a server C. It is defined as follows: $\mathcal{S}$ has a key-pair $k = (k_1, k_2)$ where $k_i \in \{0,1\}^{\kappa}$, $\mathcal{R}$ has a set of queries $\{y_i\}_{i \in [m]}$ and the server C has no input. $\mathcal{S}$ does not receive an output whereas $\mathcal{R}$ obtains $\{y'_{\pi(1)}, \dots, y'_{\pi(m)}\}$ where $y'_i = F'(k, y_i)$ and $\pi : [m] \to [m]$ is a random permutation. The output of C is the permutation π . Clearly, $\mathcal{R}$ cannot associate the response y'_i with the query y_i as all responses are pseudorandom. Figure 6 formally presents the ideal functionality of $\mathcal{F}^{(m)}_{\text{soprf}}$.

FUNCTIONALITY 6. (*Server-Aided Shuffled OPRF* - $\mathcal{F}^{(m)}_{\text{soprf}}$)
Parameters: $\mathcal{S}$, $\mathcal{R}$ and C, the set size m, a pseudorandom function (PRF) $F' : \{0,1\}^{2\kappa} \times \{0,1\}^{\ell} \to \{0,1\}^{\kappa}$ where $F'(k_1, k_2, x) = F(k_2, (F(k_1, x))$ where F is a PRF.
Behavior:
- Wait key $k = (k_1, k_2)$ from $\mathcal{S}$.
- Wait distinct queries $\{y_i\}_{i \in [m]}$ from $\mathcal{R}$.
- Send a random permutation $\pi : [m] \to [m]$ to C.
- Send $\{y'_{\pi(1)}, \dots, y'_{\pi(m)}\}$ to $\mathcal{R}$ where $y'_i = F'(k, y_i)$.

We first define $F'((k_1, k_2), x) = F(k_2, F(k_1, x))$ where F is a PRF. It is easy to see that F' is a PRF. In protocol $\Pi^{(m)}_{\text{soprf}}$, the $\mathcal{S}$ has the key $k = (k_1, k_2)$, so it can send k_1 to $\mathcal{R}$ and k_2 to C. We assume that the receiver and helper do not collude, thus they are unable to compute the sender's key k. Having k_1, $\mathcal{R}$ computes $Y' = F(k_1, Y)$ and sends Y' to C. The server C then computes $Y'' = F(k_2, Y')$, and applies a random permutation π on Y''. This one-round protocol takes into account the presence of a semi-honest sender/receiver.

Theorem 1. *Protocol* $\Pi^{(m)}_{\text{soprf}}$ *securely implements the functionality* $\mathcal{F}^{(m)}_{\text{soprf}}$ *in the presence of an adversary who may passively corrupt either* $\mathcal{S}, \mathcal{R}$*, or* C*.*

The formal proof of Theorem 1 is present in Appendix A.1.

PROTOCOL 7. (*Server-Aided Shuffled OPRF* - $\Pi^{(m)}_{\text{soprf}}$)
Parameters:
- Set size m; a PRF F.
- A sender $\mathcal{S}$, a receiver $\mathcal{R}$, a non-colluding semi-honest server C

Inputs:
- Sender $\mathcal{S}$ has input $k = (k_1, k_2)$
- Receiver $\mathcal{R}$ has input a set of m items $Y = \{y_1, \dots, y_m\}$
- Cloud server C has no input.

Protocol:
(1) $\mathcal{S}$ sends k_1 to $\mathcal{R}$ and k_2 to C.
(2) $\mathcal{R}$ computes $Y' = F(k_1, Y)$ and sends Y' to C.
(3) C chooses a random permutation $\pi : [m] \to [m]$. C computes $Y'' = F(k_2, Y')$ and sends a random permutation π of Y'' to $\mathcal{R}$.

3.2 Server-Aided OPPRF

The server-aided OPRF functionality involves a sender $\mathcal{S}$, a receiver $\mathcal{R}$, and a non colluding server C. It is defined as follows: $\mathcal{S}$ has a set of m_1 points $\mathcal{P} = \{(x_i, v_i)\}_{i\in[m_1]}$ with (pseudo)random $v_i \in \{0,1\}^\kappa$, and $\mathcal{R}$ has a set $Y = \{y_i\}_{i\in[m_2]}$. C has no input. Denote the set of first (resp. second) entries of the pairs in $\mathcal{P}$ by X (resp. V). $\mathcal{S}$ and C do not have an output whereas $\mathcal{R}$, for every y_i, obtains v_i iff $y_i \in X$, and some other pseudorandom value otherwise. This is denoted by $\mathcal{F}_{\text{sopprf}}$ and formally described in Functionality 8.

FUNCTIONALITY 8. (*Server-Aided OPPRF* - $\mathcal{F}^{(m_1,m_2)}_{\text{sopprf}}$)
Parameters: $\mathcal{S}$, $\mathcal{R}$ and C, m_1 the size of $\mathcal{P}$ and m_2 the number of queries.
Behavior:
- Wait for a set of m_1 points $\mathcal{P} = \{(x_i, v_i)\}_{i\in[m_1]}$ with distincts x_i's and (pseudo)random $v_i \in \{0,1\}^\kappa$, from $\mathcal{S}$.
- Wait for a set $Y = \{y_i\}_{i\in[m_2]}$ from $\mathcal{R}$.
- For every $i \in [m_2]$ set $v'_i = v_j$ if $y_i = x_j$ for some $j \in [m_1]$ and otherwise assign a random value to v'_i. Let $V' = \{v'_i\}_{i\in[m_2]}$.
- Send V' to $\mathcal{R}$.

In the protocol, $\mathcal{S}$, $\mathcal{R}$ and C invoke a *non-shuffled* version of OPRF, where $\mathcal{S}$ inputs the key $k = (k_1, k_2)$, $\mathcal{R}$ inputs Y, and a sets $Y' = \{y'_1, \dots, y'_{m_2}\}$ as a part of the output with $y'_i = F'(k, y_i)$. Then, $\mathcal{S}$ constructs an OKVS $T \leftarrow \text{Encode}(\{(x_i, F'(k, x_i) \oplus v_i\}_{i\in[m_1]})$ and sends T to $\mathcal{R}$, which outputs $w_j = y'_j \oplus \text{Decode}(T, y_j)$ for $j \in [m_2]$. In terms of the correctness, $\mathcal{R}$ obtains $v'_j = F'(k, y_j)) \oplus \text{Decode}(T, y_j)$ for all $y_j \in Y$. If $y_j = x_i$, then $\text{Decode}(T, y_i) = F'(k, x_i) \oplus v_j$, thus, $v'_j = v_i$. Otherwise, $\text{Decode}(T, y_i)$ gives $\mathcal{R}$ a pseudorandom value which makes v'_j pseudorandom as well. Since the sender can combine messages of OKVS and OPRF executions before sending them to the receiver, thus, this server-aded OPRF protocol is one-round.

Theorem 2. *Protocol* $\Pi^{(m_1,m_2)}_{\text{sopprf}}$ *securely computes the functionality* $\mathcal{F}^{(m_1,m_2)}_{\text{sopprf}}$ *in the* $\mathcal{F}^{(m)}_{\text{soprf}}$*-hybrid model, in the presence of an adversary who may passively corrupt either* $\mathcal{S}, \mathcal{R}$*, or* C*.*

The formal proof of Theorem 2 is present in Appendix A.2.

PROTOCOL 9. (*Server-Aided OPPRF* - $\Pi^{(m_1,m_2)}_{\text{sopprf}}$)
Parameters:
- Parties are sender $\mathcal{S}$, receiver $\mathcal{R}$, and a server C. Set sizes m_1, m_2. A PRF $F' : \{0,1\}^{2\kappa} \times \{0,1\}^\ell \to \{0,1\}^\kappa$ where $F'(k_1, k_2, x) = F(k_2(F(k_1, x))$ where F is a PRF.

Inputs:
- $\mathcal{S}$ has $\mathcal{P} = \{(x_i, v_i)\}_{i\in[m_1]}$ with (pseudo)random v_i's.
- $\mathcal{R}$ has the set $Y = \{y_i\}_{i\in[m_2]}$.
- C has no input.

Protocol:
(1) $\mathcal{S}$, $\mathcal{R}$, and C jointly invoke $\mathcal{F}^{(m_2)}_{\text{soprf}}$ where $\mathcal{S}$ inputs a random key $k = (k_1, k_2) \leftarrow \{0,1\}^{2\kappa}$, by which C obtains k_2 and $\mathcal{R}$ obtains k_1. Then $\mathcal{R}$ inputs Y, and obtains $Y' = \{y'_1, \dots, y'_{m_2}\}$ as output, where $y'_i = F'(k, y_i)$. *Note that we use a non-shuffled version of OPRF.*
(2) $\mathcal{S}$ constructs an OKVS over $T \leftarrow \text{Encode}(\{(x_i, F'(k, x_i) \oplus v_i\}_{i\in[m_1]})$ and sends T to $\mathcal{R}$.
(3) For every $j \in [m_2]$, $\mathcal{R}$ outputs $v'_j = y'_j \oplus \text{Decode}(T, y_j)$.

4 PSI CARDINALITY PROTOCOL

In this section we present three protocols:
- In Section 4.1, we simplify the server-aided PSI protocol of [27] and formally present their server-aided *two-party* PSI-CA protocol. Unlike previous "server-less" protocols (see Section 1.1) that are based on oblivious transfer [18] or on the Diffie Hellman proble [25, 43], which in turn are based on public-key primitives, the two-party PSI-CA protocol [27] uses only symmetric-key operations. This is possible, among other improvements, due to the replacement of their OPRF constructions with a server-aided version, which is much simpler and more efficient.
- In Section 4.2, we show an extension of the protocol to the multi-party case, where the adversary may passively corrupt (almost) any strict subset of the parties or passively corrupt the server. To the best of our knowledge, this is the first 'special-purpose' protocol for privately computing the intersection cardinality of more than two parties, for which we present interesting applications (see Section 5).
- In Section 4.3, we show that a server is not necessary when some parties are assumed to be semi-honest and non-colluding.

4.1 Server-Aided Two-Party PSI-CA [27]

We consider sender $\mathcal{S}$ and receiver $\mathcal{R}$ who want to compute the intersection size of their private sets $X = \{x_1, \dots, x_{m_1}\}$ and $Y = \{y_1, \dots, y_{m_2}\}$, respectively. To do so, they use a non-colluding, semi-honest cloud server C. The formal description is given in Protocol 10. The protocol is inspired by the size-hiding server-aided PSI of Kamara et al. [27]. For completeness, a description of their PSI protocol is given in Appendix E.

For correctness, notice that for a value $z \in X \cap Y$, the value $F(k, z)$ appears in both X' and Y'. On the other hand, if $z \notin X$ then $F(k, z) \notin X'$; and if $z \notin Y$ then $F(k, z) \notin Y'$.

Length of OPRF outputs. The length of OPRF output influences the probability of a collision within the protocol. Similar to previous work [29], it is sufficeint to have the output length of $\lambda + \log(m_1 m_2)$ to bound the probability of any spurious collision to 2^{λ}.

The protocol is one-round and extremely efficient because of the efficiency of the shuffled $\mathcal{F}_{\mathsf{soprf}}$. In terms of communication cost, it only requires $\mathcal{S}$ to send m_1 values to $\mathcal{R}$. The construction for $\mathcal{F}_{\mathsf{soprf}}$, in turn, requires only m_2 messages from $\mathcal{R}$ to C and m_2 messages back from C to $\mathcal{R}$. We prove the following:

PROTOCOL 10. (*Server-Aided Two-party PSI-CA*)

Parameters:

- The protocol runs between a sender $\mathcal{S}$, a receiver $\mathcal{R}$, and a server C. $\mathcal{S}$ and $\mathcal{R}$ have input size of m_1 and m_2, resp. A PRF $F' : \{0,1\}^{2\kappa} \times \{0,1\}^{\ell} \to \{0,1\}^{\kappa}$.

Inputs:

- Sender $\mathcal{S}$ has input $X = \{x_1, \ldots, x_{m_1}\}$
- Receiver $\mathcal{R}$ has input $Y = \{y_1, \ldots, y_{m_2}\}$
- Cloud server C has no input.

Protocol:

(1) $\mathcal{S}, \mathcal{R}$, and C jointly invoke $\mathcal{F}^{(m_2)}_{\mathsf{soprf}}$ as follows: $\mathcal{S}$ inputs a random key $k = (k_1, k_2) \in \{0,1\}^{2\kappa}$, upon which $\mathcal{R}$ obtains k_1 and C obtains k_2 and π (recall that $\pi : [m_2] \to [m_2]$ is a random permutation). Then $\mathcal{R}$ inputs Y and obtains $Y' = \{y'_{\pi(1)}, \ldots, y'_{\pi(m_2)}\}$, where $y'_{\pi(i)} = F'(k, y_i)$.
(2) $\mathcal{S}$ sends a random permutation of $X' = F'(k, X)$ to $\mathcal{R}$.
(3) $\mathcal{R}$ outputs $|X' \cap Y'|$.

