Secure Intersection with MapReduce[a]

Radu Ciucanu[1], Matthieu Giraud[2], Pascal Lafourcarde[2] and Lihua Ye[3]
[1]*INSA Centre Val de Loire, Univ. Orléans, LIFO EA 4022, Bourges, France*
[2]*LIMOS, Université Clermont Auvergne, Aubière, France*
[2]*Harbin Institute of Technology, China*
radu.ciucanu@insa-cvl.fr, {matthieu.giraud, pascal.lafourcade}@uca.fr, 16s003041@stu.hit.edu.cn

Keywords: Intersection, Database, Privacy, Security, MapReduce

Abstract: Relation intersection is a fundamental problem, which becomes non-trivial when the relations to be intersected are too large to fit on a single machine. Hence, a natural approach is to design parallel algorithms that are executed on a cluster of machines rented from a public cloud provider. Intersection of relations becomes even more difficult when each relation belongs to a different data owner that wants to protect her data privacy. We consider the popular MapReduce paradigm for outsourcing data and computations to a *semi-honest* public cloud. Our main contribution is the SI protocol that allows to securely compute the intersection of an arbitrary number of relations, each of them being encrypted by its owner. The user allowed to query the intersection result has only to decrypt the result sent by the public cloud. SI does not leak (to the public cloud or to the user) any information on tuples that are not in the final relation intersection result, even if the public cloud and the user collude i.e., they share all their private information. We prove the security of SI and provide an empirical evaluation showing its efficiency.

1 Introduction

The outsourcing of data and computation to the cloud is a frequent scenario in modern applications. While many cloud service providers with an important amount of data storage and of power computation (e.g., Google Cloud Platform, Amazon Web Services, Microsoft Azure) are available for a reasonable price, they do not usually address the fundamental problem of protecting the privacy of users' data.

We consider the problem of intersection of an arbitrary number of relations, each of them belonging to a different data owner. We rely on the popular MapReduce (Dean and Ghemawat, 2004) paradigm for outsourcing data and computations to a semi-honest public cloud. Our goal is to compute the relations' intersection while preserving the data privacy of each of the data owners. We develop a protocol based on public-key cryptography where each participant encrypts and sends their respective relation on the public cloud. The public cloud cannot learn neither the input nor the output data, but may learn only the number of tuples in the intersection. At the end of the computation, the public cloud sends the result to the final user, who only has to decrypt the received data. Moreover, if the public cloud and the user collude, i.e., the public cloud knows the user's private key, then they cannot learn other information than the intersection result. Secure intersection is a foundational primitive and have a lot of applications such that privacy-preserving data mining (Aggarwal and Yu, 2008), homeland security (Cristofaro and Tsudik, 2010), human genome research (Baldi et al., 2011), Botnet detection (Nagaraja et al., 2010), social networks (Mezzour et al., 2009), and location sharing (Narayanan et al., 2011).

1.1 Intersection with MapReduce

A protocol to compute the intersection between two relations with MapReduce is presented in Chapter 2 of (Leskovec et al., 2014). We stress that intersection between relations can be viewed as intersection between sets where elements of these sets correspond to the tuples of relations having the same schema. In Chapter 2 of (Leskovec et al., 2014), the public cloud receives two relations from their respective owner. A collection of cloud nodes has chunks of these two relations. The Map function creates for each tuple t a

[a]This research was conducted with the support of the FEDER program of 2014-2020, the region council of Auvergne-Rhône-Alpes.

key-value pair (t,t) where *key* and *value* are equal to the tuple. Then, the key-value pairs are grouped by key, i.e., key-value pairs output by the map phase that have the same key are sent to the same reducer (i.e., the application of the Reduce function to a single key and its associated values). For each key, the Reduce function checks if the considered key is associated to two values. If it is the case, i.e., tuple t is present in both relations, then the public cloud produces and sends the pair $(-,t)$ to the user. The dash value "$-$" corresponds to the empty value, we use it to be consistent with the key-value result form required by the MapReduce paradigm. Hence, all tuples received by the user correspond to the tuples that are in both relations. However, the key is irrelevant at the end of the protocol, hence we often omit to write it. We illustrate this approach with the following example considering three relations.

Example. We consider three relations: NSA, GCHQ, and Mossad. Each relation is owned by their respective data owner. These three relations have the same schema composed of only one attribute, namely "Suspect's ID". They are defined as follows: $\mathsf{NSA} = \{\texttt{F654},\texttt{U840},\texttt{X098}\}$, $\mathsf{GCHQ} = \{\texttt{F654},\texttt{M349},\texttt{P027}\}$, and $\mathsf{Mossad} = \{\texttt{F654},\texttt{M349},\texttt{U840}\}$. An external user, called *Interpol*, wants to receive the intersection of these three relations denoted Interpol. We illustrate the execution of intersection computation with MapReduce for this setting in Figure 1. First, each data owner outsources their respective relation into the public cloud. Then, the public cloud runs the map function on each relation and sends the output to the master controller in order to sort key-value pairs by key. Then, the master controller sends key-value pairs sharing the same key to the same reducer. In our example, we obtain 5 reducers since there are 5 different suspect's identities. The reducer associated to the key F654 has three values since the identity F654 is present in the three relations NSA, GCHQ, and Mossad. The reducer associated to the key M349 has two values since the identity M349 is only present in relations GCHQ and Mossad. Other reducers are associated to only one value since the corresponding suspect's identity is present in only one relation. For each reducer, the public cloud runs the reduce function and sends the tuple $(-,\texttt{ID})$ to the user if the suspect's identity ID is present is the three relations, else the public cloud sends nothing. In our example, we observe that the user *Interpol* only receives the pair $(-,\texttt{F654})$ since the suspect's identity F654 is present in the three relations NSA, GCHQ, and Mossad.

1.2 Problem statement

We assume $n+2$ parties: n data owners, the public cloud, and the external user (simply referred as user in the following). Each data owner is trusted (i.e., they dutifully follow the protocol and do not collude with other party) and outsources a relation R_i, with $i \in [\![1,n]\!]$, to the public cloud, denoted $\mathcal{C}$. We denote by $\mathcal{R}_i$ the owner of the relation R_i for $i \in [\![1,n]\!]$. A user, denoted $\mathcal{U}$, and who does not know the individual relations R_i is authorized to query the intersection of these n relations.

We assume that the public cloud is *semi-honest* (Lindell, 2017), i.e., it executes dutifully the computation task but tries to learn the maximum of information on relations R_i and on their intersection. In the original protocol (Leskovec et al., 2014), tuples of each relation are not encrypted, hence the public cloud learns all the content of each relation and the result of the intersection that it sends to the user as illustrated in Figure 1. In order to preserve the privacy of data owners, the cloud should not learn any plain input data, contrary to what happens for the original protocol.

Moreover, we assume that the public cloud can collude with the user, i.e., they share all their respective private information. We want that the user that queried the intersection of these n relations may learn nothing else than the intersection of the n relations, even in case of collusion with the public cloud.

