Volume 5 Issue I, January 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Private record matching using Secure multiparty computation protocol

S. Viswanandhne[1], M. D. Krithiga[2], P. Nandhini[3]
M.E. Scholar, CSE Department, Kumaraguru College Of Technology, Coimbatore, India

Abstract: ***Real-world entities are not always represented by the same set of features in different data sets. Therefore, matching records of the same real-world entity distributed across these data sets is a challenging task. If the data sets contain private information, the problem becomes even more difficult. Existing solutions to this problem generally follow two approaches: sanitization techniques and cryptographic techniques. A hybrid technique that combines these two approaches and enables users to trade-off between privacy, accuracy and cost. The project's main contribution is the use of a blocking phase that operates over sanitized data to filter out in a privacy-preserving manner pairs of records that do not satisfy the matching condition. This method incurs considerably lower costs than existing cryptographic techniques and yields significantly more accurate matching results compared to existing sanitization techniques, even when privacy requirements are high.***
Index Terms: Private information, security, accuracy, cost, record matching

I. INTRODUCTION

Analysis of data maintained by distinct entities is critical for all applications. For example, two businesses may wish to share data about customers with similar demographics (e.g., phone number) to increase their revenues. However, to protect their customer base, both parties want to keep data that are not part of the join result private.

A. Data mining

Data mining is a emerging field. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" cannot be extracted from it. Data mining attempts to answer this need. Data mining techniques search for interesting information without demanding a priori hypotheses. As a field, it has introduced new algorithms such as association rule learning. It has also applied known machine-learning algorithms such as inductive-rule learning (e.g., by decision trees) to the setting where very large databases are involved. Data mining techniques are used in business and research and are becoming more and more popular with time.

B. Confidentiality issues in data mining

A problem during collection of data is to maintain its confidentiality. The need for privacy is due to law or can be motivated by business interests. However, there are situations where the sharing of data can lead gain both parties. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. Addressing this issue, it can be shown that highly efficient solutions are possible.

Let P1 and P2 be parties owning large private databases D1 and D2. The parties wish to apply a data-mining algorithm to the joint database D1 and D2 without revealing any unnecessary information about their individual databases. That is, the only information learned by P1 about D2 is that which can be learned from the output of the data mining algorithm, and vice versa. No "trusted" third party is assumed who computes the joint output.

C. Privacy preserving mechanism

Aim of data mining is to construct models of real data. But the problem with data mining is that the real data is too valuable and thus difficult to get it. Thus the solution is to add privacy to those data. Hence only information that is really necessary will be published to other parties, like parties learn only average values of entries.

The goal is to match similar records that represent distinct individuals, therefore matching based on unique identifiers is not applicable. This problem is known as the record matching problem. Since record matching is a key component of data integration methodologies, it has been investigated extensively.

Two main approaches have been proposed for private matching. These are sanitization methods and cryptographic methods.

Sanitization techniques such as k-anonymization or random noise addition involve a tradeoff between accuracy and privacy.To achieve privacy Cryptographic techniques do not sacrifice accuracy. The algorithms that are used to private data are converted to binary circuits with private inputs. Then, using SMC protocols, accurate results are obtained.

SMC protocols guarantee that only the final result and any information that can be inferred from the final result and the input is revealed .These protocols have security parameters like encryption key sizes .

II. LITERATURE SURVEY

A. *Multidimensional k-anonymity*

K-Anonymity proposed as a algorithm for protecting privacy in microdata publishing, and numerous recoding "models" have been considered for achieving k-anonymity. A multidimensional model is proposed gives an extra degree of flexibility. This flexibility leads to higher-quality anonymizations, as measured both by general-purpose metrics and more specific notions of query answerability.

A number of organizations publish microdata for applications such as demographic and public health research. In order to protect privacy, known identifiers must be removed. This process should combine certain other attributes with external data to uniquely identify individuals. For example, an individual might be "re-identified" by joining the released data with another database on Age, and Sex.

The primary goal of k-anonymization is to protect the privacy of the individuals to whom the data pertains. However it is important that the released data remain as "useful" as possible. Many recoding models have been proposed for k-anonymization, and often the "quality" of the published data is dictated by the model that is used. Greedy algorithm for k-anonymization approach have advantages: The greedy algorithm is more efficient than proposed optimal k-anonymization algorithms for single-dimensional models. The time complexity of the greedy algorithm is *O (nlogn),* and in worse case for optimal algorithm is exponential . The greedy multidimensional algorithm produces higher-quality results than optimal single dimensional algorithms.