THEOREM 3. *Protocol 10 securely implements the Functionality 4 ($\mathcal{F}_{\mathsf{PSI-CA}}$) with $n = 2$ in the $\mathcal{F}_{\mathsf{soprf}}$-hybrid model, in the presence of an adversary who may passively corrupt either $\mathcal{S}$, $\mathcal{R}$, or C.*

The formal proof of Theorem 3 is present in Appendix A.3.

4.2 Server-Aided Multi-Party PSI-CA

In this section, we assume that all parties have the same set size m. Protocol 11 may be seen as if we have one receiver, who is P_1, and multiple senders, who are $P_2, \ldots, P_n$. The role of the server is to shuffle PRF results from the senders before delivering them to the receiver. As a simplification to Protocol 11, suppose that we want the receiver to obtain $n-1$ shares of zero for each of its items that is in the intersection. This can be done by querying the senders on each of their items and collecting the results. Each sender programs the responses such that if the query is on one of its items, then it responds with its (pseudorandom) share of zero, otherwise, it responds with some other pseudorandom value. Given the senders' responses on a query, if they sum up to zero then the receiver knows that its query is in the intersection. Since the server shuffles the responses to the queries, the receiver does not know, for a given set of responses which are shares of zero, to which query it is associated, thus, the output leaks nothing but the intersection size. Formally,

(1) $P_2, \ldots, P_n$ (the senders) generate keys for a zero sharing function S, so P_i obtains K_i such that for every x it holds that $\bigoplus_{i=2}^{n} S(K_i, x) = 0$.
(2) P_1 (the receiver) sends to the server its queries X_1.
(3) The server runs an OPPRF instance with every sender, using the queries X_1. A sender P_i ($i \in [2, n]$) programs the responses such that on query $x \in X_i$ the response is $S(K_i, x)$ whereas on any other query the response is another pseudorandom value.
(4) The server obtains the set $Y'_{i \in [2,n]}$, of $n-1$ OPPRF responses, on every query $x_i \in X_1$. It chooses a random permutation $\pi : [m] \to [m]$ and sends to P_1 the set $\{Y'_{\pi(1)}, \ldots, Y'_{\pi(m)}\}$.
(5) P_1 checks for every response set Y_i whether its values are valid shares of zero. If so, it adds 1 to the cardinality.

In the above simplification, there are several security issues: first, the server learns P_1's queries in the clear; second, the server mediates all PRF responses and therefore it learns whenever there is a set of responses that are valid shares of zero, thus it can learn the intersection size as well; third, if the receiver colludes with one of the senders, together they can reverse the server's permutation on items that are in the intersection and by that leak the intersection itself (rather than only its size).

PROTOCOL 11. (*Server-Aided Multi-Party PSI-CA*)

Parameters:

- The protocol runs between parties $P_1, \ldots, P_n$ for $n > 2$, and a cloud server C. A PRF $F : \{0,1\}^{\kappa} \times \{0,1\}^{\ell} \to \{0,1\}^{\kappa}$. A PRG $G : \{0,1\}^{\kappa} \to \{0,1\}^{\star}$.

Inputs:

- P_i has $X_i = \{x_{i,1}, \ldots, x_{i,m}\}$.
- Cloud server C has no input.

Protocol:

(1) Parties $P_2, \ldots, P_n$ invoke $\mathcal{F}_{\mathsf{ZS}}$ (Functionality 3 with no input) and each party P_i obtains the key K_i for a sharing function S.
(2) P_1 chooses a random PRG seed $\gamma_0 \leftarrow \{0,1\}^{\kappa}$ and sends it to P_2. Both P_1 and P_2 generate m values $\Gamma = (\gamma_1, \ldots, \gamma_m) \leftarrow G(\gamma_0)$ where $\gamma_i \in \{0,1\}^{\kappa}$.
(3) P_1 chooses a random PRF key k and sends it to $P_2, \ldots, P_n$.
(4) Party P_i for $i \in [2, n]$ computes the set of points $\mathcal{P}_i$ where:
   - For $i = 2$, $\mathcal{P}_2 = \{(F(k, x_{2,j}), S(K_2, x_{2,j}) \oplus \gamma_{\pi(j)})\}_{j \in [m]}$ where $\pi : [m] \to [m]$ is a random permutation chosen by P_2.
   - For $i \in [3, n]$, $\mathcal{P}_i = \{(F(k, x_{i,j}), S(K_i, x_{i,j}))\}_{j \in [m]}$.
(5) P_1 sends $X'_1 = F(k, X_1) = \{F(k, x_{1,j})\}_{j \in [m]}$ to C.
(6) C and P_i (for every $i \in [2, n]$) invoke $\mathcal{F}_{\mathsf{opprf}}$, where P_i acts as a sender with input $\mathcal{P}_i$ and C acts as a receiver with input X'_1. C obtains the result $y_{i,j}$ on the query $x_{1,j}$.
(7) For every $j \in [m]$, C computes $w_j = \bigoplus_{i=2}^{n} y_{i,j}$ and sets W to be a random permutation of $\{w_1, \ldots, w_m\}$. C sends W to P_1.
(8) P_1 outputs $|W \cap \Gamma|$.

The first issue is easily solved by having all parties $P_1, \ldots, P_n$ agree on a PRF key k, so instead of computing $|\bigcap_{i=1}^{n} X_i|$ their objective is to compute $|\bigcap_{i=1}^{n} F(k, X_i)|$. This way, the server does not know P_1's set. Hiding the intersection size from the server (the second issue above) is trickier. We solve it by having P_1 and P_2 agree on a set of random values $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ so that instead of programming the responses with the 'zero shares', on a value $x \in X_2$, P_2 programs the response $S(K_2, x) \oplus \gamma$ for some $\gamma \in \Gamma$. Now, for items that are in the intersection, the server C sees a set of responses that constitutes a valid share of some $\gamma \in \Gamma$, but since the C does not know Γ, this looks random indistinguishable from the responses on values that are not in the intersection. Finally, we propose a protocol under a relaxed setting, that solves the third issue above. Concretely, the protocol is secure as long as P_1 and P_2 do not collude. This is done by adding one step to the above description: before the server forwards the responses set W to P_1, it sums its items and forwards only the sum to P_1. This means that now P_i ($i \geq 3$) could not trace back and learn the intersection itself. This is formally presented in Protocol 11. The initial five steps of the protocol need only one round of communication as the involved parties can consolidate their messages prior to transmission. The OPPRF [30] demands two rounds. Step 7 involves a single round. Consequently, this server-aid protocol requires a total of four rounds in entirety.

Theorem 4. *Protocol 11 securely implements the Functionality 4 ($\mathcal{F}_{\text{PSI-CA}}$) for arbitrary n, in the ($\mathcal{F}_{\text{opprf}}$, $\mathcal{F}_{\text{ZS}}$)-hybrid model, in the presence of an adversary who may passively corrupt any subset of $\{P_1, P_3, \ldots, P_n\}$ or $\{P_2, P_3, \ldots, P_n\}$, or passively corrupt the server C.*

We note that, in our protocol, parties use zero shares to mask their actual input. This step is similar to the one in [22]. The formal proof of the Theorem 4 is present in Appendix A.4.

PROTOCOL 12. (*Multi-Party PSI-CA*)

Parameters:

- The protocol runs between parties $P_1, \ldots, P_n$ for $n > 2$, and a cloud server C. A PRF $F : \{0,1\}^\kappa \times \{0,1\}^\ell \to \{0,1\}^\kappa$.

Inputs: P_i has $X_i = \{x_{i,1}, \ldots, x_{i,m}\}$.

Protocol:

(1) Parties $P_2, \ldots, P_n$ invoke $\mathcal{F}_{\text{ZS}}$ (Functionality 3) and each party P_i obtains the key K_i for a sharing function S.
(2) P_1 chooses a random $s \leftarrow \{0,1\}^\kappa$ and sends it to P_2.
(3) P_2 chooses a random $k \leftarrow \{0,1\}^\kappa$ and sends it to $P_3, \ldots, P_n$.
(4) Party P_i for $i \in [2, n-1]$ computes the set of points $\mathcal{P}_i$ where:
  - $\mathcal{P}_2 = \{(F(k, x_{2,j}), S(K_2, x_{2,j}) \oplus F(s, x_{2,j}))\}_{j \in [m]}$.
  - For $i \in [3, n-1]$, $\mathcal{P}_i = \{(F(k, x_{i,j}), S(K_i, x_{i,j}))\}_{j \in [m]}$.
(5) P_n and P_i (for every $i \in [2, n-1]$) invoke an instance of the server-aided OPPRF $\mathcal{F}^{(m,m)}_{\text{sopprf}}$ where:
  - P_i acts as a sender with input $\mathcal{P}_i$,
  - P_1 acts as a cloud server with no input.
  - P_n acts as a receiver with input $X'_n = F(k, X_n)$. P_n obtains the result $y_{i,j}$ on the query $x_{n,j}$.
(6) For every $j \in [m]$, P_n computes $w_j = \bigoplus_{i=2}^{n-1} y_{i,j} \oplus S(K_n, x_{n,j})$. Then, P_n sets W to be $\{w_1, \ldots, w_m\}$.
(7) P_1 and P_n invoke the server-aided $\mathcal{F}_{\text{PSI-CA}}$ functionality with P_2 as a server, where
  - P_n acts as a sender with input W
  - P_2 acts as a cloud server with no input
  - P_1 acts as a receiver with input $V = F(s, X_1)$, and obtains $|W \cap V|$.

4.3 Multi-party PSI-CA

We now describe our "server-less" multi-party PSI-CA protocol. The main idea is to convert the problem of n-party server-aided PSI-CA to the problem of $(n-1)$-party with the use of an untrusted party P_n who, however, has a private input set X_n. Recall that in the server-aided PSI-CA protocol, the cloud server C has no input, but obtains from P_1 the PRF values $F(k, X_1)$ which are used to invoke an OPPRF with parties $P_{i \in [2,n]}$. In the problem of $(n-1)$-parties, however, party P_n (who plays the role of C) *does have* input X_n. Thus, P_n can compute its PRF values $F(k, X_n)$ on its own since it knows k. Similar to the server-aided version, P_n computes the exclusive-or of the OPPRF results and its zero share $S(K_n, x_{n,j})$, with the j-th result denoted by w_j. Note that w_j is equal to γ_j if all parties $P_{i \in [2,n]}$ has $x_{n,j}$, otherwise, w_j is random. At this point, if P_n sends all values w_j to P_1, P_1 can only compute the intersection size of $n-1$ sets $\bigcap_{i=2}^n X_i$ since there was nothing to do with the input set X_1.

Instead, to have P_1 output $|\bigcap_{i=1}^n X_i|$, we propose the following steps. Instead of using a random set Γ in Step (2) of Protocol 11, P_2 uses PRF to compute $\gamma_j \leftarrow F(s, x_{2,j})$ where s is known by only P_1 and P_2. We observe that if $x_{1,j}$ is an intersection item, the corresponding PRF value $F(s, x_{1,j})$ should be equal to a value w_k hold by P_n because of $w_k = \gamma_k = F(s, x_{2,k}) = F(s, x_{1,j})$. Therefore, the intersection size $|\bigcap_{i=1}^n X_i|$ can be computed by counting how many PRF values $F(s, x_{1,j})$ are in the set $W = \{w_1, \ldots, w_n\}$. P_1 and P_n can do this by invoking a two-party PSI-CA, where P_1 acts as a receiver with an input set $\{F(s, X_1)\}$ and P_n acts as a sender with an input set W.

We implement the two-party PSI-CA using our server-aid protocol described in Protocol 10 in which any party $P_{i \in [2,n-1]}$ (say P_2) can play the role of the cloud server. The two party PSI-CA Protocol 10 requires that both sender and receiver do not collude with the semi-honest server. Thus, in our multi-party protocol, we assume that P_2 is semi-honest and non-colluding with both P_1 and P_n. In addition, given this assumption, we can improve the performance of our multi-party OPPRF. Particularly, unlike Protocol 11 in the above section, we use our server-aided OPPRF construction described in Section 3.2 to execute an OPPRF instance between P_n and each $P_{i \in [2,n-1]}$, where P_1 plays the role of the OPPRF server (thus, P_1 is non-colluding). We formally present our server-less multi-party PSI-CA in Protocol 12, and its security statement below (see the formal proof in Appendix A.5). Similar to our server-aided PSI-CA, this server-less protocol is 4-round.