1.3 Contributions

We revisit the standard protocol for the computation of intersection with MapReduce (Leskovec et al., 2014) and propose a new protocol called SI (for Secure Intersection with MapReduce) that satisfies our aforementioned problem statement. More precisely:

- Our protocol SI guarantees that the user who queries the intersection of the n relations learns only the final result, i.e., tuples that are present in the n relations. Moreover, the public cloud does not learn information about the input data that belongs to the data owners, it learns only the cardinal of each relation and the cardinal of the intersection. SI also satisfies the problem setting in the presence of collusion between the user and the public cloud. We prove in the aforementioned properties in the random oracle model.
- To show the practical scalability of SI, we present experimental results using the MapReduce open-source implementation Apache Hadoop 3.2.0.
- Our protocol SI is efficient from both computation and communication points of view. The over-

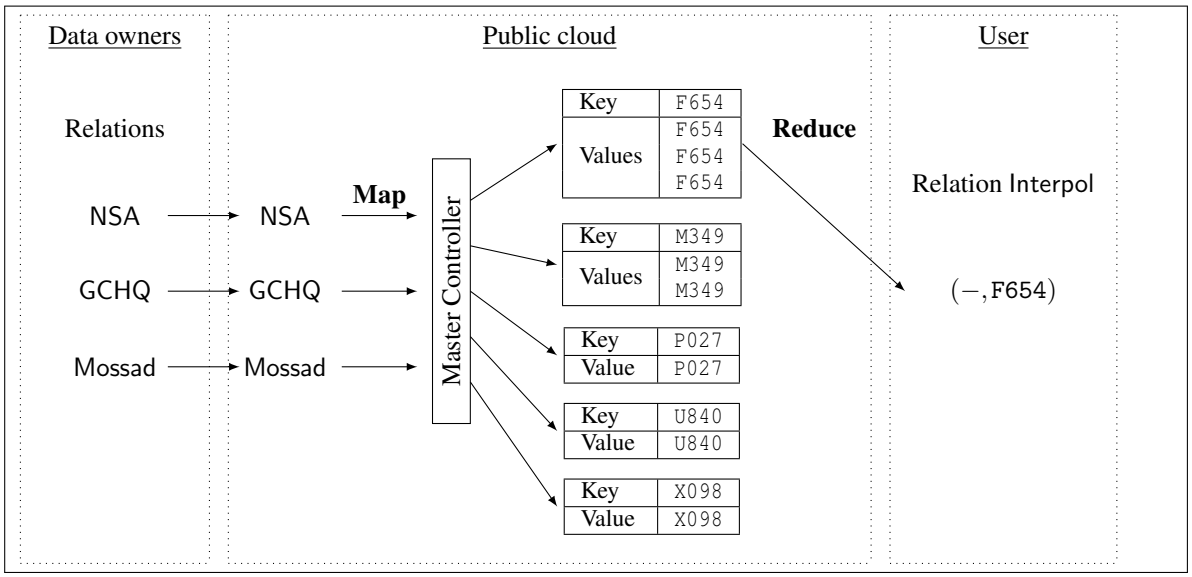

Figure 1: Example of intersection with MapReduce between three relations. First, data owners outsource their respective relation on the public cloud. The public cloud runs the Map function, then the Reduce function verifies if a key is associated to a list of three values. If that case, only the reducer associated to the key `F654` has three values. Hence, the public cloud sends the tuple $(-,\texttt{F654})$ to the user.

Protocol	*Comp. cost*	*Com. cost*
Standard	$2nN$	$(n+1)N$
SI	$C_{\mathcal{E}} \cdot N$ $+C_f \cdot (3n-2)$ $+C_{\oplus} \cdot (2N \cdot (n-2))$	$(n+1)N$

Figure 2: Summary of results in big-O. Let $N = \max(|R_1|,\ldots,|R_n|)$. Let C_f (resp. $C_{\mathcal{E}}$, $C_{\oplus}$) be the computation cost of a pseudo-random function evaluation (resp. asymmetric encryption, xor operation).

head for the computation complexity is linear in the number of tuples by relation while the communication complexity is the same as in the standard protocol (Leskovec et al., 2014). Our technique is based on classical cryptographic tools such that pseudo-random function, asymmetric and one-time-pad encryptions. We summarize in Figure 2 the trade-offs between computation and communication costs for our secure protocol SI vs the standard MapReduce protocol computing the intersection of $n \geq 2$ relations. In our communication cost analysis, we measure the total size of the data that is emitted from a map or reduce node.

1.4 Related Work

As previously mentioned, a relation can be seen as a set where tuples of the relation are the elements of the set. *Private Set Intersection* (PSI) refers to the cryptographic primitive where two parties compute the intersection of their respective sets, such that minimal information is revealed in the process. It was introduced by Freedman et al. (Freedman et al., 2004). The aim of such a primitive is to allow the two parties to learn the elements common to both sets and nothing else. Such primitives where neither party has any advantage over the other and where all parties know the intersection are called *mutual PSI* (Cristofaro et al., 2010). On the contrary, primitives where only one party learns the intersection of the two sets while the other learns nothing are called *one-way PSI* (Cristofaro et al., 2010). A natural PSI extension is called *PSI with Data Transfer* (PSI-DT) (Jarecki and Liu, 2010). In PSI-DT, one or both parties have data associated with each element of their respective sets. Thus, the intersection must be transferred with the associated data. Contrary to these approaches, our protocol SI does not reveal any information on the intersection to the data owner. Only the user knows the intersection.

The seminal work (Freedman et al., 2004) uses two-party computation and partial homomorphic encryption allowing two owners to securely compute the intersection of two sets. The proposed protocol is proven against *semi-honest* adversaries in the standard model and also proven for a malicious adversary in random oracle model. Authors consider one client and one server, each of them owning a secret dataset where the client sends polynomial coefficients associated to her dataset in an encrypted way to the server. At the end of the protocol the server knows which ele-

ments are shared with the server while the later learns nothing. In our protocol SI, we consider an arbitrary number of clients owning different relations and using a *semi-honest* public cloud to send the intersection of the relations to the user.

Following this work, (Hazay and Nissim, 2010) proposed an improved construction considering the presence of a malicious adversaries in the standard model. Contrary to us, the complexity of this construction still remains not linear in the number of elements in sets. (Cristofaro et al., 2010), and (Kissner and Song, 2005) proposed protocols for mutual PSI with linear complexity while in our protocol the user does not have any set to intersect. The scheme proposed by (Cristofaro et al., 2010) considers a malicious adversary using zero-knowledge proofs. Their scheme requires that the user performs computations at the beginning and at the end of the protocol while she has only to decrypt the final result in our protocol. In the scheme of (Kissner and Song, 2005), each data owner learns the result of the intersection. The final result is represented by a polynomial to decrypt in group. In order to obtain the elements of the intersection represented by the polynomial, each data owner has to perform additional computation for each element of her inputs while in our protocol, the user has only to decrypt the result sent by the public cloud. In this paper, we consider semi-honest adversary and prove the security of our protocol in the random oracle model. As remarked above, the intersection computed by our protocol is only known by the user and not by the data owners and the public cloud.

1.5 Outline

We introduce the needed cryptographic tools in Section 2. We recall the standard MapReduce set intersection protocol (Leskovec et al., 2014) and present our secure protocol SI in Section 3. In Section 4, we show experimental evaluations of SI considering intersection between two relations on different number of tuples, and considering intersection between different number of relations. We prove in Section 5 the security properties of SI in the random oracle model.

2 Cryptographic Tools

We start by recalling the definitions of negligible function, then definitions of public-key encryption scheme and of pseudo-random function used in SI.

Definition 1 (Negligible function (Boneh and Shoup, 2017))**.** *A function* $\varepsilon : \mathbb{N} \to \mathbb{R}$ *is called* negligible *if for all* $c > 0$ *there exists* $t_0 \in \mathbb{N}^*$ *such that for all integers* $t \geq t_0$*, we have* $|\varepsilon(t)| < 1/t^c$.

Definition 2 (Public-key encryption)**.** *Let* η *be a security parameter. A* public-key encryption *(PKE) scheme consists of three algorithms.*

- *The randomized* key generation algorithm $\mathcal{G}$ *takes the security parameter to return a public/secret key pair* (pk, sk).
- *The* encryption algorithm $\mathcal{E}$ *takes a public key pk and a plaintext m to return a ciphertext c. We denote by* $\mathcal{E}(pk, m)$ *the encryption of m.*
- *The deterministic* decryption algorithm $\mathcal{D}$ *takes a secret key sk and a ciphertext c to return a corresponding plaintext m or a special symbol* $\perp$ *indicating that the ciphertext was invalid. We denote by* $\mathcal{D}(sk, c)$ *the decryption of c.*

Let $\Pi = (\mathcal{G}, \mathcal{E}, \mathcal{D})$ be a PKE scheme, and $\mathcal{A}$ be a probabilistic polynomial time adversary. $\mathcal{A}$ has access to the oracle $\mathcal{E}(pk, LR_b(\cdot,\cdot))$ taking (m_0, m_1) as input and returns $\mathcal{E}(pk, m_0)$ if $b = 0$, $\mathcal{E}(pk, m_1)$ otherwise. Π is *indistinguishable under chosen-plaintext attack* (IND-CPA) if the advantage of the adversary $\mathcal{A}$ against the IND-CPA experiment defined by:

$$\mathsf{Adv}^{\text{ind-cpa}}_{\Pi,\mathcal{A}}(\eta) = \big|\Pr[1 \leftarrow \mathsf{Exp}^{\text{ind-cpa-1}}_{\Pi,\mathcal{A}}(\eta)] - \Pr[1 \leftarrow \mathsf{Exp}^{\text{ind-cpa-0}}_{\Pi,\mathcal{A}}(\eta)]\big|,$$

is negligible for any polynomial-size $\mathcal{A}$.

