LAYERED PRIVACY PRESERVING APPROACH TO PRIVATE RECORD MATCHING

V.C. Mini Misha*

S. Muthu Kumar **

Abstract—Record matching represent similar records based on unique identifiers is tedious; this problem is known as record matching problem. Existing solutions to this problem are sanitization techniques and cryptographic techniques. In this paper, a new hybrid technique is used that combines these two approaches; but, operates over a new technique called secured authentication model which help the user's to trade off between privacy, accuracy, and cost. The secured authentication model called Trust based authentication model, which is used to classify the user and allow only the authenticated user for record matching. This will make the system be more secured. In this approach, the sanitized data is given to the blocking phase to remove the unwanted records in privacy preserving manner. Also prove that the participants of this protocol only learn about their results. This method is cheaper, and more accurate compared to cryptographic, and sanitization techniques, even when the privacy requirements are also high.

***Keywords*:-** Anonymization, Authentication, Privacy, Record Matching, Security.

* PG Scholar

** AP, Department of Information Technology, PSN College Of Engineering And Technology, Melathediyoor, Tirunelveli, Tamilnadu, India.

I. INTRODUCTION

Analyzing the information maintained by any distinct entities is difficult. For example, if any two competitors wish to match their information, without giving their private information is impossible and tedious task. Therefore, matching the records of similar type using the Social Security Number is not applicable. This type of problem for record matching is called as record matching problem.

Data integration technologies are the key component for private record matching. The various data integration methodologies are greatly discussed in [1]. However, the data mining techniques are introduced; privacy related concerns to sharing of individual information have reformulation of the problem; therefore, introducing the concept of private record matching. Private record matching is an challenging task, uniquely identifying the information is impossible, and matching the records are performed using attributes like age, occupation, etc. [1].

Certainly, such information not always be consistent across data sets (e.g., weight of a person may vary between two admissions to different hospitals). So, it is important for detect methods that have the ability to privately match records through a distance based condition [2], [3]. Therefore, two main approaches are used for private record matching. These are sanitization methods that perturb private information to not easily understand individual identity [4] - [8] and cryptographic methods that trust on Secure Multi-party Computation (SMC) protocols [9].

Sanitization techniques such as k-anonymization [5] and random noise addition [7], [8] involve a balanced tradeoff between accuracy and privacy. There are some cryptographic properties or features that are used to create privacy preserving protocols. These properties include additive homomorphism encryption property and commutative encryption property [10]. [17], [18] explains about the data cleaning process.

Several approaches have been proposed to address the privacy protection issues in data mining applications. Heuristics based approach that protects individual information by using data perturbation method such as blocking and generalizing in [11]. These approaches have a major drawback when dealing with privacy preserving data mining problems. They trade off between the privacy of the individual information and the correctness of the data mining problems. The problem overcoming techniques are explained in [13]-[16].

Cryptographic based is an effective way to resolve the accuracy-privacy trade-off. The data mining result is very much accurate and the privacy of personal information is leaked by no means under security constraints [9]. Therefore, the privacy is ensured and accuracy is maintained when it comes to cryptographic solutions for distributed data mining applications.

The present paper describes a new concept of private record matching. This approach combines both the sanitization technique and the cryptographic technique. Both these techniques operates over an secured model called as Trust based authentication model. This model will allow only the authenticated user for record matching for the security purposes.

The trade off to this solution is along three dimensions: privacy, cost, and accuracy. This method involves four participants namely trusted party, two data holders and the querying party.

The proposed matching technique consists of five phases namely data source creation, trust based authentication, privacy preservation, query party, and blocking.

II. GENERIC IDEAS

A. Anonymization

To protect individual privacy unique identifiers are removed. In [5] it shows that the measure is not sufficient because quasi-identifier attributes can be combined with public directories. Anonymization is one solution against these attacks. By generalizing the values of quasi-identifying attributes and removing entire records from the data set. The well known definition is k-anonymity[5] requires every combination of equivalence class value to appear atleast k times in the data set . This model has been extended by many related works in the area such as l-diversity[6], and t-closeness [12].

B. Differential privacy

In [8], every privacy protection mechanism is vulnerable to some knowledge. Instead of tailoring privacy against different types of knowledge, one should minimize the disclosure risk that arises from participation into the database. Differential privacy requires random noise to be added with each query result. The magnitude of the noise depends on the privacy parameter €, and the 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(L_1 –sensitivity):

Over any two views T_1,T_2 such that $|T_1|=|T_2|$ and T_1, T_2 differ in only one record, the L_1-sensitivity of query set $Q=\{Q_1,\ldots..,Q_q\}$ is measured as S_{L1} $(Q)=\max A_{T1,T2} \sum^{q}_{i=1}| \quad Q_i^{T1} - Q_i^{T2} |$.

