An Efficient Framework for Privacy Preservation for Big Data Applications

*A **Thesis***

submitted in partial fulfillment of the requirements for the award of degree of

Doctor of Philosophy

by

Harmanjeet Kaur

(901403011)

under the guidance of

Dr. Neeraj Kumar (Professor) &

Dr. Shalini Batra (Associate Professor)

Computer Science and Engineering Department,

Thapar Institute of Engineering and Technology, Patiala, INDIA

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology,

Patiala -147004, INDIA

January 2020

Contents

List of Figures

List of Tables

List of Algorithms

Certificate

I hereby certify that the work which is being presented in this thesis entitled "**An Efficient Framework for Privacy Preservation for Big Data Applications**", in partial fulfilment of the requirements for the award of degree of **"Doctor of Philosophy"** submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology (TIET), Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Neeraj Kumar** and **Dr. Shalini Batra**, and refers other research works which are duly listed in the reference section. The matter presented in this thesis has not been submitted for the award of any other degree of this or any other institute or university.

Harmanjeet Kaur

(Harmanjeet Kaur)

Regn. No. 901403011

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Neeraj Kumar)
Associate Professor
Computer Science & Engineering Department
TIET
Patiala, 147004
Punjab, INDIA.

(Dr. Shalini Batra)
Associate Professor
Computer Science & Engineering Department
TIET
Patiala, 147004
Punjab, INDIA.

Acknowledgments

I would like to express my sincere appreciation to my supervisors, **Dr. Neeraj Kumar** and **Dr. Shalini Batra**, for being the pillar of support and encouragement throughout my research work. Their experience, strength, tenderness and willfulness, has taught me the lessons of life, which are of immense help for me to make decisions in my every endeavor.

My sincere thanks are due to **Dr. Maninder Singh**, Professor and Head, Computer Science and Engineering Department (CSED), TIET, Patiala for providing me the necessary administrative assistance in the completion of the work. I am thankful to my Ph.D. committee members, **Dr. Rinkle Rani**, Associate Professor, CSED, **Dr. Sushma Jain**, Assistant Professor, CSED, and **Dr. Kulbir Singh**, Professor, Electronics and Communication Engineering Department, TIET, Patiala for their constructive comments and regularly ensuring the progress of my research work. I am thankful to all the **faculty** and **staff** members of CSED for their support.

I offer my deepest gratitude to my **family** and **friends** whose love and affectionate blessings have been a constant source of inspiration in making my vision a reality.

The chain of gratitude would be definitely incomplete without thanking the **Almighty**, the prime mover, for inspiring and guiding me to complete this task successfully.

Patiala

January 2020

Harmanjeet Kaur

(Harmanjeet Kaur)

Abstract

In the modern data-driven world, the actual advantage of big data can be realized if data is efficiently processed and knowledge extracted from it can serve as an important component in decision making. Data mining techniques have been used to discover interesting patterns and knowledge from large datasets. Providing all the data to data miners may provide good analytics, but it can also raise many security challenges since such data can be misused by malicious users. Thus, equilibrium should be maintained between data availability and data security as one needs to secure the confidentiality of sensitive data without affecting the efficiency of applications.

Privacy preserving data mining techniques are used to extract useful information from data without compromising the security of sensitive information contained in it. Before performing any analysis on data set, it is anonymized by encryption techniques or by removing the personally identifiable information from data sets, such that the person whom the data refers will remain anonymous. The data sets used for the data mining purpose can be centralized owned by a single owner or it can be distributed among multiple parties having horizontal, vertical or arbitrary distribution. Usage of traditional cryptographic techniques for protecting the information leads to large computation and communication overheads especially, for large datasets. The anonymization techniques have less computation and communication overheads, but there is a risk of re-identification of anonymized dataset, since a large amount of data is available and by linking the different data sources with the anonymized dataset, the probability of re-identification of data is higher.

This thesis proposes a framework for privacy preserving data mining on big data. Based on the proposed framework, two application domains have been identified. The first one is privacy preserving collaborative filtering technique used for recommendation generation in the healthcare system where data is arbitrarily distributed among multiple healthcare sites.

It is an item-based collaborative filtering technique where item-item similarity is securely computed using homomorphic encryption technique and secure scalar dot product algorithm. The second is cloud-based privacy preserving collaborative filtering technique based on naive Bayesian classifier for recommendation generation on arbitrarily distributed data among multiple parties. In this technique, conditional probability is securely calculated using proposed privacy preserving conditional probability algorithm and prior probability is securely calculated using homomorphic encryption technique. Both techniques are secure and having less computation overhead as compared to the state of art privacy preserving collaborative filtering techniques. Further, k- anonymization based on neural network and support vector machine classifiers helps in the anonymization of social network data before sharing or performing any analysis on it. The proposed technique is evaluated on different parameters: Precision, Recall, F-measure, Information loss and Average path length. Through this thesis work, it can be concluded that efficient data analytics can be performed securely for both centralized and distributed data sets without much computational overheads.

Chapter 1

Introduction

The rapid technological advancements, especially in transmission, storage, and computing techniques have led to increase in the volume of data generation from various sources which include social media, internet of things, transactions, clickstreams, logs and multimedia [2]. The analysis of these massive data-sets in a secure manner has improved services and productivity across various sectors of our society which include climate, healthcare, traffic control, fraud detection and security analysis.

1.1 Big Data

In modern-day computer terminology, huge data sets with diverse and complex structures, produced at very high speed; stored, analyzed and visualized for further processes are referred as Big Data [3].

According to recent Forbes reports [4], 2.5 quintillions are created every day and 90% of total data has been created in the last two years. The Domo's sixth annual infographic report [1] has stated that the Internet population has risen to 3.8 billions from 3.4 billion in 2016 and Amazon, YouTube, Snapchat, Facebook and Netflix are the biggest users of Inter-

net. Amazon ships 1,111 packages per minute, 4.33 million videos are watched per minute on YouTube, Netflix is streaming 97,222 hrs of videos per minute and Snapchat is sharing 2,083,333 photos per minute (Figure 1.1). As seen in Figure 1.1, a huge amount of data is generated every minute and the USA alone uses 3,138,420 gigabytes of Internet data every minute.

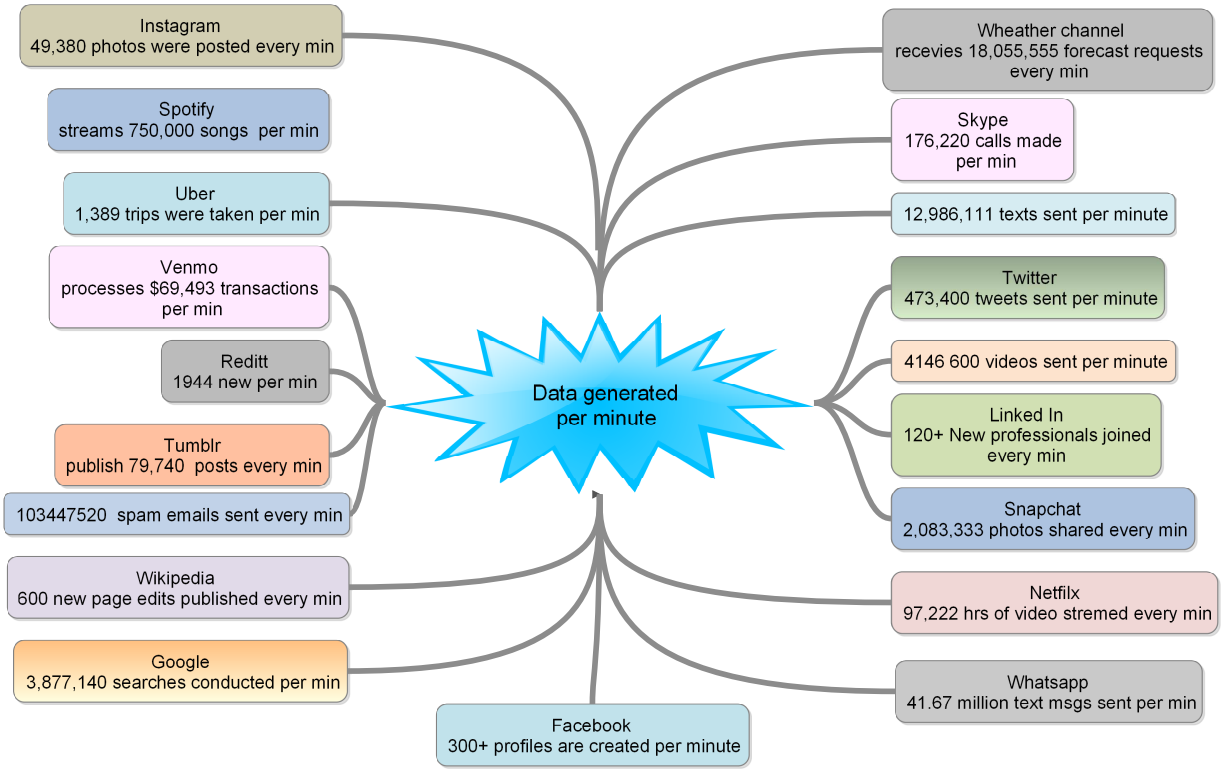


Figure 1.1: Data generation (2018) [1]

1.1.1 Big data: Definition

Big data has been defined in several ways. Laney [5] in 2001 illustrated the challenges and opportunities emerged due to the large data as a 3V model (volume, velocity, and variety). In the 3V model, *volume* refers to the quantity of data generation, *velocity* refers to the rate of data generation and *variety* represents the type of data: structured, semi-structured or unstructured. Maniyka *et al.* [6] defined big data as "the amount of data just beyond tech-

nology's capability to store, manage, and process efficiently". IDC [7] defined big data as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis". In this definition, one more V of big data has been described, *i.e.*, "*Value*". The large volume of data must be analyzed for extracting the value from it. Nowadays, many Vs have been introduced to characterize the big data. The overview of big data with the main ten V's is provided in Figure 1.2. Here, *variability* represents the data whose meaning is constantly changing. *Veracity* represents the unreliable data from some data sources, for example, the data from social networks (Facebook) [8] [9]. Unreliable data also contains useful information but there is a need of analytic tools to extract the information from this uncertain data. *Validity* refers to the accuracy and correctness of data according to its use. *Vulnerability* is the important V of big data which represent the data breach that can occur during data collection and data usage phase (data publishing, data mining, *etc*). The organizations must protect their customers' data and should be more transparent about the usage of the collected data. *Volatility* refers to the time period for the storage of data. Since the volume of the data produced is very large, there is a need to define the volatility of every data-set. Another characteristic of big data is its *visualization*. The traditional visualization tools such as spreadsheets, reports and traditional graphs cannot be used for big data due to limitations of poor scalability, functionality and response time. For developing a meaningful visualization from complex big data, tools such as tree maps, sunbursts, parallel coordinates, tree cones, *etc.*, can be used.

1.1.2 Big data mining

Big data has utility only if one extracts the knowledge and uses it for decision-making. Efficient processing and computational techniques are needed for extracting the knowledge

Figure 1.2: 10V's of Big Data

from massive data with large velocity and variety. The overall process of extracting knowledge from big data is segmented into two parts: - Data management and Data mining. Data management includes processes and techniques for efficiently acquisition, storage and retrieval of data for analysis. Data mining refers to processes and techniques used to extract the knowledge or intelligence from data.

Data management NoSQL databases are used for storing and retrieving big data which can be categorized into four parts based on their data model: *Key-value*, *Column-oriented*, *Document oriented* and *Graph data store*. In key-value data stores, data is stored as key-value pairs where each value is addressed by a distinct key. It supports the mass storage and query processing speed is better than a relational database. Column-oriented stores use the table as a data model but do not support table association. These data stores can be used for applications using large clusters to store bulk data. Most common column-oriented data

stores are Cassandra, BigTable, HyperTable, and Hbase. Document-oriented data store is of similar structure as key-value data store, but the value has semantic and it is not opaque to the system. Documents encapsulate the key-value pairs and each document has a unique key, which is used to identify the document among the collection of all documents. Commonly used document-oriented database are MongoDB, CouchDB, and Riak. Graph data stores are used to manage highly linked data-sets. These data stores are further divided into two parts: property graph data stores and RDF's framework data store. Neo4j and GraphDB are property-based graph data stores, and Sesame and Bigdata are RDF based graph stores.

Data mining A big data mining framework can be divided into three layers as [10]:-

Layer 1: *Big data mining platform*

The computer system requires access to two resources for efficient mining of data: the data and processing units. For traditional small data mining tasks, single computers having a hard disk and CPU processors are sufficient. For medium-scale data mining tasks, having large data-set that can not be fit into main memory, parallel computing and collective mining techniques can be used [11] [12]. In big data mining, since the dataset is very large, the clusters of high-performance computing nodes are required. The parallel computing programs like MapReduce and ECL are required to divide the single task into parallel components that run on computing nodes on the cluster and then merge the results received from different clusters. MapReduce is a programming paradigm for processing the large set of data distributed over cluster in parallel. The two important functions of the MapReduce paradigm are: map and reduce. The map function processes the set of data and converts it into key-value pairs, whereas the reduce function takes the output of map function as input and reduces it into a small data-set. Generally, the input and output of map and reduce tasks are stored in the

file system. The main function is called on a single master machine to pre-process the input data before map functions are called or post-process the output of reduce functions. Based upon the application, a pair of the map and reduce functions are operated once or multiple times.

Layer 2: *Big data domain knowledge and privacy*

Big data domain knowledge and privacy layer considers various aspects related to the regulations, policies, application knowledge and domain information. The two main concerns at this layer are (a) domain and application knowledge (b) data sharing and privacy. The first one provides information related to knowledge and patterns that can be identified from particular domain data and the underlying applications of a particular domain. The later provides the information regarding sensitive data that needs to be preserved before sharing the knowledge or information with other parties. Domain and application knowledge [13] provide crucial information for designing an analytic system. With the help of domain knowledge, correct features can be identified for modeling the unrevealed data (for *e. g.*, the level of T4 and T3 hormones are a better indicator of thyroid as compared to body weight). Further, it helps in the formation of achievable business objectives. For instance, in the stock market, data is generated every second in the form of buys, puts, and bids, and the mood of the market is changing continuously. The main aim of designing a prediction system for the stock market is to predict the movement of the market for the next one or two minutes. Without the proper knowledge of the domain of the stock market, it is challenging to build such a system. When the data is distributed among multiple organizations, data sharing is required to mine useful global patterns from localized data. While the intention of sharing the data is good, the concern is about the sensitive information held by different organizations such as bank transactions, medical records, *etc.* So direct exchange

of data between the parties raise privacy concerns [14]. For example, various location-based services can be provided to people after knowing their locations and their preferences, but the public revelation of data related to an individual's movement can be misused by malicious parties. To protect the privacy of the sensitive information, three main approaches are:

i) Using an access control mechanism or certification to restrict access to data so that only legitimate users can access it.

ii) Anonymizing the data, so that sensitive information is not linked to individual record. Generally, the k-anonymization privacy model is used for preserving the privacy where each record in the database is identical to the k-1 other records in terms of the quasi-identifiers. The other privacy models such as l-diversity, t-closeness and differential privacy are also used to achieve privacy. Suppression, perturbation, generalization and permutation methods are mostly used to generate the anonymized dataset. After anonymization, the data can be shared publicly without any access control mechanism.

iii) The third approach is the use of Secure Multiparty Computation (SMC) protocols like secure sum, secure log, secure comparison, secure product, *etc.*, to calculate the global function values securely from the local data of multiple parties.

Data privacy while performing the data mining has open up another research area called as privacy preserving data mining [15], where the organization having sensitive data try to perform the data mining while protecting the contained sensitive data or information in the dataset.

Layer 3: *Big data mining algorithms*

Big data mining algorithms should tackle the difficulties raised by volume, distribu-

tion, complexity and dynamic characteristics of big data. These algorithms can be divided into three stages: data fusion stage, local mining, and model fusion and global data mining. At the first stage, the sparse, heterogeneous, incomplete and uncertain data are pre-processed using data fusion techniques. In the second stage, local data mining is performed to extract local patterns. In the last stage, global knowledge is discovered from local patterns, tested and then feedback is provided to the first stage at the pre-processing level to adjust the model parameters.

1.2 Privacy Preservation in Big Data

Since big data has its potential value in healthcare, business analytics, government surveillance, and customer relationship management, it is getting attention from all the domains which include big companies, government organizations and academia [10]. Revealing all the data to data miners may provide good analytic results, but it can also raise many security issues. The value hidden in big data can also be misused by malicious users. So, a balance between big data availability and big data security needs to be maintained. The major consideration is given to 3V challenges of big data but for extracting the knowledge from data, it should be stored and mined in an efficient and secure manner. Since the data is of major importance in big data-based applications, there is a need to secure the confidentiality of sensitive data without affecting the efficiency of big data applications. Different actors involved in the knowledge discovery process (Figure 1.3) have a significant role in preserving the privacy of data.

1.2.1 Actors and their roles in preserving the privacy

In the process of knowledge discovery, the following actors can be identified:

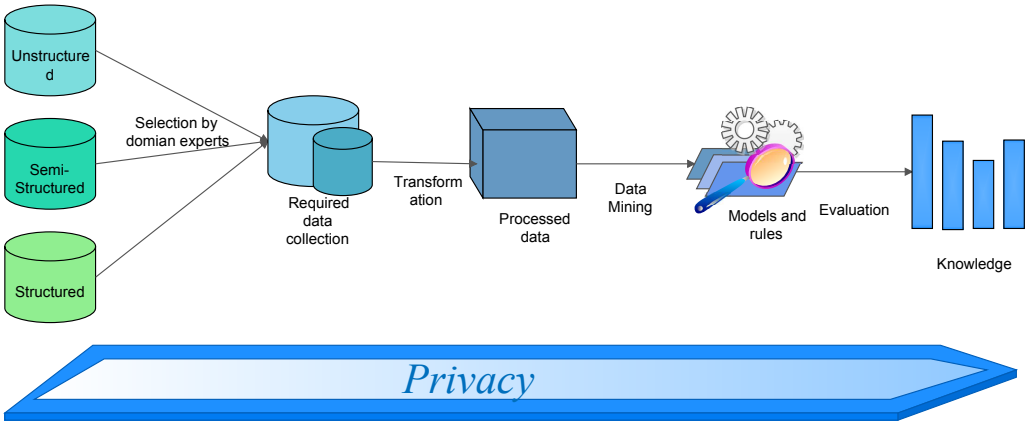


Figure 1.3: Knowledge discovery process with privacy

i) Data subjects are persons or organizations from where data is collected for analysis.

ii) The data owners/curators are the organizations or people who accumulate the data about the data subjects. They take the decision about the need of sharing the data with third parties and the privacy measures required while sharing.

iii) Data analysts perform the mining operations on the data for extracting all useful patterns from it.

iv) Data recipients are the parties who receive data mining results. If the data is published on-line, the public can be seen as the data recipient, since it is accessible to all.

Every actor should preserve the privacy of sensitive information which they own. The role of individual actors in preserving privacy are discussed in detail here:

i) Data subjects

If the data subject is individual, he possesses some data about his personal information which is also called as microdata. When the data subject provides his microdata directly to the data collector, his privacy can be compromised by an unexpected data breach of sensitive information. The preventive measures which can be taken by data subject to safeguard his sensitive information are:

(a) If microdata of the subject is very sensitive, and he does not want to provide his data, then he can decline to provide such data. He can also use the access control mechanisms to protect his sensitive data from being taken away by the data collector. While using the internet, the data subject can remove the elements of his on-line activities by deleting cookies, clearing browsers cache and usage records of applications, *etc*. Various security tools such as anti-tracking extensions, advertisement blockers and script blockers can also be used to secure the data while using the internet. The technique proposed by Mitra *et al.* [16] can also be used for protecting the users' login information in the sharing environment.

(b) In another approach, data subjects can provide his data in exchange for some benefits such as better services or monetary awards. The data subject should sign the contract with a data collector so that he will get enough compensation for any possible loss in privacy. The game theory provided in [17] can be used to formalize the contract between the data subject and data collector, where both parties want to get maximum benefits.