B. *Privacy preserving datamining*

The privacy preserving data mining primarily focuses on data analysis in such a way as to mitigate the risk of releasing some private . There are two distinct sets of problems in this . The first is problems of how two or more separate parties each with private data, may generate function of the combining their data without having to reveal it. The second focuses on how to determine whether the result of a computation alone constitutes an invasion of privacy, and if so how to mitigate the release.

C. *Secure multiparty computation*

For example two parties each having separate piece of private data which they would benefit from jointly analyzing. For example, the parties may be hospitals or any government agencies, who are supposed not to reveal their data. Performing such computations is the concern of a mature area in the PPDM literature called Secure Multi- party Computation" (SMC). The goal is to develop protocols consisting of local computations by individual parties, and the transmitting of messages between the parties.

Depending on the demands of the parties involved, one of several models of security may be appropriate. Perhaps the most well studied and rigorous formulation of a secure computation comes from cryptography. The idea is that the protocol should reveal no more information than would a fanciful "idealized" method in which the private data are presented to a completely trusted third party, who performs the computation and returns the results to each of the original parties. That is, to any specific party, the computation itself should reveal no more than whatever may be revealed by examining its input and output. To build a protocol for a particular computation, first make an assumption about the computational power to the parties. Then select a "security parameter so that for a particular party, to determine the others' private inputs becomes a computationally intractable problem e.g., public key encryption. The idea is that the parties gives their computation into a circuit consisting of wires and gates, then apply a protocol to evaluate it on their inputs. Details given are although for the time being, such a generic protocol is primarily of theoretical interest, since it is expensive for all but very small computations. An area of study is the construction of protocols for specific problems, which often result in faster and more practically applicable methods. A homomorphic encryption which allows parties to perform mathematical operations on each others' encrypted values.

D. *Private record linkage with bloom filters*

In many record linkage applications, identifiers have to be encrypted to preserve privacy. Therefore, a method for approximate

string comparison in private record linkage is needed. The idea is to store q-grams sets in Bloom filters derived from identifier values and compare them bitwise across databases. This gives the cryptographic features of Bloom filters . This method compares quite well to evaluating string comparison functions with plain text values of identifiers.

E. Privacy-preserving set operations

In many important applications, a collection of mutually distrustful parties must perform private computation over multisets. Each party's input to the function is his private input multiset. In order to protect these private sets, the players perform privacy-preserving computation; that is, no party learns more information about other parties' private input sets than what can be deduced from the result. By employing the mathematical properties of polynomials, a framework was built which is efficient, secure, and composable multiset operations: the union, intersection, and element reduction operations.

F. Blocking aware private record linkage

The problem of quickly matching records from two autonomous sources without revealing privacy to the other parties is considered. In particular, it focuses mainly to devise secure blocking scheme to improve the performance of record linkage significantly while being secure. Although there have been works on private record linkage, none has considered adopting the blocking framework. Blocking-aware private record linkage can perform large-scale record linkage without revealing privacy.

G. Anonymization

Before the release of public data set, to protect privacy, unique identifiers such as security numbers are removed. Sweeney shows in that this measure is not sufficient because quasi-identifier attributes can be combined with public directories to accurately identify individuals. Anonymization is one popular solution against such attacks. By generalizing the values of quasi-identifying attributes and removing complete records from the data set, this methods try to satisfy some definitions of anonymity. The well known of such definitions is k-anonymity, which requires combination of quasi-identifier values, so that an individual is indistinguishable within a group of size at least k.

H. Differential privacy

This work proves in that every privacy protection mechanism is vulnerable to some kind of background knowledge. Instead of tailoring privacy definitions against different types of background knowledge, one should reduce the risk of disclosure that arises from participation into a database. This notion is captured by the differential privacy protection mechanism, which addresses the case of statistical databases where users are allowed to ask aggregate queries. Differential privacy requires random noise to be added to each query result. The magnitude of the noise depends on the privacy parameter ϵ, and sensitivity of the query set Q. Denoting the response to query Q over data set T with Q^T, sensitivity is defined as follows:

Definition 1 (L1-sensitivity [30]). Over any two views T1, T2 such that $|T1| \times |T2|$ and T1, T2 differ in only one record, the L1-sensitivity of query set Q = {Q1, Qi} is measured as

$S_{L1}(Q) = \max_{T1,T2} \sum_{i=1}^{q} | Q_i^{T1} - Q_i^{T2} |$

Theorem 1 gives a sufficient condition for a statistical database to satisfy differential privacy: Theorem 1. Let Q be a set of queries answered by a statistical database, and denote by $S_{L1}(Q)$ the L1-sensitivity of Q. Then, differential privacy with parameter can be achieed by adding to each query result random noise X.