III. RECORD MATCHING

.

Record matching is defined as the process of identifying record pairs, across two data sets, that corresponds to the same real-world entities. The problem arises here involves building a classifier that accurately classifies pairs of record as “match” or “nonmatch”. The classifier is assumed to be available in record matching problem. Let consider an matching scenario with three participants. These are data holders A and B with data sets T and V, and querying party QP which provides classifier for identification. Without the loss of generality, T and V be represented as relations. Assume that these relations have same schema, $T(A_1,\ldots.,A_d)$ and $V(A_1,\ldots.,A_d)$. For t € T and v € V, the join condition, which is the decision rule DR that returns true if d_i(t. A_i, v. A_i) $\leq \Theta_i$ for all attributes ($1 \leq I \leq d$) and false otherwise.

IV. MATCHING APPROACH

This technique combines both sanitization technique with cryptographic technique in five steps. The first step, Data Source Creation, which would create attribute, partition the attribute and finally, create the data source. The second step is the trust based authentication, which is used to create user information and social network. The third step is privacy preservation, which would partition the data, and provide cryptographic functions. The fourth step is the querying party, which would check the authenticated user, and allow them to view the results. The fifth step is the blocking phase, which would used to view query, and perform matching process.

A. *Data Source Creation*

A data source, also called a data file, is a simple collection of records that store data. Data site in turn used to store a enterprise database, data files including non automated data. Enterprises maintain employee records in a data warehouse because they want a repository of all related data as well as a main repository of the business organization's historical data. In this step, it performs two main operations one is create attribute, and the other is create datasource.

In attribute creation operation, attributes are created with the specified name, and returns the attribute as an attribute object. Attribute is created for the different dataholders. It includes DatasourceId, DataHoldername, Attributes.

In create datasource operation, the data source is generated for their corresponding attributes for the different data holders. It would includes DatasourceId, Datasourcename, Attributes, and AttributeValue.

B. *Trust Based Authentication*

Trust based authentication process will consider the user authentication to the social network. It performs two operations one is user information, and the other is social network. The functions are as follows:

In user information operation, user registration is one time identification procedure that enables to obtain information necessary for allowing the user to enter the matching phase. In general, any person can become a registered user by providing some credentials, usually in the form of a username and password. After that, one can access information and privileges unavailable to non-registered users simply called as guests.

In the social network process, the registered users can communicate and collaborate with external stakeholders. Social networks can enable access to and sharing of information that is essential.

C. *Privacy Preservation*

Privacy preserving technique is created after the introduction of powerful data mining techniques, privacy concerns related to sharing of individual information have pushed research towards the reformulation of the problem and the development of new solution. Privacy preservation performs two basic operations one is data partitioning, and the other is cryptographic functions.

In data partitioning, each data holder independently partitions its records according to some privacy-preserving mechanism (e.g., k-anonymization). The outcome is a set of smaller partitions whose extents are hyper-rectangles in the multidimensional attribute space.

Algorithm 1. Anonymization For Partitioning
Require: Set E of equivalence classes

```
p ← ϕ
for all Equivalence classes e € E do
    for all Quasi-identifier attributes $A_i$ € $A^{qid}$ do
        $min_i$ ← min. value over $A_i$ in records of e
        $max_i$ ← max. value over $A_i$ in records of e
    end for
    Generate a partition p that represents e
    p. Region = $[min_1,max_1]x…x[min_{|A^{qid}|}, max_{|A^{qid}|}]$
    p.Points = all records of e
    Insert p into P
end for
Return the set of partitions, P
```

In the cryptographic functions, the records are encypted using cryptographic techniques, which is used for matching private records.

Algorithm 2. Partition Perturbation in Statistical Database
Require: Set of partitions P, Laplace noise generator Lap(λ)

```
Create an empty partition $P_{sub}$ for suppressed records
Set Region( $P_{sub}$ ) to the entire domain
for all Partitions p € P do
    r ← round (Lap(λ))
    if r > 0 then
        Add r fake records to partition Points(p)
    else
        r ← r x (-1)
        for I = 1 to min ( r, |Points(p) |)  do
            Select t € Points(p) uniformly at random
            Suppress t by moving it to Points( $P_{sub}$ )
        end for
        if r > |Points ( p )| then
            Add r - |Points(p)| fake rec.s to Points ( $P_{sub}$ )
        end if
    end if
end for
Encrypt all records in all partitions
```

D. Querying Party

In this step, the authentication is checked using their credentials which is created in the trust based step. Then the user is allowed to send query and view the matching results generated in the blocking phase.

