Hindawi Publishing Corporation
Journal of Engineering
Volume 2014, Article ID 470416, 10 pages
http://dx.doi.org/10.1155/2014/470416

Research Article

Analysis and Evaluation of Schemes for Secure Sum in Collaborative Frequent Itemset Mining across Horizontally Partitioned Data

Nirali R. Nanavati, Prakash Lalwani, and Devesh C. Jinwala

Computer Engineering Department, Sardar Vallabhbhai National Institute of Technology, Surat 395007, India

Correspondence should be addressed to Nirali R. Nanavati; nirali1111@gmail.com

Received 25 August 2014; Accepted 10 November 2014; Published 30 November 2014

Academic Editor: Jiun-Wei Horng

Privacy preservation while undertaking collaborative distributed frequent itemset mining (PPDFIM) is an important research direction. The current state of the art for privacy preservation in distributed frequent itemset mining for secure sum in a horizontally partitioned data model comprises primarily public key based homomorphic schemes which are expensive in terms of the communication and computation cost. The nonpublic key based existing state-of-the-art scheme by Clifton et al. used for secure sum in PPDFIM is efficient but prone to security attacks. In this paper, we propose Shamir's secret sharing based approaches and a symmetric key based scheme to calculate the secure sum in PPDFIM. These schemes are information theoretically secure under the standard assumptions. We further give a detailed theoretical and empirical evaluation of our proposed schemes for PPDFIM using a real market basket dataset. Our experimental analysis also shows that our schemes perform better in terms of the execution cost compared to the public key based scheme for secure sum in PPDFIM.

1. Introduction

With numerous participants mining the data to gain insightful information useful to themselves, there is an inclination to share this information [1, 2]. With the increase in competition in businesses, it has also become essential to know how the competitors are performing. The primary concern in such a scenario is that each of the competitors does not want to disclose their individual data. Hence, privacy preservation is an important concern wherein collaborative distributed data mining needs to be undertaken.

Privacy preservation in distributed data mining (PPDDM) is a significant secure multiparty computation (SMC) problem among other SMC problems [3–5]. SMC helps in knowing how the competitors are performing without compromising on either party's privacy. The issue of SMC is such that only the data mining results of each of the sites that satisfy a certain function are known in the cumulative data. The confidential data of the collaborating parties remains private.

In this paper, we focus on improving the state of the art of the privacy preserving techniques for PPDFIM (which is a subset of the area of PPDDM) in a horizontally partitioned or homogenous data model [6] considering semihonest adversaries as shown in Figure 1.

Some important application scenarios of PPDFIM include medical data, market basket data, network data, data gathered by government agencies, and media related data [6]. An example of a PPDFIM application scenario using market basket data is shown in Figure 2. Once the globally frequent itemsets are found, the secure sum subprotocol is repeated to find the globally frequent association rules for the problem of privacy preserving distributed association rule mining (PPDARM).

An efficient scheme is required in PPDFIM as the size of the data involved is often huge. Hence, PPDFIM involves computations that are repetitive in nature and occur in several rounds of data passing and thus requires an efficient scheme to do so for all practical purposes.

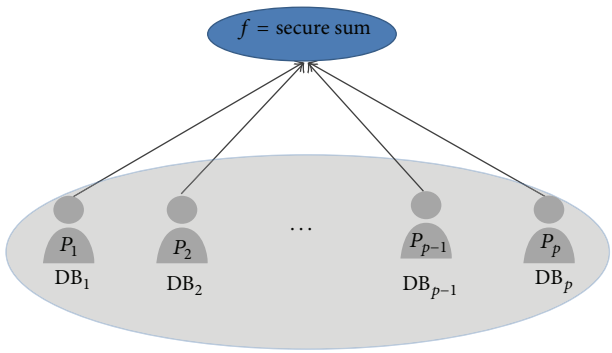

Figure 1: Semihonest adversary model for secure sum in PPDFIM.

The current state-of-the-art protocols for finding secure sum for PPDFIM in a homogenous model involve public key based schemes [7]. However, the security of these schemes is based on computationally hard problems. They are expensive in terms of the computation and communication cost as they work with large numbers (in the order of 1000s of bits) [8]. The existing *naïve* state-of-the-art scheme by Kantarcioglu and Clifton [6] without using public key homomorphism used for secure sum in PPDARM is efficient but prone to security attacks by an external adversary in a semihonest model. In [6], the authors propose a secure sum scheme that is performed using the same key for encryption and decryption. However, this scheme can reveal each party's data in case an outsider is eavesdropping on the consecutive communication channels or if the parties are colluding. Hence, it has to use confidential channels to avoid these passive attacks which would in turn make the scheme more expensive. In [9], the authors propose a game theoretic scheme to avert the insider collusion attack in the secure sum protocol but they do not discuss the outsider attacks. As an alternative if we use the secure sum scheme proposed in [7] which uses public key homomorphic encryption, the overall execution cost increases.

Our schemes are secure against the passive attacks unlike [6]. Also the proposed schemes are more efficient than the public key secure sum scheme in [7].

We propose schemes for secure sum for PPDFIM based on information theoretically secure protocols. These protocols have the highest level of security wherein the adversary simply does not have enough information to break the encryption [10]. These protocols do not have to work with very large numbers which in turn reduces the computation and communication cost.