I. Private record matching

Record matching has been studied for more than four decades. However, few methods for private record matching have been investigated. Most studies in the field focus on private matching of string attributes (e.g., names and addresses). Now the focus is rather on numerical and categorical attributes. Closely related to this work, Al-Lawati et al, propose a secure blocking scheme to reduce costs. The approach has the disadvantage to work only for a specific comparison function. Also, as the focus is mainly on efficiency, the effectiveness of the approach has not been assessed.

Several approaches investigated the secure set intersection problem. Such methods deal with exact matching and are too expensive to be applied to large databases due to their heavy reliance on cryptography. Agrawal et al. formalize a notion of private information sharing across databases that relies on commutative encryption techniques, leading to several protocols.

However, the privacy definition considered before was limited to k-anonymity. Extending this approach to novel privacy preserving

partitioning methods. We show that any anonymization method based on generalization and suppression fits this description. The work in extends the hybrid approach to differentially private databases. Anonymization within is replaced by a partitioning step that generates sanitized views of the input data sets through aggregate querying. Even though provides a privacy definition similar to the security definitions of SMC, adherence to this definition is not inspected.

Record matching is the process of identifying record pairs, across two input data sets, that correspond to similar (or the same) real-world entities. In essence, the problem consists of building a classifier that accurately classifies pairs of records as "match" or "nonmatch." In the private record matching problem, an accurate classifier is assumed to be available. Therefore, private record matching methods focus on classifying all record pairs within the input data sets privately, accurately, and efficiently. We consider a matching scenario with three participants. These are data holder's parties A and B with the data sets T and V, respectively, and a querying party QP that provides the classifier for identification of matching record pairs. In a real-world application, A and B could be hospitals and QP a researcher trying to match patients with similar characteristics such as geographical location, age and sex.

Without loss of generality, let T and V be represented as relations. Let us also assume that these relations have the same schema, T (A_1,.. Ad) and V (A_1, . . ., Ad). If not, schemas of T and V can be matched using private schema matching techniques. Given matching thresholds $\theta_i \geq 0$ and distance functions di , $Dom(T.A_i) \times Dom(V.A_i) \rightarrow R+$, defined over domains of corresponding attributes of T and V , record matching can be expressed as a join of T and V . For $t \in T$ and $v \in V$, the join condition is a decision rule DR that returns true if $d_i(t.A_i, v.A_i) \leq 0$ for all attributes ($1 \leq i \leq d$) and false otherwise. Formally,

$$DR(t,v) = \begin{cases} \text{true}, & \text{if } 1 \leq i \leq d;\ d_i(t.A_i, v.A_i) \leq \theta_i \\ \text{False}, & \text{otherwise} \end{cases}$$

Our task is to identify decision rule in a privacy preserving manner such that the result will be available to the querying party QP and private records of the data holders that do not satisfy the join condition are not disclosed.

Private matching is a challenging problem, as in many cases uniquely identifying data may not be available, and matching is performed based on attributes like age, sex, etc. Furthermore, such information may not always be completely consistent across data sets e.g., the weight of a patient may be different between two admissions to different hospitals. Therefore, it is important to devise methods that are capable of privately matching records through a distance-based condition, rather than simple equi-joins computed using cryptographic hashes.

Two main approaches have been proposed for private matching.

Sanitization methods

Cryptographic methods

Sanitization methods that perturb private information to obscure individual identity cryptographic methods that rely on Secure Multi-party Computation (SMC) protocols.

There are many limitations in the existing methods:

the cost of each individual operation is very high. Consequently, no techniques are able to provide a solution addressing all relevant application requirements with respect to privacy, accuracy and cost.

Sanitization techniques include *k- anonymization* and *random noise addition* involve a trade-off between accuracy and privacy. Consequently this provides less accurate results.

