Secure Association Rule Mining on VerticallyPartitioned Database

Ms. Nisha Gupta (Research Scholar)

Dr. Kalpana Sharma (Assistant Professor)

Computer Science & Engineering Department, Bhagwant University Ajmer (Rajasthan), India

Abstract:In this paper context of Data mining entails the discovery of unexpected but reusable knowledge from large unorganizeddatasets. In recentyears, organizations in different fields have been required to collaborate to create new value. However, datamining among and within organizations has raised privacy and confidentiality concerns.

In our proposal,parties can notcontribute tosomething other than the number of records, as well as the candidate item set. This studyfocuses on the private-set intersection as a substitute of the scalar product and shows that this intersection enablesorganizations to execute ARM on vertically divided data, allowing elastic informationcontribution whilepreserving privacy without escalating communication and totalling costs. Besides we spotlight on thestatement that the number of protocol rounds among parties can be reduced and present three use cases in whichthe proposed scheme works more effectively than the alive schemes..

Keywords-Privacy preserving, association rules mining, association rule hiding, frequent itemsets, private set intersection.

Introduction: The ubiquity of internet-of-things (IoT) devices and the peoplewho use them has generated tremendous amounts of dataWorldwide. Furthermore, cloud storage penetrationand increases in network speed make it possible to storeand distribute the data as they grow. Cisco foresees a massiveincrease in internet traf_c, projecting 4.8 ZB per yearby 2022 [2]. Data mining employs and exploits the discoveryand management of reusable and perhaps unexpected knowledgefrom large unorganized datasets. Many algorithms have been designed for efficient and automatic analysis of thesedata so that users can accumulate, assimilate, interpret, andunderstand the knowledge obtained. ARM is one of the majority common data-mining algorithms.

ARM is used to streamlinesales, optimize e-commerce advertisements, and mitigatesoftware development obstacles, among many other applications.Generally, ARM has been used to aggregate data intoone location and then mine those data [3]_[7]. When data are merged and mined, confidential information can easily beaccumulated. It becomes necessary to shield the isolation ofsuch data by protecting them from unauthorized exploitation.A privacy-preserving association rule-mining (PPARM) processhas thus been proposed. It performs data mining whileupholding data security and privacy requirements. Therefore,PPARM has attracted widespread attention as a technologyfor data sanctuary and isolation protection.Several schemes have been proposed to implement miningin distributed data environments. These schemes havebeen broadly separated into those that use secure multi-partycomputation (SMC) and those that use cryptographic techniques.A two-party secure-computation scheme was proposedto directly execute a computation protocol with inputfrom two parties without the help of a third party [8].

SMC extends secure two-party computation to any party,allowing them to execute desired computations without sharinginput values. No communication apart from the protocolis required. Therefore, it is necessary to design the protocolappropriately according to its purpose. Cryptographictechnique_based schemes use encryption (e.g., homomorphic),and, because mining is performed with encrypteddata, users lacking the secret key cannot obtain any usefulinformation.

Data privacy is protected, allowing data ownersto delegate mining activities to third parties. Decryption iscertainly required to obtain results, and SMC and cryptographictechnique_based schemes have their strengths andweaknesses. Thus, we choose a scheme that suits our use case.Distributed data environments can follow many patterns. Thetwo main types of distributions are horizontal and vertical.Horizontal distribution:With horizontal distribution, differentparties collect various recordsets to determine commonattributes. Their databases are horizontally distributed so thatthe columns are the same, but the rows are different.

Data owners often outsource data storage and miningtasks in ARM. Data owners need to store data on a cloud serveror request mining tasks from a third-party service providerbecause their expertise, resources, and storage are insufficient for the amount of data accumulated and computationalresources required for mining.When delegating data storage or mining to a third party,data privacy is important due to various security threats.

However, even if data are encrypted, or an appropriate algorithm is used, confidential information and mining resultsmay be leaked because the data are entrusted to a thirdparty. In addition, cloud services have risks such as settingerrors, unauthorized access, and API vulnerabilities. Morethan 70 million records were stolen or leaked in 2018 dueto poorly con_gured Amazon Simple Storage Service(Amazon S3) buckets [36]. In 2019, UpGuard reported thatdata were open due to a setting error of Amazon S3 [37].To comply with industry standards, data owners also needto understand access control and where data are stored. However,it may take time to check the settings, or the serviceprovider may not provide accurate settings when using acloud service. An (ISC)2survey found that one in four organizations had experienceda cloud security incident in the preceding 12 months .

According to the survey, the biggest challenges for cloudsecurity are data loss (64%) and data privacy (62%).

In particular, it is difficult for multiple companies to usecloud services together. The companies would not want touse a private cloud because of the additional costs of designand operation. Even when using a public cloud, it is necessaryto use a provider with no stake in the data owners to reducethe risk of collusion between the data owner and the cloudservice provider. Therefore, this study proposes a schemein which ARM can be performed only by the data ownerswithout outsourcing data storage and mining.

This study assumes that ARM among organizations existsin different industries, where each organization has data andshares their results. Because not every party should sharetheir data, we consider SMC-based schemes.

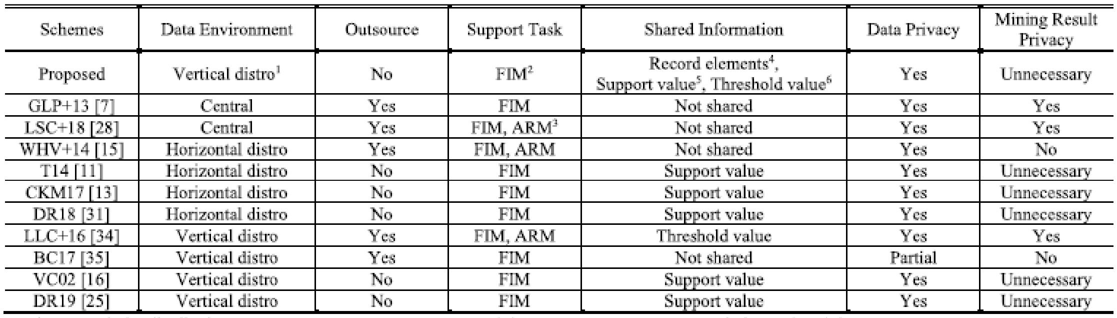
Vertical datadistribution is used because each party has different data. Theproposed ARM scheme (i.e., VC02) uses a scalar product forinformation sharing among parties [16]. With VC02, partiescan share the number of common records without unnecessaryinformation exchange. Conversely, parties cannot shareanything other than the number of common records. Therefore,we concentrate on the private-set intersection insteadof the scalar product. This study shows that the private-setintersection enables the execution of ARM on vertically partitioneddata without changing communication and computationcosts.