Definition 3 (Pseudo-random function)**.** *Let* η *be a security parameter. A* pseudo-random function *(PRF) F is a deterministic algorithm that has two inputs:* $k \in \{0,1\}^{\ell(\eta)}$ *(where* $\ell(\cdot)$ *is a polynomial function), and* $x \in \mathcal{X}$*. Its output is* $y := F(k, x) \in \mathcal{Y}$*. We said that F is defined over* $(\{0,1\}^{\ell(\eta)}, \mathcal{X}, \mathcal{Y})$.

In the rest of the paper, we assume that data of participants are included in $\mathcal{X}$, and the size of $\mathcal{Y}$ is larger enough to avoid collisions. Moreover, we denote by $f_k(\cdot) = F(k, \cdot)$ an instance of F.

Let F be a pseudo-random function, $\mathcal{A}$ be a probabilistic polynomial time adversary, $\mathsf{Func}[\mathcal{X}, \mathcal{Y}]$ be the space of functions defined over domain $\mathcal{X}$ and codomain $\mathcal{Y}$, and $b \in \{0,1\}$. If $b = 0$, the pseudo-random function F is used. Otherwise a random function from $\mathsf{Func}[\mathcal{X}, \mathcal{Y}]$ is used. We define the advantage of the adversary $\mathcal{A}$ against the PRF experiment by:

$$\mathsf{Adv}^{\text{prf}}_{F,\mathcal{A}}(\eta) = \big|\Pr[1 \leftarrow \mathsf{Exp}^{\text{prf-1}}_{F,\mathcal{A}}(\eta)] - \Pr[1 \leftarrow \mathsf{Exp}^{\text{prf-0}}_{F,\mathcal{A}}(\eta)]\big|.$$

We said that F is a *secure* pseudo-random function if this advantage is negligible.

3 MapReduce Intersection

We consider $n \geq 2$ data owners, each of them owning a relation. These n relations have the same schema and are denoted $R_1, \ldots, R_n$. We first recall in Section 3.1 the standard MapReduce protocol to perform the intersection of n relations, i.e., a simple generalization of the binary protocol presented in Chapter 2 of (Leskovec et al., 2014). This protocol obviously does not verify privacy properties of our problem setting since the public cloud learns all tuples of each relation sent by the respective data owner, and the intersection of these n relations sent to the user. Then we present in Section 3.2 our secure protocol denoted SI that computes the intersection of n relations using MapReduce. We prove that contrary to the standard protocol, SI guarantees that the public cloud learns only cardinals of relations R_i for $i \in [\![1,n]\!]$. Moreover, if the public cloud and the user collude, then they learn the intersection of these n relations that the user still knows, and cardinals of relations R_i for $i \in [\![1,n]\!]$ that the public cloud still knows, and nothing else.

3.1 Standard MapReduce Intersection

In the standard protocol (Leskovec et al., 2014), the Map function creates for each tuple t of each relation R_i, with $i \in [\![1,n]\!]$, a key-value pair where the key and the value are equal to the tuple t. For a key t, the associated reducer receives a list of tuples t. Hence, if a tuple t is only present in one relation, the reducer receives a collection only composed of one tuple t. On the contrary, if a tuple t' is present in all the n relations, the reducer receives a collection of n tuples equal to t'. If a key t is associated to a collection of n tuples t, then the Reduce function produces the key-value pair $(-,t)$ and sends it to the user. Otherwise, it produces nothing. All key-value pairs outputted by the Reduce function constitute the result of the intersection of the n relations. We present the standard protocol (Leskovec et al., 2014) computing the intersection protocol with MapReduce in Figure 3.

We now consider a *semi-honest* public cloud performing the intersection of n relations with MapReduce. In such a scenario, the public cloud learns all the content of each relation along with the intersection of these n relations.

3.2 Secure MapReduce Intersection

In order to compute intersection with MapReduce in a privacy-preserving way between $n \geq 2$ relations, our protocol uses pseudo-random function, asymmetric and one-time encryptions. We denote by F a secure

Map function:
// *key*: id of a chunk of R_i
// *value*: collection of tuples $t \in R_i$
foreach $t \in R_i$ **do**
  `emit` (t,t).

Reduce function:
// *key*: tuple $t \in \cup_{i=1}^{n} R_i$
// *values*: collection of tuples t
$L = [\,]$
foreach $v \in$ `values` **do**
  $L \leftarrow L \cup \{v\}$
if $|L| = n$ **then**
  `emit` $(-,t)$.

Figure 3: MapReduce protocol to compute the intersection of n relations.

pseudo-random function defined over $(\mathcal{K}, \mathcal{X}, \mathcal{Y})$ and by $\Pi = (\mathcal{G}, \mathcal{E}, \mathcal{D})$ an IND-CPA asymmetric encryption scheme. We also assume that the length of values outputted by the pseudo-random function is equal to the length of ciphertext outputted by the asymmetric encryption scheme. In practice, we can use the *Advanced Encryption Standard* (AES) (Daemen and Rijmen, 2002) with the *Cipher Block Chaining* mode on padded message in order to obtain a ciphertext of the same length than the ciphertext obtained with the asymmetric encryption.

3.2.1 Preprocessing of our Secure Protocol SI

Before outsourcing their relation to the public cloud, data owners perform a key setup and a preprocessing on their respective relation R_i to obtain a *protected* relation denoted R_i^*. We present the key setup and the preprocessing phase in Figure 4.

First, we need a secret key $k_1 \in \mathcal{X}$ that is shared between the n data owners. Moreover we need $n-1$ other secret keys $k_i \in \mathcal{K}$ (for $2 \leq i \leq n$) such that $k_i \neq k_j$ for $i \neq j$. Key k_i with $i \geq 2$ is shared between the owner of relation R_1 and the owner of relation R_i. Hence, the owner of relation R_1 has a set of secret keys equals to $\{k_1, k_2, \ldots, k_n\}$ while the owner of relation R_i (for $2 \leq i \leq n$) has a set of secret keys equals to $\{k_1, k_i\}$. We stress that the choice of owner of relation R_1 knowing all the secret keys is arbitrary, and we call the associated relation, i.e., R_1, the *main* relation.

The aim of this preprocessing is to protect owners' data in order to avoid the public cloud to learn tuples of each relation and the result of the intersection sent to the user. Moreover, this preprocessing is in agreement with the MapReduce paradigm. Indeed, each protected relation R_i^* is composed of tuples under the

Preprocessing:
// *input*: relation R_i with $i \in [\![1,n]\!]$
// *outputs*: protected relation R_i^*
for $1 \leq i \leq n$ **do** $k_i \xleftarrow{\$} \mathcal{K}$;
$R_i^* \leftarrow \emptyset$;
if $i = 1$ **then**
  foreach $t \in R_1$ **do**
    $R_1^* \leftarrow R_1^* \cup \{(f_{k_1}(t), (\mathcal{E}(pk,t) \oplus_{j=2}^{n} f_{k_j}(t)))\}$
else
  foreach $t \in R_i$ **do**
    $R_i^* \leftarrow R_i^* \cup \{(f_{k_1}(t), (f_{k_i}(t)))\}$
return R_i^*.

Figure 4: Preprocessing of our secure protocol SI run by each data owner.

key-value pair form.

First of all, each key of pairs of R_i^* is a pseudo-random evaluation of a tuple using the secret key k_1 known by each data owner. Since a pseudo-random function is deterministic, equal tuples share the same value of key. Hence, the map phase sends these key-value pairs to the same reducer as expected.