Hence, for the problem of undertaking secure sum in PPDFIM, we first propose a scheme based on Shamir's secret sharing (with a no third party (NoTP) model [11–13] and a semihonest trusted third party (STTP) model [14]). We observe a high communication cost in these schemes with increase in the number of parties. Hence, for scenarios with larger number of parties, we propose a symmetric key based secure sum scheme based on [15]. We also compare the two schemes with the Paillier based secure sum scheme proposed in [7]. The Paillier based scheme has the least message expansion wherein additive homomorphic schemes are concerned. We also analyse the pros and cons of each of these schemes in a semihonest model since none of the existing works in literature show a comparative analysis between these schemes for secure sum in PPDDM.

We observe that the Shamir's schemes proposed are more efficient in terms of execution cost up to fifteen parties in our setup after which the symmetric key based scheme performs better. Hence, depending on the number of parties, the more efficient scheme could be chosen for the secure sum protocol. Our schemes for finding the secure sum schemes can also be extended to k-means clustering and naïve Bayes classifier.

Thus, we summarize our contribution as follows.

(i) We propose Shamir's scheme (NoTP and STTP model based on [14]) for secure sum in PPDFIM.

(ii) We propose a symmetric key based scheme based on [15] for secure sum in PPDFIM.

(iii) Compare these schemes with the state-of-the-art secure sum scheme using Paillier homomorphic encryption [7].

(iv) There are detailed theoretical analysis and empirical evaluation of these schemes on a real retail dataset [16] from a Belgian based store.

2. Related Work

The PPDFIM algorithms are classified based on the privacy preserving techniques used. The privacy preserving techniques can be based on perturbation or cryptography. The perturbation based techniques are data or knowledge hiding [17, 18], which involves suppression of the sensitive data, randomization that has schemes for distorting the data, summarization wherein only the summary of the data is revealed. These perturbation based techniques lead to loss of accuracy in the cumulative result [19].

Hence, we primarily focus on the cryptographic techniques that do not compromise on the accuracy of the results. When cryptographic techniques [20, 21] are used, the plaintext is transformed to ciphertext. In these cryptographic techniques, the collaborative parties know only the output of the global function on their cumulative data and not the individual secret values.

The state of the art [6, 7, 11, 13] for finding the cryptographic techniques of secure sum in PPDDM involves public key schemes and secret sharing schemes. The secret sharing schemes have not been formalized particularly in PPDFIM in a homogenous setup. Also, symmetric key based schemes have not been formalized in PPDDM.

The state-of-the-art naïve secure sum scheme proposed by [6] is prone to outsider attacks and collusion attacks. Hence, the authors of [6] proposed an enhanced version using public key schemes [7]. However, the public key based scheme incurs a higher computation cost [7].

Nanavati and Jinwala [13] proposed efficient Shamir's scheme for finding global cycles in temporal association rules. Hence, in this paper we further apply Shamir's additive secret sharing scheme without a third party and with a semihonest

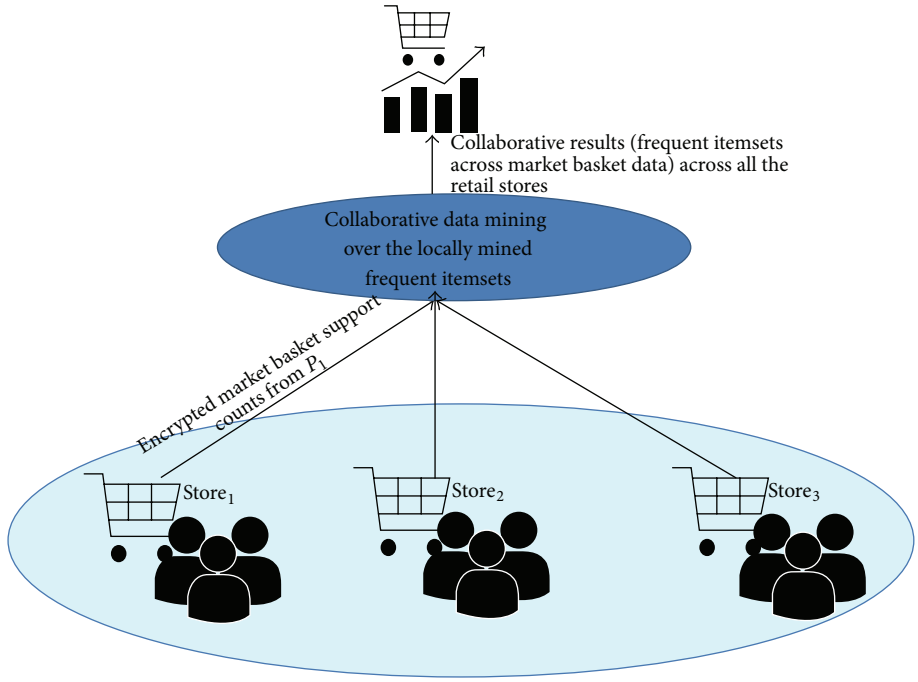

Figure 2: Collaborative market basket analysis. An application scenario for PPDFIM.

third party [14] to the problem of finding of the secure sum in PPDFIM.