Cryptographic techniques do not sacrifice accuracy to achieve privacy.

III. SYSTEM DESCRIPTION

A. *Hybrid approach*

A novel method is proposed to address private record matching by combining cryptographic and sanitization techniques. This work is the first systematic approach in this direction. The three participants assumed in this method are the two data holders, with the data sets to be matched, and the querying party, who provides the matching condition, also called the "decision rule".

The proposed private matching technique consists of three phases:

Partitioning: Each data holder independently partitions its records according some privacy-preserving mechanism e.g., k-anonymization, ∈-differential privacy. The result is a set of smaller partitions .

Blocking: All pairs of partitions from the data holders are input to a blocking decision rule. By looking at the regions covered by the partitions, the blocking decision rule outputs either match, non-match, or unknown. Only records within pairs of partitions labelled unknown are input to the costly SMC step

SMC: Pairs of records that are still not labelled are matched using cryptographic protocols. Matching record pairs are then added to

Volume 5 Issue I, January 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

the result.

If the input data sets are too large, it is required to label significant amounts of record pairs using cryptographic techniques. Since cost of the private record matching process is not known in advance, data holder parties might be not willing to participate. That is why limiting the costs of cryptographic techniques is considered.

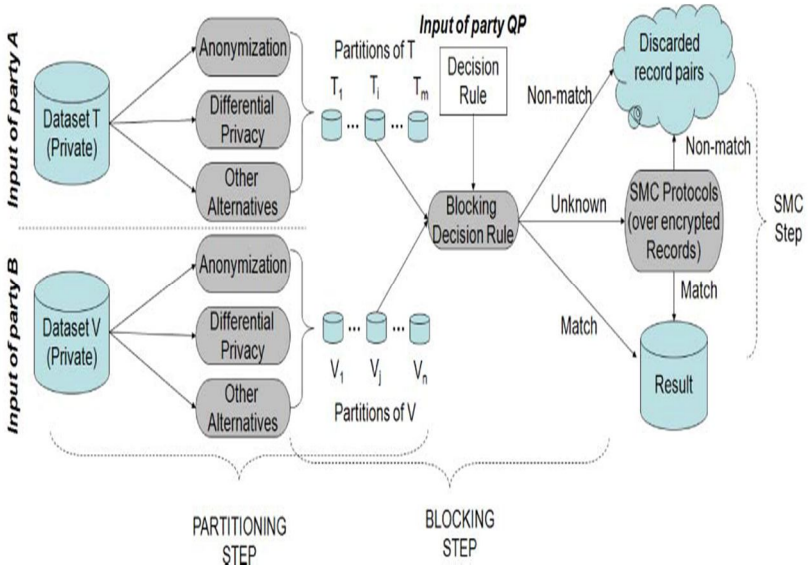

Fig. 1 Overview of the Hybrid Model

When the upper bound imposed on SMC costs is too low, some record pairs might remain unlabelled at the end of SMC. In order not to reveal irrelevant pairs, they are labelled as non matches. This precaution degrades recall since some of those unlabelled record pairs might actually be matching. Fortunately, based on the sanitized views output at the end of the partitioning step, pairs that are more likely to match can be given priority during the SMC step.

The hybrid approach combines sanitization methods with cryptographic methods in three steps. The first step, partitioning, produces sanitized views of the input data sets through perturbation. The second step of the hybrid approach is the blocking step, where pairs of partitions produced in the partitioning step are compared against one another based on the regions covered by each partition. The third step, namely the SMC step, labels any pairs of records that were not classified as match or nonmatch in the blocking step.

B. Partitioning step

A partition p consists of a set of points, Points (p) and a d-dimensional hyper-rectangle Region(p) such that for all t ϵ Points(p) $\Rightarrow$ t ϵ Region(p). In other words, every point in Points (p) should be contained by the region of partition p. The interval covered by a region r on dimension A_i is denoted as $[x_i, y_i]$, where x_i is the lower bound on attribute A_i and y_i is the upper bound.Given data set D, a partitioning algorithm outputs a set of partitions $P^D = \{p_1,\ldots p_k\}$.