E. Blocking Phase

The various functions which are involved in blocking
phase are namely view query and, matching. The view query is used to view the user's query request. The matching process involves edit distance, SMC and, finally displays the result.

In edit distance, the partition p consists of a set of points, Points(p) and a d-dimensional hyper-rectangle Region (p) which are used to find the infimum as well as the supremum distance between any pair of records within R1 and R2 over the i^{th} dimensions. Here, the return values M, N, and U refers to match, nonmatch, and unknown records respectively.

Whenever, an accurate decision will not be drawn, or the record pair which are labeled as U. Such records in this region will be privately labeled using SMC protocols. In this paper, SMC protocol uses hash function for the cryptographic techniques. The main aim of the Secure Multiparty Computation protocol is, for the participating parties to securely perform some functions of their distributed and private inputs. One way of approaching the computation to be secured is to provide a list of security properties that should be preserved.

The first property is privacy or confidentiality. A naïve attempt at formalizing privacy would be to require that each party learns nothing about their partie's inputs, even if it behaves maliciously. Therefore, the privacy requirement is usually formalized by saying that the only information learned by the parties in the computation is specified by the output function.

Another important property is correctness, this states that the parties output is really defined by the function. If the correctness is not guaranteed, then a malicious party may be able to receive the specified decision tree while the honest party receives a tree that is modified to provide misleading information. Here, the process of deciding which security property is required must be reevaluated.

In the ideal execution, the parties all send their inputs to the trusted party. The trusted party then computes the function for these inputs and send each party only its specified output.

Algorithm 3. Blocking Protocol

```
Require: $T =\{T_i\}_{1 \leq i \leq m}$ U T and $V = \{V_j\}_{1 \leq j \leq n}$ U V
    for all Partitions $T_i$ € T do
for all Partitions $V_j$ € V do
    if BDR( Region ($T_i$) , Region ($V_j$) ) = U then
        Privately match Points ($T_i$) X Points ($V_j$)
    else if BDR (Region ($T_i$) , Region ($V_j$) ) = M then
        Add  Points ($T_i$) X Points ($V_j$) to the result
    end if
end for
end for
```

Finally, the result is displayed, and it is viewed by the authenticated query party.

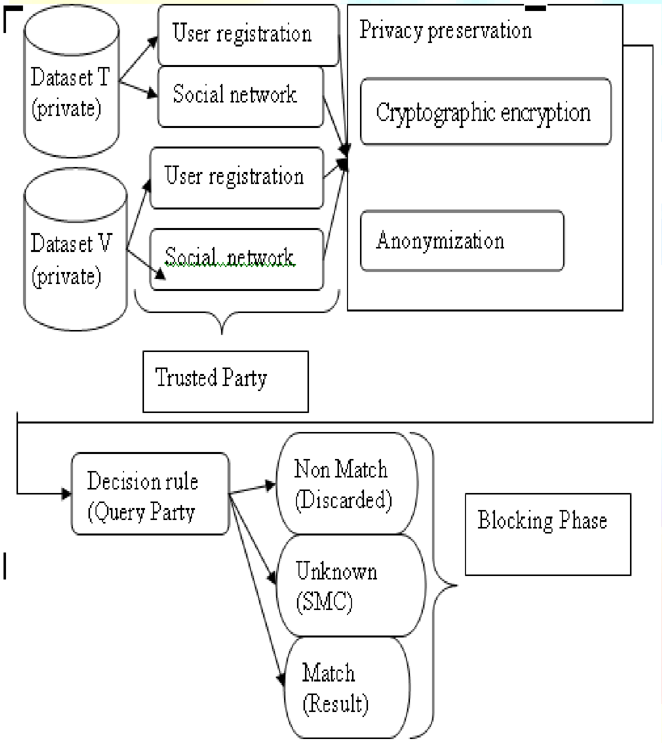

Figure 1: The Proposed System

V. SECURITY ANALYSIS

A) Data Holder Parties

The roles of data holder parties A and B are symmetric. Without loss of generality, let discuss over party A. The input of A is suppressed to data set T with an partitioning algorithm(e.g., anonymization)The view $\text{view}_A^{\prod}$ of A contains the messages A sends and receives during execution of the protocol $\prod$:

- Statistical queries A receives from QP.
- Noisy responses to these queries.
- Encrypted records sent to QP at the end of the partitioning.

The output $\text{output}_A^{\prod}$ of A is an empty set and it is independent of $\text{view}_A^{\prod}$ the formal simulation is skipped.

B) *Trusted Party*

Since, it allows only the authenticated user's,it prevents the confidentiality of the records and does not allow the unauthenticated user for record matching, to improve privacy preservation.