(c) In the third approach, if the data subject is not willing to provide his data and cannot inhibit the access to his data, he can perturb his data by using the various tools such as sockpuppets [18], MaskMe [19] or by using a fake identity to create phony information [20].

ii) Data curator

A data curator collects data from data subjects in order to perform the subsequent data mining operations which may contain the sensitive information of data subjects. Generally, the data collector uses data perturbation techniques to transform the original data in order to prevent the leakage of sensitive information about individuals. But, the modification of data can reduce the data utility. Therefore, the transformation of the

collected data should be done in such a way that modified data has sufficient utility for a desired data mining task, otherwise collecting the data will be a wasted effort. For removing the sensitive information from data records, the data curator needs to identify the types of attributes of the records. An attribute can have one of the following four types:

(a) *Explicit Identifiers* (EI) are attributes that uniquely identify a data subject from the set of records. These attributes should be removed from the database for de-identification of the data subjects from the set of records. But only the removal of explicit identifiers is not sufficient to ensure privacy.

(b) *Quasi-Identifiers* (QI) are attributes that do not necessarily disclose an individual's identity when used alone, but in combination with external databases can be used to identify the data subjects.

(c) *Sensitive Attributes* (SA) are confidential data items of the data subjects, that one does not want to disclose to other third parties.

(d) *Auxiliary Information* is non-confidential data items that bear no privacy risk and do not fit into any of the above categories.

Before publishing the data or using it for data mining purposes, explicit identifiers need to be removed and quasi-identifiers need to be modified for protecting the individual's identity and its sensitive attribute values from leakage. The quasi-identifiers are modified using various anonymization techniques. Different privacy models used for anonymization are k-anonymization, l-diversity, t-closeness, *etc*. For achieving these privacy models, various methods like generalization, suppression, anatomization, permutation, and perturbation are used.

iii Data miner: The data miner uses various data mining algorithms on collected data to

extract valuable knowledge. The two main privacy issues faced by the data miner are: first, if raw data provided to data miner contains the sensitive information of data subjects, then data breach can happen by directly observing the data. Second, by applying the various powerful data mining techniques, data miner can obtain sensitive information from the underlying data. For the first issue, it is a data curator's responsibility to remove sensitive data from the database before providing it to the data miner. For the second issue, data miners should use privacy preserving data mining techniques so that sensitive information of data subjects does not appear in data mining results. Due to the use of privacy measures, data utility can be reduced. So, data miners requires to maintain a balance between privacy and data utility.

iv Data recipient

The data miner provides the data mining results to the decision-maker (data recipient), so that he can use the provided information for business growth or achieving his objectives in the efficient ways, such as increasing business profit, hiring efficient employees, an efficient treatment plan for patients, *etc*. The two main privacy concerns of data recipients are: first, the data mining results should be provided to the data recipient directly, because if it is revealed to some other party, then a huge loss can occur to a data recipient. For this, data recipient can sign a contract with data miner for not revealing the results to any other third party without his/her consent. The second concern is the credibility of data mining results. For example, if the data recipient receives the results from the third party (transmitter) rather than directly from the data miner, he might become skeptical about the credibility of the results. For checking the credibility of the received data mining result, data recipient can use various techniques of data provenance, credibility analysis of web information, *etc*.

1.3 Privacy Preserving Data Mining

To facilitate the privacy of individual's data or information along with mining of useful information, an approach called Privacy Preserving Data Mining (PPDM) [15] has been proposed. The main aim of protecting the privacy in data mining task is to lessen the probability of misuse of data with no loss in the accuracy of results, *i.e.*, results produced in the presence of privacy preserving technique should not vary from those produced without it. The core of PPDM is to perturb the data in such a way that valuable information should be extracted without revealing the sensitive data or sensitive information contained in it. Figure 1.4 gives a general layout of PPDM techniques. The two main objectives of PPDM techniques are:

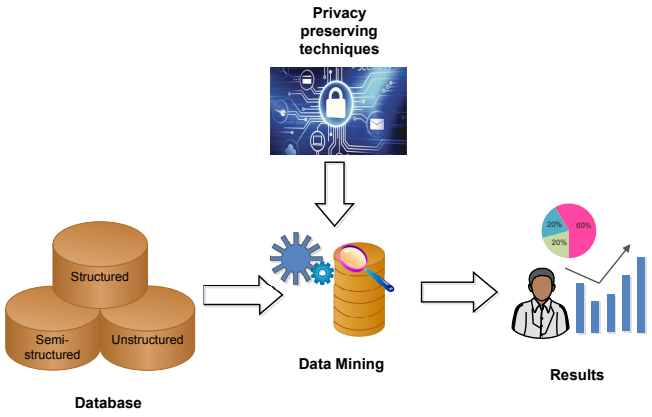


Figure 1.4: Overview of PPDM

- The sensitive raw data (for *e.g.*, individual's identification number, name, and mobile number) should not be directly used for mining.
- The sensitive mining results, whose revelation will compromise the privacy of the entity should be eliminated.

These objectives can be further divided into seven parts [21] as shown in Figure 1.5. The first objective, protection of sensitive raw data, can be divided into 5 parts. First part represents the public dataset that does not contain any sensitive data and it can be used by anyone

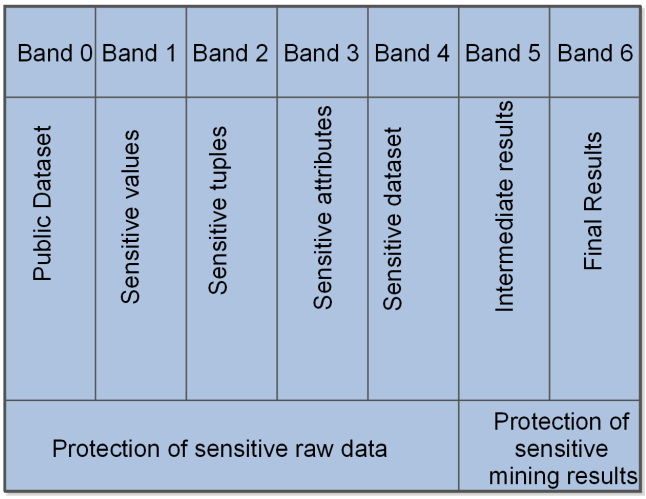


Figure 1.5: Objectives of PPDM techniques

without any changes. The second part represents the dataset which contains the few sensitive values which should be hidden before performing any analysis on it. The third part represents the dataset having sensitive records that need to be protected, for *e.g.,* the records containing information of a very important person need to be protected so that information about such person is not leaked. The fourth part represents the sensitive attributes that need to be protected, *e.g.,* the salary in a dataset. The fifth part represents the sensitive data which need to be protected. It is a hybridization of the third and fourth parts. The second objective, protection of sensitive mining results can be divided into two parts: sixth and seventh, where the sixth part represents the protection of sensitive intermediate results. During the application of PPDM techniques, some intermediate results may contain sensitive information which should be hidden. The seventh part represents the sensitive mining results which may produce after application of data mining techniques and need to be protected. For example, after the application of association rule mining techniques, some sensitive rules can get generated that need to be hidden. The PPDM techniques achieve one or all objectives depending upon the application's requirements. If traditional cryptographic techniques are used for privacy, they can have large overheads for big data. Anonymization techniques have

a risk of re-identification of anonymized dataset since a large amount of data is available and by linking the different data sources with anonymized data-set, there are more chances of retrieval of actual information.

The PPDM techniques are evaluated on different parameters such as performance, privacy, data utility, *etc*. A PPDM technique can outperform other techniques on specific parameter like privacy or performance, but the majority of the proposed PPDM techniques perform better for one parameter or other, not for all parameters.

1.3.1 Different dimensions for classification of PPDM techniques

The five dimensions for classification of PPDM techniques [22] are:

Data distribution: The datasets can be centralized having a single data owner or it can be distributed among multiple parties having multiple owners. Different PPDM techniques are required depending upon the data distribution. Some algorithms carry out the PPDM on centralized data while others work well on distributed data. The goal of distributed data mining is to compute the global model from data partitioned among multiple participants. Privacy preserving distributed data mining techniques are used for co-operative computations of the global model without revealing the individual datasets of co-operating parties to each other. Distributed data comprise horizontal, vertical or arbitrary partitioned data.

Data modification: Data modification techniques are used to alter original database record before release, for accomplishing the security assurance goal. The methods used to perform data modification are perturbation, generalization, aggregation, swapping, merging, blocking, sampling or mix of these. The data perturbation methods are used to protect the confidential data by adding random noise to original data such that aggregated values of original data and perturbed data are the same. In the blocking method, sensitive data is protected by replacing some attributes value with a specific symbol like "?". In the aggregation method,

aggregated results are used for data mining, instead of raw data for protecting privacy. For the swapping method, the values of the attributes are exchanged within the same dataset, such that original statistic of the dataset is maintained.

Data mining algorithms: Different patterns can be identified by using different types of data mining tasks such as classification, association rule mining, and clustering. After applying the privacy preserving measure, the modified dataset should support all data mining tasks that can be executed on the original dataset. But, the modified dataset supports only some data mining tasks as data utility is reduced after modification. Therefore, in this dimension, PPDM techniques are classified based on data mining tasks they support, which include privacy preserving association rule mining, privacy preserving classification, privacy preserving clustering, *etc*.

Data or rules hiding: In some cases, original data with sensitive information needs to be protected, while in other cases, the rules generated after performing the mining task like association rule mining or classification on the original data need to be protected. In this dimension, PPDM techniques are classified according to the privacy requirement of application, *i.e.*, does raw data need protection or rules need protection.

Privacy protection: The privacy protection techniques are used for selective modification of data such that the privacy-utility balance is maintained. These techniques can be classified into three types:

i) *Heuristic-based techniques* like adaptive modification which modify only selected values to maintain the privacy-utility balance.

ii) *Cryptography based techniques* like homomorphic encryption and secure multi-party computation techniques which perform the operations on encrypted data.

iii) *Reconstruction based techniques* that reconstruct the original distribution of data from

randomized data.

1.4 PPDM Applications

Privac preserving data mining has various applications in healthcare, recommender system, genomic privacy, homeland security applications, social networks, customer transaction, *etc*. Some of them are discussed in this section.

1.4.1 Healthcare

The huge data is generated by healthcare industry through record keeping, compliance and regulatory requirements, and patient care which includes Patient Health Record (PHR), Electronic Health Record (EHR), physician's notes, laboratory tests, medical imaging, sensor's data, prescriptions, insurance, and other administrative details. For extracting rules, patterns and trends, big data analytics should be performed on this data. The analytics results can help in providing low cost and effective treatments to patients. Since the healthcare data contains the personal information of patients, the data analytics on healthcare data should be performed in privacy preserving way such that no personal or sensitive information of the patient is leaked. Various systems have been designed for protecting sensitive information in healthcare data before performing the data analytics on it. Some of them are discussed below.

Sweeney [23] developed a system for removing the personal information of patients from the clinical notes and letters called the Scrub system. The clinical notes and letters contained data (patient's name, address, phone numbers, family members details, *etc.*) in textual form. The Scrub system used the number of algorithms in parallel to find the text having personal information of the patient. The traditional methods used global search and replace procedure

for de-identification of data. The Scrub system identified the 99% of personal information in the text which is much better than the traditional method whose accuracy is 30-60%.

Sweeney [24] introduced the Datafly System for performing the privacy preservation transformations on medical records stored in the database while publishing the data or performing the data mining. The stored data was in a multidimensional format having explicit identifiers as well as quasi-identifiers along with the other data. The explicit identifiers such as social security number and name were removed directly. For preventing the identification of data using quasi-identifiers (age, sex, pin code) k-anonymity approach was used. The minimum bin size was set for each field within a range of 0 and 1, where 0 means providing the original data and 1 means maximum generalization of underlying data. These values were set according to the recipient of the data. The values in records were generalized according to the bin size of the fields. Further, the outliers were also removed for the de-identification of data.

The different healthcare providers, who have the patient's information, work collaboratively for performing the data analytics on integrated data using cloud. The cloud service providers act as the third party, which provides the storage and other computing resources for performing the data analytics. Various techniques have been proposed to support such approaches. Narayan [25] provided a secure system based on attribute-based cryptography for sharing the patients' data among healthcare providers in a secure manner. Jafferi *et. al.* [26] introduced a secure management system for PHR using digital rights management system where contents are encrypted using content key encryption scheme and license containing policies to access data were issued by owners. The users possessing valid licenses were authorized to decrypt and use the content.

1.4.2 Recommender system

With the advancement in technology for storing, accessing and managing the massive data, a lot of information is available on the internet for every topic, termed as "information overload". To search for desired information in the pool of data for identifying the relevant information is a daunting task. Recommender systems are a solution to the information overload problem, which instinctively provides information to users according to their interests. They help in retaining customers, increasing sales, increasing advertisement revenues, *etc*. Collaborative Filtering (CF) [27] is a prevalent technique used in recommender systems for estimating user preferences. CF schemes provide users a list of items according to their preferences from the millions of available items in the system, which saves the users effort and time. Along with their benefits, CF schemes also have some issues like scalability, accuracy, privacy and shilling attacks. CF schemes use the customer's preferences for various products to generate the recommendation for active users. Many customers do not share their true preferences to recommender systems since they feel that such systems are incapable of protecting their data. The customers look for the system which protects their data and does not provide it to third parties without their consent. The fake user data generates poorer recommendations having low accuracy. The collection of true and accurate data is easier if the system has well-defined privacy measures. Protection of individual data while generating the recommendations can be done by using various Privacy Preserving Collaborative Filtering (PPCF) techniques [28]. PPCF schemes can be divided into two types: PPCF for centralized dataset and PPCF scheme for distributed dataset. The privacy scenario of the centralized dataset for PPCF schemes is explained above. When the dataset is distributed between multiple parties, they collaborate with each other for generating predictions using PPCF schemes on integrated data such that their sensitive data is protected. The motivation

behind collaboration is providing a more useful and richer CF service using integrated data of multiple parties.

1.4.3 Genomic privacy

Due to the technological advancements, it is easy to perform DNA sequencing and interpret the genomic information. The DNA database contains the uniquely identifying information of individuals, so privacy preserving approach needs to be used while performing analysis on it. The databases of collected DNA can be used for research purposes, medical use and law enforcement, *etc*. Removal of explicit identifiers from the database will not prevent re-identification of data, so other privacy preserving techniques like perturbation, anonymization or differential privacy are required. Malin and Sweeney [29] build a software called CleanGene for identifying the individuals from DNA entries by using publicly available information. The proposed approach had the accuracy of $98-100\%$ in the identification of individuals. The genetic profile was constructed from information like sex, genetic diseases, location of DNA collection, which can be used for identification. In another study, Malin [29] introduced a method for privacy preservation of DNA database by anonymizing DNA sequencing using generalization lattices, such that DNA entry in the database cannot be determined from $(k-1)$ other entities. The other methods for re-identification of individuals by examining the visit of patients are discussed in [30], [31] where patients visit patterns are combined with public information for unique identification. Malin [32] provided various methods to preserve the privacy of individuals in such a scenario.

1.4.4 Bioterrorism applications

In bioterrorism applications, medical data is analyzed for detecting bioterrorism attack. Usually, biological agent generates symptoms such as influenza-like illness, respiratory distress,

gastrointestinal symptoms, and febrile hemorrhagic syndromes, *etc*. Without prior knowledge of the bioterrorism attack, doctors may diagnose the patient with some related disease. For detecting bioterrorism attack, the number of cases related to particular disease needs to be identified and if it deviates from the usual number, then it can be considered as bioterrorism attack. Therefore, data related to diseases need to be reported to health agencies for surveillance of bioterrorism attack. Sweeney [33] proposed a solution for privacy preserving bioterrorism surveillance such that initially limited revelation was provided for data and if there was any suspicious activity, then access to all data was provided for identification of the bio-terrorism attack.

1.4.5 Social networking

Social networking sites such as Facebook, MySpace, Instagram, *etc.*, are web applications used by people to do social interactions with other people such as friends or relatives or people having similar interests, career or activities [34]. Every social networking site has some specific functionality such as career-oriented, social interaction, *etc*. According to Thelwall [35], these sites can be classified into three types:

- *Socializing social network services* mainly used for doing social interactions with existing friends and relatives (*e.g.*, Facebook).
- *Career-oriented social network services* used by professionals for interpersonal communication (*e.g.*, LinkedIn)
- *Social navigation social network services* used for finding specific information or resources (*e.g.*, Goodreads for books)

The social networking services have several benefits such as social interaction, on-line advertisement, web hosting service for users creation, finding job and internship opportunities by

connecting with other professionals and groups, used by activists for grassroots organization and curriculum, learning and professional use. With a lot of benefits of social networking sites, some issues which arise by the use of these sites include spamming, trolling, privacy, risk of child safety, unauthorized access and psychological effects due to excessive use.

1.4.5.1 Privacy in social network

Privacy concerns among users of social networking sites are rising due to the presence of a lot of personal information on these sites. The large organizations or governmental bodies can use personal information for creating the individual's profile, and then use that information for policy planning and decision-making. Furthermore, social network organization may retain the information which was altered or deleted by the users and then pass it on to third parties for analysis purposes. For example, social networking site Quechup used email addresses of its users for spamming operation [36]. Some potential dangers of leaking private information are:

- *Identity fraud:* The personal information provided on social network can be used for extracting the personal identity of individuals such as social security numbers [37]. The researchers of Carnegie Mellon University have showed in their study that it is possible to extract SSN numbers of individuals by using social networks and other on-line databases [38].

- *Employment-related information*: The employers analyzed the personal information of individuals, such as their posts, photos, activities, and groups joined, *etc.*, on social networking sites for screening the potential candidates [39]. Their major concerns are about getting information related to alcohol or drug use, communication skills, unseemly photos and videos, and defaming about the previous employers.

- *Preteens and teens related information*: Among all age groups, preteens and teenagers are more vulnerable victims to social networking sites. They share their personal information, locations, activities, events, photos, and videos willingly or under the peer pressure without knowing the consequences. They tend to share this information because they do not want to be left out or judged by other adolescents who are practicing these sharing activities. By sharing too much personal information they can become a victim of cyberbully, stalking, *etc*. In the future, such information could potentially harm them when pursuing job opportunities.

- *Stalking:* It is surveillance or monitoring of an individual without his/her consent. On the social networking sites, people share their photos, videos, address, *etc.*, with other on-line friends without ever meeting them once, which can make the victim of stalking very easily.

1.5 Thesis Organization

The thesis has been structured in the following chapters:

Chapter 1: Introduction: This chapter provides an overview of big data, big data mining and privacy preservation in big data mining. Further, PPDM along with its various dimensions are discussed. Various applications of PPDM have been discussed at the end of the chapter.

Chapter 2: Literature Review: This chapter provides a comprehensive review of PPDM techniques based on the data mining task classification. PPDM techniques related to association rule mining, classification, and clustering are discussed. Further, anonymization based PPDM, privacy preserving collaborative filtering and cloud-based PPDM are discussed. This chapter concludes with the problem formulation and objectives of the thesis.

Chapter 3: Proposed framework: Privacy Preserving Data Mining in Big data: In this chapter, the framework is proposed for secure data mining in big data and explained thoroughly. Further, its application on the healthcare recommender system is discussed. The PPCF scheme for generating the healthcare recommendation on arbitrarily distributed data among multiple parties is proposed. The proposed scheme has been analyzed on various parameters like privacy, accuracy, coverage, and performance.

Chapter 4: ClaMPP: A Cloud-based Multi-party Privacy Preserving Classification Scheme for Distributed Applications: In this chapter, a cloud-based multi-party privacy preserving classification scheme has been proposed. The proposed scheme uses the naive Bayesian classifier for generating the predictions on arbitrarily distributed data among multiple e-commerce companies.

Chapter 5: K-Neuro SVM based Privacy Preservation technique for Social Network: In this chapter, K-NeuroSVM based privacy preservation technique has been proposed for the social network graph. The proposed technique is evaluated on various parameters which include Average path length, Precision, Recall, F-measure and Information loss.

Chapter 6: Conclusion and Future Scope: This chapter concludes the thesis by making a brief statement of the proposed research work. Major contributions of the thesis are also provided along with the future directions for working in this area.