Moreover, each value of key-value pairs of the protected relation R_1^* is equal to the encryption of the tuple using the asymmetric encryption scheme Π with the user public key pk xored by $n-1$ pseudo-random evaluations of the tuple using secret keys $k_2, \ldots, k_n$. More precisely, for each tuple $t \in R_1$, the preprocessing computes the key-value pair equals to $(f_{k_1}(t), \mathcal{E}(pk,t) \oplus_{j=2}^{n} f_{k_j}(t))$. Hence, when the public cloud receives such key-value pairs and colludes with the user, it cannot learn the value of tuples since the asymmetric encryption is protected by pseudo-random evaluations, and secret keys $k_1, \ldots, k_n$ are not known by the public cloud.

3.2.2 Map and Reduce Phases of SI Protocol

The preprocessing presented in Figure 4 outputs an encrypted relation whose tuples are of the key-value pair form. Hence, once the public cloud receives the n encrypted relations R_i^* (for $i \in [\![1,n]\!]$) from the data owners, it runs the Map function that is simply the identity function.

After the grouping by key, the Reduce function checks if the current key $f_{k_1}(t)$, for $t \in \cup_{i=1}^{N} R_i$, is associated to a list of n values. If that is the case, it means that the n relations contain the tuple associated to the current key. Then the Reduce function uses these n values to perform an exclusive or, and obtains the asymmetric encryption of the tuple $\mathcal{E}(pk,t)$ due the property of the exclusive or.

Finally, the Reduce function produces the key-value pair $(-, \mathcal{E}(pk,t))$ and sends it to the user. The output of the Reduce function is in a key-value form to be consistent with the MapReduce paradigm since at the end of the SI protocol keys are irrelevant. All key-value pairs outputted by the Reduce function constitute the intersection of the n relations. The user has only to decrypt each value of key-value pair using her secret key in order to obtain the intersection in plain form. Our protocol SI is described in Figure 5.

Map function:
// *key*: id of a chunk of R_i^* with $i \in [\![1,n]\!]$
// *values*: collection of $(f_{k_1}(t), \mathcal{E}(pk,t) \oplus_{j=2}^{n} f_{k_j}(t))$
// or $(f_{k_1}(t), f_{k_j}(t))$ with $j \in [\![2,n]\!]$
foreach $(k,v) \in$ *values* **do**
  `emit` (k,v)

Reduce function:
// *key*: $f_{k_1}(t)$ such that $t \in \cup_{i=1}^{n} R_i$
// *values*: collection of $\mathcal{E}(pk,t) \oplus_{j=2}^{n} f_{k_j}(t)$ or $f_{k_j}(t)$
// with $j \in [\![2,n]\!]$
$L \leftarrow [\,]$
foreach $v \in$ *values* **do**
  $L \leftarrow L \cup \{v\}$
if $|L| = n$ **then**
  $\mathcal{E}(pk,t) \leftarrow \mathcal{E}(pk,t) \oplus_{j=2}^{n} f_{k_j}(t) \oplus_{j=2}^{n} f_{k_j}(t)$
  `emit` $(-, \mathcal{E}(pk,t))$

Figure 5: Map and Reduce functions of SI.

Example. We illustrate our SI protocol following the example presented in Section 1. First, we perform the preprocessing on relations: NSA, GCHQ, and Mossad. We consider relation NSA as the main relation. The three data owners share the secret key k_1, data owners of relations NSA and GCHQ share a secret key k_2, and data owners of relations NSA and Mossad share a secret key k_3. Hence, after the preprocessing phase, we obtain three protected relations denoted NSA*, GCHQ*, and Mossad* as illustrated in Figure 7.

We now illustrate the execution of intersection computation with MapReduce using our secure protocol SI in Figure 7.

3.2.3 Proof of Correctness

We say that the protocol SI is *correct* if for $n \geq 2$ relations $R_1, R_2, \ldots R_n$, SI returns the correct intersection of the $n \geq 2$ relations, i.e., the encrypted relation composed of pairs $(-, \mathcal{E}(pk,t))$ such that $t \in R$, where $R := \cap_{i=1}^{n} R_i$.

Relation NSA*
(Owner knows secret keys k_1, k_2, and k_3.)

Suspect's Identity
$(f_{k_1}(\texttt{F654}), \mathcal{E}(pk, \texttt{F654}) \oplus f_{k_2}(\texttt{F654}) \oplus f_{k_3}(\texttt{F654}))$
$(f_{k_1}(\texttt{U840}), \mathcal{E}(pk, \texttt{U840}) \oplus f_{k_2}(\texttt{U840}) \oplus f_{k_3}(\texttt{U840}))$
$(f_{k_1}(\texttt{X098}), \mathcal{E}(pk, \texttt{X098}) \oplus f_{k_2}(\texttt{X098}) \oplus f_{k_3}(\texttt{X098}))$

Relation GCHQ*
(Owner knows secret keys k_1 and k_2.)

Suspect's Identity
$(f_{k_1}(\texttt{F654}), f_{k_2}(\texttt{F654}))$
$(f_{k_1}(\texttt{M349}), f_{k_2}(\texttt{M349}))$
$(f_{k_1}(\texttt{P027}), f_{k_2}(\texttt{P027}))$

Relation Mossad*
(Owner knows secret keys k_1 and k_3.)

Suspect's Identity
$(f_{k_1}(\texttt{F654}), f_{k_3}(\texttt{F654}))$
$(f_{k_1}(\texttt{M349}), f_{k_3}(\texttt{M349}))$
$(f_{k_1}(\texttt{U840}), f_{k_3}(\texttt{U840}))$

Figure 6: Protected relations NSA*, GCHQ*, and Mossad* after the preprocessing phase of our secure protocol SI.

Lemma 1. *Assume that the pseudo-random function family F perfectly emulates a random oracle, then protocol* SI *is correct.*

Proof. Let $R_1, R_2, \ldots R_n$ be n relations. Let $R_1^*, R_2^*, \ldots R_n^*$ be the corresponding encrypted relations computed by the preprocessing phase of SI. We set $R := \cap_{i=1}^n R_i$.

For each $t \in R$, there exists a key-value pair in relation R_1^* of the form $(f_{k_1}(t), \mathcal{E}(pk,t) \oplus_{i=2}^n f_{k_i}(t))$, and a key-value pair in relation R_j^*, with $2 \leq j \leq n$, of the form $(f_{k_1}(t), f_{k_j}(t))$. Following the MapReduce paradigm, the n values are sent to the same reducer that sums the corresponding values. Thus, for each key $f_{k_1}(t)$, with $t \in R$, we obtain:

$$\mathcal{E}(pk,t) \oplus_{i=2}^n f_{k_i}(t) \oplus_{i=2}^n f_{k_i}(t) = \mathcal{E}(pk,t) \; .$$

Hence, for each $t \in R$, reducer associated to the key $f_{k_1}(t)$ emits the pair $(-, \mathcal{E}(pk,t))$ to the user. Moreover, for each $t \in \left(\cup_{i=1}^n R_i\right) \setminus R$, the reducer associated to the key $f_{k_1}(t)$ does not output the pair $(-, \mathcal{E}(pk,t))$ since it is associated to less than n values. Finally, SI produces pairs $(-, \mathcal{E}(pk,t))$ such that $t \in R$ corresponding to the intersection of relations $R_1, R_2, \ldots, R_n$ which concludes the proof. □

4 Experimental Results

We present an experimental comparison between the standard MapReduce set intersection protocol (Leskovec et al., 2014), and our secure protocol SI. We run two types of experiments, where we vary two different parameters: the *number of tuples per relation* for a fixed number of 2 relations (Section 4.2), and the *number of intersected relations* (Section 4.3).

4.1 Settings

We have done all computations on a cluster running Ubuntu Server 16.04 LTS[1] with Hadoop 3.2.0 [2] using Java 1.8.0 [3]. We use the Hadoop streaming utility and implement the Map and Reduce functions in Golang 1.6.2 [4]. The cluster is composed of one master node and of ten slave nodes. Each node has four CPU cadenced to 2.4 GHz, 80 GB of disk, and 8 GB of RAM.