Modern symmetric key encryption schemes have been formalized in wireless sensor networks [15, 22] and smart metering [23] where the number of participating entities is large. However, PPDFIM using a secure symmetric homomorphic encryption scheme has not yet been proposed for multiparty secure addition. The symmetric scheme proposed in [15] is similar to the one time pad. It is argued that for an equivalent level of security, asymmetric schemes are generally less efficient than symmetric ones. With proper key management this scheme provides unconditional security and is highly efficient [15]. It also overcomes the attacks in [6].

The other alternative for efficient public key schemes is the Elliptic curve based ciphers (ECC) based additive schemes. However, the commonly used ECEG scheme introduces an issue in which the message text must be mapped on the EC [24].

Hence, for the problem of undertaking secure sum in PPDFIM, we propose an efficient symmetric key based scheme based on [15] where the keys are generated using pseudorandom functions in a semihonest model.

To the best of our knowledge, none of the works in literature give a comparative analysis between these schemes for PPDFIM. Hence, we further show a comparative analysis of the symmetric key based scheme with the secure sum scheme based on Paillier public key homomorphic scheme and the information theoretically secure Shamir's secret sharing scheme in the NoTP model and the STTP model. These schemes are analysed for the horizontally partitioned real market basket retail [16] dataset.

3. Preliminaries

We discuss the theoretical background for our approach in this section.

3.1. Distributed Frequent Itemset Mining. This is an interesting problem which is applied to distributed environments. For our scenario of distributed frequent itemset mining, the global support for itemset $\{A, B\}$ can be calculated using the following generic formulas [6]:

$$\mathbf{Support_{AB}} = \frac{\sum_{i=1}^{\text{sites}} \text{support\_ct\_}AB\,(i)}{\sum_{i=1}^{\text{sites}} \text{db\_size}\,(i)}. \quad (1)$$

The Apriori algorithm [6] can easily be converted to the distributed scenario by using this lemma: if a rule has support $>$ k% globally, it must have support $>$ k% on one of the individual sites. However, our problem statement requires that this global Support_{AB} be found without the other parties knowing about the support counts of the other parties, namely, support_ct_$AB(i)$.

3.2. Data and Adversary Model for Our Proposed Schemes. Databases can generally be partitioned into two different categories: horizontally partitioned and vertically partitioned databases.

Horizontally partitioned databases have the same schema across all the partitions. However, the number of records may vary and each site has information on different entities.

Further, we explain the semihonest adversary model used for our proposed schemes. The semihonest adversary model comprises parties that are honest but curious. They

follow the protocol faithfully but can try to infer the secret information of the other parties from the data they see during the execution of the protocol [25]. The semihonest model is a very simple and efficient model.

We further discuss the third party based models to find the secure sum. The major advantage of the third party based model is it limits the amount of interactivity in the protocol and increases the efficiency.

We use a semihonest trusted third party (STTP) in our schemes. An STTP may try and infer information from the protocol run but does not misbehave on its own [26]. In realistic setup institutions like the bank or the cloud play the role of a STTP [27]. STTPs are worthy of study because, unlike trusted third parties, they are practical and are not naïve like the trusted third party (TTP) model. We use a STTP in our schemes using secret sharing using a STTP based on [14] and the symmetric key based scheme using a STTP based on [15].

3.3. PPDFIM: Horizontally Partitioned Databases. The primary classification of the algorithms for PPDFIM is on the basis of the type of partitioning. Based on the different partitioning, different subprotocols are used for privacy preservation.

The seminal work [6] explains that in horizontally partitioned databases; primarily two phases are required for PPDFIM. The two phases are as follows: the first is discovering candidate itemsets (those that are frequent on one or more sites). The second phase is determining which of the candidate itemsets meet the global support/confidence thresholds.

The first phase uses commutative encryption to find the candidate itemsets. The subprotocol is called the secure set union [6]. However, this is not our focus of study. In the second phase (which is our focus of study), each of the locally supported itemsets is tested to see if it is supported globally [6]. For example, the itemset $\{A, B, C\}$ is known to be supported at one or more sites. Each party has computed their local support. Further, the secure sum subprotocol is used to find the global support count of the itemset and hence to find if the itemset is frequent or not frequent globally which is the problem of PPDFIM. The same protocol is repeated to find the global confidence count and hence to find the globally frequent association rules. We focus on the secure sum subprotocol in our work.

3.4. Shamir's Secret Sharing Scheme for SMC. This is a collusion resistant algorithm based on Shamir's secret sharing [11–13] technique which is inherently information theoretic. We are studying Shamir's $[m, m]$ secret sharing scheme which is additively homomorphic in nature. This scheme is used by us for the secure sum protocol in our application.

Shamir's secret sharing scheme has been proposed for the vertically partitioned model in [11]. We proposed Shamir's secret sharing in [13] for global cycle detections. However Shamir's secret sharing has not been applied to the generic problem of secure sum in PPDFIM. We proposed secret sharing using a STTP in [14] to improve on the efficiency of the NoTP model. In this scheme the STTP is used in Phase 2 of the protocol specified in Algorithm 1 to find the sum of shares. We analyse and compare both of these schemes along with the symmetric key based scheme and the public key based additive homomorphic scheme for secure sum in PPDFIM.