Chapter 2

Literature Review

Data mining is the process of extracting the hidden patterns, knowledge and information from massive data-sets [40], which are very useful for the number of applications. But, due to the excessive use of personal data of people in the data mining process, there has been an increasing concern about the privacy threat posed by the enormous data [41]. The individual's privacy can be violated due to the illegitimate access to his personal data, the use of personal data for other purposes without his consent, discovery of undesired information during the data mining process, *etc*.

To manage these privacy issues in the data mining process, a sub-discipline of data mining, named as privacy preserving data mining has been proposed by the research community in the last few years. The main aim of PPDM is to protect sensitive data or information from unwanted disclosure, while preserving the utility of the data for data mining process. After the cutting edge research of Agrawal *et al.* [15], various techniques for PPDM have been introduced [42], [43], [44], [45]. The classification of PPDM techniques can be done according to different dimensions: data distribution, data modifications, data mining tasks, data or rules hidden and privacy protection, as discussed in the first chapter (section no. 1.3.1).

2.1 Classification of PPDM techniques

The PPDM techniques have been classified into three main categories according to data mining tasks: privacy preserving association rule mining, privacy preserving classification and privacy preserving clustering. These classifications are further divided according to their data distribution: centralized or distributed. In centralized distribution, entire data is stored at a single location and owned by a single party. The sensitive information of data providers is protected in this distribution using various techniques like randomization, data perturbation, cryptography, *etc*. The data can be distributed among multiple parties having horizontal, vertical or arbitrary partitioning. In Horizontal Data Distribution (HDD), multiple parties hold different entities having the same attributes. For example, analysis of patients' data held by different hospitals for diagnose and treatment of diseases. In Vertical Data Distribution (VDD), multiple parties hold the same entities but different attributes. For example, analysis of the data held by the hospital and environmental department of the same location or areas for prevention of the outbreak of disease. The Arbitrary Data Distribution (ADD)[46] is generalized distribution, where different parties can have different entities or different attributes. The technique designed for ADD can be applied to both HDD and VDD.

2.1.1 Privacy Preserving Association Rule Mining (PPARM)

Association rule mining techniques are used to discover the frequent patterns, correlations, and associations among the items of a dataset.

Mathematically, it can be represented as:

$I = i_1, i_2 ... i_n$ represent itemset

$T = t_1, t_2 ... t_m$ represent transaction such that $t_i \subset I$

$X \mapsto Y$ represent association rule such that $X \subset I$, $Y \subset I$ and $X \cap Y = \phi$

Here X and Y are antecedent and consequent, respectively. Both antecedent and consequent can have single or multiple items. The support and confidence are two measures used to find important association rules. The support count of X is the total number of transactions that contain X, represented as:

$\delta(X) = |t_i|X \subseteq t_i, t_i \in T|$

The support of association rule $X \mapsto Y$ is the percentage of the total number of transaction that contains itemset $X \cup Y$ and confidence is the probability of occurring itemsets in Y in the transactions containing X. Mathematically, it can be represented as:

$sp(X \mapsto Y) = \frac{\delta(X \cup Y)}{n}$

$conf(X \mapsto Y) = \frac{\delta(X \cup Y)}{\delta(X)}$

PPARM is the mining of association rules from the database such that no rule will leak the sensitive information of individuals or organizations.

The main objectives of PPARM are:

i) All the sensitive rules which can be extracted from the original database at a specified threshold should not be extracted from the sanitized database at the same or high threshold level.

ii) All the non-sensitive rules which can be extracted from the original database should also be extracted from the sanitized database at the same or high threshold level.

iii) No new rules should appear while mining the sanitized database at the same or higher threshold.

There are two main approaches for PPARM:

a. Sensitive frequent itemset hiding

b. Sensitive association rule hiding.

In the first approach, sensitive frequent itemsets are hidden, and they do not appear in association rules. Hence, all discovered rules are non-sensitive. In the second approach, after mining the association rules, the techniques are used to hide sensitive association rules. These approaches for PPARM can be further divided into two types based on data distribution: centralized and distributed.

Centralized PPARM Centralized PPARM techniques can be segmented into four types: *heuristic-based, border-based, exact and reconstruction based approaches*. In the heuristic-based approaches, the selected transactions are modified for hiding the sensitive association rules. In the border based approaches, the border of lattice is redefined in order to convert the sensitive frequent itemsets into infrequent itemsets. In the exact approaches, hiding of sensitive rules or frequent itemsets is defined as a constraint satisfaction problem which is further resolved by integer or linear programming. The details of these approaches are provided below:

i) *Heuristic-based approaches*: In heuristic-based approach, the appropriate datasets are found for data modification. Since the optimal selection of dataset for data modification is an NP-Hard problem, heuristics are used for the same. Based on the data modification techniques, these approaches can be of two types: *distortion based* and *blocking based* techniques. In the distortion based method, a selected set of 1-values are changed to 0-values for lowering the support of sensitive itemsets, whereas, in data blocking techniques, existing values are replaced by "?" or unknown values to fuzzify the original values.

ii) *Border based approaches*: The itemsets which separate the frequent itemsets from infrequent itemsets form the border and this border is modified to hide the sensitive rules. Different algorithms follow a different methodology for enforcing the new and

revised borders in the modified database.

iii) *Exact approach*: In this approach, non-heuristic algorithms are used which formulates the hiding process by constraint specification problem or optimization problem. The linear or integer programming is used to solve these problems. The algorithms used in this approach provide the optimal solution with no side effects.

iv) *Reconstruction based approaches*: The data is perturbed to sanitize the sensitive association rule and then aggregate distributions are reconstructed to mine non-sensitive association rules.

Dasseni *et al.* [47] provided three algorithms to hide the sensitive rules, where the first two algorithms were based upon reducing the confidence of rule and the third algorithm was based upon reducing the support of the rule. It was shown that the first algorithm had performed slightly better than the second. The execution time of the first two algorithms was linear to the size of the database, whereas, for the third algorithm, it was directly proportional to the size of database. In another study, Saygin *et al.* [48] provided a framework for PPARM on the centralized dataset. For hiding the sensitive rules, they used a data blocking method. The new metrics (sensitivity level, support interval and confidence interval) were defined and three algorithms: CR, CH, and GIH were proposed based on rule hiding by reducing the minimum support and rule hiding by reducing the minimum confidence. The experimental results had shown that CR had the least side effects followed by GIH. Oliveira and Zaiane [49] introduced a method for balancing privacy and knowledge discovery in association rule mining. The proposed sanitizing algorithms required only two scans of database irrespective of the size of the database and a number of sensitive rules need to be hidden, so it was more scalable and efficient as compared to techniques mentioned in [47], [48]. They had also considered the effect of sanitizing procedure on legitimate rules that should

not be hidden, whereas the previous two techniques [47], [48] did not consider this factor. Further, Evfimievski *et al.* [50] discussed that privacy breaches occurred due to the use of uniform randomization and introduced a class of randomization operators (per-transaction, item-variant, select-size, cut-and-paste and mixing randomization) which provide more security during PPARM. The formulae for an unbiased support estimator and its variance had been derived to find the support counts from randomized dataset and perform PPARM. A number of PPARM techniques have been proposed since then. Some recent studies are discussed in Table 2.1.2.1.

Distributed PPARM The secure multi-party computation techniques are used for PPARM on distributed database among multiple parties. The transactions can be distributed horizontally, vertically or arbitrarily among multiple parties. The different parties want to learn global association rules collaboratively, without sharing their confidential data. So, they protect their data before sharing and then perform the computations on protected data to generate association rules. In early work, Vaidya and Clifton [51] presented a PPARM technique for vertically distributed data between two parties based on secure scalar product protocol. It was shown that the proposed protocol was secure and had quite less communication cost. Further, Kantarcıoglu and Clifton [52] introduced a PPARM technique for horizontally distributed data among three or more parties. The two phases were proposed to discover globally frequent item set. In the first phase, commutative encryption was used by each party to encrypt its own itemset and the itemsets (already encrypted) it received from other parties. These encrypted values were passed to each party, which decrypted them to obtain the complete set. In the second phase, each party added a random value to its support count and passed it to another party. The last party performed a secure comparison using Yao's generic method [53] with the first party to determine whether the final result was greater than the threshold value plus the random value. In another study, Zhan et al. [54] provided a solution

for vertically distributed data between multiple parties (two or more than two parties) using homomorphic encryption, whereas previous study Vaidya and Clifton [51] were provided for two parties only. Some recent PPARM techniques for distributed datasets are provided in Table 2.1.2.2.

2.1.2 Privacy Preserving Classification (PPC)

Classification, a two-way step is a data mining technique used to analyze the data-set and construct the set of rules to classify the target data into predefined classes or groups [55]. In the first step, a model is build based on relationships between values of predictor and values of target. In the next step, the build model is used to predict the class of target value. For example, credit risks of loan applicants as low, high or medium, can be predicted by using classification algorithms. The most commonly used classification models are Decision tree, Support vector machine, Neural network, Bayesian network, *etc.*

2.1.2.1 Decision tree

The decision tree learning model consists of a flowchart like a tree structure, where each internal node represents one of the input variables, edges from internal nodes represent the outcomes of test on internal nodes, and leaf nodes represent the class labels [55]. The path from root to a leaf node represents the values of input variables which lead to target value represented by the leaf node. Various privacy preserving techniques for realizing the decision tree classifier have been proposed [15], [56], [57]. Aggarwal [57] proposed a secure decision tree classifier on the centralized dataset using randomization and reconstruction techniques. The privacy of original data was preserved by using a randomizing function based on the uniform or Gaussian distribution. For building a classifier, original data distribution was reconstructed from the perturbed data distribution. Du and Zhan [58] proposed a random-

Table 2.1: Overview of centralized PPARM techniques

Paper	Association rule mining technique	Privacy Preserving technique	Advantages	HD	GR	LR	Computation complexity
Pathak *et al.* [59]	Distributed apriori algorithm was used on the database divided into *n* cluster	Heuristic based technique using the concept of impact factor of transaction on sensitive rule	–	zero	less	less	O(st) where *st* was time required to find impact factor of transaction on a cluster.
Domadiya and Rao [60]	Apriori algorithm [61]	Heuristic based approach named as MDSRRC	Proposed approach was better than algorithm DSRRC and it had considered multiple items on R.H.S of rule	0%	0%	36%	–
Le *et al.* [62]	Apriori algorithm [63] ś association rule mining algorithm	Heuristic algorithm based on intersection lattice theory	Proposed algorithm-ILARH has minimize the side effects as compared to Moustakides and Verykios [64] technique	0%	less than Moustakides and Verykios [64] technique	less than Moustakides and Verykios [64] technique	–
Shah *et al.* [65]	Apriori algorithm [61]	Two algorithms: ADSRRC and RRLR based on heuristic approach	The proposed algorithms overcame the shortcomings of existing algorithm DSRRC and outperform it *w.r.t.* data quality and side effects.	0%	0%	ADSRRC-36.3% RRLR-22.7%	–
Jain *et al.* [66]	Representative rules were found using approach proposed by Kryszkiewicz [67]	Heuristic approach based on Data distortion technique	The position of sensitive items were changed in the transactions to hide the sensitive rules and all sensitive rules were hidden in minimum number of passes as compared to other related techniques	0%	–	–	–

Paper	Association rule mining technique	Privacy Preserving technique	Advantages	HD	GR	LR	Computation complexity
Lou *et al.* [68]	Apriori algorithm	DPQR algorithm	In the proposed approach both data perturbation and query restriction strategies were used to increase the privacy level. The time efficiency was improved by dividing the matrix into blocks.	–	–	–	$O(2^k)$-inverse matrix $O(3^n n)$-counting process of generated n-itemset.
Hong *et al.* [69]	Apriori algorithm	SIF-IDF algorithm based on greedy approach	The degree of similarities between sensitive itemsets and transactions were computed. The transactions were sorted in decreasing order, and then they were processed to hide the sensitive itemsets.	0%	0%	Less than Wu *et al.* [70] greedy based approach	More than Wu *et al.* [70] greedy based approach
Le *et al.* [71]	Apriori algorithm [61]	Heuristic approach based on the intersection lattice of frequent itemsets	In the proposed approach, victim item is identified such that the removal of that item cause the least effect on frequent itemsets, and then victim item was removed from the minimum number of selected transactions such that distortion in original database was less as much as possible.	0%	0%	8%	1500 sec approx for 5 SARs
Lin *et al.* [72]	Apriori algorithm [61]	HMAU algorithm based on data distortion technique	In proposed approach, three parameters: hiding failure, missing rules and artificial rules, which have assigned values according to user's requirements, were used to evaluate whether transaction was required to be deleted or not.	0%	0%	0%	700 sec approx for 10% FI

ization response technique called a Multivariate Randomized Response technique for the multiple-attributes dataset.

They used this technique to perturb the original dataset and build a decision tree classifier using a modified Iterative Dichotomiser 3 (ID3) algorithm, on perturbed dataset. The accuracy of decision tree classifier is good when the randomization parameter has value in the range $[0.6, 1]$ and $[0, 0.4]$. Agrawal and Srikant [73] proposed a solution for building a privacy preserving decision tree classifier or naive Bayes classifier on the user's perturbed data. The user's data was perturbed at the personal system using the normal distribution, Gaussian distribution and by swapping values. From the aggregated perturbed data, the original data distribution was reconstructed and used for building a classification model. Few studies have been proposed for the distributed dataset. Lindell and Pinkas [15] proposed a solution for building a privacy preserving decision tree classifier based on the ID3 algorithm on horizontally distributed data using oblivious transfer protocol between two parties. Xiao *et al.* [56] proposed a privacy preserving ID3 algorithm for building a decision tree classifier based on SMC protocols for horizontally distributed data between multiple parties. Various SMC protocols based on the data disguise techniques were proposed: a secure two-party multiplication protocol, the secure two-party and multi-party reverse multiplication protocols, a secure multi-party xLogx protocol, and a secure multi-party FindMax protocol. The proposed method was more efficient than the solution proposed by Lindell and Pinkas [15]. Du and Zhan [74] discussed a privacy preserving decision tree classifier on vertically partitioned data-set between two parties, based on secure scalar dot product protocol. For achieving good performance, a semi-trusted third party was used. The communication cost of the proposed protocol is about $2n$, whereas the computation cost is $O(n)$. In this work, the assumption was made that both parties should have the class attributes.

2.1.2.2 Support vector machine

Support Vector Machine (SVM) is one of the finest available classifiers which is usually used for binary classification. The bifurcation is done on the basis of kernel function which divides the entire section into two regions. SVM basically finds the best possible linear decision surface concerning the two classes. The biased arrangement of the support vectors is generally known as the decision surface. In other words, the nature of the margin among the two classes is decided by the support vectors. Primarily for a listed class, SVM determines a hyperplane or a surface which increases the margin among the positive and negative observations. As depicted in Figure 2.1a, the boxes and circle are the support vectors that are primarily the observations that support hyperplane on both sides. The hyperplane is simply defined as a line in 2D, whereas in 3D, it is defined as plane. In greater dimensions *i.e.*, more than 3D, it is entitled as hyperplane. Margin is well-defined as the space among the hyperplane and the nearest data point. If that space is doubled up, it would be identical to the margin. SVM determines a hyperplane or a surface which increases the margin among the positive and negative observations for a designated class. Support Vector Machines are adaptable for different decision function. SVMs are very effective when very high dimensional spaces are concerned. Also, when the numbers of dimensions turn out to be bigger than the current number of samples, then also SVM appeared to be very effective. Various studies have been proposed for PPC using SVM [75], [76], [77], [78], [79]. Recently, Maekawa *et al.* proposed a PPC scheme using SVM based upon protecting the extracted features of data by using the random unitary transformation. The properties of the protected features were used for secure computation of the SVM model by using the existing algorithms. The performance of the proposed secure SVM model was same as for unprotected SVM model. Further, Hyma *et al.* [80] presented a PPC scheme using SVM for the centralized dataset based upon the heterogeneous data distortion technique. Three parameters:

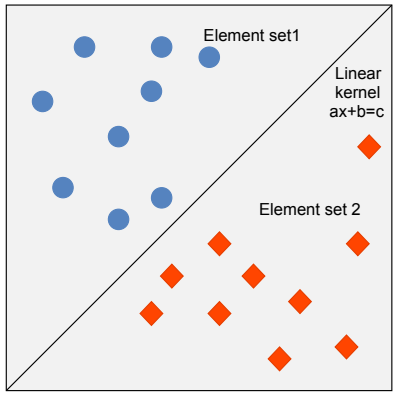
Table 2.2: Overview of distributed PPARM techniques

Paper	Data Distribution	Association rule mining technique	Privacy Preserving technique	Scenario	Communication complexity	Computation complexity
Keshavamurthy *et al.* [81]	HDD, VDD and ADD	Genetic algorithms	Trusted third party was used	For privacy preserving, concept of trusted third party with two offsets is used. The privacy of local data is preserved by collaborating parties at its own sites and privacy of global data is preserved by trusted third party.	–	–
Lakshmi and Rani [82]	HDD	–	Double hash based secure sum technique	A PPARM technique was proposed for HDD in multiple parties ($n > 2$) without trusted third party. Each party has found the global association rules from global frequent item sets, which were calculated from local frequent itemsets provided by each party in disguised form	$O(n^2)$	–
Lakshmi and Rani [83]	HDD	Apriori algorithm	Sign based secure sum cryptography technique	A model was presented for PPARM on HDD based on a cryptography technique using trusted third party. The third party starts the process and prepare the merged list of item sets. All parties have calculated the partial and total support of these item sets. Based on this information third party has found global frequent item sets.	less	less
Kaosar *et al.* [84]	HDD	Apriori algorithm,	Fully homomorphic encryption technique	A secure comparison method of integers using fully homomorphic encryption technique was proposed and based on that PPARM technique was presented.	less than the Yao's [85]	Encryption time 0.04s, Evaluation time 7s, Decryption time 1.1×10^{-4}.

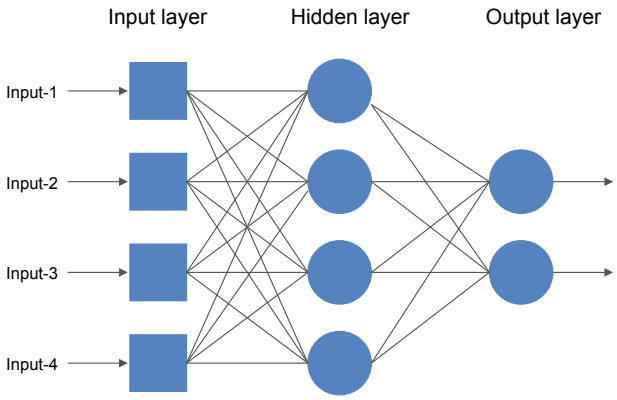
Paper	Data Distri-bution	Association rule mining technique	Privacy Preserving technique	Scenario	Communication complexity	Computation complexity
Nguyen *et al.* [86]	HDD	–	Homomorphic encryption	They have used maximum frequent itemset approach for finding the candidate set and homomorphic encryption scheme for computing global support. Two schemes were proposed: Enhanced M.Hussein *et al.* Scheme (EMHS) and MHS+Pai scheme.	less than Hussein *et al.* [87] scheme	–
Tassa [88]	HDDl	Fast Distributed Mining algorithm	Two novel secure multi-party algorithms: one for computing the union of subsets and another for finding set inclusion	Two algorithms were proposed: First algorithm for computing the union of subsets held by multiple parties, and second algorithm to find the inclusion of element held by one party into the subset held by another party.	$4(K+1)$ communication rounds, $(M^2+M-1)(K+1)$ messages, $(M^2-2)(log_2 M)n+2n\|h\|+(M-1)Bl$ bits of communication	$\Theta(Mnlog_2 M)$ bit operations
Abdel Wahab *et al.* [89]	HDD	–	ε-differential privacy	The three entities were considered in the proposed approach: data consumers, data providers and data miner. The privacy of sensitive information of each data provider was protected against other data providers and also from the inference attacks from data consumer's query. The data consumer query was also protected from data providers.	–	**Efficiency w.r.t attributes in query** For 45222 no. of records and 8 no. of specializations and 6 no. of attributes in data consumer query execution time is 1 sec. **Efficiency w.r.t specializations** 2 secs for 8 n. of specalizations and 40000 records.