According to our SI protocol, we use as pseudo-random function the *Advanced Encryption Standard* (AES) symmetric encryption scheme in *Cipher Block Chaining* (CBC) encryption mode with a key size of 128 bits. For the purpose of our protocol, the initial vector is fixed and common to all data owners. For the asymmetric encryption scheme, we use the RSA-OAEP scheme with a key size of 2048 bits and SHA-256 as hash function.

For each experiment, we report average CPU times over 8 runs. Since the cluster environment is not isolated from other machines of the netword, we do not give measures for the communication cost.

4.2 Number of Tuples

In the experiment on the number of tuples, we consider two relations of the same schema composed of only one attribute. These two relations have the same cardinal C and share $C/2$ elements. Tuples of the first relation consist in all integers from 1 to C, while tuples of the second relation consist in all integers from $C/2$ to $C + C/2$. We run the original protocol (Leskovec et al., 2014) and our secure protocol SI on couples of relations of cardinal $500,000$ to $3,000,000$, by step of $250,000$.

We remark in Figure 8 that the computation complexity of our secure protocol is linear as determined in the complexity study (cf. Figure 2).

4.3 Number of Intersected Relations

In the experiment on the number of intersected relations, we consider intersection between different number of relations of the same schema composed of

[1] https://www.ubuntu.com/
[2] https://hadoop.apache.org/
[3] https://java.com/en/
[4] https://golang.org/

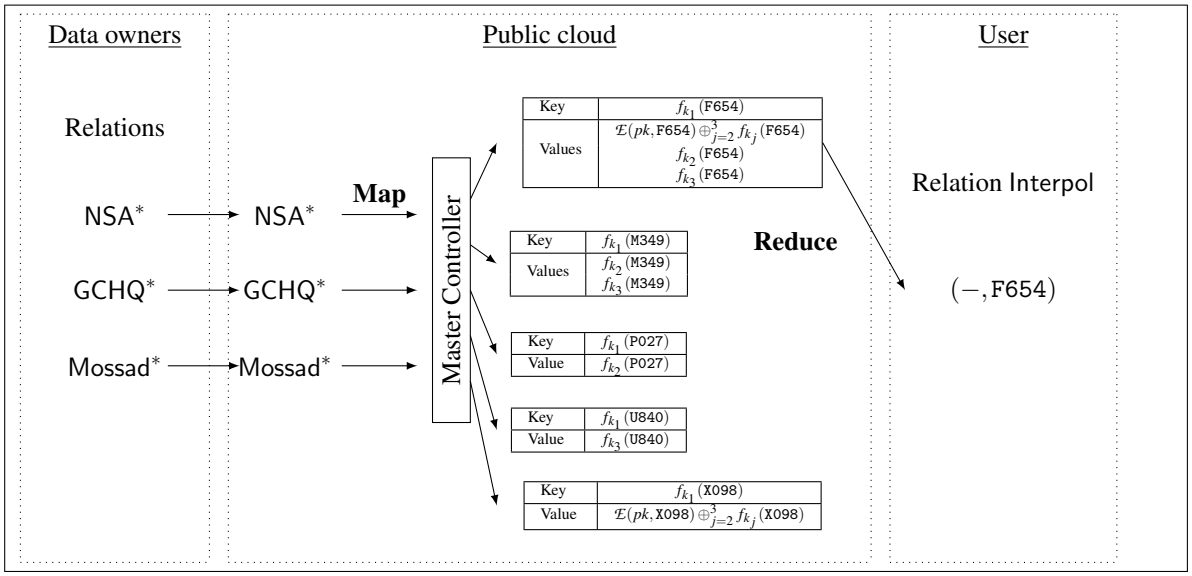

Figure 7: Example of intersection with MapReduce between three relations using our secure protocol SI. First, data owners outsource their respective protected relation on the public cloud. The public cloud runs the Map function, then the Reduce function verifies if keys are associated to a list of three values. In that case, the public cloud only sends one encrypted tuple from the reducer associated to the key $f_{k_1}(\texttt{F654})$ since it is the only one that contains three values. This tuple is $(-, \mathcal{E}(pk, \texttt{F654}))$ equals to $(-, (\mathcal{E}(pk, \texttt{F654}) \oplus_{j=2}^{3} f_{k_j}(\texttt{F654})) \oplus_{j=2}^{3} f_{k_j}(\texttt{F654})$.

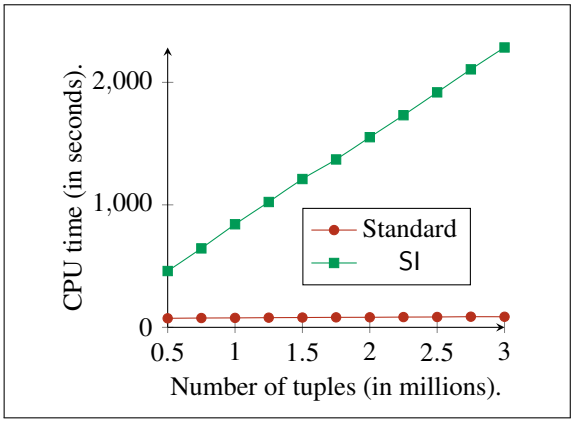

Figure 8: CPU time of the standard MapReduce protocol to compute the intersection between two relations.

only one attribute. We start by computing the intersection of 2 relations to finish with the intersection of 10 relations. In each case, relations have 500,000 tuples and shares 250,000 tuples. In practice, for the intersection of $2 \leq n \leq 10$ relations, tuples of the first relation consist in integers from 1 to 500,000, and tuples of the i-th relation, with $2 \leq i \leq n$, consist in integers from 1 to 250,000 and integers from $i \cdot 250,000 + 1$ to $(i+1) \cdot 250,000$.

We compare the standard protocol (Leskovec et al., 2014) and our secure protocol SI for the experiment on the number of intersected relations. As shown in Figure 9, the computation complexity of our secure protocol is linear as determined in the complexity study (cf. Figure 2). We observe that the computation complexity is less compare to the experiment on the number of tuples. Indeed, when we run the SI protocols with 10 relations of 500,000 tuples (i.e., a total of 5,000,000 tuples), the CPU time is approximately equals to 550 seconds while the CPU time for the intersection of 2 relations of 2 millions (i.e., a total of 4,000,000 tuples) is approximately equals to 1,500 seconds. This is due to number of common elements of each relation. In the case of the intersection of 10 relations (each composed of 500,000 tuples), relations share 250,000 while in the case of the intersection of the 2 relations (each composed of 2,000,000 tuples), relations share 1,000,000 tuples. Hence, the Reduce function has to performs a large number of exclusive or on 2048-bits strings.

5 Security Proof

In this section, we provide a formal security proof of our SI protocol with $n \geq 2$ data owners that computes the intersection of n relations and considering *semi-honest* adversaries. The n data owners respectively own a relation denoted R_i for $1 \leq i \leq n$ such that each relation R_i is constituted of N_i unique tuples, and all relations have the same schema. Moreover, the n

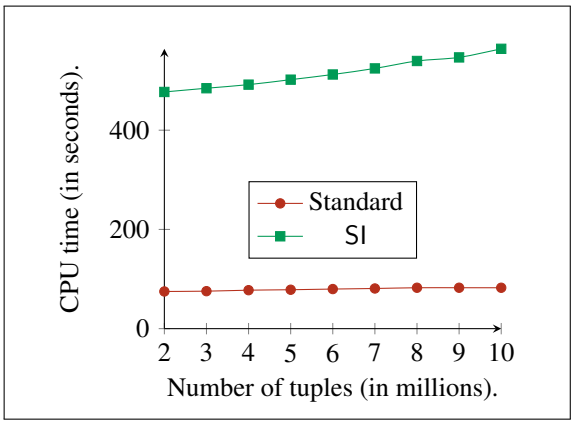

Figure 9: CPU time of the experiment on the number of intersected relation on the original protocol(Leskovec et al., 2014) and of our secure protocol SI.

relations share N tuples in common, in other terms, we have $R := \cap_{i=1}^{n} R_i$ with $\#R = N$.