4. The Problem Statement

We consider a coopetitive scenario of homogenous or horizontally partitioned databases where there are "p" parties that are semihonest. These parties aim to collaboratively find the itemsets that are frequent globally without disclosing their identities.

Hence, the aim is to calculate the global support counts of the candidate itemsets privately for PPDFIM which is our case study. Once the same procedure is repeated using the confidence values, we get the global association rules.

To formalize our problem, let there be a set of "t_i" transactions and maximum n items at each of the partitions "i" (where $0 < i \leq p$) and each transaction has a subset of "n" items. N is the set of candidate itemsets for whom the total support count needs to be calculated.

Let the schema at each of the sites be of the form ⟨Transaction_id items_bought⟩. The parties undertake secure union of the locally frequent itemsets exceeding minSupp as shown in [6]. This gives the output as the candidate itemsets. Now the global support needs to be calculated using the secure sum subprotocol for the support counts of the candidate itemsets which is our focus of study.

We propose Shamir's scheme (NoTP and STTP model based on [14]) for secure sum in PPDFIM. We further propose the symmetric key based scheme based on [15] which has not been formalized in PPDDM as has only been proposed for wireless sensor networks in [15]. We compare these schemes with the public key based homomorphic scheme in [7].

The notations used in the proposed algorithm by us are shown in the notations section.

5. Security Model

The security and trust model of our scenario depends on the protocol used. We discuss a comparative analysis of Shamir's secret sharing scheme (NoTP and STTP model) [14] with the proposed symmetric key based scheme and the secure sum scheme based on Paillier homomorphic encryption [7] for secure sum in PPDFIM.

We consider a semihonest adversary model for our scenario. In the NoTP based Shamir's secret sharing scheme, we use a mesh topology for the participating semihonest parties. For the STTP based scheme proposed by us in [14] we use a mesh topology for Phase 1 and a star topology in Phase 2. The secret sharing based schemes for PPDFIM are discussed in Algorithm 1. The public key based secure sum scheme [7] is based on a ring topology.

In the symmetric key based scheme (Algorithm 2) which has not been formalized in PPDDM, the topology is a ring topology and all the parties are assumed to be semihonest.

Setup:
The common random numbers $X = \{x_1, \ldots, x_p\}$ are distinct publically available numbers in a finite field **F** of size P where P is a prime number and secret $V_i < P$ $(0 < i \leq p)$.
The coefficients $\{a_1, \ldots, a_{p-1}\} < P$.

(1) **for** each global candidate itemset $j = 1$ to N
(2) **for** each party P_i, $(i = 1, 2, \ldots, p)$ **do**
(3) each party selects a random polynomial $q_i(x) = a_{p-1}x^{p-1} + \cdots + a_1x^1 + V_{ij}$
(4) compute the share of each party P_y $(y = 1, 2, \ldots, p)$, where share$(V_{ij}, P_y) = q_i(x_i)$
(5) **for** $y = 1$ to p $(i \neq y)$ **do**
(6) send share(V_{ij}, P_y) to party P_y
(7) receive the shares share(V_{ij}, P_y) from every party P_y
(8) **end for**
(9) compute Sum$(x_i) = q_1(x_i) + q_2(x_i) + \cdots + q_p(x_i)$
(10) **end for**
(11) ***if*** *NoTP model is used*
(12) **for** each party P_i, $(i = 1, 2, \ldots, p)$ **do**
(13) **for** $y = 1$ to p $(i \neq y)$ **do**
(14) send Sum(x_i) to party P_y
(15) receive the results Sum(x_i) from every party P_y
(16) solve the set of p equations to find the sum of $\sum_{i=1}^{p} V_{ij}$ secret values.
(17) **end for**
(18) **end for**
(19) **end if**
(20) ***else if*** *STTP model is used*
(21) **for** each party P_i, $(i = 1, 2, \ldots, p)$ **do**
(22) send Sum(x_i) to STTP
(23) solve the set of p equations at STTP to find the sum of $\sum_{i=1}^{p} V_{ij}$ secret values.
(24) **end for**
(25) **for** each party P_i, $(i = 1, 2, \ldots, p)$
(26) STTP sends the final list of the computed sum of secret values to each party
(27) **end for**
(28) **end if**
(29) **end for**

Algorithm 1: Proposed algorithm for PPDFIM using Shamir's secret sharing based on NoTP model and STTP model.

There is also a STTP for the protocol that does not collude with any of the parties. It does not require confidential channels like the naïve secure sum scheme proposed by [6]. Our proposed algorithm is resistant to all passive attacks by semihonest adversaries.

6. The Proposed Algorithms

In this section, we propose secure encryption schemes that allow efficient additive aggregation of encrypted values for PPDFIM.

Shamir's secret sharing based scheme based on [14] for secure sum for PPDFIM with and without the third party is given in Algorithm 1.

