# Privacy Preserving Issues In Datamining

Aasif Hasan*

***Abstract***

*This paper talks about threats to privacy that can occur through data mining and then view the privacy problem as a variation of the inference problem in databases and paper addresses the issue of privacy preserving data mining. Specifically, it considers a scenario in which two parties owning confidential databases wish to run a data-mining algorithm on the union of their databases, without revealing any unnecessary information. The above problem is a specific example of secure multi-party computation and assuch, can be solved using known generic protocols. Privacy preserving mining of distributed data has numerous applications. Paper suggests that the solution to this is a tool-kit of components that can becombined for specific privacy preserving data mining applications. Research in secure distributed computation, which was done as part of a larger body of research in the theory of cryptography, has achieved remarkable results. It was shown that non-trusting parties could jointly compute functions of their different inputs while ensuring that no party learns anything but the defined output of the function.*

**1. Overview:**

1. **What is Data Mining?** Extracting implicit un-obvious patterns and relationships from a warehoused of data sets.
2. **This information can be useful to increase the efficiency of the organization and aids future plans.**
3. **Can be done at an organizational level?**
   By Establishing a data Warehouse
4. **Can be done also at a global Scale?**
   By using Distributed Data Mining
5. **Defining Privacy?**

The common definition of privacy in the cryptographic community limits the information that is leaked by thedistributed computation to be the information that can be learned from the designated output of the computation. Although there are several variants of the definition of privacy, for the purpose of this discussion we use the definition that compares the result of the actual computation to that of an "ideal" computation: Consider first a party that is involved in the actual computation of a function (e.g. a data mining algorithm). Consider also an "ideal scenario", where in addition to the original parties there is also a "trusted party" who does not deviate from the behavior that we prescribe for him, and does not attempt to cheat. In the ideal scenario all parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties.

**6. What is Privacy Preserving?**

Consider a scenario in which two or more parties owning confidential databases wish to run a data mining algorithm on the union of their databases without revealing any unnecessary information. **eg:-** Consider separate medical institutions that wish to

*Department of Computer Science, Kailash Chand Bansal College of Technology, Bhopal

conduct a joint research while preserving the privacy of their patients. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. Note that we consider here a distributed computing scenario, rather than a scenario where all data is gathered in a central server, which then runs the algorithm against all data. (The central server scenario introduces interesting privacy issues, too, but they are outside the scope of this paper.)

**7. How much Privacy?**

It is obvious that if a data-mining algorithm is run against the union of the databases, and its output becomes known to one or more of the parties, it reveals something about the contents of the other databases. (For example, if a researcher from a medical institution learns that the overall percentage of patients that have a certain symptom is 50%, while he knows that this percentage in his population of patients is 40%, then he also learns that more than 50% of the patients of the other institutions have this symptom.) This leak of information is inevitable, however, if the parties need to learn this output.

**8. Adversarial behavior**

Privacy preserving protocols are designed in order to preserve privacy even in the presence of adversarial participants that attempt to gather information about the inputs of their peers. There are, however, different levels of adversarial behavior. Cryptographic research typically considers two types of adversaries: A semi honest adversary (also known as a passive, or honest but curious adversary) is a party that correctly follows the protocol specification, yet attempts to learn additional information by analyzing the messages received during the protocol execution. On the other hand, a malicious adversary may arbitrarily deviate from the protocol specification.

## 2. Motivation

**Challenges**

1. Privacy Concerns
2. Proprietary information disclosure
3. Concerns about Association breaches
4. Misuse of mining
5. These Concerns provide the motivation for privacy preserving data mining solutions.

**Advantages of Privacy Preservation**

1. Protection of personal information
2. Protection of proprietary or sensitive information
3. Enables collaboration between different data owners (since they may be more willing or able to collaborate if they need not reveal their information)
4. Compliance with legislative policies

**Approaches to Preserve Privacy**

1. Restrict Access to data (Protect Individual records)
2. Protect both the data and its source:

2.1 Secure Multi-party computation (SMC)

2.2 Input Data Randomization
3. There is no such one solution that fits all purposes

**Tools for Privacy Preserving Data Mining**

There are two kind of approaches used in Data Mining for Preserving data.
1. Centralized
2. Distributed

**Privacy Preserving Central Data Mining**

The first approach assumes the data is Centralized. Data mining has operated on a data-warehousing model of gathering all data into a central site, then running an algorithm against that data. Privacy considerations may prevent this approach (as show in figure). For example, the Centers for Disease Control may want to use data mining to identify trends and patterns in disease outbreaks, such as understanding and predicting the progression of a flu epidemic. Insurance companies have considerable data that would be useful but are unwilling to disclose this due to patient privacy concerns. An alternative is to have each of the insurance companies provide some sort of statistics on their data that cannot be traced to individual patients, but can be used to identify the trends and patterns of interest to the CDC.