The contributions of this study are as follows:

1- The parties can exchange information specified bythe parties and execute the ARM without outsourcingthe ARM.

2- Flexible information exchange, such as the number ofcommon records and record elements, including iteamsets,and whether or not thresholds are exceeded, can be accomplished.

3- Information exchange uses a private-set inter section ,and only information determined in advance is shared.RelatedWorks. Table 1 summarizes the characteristics ofrelated research, including the proposed scheme.



Schemes	Data Environment	Outsource	Support Task	Shared Information	Data Privacy	Mining Result Privacy
Proposed	Vertical distro[1]	No	FIM[2]	Record elements[4], Support value[5], Threshold value[6]	Yes	Unnecessary
GLP+13 [7]	Central	Yes	FIM	Not shared	Yes	Yes
LSC+18 [28]	Central	Yes	FIM, ARM[3]	Not shared	Yes	Yes
WHV+14 [15]	Horizontal distro	Yes	FIM, ARM	Not shared	Yes	No
T14 [11]	Horizontal distro	No	FIM	Support value	Yes	Unnecessary
CKM17 [13]	Horizontal distro	No	FIM	Support value	Yes	Unnecessary
DR18 [31]	Horizontal distro	No	FIM	Support value	Yes	Unnecessary
LLC+16 [34]	Vertical distro	Yes	FIM, ARM	Threshold value	Yes	Yes
BC17 [35]	Vertical distro	Yes	FIM	Not shared	Partial	No
VC02 [16]	Vertical distro	No	FIM	Support value	Yes	Unnecessary
DR19 [25]	Vertical distro	No	FIM	Support value	Yes	Unnecessary

TABLE 1. A comparison of main features in related PPARM schemes.

Classified by data environment and labeled Central,Horizontal distribution, and Vertical distribution. In a centralenvironment, if a data owner does not outsource, privacyprotection is not important because only the data ownerknows the raw data and mining results. Therefore, in general,outsource is assumed for PPARM in the central environment.

In a distributed environment, sharing/merging data and miningtasks may be outsourced to a trusted third party or maybe processed among parties without outsourcing.ARM can be divided into the task of extractingfrequent itemsets, calculating support, and confidence(see section II.A) followed by comparing them to thresholdsfor the generation of association rules. Table 1 shows the former task as frequent itemset mining (FIM) and the latter asARM. Given the frequent itemsets and their support values,data owners can generate association rules, so PPARM generally focuses on FIM. A Scheme that can execute ARMwithout sharing the result of FIM has been proposed.If the target dataset is held by multiple data owners, the dataowners may share data for mining. In ARM, to determineif a candidate itemset is a frequent itemset, the data ownerneeds to know whether the support value of the candidateitemset exceeds the threshold value. Therefore, the supportvalues of the candidate itemsets are shared, or only theresults of whether or not they exceed the threshold value are shared to keep the support value secret. Information sharing

is unnecessary when the data owner does not perform mining;thus, schemes [15], [35] that implement ARM tasks withoutsharing information by the data owner have been proposed.

The raw data information held by the data owners mustbe kept secret from others, including other data owners.

Most schemes consider data privacy, but the partial databasecontents are known to the data outsourcing destinationin BC17 [35]. Furthermore, when data are outsourced andmined by someone other than the data owners, the data ownerswant to protect the privacy of the mining results. Becausethe mining results are the property of the data owners, leakingthem would impair pro_ts. For example, when building asales promotion plan from sales data mining, leakage ofmining results to competitors can adversely affect sales. Someschemes require the cloud server to know the mining result,in which case, the privacy of the mining result is not protected.

Our proposed scheme can perform FIM without outsourcingin a vertically distributed environment. The major differencefrom other schemes is that data owners can share various information.

II. PRELIMINARIES

A. ARM

ARM was originally proposed to and relationships amongitems from supermarket transaction data. The ARM problem can be formally stated . Let J be a set of all items,and database D consists of a set of transactions over J . LetT D .tid; I / be a transaction over J . The tid is the transaction identiuer and is defined in D to make the transaction unique.I is a set of items from J . I 2 J and I 6D ;.When theprobabilities of itemset X or Y in the transactions are 30%and 10%, respectively, the probability that both itemsets are included in the transaction is predicted to be 3%. If both itemsetshave a 15% chance of being included in the transaction,X and Y will be related. However, because it is necessary toindicate whether there is a high probability of buying Y whenbuying X or a high probability of buying X when buying Y ,the association rules have a direction. Let X) Y be an associationrule with antecedent X and consequent Y . The supportand con_dence are used to evaluate the association rules. Thesupport of rule X) Y is de_ned as the ratio of transactionsincluding X and Y as a whole: Support .X) Y / D_X [Y / = jDj, where _ .X/ indicates the number of transactionsthat satisfy condition X. The confidence of rule X) Y is defined as the value obtained by dividing the number oftransactions including X and Y by the number of transactionsincluding X. Con_dence .X H) Y / D _ .X [Y / =_ .X/.Given database D and two threshold values, minsupp and con_dence .X) Y / > minconf in D.

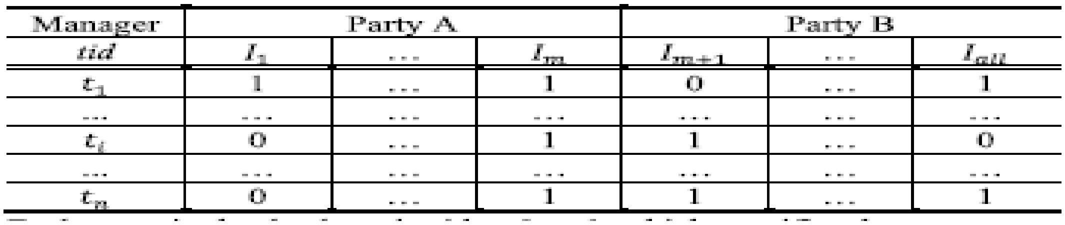
B. APRIORI ALGORITHM

ARM can be divided into two phases. The first phase _ndsfrequent itemsets that exceed minsupp. The second phase ands itemsets that exceed minconf from the frequent itemsetsdetermined in the _rst phase. The number of rule candidatesincreases sharply with the increasing number of items in thedatabase. An apriori algorithm is proposed to _nd frequentitemsets ef_ciently. It skips the calculation by using the featurethat the support of an itemset is less than or equal to thesupport of the sub-itemset. The sets of transactions are treatedas a database with n rows and m columns to execute thealgorithm more ef_ciently.