We start by modeling our SI protocol that computes the intersection of n relations by $n+2$ parties. The owner of the *main* relation R_1 is denoted $\mathcal{R}_1$, while others owners are denoted $\mathcal{R}_i$ for $2 \leq i \leq n$ respectively. The public cloud is denoted $\mathcal{C}$, and the user is denoted $\mathcal{U}$. Parties use respective inputs $I = (I_{\mathcal{R}_1}, \ldots, I_{\mathcal{R}_n}, I_{\mathcal{C}}, I_{\mathcal{U}})$ and a function $g = (g_{\mathcal{R}_1}, \ldots, g_{\mathcal{R}_n}, g_{\mathcal{C}}, g_{\mathcal{U}})$ such that:

- $\mathcal{R}_1$ has the input $I_{\mathcal{R}_1} := (pk, k_1, \ldots, k_n, R_1)$ where pk is a public key of the cryptosystem Π , $k_i \in \mathcal{K}$ for $1 \leq i \leq n$ are secret keys for the pseudo-random function F, and R_1 is the relation owned by $\mathcal{R}_1$. The party $\mathcal{R}_1$ outputs $g_{\mathcal{R}_1}(I) = \perp$ (where $\perp$ denotes that the function returns nothing) since $\mathcal{R}_1$ does not learn any information.
- $\mathcal{R}_i$ for $2 \leq i \leq n$ has the input $I_{\mathcal{R}_1} := (pk, k_1, k_i, R_i)$ where pk is a public key of the cryptosystem Π, $k_1 \in \mathcal{K}$ and $k_i \in \mathcal{K}$ are secret keys for the pseudo-random function F, and R_i is the relation owned by $\mathcal{R}_i$. The party $\mathcal{R}_i$ outputs $g_{\mathcal{R}_i}(I) = \perp$ since it does not learn any information.
- $\mathcal{C}$ has the input $I_{\mathcal{C}} := pk$ where pk is a public key the cryptosystem Π. It returns $g_{\mathcal{C}}(I) = (N_1, \ldots, N_n, N)$ because the party $\mathcal{C}$ learns the number of tuples in each relation R_i for $1 \leq i \leq n$ and the number of tuples common to these n relations.
- $\mathcal{U}$ has the input $I_{\mathcal{U}} := (pk, sk)$ where (pk, sk) is a key pair of the cryptosystem Π. It returns $g_{\mathcal{U}}(I) = R$ where $R = \cap_{i=1}^{n} R_i$, i.e., the result of intersection between the n relations computed by the public cloud.

We prove that our protocol SI securely computes the intersection of n relations in the presence of static semi-honest adversaries even if the two parties $\mathcal{C}$ and $\mathcal{U}$ collude, i.e., if the public cloud and the user share all their respective public and private information. In other terms, we prove that the public cloud and the user do not learn extra information than they have even if the user share her secret key sk with the public cloud. The security proof is given in Theorem 1.

Theorem 1. *Assume F is a secure pseudo-random function and that Π is an IND-CPA asymmetric encryption scheme, then the SI protocol securely computes the intersection of n relations in the presence of static semi-honest adversaries even if parties $\mathcal{C}$ and $\mathcal{U}$ collude.*

The security proof for Theorem 1 is decomposed in Lemma 2 for parties $\mathcal{R}_i$ (with $1 \leq i \leq n$), and in Lemma 3 for the collusion between parties $\mathcal{C}$ and $\mathcal{U}$.

Lemma 2. *Let η be a security parameter. There exists probabilistic polynomial-time simulators and $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}$ for $1 \leq i \leq n$ such for all inputs $I = (I_{\mathcal{R}_1}, \ldots, I_{\mathcal{R}_n}, I_{\mathcal{C}}, I_{\mathcal{U}})$, we have:*

$$\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}(1^{\eta}, I_{\mathcal{R}_i}, g_{\mathcal{R}_i}(I)) \overset{\mathsf{c}}{\equiv} \mathsf{view}^{\mathsf{SI}}_{\mathcal{R}_i}(I, \eta) \ .$$

Proof. Consider that $\mathcal{R}_i$ (with $i \in [\![1, n]\!]$) is corrupted, we observe that $\mathcal{R}_i$ receives no output and no incoming message from other parties. Thus, we merely need to show that a simulator can generate the view of party $\mathcal{R}_i$ from its inputs.

We start, with the data owner of the main relation, i.e., party $\mathcal{R}_1$. In the protocol, $\mathcal{R}_1$ receives the public key pk of the user, the n secret keys $k_1, \ldots, k_n$ used with the pseudo-random function F, and the relation R_1. Formally, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_1}$ is given $(pk, k_1, \ldots, k_n, R_1)$, and works as follows:

1. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_1}$ runs the preprocessing phase (cf. Figure 4) on input $(pk, k_1, \ldots, k_n, R_1)$. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_1}$ obtains the protected relation R_1^* associated to the relation R_1.
2. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_1}$ sends R_1^* to the party $\mathcal{C}$.

In the same way, parties $\mathcal{R}_i$ (with $i \in [\![2, n]\!]$) receives the public key pk of the user, the secret keys k_1 and k_i for the pseudo-random function F, and the corresponding relation R_i. Formally, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}$ (for $2 \leq i \leq n$) is given (pk, k_1, k_i, R_i), and works as follows:

1. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}$ runs the preprocessing phase (cf. Figure 4) with input $I_{\mathcal{R}_i} = (pk, k_1, k_i, R_i)$ and obtains the protected relation R_i^* associated to the relation R_i.
2. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}$ sends R_i^* to the party $\mathcal{C}$.

We remark that $\mathcal{S}_{\mathcal{R}_i}$, for all $i \in [\![1, n]\!]$, uses exactly the same algorithm as the real protocol SI, then it

describes the same distribution as $\mathsf{view}^{\mathsf{SI}}_{\mathcal{R}_i}(I)$, i.e., we have for $1 \leq i \leq n$:

$$\mathcal{S}^{\mathsf{SI}}_{\mathcal{R}_i}(1^\eta, I_{\mathcal{R}_i}, g_{\mathcal{R}_i}(I)) \stackrel{\mathsf{c}}{\equiv} \mathsf{view}^{\mathsf{SI}}_{\mathcal{R}_i}(I,\eta)\ ,$$

which concludes the proof. □

Lemma 3. *If F is a secure pseudo-random function and Π is an IND-CPA asymmetric encryption scheme, then there exists a probabilistic polynomial-time simulator $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ such for all inputs $I = (I_{\mathcal{R}_1}, \ldots, I_{\mathcal{R}_n}, I_{\mathcal{C}}, I_{\mathcal{U}})$, we have:*

$$\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta, I_{\mathcal{C}}, g_{\mathcal{C}}(I)), (1^\eta, I_{\mathcal{U}}, g_{\mathcal{U}}(I))) \stackrel{\mathsf{c}}{\equiv} \mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta)\ .$$

$\mathsf{OPRF}(j,x)$:
if $T[j,x] = \emptyset$ **then** $T[j,x] \stackrel{\$}{\leftarrow} \{0,1\}^{|\mathcal{Y}|}$;
return $T[j,x]$.

Figure 10: Random oracle OPRF.

Proof. Let $\eta \in \mathbb{N}$ be a security parameter. Before to build $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ that computes a distribution that can be simulated perfectly, we use the hybrid argument to build hybrid simulators denoted $\mathcal{S}^{H_i-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ for $1 \leq i \leq n$. The simulator, $\mathcal{S}^{H_i-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ works as SI but each evaluation of the pseudo-random function performed by parties $\mathcal{R}_j$ for $1 \leq j \leq i$ are substituted using the random oracle OPRF presented in Figure 10. We stress that when an entry is being accessed for the first time, it is chosen at random and then used thereafter.