Theorem 5. *Protocol 12 securely implements the Functionality 4 ($\mathcal{F}_{\text{PSI-CA}}$) for arbitrary n, in the ($\mathcal{F}_{\text{sopprf}}$, $\mathcal{F}_{\text{ZS}}$, $\mathcal{F}_{\text{PSI-CA}}$, server-aided two-party PSI-CA)-hybrid model, in the presence of an adversary who may passively corrupt any subset from $\{P_3, \ldots, P_n\}$ or passively corrupt P_1 or P_2 (i.e. P_1 and P_2 are non-colluding).*

5 APPLICATIONS

We demonstrate that our PSI-CA can be used for several privacy-preserving applications by implementing two running example applications which are built on the two-party PSI-CA protocol [27] and multi-party PSI-CA protocols, respectively.

5.1 Secure Dot Product Construction

Given a secure protocol for computing the cardinality of the intersection of the parties' sets, the protocol for dot product is simple. Let x_i be an m-element binary vector of party P_i, and let $A_i = \mathbf{idx}(x_i)$. It is easy to see that the dot product of the x_i's is exactly the cardinality of the intersection of the A_i's, that is, $\sum_{j=1}^m \prod_{i=1}^n x_i[j] = |\bigcap_{i=1}^n A_i|$. Thus, to securely compute the dot product, we can use the PSI-CA functionality described in the previous section. Note that even though the input size is $O(m)$, the communication complexity of the protocol is only $O(t)$, which makes it extremely efficient when

$t = o(m)$, where t is the upper bound on the Hamming weight of the vectors.

One subtle issue is that in the PSI-CA protocols the parties know the number of elements in each other's set, which leaks more information than required. Here, we assume that there is a known upper bound, t, on the Hamming weight of the vectors X_i's, and require that the parties' input to the PSI-CA contains exactly t items. That is, if the Hamming weight of X_i is $t' < t$ then P_i adds random "dummy" items to its input to the PSI-CA. Formally, for a given upper bound t, P_i inputs A_i to the PSI-CA where $A_i \leftarrow \mathbf{idx}'(X_i, t)$ and $\mathbf{idx}'(X, t)$ is defined as follows: let t' be the Hamming weight of X, set $A = \mathbf{idx}(X)$, pick $t - t'$ random values $D = \{d_1, \ldots, d_{t-t'}\}$ from the domain $\mathcal{D} = \{m + 1, \ldots, 2^{\lambda+\log(t)} + m\}$ and output $A = A \cup D$. The choice of the domain $\mathcal{D}$ allows the collision probability of dummy items to be negligible and equals to $2^{-\lambda}$.

The formal description is given in Protocol 15 in Appendix D. Note that it is possible to compute dot product DotProd with or without the help of a cloud server C, so both variants are presented. The protocol's correctness, complexity and security follow directly from the underlying PSI-CA protocol presented in Section 4 with different corruption structures.

Theorem 4. *Protocol 15 securely implements the Functionality 5 ($\mathcal{F}_{\text{DotProduct}}$) in the ($\mathcal{F}_{\text{PSI-CA}}$)-hybrid model. In particular, if π is a protocol that securely computes $\mathcal{F}_{\text{PSI-CA}}$ in the presence of an adversary $\mathcal{A}$ then, when instantiated with π, Protocol 15 is secure in the presence of adversary $\mathcal{A}$ as well.*

5.2 Heatmap Computation

As stated in [4], the heatmap can be considered as a two-party computation between HHS and a mobile network operator (MNO). HHS has a list of individuals who have reported positive for the disease. MNO knows an approximated location data of their subscribers as the subscriber connects to a certain cell tower when traveling (unless the user does not have a phone or disconnects to their network provider). Mathematically, HHS generates a binary vector $x \in \mathbb{Z}_2^N$ which indicates whether the user $i \in [1, N]$ amongst N subscribed individuals has tested positive ($x[i] = 1$) or not ($x[i] = 0$). For each cell tower $j \in [1, m]$, the MNO initializes a vector y_j of n elements, where $y_j[i]$ corresponds to the i-th subscriber (say that HHS and MNO agree on the subscribers' identifier and on their positions in the vectors). If the i-th subscriber connects to a cell tower j within some period of time, then $y_j[i] = 1$, and $y_j[i] = 0$ otherwise. To learn how many positive individuals visit a certain area (e.g. the area covered by the j-th cell tower, HHS and MNO run a secure dot product protocol to obtain $x \cdot y_j$.

The solution proposed in [4] relies on HE to implement the secure dot product for the heatmap problem. Even with the HE optimizations, [4] requires $O(N)$ independent secure multiplications to compute $x \cdot y_j$ for each cell tower. Therefore, their protocol costs $O(mN)$ HE multiplications to compute secure vector-matrix multiplications $x \cdot Y$, where Y consists of m columns $y_1, \ldots, y_m$. Each element of $x \cdot Y$ corresponds to how many diagnosed subscribers visited a cell town.

In this work, we observe that the proportion of diagnosed individuals among all N subscribed individuals is usually small (e.g. $0.01 - 0.1\%$ new positive cases per day [2]), thus, the vector x is sparse. In addition, the vector y_j is also sparse due to people's localized travel habits. Therefore, the heatmap computation is a perfect application for our DotProd where the input vectors are sparse. By applying DotProd, we show that the computational complexity of the dot product in the heatmap example can be reduced from $O(N)$ to $O(t)$, where t is the maximum between the upper bound on the number of new positive test cases and the upper bound on the number of individuals visiting a geographical area covered by a cell tower.

Multiple MNOs. We support a heatmap computation between one HHS, P_0, and multiple MNOs, $P_1, \ldots, P_n$. For a cell tower $j \in [1, m]$, the MNO P_k ($k \in [n]$) has the vector y_j^k of N elements. $y_j^k[i]$ indicates whether a subscriber i connects to a cell tower j of the MNO P_k (we assume that the j-th cell tower of all MNOs covers the same geographical area, this should be adjusted in practice). The sum of the dot products $\sum_{k=1}^{n}(x \cdot y_j^k)$ indicates how many individuals, across different MNOs, visit a certain area. In our multi-party heatmap, if P_0 invokes DotProd with each MNO P_k where P_0's input is x and P_k's input is y_j^k, P_0 learns extra information – each term of the sum $\sum_{k=1}^{n}(x \cdot y_j^k)$. To address the issue, we modify the underlying shuffled-opprf protocol of DotProd. At the high-level idea, C computes PRF values of all MNOs $P_{k\in[n]}$, permutes them before returning to the P_0. The formal description of our multi-party heatmap computation is presented in Protocol 13.

PROTOCOL 13. (*Server-aided Heatmap Construction*)

Parameters:
- Parameters k, N, t.
- A HHS and n MNO $P_1, \ldots, P_n$, and a cloud server C
- A PRF $F : \{0,1\}^\kappa \times \{0,1\}^\star \rightarrow \{0,1\}^\kappa$

Inputs:
- A HHS P_0 has input a binary vector x of length N
- Each MNO $P_{k\in[n]}$ has input a binary matrix Y_k of size $N \times m$
- Cloud server C has no input.

Protocol $n = 1$: For each $j \in [m]$, the HHS and the MNO P_1 invoke DotProd where P_0 input is x and $P1$ input is y_j^1. The HHS outputs $x \cdot y_j^1$

Protocol $n > 1$:
1. HHS computes a set $A \leftarrow \mathbf{idx}(x)$ and pads A with dummy items to the upper-bound set size t.
2. P_0 chooses a random $s \leftarrow \{0,1\}^\kappa$ and sends it to $P_1, \ldots, P_n$.
3. For each $j \in [m]$:
   (a) $P_{k\in[n]}$ computes a set $B_k \leftarrow \mathbf{idx}(y_j^k)$, and pads B_k with dummy items to the upper-bound set size t.
   (b) Each MNO $P_{k\in[n]}$, the HHS, and the cloud server C jointly invoke a modified shuffled-OPRF:
      - P_k chooses two PRF keys $s_{k,1}, s_{k,2} \leftarrow \{0,1\}^\kappa$
      - P_k sends $s_{k,1}$ to HHS and sends $s_{k,2}$ to C
      - HHS computes and sends $A'_k = F(s_{k,1}, A)$ to C.
      - C computes $A''_k = F(s_{k,2}, A'_k)$.
      C sends a *permutation* of $A'' \leftarrow \{A''_1, \ldots, A''_n\}$ to HHS
   (c) Each $P_{k\in[n]}$ sends $B'''_k = F(s, F(s_{k,2}, F(s_{k,1}, B_k)))$ to C who sends a *permutation* of $B^\star \leftarrow \{B'''_1, \ldots, B'''_n\}$ to HHS.
   (d) HHS computes $A^\star = F(s, A'')$ and outputs $|A^\star \cap B^\star|$.

In real-world scenarios, HHS prefers to minimize bandwidth cost and computation workload on their side. Our protocol makes this happen by making use of the untrusted server. For the heatmap computation, HHS only needs to compute nmt and $2nmt$ symmetric-key operations in the two-party and multi-party settings, respectively. In terms of communication cost, HHS sends and receives $3nmt$ elements. Finally, our protocol requires only 1-round communication.

5.3 Association Rule Learning

Association rules learning (ARL) aims to discover regularities/rules between variables in transaction data. In this work, we use our DotProd protocol to mitigate information leakage in ARL when training the model on a vertical partitioning of the private database between multiple parties. We study the ARL definition in [3] and adapt it to the privacy-preserving context (see Definition 1 in Appendix B). We consider only a vertically-partitioned database since if the data is horizontally-partitioned, each party can locally compute ARL. For whom are not familiar with ARL, we provide a detailed explanation of the algorithm in Appendix C.

PROTOCOL 14. (*Privacy-Preserving ARL*)

Parameters:
- A ARL threshold τ, α attributes, empty lists $L_n, \dots, L_\alpha$.
- n parties: $P_1, \dots, P_n$.
- An DotProd functionality described in Functionality 5.
- An `apriori-gen` algorithm described in Figure 1.

Inputs: $P_{i\in[n]}$ has input a vertically-partitioned database $T_{i\in[n]}$.

Protocol:
(1) $P_{i\in[n]}$ locally computes a list L_1^i of frequent itemsets that has only 1 attribute.
(2) $P_{i\in[n]}$ invoke a DotProd with each attribute input $j_i \in L_1^i$, and add j_i into a published list L_n if the output of the DotProd is great than τ (e.g. a sum of element-wise products of multiple sparse binary vectors $T[j_i]$ as $\sum_{v=1}^{m}\prod_{i=1}^{n} T[j_i][v] > \tau$)
(3) For $k = n+1$ to α, if L_k is empty, the parties do the following:
  (a) $P_{i\in[n]}$ locally computes $C_k = \texttt{apriori-gen}(L_{k-1})$.
  (b) For each candidate $c \in C_k$, let $J = \{j_1, \dots, j_m\}$ be a set of attributes in c.
    - Assume that each $P_{i\in[n]}$ have h_i attributes $J_i = \{j_{i_1}, \dots, j_{i_{h_i}}\}$. P_i locally computes an element-wise product of multiple binary vectors $T[j_{i_v}]$ as $X_i \leftarrow \prod_{v=1}^{h_i} T[j_{i_v}]$
    - Parties invoke a DotProd execution:
      – P_i inputs X_i.
      – P_1 obtains the output s, and adds c to L_{k+1} if $s > \tau$

Privacy-preserving ARL (PPARL) consists of two subproblems (see Appendix B). The second subproblem can be publicly solved since the frequent itemsets are a part of the ARL result. According to [44], one can reduce the first subproblem of PPARL to securely computing the dot products of the binary vectors with minor leakage information. For simplicity, consider the candidate itemset has only two attributes. Let x and y represent columns in the database. i.e., $x[i] = 1$ iff row i has value 1 for attribute X (similar for y and Y). Each party P_1 and P_2 holds a vertically-partitioned database of the transaction x and y respectively. The dot product of two m-element vectors x and y as $x \cdot y = \sum_{i=0}^{m} x[i]y[i]$ is the support count which indicates how many times the itemset XY appears in the joint transaction set. The dot product computation requires the joint database from both parties, thus, it should be computed in a privacy-preserving manner. Given $s \leftarrow x \cdot y$, the parties can check whether the obtained support count is greater or equal to the threshold τ. If yes, the candidate itemset is a frequent itemset. In the ideal world, if $s < \tau$, the exact value of s is not revealed to the parties. Thus, the information is considered as leakage information in our PPARL scheme as well as previous work [16, 44]. Note that [16, 44] reveal more information than ours - they leak indexes that $x[i] = y[i] = 1$ (i.e. intersection items).