Each row and column correspondto a transaction and an item, respectively. Each entry in thedatabase is either 0 or 1, specifying the presence or absenceof an item. If the i-th row and j-th column in the databasecorrespond to transaction ti and item Ij, respectively, thenthe j-th entry in row i, denoted by ti [j], indicates whether ticontains Ij.

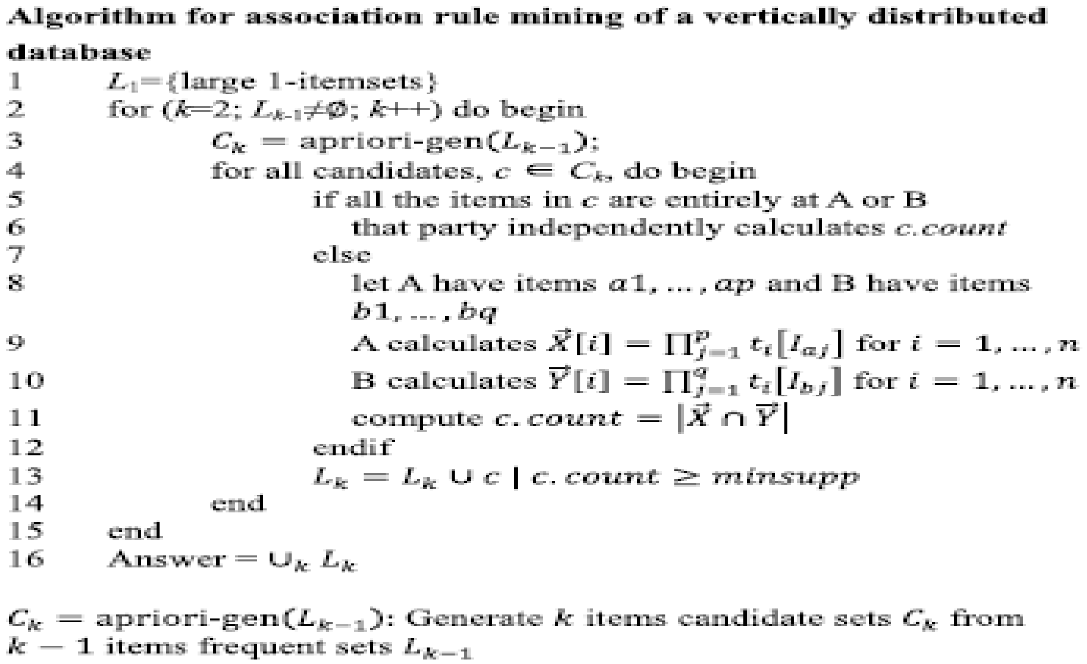
C. DISTRIBUTED ARM

This study considers a vertically distributed database.Database D is split vertically into two sets of columns.D includes all items Iall and is divided into DB1 witha column set of items I1 to Im and DB2 with a columnset of items ImC1 to Iall . Table 2 shows an exampleof the divided database. Parties A and B manageDB1 and DB2, respectively, and cannot browse the contentsof the other DBs.



Manager	Party A			Party B		
tid	I_1	...	I_m	I_{m+1}	...	I_{all}
t_1	1	...	1	0	...	1
...	...	...	...	...	...	...
t_i	0	...	1	1	...	0
...	...	...	...	...	...	...
t_n	0	...	1	1	...	1

TABLE 2. Divided transaction database.



Algorithm for association rule mining of a vertically distributed database

```
$L_1$={large 1-itemsets}
for ($k$=2; $L_{k-1}$≠∅; $k$++) do begin
        $C_k$ = apriori-gen($L_{k-1}$);
        for all candidates, $c \in C_k$, do begin
                if all the items in $c$ are entirely at A or B
                    that party independently calculates $c.count$
                else
                    let A have items $a1, \ldots, ap$ and B have items
                      $b1, \ldots, bq$
                    A calculates $\vec{X}[i] = \prod_{j=1}^{p} t_i[I_{aj}]$ for $i = 1, \ldots, n$
                    B calculates $\vec{Y}[i] = \prod_{j=1}^{q} t_i[I_{bj}]$ for $i = 1, \ldots, n$
                    compute $c.count = |\vec{X} \cap \vec{Y}|$
                endif
                $L_k = L_k \cup c \mid c.count \geq minsupp$
        end
end
Answer = $\cup_k L_k$
```

C_k = apriori-gen(L_{k-1}): Generate k items candidate sets C_k from $k-1$ items frequent sets L_{k-1}

ALGO 1. This algorithm extends the apriori algorithm to the vertically distributed database.

necessary to calculate the number of transactions, including the candidate itemset. If the candidate itemset is composedof only A and B items, the entities mutually calculate thenumber of transactions. Conversely, if the candidate itemsetexists across A and B, it is necessary to exchange informationwith each other. First, each party generates an *n*-dimensionalvector for each item in the candidate itemset. If the number of transactionsis larger than *minsupp*, it is regarded as a frequentitemset. Repeat this process until there are no more candidateitemsets. Frequent and candidate itemsets need to be shared,but confidential information of each party is not disclosed in itemset sharing because the frequent itemsets are shared onlyin the final step.

D. SECURITY MODEL

This study assumes a semi-honest adversary because the usecase entails communication among trusted organizations.

1) SEMI-HONEST ADVERSARIES

In this model, both parties follow the actions they are supposedto take according to the protocol. However, they can tryto deduce more information from the data they obtain during execution.We follow [42] and assume that participants in ourprotocol include a receiver, *R*, and a sender, *S*. The receiver,*R*, executes the protocol and receives the _nal result. Thesender, *S*, provides information per the protocol. The definitions of receiver and sender security are as follows:

2) THE SECURITY OF THE RECEIVER_INDISTINGUISHABILITY

This security requires that the sender cannot distinguish between *R* and *S*, even if the inputs of the receiver aredifferent.

3) THE SECURITY OF THE SENDER COMPARISON TO THEIDEAL MODEL

This security requires that the receiver cannot obtain more information than specified. An ideal implementation will formalize that definition. The ideal model implements athird party that obtains inputs from both parties and outputsthe result. The security of the sender requires that theoutput of the protocol is indistinguishable from the idealimplementation.

III. PROPOSED SCHEME

A. OVERVIEW

The algorithm revealed in Figure 1 does not bring up howto estimate the number ofjunction from the columnvectors derived by each party.