We start by showing how to build the simulator $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$:

1. $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ generates n relations $R_1, \ldots, R_n$ of same schema such that $\#R_i = N_i$ for $1 \leq i \leq n$, and $\#R = N$.
2. $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ generates n secret keys $k_1, \ldots, k_n$ for the pseudo-random function, and a key pair (sk, pk) for the asymmetric encryption scheme Π according to the security parameter η.
3. For each tuple $t \in R_1$, $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathsf{OPRF}(k_1,t), \mathcal{E}(pk,t) \oplus F(k_2,t) \oplus \cdots \oplus F(k_n,t))$ to construct R^*_1, the protected relation associated to R_1.
4. For each tuple $t \in R_i$ (with $2 \leq i \leq n$), $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathsf{OPRF}(k_1,t), F(k_i,t))$ to construct R^*_i, the protected relation associated to R_i.
5. For each tuple $t \in R$, $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathcal{E}(pk,t), \mathcal{E}(pk,t))$ and stores it into a list S.
6. Finally, simulator $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ outputs: $(R_1, \ldots, R_n, R^*_1, \ldots, R^*_n, R, S)$.

Assume by contradiction that there exists a non-uniform probabilistic-polynomial time distinguisher D such that for all inputs I, we have:

$$\Big|\Pr[D(\mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta)) = 1] - \Pr[D(\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta, I_{\mathcal{C}}, g_{I_{\mathcal{C}}}(I)), (1^\eta, I_{\mathcal{U}}, g_{I_{\mathcal{U}}}(I)))) = 1]\Big| = \mu(\eta)\ ,$$

where μ is a non-negligible function in η.

We construct a non-uniform probabilistic-polynomial time guessing algorithm $\mathcal{A}$ that uses D to win the PRF experiment against the pseudo-random function F. Algorithm $\mathcal{A}$ works as follows:

1. $\mathcal{A}$ generates n relations $R_1, \ldots, R_n$ of same schema such that $\#R_i = N_i$ for $1 \leq i \leq n$, and $\#R = N$.
2. $\mathcal{A}$ generates n secret keys $k_1, \ldots, k_n$ for the pseudo-random function F, and a key pair (sk, pk) for the asymmetric encryption scheme Π according to the security parameter η.
3. For each tuple $t \in R$, $\mathcal{A}$ computes $\mathcal{E}(pk,t)$ and stores the value into a list S such that $S[t] = \mathcal{E}(pk,t)$.
4. For each tuple $t \in R_1 \setminus R$, $\mathcal{A}$ computes: $(f_b(k_1,t), \mathcal{E}(pk,t) \oplus F(k_2,t) \oplus \cdots \oplus F(k_n,t))$ to construct R^*_1, the protected relation associated to R_1. If $t \in R$, then $\mathcal{A}$ computes $(f_b(k_1,t), S[t] \oplus F(k_2,t) \oplus \cdots \oplus F(k_n,t))$ and stores the tuples into R^*_1.
5. For each tuple $t \in R_i$ (with $2 \leq i \leq n$), $\mathcal{A}$ computes $(f_b(k_1,t), F(k_i,t))$ to construct R^*_i, the protected relation associated to R_i.
6. $\mathcal{A}$ invokes D on input: $(R_1, \ldots, R_n, R^*_1, \ldots, R^*_n, R, \{(S[t], S[t]) : t \in R\})$.

First, we remark that:

$$\Pr[\mathsf{Exp}^{\text{prf-0}}_{F,\mathcal{A}}(\eta) = 1] = \Pr[D(\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta, I_{\mathcal{C}}, g_{\mathcal{C}}(I)), (1^\eta, I_{\mathcal{U}}, g_{\mathcal{U}}(I)))) = 1]\ .$$

Indeed, when $b = 0$ the view that $\mathcal{A}$ uses as input for D is computed as in the simulator $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. Then the probability that the experiment $\mathsf{Exp}^{\text{prf-0}}_{F,\mathcal{A}}$ returns 1 is equal to the probability that the distinguisher D returns 1 on input computed by the simulator $\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. On the other hand, we have:

$$\Pr[\mathsf{Exp}^{\text{prf-1}}_{F,\mathcal{A}}(\eta) = 1] = \Pr[D(\mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta)) = 1]\ .$$

When $b = 1$, the view that $\mathcal{A}$ uses as input for D is computed as in the real protocol SI. Then the probability that the experiment $\mathsf{Exp}^{\text{prf-1}}_{F,\mathcal{A}}$ returns 1 is equal to the probability that the distinguisher D returns 1 on input computed as in the real protocol.

It therefore follows that:

$$\begin{aligned}\mathsf{Adv}^{\text{prf}}_{F,\mathcal{A}}(\eta) &= \big|\Pr[\mathsf{Exp}^{\text{prf-1}}_{F,\mathcal{A}}(\eta)=1]-\Pr[\mathsf{Exp}^{\text{prf-0}}_{F,\mathcal{A}}(\eta)=1]\big| \\ &= \big|\Pr[D(\mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta))=1] \\ &\quad -\Pr[D(\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1]\big| \\ &= \mu(\eta)\,,\end{aligned}$$

which is non-negligible. However, we assume that F is a secure pseudo-random function, hence, it does not exist D such that:

$$\big|\Pr[D(\mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta))=1]- \Pr[D(\mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1]\big|$$

is non-negligible. Hence, we have:

$$\mathsf{view}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}(I,\eta) \stackrel{\mathrm{c}}{\equiv} \mathcal{S}^{H_1-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I,\eta)),(I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I,\eta)))\,.$$

The construction of simulator $\mathcal{S}^{H_i-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ for $2\leq i\leq n$ is very similar. The only difference is that evaluations of the pseudo-random function F are replaced using the oracle OPRF. We prove as previous that we have for $1\leq i\leq n-1$:

$$\begin{aligned}&\mathcal{S}^{H_i-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{\mathcal{C}}(I)),(1^\eta,I_{\mathcal{U}},g_{\mathcal{U}}(I))) \stackrel{\mathrm{c}}{\equiv} \\ &\mathcal{S}^{H_{i+1}-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{\mathcal{C}}(I)),(1^\eta,I_{\mathcal{U}},g_{\mathcal{U}}(I)))\,.\end{aligned}$$

Finally, we show of to build the simulator $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. The difference between $\mathcal{S}^{H_n-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ and $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ is that $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ substitutes Π encryption of real values by Π encryption of random values of the same size. More formally, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ works as follows:

1. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ generates n relations $R_1,\ldots,R_n$ of same schema such that $\#R_i=N_i$ for $1\leq i\leq n$, and $\#R=N$.
2. $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ generates n secret keys $k_1,\ldots,k_n$ according to the security parameter η.
3. For each tuple $t\in R$, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ picks a random string bits r of the same length than t, computes $\mathcal{E}(pk,r)$ and stores it into $S[t]$.
4. For each tuple $t\in R_1\setminus R$, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathsf{OPRF}(k_1,t),\mathcal{E}(pk,t)\oplus\mathsf{OPRF}(k_2,t)\oplus\cdots\oplus\mathsf{OPRF}(k_n,t))$ to construct R_1^*, the encrypted relation associated to R_1. If $t\in R_1\cap R$, then $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathsf{OPRF}(k_1,t),S[t]\oplus\mathsf{OPRF}(k_2,t)\oplus\cdots\oplus\mathsf{OPRF}(k_n,t))$.
5. For each tuple $t\in R_i$ (with $2\leq i\leq n$), $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ computes $(\mathsf{OPRF}(k_1,t),\mathsf{OPRF}(k_i,t))$ to construct R_i^*, the encrypted relation associated to R_i.
6. Finally, $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$ outputs $(R_1,\ldots,R_n,R_1^*,\ldots,R_n^*,R,S)$.

Now we show that we have:

$$\begin{aligned}&\mathcal{S}^{H_n-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))) \stackrel{\mathrm{c}}{\equiv} \\ &\quad\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I)))\,.\end{aligned}$$

Let η be the security parameter. Assume there exists a non-uniform probabilistic-polynomial time distinguisher D such that for all inputs I, we have:

$$\begin{aligned}&\big|\Pr[D(\mathcal{S}^{H_n-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1] \\ &-\Pr[D(\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1]\big| \\ &=\mu(\eta)\,,\end{aligned}$$

where $\mu(\cdot)$ is a non-negligible function in η.