For the proposed symmetric key based scheme based on [15] only one modular addition is necessary for cipher text aggregation. The security of the scheme is based on the pseudorandom number generator (PRNG), a standard cryptographic primitive. The idea is to perform a modular addition of a classic stream cipher with the secret. Every party uses a different pseudorandom stream as mentioned in [15]. We assume that the seed for the PRNG has been exchanged with the STTP using a public key, key exchange protocol and since we are using a pseudorandom stream we do not need to exchange the keys repetitively but after periodic intervals. The details of the key exchange are however not our scope of study for this protocol. In [15], the authors promise security with small cipher text sizes. We propose a PPDFIM algorithm using the symmetric scheme in Algorithm 2.

7. Theoretical Analysis

Given below is the theoretical analysis of the secret sharing and the symmetric key based approaches for finding the global support for the problem of PPDFIM. The public key based scheme for PPDFIM has been analysed in [7]. The summarized theoretical analysis for all the schemes for secure sum in PPDFIM is given in Table 1.

7.1. Correctness Analysis. We assume that the party P_i $(0 < i \leq p)$ has V_{ij} $(0 < j \leq N)$ values corresponding to the support counts of the candidate frequent itemsets N. For the secure sum protocol in PPDFIM, our goal is to find the value of $\sum_{i=1}^{p} V_{ij}$ for all j.

Setup: Each party exchanges its seed with the STTP securely.
$M \geq p * \max\{V_i\}$ (V_i is the plain text at the ith party) $(0 < i \leq p)$; $V_i \in Z_M$; $K_i \in Z_M$.
Frequentitems = ∅.
Require: $p \geq 2$.
(1) **for** each global candidate itemset $j = 1$ to N
(2) **for** each party P_i $(i = 1, 2, \ldots, p)$
Encryption phase:
(3) Each party has a secret $V_{ij} \in Z_M$ where M is the modulus
(4) Each party generates a key stream $K_{ij} \in Z_M$ where M is the modulus
(5) Ciphertext $C_{ij} = V_{ij} + K_{ij} \bmod M$
(6) **end for**
Aggregation:
(7) $\text{cipherSum}_j = 0$
(8) **for** $i = 1$ to p
(9) $\text{cipherSum}_j = \text{cipherSum}_j + C_{ij} \bmod M$
(10) P_i sends cipherSum_j to P_{i+1}*
(11) **end for**
Decryption at STTP:
(12) $\text{Sum}_j = \text{cipherSum}_j$
(13) $\text{Sum}_j = \text{Sum}_j - \sum_{i=1}^{p} K_{ij} \bmod M$
(14) P_t broadcasts the value of Sum_j to all the parties
(15) if (Sum_j > global support threshold)
(16) Frequentitems = Frequentitems U Itemset_j
(17) **end for**

(The same procedure is followed for finding the frequent association rules
once the frequent itemsets are deciphered in case of PPDARM.)
* Pp sends the sum value to P_t which is a semi-honest trusted third party (STTP).
(i) If $\text{Sum} = \sum_{i=1}^{p} V_i > M$, decryption produces Sum < M. In practice, M must be chosen as $M = p * \max\{V_i\}$.

Algorithm 2: Proposed algorithm for secure sum in PPDFIM using symmetric keys.

Table 1: Comparative analysis for schemes to evaluate secure sum for PPDFIM.

Symmetric key based scheme based on [15]	Shamir's secret sharing (NoTP model) based on [11, 14]	Shamir's secret sharing (STTP model) based on [14]	Paillier based scheme for secure sum [7]
$O(N * p)$ modular additions and generation of $O(N * p)$ modular addition pseudorandom streams	Computation is done in the same field as the secret.	Computation is done in the same field as the secret.	It requires expensive operations: modular exponentiation of large numbers (1000s of bits).
Communication cost: **O(N * p)—Phase 1 and Phase 2**	Communication cost: **O(N * p^2)—Phase 1 and Phase 2**	Communication cost: **O(N * p^2)—Phase 1; O(N * p)—Phase 2**	Communication cost: **O(N * p) —Phase 1 and Phase 2**
Information theoretically secure	Information theoretically secure	Information theoretically secure	Computationally secure

For the secret sharing based scheme (NoTP and STTP), the $\sum_{i=1}^{p} V_{ij}$ is calculated at each of the parties and at the STTP, respectively. Thus, the semihonest parties know only the sum of the secrets (global support count for candidate itemsets) and their original secrets. They do not know the secrets of the other parties.

The analysis of the secure sum scheme based on additively homomorphic Paillier scheme can be found in [7]. Now for the symmetric key based protocol proposed for PPDDM using the ring topology, each party generates the ciphertext C_{ij} by adding its plaintext to the key stream generated by using a pseudorandom number generator. The cipher text is sent to the $(i + 1)$th party which adds its ciphertext to it until the last party is reached. This pth party sends the final sum to the STTP. The STTP being aware of the pseudorandom generator and the seed subtracts the sum of the key streams using the equation $\text{Sum}_j - \sum_{i=1}^{p} K_{ij} \bmod M$. This will in turn give the value of $\sum_{i=1}^{p} V_{ij}$ which is the global support count. Hence, the correctness of the protocol is verified.