This allowed the two parties toshare only the number of transactions, as well as the candidateitemset. It is enough to simply extract the association rules.However, if the party wants to use rules other than thoseof association, the shared information may be insuficient.Therefore, it is desirable that the parties flexibly changethe information to be shared according to their purposes.Thus, our scheme enables flexible information sharing via the replacement of the scalar product from VC02 [16] witha private matching from FNP04 . Our scheme has thefollowing advantages:

i In addition to the number of transactions that include thecandidate itemset, parties can share which transactionsinclude the candidate itemset and whether the numberof transactions, including those of the candidate itemset,exceeds a threshold. Each party can choose whatinformation they want to share.

ii The information shared by each party is not known tothe other party.

iii Because the anticipated scheme can share the same informationwhen using a scalar product, it is also possibleto ensure that the same rules are generated in advance.

B. PRIVATE SET INTERSECTION

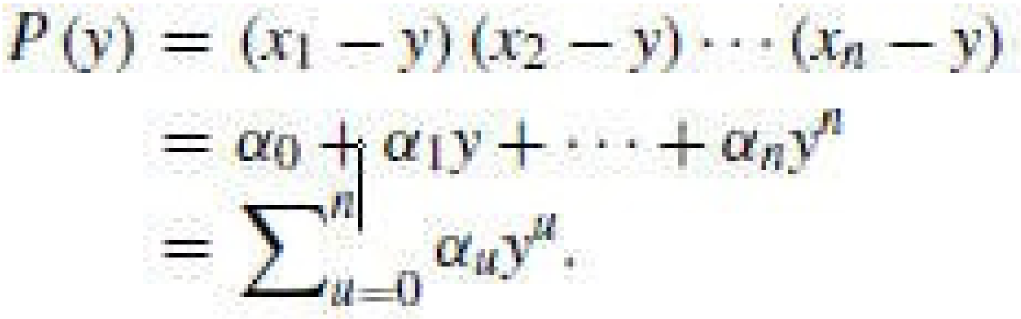
As mentioned in section II.C, each party needs to exchangeinformation and calculate the required value on line 11 ofFigure 1. In this study, we propose to implement line 11 using the private set intersection.

In other words, we replace line 11with a protocol where the inputs are *X* and *Y* , generated onlines 9 and 10, and the output is *c*:*count*. Furthermore, privateset intersection allows output other than *c*:*count*.In this paper, the protocol that outputs only the number oftransactions, including the candidate itemset, such as scalarproduct computation, is called *private cardinality matching*.Similarly, the protocol that outputs the elements of the transactions,including the candidate itemset, for which the sum of the elements has the same result, is called *private matching*.Additionally, the protocol that outputs the result of whetherthe number of transactions, including the candidate itemset,exceeds a threshold is called *private matching for cardinalitythreshold*. We show the flow of the protocol in Figure 2.

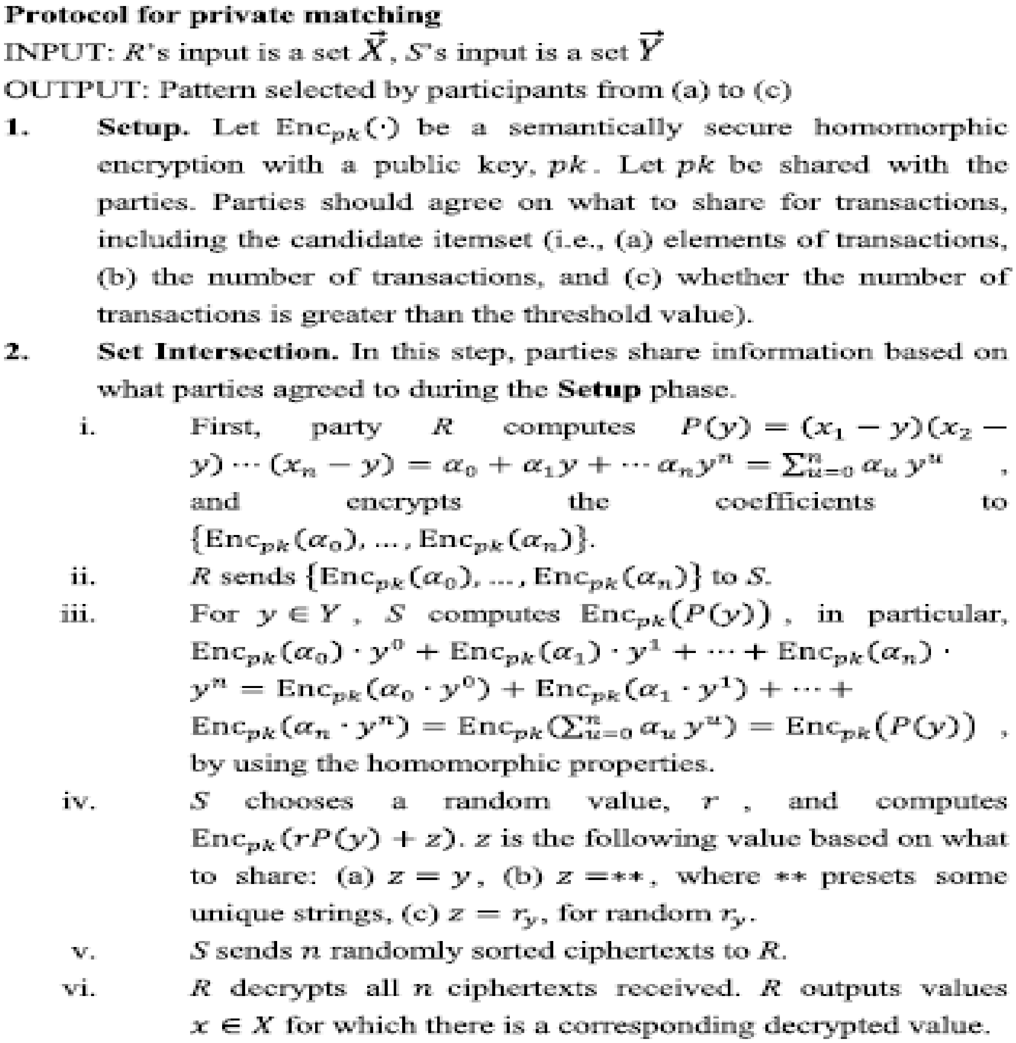
Each party can change its shared information by changingthe value sent during step iv of the Set Intersection phase.Because private matching is the basic pattern, private matchingis described prior to the other two patterns. Inputs ofprotocols are E*X* and E*Y* . However, ``1'' and ``0'' in the columnvector is replaced with the corresponding *tid* and randomvalue not included in *tid*, respectively. Protocol output hasthree patterns: (a) elements of transactions, (b) the numberof transactions, and (c) whether the number of transactions isgreater than the threshold value. We assume participants inour protocol are a receiver, *R*, and a sender, *S*.