Paper	Data Distri-bution	Association rule mining technique	Privacy Preserving technique	Scenario	Communication complexity	Computation complexity
Nana-vati and Jinwala [90]	HDD	Interleaved algorithm proposed by Ozden *et al.* [91],	Homomorphic encryption and Shamir's secret sharing	Two algorithms were proposed to detect partial and total global cycles while mining the cyclic association rules from the HDD among multiple parties.	Efficient private matching: $O(pk)$, Secret sharing scheme: $O(kP^2)$	Cyclic association rule generation: $O(n \times max(t_i) \times 2_n \times x)$.
Zhang *et al.* [92]	HDD	Fast distributed mining [93]	Commutative encryption scheme	A PPARM technique on HDD between two parties was proposed based on novel secure division computation protocol using commutative encryption scheme.	Protocol I: $O(\|CG_{i(k)}\| \times B)$ Protocol II: $O(\|CG_{i(k)}\| \times B)O(S_t)$	Protocol I: $O(\|GL_{i(k-1)}\|^2) + O(\|CG_{i(k)}\| \times T_c)$ Protocol II: $O(S_c)$
Lakshmi and Rani [94]	VDD	–	Cryptography (encryption, decryption)	A PPARM model for VDD among multiple parties was proposed using a third party (data miner) based on semi-honest privacy model. The proposed model is based on cryptographic (encryption, decryption) and scalar dot product techniques.	$O(n)$ data transfers	$O(m)$ number of encryption and decryption
Kumara-Swamy *et al.* [95]	VDD	C 5.0 data mining algorithm	Commutative RSA algorithm	A Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM) approach was proposed for PPARM on vertically partitioned data set among multiple parties. The semi-honest trust model was used.	$O(2(n-1)F)$ where n is number of parties and F is total bits transferred	Execution time was 0.5 sec for spam e-mail dataset

privacy choice of the data subject, domain privacy, and correlation were considered before modification of the original data. Based upon these parameters, modification of data was done by adding noise for continuous data and by the generalization process for discrete data. The proposed scheme was efficient and had good accuracy.



(a) SVM



(b) Layered architecture of neural network

Figure 2.1: SVM and Neural Network

2.1.2.3 Neural network

According to Gurney [96] "A Neural Network (NN) is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing power of the network is stored in the inter-element connection strengths, which are computed by the process of adaptation to, or learning from a set of training patterns". The NN contains three types of layers: input layer, hidden layers, and output layer.

The layer which receives the input from the external source is called an input layer. The layer which produces the final result of NN is called an output layer. There can be zero or more layers between the input and output layers, which are called hidden layers. In the hidden layer, NN changes the input data with some threshold value as per the layering and the normal transformation for this layer can be represented as:

$$w_1 = ax + b \tag{2.1}$$

where w_1 is the changed weight of the input data, a and b are arbitrary constants. The primary elements of all Artificial Neural Networks (ANN) are closely interlinked artificial neurons imitated by multiple nodes where each node accepts an input, implements a function to it and after that transfers the output. At every node, the output is referred as its activation or node value. As shown in Figure 2.1b, in the input layer, the input values are weighted, *i.e.*, each input value is multiplied with individual weight. In the hidden layer, neurons implement sum function which aggregates all the weighted inputs. In the output layer of ANN, the aggregation of previously weighted inputs is carried through activation function in order to generate the output value. A lot of studies have been done for PPC by using neural networks on the centralized as well as distributed datasets [97], [98], [99], [100]. Recently, Chabanne *et al.* [101] introduced a PPC technique by using deep neural networks for non-linear layers greater than two. The proposed technique based upon the main ideas of cryptonets [102] and batch normalization principle [103] and had better accuracy than cryptonets. Its accuracy was close to the accuracy obtained on the unprotected data. Abadi *et al* [104] presented PPC scheme for neural networks by using differential privacy. The proposed scheme had three main components: differentially private stochastic gradient descent algorithm for training the model, moments accountant for estimating the privacy costs and hyperparameter tuning

for balancing the privacy, accuracy, and performance.

2.1.2.4 Naive Bayesian Classification (NBC)

The Naive Bayes Classifier, based on Bayes' theorem, is a simple and efficient probabilistic classifier for binary and multi-class classification problems. As described in [105], NBC used for the classification tasks where each instance r is described by the attribute values $< r_1, r_2 ... r_n >$ and the target function $f(c)$ can have any value from some finite set C. The target function $f(c)$ is trained by using the number of training samples and based upon that new instance is classified. The Bayesian approach for classifying the new instance is to assign the most probable target value, v_{MPE}, given the attribute values $< r_1, r_2 ... r_n >$ that describe the instance.

$$v_{MPE} = argmax_{c_j \in C}(P(c_j | r_1, r_2 ... r_n)) \tag{2.2}$$

Using Bayes theorem

$$v_{MPE} = argmax_{c_j \in C} \frac{P(r_1, r_2 ... r_n | P(c_j)) P(c_j)}{P(r_1, r_2 ... r_n)} = argmax_{c_j \in C} P(r_1, r_2 ... r_n) P(c_j) \tag{2.3}$$

To simplify it further, it is assumed that attributes values are conditionally independent. Hence,

$$v_{NBC} = argmax_{c_j \in C}(P(c_j) \prod_i P(\frac{r_i}{c_j})) \tag{2.4}$$

where v_{NBC} represents the predicted value by the naive Bayes classifier. The conditional probabilities $P(r_i | c_j)$ and prior probabilities $P(c_j)$ are required to be calculated from the training set. Various techniques have been proposed for Privacy Preserving Naive Bayesian Classification (PPNBC) which have been discussed below.

Centralized PPC using NBC

Zhan *et al.* [106] introduced a PPNBC scheme based on a multivariate randomized response technique. They modified the naive Bayesian classifier algorithm such that it can handle disguised data and then compared the accuracy of classifiers built on disguised data and undisguised data. It was shown that when the randomization parameter θ was selected between $[0, 0.4]$ and $[0.6, 1]$ accuracy of the proposed PPNBC classifier was closed to classifier built on original data. Agrawal and Srikant [107] proposed a method for building NBC from perturbed data provided by internet users. After constructing an NBC from perturbed data, the model was provided to internet users for performing classification. Zhang *et al.* [108] proposed a new method for randomization, named as Randomized Response with Partial Hiding (RRPH). The NBC model was constructed on disguised data with RRPH method. For the proposed scheme, the selection of random parameter p between the interval $[0.35, 8]$ provides a good balance between privacy and mining accuracy. Vaidya [109] *et al.* provided a differentially private NBC on the centralized dataset, which can be placed on a cloud as a Platform-as-a-Service model. The parameters ε and n need to be adjusted according to the dataset, where n represents the total number of attributes. The time required to make proposed differentially private NBC was almost the same as required for making NBC from original data. Liu *et al.* [110] presented a secure patient-centric clinical decision support system by using PPNBC. The additive homomorphic proxy aggregation scheme was developed for building an NBC model without leaking the sensitive information of patients. Further, a secure Top-K protocol was presented for retrieving the top-k disease names by patients according to their preferences where Paillier homomorphic encryption scheme and secure multiplication protocol were used for achieving privacy. Gao *et al.* [111] introduced a PPNBC on the centralized dataset which was secure against substitution and comparison attack. The proposed techniques used the double-blinding technique combined with additively

homomorphic encryptions and oblivious transfers to preserve the privacy of both parties (user requesting a classification service and server holding a classification model). It has less on-line overhead computation and communication costs.

Distributed PPC using NBC

Kantarcıoglu *et al.* [112] proposed a PPNBC scheme for horizontally distributed dataset between multiple parties based on secure sum and logarithmic protocol and the semi-honest attack model was considered. Vaidya *et al.* [113] provided a PPNBC scheme for horizontally or vertically distributed data between multiple parties. They assumed the semi-honest privacy model and various SMC protocols (secure sum, secure logarithmic, one-out-of-two oblivious transfer, secure scalar dot product) were used. Zhan *et al.* [114] presented a PPNBC for horizontally distributed data between multiple parties by using homomorphic encryption scheme. The proposed scheme was secure and efficient in terms of computation and communication cost, having upper bound for communication complexity as $\alpha(m^2+2n+3t+2)$ and for computation complexity as $2N+m2+(5+n)t+n+m+1$, where α, m, n, and N represent the count of bits for each transmitted element, number of classes, number of parties, count of random numbers generated by a party and the total number of records respectively. Yi and Zhang [115] proposed a PPNBC for horizontally distributed data between multiple parties using a homomorphic encryption scheme. First, they provided the two-party protocol for PPNBC and then extended it to multiple parties using a semi-trust mixer model. In the semi-trusted mixer model, party sent messages to two semi-trust mixers, which execute the two-party protocol and send the classification result to other parties. The proposed protocol was secure and had good efficiency in terms of computation and communication cost as compared to the protocol proposed by Kantarcıoglu *et al.* [112].

2.1.3 Privacy Preserving Clustering (PPCl)

Clustering is the process of combining a set of similar data objects into a group called cluster. The objects of a cluster have high similarity, and they are very dissimilar to objects in other clusters. The similarities and dissimilarities between objects are generally calculated using different distance measures such as Euclidean, Manhattan, Hamming distance, *etc.* Clustering methods can be classified into partitioning methods, hierarchical methods, density-based methods, *etc*. Based upon data distribution recent studies of privacy-preserving clustering can be divided into two types: schemes based on perturbation for centralized dataset and schemes based on SMC for distributed datasets.

Centralized PPCl

Oliveira and Zaıane [116] presented two data transformation techniques: Object Similarity-Based Representation (OSBR) and Dimensionality Reduction-Based Transformation (DRBT). OSBR transformation was used for privacy preserving clustering on centralized data, whereas DRBT was used for privacy preserving clustering over centralized data and vertically distributed data. Parameswaran and Blough [117] introduced a data obfuscation technique named as Nearest Neighbor Data Substitution (NeNDS) and its hybrid version using Geometric transformation named as GT-NeNDS. Authors defined the property, named *reversibility*, was to check the robustness of data obfuscation techniques. The GT-NeNDS technique had cluster preservation, high displacement and difficult to reverse properties. Hajian and Azgomi [118] presented a PPCl technique for a centralized dataset using haar wavelet transform and scaling data perturbation techniques. Li and Zhang [119] presented a novel solution for PPCl over centralized data using Double-Reflecting Data Perturbation Method (DRDP) and Rotation Based Translation (RBT) method. The proposed solution used hybridization of DRDP and RBT methods, named as Hybrid Data Transformation (HDT),

for perturbing the sensitive numerical attributes. It preserves the statistical features of data while performing clustering. The variance between actual and perturb value was used to quantify the privacy, and misclassification error was used to determine the effect on privacy on the accuracy. Rajalaxmi and Natarajan [120] proposed the two data transformation methods to perturb the sensitive categorical data while performing cluster analysis. These were: Hybrid Data Transformation using Translation and Rotation (HDTTR) and Hybrid Data Transformation using Scaling and Rotation (HDTSR). In HDTTR, sensitive categorical attributes were perturbed using additive noise followed by a rotation, whereas in HDTSR sensitive categorical attributes were perturbed using multiplicative noise followed by a rotation. They also used variance for quantifying the privacy and miss classification error for determining the effect of privacy on the accuracy. Liu and Xu [121] proposed a PPCl scheme that used hybridization of random response technique and geometric transformation method for data perturbation of numerical data. The proposed scheme was efficient with high privacy and accuracy of mining results. Four different types of geometric transformation used were: translation, scaling, rotation, and reflection. The four random parameters which correspond to probabilities of these four geometric transformations were used in the proposed algorithm to select the different geometric transformation, which perturbs the sensitive numerical data.

Distributed PPCl

Merugu and Ghosh [122] introduced a secure framework for distributed clustering for the unsupervised and semi-supervised case. First, the generative models were trained on local data and then center combiner integrate these models based upon the received parameters of generative models. The artificial samples were generated using Markov chain Monte Carlo techniques and then the combined global model was fitted to these samples. The most important algorithm in clustering is k-means. Various privacy preserving k-means algorithm were proposed on distributed data. Vaidya and Clifton [123] introduced a solution for PPCl

over vertically partitioned dataset among multiple parties. The entities of each party were kept confidential, and a homomorphic encryption scheme and secure permutation [124] were used to find out the secure distance between parties. The three non-colluding parties were required for the execution of the protocol and computational costs were also very high. Further, Samet *et al.* [125] introduced an algorithm, which can be applied to two parties and there was no need for non-colluding parties. The proposed algorithm was based on novel secure multi-party addition protocol and secure sum [126] protocol. However, the protocol reveals the number of entities in clusters. Doganay *et al.* [127] presented a PPCl scheme using an additive secret sharing scheme instead of a homomorphic encryption scheme to reduce communication and computation costs. The overhead costs were reduced, but they end up using four non-colluding parties instead of three [123]. For horizontal distribution, Jha *et al.* [128] proposed the two protocols for PPCl on the dataset divided between two parties. First protocol named OPE, was based on oblivious polynomial evaluation and the second protocol named DPE, was based on homomorphic encryption. The entities of two parties were kept confidential, while intermediate centers of clusters were revealed during the execution of the above protocols. The semi-honest privacy model was considered. Further, Samet *et al.* [125] provided the solution for the multi-party environment. The secure division protocol was used, which protect the entities of each party as well as does not reveal the number of entities in clusters. But the intermediate centers of clusters were also revealed in this protocol. Jagannathan and Wright [46] introduced a PPCl for ADD between two parties. The secure scalar product protocol was used for calculation of distance and Yao's garbled circuit was used for secure comparison and calculation of intermediate centres of clusters. The proposed protocol has high computation cost, since security primitives were executed separately for each entity. In this protocol, the division was considered as the multiplication of inverse, which was not a correct implementation of the k-means algorithm. Su *et al.* [129]

proposed a solution for ADD using secure scalar product protocol, secure oblivious polynomial evaluation protocol, and secure approximation technique. It was more secure and accurate in mining results than protocol proposed by Jagannathan and Wright [46] but has some drawbacks such as leakage of the intermediate assignments of objects to clusters and the count of objects in each cluster to different parties. Further, Bunn and Ostrovosksy [130] introduced a PPCl protocol that had secure sub-protocols based on Paillier cryptosystem for each operation of k-means clustering algorithm. The proposed protocol was more secure as it does not leak intermediate centers and clusters assignment, but it leaks the number of iterations of k-means algorithms.

2.2 Anonymization based PPDM

Anonymization is a process of removing the personally identifiable data from the original data through different methods such as generalization, suppression, data removal, permutation, swapping, *etc*. [131]. The k-anonymity method [132] is treated as the classical anonymization method and the majority of the studies are based on k-anonymity. Few studies are based on its improved methods like l-diversity, t-closeness, k^m-anonymization, (α,k) anonymity, *etc.* [133].

The aim of the k-anonymity method is to generalize the values of quasi-identifiers in an original table such that every record in the modified table is identical to at least k-1 other records in terms of the quasi-identifiers. If an attacker knows the quasi-identifier values of target individual, the probability of identification of the target's record from the k-anonymized table will not exceed 1/k. Poovammal and Ponnavaikko [134] introduced a task-independent anonymization technique for protecting the sensitive raw data such that the utility of data was preserved for mining purpose. In most of the privacy preserving general-

ization methods, loss of information is due to generalization (transformation) of QI attributes and SA attributes. Authors gained both privacy and no information loss, by only transforming part of the QI and SA attributes. Zhu and Peng [135] proposed a modified entropy l-diversity model to the address privacy of medical data where more detailed attacking conditions and characteristics of medical information were taken into consideration. Approaches that address specific problems were also developed using the anonymization method. The k-anonymity based method illustrated by Matatov *et al.* [136] was used to search for optimal feature set partitioning and the one discussed by Funga *et al.* [137] was used for cluster analysis. Zhu *et al.* [138] proposed a data reconstruction scheme for the k-anonymization of the dataset before performing the predictive analytics on it. The aggregation masking method was used for protecting the sensitive numerical data and the swapping method was used protecting the sensitive nominal data. The good subset from the masked data was found by using the genetic algorithm and then it was replaced in the original dataset, such that the modified dataset satisfies the k-anonymity constraint.

2.2.1 Social network privacy using anonymization techniques

In recent years, social networking sites are being used by people from every age group. For finding the interesting social patterns and knowledge from data shared on these sites, the social network analysis has become very significant for organizations. To perform the analysis, organizations that own the social network applications may need to provide their data to third parties. But, even after the removal of the explicit identifiers from the dataset, which is called as naive anonymization, the publication of the network data may leak sensitive information about individuals, such as one's intimate relationships with others. So, before publishing the data, it should be completely anonymized. The k-anonymization and its variants such as the k-neighborhood-l-diversity model [139], k-isomorphism [140] and k-degree l-diversity

model [141] are used for properly anonymizing the social network data. The next paragraph discusses this in detail.

Recently, a lot of research work has been done in social network privacy. The social network data can have two types of attacks: active and passive attacks. In the active attack, when the organization is collecting the social network data, the attacker embeds the special type of sub-graphs in the original graph. The attacker then re-identifies those sub-graphs and nodes connected to them from the published social network graph. In a passive attack, the attacker uses background information like vertices degree, neighborhood information, connections between vertices, *etc*. The two privacy models k-anonymity and k-anonymity l-diversity have been used to protect the node re-identification and combination of node re-identification and the sensitive attribute of nodes, respectively.

Fung *et al.* [142] introduced a k-anonymization technique for a social network dataset with the aim of protecting the privacy of common sharing patterns (s-patterns), which was used for knowledge discovery for marketing and consumer behavior analysis process. The experimental evaluation has been done on three real-life datasets: Gnutella05, Gnutella08 and Adult. For computing, the data utility of s-patterns, the transformation of the common s-patterns before and after anonymization were measured and A+ tool called MAFIA was used to discover the frequent s-patterns. Zhou and Pei [141] presented a k-degree l-diversity model for protecting the privacy of social network data before publishing. They implemented distinct l-diversity and recursive (c, l) diversity privacy model and the proposed KD3D model was tested with CORA and DBLP data sets and the KD5D model was tested on the Arnet dataset. The experimental results demonstrated that the addition of noisy nodes was reduced as compared to the previous work by edge editing only. Yuan and Chen [143] identified some significant privacy attacks, named as neighbourhood attacks. They identified the effect of the parameters, namely, α, β and γ on the results obtained after anonymization and concluded

Table 2.3: Related studies for social network privacy

Paper	Privacy model	Approach	Experimental results
Liu and Terzi [147]	k-degree anonymous model	The k-degree anonymization model for graph is defined and then implemented using dynamic programming (Greedy algorithm). Edge addition and deletion operations are performed to achieve k-degree anonymity	The original and anonymized graph are compared on parameters Cluster coefficient, APL, Edge intersection and Power law distribution. It has been concluded that k-anonymization on graph is more challenging than tabular data and the single change in node or edge has effect on whole network graph.
Zou *et al.*[148]	k-automorphism model	An algorithm K-Match (KM) was proposed which ensures k-automorphism. The proposed k-automorphism privacy model protects the network graph against multiple structural attacks (degree attack, sub-graph attack, 1-neighbour attack and hub fingerprint attack). The dynamic release of network graph is also addressed.	The proposed algorithm evaluated on real dataset (Prefuse) and synthetic datasets (ER and SF) on parameters: Total degree difference, Average shortest path and Cluster coefficient. The utility of proposed anonymous graph is less as compared to other related techniques which deal with the single structural attack.
Zhou and Pei [139]	k-neighbourhood-l-diversity model	The neighbourhood attack is defined for graph. The traditional k-anonymous l-diversity model of tabular data is extended to k-neighbourhood l-diversity model for graph. It has been shown that finding optimal k-neighbourhood l-diverse graph is NP Hard problem and its practical solution is given by using greedy algorithm.	The anonymized graph contains all vertices and edges of original graph. The performance of aggregate network queries on the anonymized graph has been analysed, which shows high accuracy.