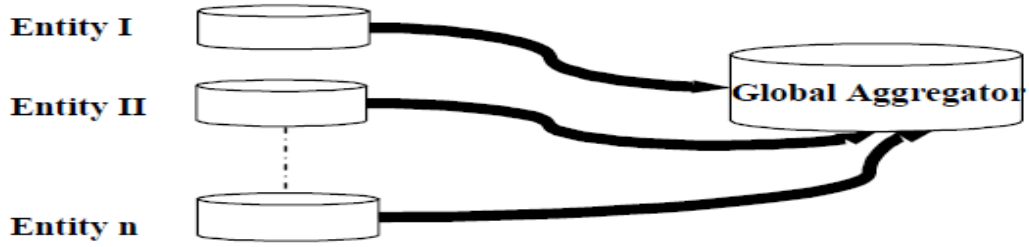


**Fig** : Centralized Data Mining Architecture

**Privacy Preserving Distributed Data Mining:** The second approach assumes the data is distributed between two or more sites, and these sites cooperate to learn the global data mining results without revealing the data at their individual sites, with a method that enabled two parties to build a decision tree without either party learning anything about the other party's data, except what might be revealed through the final decision tree. In the distribution approach suggests an answer: build a tool-kit of privacy-preserving distributed computation techniques, that can be assembled to solve specific real-world problems. If such component assembly can be simplified to the point where it qualifies as development rather than research, practical use of privacy-preserving distributed data mining will become widely feasible.

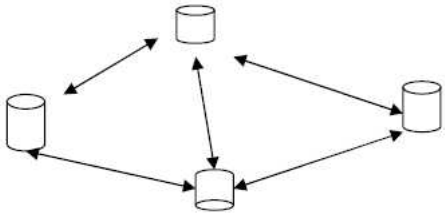

**Fig** : Distributed Data Mining Architecture

We will discuss a formalism that enables us to capture and analyze what is meant by privacy preserving distributed data mining.

**Secure Multiparty Computation:** The basic idea of Secure Multiparty Computation is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. One way to view this is to imagine a trusted third party - everyone gives their input to the trusted party, who performs the computation and sends the results to the participants. Now imagine that we can achieve the same result without having a trusted party. Obviously, some communication between the parties is required for any interesting computation - how do we ensure that this communication doesn't disclose anything? The answer is to allow non determinism in the exact values sent in the intermediate communication (e.g., encrypt with a randomly chosen key), and demonstrate that a party with just its own input and the result can generate a predicted" intermediate computation that is as likely as the actual values. However the general method given does not scale well to data mining sized problems.

**3. Techniques:** We present here two efficient methods for privacy-preserving computations that can be used to support data mining. Not all are truly secure multiparty computations -in some; information other than the results is revealed -but all do have provable bounds on the information released. In addition, they are *efficient*: the communication and computation cost is not significantly increased through addition of the privacy preserving component.

**4. Cryptographic Results: Secure Function Evaluation:** We describe here results of a body of cryptographic research that shows how separate parties can jointly compute any function of their inputs, without revealing any other information. As we argued above, these results achieve maximal privacy that hides all information except for the designated output of the function. This body of research attempts to model the world in a way which is both realistic and general. While there are some aspects of the “real world” that are not modeled by this research, the privacy guarantees and the generality of the results are quite remarkable.

**5. The main building block oblivious transfer:** Oblivious transfer is a basic protocol that is the main building block of secure computation. It might seem strange at first, but its role in secure computation should become clear later. (In fact, it was shown that oblivious transfer is sufficient for secure computation in the sense that given an implementation of oblivious transfer, and no other cryptographic primitive, one could construct any secure computation protocol.) The protocol involves two parties, the sender and the receiver. The sender’s input is a pair (x0, x1) and the receiver’s input is a bit $\sigma \in \{0, 1\}$. At the end of the protocol the receiver learns x_ (and nothing else) and the sender learns nothing. In other words, if we use the notation (input A, input B) => (output A, output B) to define the result of a function, then oblivious transfer is the function $((x0, x1), \sigma) \Rightarrow (\lambda, x\_)$, where $\lambda$ is the empty output.

**6. Randomization Approach:** The problem of building classification models over randomized data was addressed. Each client has a numerical attribute Xi e.g. age, and the

server wants to learn the distribution of these attributes in order to build a classification model. The clients randomize their attributes Xi by adding random distortion values ridrawn independently from a known distribution such as a uniform distribution over a segment. The server collects the values of Xi + ri and reconstructs the distribution of the Xi ’s using a version of the Expectation Maximization (EM) algorithm that provably converges to the maximum likelihood estimate of the desired original distribution. The goal is to discover association rules over randomized data. Each client has a set of items (called a transaction), e.g. product preferences, and here the server wants to determine all item sets whose support (frequency of being a subset of a transaction) is equal to or above a certain threshold. To preserve privacy, the transactions are randomized by discarding some items and inserting new items, and then are transmitted to the server. Statistical estimation of original supports and variances given randomized supports allows the server to adapt Apriori algorithm to mining item sets frequent in the non-randomized transactions by looking at only randomized ones.

**7. Problem:** Computation Overhead, number of exchanged messages *O(n*m)*And other techniques are :

- Secure Size of Set Intersection
- Scalar Product
- Multiparty cooptation
- Secure sum

**8. Applications**

We now demonstrate how the above protocols can be used to make several standard data mining algorithms into privacy preserving distributed data mining algorithms.