1) PRIVATE MATCHING

This section presents the patterns sharing the most information.Here, the parties share elements of transactions, includingthe candidate itemset. First, *R* generates a polynomial,*P* .*y*/, whose root is the input E*X* D f*x*1 *xn*:



$$\begin{aligned} P(y) &= (x_1 - y)(x_2 - y)\cdots(x_n - y) \\ &= \alpha_0 + \alpha_1 y + \cdots + \alpha_n y^n \\ &= \sum_{u=0}^{n} \alpha_u y^u . \end{aligned}$$



Protocol for private matching

INPUT: R's input is a set $\vec{X}$, S's input is a set $\vec{Y}$

OUTPUT: Pattern selected by participants from (a) to (c)

1. **Setup.** Let $\mathrm{Enc}_{pk}(\cdot)$ be a semantically secure homomorphic encryption with a public key, pk. Let pk be shared with the parties. Parties should agree on what to share for transactions, including the candidate itemset (i.e., (a) elements of transactions, (b) the number of transactions, and (c) whether the number of transactions is greater than the threshold value).
2. **Set Intersection.** In this step, parties share information based on what parties agreed to during the **Setup** phase.
   - i. First, party R computes $P(y) = (x_1 - y)(x_2 - y)\cdots(x_n - y) = \alpha_0 + \alpha_1 y + \cdots \alpha_n y^n = \sum_{u=0}^{n} \alpha_u y^u$, and encrypts the coefficients to $\{\mathrm{Enc}_{pk}(\alpha_0), \ldots, \mathrm{Enc}_{pk}(\alpha_n)\}$.
   - ii. R sends $\{\mathrm{Enc}_{pk}(\alpha_0), \ldots, \mathrm{Enc}_{pk}(\alpha_n)\}$ to S.
   - iii. For $y \in Y$, S computes $\mathrm{Enc}_{pk}(P(y))$, in particular, $\mathrm{Enc}_{pk}(\alpha_0)\cdot y^0 + \mathrm{Enc}_{pk}(\alpha_1)\cdot y^1 + \cdots + \mathrm{Enc}_{pk}(\alpha_n)\cdot y^n = \mathrm{Enc}_{pk}(\alpha_0 \cdot y^0) + \mathrm{Enc}_{pk}(\alpha_1 \cdot y^1) + \cdots + \mathrm{Enc}_{pk}(\alpha_n \cdot y^n) = \mathrm{Enc}_{pk}(\sum_{u=0}^{n} \alpha_u y^u) = \mathrm{Enc}_{pk}(P(y))$, by using the homomorphic properties.
   - iv. S chooses a random value, r , and computes $\mathrm{Enc}_{pk}(rP(y) + z)$. z is the following value based on what to share: (a) $z = y$, (b) $z =**$, where $**$ presets some unique strings, (c) $z = r_y$, for random r_y.
   - v. S sends n randomly sorted ciphertexts to R.
   - vi. R decrypts all n ciphertexts received. R outputs values $x \in X$ for which there is a corresponding decrypted value.

Flexible information sharing between R and S withoutrevealing information.*R* expands the polynomial, encrypts the coefficientsusing homomorphic encryption, and sends fEnc*pk* (_0); Enc*pk* (_*n*)g to *S*. *S* generates a ciphertext of *P .y/* usinghomomorphic properties. Then, *S* generates a value to sendto *R* by using random number *r* for all elements *y* of E*Y* ,as follows:Enc*pk* .*P* .*y*// _Enc*pk* .*r*/CEnc*pk* .*y*/ D Enc*pk* .*rP* .*y*/ C *y*/,

and *P* .*y*/ D 0 when *y* is included in the input of *R*. Thus,the calculation result of *S* is Enc*pk* .*y*/. On the other hand,it will be a random value when *y* is not included in theinput of *R*. Because *R* can decrypt the ciphertext, *R* decryptsall values sent from *S,* and obtains the *tid*, including thecandidate itemset. Finally, *R* shares its results with *S*.

2) PRIVATE CARDINALITY MATCHING

This protocol pattern limits the information to be shared.Protocols do not share elements of transactions. Instead, theyshare only the number of transactions. In other words, thispattern shares the same information as the scalar product usedin VC02. We can implement this pattern with only a smallchange to the behaviour of private matching. In the step where*S* calculates ciphertext, *S* enters a unique string, insteadof *y*. *R* decrypts the ciphertext received from *S* and counts thenumber of cipher texts from which was obtained. The totalnumber of this ciphertext matches the number of transactions,including the candidate itemset.

IV. ANALYSIS

A. SECURITY ANALYSIS

The output of the algorithm in Figure 1 includes itemsetshaving *k* elements that are candidates for frequent itemset.If *L*1 is to be shared with the other party, they exchangeinformation only on line 11. Therefore, the security of theanticipated scheme depends on the data-sharing protocol. Theproposed scheme uses FNP04 as the information-sharingprotocol, which ensures the security claims in the semi-honestmodel as follows:

Lemma 1 (Correctness)*:* The protocol participants canobtain the desired result only if *R* and *S* share a common valueby calculating according to the protocol for private matching.In particular, *R* can get the encrypted *y* during step vi duringthe Set Intersection phase if *x* D *y*. Otherwise, *R* obtains theencrypted random number.

Lemma 2 (security of R is preserved): Based on the inputof *R* to the protocol, *R* calculates the polynomial *P* .*y*/ andsends the information to *S*. Specially, *R* calculates *P* .*y*/by using the input value as a root. It encrypts the coefficientof the *P* .*y*/ and sends *S* the encrypted value. Because theinformation obtained by *S* is only the encrypted coefficient, *S*cannot distinguish the input of *R*. Thus, when the encryptionscheme is secure, the privacy of *R* is protected.