C. Blocking step

Given two regions R1 and R_2, let $d_i^{inf}(R_1, R2)$ denote the infimum distance between any pair of records within R_1 and R_2 over the i^{th} dimension. Formally,

$d_i^{inf}(R_1,R_2) = inf_{t\epsilon\, R1, v\epsilon\, R2}\ (d_i(t, v))$

By definition, $d_i^{inf}(R_1,R_2)$ is the greatest lower bound on the distance. If $d_i^{inf}(R_1, R_2) > \theta_i$ for some $1 \leq i \leq d$, then no two points from R1 and R2 can match. The supremum distance is defined similarly as:

$d_i^{sup}(R_1,R_2) = sup_{t\epsilon\, R1, v\epsilon\, R2}\ (d_i(t, v))$

By definition, $d_i^{sup}(R_1, R_2)$ limits from above the maximum distance between two arbitrary points of R_1 and R_2. If these distance values never exceed the threshold for any attribute, then all points within R1 × R2 should match.

Based on infimum and supremum distance functions, the blocking decision rule BDR (R1, R2) can be defined as

Volume 5 Issue I, January 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

$$BDR(R_1, R_2) = \left\{ \begin{array}{ll} \text{N if } \exists\ 1 \le i \le i \le d,\ d_i^{inf}(R_1, R_2) > \theta_i \\ \text{M if } \forall\ 1 \le i \le i \le d,\ d_i^{sup}(R_1, R_2) \le \theta_i \\ \text{U} \qquad\qquad \text{otherwise} \end{array} \right\}.$$

Here, the return values M, N and U refer to match, nonmatch, and unknown, respectively. Not all pairs of regions can be classified as M and N. Whenever an accurate decision cannot be drawn, the pair is labelled U. Records in such regions will be labelled privately by SMC protocols.

D. Overall protocol for blocking

Let $\{T_i\}_{1 \le i \le m}$ (respectively $\{V_j\}_{1 \le j \le n}$) be the set of partitions extracted from data set T (respectively V). Algorithm describes the overall protocol for the blocking step. For every partition $\{T_i\}_{1 \le i \le m}$ of T and $\{V_j\}_{1 \le j \le n}$ of V, the blocking decision rule BDR is evaluated. In step 3, record pairs that will be labelled with SMC protocols are identified. Step 6 inserts matching record pairs to the result set.

Protocol for the blocking step
Require: T = $\{T_i\}_{1<i<m}$ U T and V = $\{V_j\}_{1<j<n}$ U V
for all Partitions Ti 2 T do
for all Partitions Vj 2 V do
if BDR (Region (T_i), Region(V_j))= U then
Privately match Points (T_i) _ Points (V_j)
else if BDR(Region(T_i), Region(V_j))= M then
Add Points(T_i) _ Points(V_j) to the result
end if, end for,end for
Assuming that step 6 only marks the pair (T_i,V_j) as M and that step 4 is performed in the SMC step, Algorithm terminates in O(m × n) time.

E. SMC step

Considering each partition as a small data set by itself, any existing solution for privacy preserving record matching can be applied to match the set of non-blocked partition pairs. In classical SMC protocols, using some cryptographic assumptions, it can be proven that only the final results and anything that could be inferred by looking at the final results are revealed. This method provides security guarantees which are slightly different from the security guarantees provided by the generic SMC protocols. Implicitly it is assumed that disclosure of the output of our privacy preserving partitioning algorithms does not violate privacy. This is reflected in the privacy definition , where the goal is to reveal only the final record matching result, the privacy preserving partitioning of the data sets and anything that can be inferred from the result and the partitioned data sets. Since the blocking step only depends on pairs of partitions, it satisfies the goal stated above. In other words, anything revealed during the blocking step could be inferred from the partitioned data sets.

1) *Basic SMC protocol for record matching:* For each pair of records that is not blocked, there is a need to securely learn whether such a pair actually matches or not. In other words, for each possibly matching record pair and for each attribute, we need to securely calculate whether $d_i(t.A_i, v.A_i) \le \theta_i$ is satisfied. Such a secure calculation is possible using generic SMC circuit evaluation techniques. Also recently many protocols have been proposed using special encryption functions such as commutative encryption and homomorphic encryption. Either these protocols, or any other SMC technique that can securely compute d(t, v) could be used in the SMC step.
2) *Limited SMC budget:* Efficiency of a blocking scheme is measured by the reduction ratio (RR) metric. Given a baseline comparison space S, reduction ratio is the fraction of savings from the comparison space attained by the blocking scheme. The results are compared to the benchmark solution that privately evaluates DR over all record pairs in the Cartesian product T × V .Therefore, $|S| = |T \times V| = |T| \times |V|$.

hen, RR = 1 - $\frac{\text{number of secure decision rule evaluations}}{|T| \times |V|}$

When the input data sets T and V are large, even after considerable reduction in comparison space, the cost of applying our

solutions might be higher than the amount anticipated by the participants. In order to prevent high costs from hampering the record matching process, an extension to the methods are discussed where participants can determine SMC budget.