We construct a non-uniform probabilistic-polynomial time guessing algorithm $\mathcal{B}$ that uses D to win the $\mathsf{IND\text{-}CPA}$ experiment. Algorithm $\mathcal{B}$ works as follows:

1. $\mathcal{B}$ generates n relations $R_1,\ldots,R_n$ of same schema such that $\#R_i=N_i$ for $1\leq i\leq n$, and $\#R=N$.
2. $\mathcal{B}$ generates n secret keys $k_1,\ldots,k_n$ according to the security parameter η.
3. For each tuple $t\in R$, $\mathcal{B}$ picks a random string-bits of same length than t, computes $\mathcal{E}(pk,LR_b(r,t))$ and stores the result into $S[t]$.
4. For each tuple $t\in R_1\setminus R$, $\mathcal{B}$ computes $(\mathsf{OPRF}(k_1,t),\mathcal{E}(pk,t)\oplus\mathsf{OPRF}(k_2,t)\oplus\cdots\oplus\mathsf{OPRF}(k_n,t))$ to construct R_1^*, the encrypted relation associated to R_1. If $t\in R\cap R_1$, then $\mathcal{B}$ computes $(\mathsf{OPRF}(k_1,t),S[t]\oplus\mathsf{OPRF}(k_2,t)\oplus\cdots\oplus\mathsf{OPRF}(k_n,t))$ and stores the tuples into R_1^*.
5. For each tuple $t\in R_i$ (with $2\leq i\leq n$), $\mathcal{B}$ computes $(\mathsf{OPRF}(k_1,t),\mathsf{OPRF}(k_i,t))$ to construct R_i^*, the encrypted relation associated to R_i.
6. $\mathcal{B}$ invokes D on input $(R_1,\ldots,R_n,R_1^*,\ldots,R_n^*,\{(S[t],S[t])\colon t\in R\})$.

First, we remark that:

$$\Pr[\mathsf{Exp}^{\text{ind-cpa-0}}_{\Pi,\mathcal{B}}(\eta)=1]= \Pr[D(\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1]\,.$$

When $b=0$, the view that $\mathcal{B}$ uses as input for D is computed as in the simulator $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. Then the probability that the $\mathsf{IND\text{-}CPA}$ experiment returns 1 is equal to the probability that the distinguisher D returns 1 on inputs computed as in the simulator $\mathcal{S}^{\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. On the other hand, we have:

$$\Pr[\mathsf{Exp}^{\text{ind-cpa-1}}_{\Pi,\mathcal{B}}(\eta)=1]= \Pr[D(\mathcal{S}^{H_n-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}((1^\eta,I_{\mathcal{C}},g_{I_{\mathcal{C}}}(I)),(1^\eta,I_{\mathcal{U}},g_{I_{\mathcal{U}}}(I))))=1]\,.$$

Indeed, when $b=1$ the view that $\mathcal{B}$ uses as input for D is computed as in the simulator $\mathcal{S}^{H_n-\mathsf{SI}}_{\mathcal{C},\mathcal{U}}$. Then the

probability that the IND-CPA experiment returns 1 is equal to the probability that the distinguisher D returns 1 on inputs computed as in the simulator $\mathcal{S}_{C,\mathcal{U}}^{H_n-\mathsf{SI}}$.

Finally, we evaluate the probability that $\mathcal{B}$ wins the experiment:

$$\begin{aligned}\mathsf{Adv}_{\Pi,\mathcal{B}}^{\text{ind-cpa}}(\eta) &= \big|\Pr[\mathsf{Exp}_{\Pi,\mathcal{B}}^{\text{ind-cpa-1}}(\eta)=1] \\ &\quad -\Pr[\mathsf{Exp}_{\Pi,\mathcal{B}}^{\text{ind-cpa-0}}(\eta)=1]\big| \\ &= \big|\Pr[D(\mathcal{S}_{C,\mathcal{U}}^{H_n-\mathsf{SI}}((1^\eta,I_C,g_{I_C}(I)),(1^\eta,I_\mathcal{U},g_{I_\mathcal{U}}(I))))=1] \\ &\quad -\Pr[D(\mathcal{S}_{C,\mathcal{U}}^{\mathsf{SI}}((1^\eta,I_C,g_{I_C}(I)),(1^\eta,I_\mathcal{U},g_{I_\mathcal{U}}(I))))=1]\big| \\ &= \mu(\eta)\,,\end{aligned}$$

which is non-negligible. However, we assume that Π is IND-CPA. Hence, we have:

$$\begin{aligned}&\mathcal{S}_{C,\mathcal{U}}^{H_n-\mathsf{SI}}((1^\eta,I_C,g_{I_C}(I)),(1^\eta,I_\mathcal{U},g_{I_\mathcal{U}}(I))) \overset{\mathsf{c}}{\equiv} \\ &\quad \mathcal{S}_{C,\mathcal{U}}^{\mathsf{SI}}((1^\eta,I_C,g_{I_C}(I)),(1^\eta,I_\mathcal{U},g_{I_\mathcal{U}}(I)))\,.\end{aligned}$$

By transitivity, we have:

$$\mathsf{view}_{C,\mathcal{U}}^{\mathsf{SI}}(I,\eta) \overset{\mathsf{c}}{\equiv} \mathcal{S}_{C,\mathcal{U}}^{\mathsf{SI}}((1^\eta,I_C,g_{I_C}(I)),(1^\eta,I_\mathcal{U},g_{I_\mathcal{U}}(I)))\,.$$

which concludes the proof. □

6 Conclusion

We have presented an efficient privacy-preserving protocol using the MapReduce paradigm to compute the intersection between an arbitrary number of relations. In fact, in the standard protocol (Leskovec et al., 2014), the public cloud performing the computation learns all tuples of the data owners along the intersection result that it sends to the user. In our protocol SI, the public cloud cannot learn such information on the input sets. Moreover, if the cloud and the user collude, then they cannot learn more than the result of the intersection. If no such a collusion exists, then the public cloud only learns cardinals of the relations sent by the data owner, and the cardinal of their intersection. We have compared the standard and our secure approach SI with respect to three fundamental criteria: computation cost, communication cost, and privacy guarantees. We also implemented SI with the Hadoop framework and presented empirical results showing the scalability of SI.

Looking forward to future work, we plan to study secure set intersection with MapReduce while considering a malicious public cloud, i.e., the public cloud can perform any operations on data that it process.

REFERENCES

Aggarwal, C. C. and Yu, P. S. (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining*.

Baldi, P., Baronio, R., Cristofaro, E. D., Gasti, P., and Tsudik, G. (2011). Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In *CCS*.

Boneh, D. and Shoup, V. (2017). *A Graduate Course in Applied Cryptography*.

Cristofaro, E. D., Kim, J., and Tsudik, G. (2010). Linear-Complexity Private Set Intersection Protocols Secure in Malicious Model. In *ASIACRYPT*.

Cristofaro, E. D. and Tsudik, G. (2010). Practical private set intersection protocols with linear complexity. In *FC*.

Daemen, J. and Rijmen, V. (2002). *The Design of Rijndael: AES*. Information Security and Cryptography. Springer.

Dean, J. and Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. In *OSDI*.

Freedman, M. J., Nissim, K., and Pinkas, B. (2004). Efficient Private Matching and Set Intersection. In *EUROCRYPT*.

Hazay, C. and Nissim, K. (2010). Efficient Set Operations in the Presence of Malicious Adversaries. In *PKC*.

Jarecki, S. and Liu, X. (2010). Fast secure computation of set intersection. In *Security and Cryptography for Networks*.

Kissner, L. and Song, D. X. (2005). Privacy-preserving set operations. In *CRYPTO*.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of Massive Datasets, 2nd Ed*. Cambridge University Press.

Lindell, Y. (2017). How to simulate it - A tutorial on the simulation proof technique. In *Tutorials on the Foundations of Cryptography*.

Mezzour, G., Perrig, A., Gligor, V. D., and Papadimitratos, P. (2009). Privacy-preserving relationship path discovery in social networks. In *CANS*.

Nagaraja, S., Mittal, P., Hong, C., Caesar, M., and Borisov, N. (2010). Botgrep: Finding P2P bots with structured graph analysis. In *USENIX*.

Narayanan, A., Thiagarajan, N., Lakhani, M., Hamburg, M., and Boneh, D. (2011). Location privacy via private proximity testing. In *NDSS*.