Lemma 3 (security of S is preserved): The ideal modelassumes a third party who takes the input E*X* of *R* and the inputE*Y* of *S* and outputs the result of protocols. In this case, the

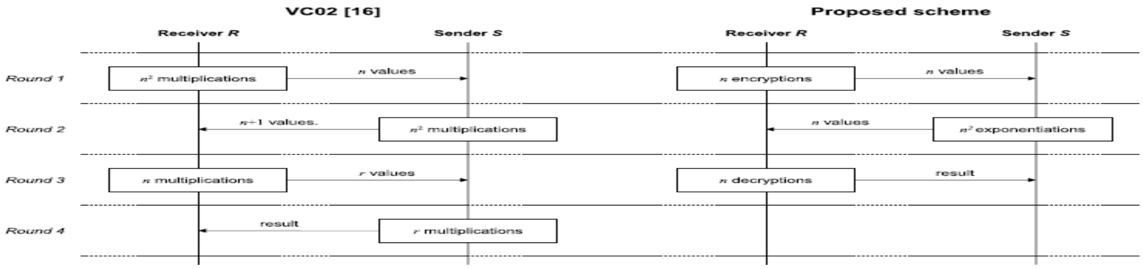


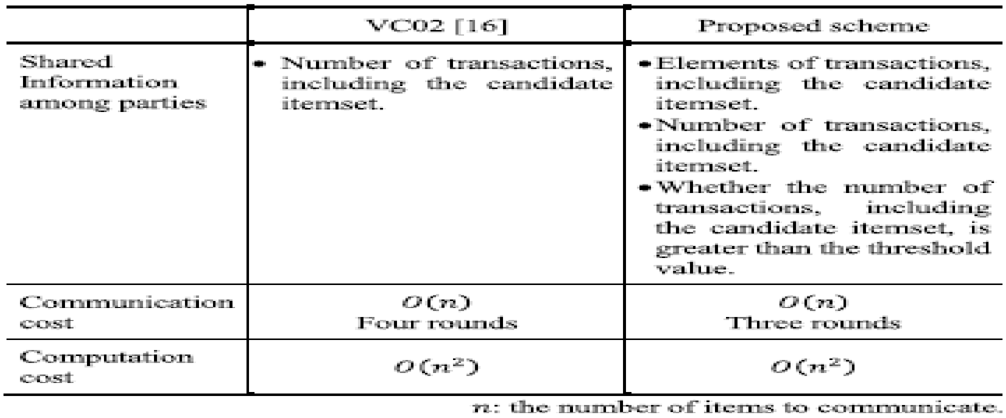
FIGURE 3. Protocols for sharing information for VC02 and the proposed scheme are illustrated. The content of each box shows the computational
overhead of each round. The label on each arrow shows the message to be transmitted.information that R can obtain from the third party is only theresult of $X \setminus Y$. In the real model, R can only obtain Enc .y/or Enc .r/, where r is a random value, so the information that

can be obtained is indistinguishable from that of the idealmodel; thus, the privacy of S is protected.

B. COMMUNICATION/COMPUTATION ANALYSIS

Table 3 shows the results of a comparison between VC02 andthe proposed scheme. The key difference is the type of informationthat can be shared among parties. VC02 can share only one type of information, but the proposed scheme can sharethree types. There is no difference in the amount of communicationper round, but there is a difference in the number

of communication rounds among parties. VC02 requiresfour rounds of communication until the results are obtained,whereas the proposed scheme requires three rounds.Figure 3 shows the _ows of protocols for informationsharing. In VC02, two parties, simply referred to as A and B,execute a four-round protocol on line 11 of Figure 1 to performscalar product. In this study, for the sake of consistency,the party that starts the protocol is called a receiver R, and theother party is called a sender S. First, R sends a message having$n$ values to S, and S responds with a message composedof n C 1 values. In response, R sends a message consistingof r values, where r is a random number determined by $S$and satis_es $n > r$. Finally, S calculates the _nal result andsends it to R. Communication overhead is proportional to $n$and can be expressed as $O(n)$. In the protocol of the proposedscheme, n ciphertexts and the _nal result are transmitted forthree rounds. Communication cost overhead of the proposedscheme is $O(n)$ as with VC02.In VC02, the calculation for generating n values that$R$ and S send _rst requires the largest calculation cost



	VC02 [16]	Proposed scheme
Shared Information among parties	• Number of transactions, including the candidate itemset.	• Elements of transactions, including the candidate itemset. • Number of transactions, including the candidate itemset. • Whether the number of transactions, including the candidate itemset, is greater than the threshold value.
Communication cost	$O(n)$ Four rounds	$O(n)$ Three rounds
Computation cost	$O(n^2)$	$O(n^2)$

n: the number of items to communicate.

TABLE 3. Comparison of schemes.

Because *n* multiplications are needed to make one value,a total of *n*2 multiplications are required. Thus, the overheadof the calculation cost is *O*(*n*2). In the proposed scheme,*R* computes the coef_cients of the polynomial using interpolationand performs *n* instances of encryption and decryption.*S* has the largest computation overhead, and the processingof step iii during the Set Intersection phase requirea computation overhead of *O*(*n*2) exponentiations. In thisround, the computation overhead of Enc*pk .P .y//* is *O*(*n*)exponentiations because it is indispensable to compute *yn*.In addition, *n* multiplications of homomorphically encryptedvalues are required. Because these multiplications are actually implemented as exponentiations, the total overheadis *O*(*n*2) exponentiations. Freedman *et al.* [46] mentionreduction of the computational overhead and explain thatusing a hash function reduces the overhead to *O*(*n* ln ln *n*).However, the reduction of the computation cost is outside thescope of this study because the proposed scheme has featuresother than the computation cost.