Paper	Privacy model	Approach	Experimental results
Cheng *et al.* [140]	k-isomorphism	The basic algorithm and its refinement are proposed to achieve k-isomorphism privacy model in graph. The embedded subgraph structural attack is considered. A graph is first divided into k subgraphs having same vertices and then by using edge addition and deletion, pairwise graph isomorphism is obtained.	The proposed algorithm is evaluated on 5 dataset from 3 applications named as: HEP-Th, EUemail and LiveJournal. It protects the NodeInfo and LinkInfo against embedded subgraph structural attack and the utility of anonymized graph is also preserved.
Ying and Wu [149]	Edge anonymity	The investigation has been done on the affect of edge randomization on graph properties, specifically spectrum of graph. The two methods : Sptr Add/Del and Sptr Switch are proposed for edge randomization while preserving the spectrum of graph.	The experiments are conducted on political book graph and political blogosphere datasets. The utility of perturbed graph is improved with some loss in privacy.
Yuan *et al.* [150]	k-degree l-diversity anonymity model	The k-degree l-diversity anonymity model is defined to protect structural information as well as sensitive labels of node. A new method of noisy node addition is proposed for achieving k-degree l-diversity anonymization.	The proposed technique is evaluated on three datasets Arnet, Cora and DBLP. It has been demonstrated that noisy node addition method produce less distortion in graph properties as compared to only edge editing method.

that a person can trade-off between adding edges and deriving labels by modification of these three parameters. Krishna and Murthy [144] presented k-degree l-diversity privacy model based on the noisy node addition method such that there is less distortion in the original properties of the graph. They explained both distinct l-diversity and recursive assortment for the proposed k-degree l-diversity privacy model. The experimental results illustrated that the proposed technique had improved results compared to the preceding work using edge editing only. Vasudevan *et al.* [145] proposed two privacy preserving approaches: the first one protected data privacy using an extended role-based access control mechanism and the second approach used cryptographic techniques. Xu *et al.* [146] proposed a trust-based mechanism based on bandit approach for collaborative privacy management in social networks. In the proposed scheme, aggregated opinions of all the involved users were considered before posting any item on social networking sites. A brief overview of some studies related to social network privacy is given in Table 2.2.1.

2.3 PPCF

PPCF is one of the applications of PPDM. Various PPCF schemes have been proposed for the centralized dataset. Polat and Duo [151] used Randomized Perturbation Techniques (RPTs) to overcome the privacy problem in both correlation and Singular Value Decomposition (SVD) based CF schemes. Bilge and Polat [152] discussed the methods to generate predictions with more accuracy by utilizing Discrete Wavelet Transformation (DWT) technique while preserving the individual users' privacy. For privacy preservation, users' data was masked by using RPTs. They tried to enhance the scalability and accuracy of PPCF schemes by applying k-means, fuzzy c-means, self-organizing map and bisecting k-means clustering on CF schemes while preserving the user's privacy [153], [154]. Parra *et al.* [155]

introduced an architecture for preventing user profiling based on their ratings, by using a perturbation technique called suppression of ratings. The proposed approach enhanced privacy at the cost of degradation of accuracy.

For the distributed dataset, Polat and Du [28] proposed a solution for generating recommendations for a single item from the VDD distribution between two parties while preserving their sensitive information. The proposed approach was based on the homomorphic encryption technique. Further, Polat and Du [156] introduced the techniques for generating top-n recommendations from HDD or VDD distribution among multiple parties. The proposed techniques were based on perturbation techniques; hence computation overhead was less as compared to other techniques based on cryptographic techniques. Kaleli and Polat [157] proposed the PPCF technique for HDD on multiple parties based on clustering . Different parties perform clustering of their distributed data off-line securely and then generate recommendations online, based on the k-nearest neighbor approach. Further, Kaleli and Polat [158] discussed a scheme based on NBC for generating predictions on distributed data while preserving privacy. The scheme was based on binary ratings which determines whether user will like or dislike an item. Yakut and Polat [159] investigated on how to generate referrals from HDD or VDD distribution of ratings between two parties using the SVD based method. The proposed approach was secure, generated accurate predictions and did not introduce overhead on-line computation costs. Jeckman *et al.* presented a PPCF scheme based on cryptographic techniques on HDD such that there was no accuracy loss due to privacy measure, but it has high computation costs [160]. Further, Kikuchi *et al.* presented the PPCF protocol on HDD among two parties based on cryptographic protocols [161]. For improving the computation time of protocol, a similarity measure based on the aggregation of locally computed similarity was used. All the schemes discussed in this section are for HDD and VDD, whereas ADD is more generalized distribution. Yakut and Polat [162], [163], [164]

proposed three different schemes for ADD between two parties whose overview is provided in Table 2.4. The proposed techniques have high computation cost due to the use of a homomorphic encryption technique. Memis and Yakut [165] proposed the PPCF technique for resolving overlapped rating problems, but they did not consider the multiple parties and the computation cost of the off-line model generation process is also high .

While exploring the medical domain, the two primary studies have been considered: Katezbeisser and Petkovic [166] and Hoens *et al.* [167]. Katezbeisser and Petkovic [166] proposed a privacy preserving medical recommender system for centralized architecture, which encodes the patient health data into the binary vector and recommends a suitable doctor by using matching protocol. Hoens *et al.* [167] proposed a medical recommendation system while maintaining patient privacy. Two architectures secure processing and anonymous contributions were proposed such that both maintain the patient's privacy about who was contributing the ratings and who was inquiring. The secure processing architecture has high computation cost whereas, in anonymous contributions, architecture security can be jeopardized if available ratings are less.

2.4 PPDM based on Cloud

The advent of cost-effective cloud services provides a great opportunity for organizations to reduce their cost and increase productivity. The company lacking in expertise or computational resources can outsource its mining needs to a cloud, but the outsourced data may contain sensitive information that cannot be directly stored or mined on the cloud. To protect the sensitive data, privacy preserving techniques are required, while storing and mining the data on the cloud. Various techniques have been proposed for PPDM through the usage of the cloud. Yuan and Yu [168] provided secure back-propagation neural network learning

on ADD among multiple parties on cloud environment based on the doubly homomorphic encryption technique. In the proposed technique, all parties locally encrypt their data and stored it on cloud. The cloud then executes the most expensive operations related to back-propagation neural network on encrypted data and provides the result to all parties in an encrypted form. The proposed technique was scalable and had minimal computation and communication costs for each party. Yi *et al.* [169] discussed the PPARM technique on the encrypted data stored in the cloud. Wei *et al.* [170] presented a *SecCloud* protocol for storage security as well as computation security. The *SecCloud* protocol used designated verifier signature, batch verification, and probabilistic sampling techniques. Lin and Deng [171] proposed an algorithm, named as, Cloud-based Association Rule Mining (CARM) for secure frequent itemset mining on cloud environment by using the parallel algorithms. The CARM algorithm was scalable and had less computation costs. Jindal and Dave [172] introduced a data privacy protocol for the distributed cloud environments. The proposed protocol protected the user's information from unauthorized access, malicious users and cloud owners.

The proposed technique used the $n(n \geq 2)$ semi-honest servers, three solutions were provided based on distributed Elgamel cryptosystem to attain privacy at item-level, transaction-level, and database-level respectively. Kao *et al.* [173] introduced a reversible contrast mapping algorithm for preserving the privacy of data stored in the cloud while performing the data mining on it. Li *et al.* [174] presented a privacy preserving frequent itemset mining and PPARM techniques for vertically partitioned datasets using the cloud. The efficient homomorphic encryption and secure comparison scheme were proposed for maintaining data privacy. The proposed scheme was secure and more efficient in terms of computation time w.r.t other related techniques. Lai *et al.* [175] introduced a semantically secure solution for outsourcing the association rule mining on cloud-based upon a fully secure symmetric-key

Table 2.4: Comparison of PPCF schemes for ADD

Scheme	Data Distribution	Multi-party	Privacy preserving method	Advantages	Limitations
Yakut and Polat [162]	CDD	No	Privacy preserving CF based on homomorphic encryption, addition of default and fake votes	Less online-computation cost and balance is maintained between parameters accuracy, privacy and efficiency	High computational cost of off-line model generation, scheme can not be applied on ADD.
Yakut and Polat [163]	ADD	No	Privacy preserving CF based on homomorphic encryption, addition of default and fake votes	Scheme is applied on generalized partitioning ADD and balance is maintained between parameters accuracy and privacy	High online and off-line computation and communication cost
Yakut and Polat [164]	ADD	No	Privacy preserving NBC based CF by using homomorphic encryption, addition of default and fake votes	Less on-line-computation cost and balance is maintained between parameters accuracy, privacy and efficiency	High off-line computational cost, scheme can not be applied on numeric ratings.
Memis and Yakut [165]	ADD	No	Privacy preserving CF for overlapped ratings by using homomorphic encryption, addition of default and fake votes	The effect of overlapped ratings is taken into account	High computation cost of off-line model generation process.

predicate-only encryption scheme. Hussain *et al.* [176] introduced a secure cloud based scheme called incentives-based vehicle witnesses as a service (IVWaaS), which employs vehicles moving on the road as the witnesses to designated events. Further, Giannotti *et al.* [177] presented a novel encryption scheme named Robfrugal for PPARM on cloud. The cloud service provider acted as a third party which mined the association rules for the data owner. The items and association rules were considered as private data of data owner (organization). The third-party performed mining on encrypted data using the Robfrugal technique and provided the pattern to the data owner. The data owner extracted the actual patterns from the received patterns from the third party.

2.5 Motivation

The data has a vital role in all big data-based applications. Various data analysis techniques are applied to data, which makes maintenance of privacy of the data the most important component. The PPDM techniques are used to facilitate the privacy and security of an individual's or organization's information along with the mining of useful information. While applying PPDM techniques the data miner needs to maintain a balance between privacy and utility. The implications of privacy and data utility vary with the characteristics of data and the purpose of the mining task. As data types become more complex and new types of data mining applications emerge, finding appropriate ways to quantify privacy becomes a challenging task that needs to explore further. Considering the current scenario, the definition of privacy is essentially personalized. There is a need for introducing personalized privacy in existing privacy preserving data mining or privacy preserving data publishing techniques and developing the practical personalized anonymization methods. Further, there is also a need for quantifying an individual's privacy preference. Currently, the related PPDM techniques

either provide a high level of data privacy with high overhead costs (Computation and Communication costs) or jeopardize the privacy of data. If traditional cryptography techniques are used for privacy, they can have large overheads for large datasets. Usage of anonymization techniques has a risk of re-identification of anonymized dataset. Since a large amount of data is available and the linking the different data sources with the anonymized dataset can lead to the re-identification of data. Therefore, a lot of research efforts are required for the security and privacy challenges of big data. Here, in this thesis, the focus is on privacy preserving data mining technique for big data, which is secure, has high utility and which is efficient in terms of computation costs.

2.6 Research Objectives

The objectives of the proposed work are as:

i) To analyse and explore various Privacy Preserving Data Mining (PPDM) techniques and analyse their applicability in various data mining domains.

ii) To propose the techniques for privacy preservation of sensitive information.

iii) To test and compare the proposed techniques using various evaluation metrics (like accuracy, computation time) and existing techniques respectively.

Chapter 3

Proposed Framework: Privacy Preserving Data Mining in Big Data

This chapter presents the architecture of the proposed framework Privacy Preserving Data Mining in Big Data (PPDMB) for applications like healthcare, recommender system, bioterrorism, *etc*. The architecture and layered view of proposed framework PPDMB and its application on the healthcare recommender system is presented in this chapter.

3.1 Architecture of PPDMB

The architecture of PPDMB framework is divided into three layers.

3.1.1 Layered view

Layer 1 : *Data collection and pre-processing phase:* Since the collected data is not in the standard format, it needs to be cleansed and standardized before performing any analysis on it. Various phases of data pre-processing are:

(a) Data cleansing: Once the data is collected, it should be cleansed and transformed

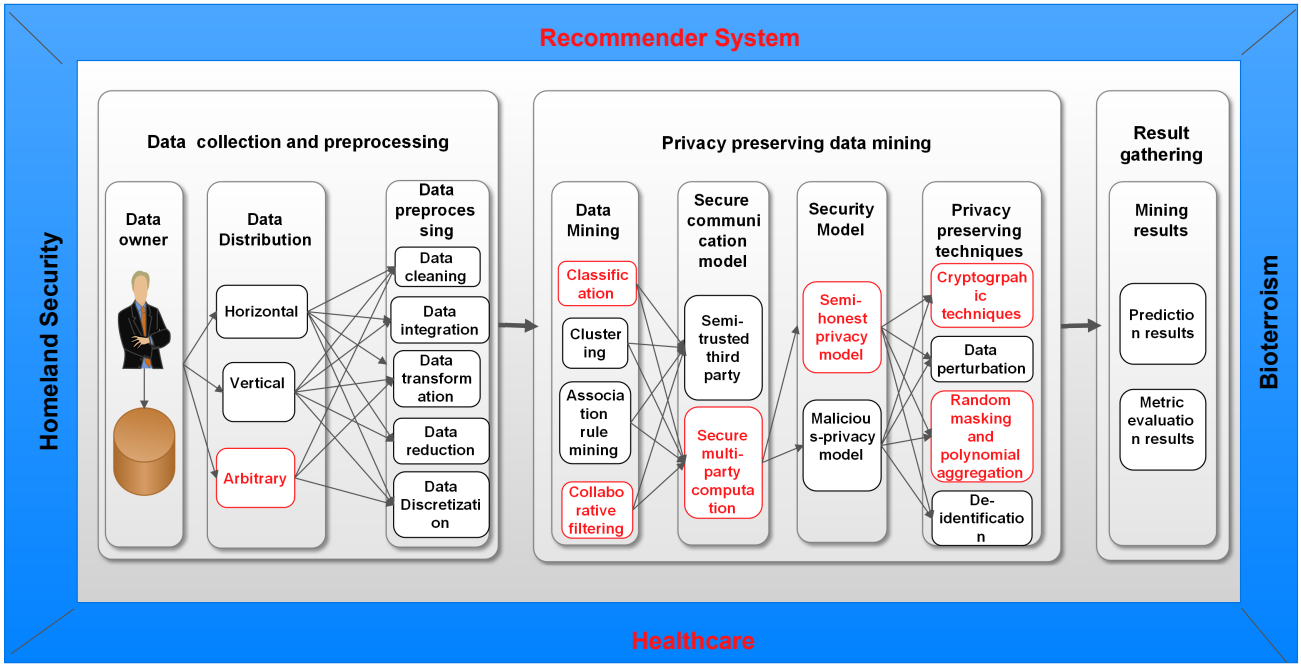


Figure 3.1: A framework for privacy preserving data mining in big data

into useful and appropriate form by data preparator. In the proposed framework, missing data is removed from the dataset depending upon the threshold value.

(b) Data integration: In data integration, data with different representations are combined and data conflicts are resolved for *e.g.*, attribute values having different representations and scales in different databases.

(c) Data transformation: This process is used for rectifying any inconsistencies in the dataset for *e.g.,* "Attribute Names inconsistency". In some scenarios given data elements might be referred by different names in the different databases. In the proposed work, given data element is refereed by the same name in all databases.

(d) Data reduction: It is done to present a reduced representation of the original data having similar analytical results. The data reduction techniques include simple aggregation, tabulation or more sophisticated techniques like principal component analysis, factor analysis, *etc.*

(e) Data discretization: It is performed to reduce the number of values of a continuous attribute by dividing them into the range of attribute intervals.

Layer 2 : Privacy preserving data mining is the process of performing data mining in a secure manner. Various data mining techniques include classification, clustering, association rule mining and CF (Discusses in Chapter 2 Section 2.1, 2.3). In the proposed work, CF is used for generating the predictions in healthcare and e-commerce recommender systems.

For distributed applications, data analytics needs to be performed in a secure manner by using the collective data of all parties. Two secure communication models for this are trusted third party model and SMC model.

- Trusted third party model: In this model, all parties provide their data to trusted third party, which performs the analysis on all data and provides the results to all parties. The trusted third party does not share any information, of an individual party, with other parties. Identifying such a trusted third party in a real-world scenario is quite a cumbersome task.
- SMC model: This model is used by parties to perform analysis of their collective data in a secure manner without using any trusted third party. The two main requirements of SMC protocols are privacy and correctness [178] *i.e.* parties should learn only about their output during the execution of protocols and each party should receive the correct results.

In the proposed work SMC model is considered. The two types of privacy models considered under SMC are the semi-honest privacy model and the malicious privacy model [178].

i) *Semi-honest privacy model*: In this model, all parties follow the protocol speci-

fication honestly, but try to gain the information from the intermediate and final results obtained during the protocol execution. It is also called an honest but curious or passive privacy model.

ii) *Malicious privacy model*: In this model, parties can deviate from protocol specifications to obtain the sensitive information of other parties during protocol execution, also called active privacy model. It is highly secured since no adversarial attack can succeed in this model.

The semi-honest privacy model is considered in our work since protocols for this model are efficient, prevent the inadvertent leakage of sensitive information between the parties and hence serve as main step toward designing a highly secure solution.
Various privacy preserving techniques used for analytic on distributed data are Cryptographic techniques, Randomization, Random masking, polynomial aggregation, data perturbation, *etc.*

(a) Cryptographic techniques: Generally public key encryption and homomorphic encryption techniques are used for distributed PPDM. The generic construction paradigms used with these cryptographic primitives are oblivious transfer, oblivious polynomial evaluation, secret sharing scheme, and zero-knowledge proofs.

(b) Data perturbation: In these techniques, random noise is added to data for protecting the privacy of data. The noise added is sufficiently large so that individual values cannot be recovered. Two main methods used for the randomization process are probability distribution method and the value distortion method [42].

   i. Probability distribution method: The original data is replaced by the same distribution's sample data or by the distribution itself.

   ii. Value distortion method: The random noise is added to the attribute values

using additive or multiplicative random noise.

(c) The random masking and polynomial aggregation techniques are used in this method to mask the attribute values using random numbers and the required values are calculated from the masked data using polynomial aggregation technique.

(d) De-identification: It is a traditional technique used to preserve the privacy of data by using techniques like swapping, generalization, suppression, removal of data, *etc*. Since a large amount of data is available so chances of re-identification of data are increasing. To lessen the effects of re-identification of data, more rigorous privacy models k-anonymity, l-diversity, t-closeness, and their variants have been introduced.

Layer 3 : In this layer, predictions are received after applying the PPDM techniques in the previous layer. These results are evaluated based on various metrics which include accuracy, security, computation and communication cost.

In the research work carried out so far in the domain of the privacy preservation of distributed big data analytics, PPDM techniques based on SMC protocols are used. The SMC protocols use the cryptographic techniques for the secure computation of function based upon the distributed data among multiple parties. Due to the use of cryptographic techniques, computation overhead is very high. Some SMC protocols have used the data perturbation techniques, which cause some loss in accuracy. In PPDMB, random masking and polynomial aggregation techniques are used for SMC protocols and the homomorphic encryption technique is used for computation of secure sum only, which results in less computation overhead and negligible accuracy loss.

The proposed framework is implemented on two applications: one is PPCF for healthcare recommender system which is discussed in this chapter and another one is cloud-based

privacy preserving naive Bayesian classification scheme for general recommender system which is discussed in Chapter 4.

3.2 EMS-PPCF: An Efficient Multi-party Scheme for Privacy Preserving Collaborative Filtering for Healthcare Recommender System

The tremendous growth in Information technology has revolutionized the healthcare industry in a big way. Patients use the Internet for getting information about diseases, diagnoses, treatments, physicians, and hospitals. One of the major concerns of the patients is choosing an appropriate physician for the treatment of disease. A huge amount of healthcare data presently available all over the world can be used to provide better healthcare services to patients facing various aging problems. Since the extraction of useful personalized information from the large amounts of data is a difficult task, one of the available options to ease this information overload is to use healthcare recommender systems. Ailing persons can ask their friends and family for advice on where to seek treatment. The recommendations produced by them are trustworthy, but the likelihood that friends and family have the same medical history as that of the patient is quite low. The other option for patients is to find the information and/or ratings of the physicians through the Internet. Since the medical data is generally treated as confidential, these ratings might be sparse and inaccurate. If the privacy of a patient, who is contributing to the recommender system, is maintained and different hospitals or health centers having healthcare data work collaboratively in a secured manner, then reliable and accurate predictions can be generated for the physician. Different sites, having ADD of healthcare services at various nodes can collaborate with each other to gen-

erate customer's preferences leading to mutual advantage and overcoming the issues related to insufficient ratings. Few studies have proposed PPCF on ADD, but they have considered only two parties and the computation cost of the off-line model generation process is also high since they use the homomorphic encryption technique.