Similar to RR, we represent SMC budget as a fraction of the Cartesian product size:

$$\text{SMC Budget} = \frac{\text{max. number of secure decision rule evaluations}}{|T| \times |V|}$$

The number of record pairs that were not labelled after the blocking step, hence must be labelled in the SMC step, is $(1 - RR) \times |T| \times |V|$. If $1 - RR \leq$ SMC budget, then there is no challenge in enforcing the limits over SMC operations because the budget meets or exceeds the need. However, when $1 - RR >$ SMC budget, then some record pairs cannot be properly labelled.

In order to prevent disclosure of irrelevant record pairs, we assume that all such records are excluded from the result set (i.e., assumed to be non-matching record pairs). Whenever SMC budget is insufficient, record pairs should be chosen carefully to maximize the number of matching record pairs found in the SMC step. This notion is captured by the recall measure. Let H be some heuristic that guides us in selecting the record pairs toward which the SMC budget is spent. Then, the recall of H, denoted $Recall_H$, is the fraction of matching record pairs that H can identify in the SMC step. Formally, denoting matching record pairs by the set n_M, the recall of heuristic H is

$$Recall_H = \frac{\text{number of matching pairs found by H}}{|n_M|}$$

A naive approach to enforce SMC budget would be choosing a random subset of unlabelled record pairs. Yet, it makes more sense to use the information contained in partition regions. Below we discuss various heuristics that help identify possibly matching record pairs. Among these, the heuristic that has the maximum recall should be favoured. Selection Heuristics Our heuristics rank pairs of partitions. In the SMC step, pairs are processed according to these ranks. If SMC budget is low, then low-ranked pairs may be excluded and automatically labelled as "non-match". The heuristics are outlined below. An empirical evaluation of these heuristics is provided.

a) *Minimum comparison cost first:* In this heuristic, partitions of data set T are sorted with respect to the number of secure DR evaluations required to find all matching records of V . Then, the partitions are processed in ascending order. The idea is maximizing the fraction of records of T that are matched against V. H1- would be advantageous if the partitions were weighted based on some criteria. For example, partitions that contain individuals of a certain age group may be given priority over others.
b) *Minimum volume partition first:* In this heuristic, partitions p of T are sorted with respect to the volume of their regions, Region(p). Then, partitions are processed in ascending order. Considering records as random variables supported over their partition regions, this heuristic assumes that lower volumes imply less uncertainty in estimating the actual value of a record. Based on this idea, partitions with the smallest region are processed first.
c) *Partition pair*(p1, p2) with maximum Region(p1) ∩ Region(p2) volume first. This heuristic assumes the volume of the intersection between partition regions is an accurate indicator of possibly matching records. Therefore, pairs are ordered based on normalized intersection volumes and processed in descending order.

F. Privacy definition

Privacy guarantees of SMC techniques can be proven under reasonable assumptions. We believe that a similar theoretical framework is needed for our hybrid approach. To this end, we extend the basic definitions and techniques used in SMC so that they apply to our hybrid framework. It is focussed on security/privacy definitions of the semihonest model, where each party reveals some sanitized information about its data.

In the semihonest model a computation is secure if a party's view during protocol execution can be effectively simulated based on its input and output. This does not imply that all private information is protected. Under this definition, disclosure of any information that can be deduced from the final result is not a violation.

We extend the basic model by including sanitized data in the form of anonymized data sets that satisfy differential privacy definitions. We assume that such sanitized data are public and can be accessible by all participants. Formally, let $\bar{a} = (a_1, \ldots, a_z)$ be the sanitized data .