V. USE CASES

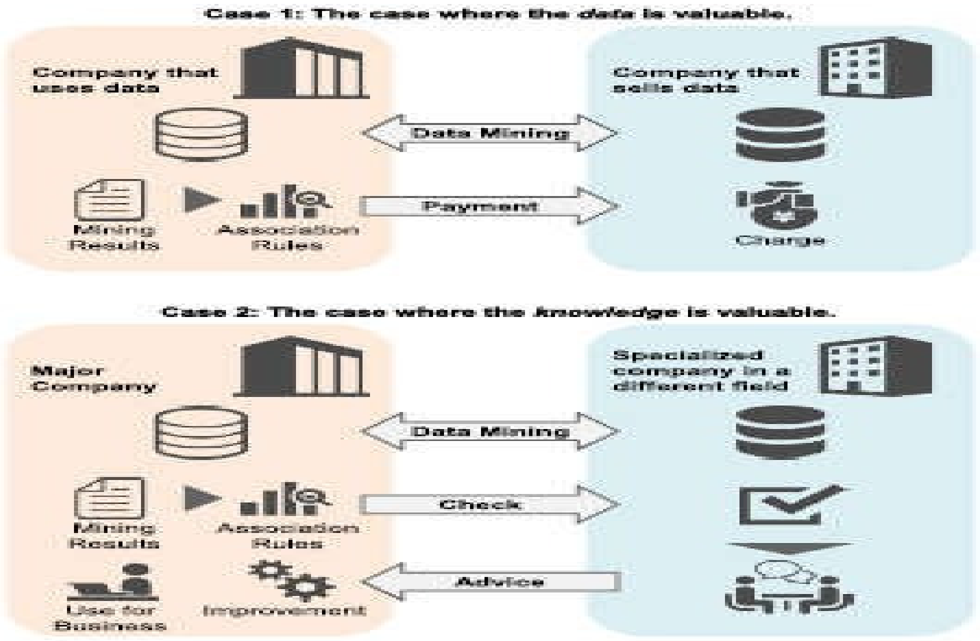
The major difference between protocols of the proposedscheme and VC02 is the number of rounds. Due to the difference,*R* and *S* execute the final round in the proposed schemeand VC02, respectively. In the final round, the party sharesthe final result of the protocol with the other in both schemes.Therefore, if the recipient of the final round does not requirethe final result, the final round can be omitted. The recipientof the final round is *S* in the proposed scheme and *R* in VC02.Because *R* starts the protocol, *R* should need the results of the protocol. On the other hand, we believe that there arecases where *R* does not require the results. In these cases,*R* expects another benefit by cooperating with the protocoland providing its own data. We present three use cases for it.We organize use cases based on the concept of the datainformation knowledge wisdom (DIKW) hierarchy .

This is a framework for interpreting information and consistsof four layers: *Data*, *Information*, *Knowledge*, and *Wisdom*.The four layers are defined as follows according to Data are defined as symbols that represent properties ofobjects, events, and their environment. They are of nouse until they are in a useable form.

If we replace raw data, such as sales data, with ``*data*,"frequent itemsets with ``*information*," and association ruleswith ``*knowledge*," we can consider ARM with this concept.Based on this concept, the proposed scheme and VC02 areschemes for safely sharing *information*. However, in a casewhere the emphasis is on something other than *information*,sharing of *information* is unnecessary, and the protocol could omit the final round. Therefore, Figure 4 shows each use casethat has value in *data*, *knowledge*, and *wisdom*.

Case 1: The case where the *data* are valuable.If a company has valuable data, it will receive a paymentfor providing the data and cooperating with the mining. Thecompany may be specialized in collecting data and sellingdata to the companies that need data. It may be unnecessaryto share the final result of the protocol when demanding aremuneration other than information for providing data.

Case 2: The case where the *knowledge* is valuable.The final result of the protocol in the proposed schemeand VC02 is frequent itemsets. Parties know which itemsare likely to appear in the same data from the frequentitemsets but do not know which items are appropriate as anantecedent or a consequent. Parties cannot properly use the



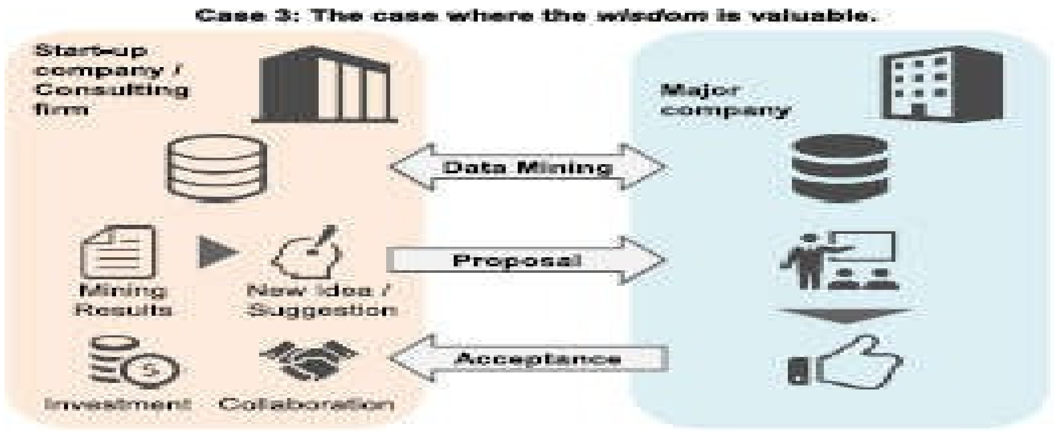


FIGURE 4. Use cases that have value in data, knowledge, and wisdom.mining results until they understand the relationship betweenthe antecedent and the consequent.We consider the concept of open innovation . Supposethat a major company asks for cooperation from a specialized company in a different field. The major company wantsto develop new products or services by combining its dataand specialized data of the specialized company. However, the major company cannot understand the details of thedata because the data of the specialized company is from a different field. The major company asks the specializedcompany for advice. Association rules may help the specializedcompany give appropriate advice because relationshipsamong items are more intuitive. A form, such as knowledge,is preferred for these use cases because that form is easy for people to understand.

Case 3: The case where the *wisdom* is valuable.We consider the case of combining the data of a majorcompany and the data of a startup company to create a newbusiness. The major company wants services and solutionsbased on new ideas the startup company has derived frommining results. In other words, the major company requiresthe startup company to propose their own wisdom, not informationor knowledge. When the major company is impressedwith the wisdom, it offers funding and collaboration opportunitiesto the startup company.

In addition, we could make this case by replacing thestartup company with a consulting . Mining is carriedout using the unique data held by each of them as input. Theconsulting _rm sublimates the information and knowledgeobtained from mining into wisdom and proposes businessimprovements and management strategies to the client company.

In other words, the client company pays the consultingfee for the wisdom proposed by the consulting company.These use cases require unique and personal insights ratherthan the information and knowledge required for mining.Therefore, the proposed scheme is more appropriate thanVC02 for the aforementioned use cases.

VI. CONCLUSION AND FUTURE WORK

This paper presented a secure ARM scheme for verticallypartitioned data. The proposed scheme enables flexible informationsharing by using private-set intersections. Comparedwith existing ARM schemes that use scalar products, ourcommunication and calculation costs were comparable, andmultiple information sharing patterns were achieved. Furthermore,the number of protocol rounds could be reduced fromthe existing scheme. Focusing on this point, we presented usecases for which the proposed scheme works effectively.