In this work, we consider n-party setting with *global* rules where every vertically-partitioned transaction database $T_{i\in[n]}$ has at least one item in the frequent itemset. Protocol 14 presents our PPARL construction which closely follows the Apriori algorithm [3, 44]. The first two steps aim to find a list of itemsets that (1) appear in the transaction set T at least τ times; and (2) every party has at least one attribute in the itemset. We denote the obtained list to be L_n. Given L_n, the party locally computes a list of candidates C_{n+1} for itemsets of size $n+1$ using the `apriori-gen` algorithm [3]. At the high-level idea, the function `apriori-gen` is done by generating a superset of possible candidate itemsets and pruning this set. We present the `apriori-gen` algorithm in Figure 1, and refer the reader to [3] for more detail. Note that `apriori-gen` is computed on the public list L_n, thus it leaks no additional information. The parties jointly execute Step (3) to compute $L_{t>n}$ until it is empty.

6 IMPLEMENTATION AND PERFORMANCE

We evaluate the performance of our PSI-CA (or DotProd) protocols and estimate the performance of heatmap computation and ARL. Protocols are evaluated under different network settings, number of parties, and input set sizes to demonstrate their scalability. Our implementation is available at https://github.com/asu-crypto/mpsica.

Choice of Parameters. We run experiments on a single machine 2× 36-core Intel Xeon 2.30GHz CPU and 256GB of RAM and simulated network using the Linux *tc* command. We consider two network settings: the LAN setting has 0.02ms round-trip latency and 10 Gbps network bandwidth; the WAN setting has 96ms round-trip latency and 200 Mbps network bandwidth. In our implementation, each party uses a separate thread to communicate with other parties. The computational security parameter $\kappa = 128$ and the statistical security parameter $\sigma = 40$. The number of parties is in a range of $\{2, 4, 8, 16\}$. The set size m of PSI-CA or the upper-bound Hamming weight t of DotProd is in $\{2^{12}, 2^{16}, 2^{20}, 2^{24}\}$.

Choice of PRF, OPPRF, and OKVS. We instantiate the PRF F using AES-NI. We use OKVS and OPPRF as a black box in the implementation. Our implementation uses the table-based OPPRF code from [30]. While there are different OKVS constructions [22], we choose the most efficient Encode and Decode of 3-cuckoo PaXoS data structure. The number of bins in the cuckoo table is $1.3m$ with 3 hash functions.

PSI-CA and DotProd protocols. Recall that the steps of PSI-CA and DotProd protocols are similar, except for a small cost overhead in Step (1) of DotProd where each party locally computes a function $\mathbf{idx}()$. In the DotProd protocol, we assume that there is a known upper bound, t, on the Hamming weight of the party's input vector X. To implement DotProd using PSI-CA, we require that the parties' input to the PSI-CA contains exactly t items. Thus, we only report the detailed computational and communication performance results of our PSI-CA protocols for the set size m. It indicates that the DotProd protocols are evaluated with the upper bound $t = m$.

6.1 Performance of Two-party Protocols

PSI-CA Protocol. We evaluate the two-party PSI-CA protocol [27] (Protocol 10) in the LAN and WAN settings. The aim is to evaluate its performance in comparison to existing work and determine the most effective candidate for our multi-party protocols, heatmap computation, and ARL among the two-party PSI-CA options. We consider both balanced and unbalanced set sizes as our heatmap computation is built on the asymmetric two-party PSI-CA. In Protocol 10, the parties do not need to involve in the entire protocol's computation. The sender $\mathcal{S}$ send $F((k_1, k_2), X)$ and k_1 to the receiver, send k_2 to the server C at the same time and complete its computation. Similarly, the C does not need to be online during the whole process. Instead, the C start its computation when receiving the $\mathcal{S}$'s key PRF k_2 and the set of receiver's queries. Therefore, we report the performance of each participant separately in Table 6 (in Appendix). We find that Protocol 10 scales well in the experiments as it contains only AES calls. For instance, the total run time with the input set size $m_1 = m_2 = 2^{20}$ is only 1.5 seconds.

Two-party PSI-CA Comparison. Both DH-based and delegated PSI-CA [18] protocols are secure against a semi-honest adversary, but the latter requires two non-colluding servers. Note that one can use the protocol proposed in [33] to implement PSI-CA, however, the protocol is much expensive compared to DH-based PSI-CA. The PSI-CA implementation of [15, 43] is not available[2], thus we exclude them from the comparison. In addition, we compare Protocol 10 with ROOM-based protocol [41]. The two-party DotProd of [41] consists of two expensive steps: ROOM and a generic dense matrix multiplication. We only report the performance of ROOM in settings where [41] performs best. We use DH-based PSI code implemented by [39] with the fastest Curve25519 implementation from `libsodium`. For a fair comparison, we run the implementation of delegated PSI-CA [18] and DH-based PSI on the same benchmark machine and network settings. Note that [18] only provides the implementation of their protocol building blocks, thus, there are no performance results on the WAN setting. The times[3] for ROOM are taken from [41, Figure 17] and [31, Table 2], initially provided for a database 50, 000 and a number of queries 5, 000 and 50, 000. Table 4 presents the performance of each PSI-CA protocol. When comparing the protocols, we find that the running time of Protocol 10 is $10 - 100\times$ faster than that of the prior works. In addition, our protocol requires $2 - 5\times$ less bandwidth cost compared to them.

Performance of Heatmap Computation. In the two-party setting, executing the heatmap computation essentially involves multiple DotProd or PSI-CA executions. Similar to [4], we want to evaluate our protocol for smaller nation-states such as New York City or Singapore which has a population around $N = 2^{23}$. Concretely, we consider a case in which the MNO has a matrix Y of size $N \times m$ and the HHS has a vector x of N, where $N = 2^{23}$ and $m = 2^{15}$. The parties need to perform m DotProd instances as $x \cdot y_{j \in [m]}$, where y_j is the j^{th} column of Y. Recall the x and y_j are binary vectors that indicate whether an individual tested positive to COVID-19, and whether this individual visited a place nearby the network town y_j, respectively. Among $N = 2^{23}$, we assume that there are $t_2 = 2^{12}$ new positive cases per day [2], and each patient visits 4 places per day on average. We run $m = 2^{15}$ instances of Protocol 10 with the MNO's set size $t_1 = 2^{14}$ and the HHS's set size $t_2 = 2^{12}$, and find that our protocol costs about 10 minutes using a *single* thread. On the other hand, [4] reports about 90 minutes but using 96 threads and stronger benchmark machine [4]. Therefore, we estimate that our protocol is at least $50\times$ faster than [4]. It is due to the fact that our protocol is based on symmetric-key operations while [4] heavily relies on public-key operations. In addition, [4] requires that the participants agree on database indices (i.e. data alignment before running heatmap computation). Using PSI-CA, we can remove this requirement. The party's input can be a set of patient/visitor ids (instead of the vector/matrix).

Performance of ARL. Based on the DotProd performance, we *estimate* the performance of our ARL. In two-party setting, each party $P_{i \in [2]}$ locally computes a list L_1^i of frequent itemsets that has only one attribute. The parties sequentially invoke DotProd to compute lists L_k of frequent itemsets that has exactly k attributes where L_{k+1} is empty (say L_{m+1} is empty). Assume that each attribute/vector in $L_{j \in [2,m]}$ has a Hamming weight t_j. Also, assume that each C_j has $|C_j|$ candidates. The performance of our ARL is $\sum_{i=2}^{m} |C_j| [\Pi_{\text{DotProduct}}^{(t_j,2)}]$, where $[\Pi_{\text{DotProduct}}^{(t_j,2)}]$ is the cost of two-party DotProd with Hamming weight t_j. According to Table 6, we estimate that our ARL would take under hours to compute ARL of the database with million records.

6.2 Performance of Multi-party Protocols

PSI-CA Protocol. The running times and communication overhead of our server-aided multiparty PSI-CA are shown in Table 5 (Appendix). The protocol is asymmetric with respect to the server, the receiver P_1 and other parties $P_{i \in [2,n]}$, thus, we report the performance results of these parties separately. In our protocol, the workload of the receiver is light as it only requires to call m AES instances. The majority of the receiver's running time is to wait for other parties to finish their work. For example, P_1 takes 34.74 seconds to compute PSI-CA (or DotProd) with $n = 8$ and $m = 2^{20}$ (or $t = 2^{20}$) in the LAN setting. Also, the server plays the role of the receiver in most OPPRFs, his communication cost is highest amongst other participants. For $n = 8$ and $m = 2^{20}$ (or $t = 2^{20}$), the PSI-CA (or DotProd) requires 3305 MB on the server's side.

Table 1 presents the performance of our "server-less" multiparty PSI-CA protocol in both LAN and WAN settings. Similar to the server-aided protocol, we separately report the performance results of P_1, P_2, P_n and other parties $P_{i \in [3,n-1]}$. Unlike server-aided protocol, this protocol only relies on OKVS (i.e. makes use of symmetric-key operations only). We find that our protocol scales to large input sets (e.g. $m = 2^{20}$) with a large number of participants (e.g. $n = 16$). For $n = 16$ and $m = 2^{20}$ (or $t = 2^{20}$), our protocol requires only 6 seconds with the total communication cost 1GB.

[2][43] requires a non-colluding server that is similar to Protocol 10, but their protocol heavily replies on DH based PSI. [15] requires two non-colluding senders, each holds an identical input set.
[3]Unknown benchmark machine

[4]an c5.24xlarge AWS EC2 instance (96 vCPU @ 3.6 GHz, 192 GiB RAM)

Table 1: Run time (in second) and communication cost (in MB) of our "server-less" multiparty PSI-CA protocols for n parties on sets of size m.

	m	$n=4$				$n=8$				$n=16$			
		P_1	P_2	$P_{(3:n-1)}$	P_n	P_1	P_2	$P_{(3:n-1)}$	P_n	P_1	P_2	$P_{(3:n-1)}$	P_n
Runtime LAN (second)	2^{12}	0.08	0.07	0.06	0.07	0.08	0.07	0.06	0.07	0.08	0.08	0.06	0.08
	2^{16}	0.40	0.37	0.21	0.33	0.41	0.38	0.22	0.34	0.43	0.39	0.23	0.36
	2^{20}	7.00	5.69	3.99	6.38	7.31	5.77	4.26	7.02	7.74	6.43	4.60	7.16
Runtime WAN (second)	2^{12}	1.52	1.33	0.06	0.75	1.74	1.54	0.06	0.96	1.74	1.55	0.06	0.97
	2^{16}	4.21	3.61	0.98	2.37	4.60	4.00	1.17	2.76	6.09	5.49	1.46	4.26
	2^{20}	22.59	21.24	11.32	20.20	34.74	33.34	19.86	32.34	60.62	59.27	36.72	58.26
Comm. Cost (MB)	2^{12}	0.52	0.28	0.16	0.71	1.02	0.28	0.16	1.85	2.02	0.29	0.16	4.12
	2^{16}	8.27	4.54	2.54	11.35	16.27	4.54	2.54	29.51	32.27	4.54	2.54	65.83
	2^{20}	132.32	72.64	40.64	181.60	260.32	72.64	40.64	472.16	516.32	72.64	40.64	1053.28

Table 2: Run time (in second) and communication cost (in MB), and round number of[10] and our protocols for 4 parties and no collusion. Each party has a set size m. The numbers of [10] are for PSI itself (not, PSI-CA).

	PSI [10] (server-less, semi-honest)			PSI-CA Protocol 11 (server-aided, semi-honest)			PSI-CA Protocol 12 (server-less, semi-honest)		
m	2^{12}	2^{16}	2^{20}	2^{12}	2^{16}	2^{20}	2^{12}	2^{16}	2^{20}
LAN	0.23	1.6	23.8	0.19	1.38	19.65	0.07	0.40	7.00
WAN	1.9	7	69.6	1.89	6.90	106.08	1.52	4.21	22.59
Comm.	3.2	49.4	790.2	3.41	53.86	967.32	0.84	13.35	213.6
Rounds	10			4			4		

Table 3: Run time (in second), communication cost (in MB) of [34] and our server-less protocol for n parties. Each party has a set size m.