In this section, an Efficient Multi-party Scheme for PPCF for healthcare recommender system, named as EMS-PPCF, is proposed for ADD among multiple healthcare sites. Two phases have been considered: off-line model generation and online prediction generation. Three protocols have been considered for privacy preservation and analysis of each protocol is done separately. The Paillier homomorphic encryption system is used to calculate the length of vector X securely. Analysis of the EMS-PPCF has been done for security, accuracy, coverage, and performance on healthcare and Movieslens datasets.

3.2.1 Background

Patients can submit their symptoms to the healthcare recommender system which can identify the disease from symptoms and recommend the physicians to them. The symptoms of the patients like blood pressure, blood sugar, heartbeat, and temperature can be recorded by using wireless body sensor nodes, which can transfer patient's information to a smartphone via Wireless Body Area Network (WBAN). Further, the smartphone submits this information to the healthcare recommender system through mobile networks or Wi-Fi. The recommender system recommends the physician to patient after diagnosis of disease from symptoms. The overview of the health monitoring scenario using WBAN is depicted in Figure 1. Patients submit their feedback or ratings about physicians to healthcare centers after receiving the treatments. It is assumed that $H_1, H_2, ...H_n$ represent n healthcare recommender systems, where $S_1 = (P_1, P_2...P_{h1})$, $S_2 = (P_1, P_2...P_{h2})$,...$S_n = (P_1, P_2...P_{hn})$ are set of patients registered with healthcare recommender systems $H_1, H_2, ...H_n$ respectively, such that $S_1 \cap S_2 ... \cap S_n \neq \emptyset$,

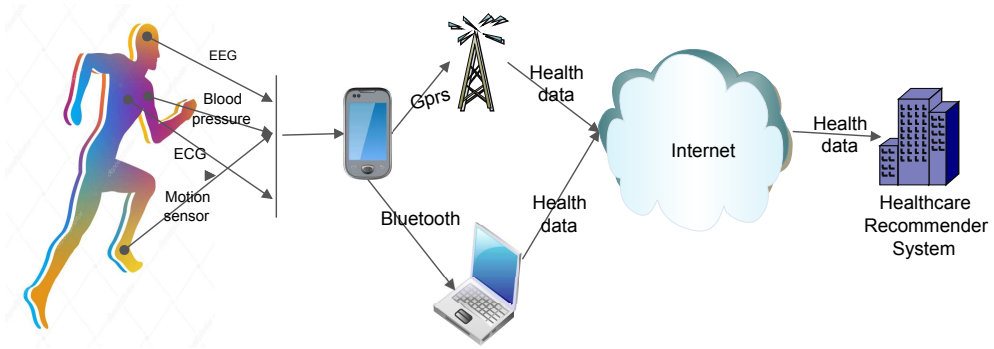


Figure 3.2: Health-monitoring scenario via WBAN

as shown in Figure 3.3. The patient's ratings can be distributed horizontally, vertically or arbitrarily among different healthcare centers, where arbitrary distribution is more generalized as compared to horizontal or vertical distribution. For providing accurate and reliable recommendations, all ratings are needed to be considered. Among the various techniques used in

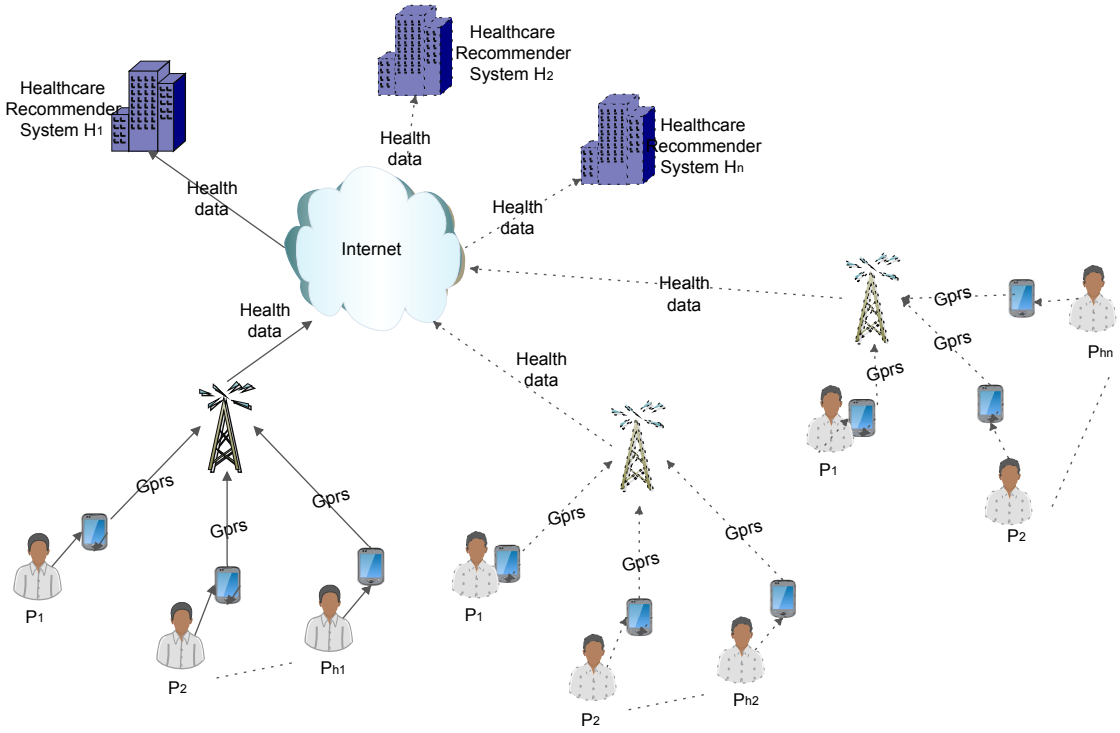


Figure 3.3: Healthcare recommender system architecture

recommender systems, CF is the most popular and conventional technique used to generate recommendations. The user-based CF technique tries to identify those users who have past

rating patterns similar to that of the current user and then use their ratings of other items to generate the predictions for the current user. In the item-based CF technique, similarities are found among items instead of the users. These techniques are more scalable and produce high-quality recommendations as compared to user-based CF techniques. In memory-based CF, entire data is used directly to generate predictions, whereas in model-based CF, all data is used to create model off-line and then that model is used to generate the prediction. Model-based CF techniques are more scalable than memory-based CF techniques because they have less online computation cost as compared to memory based CF techniques. Since a large number of ratings are used in the prediction generation process, model-based recommender systems are preferred over memory-based. Various techniques have been proposed for PPCF on distributed data. In schemes proposed for ADD by Yakut and Polat [162], [163], [164], default or fake ratings were added for preserving the privacy of data of different parties. Due to the addition of default or fake ratings, there can be some accuracy losses, and the predictions generated from those are not based on the actual ratings only. Further, these schemes were proposed for only two parties. If multiple parties can participate, then accuracy and coverage during the prediction generation process will be more. Since these models need to be regularly updated due to frequent addition of users in recommender systems, its computation time should be less. However, these schemes were based on model-based CF and used homomorphic encryption techniques for the model generation process, which has high computational cost. In our proposed work, a secured CF technique is presented for healthcare recommendation generation on ADD among multiple parties, which maintains privacy and requires less computation time.

3.2.2 System and security models

3.2.2.1 System model

In EMS-PPCF, patients are registered with the healthcare system using a secure cross-domain handshake scheme provided by He *et al.* [179]. The key establishment and key sharing mechanism for registration of patients to the healthcare system is done through Trusted Authority (TA) which generates the master key and private keys for healthcare centers. The healthcare centers register with TA and obtain their private keys and use them further to generate the private keys for their patients. A particular healthcare center collaborates with other healthcare centers using its private key. After registering, patients submit their ratings and feedback to the healthcare system, as shown in Figure 3.3.

EMS-PPCF for secure recommendation generation on arbitrarily distributed healthcare data is dependent upon the model-based item-item CF technique. In model-based item-item CF technique, similarities between item vectors are found, and these item-item similarities are used as a model to generate the prediction. Different similarity measures like cosine similarity, adjusted cosine similarity, and Pearson similarity can be used to measure item-item similarity weight. In the EMS-PPCF, the most commonly used similarity measure, *i.e.*, cosine similarity is used to measure item-item similarity. The cosine similarity weight $w_{i,j}$ between items i and j is defined as [180]:

$$w_{i,j} = \frac{\vec{i}.\vec{j}}{||\vec{i}||||\vec{j}||} \tag{3.1}$$

where $\vec{i}.\vec{j}$ represents dot product between item vector i and j and $||\vec{i}||$ represents the length of item vector i. After finding the item-item similarities, online prediction $Pred_{a,i}$ is generated

for N nearest neighbor of item i for user a as [163]:

$$Pred_{a,i} = \frac{\sum_{j \in N}(w_{i,j} \times r_{a,j})}{\sum_{j \in N} |w_{i,j}|} \tag{3.2}$$

where $r_{a,j}$ represents the rating of user a for item j.

The system model of EMS-PPCF is shown in Figure 3.4. Let $A_1, A_2 ... A_p$ represent the parties that want to collaborate for generating the recommendations. The complete dataset D is distributed arbitrarily between different parties (p), such as $D = D_{A_1} || D_{A_2} || D_{A_3} ... D_{A_p}$ where $D_{A_z} \in A_z$. Each party A_z has its item set I_{A_z} such that $I_{A_z} \in D_{A_z}$. They share their item set I_z with each other in the secured form to calculate item-item similarities between them. The item set I_{A_z} represents the rating matrix in the form of item vectors for party A_z. In EMS-PPCF, it is assumed that user id and item label are publicly known to each party. This helps in the synchronization between different parties. Only the ratings are shared in the secure form with each other. These item similarities act as an off-line model $Model_{A_z}$ for each party A_z which are used by them to provide the prediction to the user. As shown in Figure 3.4, when a party, say A_1, receives a query $q_{a,i}$ from the user a about item i, it generates the prediction $Pred_{a,i}$ using its off-line model $Model_{A_1}$ and sends it to the user a. Here, party A_1 is Master Party (MP) and all other parties $A_2 ... A_p$ are Collaborating Parties (CP).

Synchronization among multiple parties As shown in Figure 3.4, different parties collaborate with each other for generating the recommendations for the user a. The synchronization among computations performed by multiple parties is required for appropriate computations. The flowchart in Figure 3.5 shows steps for synchronization between two parties which can be extended to multiple parties. Levchin *et al.* [181] have provided synchronization technique using synchronization server which has been modified in the proposed work

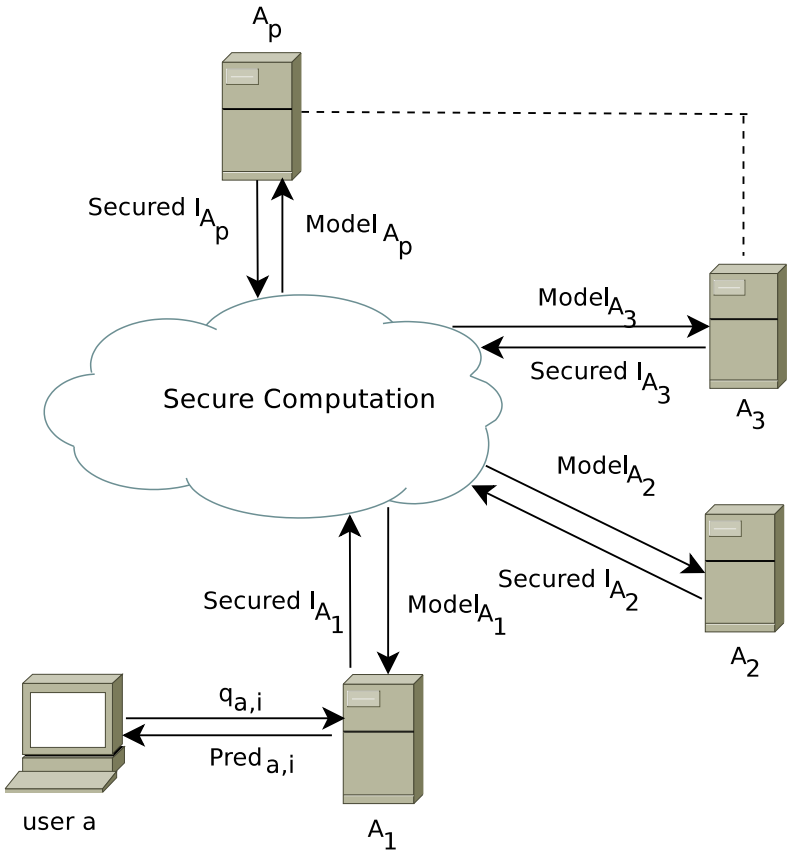


Figure 3.4: System model of EMS-PPCF

without synchronization server. Two parties A_i and A_j, check each other's identities and if they identify each other, then only they start transaction among themselves. The first party A_i synchronizes with party A_j and sends its transaction. The party A_j processes the transaction, synchronizes with party A_i, submits its value and closes the transaction.

Attack models During collaboration among multiple parties for the knowledge discovery process, various types of attacks can be performed for extracting the sensitive information of parties. Some of them are defined as:

1. A_1*: Collaborating parties collude with each other for capturing MP's data*: In this attack, $(p-1)$ collaborating parties collude with each other for extracting the MP's private data. Since MP is interacting with the user a, it contains the partial results from

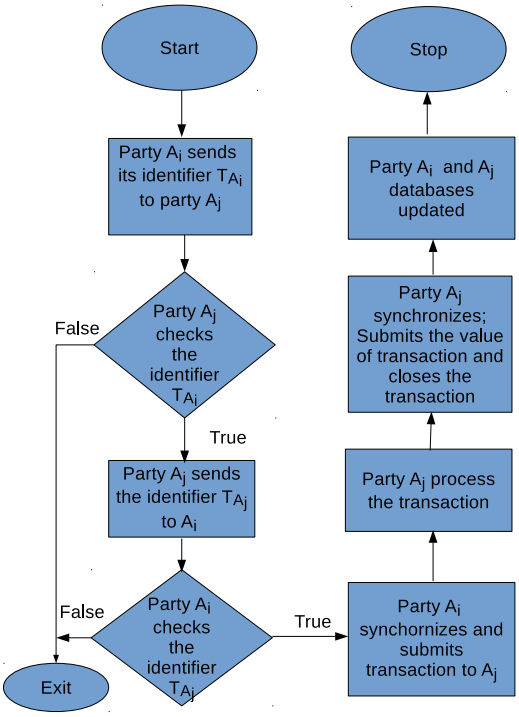


Figure 3.5: Flowchart for synchronization between two parties

all parties. MP combines its results to partial results and generates the prediction for the user a.

2. A_2*: Collaborating parties and MP collude with each other for capturing the target party's data*: In this attack, $(p-2)$ collaborating parties and master party collude with each other for extracting the private data of target party. Since the target party is a collaborating party, it does not have partial results of all other parties.

3.2.2.2 Security model

The different parties involved in the prediction generation process follow the semi-honest model of privacy, *i.e.*, different parties will follow the protocol honestly; however, they can try to guess the confidential information from intermediate and final results. In EMS-PPCF, ratings, information about rated items and non-rated items are considered as confidential

Table 3.1: Comparison of schemes for calculating secure sum

Scheme	Computation complexity	Communication complexity	Fault tolerance	Security	Topology required
Secret sharing	Low	High	Yes	High	No
Paillier	High	Low	Yes	High	No
Data perturbation	Low	Low	No	High	Yes

data of the user. Therefore, all protocols should be secure enough such that they will not reveal any confidential information while exchanging intermediate and final results among different parties.

3.2.3 Proposed work

EMS-PPCF on ADD consists of two phases: off-line model generation and online prediction generation. In the off-line model generation phase, the item-item similarity of the distributed data among different companies is calculated in privacy preserving way by using protocol I and II. Protocol I is used to calculating the dot product between item vectors securely using random masking and aggregation technique [182]. Protocol II is used to calculate the length of item vector securely using Paillier homomorphic encryption technique. Additive property of homomorphic encryption is used since only the secure sum is required to calculate the length of item vector distributed among multiple parties. The comparison of different methods for calculating the length of item vector is provided in Table 3.1. As shown in Table 3.1, the communication cost of the secret sharing scheme is very high as compared to homomorphic encryption whereas, for data perturbation scheme, fault tolerance is zero and parties need to be connected in some topology (*e.g.,* ring). Due to all these issues, the homomorphic encryption scheme has been used in our proposed work. After that, each party A_z has its own $Model_{A_z}$ consisting of similarity weights of items. In the online phase, the party receives a

query of prediction generation from the user a for an item i. The party will use the model generated off-line to provide the prediction to the user.

3.2.3.1 Off-line phase

In the off-line phase, item-item similarities are calculated and stored as an off-line model. For calculating item-item similarity weight using Eq. (3.1), the dot product between item vectors X and Y is calculated. Since data is arbitrarily distributed between parties, the dot product between vectors X and Y is represented as:

$X.Y = X^{A_1}.Y^{A_1} + X^{A_1}.Y^{A_2}...X^{A_1}.Y^{A_p}$

$+X^{A_2}.Y^{A_1} + X^{A_2}.Y^{A_2}...X^{A_2}.Y^{A_p}$

.

.

.

$+X^{A_p}.Y^{A_1} + X^{A_p}.Y^{A_2}...X^{A_p}.Y^{A_p}$

where X^{A_z} represents the component of vector X held by party A_z *i.e.* $X^{A_z} \in I_{A_z}$. The components of item vectors X and Y, owned by the same party, *i.e.*, $X^{A_1}.Y^{A_1}$, $X^{A_2}.Y^{A_2}$... $X^{A_p}.Y^{A_p}$, are calculated by them without any privacy measure. Whereas, the components of item vectors X and Y owned by different party *i.e.* $X^{A_1}.Y^{A_2}$, $X^{A_1}.Y^{A_3}$, $X^{A_1}.Y^{A_p}$ *etc.* are calculated by the collaboration of different parties with privacy measures. Protocol I is used to calculating these values securely.

Protocol I: Calculating scalar dot product $X.Y$ ***securely***

1. For each party A_z where $1 \leq z \leq p$

   (a) Calculate dot product between components X^{A_z} and Y^{A_z} owned by A_z, at its site as:

   $U^o_{A_z} = X^{A_z}.Y^{A_z}$

(b) Calculate summation of the dot product between cross wide components $U^c_{A_z}$ using PPSDPC as:

$U^c_{A_z} = PPSDPC(A_z)$

where PPSDPC is algorithm 1 named as Privacy Preserving Scalar Dot Product Computation.

(c) Calculate sum of $U^o_{A_z}$ and $U^c_{A_z}$ as:

$U_{A_z} = U^o_{A_z} + U^c_{A_z}$

end for

2. Calculate dot product $X.Y$ as: $X.Y = \sum_z U_{A_z}$

In algorithm PPSDPC, the party A_z masks its item vector X^{A_z} using the random numbers and then send those masked values to other collaborating parties, A_c, such that $1 \leq c \leq p$ and $c \neq z$. These parties perform the computations on these masked values using their corresponding item vector Y^{A_c} and send the summation of these values to the party A_z. The party A_z calculates the summation of the dot product between the cross wide components $U^c_{A_z}$ using the received values Ds_{A_c} from other collaborating p arties A_c. Then, as shown in step 2 of protocol I, the dot product $X\dot{Y}$ is calculated using these values.

For calculating item-item similarities, length of item vector X *i.e.,* $||X||$ needs to determined where $(X : 1,2..m)$. The protocol II is provided to calculate the length of item vector X securely which has been distributed among multiple parties.