In the future, we plan to perform stricter security analyseson the wide use of privacy-preserving data-mining techniques.Furthermore, we plan to investigate other algorithmsand consider more ef_cient data-mining techniques.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning. *The Digitization of the WorldFrom Edge to Core*. IDC. Accessed: Sep. 30, 2019. [Online]. Available:https://www.seagate.com/_les/www-content/our-story/trends/_les/idcseagate-dataage-whitepaper.pdf

[2] C. Public. *Cisco Visual Networking Index: Forecast and Trends,2017_2022*. Cisco. Accessed:Sep.30,2019.[Online].Availablehttps://www.cisco.com/c/en/us/solutions/collateral/ service-provider/visual-networking-index-vni/white-paper-c11-741490.pdf

[3] W. K. Wong, D. W. Cheung, B. Kao, N. Mamoulis, and E. Hung,`Security in outsourcing of association rule mining," in *Proc. 33rdInt. Conf. Very Large Data Bases(VLDB)*, Vienna, Austria, 2007,pp. 111_122.

[4] I. Molloy, N. Li, and T. Li, ``On the (In)Security and (Im)Practicality ofoutsourcing precise association rule mining," in *Proc. 9th IEEE Int. Conf.Data Mining*, Washington, DC, USA, Dec. 2009, pp. 872_877.

[5] C.-H. Tai, P. S. Yu, and M.-S. Chen, ``K-support anonymity based onpseudo taxonomy for outsourcing of frequent itemset mining," in *Proc.16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*,Washington, DC, USA, 2010, pp. 473_482.

[6] F. Giannotti, L. V. S. Lakshmanan, D. Pedreschi, H. Wang, andA. Monreale, ``Privacy-preserving data mining from outsourceddatabases," in *Computers, Privacy and Data Protection: an Elementof Choice*, S. Gutwirth, Y. Poullet, P. De Hert, R. Leenes, Eds. Dordrecht,The Netherlands: Springer, vol. 2011, pp. 411_426.

[7] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, andH.Wang, ``Privacy-preserving mining of association rules from outsourcedtransaction databases,'' *IEEE Syst. J.*, vol. 7, no. 3, pp. 385_395, Sep. 2013.

[8] A. C. Yao, ``Theory and application of trapdoor functions,'' in *Proc. 23rdAnnu. Symp. Found. Comput. Sci. (sfcs)*, Nov. 1982, pp. 80_91.

[9] M. Kantarcioglu and C. Clifton, ``Privacy-preserving distributed miningof association rules on horizontally partitioned data,'' *IEEE Trans. Knowl.Data Eng.*, vol. 16, no. 9, pp. 1026_1037, Sep. 2004.

[10] A. Schuster, R. Wolff, and B. Gilburd, ``Privacy-preserving associationrule mining in large-scale distributed systems,'' in *Proc. IEEEInt. Symp. Cluster Comput. Grid*, Chicago, IL, USA, Apr. 2004,pp. 411_418.

[11] T. Tassa, ``Secure mining of association rules in horizontally distributeddatabases,'' *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 970_983,Apr. 2014.

[12] X. Juan and Z. Yanqin, ``Application of distributed oblivious transferprotocol in association rule mining,'' in *Proc. 2nd Int. Conf. Comput. Eng.Appl.*, Washington, DC, USA, 2010, pp. 204_207.

[13] H. Chahar, B. N. Keshavamurthy, and C. Modi, ``Privacy-preserving distributedmining of association rules using Elliptic-curve cryptosystem andShamir's secret sharing scheme,'' *Sadhan a*, vol. 42, no. 12, pp. 1997_2007,Dec. 2017.

[14] C. N. Modi and A. R. Patil, ``Privacy preserving association rule miningin horizontally partitioned databases without involving trusted third party(TTP),'' in *Proc. 3rd Int. Conf. Adv. Comput., Netw. Inform. (ICACNI)*.New Delhi, India: Springer, 2015, pp. 549_555.

[15] O. A. Wahab, M. O. Hachami, M. Vivas, G. G. Dagher, and A. Zaffari,``DARM: A privacy-preserving approach for distributed association rulesmining on horizontally-partitioned data,'' in *Proc. 18th Int. Database Eng.Appl. Symp.*, Porto, Portugal, 2014, pp. 1_8.

[16] J. Vaidya and C. Clifton, ``Privacy preserving association rule miningin vertically partitioned data,'' in *Proc. 8th ACM SIGKDD Int. Conf.Knowl. Discovery Data Mining (KDD)*, Edmonton, AB, Canada, 2002,pp. 639_644.

[17] S. Zhong, ``Privacy-preserving algorithms for distributed mining of frequentitemsets,'' *Inf. Sci.*, vol. 177, no. 2, pp. 490_503, Jan. 2007.

[18] J. Vaidya and C. Clifton, ``Secure set intersection cardinality with applicationto association rule mining,'' *J. Comput. Secur.*, vol. 13, no. 4,pp. 593_622, Oct. 2005.

[19] X. Ge, L. Yan, W. Shi, and J. Zhu, ``Privacy-preserving distributed association

rule mining based on the secret sharing technique,'' in *Proc.2nd Int. Conf. Softw. Eng. Data Mining*, Chengdu, China, Jun. 2010,pp. 345_350.

[20] R. Kharat, M. Kumbhar, and P. Bhamre, Eds., ``Ef_cient privacy preservingdistributed association rule mining protocol based on randomnumber,'' in *Proc. Intell. Comput., Netw., Inform.*, New Delhi, India, 2014,pp. 827_836.

[21] B. Rozenberg and E. Gudes, ``Association rules mining in verticallypartitioned databases,'' *Data Knowl. Eng.*, vol. 59, no. 2, pp. 378_396,Nov. 2006.

[22] H. Hammami, H. Brahmi, S. Ben Yahia, and I. Brahmi, ``Using homomorphicencryption to compute privacy preserving data mining in a cloudcomputing environment,'' in *Proc. Eur., Medit., Middle Eastern Conf. Inf.Syst. (EMCIS)*, Coimbra, Portugal, 2017, pp. 397_413.

[23] M. Waddey, P. Poncelet, and S. Ben Yahia, ``A novel approach for privacymining of generic basic association rules,'' in *Proc. ACM 1st Int.WorkshopPrivacy Anonymity Very Large Databases (PAVLAD)*, Hong Kong, 2009,pp. 45_52.

[24] B. Wang, Y. Zhan, and Z. Zhang, ``Cryptanalysis of a symmetric fullyhomomorphic encryption scheme,'' *IEEE Trans. Inf. Forensics Security*,vol. 13, no. 6, pp. 1460_1467, Jun. 2018.

[25] N. Domadiya and U. P. Rao, ``Privacy preserving distributed associationrule mining approach on vertically partitioned healthcare data,'' in *Proc.Procedia Comput. Sci.*, Fez, Morocco, 2019, pp. 303_312.