7.2. Complexity Analysis. For the symmetric key based scheme, the complexity analysis involves the communication and the computation costs. If we consider "p" parties and "N" as the list of candidate frequent itemsets, the communication

cost of our scheme to find globally frequent itemsets is $O(p * N)$. This is because the first phase involves a ring topology and then finally the sum is broadcasted to all the parties. We do not consider the communication cost for the exchange of the seeds with the STTP after periodic intervals of time as the seed exchange is out of our scope of study. The computation cost involves $O(p * N)$ modular additions for p parties and N global candidate itemsets and the cost of generating $O(p * N)$ pseudorandom streams.

The communication cost for Shamir's scheme (NoTP model) is $O(N * p^2)$ for each of the phases. However, the communication cost for Shamir's scheme (STTP model) is $O(N * p^2)$ for the first phase and $O(N * p)$ for the second phase. Along with the communication cost there would also be the computation cost of generating the random polynomial, p^2 polynomial evaluations, and $p(p-1)$ additions and solving the equations with unknowns to find the sum $\sum_{i=1}^{p} V_{ij}$.

For the Paillier based secure sum scheme the communication cost is $O(N * p)$ as the topology is a ring topology. As far as the computation cost is concerned it involves Paillier's additive homomorphic encryption of $O(p * N)$ values and for the leader to decrypt $O(N)$ values.

7.3. Security Analysis. Our protocol using symmetric keys to find the secure sum for PPDFIM is semantically secure and preserves the privacy of the participants.

The privacy of the participants is preserved as the parties do not know the support counts of each other as they merely get the ciphertext. The STTP gets the value of $\sum_{i=1}^{p} V_{ij}$ and not the individual support counts and hence the privacy is preserved.

As far as the security is concerned in the symmetric key based scheme, if there is an eavesdropper/outsider he is only able to see the encrypted data and cannot get the hold of the actual values. Our protocol is as secure as the one time pad. Our scheme using Shamir's secret sharing is also information theoretically secure [10].

This is not the case with the secure sum protocol mentioned in [6] without using public keys as in that case the eavesdropper is able to predict the individual values.

However, as far as collusion of the symmetric key based scheme is concerned, even if the parties collude they will not get the data of the other noncolluding parties as they only have the cipher texts. Also the security of the scheme is based on the indistinguishability property of a pseudorandom function (PRF) and the lack of randomness in those generators or in their initialization vectors is disastrous for the protocol [15]. As a result, keys should be changed on a regular basis and kept secure during distribution [15].

The security analysis of Shamir's scheme with the NoTP and STTP model is based on the assumption that each party or an outsider cannot get all p shares of the secret. This property in turn makes the scheme information theoretically secure.

The Paillier based protocol is computationally secure because any party other than the trusted leader site cannot decrypt the encrypted sum value as explained in [7]. Also since the values are encrypted, the outsiders cannot know the individual secrets.

8. Performance Evaluation

In this section we give the details of the methodology of evaluation, the simulator used, the metrics, inputs of evaluation, and the datasets used.

8.1. Methodology of Evaluation. For our scenario where there are distributed coopetitive parties involved, we have shown our experimental results using the real retail [16] dataset for scenarios up to 20 parties.

We model our multiparty scenario by randomly dividing the data among all the parties using horizontal partitioning. We model four schemes for comparison in a PPDFIM scenario being Paillier based secure sum, proposed Shamir's additive secret sharing (STTP and NoTP model), and the proposed symmetric key based scheme for secure sum in PPDFIM.

The schemes are implemented in Java for PPDFIM on a noncloud single machine using the simulator SimJava [28] based on multithreading. We used an Intel Core i5 CPU with 6 GB RAM and 2.5 GHz speed and a 64-bit Operating System for our implementation.

Once the data set is generated, we have implemented frequent itemset mining at each of the parties. The next step is to calculate the secure union to find the candidate itemsets using the Pohlig-Hellman scheme as mentioned in [6]. The choice of the scheme and the secure union subprotocol is not our focus of study. The individual support counts are then communicated among the parties privately using the symmetric key based scheme proposed in this paper, the secure sum scheme [7], and the noncollusive secret sharing schemes proposed in [14]. Finally we are able to decipher the globally frequent itemsets that have a cumulative count greater than the global support threshold. The methodology of evaluation is given in Figure 3.

8.2. Details of Simulation. The SimJava simulator [28] is used for the simulation of a distributed setup using event simulations in Java. We have implemented scenarios to calculate global support counts privately for coopetitive setups up to 20 parties.

8.3. Dataset. We consider a market basket data from a Belgian based store [16] with 17000 items and about 28K transactions (dataset 1) at each site. We consider a scenario of maximum 20 coopetitive parties for the test application. The aim is to predict the globally frequent itemsets among the coopetitive parties privately.

8.4. Inputs of Evaluation. We evaluate our algorithm based on the following inputs.

(i) We have carried out our experiments on a retail dataset [16] for the same number of candidate itemsets.