Let $X = x_1, x_2, x_3 ... x_n$ then

$||X|| = \sqrt{X.X} = \sqrt{(x_1^2 + x_2^2 + ... + x_n^2)}$

Since X is distributed among different parties, $||X||$ is defined as:

$||X|| = \sqrt{X^{A_1}.X^{A_1} + X^{A_2}.X^{A_2} ... X^{A_p}.X^{A_p}} = \sqrt{(\sum_{z_1} {x_z^{A_1}}^2 + \sum_{z_2} {x_z^{A_2}}^2 + ... + \sum_{z_p} {x_z^{A_p}}^2)}$

where X^{A_y} is component of X held by party A_y and $\sum_{z_y} {x_z^{A_y}}^2$ represents the summation of

square of x_z over the total number of elements z_y of vector X^{A_y}. The Paillier homomorphic encryption system is used to calculate the length of vector X securely [183] by using the additive property of homomorphic encryption.

Protocol II: Calculate the length of item vector X.

To find the length of item vector X securely, homomorphic encryption technique is used. The party, who receives the query from a user, generates public and private keys. It keeps the private key with itself and shares the public key with all other parties using secure communication channel. The detailed process of calculating vector length $||X||$ is defined as:

For each item vector X such that $X : 1,2,...m$.

1. Party A_1 generates homomorphic encryption public and private keys pk and sk respectively and share the public key pk with all other parties.

2. Party A_1 computes

   $sum_{A_1} = {x_1^{A_1}}^2 + {x_2^{A_1}}^2 + ...{x_{z_1}^{A_1}}^2 = \sum_{z_1} {x_z^{A_1}}^2$

   $Esum_{A_1} = E_{pk}(sum_{A_1})$

3. Each party A_i $(1 < i \leq p)$, receives $Esum_{A_{i-1}}$ from party A_{i-1} and compute

   $sum_{A_i} = \sum_{z_i} {x_z^{A_i}}^2$

   $Esum_{A_i} = E_{pk}(sum_{A_i}) Esum_{A_{i-1}}$

4. Party A_p sends $Esum_{A_p}$ value to party A_1.

5. Party A_1 computes Sum and $||X||$ and sends those values to all parties.

   $Sum = D_{sk}(Esum_{A_p})$

   $||X|| = \sqrt{Sum}$

After execution of protocol I, each party A_z has dot product $X.Y$ between items X and Y, where $(X : 1,2,...m)$ and $(Y : 1,2...m)$. They also have vector length $||X||$ for each item. By

Algorithm 3.1 Privacy Preserving Scalar Dot Product Computation

procedure PPSDPC(A_z)
Step-1: Operations performed by A_z
Choose the two large prime numbers p_1 and p_2 such that $|p_1| = k_1$, $|p_2| = k_2$ and $k_2 > (nt^2+1)p_1^2$.
Set $S = 0$ and choose n positive random numbers $(u_1, u_2, u_3 ... u_n)$ such that $t\sum_{z=1}^{n} u_z < p_1 - tn$.
    for each element $x_y^{A_z} \in X^{A_z}$ **do**
        choose a random number ran_y such that $|ran_y p_2| = k_3$
        calculate $s_y = ran_y p_2 - u_y$
        if $x_y^{A_z} \neq 0$ **then**
            $C_y = x_y^{A_z} p_1 + u_y + ran_y p_2$
            $S = S + s_y$
        else
            $C_y = u_y + ran_y p_2$
            $S = S + s_y$
        end if
    end for
Keep p_2 and S confidential, and send ($p_1, C_1, C_2, C_3 ... C_y ... C_n$) to other collaborating parties A_c such that $1 \leq c \leq p$ and $c \neq z$
Step-2: Operations performed by each collaborating party A_c
    for each party A_c **do**
        for each element $y_q^{A_c} \in Y^{A_c}$ **do**
            if $y_q^{A_c} \neq 0$ **then**
                $D_q = y_q^{A_c} p_1 C_q$
            else
                $D_q = C_q$
            end if
        end for
        Compute Ds_{A_c} where $Ds_{A_c} = \sum_{q=1}^{n} D_q$
    end for
Send value of Ds_{A_c} of each party A_c to party A_z.
Step-3: Operations performed by party A_z
    for each Ds_{A_c} value **do**
        $I = \left(Ds_{A_c} + S\right) \mod p_2$
        $I_c = \frac{I - I \mod p_1^2}{p_1^2}$
    end for
$U_{A_z}^c = \sum_{A_c} I_c$
return $U_{A_z}^c$
end procedure

using that values, each party A_z creates a model $Model_{A_z}$ by calculating similarity weight between item X and Y using Eq. (3.1).

3.2.3.2 On-line phase

During the on-line phase, each party A_z generates a prediction for the user a and item i by using its model $Model_{A_z}$, created in the off-line phase. Protocol III is used to generate the prediction.

Protocol III: Generate the prediction for user a of item i

1. Based on ratings available for user a with party A_z, it calculates summation of product of user rating $r_{a,j}$ for item j and corresponding similarity weight $w_{i,j}$ between items i and j , and summation of similarity weight $w_{i,j}$ between items i and j for s number of the nearest neighbors of item i.

$$Sum_{rcs} = \sum_{s} w_{i,j} \times r_{a,j} \tag{3.3}$$

$$Sum_{cs} = \sum_{s} w_{i,j} \tag{3.4}$$

2. The party A_z calculates prediction $Pred_{a,i}$ by Eq. (3.2), Eq. (3.3) and Eq. (3.4) as:

$$Pred_{a,i} = \frac{Sum_{rcs}}{Sum_{cs}} \tag{3.5}$$

Collaboration among different parties Three protocols have been proposed in this work. In protocol I and II, different parties collaborate with each other for generating the off-line model. More precisely, in protocol I, the master party performs random masking on its item vectors and send them to all other parties. The other parties perform some computation on the masked values using their item vectors and send the aggregated results to the master

party. From these results, the master party calculates the dot product between item vectors. In protocol II, different parties calculate the length of item vector collaboratively in a secure manner using the homomorphic encryption technique. The master party generates the public and private keys and sends the public key to all other parties. The master party encrypts its share of the length of item vector and sends it to the next party. The next party encrypts its share of length of item vector and performs the multiplication of its encrypted value with the value received from the previous party. This process is repeated by all parties. In the end, the master party receives the encrypted value which is decrypted to obtain the length of item vector. The master party has dot product between item vectors and length of item vectors. By using these values, the master party calculates the similarity between item vectors. In this way, different parties collaborate with each other for computing the similarities between their item vectors.

3.2.4 Privacy analysis

Various protocols have been used in EMS-PPCF for generating prediction securely. The protocol I and protocol II are used to generate an off-line model, whereas protocol III is used to generate prediction online. The privacy analysis of each protocol is done separately in the next subsections:

3.2.4.1 Analysis of protocol I

In protocol I, first party A_z ($1 \leq z \leq p$) calculates the scalar dot product between components of item vectors X and Y owned by them, *i.e.,* $X^{A_z}.Y^{A_z}$ at its site. Since no data is leaked when operations are performed on their own data at their site, no privacy measure is required during these operations. The scalar dot product between cross-wide components X^{A_z} of party A_z and Y^{A_y} of other parties such that $1 \leq y \leq p$ and $y \neq z$ are calculated securely using PPSDPC. In

this algorithm, the party A_z masks its component X^{A_z} with random numbers and large prime numbers and sends it to other parties A_y. They perform some operations on masked data and their component Y^{A_y} as explained in PPSDPC, and then send the masked results to party A_z. It calculates the value of the scalar product from these masked values.

3.2.4.2 Privacy analysis of PPSDPC

As mentioned in algorithm PPSDPC, each element $x_y^{A_z} \in X^{A_z}$ is converted into $C_y = x_y^{A_z} p_1 + u_y + ran_y p_2$ or $C_y = u_y + ran_y p_2$ depending upon whether $x_y^{A_z} \neq 0$ or $x_y^{A_z} = 0$ respectively. Since every element $x_y^{A_z}$ is masked by random numbers u_y and ran_y and prime number p_2, it is difficult to find whether C_y is formed by $x_y^{A_z} p_1 + u_y + ran_y p_2$ or $u_y + ran_y p_2$. Every time different random numbers u_y and ran_y are generated, it is difficult to link C_y and $C_y^{'}$ (calculated using random numbers $u_y^{'}$ and $ran_y^{'}$). Therefore, each element $x_y^{A_z} \in X^{A_z}$ is secured. For other collaborating parties A_c, each element $y_q^{A_c} \in Y^{A_c}$ is converted into $D_q = y_q^{A_c} p_1 C_q$ or $D_q = C_q$ depending upon whether $y_q^{A_c} \neq 0$ or $y_q^{A_c} = 0$. This operation does not hide the value of $y_q^{A_c}$. But when summation is performed on all values of D_q of party A_c *i.e.* $Ds_{A_c} = \sum_{q=1}^{n} D_q$, it hides all values of D_q and hence $y_q^{A_c}$. The party A_z calculates the value of $U_{A_z}^c$ from the received value Ds_{A_c} of other collaborating parties. Hence, the calculation of $U_{A_z}^c$ is secured. By using the value $U_{A_z}^c$, party A_z calculates scalar product $X.Y$ between vector X and Y.

3.2.4.3 Analysis of protocol II

This protocol is used to the find length of item vector ($||X||$) securely, distributed among multiple parties, using Paillier homomorphic encryption technique. The party A_1 generates the public and private keys and shares the public key with all other parties. All parties A_i encrypt their value sum_{A_i}. The party A_1 sends its encrypted value to party A_2. The party

A_2 performs the addition on encrypted values and sends it to the next party A_3 and so on. In the end, party A_p sends addition on encrypted values to party A_1. Party A_1 decrypts the encrypted value, calculates the length of item vector and shares it with all other parties. As discussed in [183], ciphertext generated by the homomorphic encryption technique is semantically secure, *i.e.*, other parties are not able to infer any value from the encrypted text. Hence, protocol II is secured.

3.2.4.4 Analysis of protocol III

During the online prediction generation process, user a submits the query of prediction generation for item i to party A_z. The party A_z generates the prediction based on its model and sends the result to user a. Since other parties are not involved in this process, the protocol is secure.

3.2.5 Experimental results and analysis

The EMS-PPCF is experimentally evaluated for two case studies: Healthcare recommender system and Movies recommender system. The experiments are conducted on a PC having processor Intel core i7 and 8GB RAM. The scheme is implemented in Java with parameter settings for PPSDPC as $k_1 = 256$, $k_2 = 524$ and $k_3 = 1024$, and is evaluated on three parameters: accuracy, coverage, and performance. The performance of EMS-PPCF in terms of offline and online computation cost is given in Section 3.2.6.

Resilience against attack models Defense mechanism of the EMS-PPCF against attacks defined in Section 3.2.2.1 are:

i) A_1: In EMS-PPCF, the dot product between the item vectors of MP and CPs is calculated using algorithm 3.1. As explained in algorithm 3.1, the item vector of MP is masked

using random numbers and large prime numbers. So, colluding CPs will not get the item vector of MP, even if all CPs collude with each other. During the calculation of length of item vector using protocol II, the colluding parties will not know anything except the length of item vector of MP. Our main privacy constraint is to preserve ratings and information about whether the item is rated or non-rated, which is preserved in this case.

ii) A_2: In this attack model, $(p-2)$ CPs and MP collude with each other against a single collaborating party called target party. While calculating the dot product between item vectors of MP and target party using algorithm 3.1, the colluding parties will know only aggregated results of the target party. From these aggregate values, they will not be able to infer the information about whether the item is rated or non-rated and the exact rating of an item. As explained in the previous paragraph, during the calculation of the length of item vector, the colluding parties will know only the length of item vector of target party, nothing else.

Data tampering : Data tampering can be defined as an act of modifying the data intentionally. It can be done on data at rest or in transit. In the proposed work, semi-honest privacy model is considered, so the parties will not tamper their data deliberately. Even if they do, dishonest parties will not be able to guess the target party's private data as discussed in *Resilience against attack models*. But, the accuracy of the prediction generation process will get affected. For prevention against outside attackers, firewalls and access control mechanisms are used. The data is transmitted over secure channels and data is always in encrypted or masked form during transmission.

The accuracy of EMS-PPCF is calculated as Mean Absolute Error (MAE):

$$MAE = \frac{1}{h}\sum_{i=1}^{h}|r_i - P_i| \tag{3.6}$$

where h is the number of test ratings, r_i is actual rating and P_i is predicted rating.

Coverage of CF technique is defined as:

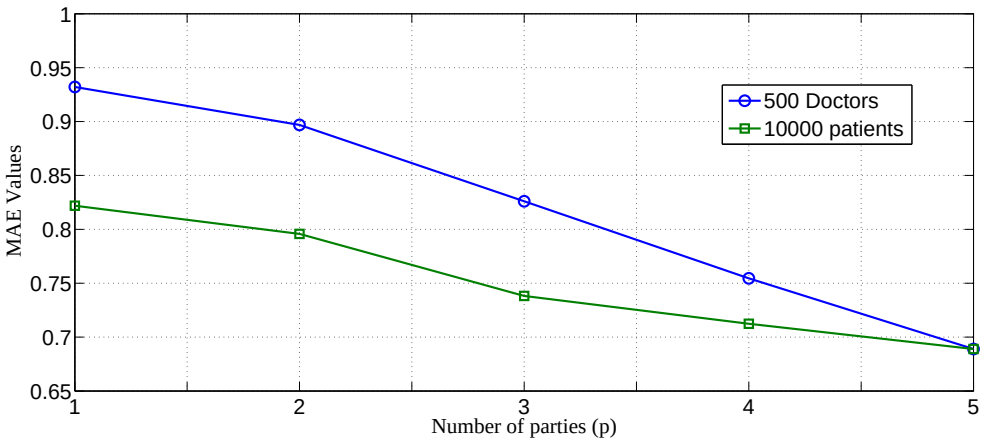
$$Cov = \frac{N_{pred}}{N_{ratings}} \tag{3.7}$$

where N_{pred} represents the count of actual predictions generated and $N_{ratings}$ represents the total count of the predictions that need to be generated.

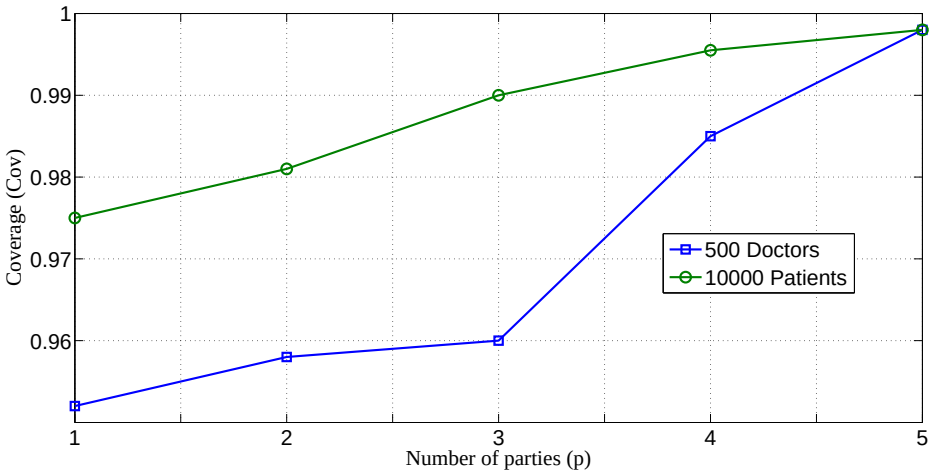
3.2.5.1 Healthcare recommender system

The simulated dataset for healthcare contains discrete ratings from 1 to 5 of 10000 patients for 500 doctors. The dataset is divided into training and testing data in 70:30 ratio respectively and 10-fold cross-validation scheme is used while evaluating the results. In Figure 3.6a, MAE values are shown for the healthcare dataset, where 10000 patient ratings for 500 doctors are divided among 5 parties. p represents the total number of parties collaborating which varies from 1 to 5 and $p = 1$ means that there is no collaboration and all parties are generating predictions individually. The $p = 2$ signifies that two parties are collaborating and likewise for $p = 3$, 4 *or* 5. Figure 3.6a depicts that MAE value is lower when all parties are collaborating, *i.e.,* $p = 5$. Lower the MAE value, more is the accuracy.

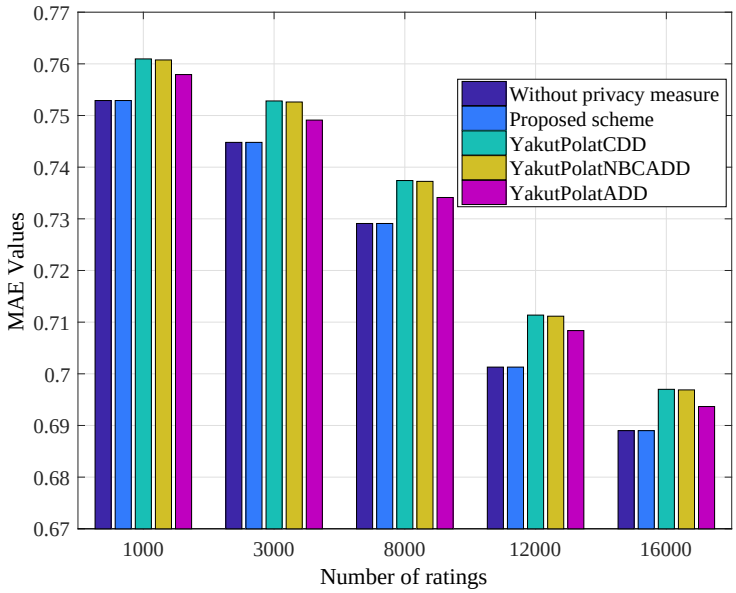
The techniques proposed by Yakut and Polat [162], [163], [164] have some accuracy losses due to the addition of default or fake ratings as shown in Table 3.3, but in the EMS-PPCF, there is no accuracy loss. The techniques Yakut and Polat [162], [163], [164] are referred as YakutPolatCDD, YakutPolatADD, and YakutPolatNBCADD, respectively in the figures of this chapter as well as in the figures of next chapter (Ch. 4). As shown in Figure 3.6c, the accuracy of our scheme for healthcare dataset is same as original integrated data having no privacy measure. For healthcare dataset, coverage is improved when parties are collaborating as compared to when no party is collaborating as shown in Figure 3.6b.