	Three-party PSI-CA [34]			Our PSI-CA Protocol 12					
(n,m)	$(8,2^{14})/(4,2^{15})$	$(8,2^{18})/(4,2^{19})$	$(8,2^{22})/(4,2^{23})$	$(8,2^{14})$	$(8,2^{18})$	$(8,2^{22})$	$(4,2^{15})$	$(4,2^{19})$	$(4,2^{23})$
LAN	0.2	3.1	74	0.15	1.79	29.39	0.23	3.48	56.28
WAN	1.8	15.8	267	2.31	10.63	131.19	2.78	12.79	159.83
Comm.	32.6	521.5	8344	7.88	187.00	1008.32	6.68	106.80	1708.80

Comparison with Prior Work. The three-party PSI-CA protocol [34] can be applied to multi-party cases by letting all the n parties secret-share their set of m items to their three parties/leaders S_1, S_2, S_3, then the three leaders jointly compute the PSI-CA output. The three leaders conduct the computation in the honest-majority model, which might achieve the similar security assumption in our server-less protocol in which P_1, P_2, P_n acts as leaders. To implement a n-party PSI-CA, each having m input items, the protocol of [34] requires to run PSI-CA on the total of mn secret-shared input items. Note that [34] only consider computing the PSI-CA for two sets, each of m items. Thus, the running time and communication cost of their protocol reported in [34, Figure 8] is for computing PSI-CA on the total of $2m$ secret-shared input items. To have a fair comparison, we report the performance of ours and [34]'s protocol for the total mn input items. For example, computing PSI-CA for $n = 2^3$ parties, each with $m = \{2^{14}, 2^{18}, 2^{22}\}$, results in the computation of the total $mn \in \{2^{17}, 2^{21}, 2^{25}\}$ elements. This is equivalent to the experiential results for the two-party PSI-CA using [34] with the set size $\{2 * 2^{16}, 2 * 2^{20}, 2 * 2^{24}\}$, which are reported in [34, Figure 8] where each party has $\{2^{16}, 2^{20}, 2^{24}\}$ input items, respectively (i.e., one needs to execute the two-party PSI-CA of [34] with each input set of $mn/2$ items). Since the implementation of [34] is not publicly available, we take numbers from the publication and have the comparison with our protocol. We present the detailed performance comparison in Table 3 [5]. Our protocol shows about 1.5× faster than [34] for sufficient large m. We also note that when these leaders servers collude, our protocol only reveals the intersection items while [34] leaks all input items to the adversary.

As far as we know, [10]'s implementation is not publicly available. Thus, we take their reported run times from [10, Table 2-5]. For the most direct comparison, we used the same configured machine (2x 36-core Intel Xeon 2.30GHz 256GB of RAM) and network settings to evaluate their and our protocols. We compare our "server-less" protocol with [10] for the case of $n = 4$, one dishonestly colluding (no collusion), each with $m \in \{2^{12}, 2^{16}, 2^{20}\}$. We show an improvement of 1.6 – 4× in the run time, and 3.5 – 4× in the bandwidth cost. We report the performance numbers in Table 2. Our server-less protocol with $n = 16$ requires only 7.74s in the LAN setting and $m = 2^{20}$ (see Table 1). From Table 2, the [10] with $n = 4$ requires 23.8s in the same setting. Our protocol with $n = 16$ is already 3.07× faster than [10] with $n = 4$, thus, we do not present the comparison of the two protocols for larger n.

Performance of Heatmap Computation. The complexity of our heatmap protocol is linear in the number of MNOs. Using the suitable parameters of the two-party heatmap where each MNO has a matrix of size $2^{23} \times 2^{15}$, and HHS has a vector of size 2^{23}, we *estimate*

[5]We estimate the running time by using linear interpolation for $m < 2^{20}$ and linear extrapolation for $m > 2^{20}$ based on the running time we have in Table 1.

that our protocol takes about one hour if there are 6 MNOs involved in the protocol execution. Note that our protocol does not reveal additional information other than the output – how many patients visit a certain area. In contrast, [4] only works in the two-party setting. In real-world scenarios, there are many MNOs. If using only their protocol where the HHS executes vector-matrix multiplication with each MNO and then computes the "global" heatmap, this solution leaks extra information – the individual result of each vector-matrix multiplication.

Performance of ARL. Similar to the two-party ARL, the performance of our multi-party ARL is $\sum_{i=n}^{m} |C_j|[\Pi_{\mathsf{DotProduct}}^{(t_j,n)}]$, where $[\Pi_{\mathsf{DotProduct}}^{(t_j,n)}]$ is the cost of n-party DotProd with Hamming weight t_j. Here, we assume that each attribute/vector in $L_{j\in[n,m]}$ has a Hamming weight t_j. According to the performance of our multi-party DotProd (or multi-party PSI-CA) shown in Table 5&1, we estimate that our ARL would take under a day to compute ARL of the database with million records.

ACKNOWLEDGMENTS

The first and the third authors were partially supported by NSF awards #2101052, #2200161, #2115075, and ARPA-H SP4701-23-C-0074.

REFERENCES

[1] 2018. Outsourcing scalar products and matrix products on privacy-protected unencrypted data stored in untrusted clouds. *Information Sciences* (2018).

[2] 2023. Covid-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/ https://www.worldometers.info/coronavirus/.

[3] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. *SIGMOD Rec.* 22, 2 (June 1993), 207–216. https://doi.org/10.1145/170036.170072

[4] Alexandros Bampoulidis, Alessandro Bruni, Lukas Helminger, Daniel Kales, Christian Rechberger, and Roman Walch. 2020. Privately Connecting Mobility to Infectious Diseases via Applied Cryptography. Cryptology ePrint Archive, Report 2020/522. https://ia.cr/2020/522.

[5] Donald Beaver. 1991. Efficient Multiparty Protocols Using Circuit Randomization. In *CRYPTO (LNCS, Vol. 576)*, Joan Feigenbaum (Ed.). Springer, 420–432.

[6] Alex Berke, Michiel Bakker, Praneeth Vepakomma, Ramesh Raskar, Kent Larson, and AlexSandy' Pentland. 2020. Assessing disease exposure risk with location histories and protecting privacy: A cryptographic approach in response to a global pandemic. *arXiv preprint arXiv:2003.14412* (2020).

[7] Abhishek Bhowmick, Dan Boneh, Steve Myers, Kunal Talwar, and Karl Tarbe. 2021. The Apple PSI System. [Online; accessed 18-Sept-2021].

[8] Elette Boyle, Niv Gilboa, and Yuval Ishai. 2015. Function Secret Sharing. In *EUROCRYPT 2015, Part II (LNCS, Vol. 9057)*, Elisabeth Oswald and Marc Fischlin (Eds.). Springer, Heidelberg, 337–367. https://doi.org/10.1007/978-3-662-46803-6_12

[9] Elette Boyle, Niv Gilboa, and Yuval Ishai. 2016. Function Secret Sharing: Improvements and Extensions. In *ACM CCS 2016*, Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM Press, 1292–1303. https://doi.org/10.1145/2976749.2978429

[10] Nishanth Chandran, Nishka Dasgupta, Divya Gupta, Sai Lakshmi Bhavana Obbattu, Sruthi Sekar, and Akash Shah. 2021. Efficient Linear Multiparty PSI and Extensions to Circuit/Quorum PSI. CCS.

[11] Melissa Chase and Peihan Miao. 2020. Private Set Intersection in the Internet Setting from Lightweight Oblivious PRF. In *CRYPTO 2020, Part III (LNCS, Vol. 12172)*, Daniele Micciancio and Thomas Ristenpart (Eds.). Springer, Heidelberg, 34–63. https://doi.org/10.1007/978-3-030-56877-1_2

[12] Ivan Damgård and Jesper Buus Nielsen. 2007. Scalable and Unconditionally Secure Multiparty Computation. In *CRYPTO 2007 (LNCS, Vol. 4622)*, Alfred Menezes (Ed.). Springer, Heidelberg, 572–590. https://doi.org/10.1007/978-3-540-74143-5_32

[13] Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. 2012. Fast and Private Computation of Cardinality of Set Intersection and Union. In *Cryptology and Network Security*, Josef Pieprzyk, Ahmad-Reza Sadeghi, and Mark Manulis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 218–231.

[14] Daniel Demmler, Thomas Schneider, and Michael Zohner. 2015. ABY - A Framework for Efficient Mixed-Protocol Secure Two-Party Computation. In *NDSS 2015*. The Internet Society.

[15] Samuel Dittmer, Yuval Ishai, Steve Lu, Rafail Ostrovsky, Mohamed Elsabagh, Nikolaos Kiourtis, Brian Schulte, and Angelos Stavrou. 2020. Function Secret Sharing for PSI-CA: With Applications to Private Contact Tracing. Cryptology ePrint Archive, Report 2020/1599.

[16] Changyu Dong and Liqun Chen. 2014. A fast secure dot product protocol with application to privacy preserving association rule mining. In *PAKDD*.

[17] Changyu Dong, Liqun Chen, and Zikai Wen. 2013. When private set intersection meets big data: an efficient and scalable protocol. In *ACM CCS 2013*, Ahmad-Reza Sadeghi, Virgil D. Gligor, and Moti Yung (Eds.). ACM Press, 789–800. https://doi.org/10.1145/2508859.2516701

[18] Thai Duong, Duong Hieu Phan, and Ni Trieu. 2020. Catalic: Delegated PSI Cardinality with Applications to Contact Tracing. In *ASIACRYPT 2020, Part III (LNCS, Vol. 12493)*, Shiho Moriai and Huaxiong Wang (Eds.). Springer, Heidelberg, 870–899. https://doi.org/10.1007/978-3-030-64840-4_29

[19] Ellis Fenske, Akshaya Mani, Aaron Johnson, and Micah Sherr. 2022. Accountable Private Set Cardinality for Distributed Measurement. *ACM Trans. Priv. Secur.* 25, 4, Article 25 (jul 2022), 35 pages. https://doi.org/10.1145/3477531

[20] Michael J. Freedman, Yuval Ishai, Benny Pinkas, and Omer Reingold. 2005. Keyword Search and Oblivious Pseudorandom Functions. In *TCC*, Joe Kilian (Ed.).

[21] Gayathri Garimella, Payman Mohassel, Mike Rosulek, Saeed Sadeghian, and Jaspal Singh. 2021. Private Set Operations from Oblivious Switching. In *Public-Key Cryptography – PKC 2021*, Juan A. Garay (Ed.). Springer International Publishing, Cham, 591–617.

[22] Gayathri Garimella, Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. 2021. Oblivious Key-Value Stores and Amplification for Private Set Intersection. In *CRYPTO 2021, Part II (LNCS, Vol. 12826)*, Tal Malkin and Chris Peikert (Eds.). Springer, Heidelberg, Virtual Event, 395–425. https://doi.org/10.1007/978-3-030-84245-1_14

[23] Shafi Goldwasser and Silvio Micali. 1982. Probabilistic Encryption How to Play Mental Poker Keeping Secret All Partial Information *(STOC '82)*.

[24] Chunqiang Hu, Ruinian Li, Wei Li, Jiguo Yu, Zhi Tian, and Rongfang Bie. 2016. Efficient Privacy-Preserving Schemes for Dot-Product Computation in Mobile Computing *(PAMCO '16)*.

[25] Mihaela Ion, Ben Kreuter, Ahmet Erhan Nergiz, Sarvar Patel, Shobhit Saxena, Karn Seth, Mariana Raykova, David Shanahan, and Moti Yung. 2020. On Deploying Secure Computing: Private Intersection-Sum-with-Cardinality. In *EuroS&P*. IEEE, 370–389.

[26] Daniel Kales, Christian Rechberger, Thomas Schneider, Matthias Senker, and Christian Weinert. 2019. Mobile Private Contact Discovery at Scale. In *USENIX*.

[27] Seny Kamara, Payman Mohassel, Mariana Raykova, and Seyed Saeed Sadeghian. 2014. Scaling Private Set Intersection to Billion-Element Sets. In *Financial Cryptography and Data Security - 18th International Conference, FC 2014, Christ Church, Barbados, March 3-7, 2014, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 8437)*, Nicolas Christin and Reihaneh Safavi-Naini (Eds.). Springer, 195–215. https://doi.org/10.1007/978-3-662-45472-5_13

[28] Lea Kissner and Dawn Xiaodong Song. 2005. Privacy-Preserving Set Operations. In *CRYPTO 2005 (LNCS, Vol. 3621)*, Victor Shoup (Ed.). Springer, Heidelberg, 241–257. https://doi.org/10.1007/11535218_15

[29] Vladimir Kolesnikov, Ranjit Kumaresan, Mike Rosulek, and Ni Trieu. 2016. Efficient Batched Oblivious PRF with Applications to Private Set Intersection. In *ACM CCS 2016*, Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM Press, 818–829. https://doi.org/10.1145/2976749.2978381

[30] Vladimir Kolesnikov, Naor Matania, Benny Pinkas, Mike Rosulek, and Ni Trieu. 2017. Practical Multi-party Private Set Intersection from Symmetric-Key Techniques. In *ACM CCS 2017*, Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (Eds.). ACM Press, 1257–1272. https://doi.org/10.1145/3133956.3134065

[31] Tancrède Lepoint, Sarvar Patel, Mariana Raykova, Karn Seth, and Ni Trieu. 2020. Private Join and Compute from PIR with Default. Cryptology ePrint Archive, Report 2020/1011. https://ia.cr/2020/1011.