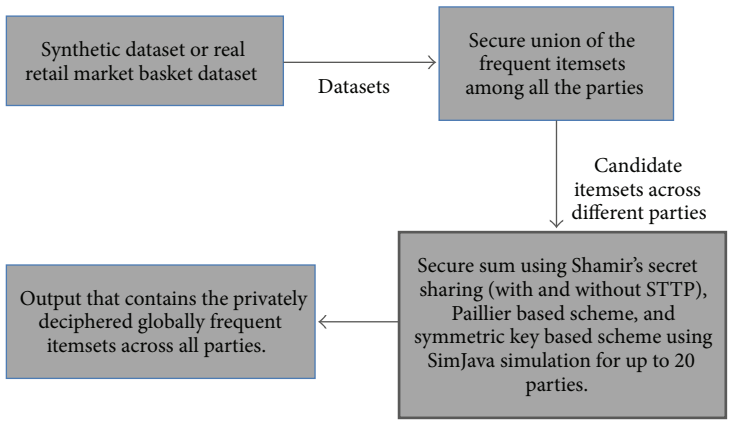

Figure 3: Methodology of evaluation for the proposed algorithms in a PPDFIM setup.

(ii) Different privacy preserving secret sharing techniques include Shamir's secret sharing [14] for semi-honest parties (using NoTP and STTP models), Paillier based secure sum scheme [7], and proposed symmetric key based scheme.

(iii) Number of participating parties: our algorithm works for a multiparty simulation with scenarios involving up to 20 parties.

The experiments were performed on the same datasets multiple times and an average of those experimental results has been taken.

9. Results and Analysis

Given below are the performance results followed by the empirical analysis of the schemes proposed by us.

9.1. Performance Results. We carry out our experiments to calculate the secure sum in the PPDFIM scenario using the dataset 1 [16]. In Figure 4, we show the performance results for the four techniques based on Paillier based secure sum, Shamir's secret sharing scheme (NoTP model and STTP model), and the proposed symmetric key based scheme. These experiments are carried out using SimJava for up to 20 participating parties.

9.2. Empirical Analysis. From Figure 4 we observe that the execution cost increases with the increase in the number of parties in the coopetitive setup. This is because the value of p increases which has an impact on the communication and the computation cost of all the setups.

Further, we observe that Shamir's secret sharing schemes perform the best in terms of the execution cost for lesser number of parties. However, as the number of parties increases the communication cost increases by $O(p^2)$ in Shamir's scheme as the topology is a mesh topology.

Shamir's schemes with NoTP and the proposed symmetric key based scheme break even at number of parties equal to 10. Shamir's scheme with STTP and the symmetric key based scheme break even at 15 parties. This is because

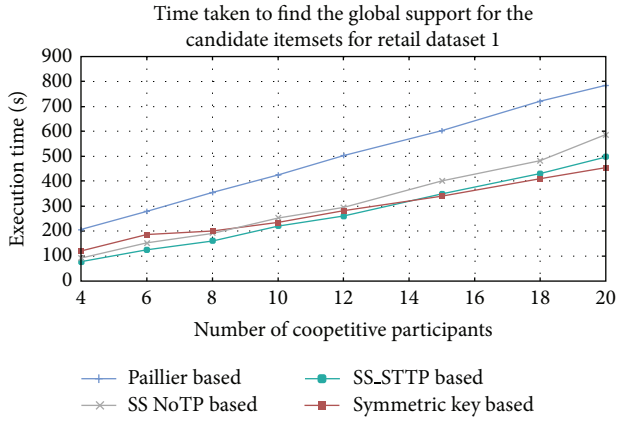

Figure 4: Graph showing the execution time for secure sum to find globally frequent itemsets for dataset [16].

even though the symmetric scheme has the overhead of the PRNG at the sender and at the third party for each of the $O(N * p)$ values, still, with increase in number of parties the communication cost increases only as $O(N * p)$. Hence, this behaviour is displayed. The Paillier based secure sum scheme for PPDFIM however shows a high execution cost compared to the other two schemes due to the mathematical operations being carried on very large numbers (1024 bits in our experiments) [8]. However, the NoTP secret sharing scheme and the Paillier based scheme would converge for higher number of parties.

10. Conclusion

We propose secure, efficient schemes for secure sum in privacy preserving distributed frequent itemset mining. The execution cost in our proposed symmetric key based scheme breaks even with Shamir's secret sharing scheme (NoTP model) at a threshold of 10 parties and with Shamir's secret sharing scheme (STTP model) at 15 parties. The execution cost of the symmetric key based schemes is lower than the

public key scheme. Also our protocol is more secure than the state-of-the-art secure sum protocol [6].

Hence, the symmetric key based information theoretically secure schemes would be ideal for scenarios where the number of participating parties is large. However, for scenarios with lesser number of parties (corporate model [29]), the information theoretically secure Shamir's scheme performs better in terms of the execution cost.

Notations

p:	Number of collaborating parties
t_i:	Set of transactions at each of the parties i $(0 < i \leq p)$
minSupp:	Global minimum support
n:	Total number of items
N:	Set of candidate itemsets for which the global support is to be calculated
STTP:	Semitrusted third party.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. Kantarcioglu and R. Nix, "Incentive compatible distributed data mining," in *Proceedings of the IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust (SocialCom/PASSAT '10)*, pp. 735–742, 2010.

[2] P. Kamakshi and A. Babu, "Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data," *Journal of Computing*, vol. 2, no. 4, pp. 115–119, 2010.