C) *Querying Party QP*

The QP provides the input decision rule, DR(t,v). The output is represented as a set of matching record pairs:

$M = T_{DR(t,v)} V$.

The following constitutes $\text{view}_{QP}^{\prod}$:

- Statistical queries issued to A and B during partitioning and the noisy responses received from A and B.
- Partitioning of T named as regions and the encrypted records belonging to each partition.
- Partitioning of V, is similar as T.

VI. RESULTS AND DISCUSSION

In this section, the proposed record matching technique is analyzed. The first step in this process is attribute and datasource creation. In this step, the data are converted as the useful ones after hiding the sensitive information. The second step is the trust based authentication, here the user's are previously registered and it allow only the authenticated user for matching process. The third step is the privacy preservation, in which the privacy of the data are provide by using anonymization, cryptographic techniques and partition perturbation. The fourth step is the querying party, where the authenticated user's query is given to the blocking phase for record matching. Finally the blocking phase, here the records are matched using the query and displays the result either using the edit distance or the SMC protocol.

VII. CONCLUSION

In this paper, a new record matching technique has been proposed. The idea relies on secured trust based authentication model operates over by combining both the sanitization as well as the cryptographic functions. While combining both the techniques, it provides lower cost compared with cryptographic technique, record pairs marked as nonmatch by edit distance process are 100% against disclosure, improves the privacy, efficiency and accuracy of the matching process since this method is applies to any privacy preserving mechanism for partitioning datasource and cryptographic technique for matching records.

REFERENCES

[1]A.K. Elmagarmid, P.G. Iprirotis, and V.S. Verykios,"Duplicate Record Detection: A Survey," IEEE Trans. Knowledge and Database Eng., vol. 19, no. 1,pp. 1-16, Jan. 2007.

[2]C. Clifton, M. Kantarciolu, A. Doan, G. Schadow, J.Vaidya, A. Elmagarmid, and D. Suciu, "Privacy-Preserving Data Integration and Sharing," Proc. Ninth ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD '04), pp. 19-26, 2004.

[3]R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private Databases," Proc. ACM SIGMOD Int'l Conf.Management of Data, PP. 86-97, 2003.

[4]B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation, "proc. 21st Int'l Conf. Data Eng. (ICDE '05), pp. 205-216, 2005.

[5]L. Sweeney, "k-Anonymity: A Model for Promoting Privacy," Int'l J.Uncertainity, Fuzziness and knowledge-Based Systems, vol. 10, no.5, pp. 557-570, 2002.

[6]A. Machanavajjhala, J. Gehrke, d. Kifer, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," proc. 22nd Int'l Conf. data Eng. (ICDE '06), p.24, 2006.

[7]R. Agrawal and R.Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 439-450, 2000.

[8]C. Dwork, "Differential Privacy," Proc. Int'l Colloquium Automata, Languages and Programming (ICALP '02), pp. 1-12,2006.

[9]O. Goldreich, "Gereral Cryptographic Protocols,"The Foundations of Cryptography, vol. 2, Cambridge univ. Press, 2004.

[10] M.J. Freedman, K.Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," Proc. Eurocrypt, 2004.

[11] X. Xiao and Y. Tao, "Output Perturbation with Query Relaxation," Proc. VLDB Endoement, vol. 1, no. 1, pp. 857-860, 2008.

[12] 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.

[13] F. Emehci, D. Agrawal, A.E. Abbadi, and A. Gulbeden, “Privacy Preserving Query Processing Using Third Parties,” Proc. 22nd Int’l Conf. Data Eng. (ICDE ’06), 2006.

[14] A. Inan, M. Kantarcioglu, E. Bertino, and M. Scannapieco, “A Hybrid Approach to Private Record Linkage,” Proc. IEEE 24th Int’l Conf. Data Eng. (ICDE ’08), pp. 496-505, 2008.

[15] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, “Private Record Matching Using Differential Privacy,” Proc. 13th Int’l Conf. Extending Database Technology (EDBT ’10), pp. 123-134, 2010.

[16] C. Clifton, M. Kantarcoglu, X. Lin, J. Vaidya, and M. Zhu, “Tools for Privacy Preserving Distributed Data Mining,” SIGKDD Explorations, vol. 4, no. 2, pp. 28-34, Jan 2003.

[17] M.G. Elfeky, A.K. Elmagarmid, and V.S. Verykios, “TAILOR: A Record Linkage Tool Box,” Proc. 18th Int’l Conf. Data Eng. (ICDE ’02), pp. 17-28, 2002.

[18] R. Schnell, T. Bachteler, and J. Reiher, “Privacy-Preserving Recod Linkage Using Bloom Filters,” BMC Medical Informatics and Decision Making, vol. 9, no. 1, p. 41, Aug. 2009.