[32] Catherine Meadows. 1995. Formal Verification of Cryptographic Protocols: A Survey (Invited Lecture). In *ASIACRYPT'94 (LNCS, Vol. 917)*, Josef Pieprzyk and Reihaneh Safavi-Naini (Eds.). Springer, Heidelberg, 135–150. https://doi.org/10.1007/BFb0000430

[33] Peihan Miao, Sarvar Patel, Mariana Raykova, Karn Seth, and Moti Yung. 2020. Two-Sided Malicious Security for Private Intersection-Sum with Cardinality. In *CRYPTO 2020, Part III (LNCS, Vol. 12172)*, Daniele Micciancio and Thomas Ristenpart (Eds.). Springer, Heidelberg, 3–33. https://doi.org/10.1007/978-3-030-56877-1_1

[34] Payman Mohassel, Peter Rindal, and Mike Rosulek. 2020. Fast Database Joins and PSI for Secret Shared Data. In *ACM CCS 2020*, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). ACM Press, 1271–1287. https://doi.org/10.1145/3372297.3423358

[35] Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. 2020. PSI from PaXoS: Fast, Malicious Private Set Intersection. In *EUROCRYPT 2020, Part II (LNCS, Vol. 12106)*, Anne Canteaut and Yuval Ishai (Eds.). Springer, Heidelberg, 739–767. https://doi.org/10.1007/978-3-030-45724-2_25

[36] Benny Pinkas, Thomas Schneider, Oleksandr Tkachenko, and Avishay Yanai. 2019. Efficient Circuit-Based PSI with Linear Communication. In *EUROCRYPT 2019, Part III (LNCS, Vol. 11478)*, Yuval Ishai and Vincent Rijmen (Eds.). Springer, Heidelberg, 122–153. https://doi.org/10.1007/978-3-030-17659-4_5
[37] Benny Pinkas, Thomas Schneider, and Michael Zohner. 2014. Faster Private Set Intersection Based on OT Extension. In *USENIX Security 2014*, Kevin Fu and Jaeyeon Jung (Eds.). USENIX Association, 797–812.
[38] Peter Rindal and Phillipp Schoppmann. 2021. VOLE-PSI: Fast OPRF and Circuit-PSI from Vector-OLE. In *EUROCRYPT 2021, Part II (LNCS, Vol. 12697)*, Anne Canteaut and François-Xavier Standaert (Eds.). Springer, Heidelberg, 901–930. https://doi.org/10.1007/978-3-030-77886-6_31
[39] Mike Rosulek and Ni Trieu. 2021. Compact and Malicious Private Set Intersection for Small Sets. CCS. https://ia.cr/2021/1159.
[40] Cynthia Rudin. 2012. MIT Lecture Notes: Machine Learning and Statistics. https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec01.pdf.
[41] Phillipp Schoppmann, Adrià Gascón, Mariana Raykova, and Benny Pinkas. 2019. Make Some ROOM for the Zeros: Data Sparsity in Secure Distributed Machine Learning. In *ACM CCS 2019*, Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz (Eds.). ACM Press, 1335–1350. https://doi.org/10.1145/3319535.3339816
[42] Babak Siabi, Mehdi Berenjkoub, and Willy Susilo. 2019. Optimally Efficient Secure Scalar Product With Applications in Cloud Computing. *IEEE Access* (2019).
[43] Ni Trieu, Kareem Shehata, Prateek Saxena, Reza Shokri, and Dawn Song. 2020. Epione: Lightweight contact tracing with strong privacy. *arXiv* (2020).
[44] Jaideep Vaidya and Chris Clifton. 2002. Privacy Preserving Association Rule Mining in Vertically Partitioned Data *(KDD)*.
[45] Jaideep Vaidya and Chris Clifton. 2005. Secure set intersection cardinality with application to association rule mining. *Journal of Computer Security* 13 (10 2005), 593–622. https://doi.org/10.3233/JCS-2005-13401
[46] Jun Zhang, Xin Wang, Siu-Ming Yiu, Zoe Jiang, and Jin Li. 2017. Secure Dot Product of Outsourced Encrypted Vectors and its Application to SVM.
[47] Youwen Zhu, Zhikuan Wang, Bilal Hassan, Yue Zhang, Jian Wang, and Cheng Qian. 2016. Fast Secure Scalar Product Protocol with (almost) Optimal Efficiency. In *Collaborative Computing: Networking, Applications, and Worksharing*, Song Guo, Xiaofei Liao, Fangming Liu, and Yanmin Zhu (Eds.). Springer International Publishing, Cham, 234–242.

A CORRECTNESS AND SECURITY PROOF

A.1 Server-Aided Shuffled OPRF

Theorem 1. *Protocol $\Pi^{(m)}_{\mathsf{soprf}}$ securely implements its functionality $\mathcal{F}^{(m)}_{\mathsf{soprf}}$ in the presence of an adversary who may passively corrupt either $\mathcal{S}$, $\mathcal{R}$, or C.*

Proof. We exhibit simulators $\mathsf{Sim}_{\mathcal{S}}$, $\mathsf{Sim}_{\mathcal{R}}$, and Sim_{C} for simulating the view of corrupt $\mathcal{S}$, $\mathcal{R}$, and C respectively which consists of the randomness, input, output, and received messages during the execution of the protocol. And then we argue the indistinguishability of the produced transcript from the real execution.

- *Passively Corrupted* $\mathcal{S}$. $\mathcal{S}$ does not receive anything during the execution of the protocol. So it is trivial to simulate his view.
- *Passively Corrupted* $\mathcal{R}$. $\mathsf{Sim}_{\mathcal{R}}$ randomly select a pair of keys (k_1, k_2) and appends the k_1 to the view. Given the PRF F, $\mathsf{Sim}_{\mathcal{R}}$ computes $Y'' = F(k_2, F(k_1, Y))$ for the input set Y and appends a permutation of Y'' to the view. Now we argue that the view output by $\mathsf{Sim}_{\mathcal{R}}$ is indistinguishable from the real one. The way that $\mathsf{Sim}_{\mathcal{R}}$ selects keys is identical to the real execution when assuming that the receiver does not collude with the server. Outputs of the PRF given different keys are computationally indistinguishable. So this simulated view is computationally indistinguishable from the real execution.
- *Passively Corrupted* C. Sim_{C} randomly selects a pair of keys (k_1, k_2) and appends the k_2 to the view. Given the PRF F, Sim_{C} computes $Y' = F(k_1, Y)$ for the randomly selected input set Y and appends it to the view. Now we argue that the view output by Sim_{C} is indistinguishable from the real one. The way that Sim_{C} selects keys is identical to the real execution when assuming that the receiver does not collude with the server. Outputs of the PRF given are computationally indistinguishable. So this simulated view is computationally indistinguishable from the real execution.

□

A.2 Server-Aided OPPRF

Theorem 2. *Protocol $\Pi^{(m_1,m_2)}_{\mathsf{sopprf}}$ securely computes functionality $\mathcal{F}^{(m_1,m_2)}_{\mathsf{sopprf}}$ in the $\mathcal{F}^{(m)}_{\mathsf{soprf}}$-hybrid model, in the presence of an adversary who may passively corrupt either $\mathcal{S}$, $\mathcal{R}$, or C.*

Proof. We exhibit simulators $\mathsf{Sim}_{\mathcal{S}}$, $\mathsf{Sim}_{\mathcal{R}}$, and Sim_{C} for simulating the view of corrupt $\mathcal{S}$, $\mathcal{R}$, and C respectively, and argue the indistinguishability of the produced transcript from the real execution.

- *Passively Corrupted* $\mathcal{S}$. $\mathsf{Sim}_{\mathcal{S}}$ simulates the view of corrupt $\mathcal{S}$, which consists of $\mathcal{S}$'s randomness, input, output, and received messages. $\mathsf{Sim}_{\mathcal{S}}$ proceeds as follows. It chooses a random key $k = (k_0, k_1) \leftarrow \{0,1\}^{2\kappa}$, calls a simulator $\mathsf{Sim}^{\mathsf{soprf}}_{\mathcal{S}}$ of the corrupt sender in the server-aided OPRF, and appends its output to the view. Since the $\mathsf{Sim}^{\mathsf{soprf}}_{\mathcal{S}}$ is trivial (i.e. $\mathcal{S}$ does not receive any messages during the execution of $\Pi^{(m)}_{\mathsf{soprf}}$), it is easy to see the view of $\mathsf{Sim}_{\mathcal{S}}$ is computationally indistinguishable from the real execution.
- *Passively Corrupted* $\mathcal{R}$. $\mathsf{Sim}_{\mathcal{R}}$ simulates the view of corrupt $\mathcal{R}$, which consists of $\mathcal{R}$'s randomness, input, output, and received messages. $\mathsf{Sim}_{\mathcal{R}}$ proceeds as follows. It calls a $\mathsf{Sim}^{\mathsf{soprf}}_{\mathcal{R}}$ with input Y and appends the output to the view. To simulate Step 3, $\mathsf{Sim}_{\mathcal{R}}$ generates m_1 random points $(x_i, v_i) \leftarrow \{0,1\}^{\ell} \times \{0,1\}^{\kappa}$, constructs an OKVS over $T \leftarrow \mathsf{Encode}(\{(x_i, v_i)\}$, and appends it to the view.
  We now argue that the output of $\mathsf{Sim}_{\mathcal{R}}$ is indistinguishable from the real execution. For this, we formally show the simulation by proceeding with the sequence of hybrid transcripts $T0; T1; T2$, where $T0$ is the real view of S, and $T3$ is the output of $\mathsf{Sim}_{\mathcal{R}}$.
  - Hybrid 1. Let T_1 be the same as T_0, except the output of server-aid OPRF execution is replaced by the output of the $\mathsf{Sim}^{\mathsf{soprf}}_{\mathcal{R}}$. It is easy to see T_0 and T_1 are computationally indistinguishable.
  - Hybrid 2. Let T_2 be the same as T_1, except the OKVS T is constructed on randomly selected points (x_i, v_i). Since the value $F'(k, x_i) \oplus v_i$ are also pseudorandom in the real execution, the two constructed OKVS tables T are computationally indistinguishable.
- *Passively Corrupted* C. Since the C only participates in the execution of server-aid OPRF as the C, the Sim_{C} is exactly the same Sim_{C} in the server-aid OPRF. According Theorem 1, it is computationally indistinguishable from the real execution and we omit the proof here.

□

Table 4: Run time (in second), communication cost (in MB), and system requirement of the two-party PSI-CA (or DotProd) protocols: DH-based PSICA [25, 32], OSN-based PSI-CA [21], Catalic [18], ROOM as a building block in DotProd [41], and Protocol 10 (a simpler variant of the [27] PSI protocol) for the sender set size m_1 and receiver set size m_2. Cells with − denote trials that are not supported by the protocol.

	DH-PSICA [25]				OSN-based PSI-CA [21]				ROOM [41]				Catalic [18]				Protocol 10			
m_1	2^{16}		2^{20}		2^{16}		2^{20}		2^{16}		2^{20}		2^{16}		2^{20}		2^{16}		2^{20}	
m_2	2^{12}	2^{16}	2^{16}	2^{20}	2^{12}	2^{16}	2^{16}	2^{20}	2^{12}	2^{16}	2^{16}	2^{20}	2^{12}	2^{16}	2^{16}	2^{20}	2^{12}	2^{16}	2^{16}	2^{20}
LAN	8.31	10.21	112.51	191.87	-	6.56	-	84.88	14.3	144.17	-	-	6.41	8.92	85.1	166.12	0.1	0.13	1.01	1.5
WAN	11.26	11.5	150.14	248.32	-	24.57	-	284.62	-	-	-	-	-	-	-	-	1.54	2.37	4.85	8.24
Comm.	2.82	4.78	46.14	77.59	-	55.49	-	1030	863	13788	878	13837	6.29	6.29	100.66	100.66	1.18	3.15	18.87	50.33
System Req.	server-less semi-honest parties												two non-colluding servers semi-honest parties/servers				one non-colluding server semi-honest parties/servers			

Table 5: Run time (in second) and communication cost (in MB) of our server-aided multiparty PSI-CA protocols for n parties on sets of size m.

	m	$n = 4$			$n = 8$			$n = 16$		
		P_1	$P_{(2:n)}$	Server	P_1	$P_{(2:n)}$	Server	P_1	$P_{(2:n)}$	Server
Runtime LAN (second)	2^{12}	0.19	0.18	0.19	0.35	0.28	0.35	0.54	0.39	0.54
	2^{16}	1.38	1.02	1.37	2.31	1.22	2.3	4.56	1.97	4.55
	2^{20}	19.65	15.6	19.39	33.86	16.8	33.61	71.17	32.36	70.89
Runtime WAN (second)	2^{12}	1.89	1.15	1.84	2.47	1.16	2.08	3.04	1.25	2.66
	2^{16}	6.9	3.18	6.01	14.07	4.1	13.28	26.09	6.01	25.3
	2^{20}	106.08	23.98	97.49	197.35	39.43	196.87	409.13	71.26	408.65
Comm. Cost (MB)	2^{12}	0.13	1.64	5.05	0.13	1.64	11.61	0.13	1.64	24.73
	2^{16}	2	25.93	79.79	2	25.93	183.51	2	25.93	390.95
	2^{20}	32	467.66	1434.98	32	467.66	3305.62	32	467.66	7046.9

Table 6: Running time (in second) and communication cost (in MB) of the two-party PSI-CA protocol [27] (Protocol 10) for the sender set size m_1 and receiver set size m_2.