[3] W. Du and M. J. Atallah, "Secure multi-party computation problems and their applications: a review and open problems," in *Proceedings New Security Paradigms Workshop (NSPIN '01)*, V. Raskin, S. J. Greenwald, B. Timmerman, and D. M. Kienzle, Eds., pp. 13–22, ACM, September 2001.

[4] O. Goldreich and A. Warning, "Secure multi-party computation," 2002, http://www.wisdom.weizmann.ac.il/~oded/pp.html.

[5] P. Bogetoft, D. Christensen, I. Damgard et al., "Secure multiparty computation goes live," in *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, pp. 325–343, Springer, Berlin, Germany, 2009.

[6] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1026–1037, 2004.

[7] M. Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-Preserving Data Mining*, vol. 34 of *Advances in Database Systems*, pp. 313–335, Springer, New York, NY, USA, 2008.

[8] T. Pedersen, Y. Saygin, and E. Savas, "Secret Sharing vs. Encryption-based Techniques For Privacy Preserving Data Mining," Sciences-New York, December, pp. 17–19, 2007.

[9] H. Kargupta, K. Das, and K. Liu, "Multi-party, privacy-preserving distributed data mining using a game theoretic framework," in *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '07)*, pp. 523–531, Springer, Berlin, Germany, 2007.

[10] "Information-theoretic Security," http://en.wikipedia.org/wiki/Information-theoretic_security.

[11] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in *Proceedings of the 2nd International Conference on Software Engineering and Data Mining (SEDM '10)*, pp. 345–350, June 2010.

[12] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[13] N. R. Nanavati and D. C. Jinwala, "Privacy preserving approaches for global cycle detections for cyclic association rules in distributed databases," in *Proceedings of the SECRYPT*, pp. 368–371, SciTePress, 2012.

[14] N. R. Nanavati, N. Sen, and D. C. Jinwala, "Analysis and evaluation of efficient privacy preserving techniques for finding global cycles in temporal association rules across distributed databases," *International Journal of Distributed Systems and Technologies, IGI Global*, vol. 5, no. 3, pp. 58–76, 2014.

[15] C. Castelluccia, A. C.-F. Chan, E. Mykletun, and G. Tsudik, "Efficient and provably secure aggregation of encrypted data in wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 5, no. 3, pp. 1–36, 2009.

[16] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "Using association rules for product assortment decisions: a case study," in *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, pp. 254–260, San Diego, Calif, USA, 1999.

[17] V. S. Verykios and A. Gkoulalas-Divanis, "A survey of association rule hiding methods for privacy," in *Privacy-Preserving Data Mining*, pp. 267–289, 2008.

[18] C.-W. Lin, T.-P. Hong, and H.-C. Hsu, "Reducing side effects of hiding sensitive itemsets in privacy preserving data mining," *The Scientific World Journal*, vol. 2014, Article ID 235837, 12 pages, 2014.

[19] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: current scenario and future prospects," in *Proceedings of the 3rd International Conference on Computer and Communication Technology (ICCCT '12)*, pp. 26–32, Allahabad, India, November 2012.

[20] A. Evfimievski and T. Grandison, *Privacy Preserving Data Mining*, IBM Almaden Research Center, 2007.

[21] C. C. Aggarwal and P. S. Yu, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-Preserving Data Mining*, vol. 34 of *The Kluwer International Series on Advances in Database Systems*, pp. 11–52, Springer US, 2008.

[22] C. Castelluccia, E. Mykletun, and G. Tsudik, "Efficient aggregation of encrypted data in wireless sensor networks," in *Proceedings of the 2nd Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous '05)*, pp. 109–117, IEEE Computer Society, July 2005.

[23] B. Vetter, O. Ugus, D. Westhoff, and C. Sorge, "Homomorphic primitives for a privacy-friendly smart metering architecture," in *Proceedings of the International Conference on Security and Cryptography (SECRYPT '12)*, pp. 102–112, SciTePress, 2012.

[24] S. Peter, K. Piotrowski, and P. Langendoerfer, "On concealed data aggregation for WSNs," in *Proceedings of the 4th IEEE*

Consumer Communications and Networking Conference (CCNC '07), pp. 192–196, Las Vegas, Nev, USA, January 2007.

[25] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *CRYPTO '00 Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology*, pp. 36–54, Springer, London, UK.

[26] J. Hong, J. Kim, M. K. Franklin, and K. Park, "Fair threshold decryption with semi-trusted third parties," *International Journal of Applied Cryptography*, vol. 2, no. 2, pp. 139–153, 2010.

[27] V. Vinod, A. Narayanan, K. Srinathan, C. P. Rangan, and K. Kim, "On the power of computational secret sharing," in *Progress in Cryptology—INDOCRYPT 2003*, vol. 2904 of *Lecture Notes in Computer Science*, pp. 162–176, Springer, Berlin, Germany, 2003.

[28] F. W. Howell and R. McNab, "Simjava: a discrete event simulation package for Java with applications in computer systems modelling," in *Proceedings of the 1st International Conference on Web-Based Modelling and Simulation*, pp. 51–56, San Diego, Calif, USA, 1998.

[29] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," in *Proceedings of the US National Science Foundation Workshop on Next Generation Data Mining*, H. Kargupta, A. Joshi, and K. Sivakumar, Eds., pp. 126–133, 2002.