**i. Association rules in horizontally partitioned data:** In a horizontally partitioned database, the transactions are distributed among n sites. The global support count of an item set is the sum of all the local support counts. An item set X is globally supported if the global support count of X is bigger than s% of the total transaction database size. The global confidence of a rule X => Y can be given as {X  Y }.*sup/X.sup*. A k-item set is called a globally large k-item set if it is globally supported.

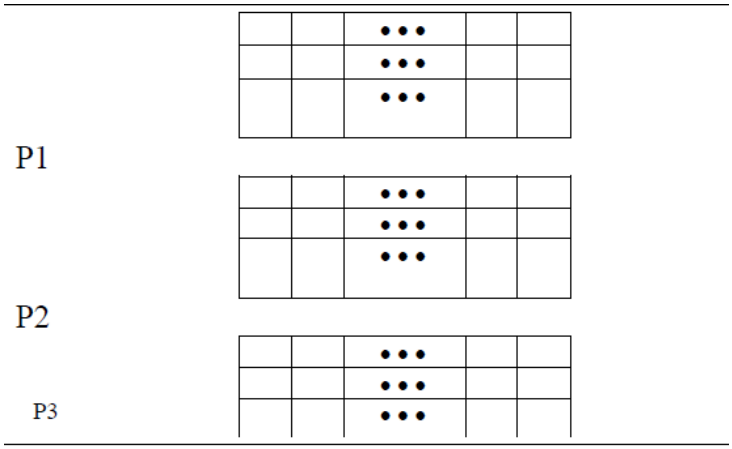


**Fig** : Horizontally partitioned

**ii. Association rules in vertically partitioned data:** Extending the existing apriori algorithm can do mining private association rules from vertically partitioned data, where the items are partitioned and each itemset is split between sites. Most steps of the apriori algorithm can be done locally at each of the sites. The crucial step involves finding the support count of an item set. If we can securely compute the support count of an item set, we can check if the support is greater than threshold, and decide whether the item set is frequent. Using this, we can easily mine association rules securely.

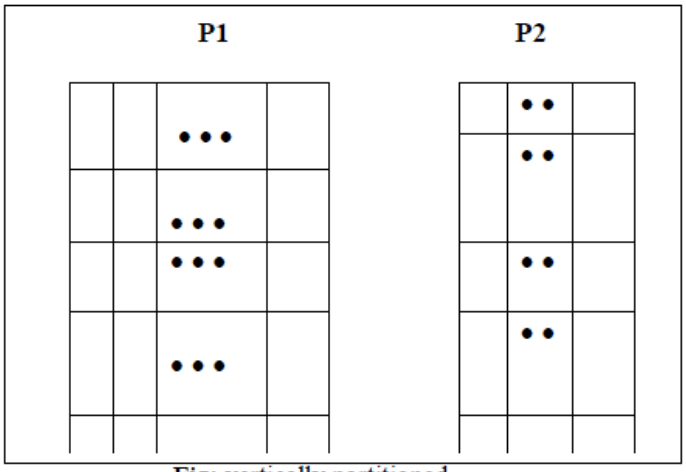


Fig: vertically partitioned

## 9. Limitation

- Scalability
- High Overhead
- Details of the trust model assumptions
    - Users are honest and follow the protocol

**Future Scope**

- Preprocessing of data for PPDM.
- Privacy-preserving data solutions that use both randomization and cryptography in order to gain some of the advantages of both.
- Policies for privacy-preserving data mining: languages, reconciliation, and enforcement.
- Incentive-compatible privacy-preserving data mining.

## 10. Concluding Remarks

1. No one solution can fit all.
2. Which area looks more promising?
3. Increasing use of computers and networks has led to a proliferation of sensitive data.
4. Without proper precautions, this data could be misused.
5. Many technologies exist for supporting proper data handling, but much work remains, and some barriers must be overcome in order for them to be deployed.

6. Cryptography is a useful component, but not the whole solution.
7. Technology, policy, and education must work together.
8. Can we create robust randomization schemes to a wide scale of applications and different distributions of data?
9. How to deal with the case of malicious participants?

## 11. References


1. Benny Pinkas, (2003) **Cryptographic techniques for Privacy Preserving Data Mining**, HP Labs, volume 4, Issue 2- p.p.12 – 19.]
2. Moheaeb Rajab, **Privacy Preserving Data Mining** – p.p. 1-20
3. R. Agrawal and R. Srikant. (1994) **Fast algorithms for mining association rules**. In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, Sept. 12-15 1994. VLDB.
4. J. S. Vaidya and C. Clifton.(2002) **Privacy preserving association rule mining in vertically partitioned data**. In the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 639-644, July 23-26.
5. Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya **Tools for Privacy Preserving Distributed Data Mining** Purdue University Department of Computer Sciences 250 N University St West Lafayette, IN 479072066 USA (clifton, kanmurat, jsvaidya)@cs.purdue.edu
6. Rebecca Wright, **Privacy-Preserving Data Mining**, Computer Science Department Stevens Institute of Technology www.cs.stevens.edu/~rwright-p.p. 1-27