		Comm.			LAN			WAN		
m_2	m_1	Receiver	Sender	Server	Receiver	Sender	Server	Receiver	Sender	Server
2^8	2^8	0.012	0.004	0.008	0.002	0.001	0.002	0.481	0.001	0.289
	2^{10}	0.025	0.016	0.008	0.002	0.002	0.002	0.482	0.002	0.29
	2^{12}	0.074	0.066	0.008	0.008	0.006	0.006	0.679	0.005	0.486
2^{12}	2^{12}	0.197	0.066	0.131	0.01	0.005	0.008	1.066	0.005	0.681
	2^{14}	0.393	0.262	0.131	0.029	0.019	0.02	1.274	0.02	0.888
	2^{16}	1.18	1.049	0.131	0.099	0.065	0.065	1.537	0.066	1.144
2^{16}	2^{16}	3.146	1.049	2.097	0.132	0.058	0.102	2.374	0.065	1.579
	2^{18}	6.291	4.194	2.097	0.315	0.209	0.203	2.924	1.42	2.097
	2^{20}	18.874	16.777	2.097	1.007	0.583	0.553	4.853	3.732	3.597
2^{20}	2^{20}	50.332	16.777	33.554	1.501	0.964	1.206	8.235	5.745	7.681
	2^{22}	100.663	67.109	33.554	4.814	2.637	4.247	19.535	15.373	18.772
	2^{24}	301.99	268.435	33.554	19.123	9.625	17.305	66.244	54.594	64.089

A.3 Server-Aided Two-party PSI-CA

THEOREM 3. *Protocol 10 securely realizes Functionality 4 ($\mathcal{F}_{\text{PSI-CA}}$) with n = 2 in the $\mathcal{F}_{\text{soprf}}$-hybrid model, in the presence of an adversary who may passively corrupt either $\mathcal{S}$, $\mathcal{R}$, or C.*

PROOF. We exhibit simulators $\text{Sim}_{\mathcal{S}}$, $\text{Sim}_{\mathcal{R}}$, and Sim_C for simulating the view of corrupt $\mathcal{S}$, $\mathcal{R}$, and C respectively, and argue the indistinguishability of the produced transcript from the real execution.

- *Passively Corrupted $\mathcal{S}$.* $\text{Sim}_{\mathcal{S}}$ simulates the view of corrupt $\mathcal{S}$, which consists of $\mathcal{S}$'s randomness, input, output, and received messages. $\text{Sim}_{\mathcal{S}}$ proceeds as follows. It chooses a random key $k = (k_0, k_1) \leftarrow \{0,1\}^{2\kappa}$, calls a simulator $\text{Sim}_{\mathcal{S}}^{\text{soprf}}$ of the corrupt sender in the server-aided OPRF, and appends its output to the view. Since the $\text{Sim}_{\mathcal{S}}^{\text{soprf}}$ does not receive any messages in the protocol, it is easy to see the view of $\text{Sim}_{\mathcal{S}}$ and the view in the real execution are identical.
- *Passively Corrupted $\mathcal{R}$.* $\text{Sim}_{\mathcal{R}}$ simulates the view of corrupt $\mathcal{R}$, which consists of $\mathcal{R}$'s randomness, input, output, and received messages. $\text{Sim}_{\mathcal{R}}$ proceeds as follows. It calls a $\text{Sim}_{\mathcal{R}}^{\text{soprf}}$ with input Y and appends the output to the view. To simulate Step 2, $\text{Sim}_{\mathcal{R}}$ generates a random set of m_1 values $X = \{x_1, ..., x_{m_1}\}$, chooses random key $k' = (k'_1, k'_2) \leftarrow \{0,1\}^{2\kappa}$, computes $X'' = F'(k, X)$ and appends it to the view. The

PRF values X'' received from the simulator are computationally indistinguishable from the random permutation of X' received from the real execution.

- *Passively Corrupted C.* Since the C only participates in the execution of server-aid OPRF as the C, the Sim_C is exactly the same Sim_C in the server-aid OPRF. According Theorem 1, it is computationally indistinguishable from the real execution and we omit the proof here.

□

A.4 Server-Aided Multi-Party PSI-CA

Correctness. We consider three cases based on whether x is in the intersection of all sets $X_{i\in[n]}$:

- Case 1: Suppose $x \in \bigcap X_{i\in[n]}$. In other words, $\forall i \in [n], \exists x_{i,j_i} \in X_i$, such that $x_{i,j_i} = x$. Thus, we have (i) all PRF values $x'_{i,j_i} = F(k, x_{i,j_i})$ are equal which is $F(k,x)$, (ii) XORing all zero shares $S(K_i, x_{j_i})$ for $i \in [2,n]$ is equal to zero. When querying the OPPRF programmed $\mathcal{P}_{i\in[3,n]}$ using the common PRF value $x'_1 = F(k,x)$, the cloud server obtains $y_{i,j}$. Based on the correctness of OPPRF, we have $y_{i,j_i} = S(K_i, x_{j_i})$. In addition, the cloud server obtains $y_{2,j_2} = S(K_2, x_{j_2}) \oplus \gamma_{j_2}$ when querying on x'_1. Therefore, the value $w_j = \bigoplus_{i=2}^{n} y_{i,j_i}$ is equal to γ_{j_2} which belongs to the set Γ known by P_1. Thus, P_1 can count how many w_j in Γ to output the intersection size.
- Case 2: Suppose x is in X_1 and is not an element in some sets $X_{i\in[2,n]}$. Some OPPRF output $y_{i,j}$ is a random value since $F(k,x)$ was never used in the OPPRF programming process. Therefore, $w_j = \bigoplus_{i=2}^{n} y_{i,j}$ is random and does not belong to the set Γ.
- Case 3: Suppose x is an element in some sets $X_{i\in[2,n]}$, but not in X_1. Some OPPRF output $y_{i,j}$ is a random value. Therefore, $w_j = \bigoplus_{i=2}^{n} y_{i,j}$ is random and does not belong to the set Γ.

Theorem 4. *Protocol 11 securely realizes Functionality 4 ($\mathcal{F}_{\mathsf{PSI-CA}}$) for arbitrary n, in the ($\mathcal{F}_{\mathsf{opprf}}$, $\mathcal{F}_{\mathsf{ZS}}$)-hybrid model, in the presence of an adversary who may passively corrupt any subset of $\{P_1, P_3, \ldots, P_n\}$ or $\{P_2, P_3, \ldots, P_n\}$ or passively corrupt the cloud server C.*

Proof. We separate the proof to the maximal collusion, from which a security to non-maximal ones can be derived. We exhibit simulators in three different cases and argue the indistinguishability of the produced transcript from the real execution.

- *Case 1: $P_1, P_3, \ldots, P_n$ are passively corrupted.* The simulator first calls the $\mathcal{F}_{\mathsf{ZS}}$ simulator $\mathsf{Sim}^{\mathsf{ZS}}$ and appends the parties keys K_i's for a zero sharing to the view of $P_1, P_3, \ldots, P_n$. In addition, the simulator runs a PRG to generate a set of Γ with key γ_0 and appends them to the view of P_1, The simulator also appends a PRF key k to the view of $P_1, P_3, \ldots, P_n$. The simulator now calls the $\mathcal{F}_{\mathsf{opprf}}$ simulator $\mathsf{Sim}_{\mathcal{S}}^{\mathsf{OPPRF}}$ with input $\mathcal{P}_i = \{(F(k, X_i), S(K_i, X_i))\}$ and appends the output to the view of $P_3, \ldots, P_n$. Then the simulator calls the $\mathcal{F}_{\mathsf{opprf}}$ simulator $\mathsf{Sim}_{\mathcal{R}}^{\mathsf{OPPRF}}$ with input $F(k, X_1) = \{F(k, x_{1,j})\}_{j\in[m]}$ for each instance of $P_3, \ldots, P_n$, receives $Y_i = \{y_{i,j}\}_{i\in[2,\ldots,n],j\in[m]}$, computes set $W = \{w_j\}_{j\in[m]}$ where $w_j = \bigoplus_{i=2}^{n} y_{i,j}$, and appends a random permutation of W to the view of P_1 ($\{y_{2,j}\}_{j\in[m]}$ are obtained by randomly choose m values from $\{0,1\}^{\ell}$). The joint view of the parties $P_1, P_3, \ldots, P_n$ is identically distributed in the simulation, and in the real execution, the messages seen by them are identically distributed and so is the output given to P_1 (who is the only party receiving output).
- *Case 2: $P_2, P_3, \ldots, P_n$ are passively corrupted.* Most of the simulation is similar to the case above. The simulator first calls the $\mathcal{F}_{\mathsf{ZS}}$ simulator $\mathsf{Sim}^{\mathsf{ZS}}$ and appends the parties keys K_i's for a zero sharing to the view of $P_2, \ldots, P_n$. In addition, the simulator runs a PRG to generate a set of Γ with key γ_0 and appends them to the view of P_2, The simulator also appends a PRF key k to the view of $P_2, \ldots, P_n$. The simulator now calls the $\mathcal{F}_{\mathsf{opprf}}$ simulator $\mathsf{Sim}_{\mathcal{S}}^{\mathsf{OPPRF}}$ with input $\mathcal{P}_i = \{(F(k, X_i), S(K_i, X_i))\}$ and appends the output to the view of $P_2, \ldots, P_n$. This concludes the simulation. The joint view of the parties is identically distributed in the simulation and in the real execution, the messages seen by them are identically distributed and these corrupted parties do not receive outputs.
- *Case 3: C is passively corrupted.* The server C has no input or output. In the protocol it receives the pseudorandom values X' from P_1 and the pseudorandom values $y_{i,j}$ for $i \in [2,n]$ and $j \in [m]$. These $m \cdot n$ values can be easily simulated by handing C $m \cdot n$ random values. The output of all parties $P_1, \ldots, P_n$ are identically distributed in the simulation and the real execution. It only remains to argue that the view of C is computationally indistinguishable in both cases, which follows from the security of the F and the OPPRF functionality.

□

A.5 Multi-party PSI-CA

Correctness. We consider three following cases based on whether x is in the intersection of all sets $X_{i\in[n]}$:

- Case 1: Suppose $x \in \bigcap X_{i\in[n]}$. In other words, $\forall i \in [n], \exists x_{i,j_i} \in X_i$, such that $x_{i,j_i} = x$. Thus, we have (i) all PRF values $F(k, x_{i,j_i}) = F(k,x)$ are equal, (ii) XORing all zero shares $S(K_i, x_{j_i}), i \in [2,n]$ is equal to zero. When querying the OPPRF points $\mathcal{P}_{i\in[2,n]}$ using the common PRF value $F(k,x)$, the party P_n obtains y_{i,j_i}. Based on the correctness of OPPRF, we have $y_{i,j_i} = S(K_i, x)$ for $i \in [3, n-1]$ and $y_{2,j_2} = S(K_i, x) \oplus PRF(s,x)$. Therefore, the value $w = \left(\bigoplus_{i=2}^{n-1} y_{i,j_i}\right) \oplus S(K_n, x)$ is equal to $PRF(s,x)$ as $\bigoplus_{i=2}^{n} S(K_i, x) = 0$. Step (7) allows P_1 to count x to output the intersection set by checking whether $w \in F(s, X)$.
- Case 2: Suppose x is in X_1 and is not an element in some sets $X_{i\in[2,n]}$. Clearly, $w \notin F(s, X_1)$ with the high probability.
- Case 3: Suppose x is an element in some sets $X_{i\in[2,n]}$, but not in X_1. The value w_j might equal to $F(s, x_2)$ for $x_2 \in X$ or random. However, $x_2 \notin X_1$, thus $w \notin F(s, X_1)$ with the high probability.

Theorem 5. *Protocol 12 securely realizes Functionality 4 ($\mathcal{F}_{\mathsf{PSI-CA}}$) for arbitrary n, in the ($\mathcal{F}_{\mathsf{sopprf}}$, $\mathcal{F}_{\mathsf{ZS}}$, $\mathcal{F}_{\mathsf{PSI-CA}}$, server-aided two-party PSI-CA)-hybrid model, in the presence of an adversary who may passively corrupt any subset from $\{P_3, \ldots, P_n\}$ or passively corrupt P_1 or P_2 (i.e. P_1 and P_2 are non-colluding).*

Proof. We separate the proof into multiple cases, depending on the adversary's corruption. As before, we assume maximal corruption and stress that the security in the case of non-maximal corruption can be easily derived. We exhibit simulators in three different cases and argue the indistinguishability of the produced transcript from the real execution.