Let $f : (\{0, 1\}^*)^z \rightarrow (\{0, 1\}^*)^z$ be a probabilistic, polynomial-time functionality, where $f_i(x_1, x_2, \ldots, x_z)$ denotes the ith component of $f(x_1, x_2, \ldots, x_z)$ and let $\prod$ be a z-party protocol for computing f. For $I = \{i_1, i_2, \ldots, i_t) \subseteq [z]$ where [z] denotes the set {1, 2, . . . , z}, we let $f_I(x_1, x_2, \ldots, x_z)$ denote the subsequence $f_{i1}(x_1, x_2, \ldots, x_z), f_{i2}(x_1, x_2, \ldots, x_z), \ldots, f_{it}(x_1, x_2, \ldots,$

x_z). Let the view of the ith party during an execution of protocol $\prod$ on $x = x_1, x_2, \ldots, x_z$, denoted by view$_i^{\prod}(x)$, be $(x_i, r_i, m_i^1, \ldots; m_i^t)$ where $r_i$represents the result of the ith party's internal coin tosses, and m_i^jrepresents the jth message received by third party. Also, given $I = i_1, i_2, \ldots, i_t$, we let view$_I^{\prod}(x)$ denote the subsequence (I, view$_{i1}^{\prod}(x)$. . . view$_{it}^{\prod}(x)$).

In this context, three parties want to compute the record matching function f(T,V ,DR) where data set T (respectively V) is the input of the first party (respectively the second party) and DR is the input of the QP. Also, it is defined as $f_1(T,V,DR) = f_2(T,V,DR) = \phi$ and $f_3(T,V,DR)$= decision rule(V)(i.e., the set of matched records). In addition, let _a be the union of the sanitized data released during the blocking step. The protocol privately computes record matching function f(T,V ,DR)if the above holds. Compared to the existing privacy definitions in the semihonest model, we assume that all sanitized data (e.g., anonymized data or differentially private statistical query results) are available to any coalition of parties. The objective of the privacy preserving protocol is to reveal nothing more than what can be inferred by all sanitized information, original inputs of the colluding parties and the final function result (here, the set of matching record pairs). In contrast to classic SMC models, the hybrid model can trade off privacy versus efficiency easily. If no sanitized data is revealed (i.e., ā=ϕ), this model will be equivalent to SMC models. On the other hand, by revealing sanitized data, it is possible to improve the efficiency of SMC protocols without sacrificing accuracy.

G. Advantages of proposed system

This hybrid approach has several advantages over existing methods, which can be summarized as follows:

Costs are lower than, and at worst equal to the costs of existing cryptographic techniques.

Allow participants to trade-off between accuracy, privacy, and costs.

This method can be applied to any privacy preserving algorithm for partitioning a data set and any cryptographic technique for matching private information.

IV. CONCLUSION AND FUTURE WORK

In this work, a novel approach is proposed that combines sanitization methods and cryptographic methods to solve the private record matching problem. Our method allows participants to trade-off between accuracy, privacy, and costs. Empirical analysis of the proposed methods performed on real-world data indicates that the hybrid approach attains significant savings in costs even at considerably high levels of privacy protection. Thus the hybrid approach allows us to compare two different datasets that includes alphanumeric characters.

A promising area of future research might be extending the idea of hybrid approaches to other privacy preserving data mining tasks. It is believed that the hybrid approach could provide substantial performance improvements for privacy preserving distributed data mining protocols.

REFERENCES

[1] A.C. Yao, "Protocols for Secure Computation," Proc. IEEE Symp.Foundations of Computer Science (CS), pp. 160-164, 1982.
[2] Bipin Joshi, Paul Dickinson, Fabio ClaudoFerracchiati, Wrox.Press, "Professional ADO.NET Programming"
[3] C. Dwork, "Differential Privacy," Proc. Int'l Colloquium Automata, Languages and Programming (ICALP '02), pp. 1-12, 2006.C.
[4] Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. Third Theory Computing Conf. (TCC), pp. 265-284, 2006.
[5] Chris Goode, John Kauffman Microsoft C#.NET Programmer's book (Tata McGraw Hill Edition),2002.
[6] James R. Groff and Paul N. Weinberg, Osborne/McGraw-Hill © 1999, "SQL: The Complete Reference".
[7] K. Le Fevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. 22nd Int'l Conf. Data Eng.
[8] N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE '07), pp. 106-115, 2007.
[9] O. Goldreich, "General Cryptographic Protocols," The Foundationsof Cryptography, vol. 2, Cambridge Univ. Press, 2004.