- *Case 1: $P_3, \dots, P_n$ are passively corrupted.* The simulator first calls the $\mathcal{F}_{\mathsf{ZS}}$ simulator $\mathsf{Sim}^{\mathsf{ZS}}$ and appends the parties keys K_i's for a zero sharing to the view of $P_3, \dots, P_n$. The simulator then appends a PRF key k to the view of $P_3, \dots, P_n$. The simulator now calls the $\mathcal{F}_{\mathsf{opprf}}$ simulator $\mathsf{Sim}^{\mathsf{OPPRF}}_{\mathcal{S}}$ with input $\mathcal{P}_i = \{(F(k, X_i), S(K_i, X_i))\}$ and appends the output to the view of $P_3, \dots, P_{n-1}$. Then the simulator calls the $\mathcal{F}_{\mathsf{opprf}}$ simulator $\mathsf{Sim}^{\mathsf{OPPRF}}_{\mathcal{R}}$ with input $F(k, X_n)$ for each instance of $P_2, \dots, P_{n-1}$, receives $Y_i = \{y_{i,j}\}_{i \in [2,\dots,n-1], j \in [m]}$, computes set $W = \{w_j\}_{j \in [m]}$ where $w_j = \bigoplus_{i=2}^{n-1} y_{i,j} \oplus S(K_n, x_{n,j})$. Then the simulator calls the two-party server-aided $\mathcal{F}_{\mathsf{PSI-CA}}$ simulator $\mathsf{Sim}^{\mathsf{PSI-CA}}_{\mathcal{S}}$ with input W, and appends the output to the view of P_n.
  This concludes the simulation. The joint view of $P_3, \dots, P_n$ is computationally indistinguishable for the simulation and in the real execution.
- *Case 2: Passively Corrupted P_2.* The simulator first calls the $\mathcal{F}_{\mathsf{ZS}}$ simulator $\mathsf{Sim}^{\mathsf{ZS}}$ and appends the key K_2's for a zero sharing to the view of P_2. It then appends the PRG seed γ_0, $\mathcal{P}_2$, and the transcript of $\mathcal{F}_{\mathsf{opprf}}$ execution with C to its view. Since the $\mathcal{F}_{\mathsf{ZS}}$ is secure, the message received from $\mathcal{F}_{\mathsf{ZS}}$ execution reveals no information to the corrupt P_2. For other views, P_2 plays as the sender and receives nothing. Thus, the view of P_2 is computationally indistinguishable for the simulation and in the real execution.
- *Case 3: Passively Corrupted P_1.* The simulator first appends a PRF key s to the view of P_1. Then the simulator calls the two-party server-aided OPPRF simulator $\mathsf{Sim}^{\mathsf{SOPPRF}}_{C}$ without input, and appends the output to the view of P_1. Finally the simulator calls the simulator $\mathsf{Sim}^{\mathsf{SPIS-CA}}_{\mathcal{R}}$ with input $V = F(s, X_1)$ and appends the output to the view of P_1. This concludes the simulation. The view of P_1 is computationally indistinguishable for the simulation and in the real execution.

□

B MULTI-PARTY ARL

Definition 1. *In the privacy-preserving ARL (PPARL) problem, there are n parties $P_1, \dots, P_n$, each holding a private vertically partitioned database of transactions $T_1, \dots, T_n$, respectively. Let $T = T_1 || \dots || T_n$ be a jointed vertically database of n parties. Let $I = \{i_1, i_2, ..., i_m\}$ be a public set of binary attributes, called items. Each transaction (row) $t \in T$ is represented as a binary vector, with $t[k] = 1$ if the transaction contains item $i_k \in I$, and $t[k] = 0$ otherwise. We say that the transaction t satisfies* $\mathbf{idx}(t)$. *Denote an association rule by $\Rightarrow$. Let $X, Y \subseteq [m]$, we consider the following association rules:*

1. *The rule $X \Rightarrow Y$ holds in T with support factor of $0 \leq s \leq 1$ iff at least s% of transactions in T satisfy $X \cup Y$*
2. *The rule $X \Rightarrow Y$ holds in T with confidence factor of $0 \leq c \leq 1$ iff at least c% of transactions in T that satisfy X also satisfy Y.*
3. *The rule $X \Rightarrow Y$ is global if every transaction in T has at least one item in $X \cup Y$.*

The goal of PPARL is to allow all parties $P_1, \dots, P_n$ to find all ***global*** *rules having high support and confidence on their jointed database T while maintaining the privacy of each individual database.*

Generally speaking, the support factor indicates how frequently the itemset appears in the dataset. The support of X with respect to T is defined as the proportion of transactions in the dataset which contains the itemset X. That is, $\mathrm{supp}(X) = \frac{|\{X \subseteq T\}|}{|T|}$.

The confidence factor indicates how often the rule $X \Rightarrow Y$ is true. The confidence value of a rule, $X \Rightarrow Y$, in a set of transactions T, is the proportion of the transactions that contain X which also contain Y. $\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$. Thus confidence can be interpreted as an estimate of the conditional probability.

Given the definitions of support and confidence factors, the method for finding an association rule [3] can be decomposed into two subproblems.

1. Find the frequent itemset: The frequent itemset is defined as the itemset that appears in the transaction set T at least τ times, where τ is predefined minimum support (also called a threshold).
2. Use the frequent itemsets to generate the association rules: For every large itemset X, find all non-empty subsets A of X. For every such subset A, output a rule of the form $A \Rightarrow (X \setminus A)$ if the ratio of $\mathrm{supp}(X)$ to $\mathrm{supp}(A)$ is at least τ.

C EXAMPLE OF THE ARL ALGORITHM

For simplicity, we consider two parties P_1 and P_2, each holding a vertical partitioned database T_1 and T_2, respectively. Assume that T_1 has 3 attributes/columns $\{a_1, a_2, a_3\}$, and T_2 has 2 attributes/columns $\{b_1, b_2\}$.

One important step of the ARL algorithm is to find all "global" frequent itemsets. For example, we want to compute how many transactions that contain 2 attributes (a_1, b_1). If the number of these transactions is greater than a threshold t, we say that (a_1, b_1) is a frequent itemset.

For a better protocol explanation. We define "global" vs "local" frequent itemset. A frequent itemset is global if each party has at least one item in the frequent itemset (this aligns with the global rule mentioned in Definition 1). A frequent itemset is local if the frequent itemset contains only items belonging to one party.

If (a_1, b_1) is a "global" frequent itemset, the attribute a_1 itself should be a "local" frequent itemset. Thus, before any interaction between parties, each party P_i needs to locally compute a list L_1^i that has only 1 attribute. For example, if the attribute a_1 appears more than or equal t times in T_1, then a_1 is a local frequent itemset, and thus a_1 is added to L_1^1. In contrast, if the attribute a_2 appears less than t times in the T_1, then a_2 is not a local frequent itemset, and thus $a_2 \notin L_1^1$. Assume that from Step 1, we have $L_1^1 = \{a_1, a_3\}$, and $L_1^2 = \{b_1, b_2\}$.

Step 2 of Protocol 14 aims to find a list L_n of "global" frequent itemsets, where each itemset has n items ($n = 2$ in the two-party setting). To do so, the parties run DotProd where the party's input is each itemset in L_1^1 and L_1^2. For example, the parties check whether

Algorithm 1 `apriori-gen`(L_t)

```
1: Find all pairs of itemsets in $L_t$ where the first $t-1$ items are identical.
       e.g., $t = 5$ and two pairs $\{a,b,c\}$, $\{a,b,d\}$
2: Union them (lexicographically) to get a list of candidates $C'_{t+1}$
       e.g., $\{a,b,c\}, \{a,b,d\} \to \{a,b,c,d\}$
3: Prune $C_{t+1} = \{c \in C'_{t+1} \mid \forall s_c \notin L_t\}$, where $s_c$ is a $t$-subsets of $c$.
4: Return $C_{t+1}$
```

Figure 1: A Simplest apriori-gen Algorithm [3, 40]

each of pairs $(a_1,b_1), (a_1,b_2), (a_3,b_1), (a_3,b_2)$ are "global" frequent items. Assume that the column a_1 is $(1,1,1,0,0)$ and the column b1 is $(1,1,1,1,0)$. The dot product $a_1 \cdot b_1$ is 3. E.g. a pair (a_1,b_1) appears 3 times in the database. If the threshold $t = 2$, the (a_1,b_1) is a "global" frequent itemset.

Step 3 of the protocol aims to find a list L_k of "global" frequent itemsets, where each itemset has k items (here, $k > 2$). For example, the parties want to check whether (a_1,a_3,b_1) is a "global" frequent itemset (in this case, $k = 3$). They first need to compute the dot product $a_1 \cdot a_3 \cdot b_1$. To do so, P_1 locally computes a dot product of a_1 and a_3 before running a secure DotProd with P_2 (see Step 3b). The function apriori-gen is for improving the computation – it helps to generate the set of *candidate* itemsets for L_k.

D OUR SECURE DOT PRODUCT PROTOCOL

See Protocol 15.

PROTOCOL 15. (*Secure Dot Product* - $\Pi^{(t,n)}_{\text{DotProduct}}$)

Parameters:
- An upper-bound t.
- n parties: $P_1, \dots, P_n$; an untrusted server C;
- A PSI-CA functionality $\mathcal{F}_{\text{PSI-CA}}$ in Functionality 4.
- A function $\mathbf{idx}' : \mathbb{Z}_2^\star \times \{0,1\}^\star \to (\{0,1\}^\star)^\star$ in Section 5.1

Inputs:
- $P_{i\in[n]}$ has $X_i = \{x_{i,1}, \dots, x_{i,m}\}$.
- Cloud server C has no input.

Protocol:
(1) Each party $P_{i\in[n]}$ computes $A_i \leftarrow \mathbf{idx}'(X_i, t)$.
(2) All parties invoke $\mathcal{F}_{\text{PSI-CA}}$ where P_i inputs A_i, C inputs nothing, and P_1 obtains the output $|\bigcap_{i=1}^n A_i|$.

E SERVER-AIDED 2-PARTY PSI PROTOCOL[27]

See Protocol 16.

PROTOCOL 16. (*Server-Aided 2-Party PSI [27]*)
Parameters: There are 2 parties P_1, P_2 and a third-party server S. P_1 and P_2 have sets X_1 and X_2 as input, respectively. The server S does not have input. Let F be a PRF, and parameter $d > 0$.

Protocol:
(1) P_1 chooses sets D_0, D_1, D_2 and a key k_1 such that $|D_0| = |D_1| = |D_2| = d$, sends them to P_2 and set $Y_1 \leftarrow X_1 \cup D_0 \cup D_1$.
(2) P_2 sets $Y_2 \leftarrow X_2 \cup D_0 \cup D_2$.
(3) P_2 chooses a random key k_2 and sends it to the server S.
(4) Party P_1 sends a shuffled version of $Y'_1 = \{F(k_1,x)\}_{x\in Y_i}$ to S.
(5) The server returns a shuffled version π of $Y''_1 = \{F(k_2,y)\}_{y\in Y'_1}$ to P_1
(6) Party P_2 sends a shuffled version of $Y''_2 = \{F(k_2, F(k_1,x))\}_{x\in Y_2}$ to P_1.
(7) P_1 computes $I = Y''_1 \cap Y''_2$ and sends the result to P_2
(8) P_2 computes $I^{-1} = \{F^{-1}(k_1, F^{-1}(k_2,x)) | \forall x \in I\}$
(9) P_2 check that I has the right form and aborts if:
  (a) Either $D_0 \not\subset I^{-1}$ or $D_2 \cap I^{-1} \neq \emptyset$
  (b) There exists $x \in X_2$ and $\alpha, \beta \in [\lambda]$ such that $x||\alpha \in I^{-1}$ and $x||\beta \notin I^{-1}$
(10) If P_2 does not abort, it notifies S who sends the shuffled function π to P_1. P_1 uses π learns the values in the set I^{-1}
(11) P_1 checks that I has the right form as in Step (9) and aborts if the check fails.
(12) The parties output distinct items in $I^{-1} \setminus D_0$.

F ZERO SHARING PROTOCOL [30]

See Protocol 17.

PROTOCOL 17. (*Zero-Sharing* - Π_{ZS} *[30]*)
Parameters: There are n parties $P_1, \dots, P_n$. There is a PRF $F : \{0,1\}^\kappa \times \{0,1\}^\ell \to \{0,1\}^\kappa$.

Protocol:
(1) Each party P_i picks a random seed $r_{i,j}$ for $j \in [i+1, n]$ and sends $r_{i,j}$ to P_j. The key K_i of party P_i is $(k_{1,i}, \dots, k_{i-1,i}, k_{i,i+1}, \dots, k_{i,n})$.
(2) To obtain its share for value x, party P_i computes

$$S(K_i, x) = \left(\bigoplus_{j<i} F_{k_{j,i}}(x)\right) \oplus \left(\bigoplus_{j>i} F_{k_{i,j}}(x)\right)$$