(a) MAE values for collaboration among parties



(b) Coverage



(c) Comparison of accuracy with other related schemes

Figure 3.6: Healthcare dataset

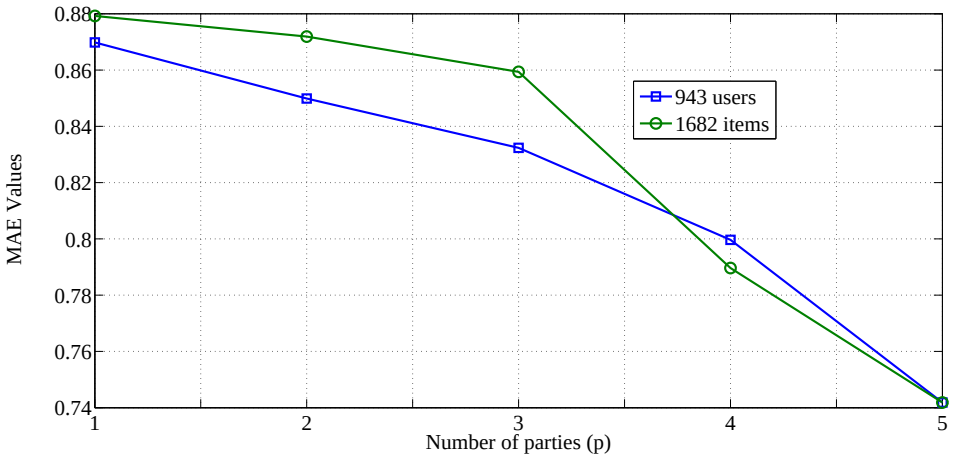
Table 3.2: Comparison of EMS-PPCF with related work in medical domain

Paper	Scenario	Data Distribution	Multi-party	Privacy preserving method	Security	Computation overhead
Katezbeisser and Petkovic [166]	Privacy preservation of patient's data while recommending doctor based on ratings stored in centralized server and forming the communities of patient's having similar health conditions.	Centralized	No	Cryptographic protocols	High	High
Hoens *et al.* [167]ACA	Privacy preservation of patient's data who is inquiring about doctor recommendation as well as who is providing the feedback about doctor using Anonymous Contribution Architecture (ACA)	Horizontal	Yes	Anonymizer	Low	Low
Hoens *et al.* [167]SPA	Privacy preservation of patient's data who is inquiring about doctor recommendation as well as who is providing the feedback about doctor using Secure Processing Architecture (SPA)	Horizontal	Yes	Secure multi-party techniques	High	High
EMS-PPCF	Different healthcare recommender systems collaborate with each other for improving their recommendations on ADD in secure manner	Arbitrary	Yes	Random masking, polynomial aggregation and homomorphic encryption	High	Low

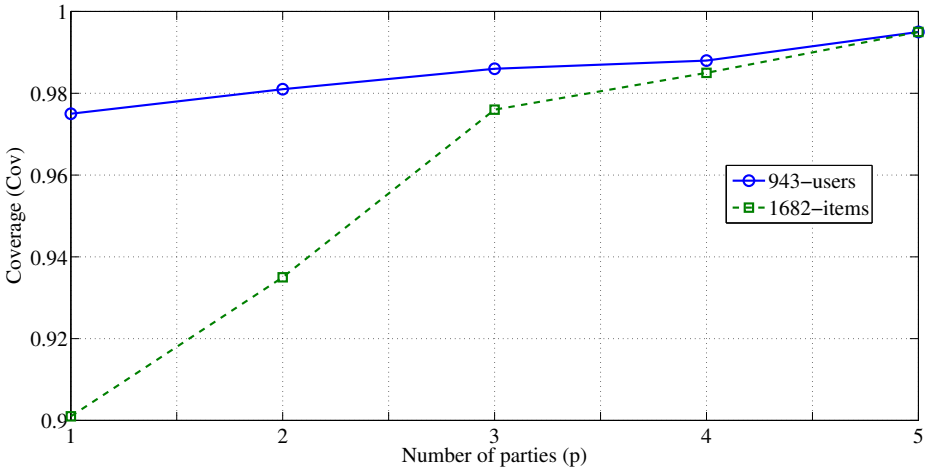
Here, 1000 test ratings are considered, and their training data is divided among 5 parties, and the value of p has the same significance as described above. As discussed earlier, in the medical domain two main studies have been discussed for privacy preserving health-care recommender system: Katezbeisser and Petkovic [166] which uses centralized data and Hoens *et al.* [167] which considers horizontally distributed data. The relative comparison of EMS-PPCF with above-mentioned schemes is given in Table 3.2.

3.2.5.2 Movies recommender system

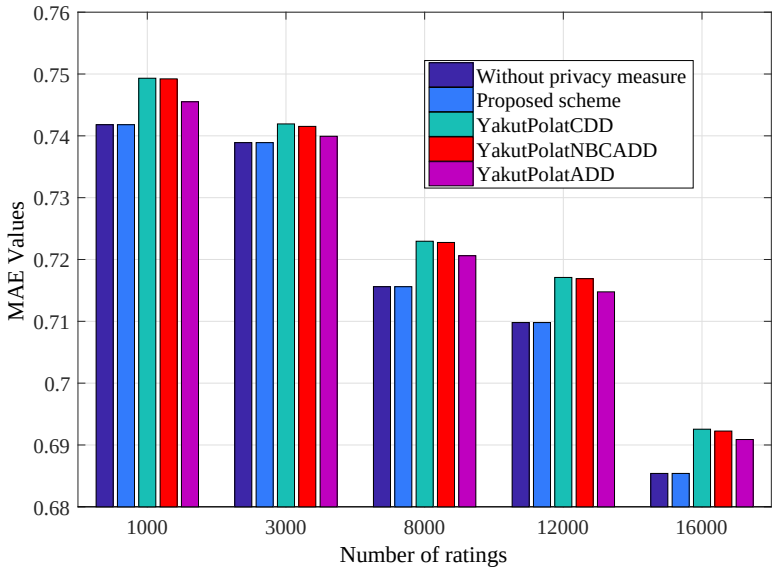
The MovieLens dataset [184] contains discrete ratings of 943 users for 1682 movies having a range from 1 to 5. It is divided into training and testing data in 70:30 ratio respectively and 10-fold cross validation scheme is used while evaluating the results. The MAE values calculated for ADD among different parties are shown in Figure 3.7a for 1000 test ratings. The value of p has the same significance as described in subsection 3.2.5.1. Figure 3.7a depicts that MAE value is lower when all parties are collaborating *i.e.,* $p = 5$ and lower the MAE value, more is the accuracy. As shown in Figure 3.7c, the accuracy of our scheme for Movielens dataset is same as original integrated data having no privacy measure, whereas other related techniques have some accuracy loss. Figure 3.7b indicates that coverage is



(a) MAE values for MovieLens



(b) Coverage



(c) Comparison of accuracy with other related schemes

Figure 3.7: MovieLens dataset

improved when parties are collaborating as compared to when no party is collaborating. Here, 1000 test ratings are considered, and their training data is divided among 5 parties as in subsection 3.2.5.1.

3.2.6 Performance evaluation

The performance of EMS-PPCF is evaluated in terms of the computation costs of off-line and online phases. The PPSDPC algorithm is used to calculate scalar dot product between two item vectors during off-line phase. So, the computation cost of calculating scalar dot product between all item vectors is $O(m^2n)$ where m represents the count of items and n represents a count of users involved in the prediction generation process. Next, protocol II is used to calculate the length of item vector $\vec{X}$ using homomorphic encryption technique. For $(HPE.key, HPE.enc, HPE.dec, HPE.mult, HPE.exp)$ homomorphic encryption system, computation cost is $O(mp)HPE.enc$, $O(mp)HPE.mul$, and $O(m)HPE.dec$, respectively where p represents the number of parties. The computation cost of online phase is same as that of centralized system which is $O(ns)$, where s is the number of similar items considered while generating the prediction for an item. The computation costs of other related schemes are given in Table 3.3. As shown in Table 3.3, theoretically computation cost of off-line as well as online phase of the EMS-PPCF is less than other related schemes. Our method is for multi-party, whereas other related schemes are for only two parties. Hence, to compare the computation costs of off-line phase of EMS-PPCF experimentally with other related schemes, we have considered only two parties. When $p = 2$, the computation overhead of EMS-PPCF for off-line phase is $O(m)HPE.enc$, $O(m)HPE.mul$, and $O(m)HPE.dec$. The benchmarks available at CRYPTO++ toolkit (https:// www.cryptopp.com/) are used to find the running time of cryptographic functions. It is observed that the model generation time of EMS-PPCF is less than other related schemes. For online phase, the scheme proposed by

Table 3.3: Comparison of EMS-PPCF with related techniques

Paper	Data Distribution	Multi-party	Privacy preserving method	Accuracy loss	Off-line Computation overhead	On-line computation overhead
Yakut and Polat [162]	Cross distribution	No	Homomorphic encryption, addition of default and fake votes	1%	$O(m^2n)HPE.enc$, $O(m^2n)HPE.dec$ and $O(m^2n)HPE.mul$	$O(nm)$
Yakut and Polat [163]	Arbitrary	No	Homomorphic encryption, addition of default and fake votes	1%	$O(m^2n)HPE.enc$, $O(m^2)HPE.mul$, $O(m^2n)HPE.exp$ and $O(m^2)HPE.dec$	$O(m_{ab})HPE.enc$, $O(m_{ab}w_{a,jb})HPE.dec$ and $O(m_{ab}w_{a,jb})HPE.exp$
Yakut and Polat [164]	Arbitrary	No	Homomorphic encryption, addition of default and fake votes	less than 1%	$O(mn)HPE.enc$, $O(m^2)HPE.dec$, $O(m^2n)HPE.mul$	$O(nm)$
EMS-PPCF	Arbitrary	Yes	Random masking and polynomial aggregation techniques	No accuracy loss	$O(mp)HPE.enc$, $O(mp)HPE.mul$ and $O(m)HPE.dec$	$O(sn)$

Note: m_{ab} and $w_{a,jb}$ represent no. of z-scores sent to other party and number of random pieces in which similarity weight $w_{a,j}$ is divided [163], respectively

Yakut and Polat [163] involves homomorphic encryption operations. So, its computational cost is very high as compared to other schemes. The other two schemes proposed by Yakut and Polat [164] [162] have computation cost $O(nm)$, whereas, for our scheme, it is $O(ns)$. Since $s < m$, the online computation cost of the EMS-PPCF is always less than other related schemes.

3.2.7 Discussion

The EMS-PPCF is secure, as it does not disclose any rating of items and information about whether an item is rated or non-rated. It has been experimentally demonstrated in the above section that accuracy and coverage of prediction generation are improved due to the collaboration of different parties, and off-line model generation computation cost, as well as online prediction generation computation cost of the proposed work, are less than other related schemes. The default and fake ratings have not been added for privacy preserving purpose, on the other hand, multi-party random masking, polynomial aggregation, and homomorphic encryption technique are used. So, there is no accuracy loss due to the introduction of privacy measures.

The main contributions of the proposed work in this chapter are:

i) The PPCF technique is proposed for healthcare recommendation generation on ADD among multiple parties.

ii) The three protocols have been proposed in our work. The similarity between items is generated using Protocol I, Protocol II calculates the length of item vectors and Protocol III generates the predictions.

iii) The EMS-PPCF is analyzed on different parameters: security, accuracy, coverage and performance and it has been experimentally demonstrated that accuracy and coverage of the EMS-PPCF is better when different parties collaborate.

In the next Chapter 4, another application of the PPDMB is discussed for the recommender system. A cloud-based multi-party privacy preserving classification scheme has been proposed for prediction generation on ADD among multiple e-commerce organizations.

Chapter 4

ClaMPP: A Cloud-based Multi-party Privacy Preserving Classification Scheme for Distributed Applications

This chapter proposes a Cloud-based Multi-party Privacy Preserving Classification Scheme, named as ClaMPP, which uses the NBC for recommendation generation on ADD among multiple parties.

4.1 Background

In the existing literature, the majority of recommendation based systems make predictions based on the available local data rating [185], [151],[186],[152], [155]. This can be considerably improved if ratings present with different e-commerce companies are taken into account as this can maximize the precision and enhance the reliability of recommendations. CF is a traditional and most commonly used method for recommendation generation based on the preference dataset of similar users. For accurate and reliable predictions, sufficient

data should be available, but newly established companies or existing companies launching a new product will not have required data and in such scenario prediction generation from a small amount of data by using CF technique may be difficult and inefficient. One of the solutions to this problem is the collaboration of data available with different companies in the form of ratings. If various companies (newly established and existing ones) collaborate with each other, reliable and accurate predictions can be generated. But for different companies their user ratings, rated items, and non-rated items are profitable assets, and they do not want to reveal it [162]. If these concerns of different companies are addressed, then they might collaborate with each other in the prediction generation process. There is a need of privacy preserving knowledge discovery technique to consolidate the data distributed among various companies. But the privacy comes with the cost of high computational overhead due to the use of cryptographic functions on massive datasets, which can be overcome by the use of cloud computing technologies. The large storage and computational power of cloud computing can be used to store massive datasets and perform high-intensive computing operations. With the help of cloud computing infrastructure, different parties belong to different regions can collaborate with each other for prediction generation as shown in Figure 4.1. Data required for prediction generation can be distributed horizontally, vertically or arbitrarily between different parties. Various techniques have been proposed for horizontal and vertical distribution. In arbitrary data distribution, ratings are distributed among various parties such that, for a particular item, some ratings are held by one party, some by another party and so on. The user-item rating matrix can be numeric or binary, depending on whether information stored is how much a user likes item or whether a user likes or dislike the item. For binary user-item rating matrix, NBC can be used to anticipate the likeness of the user for an item. To best of our knowledge, only two PPCF techniques have used NBC for prediction generation on the distributed dataset. Among them, one has considered horizontal and ver-

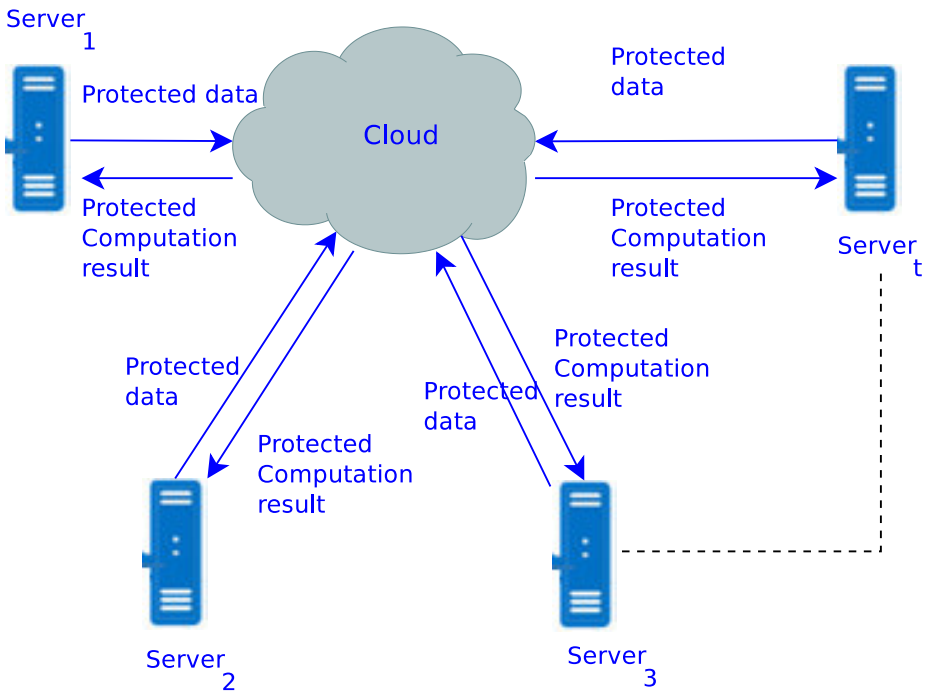


Figure 4.1: PPCF on cloud

tical data distribution among multiple parties, whereas other has considered ADD between two parties only and none of the techniques have used the cloud environment. Since ADD is generalized data distribution, the PPCF scheme should be designed which includes multiple parties collaborating with each other without data perturbation and minimum computation overhead.

4.1.1 NBC for recommendation generation

In some rating systems, users express their preferences about a product in binary form, *i.e.*, they either like a product or dislike it. When factors affecting the classification results are conditionally independent, NBC is considered as one of the best classifiers. It can directly calculate the probability of user's possible interests and calculation of similarity or distance metrics (*e.g., Pearson correlation, cosine similarity, and Euclidean distance*) is not required, whereas in other algorithms such as k-nearest neighbour there are many parameters which

need to be determined manually.

The class label of item l given its m feature values can be calculated using NBC as:

$$p(class_y|f_1, f_2...f_m) \propto p(class_y)\Pi_{i=1}^{m} p(f_i|class_y) \tag{4.1}$$

where $f_1, f_2...f_m$ represent the m features of item l, $class_y$ represents the class label of item l and the "naive" assumption is made that given the class label, features are independent of each other. Here, $p(class_y)$ represents the prior probability and $p(f_i|class_y)$ represents the conditional probability and both are calculated from training data.

In the naive Bayesian model for user-based CF, users correspond to features and entries of the rating matrix correspond to feature values. Similarly, for the naive Bayesian model for item-based CF, items correspond to features and entries of rating matrix correspond to feature values. In this proposed work, privacy preserving item-based CF naive Bayesian model is constructed for distributed data among multiple parties. The steps used for generating a prediction based on this model for centralized dataset without any privacy measure are as follows [187].

a) Let D be a user-item rating matrix recording the ratings provided by n users for m items. The rating vector $\vec{R_u}$ of user u is denoted as: $\vec{R_u} = (r_1, r_2...r_m)$, where r_i represents rating for item i.

b) Two classes of classification algorithm are: like and dislike, denoted by C_y, where y represents like or dislike. The NBC classifier assumes that given the class variable, features are independent of each other.

c) The probability of target item q belonging to class C_y, where y represents like or dislike value, is defined as:

$$p(C_y|r_1, r_2...r_m) \propto p(C_y)\Pi_{h=1}^{m} p(r_h|C_y) \tag{4.2}$$

where $p(C_y)$ is prior probability and $p(r_h|C_y)$ is conditional probability.

d) For any active user a, the algorithm predicts the class for item q, based on probabilities calculated for two classes. The item q is assigned to the class having high probability. Based on these steps, the PPCF scheme using NBC for ADD among multiple parties is proposed.

Three datasets MovieLens100K [184], MovieLens20M [184] and Jester [188] are taken in this work. For the MovieLens100K dataset, the number of feature sets is 1682 and the total feature values are 1 lakhs, for MovieLens20M, the number of feature sets is 27000 and the total feature values are 20 million and for Jester dataset, the number of feature sets is 140 and total feature values are 1.8 million. In this work, features are assumed to be independent of each other for all the three datasets.

4.1.2 Problem definition

Privacy in distributed computing means that collaborating parties should mine their distributed data in such a manner that data is not disclosed to any party and they should get final mining results only. If confidential data of one party is disclosed to other contending vendors, then contending parties might loss their business advantages, leading to financial losses. Also, legal constraints [189], [116] prevent the data collector from sharing the confidential information with other parties. In this work, the ratings, rated items and non-rated items are considered as confidential data of the organization. The privacy constraint of this work is that while exchanging intermediate or final results between different parties, any data should not be disclosed that help in inferring any confidential values. Further, semi-honest adversary model is considered, *i.e.*, collaborating parties use protocol specification, but they can use any information obtained as an intermediate or final result along with information

from the outside world to infer confidential values. The main goal of the proposed scheme in terms of privacy is to prevent collaborative parties from inferring confidential values or compromising the privacy constraints. Besides the privacy, the balance between other conflicting parameters: accuracy, and efficiency needs to be maintained. There should be negligible or less accuracy loss due to privacy measure. The communication, computation, and memory overheads that may occur due to privacy measure for on-line phase should be very less as compared to the off-line phase, such that predictions can be generated within a defined time limit. Hence, the problem definition of our work is to generate recommendations accurately on ADD among multiple parties within prescribed time while following the privacy constraints.

4.2 Proposed Work

The proposed privacy preserving NBC based prediction generation scheme considers more than two parties and is based on binary ratings. Firstly, NBC based prediction model is generated off-line by using the distributed data among multiple parties $P_1, P_2, P_3 ... P_t$ such that $t > 2$ and then different parties use the generated off-line model for providing the on-line predictions to their customers.

4.2.1 Off-line model generation

For generating off-line model, conditional probabilities $p(r_h|C_y)$ and prior probabilities $p(C_y)$ should be calculated for all item vectors. Hence, the parties $P_1, P_2, P_3 ... P_t$ collaborate with each other off-line using cloud to compute these probabilities while maintaining the privacy of their data. The privacy is ensured while calculating conditional probabilities by using an extended version of privacy preserving scalar product protocol based on random

masking and polynomial aggregation technique [182]. The homomorphic encryption technique is used for ensuring privacy while calculating prior probabilities. Additive property of homomorphic encryption is used since only the secure sum is required.

4.2.1.1 Calculation of conditional probabilities

Since the user can ask prediction for any item, a model is required which considers all possible ratings (1 and 0) for all items. Let's consider that a user u wants prediction for item l belonging to the set of items m. Assume that $\vec{K}$ and $\vec{L}$ represent the rating vectors for an item k and a target item l, respectively. Since data is arbitrarily partitioned among parties so $\vec{K} = (\vec{K_{P_1}}||\vec{K_{P_2}}||...\vec{K_{P_t}})$ and $\vec{L} = (\vec{L_{P_1}}||\vec{L_{P_2}}||...\vec{L_{P_t}})$, where $\vec{K_{P_i}}$ is rating vector for item k held by party P_i and $\vec{L_{P_i}}$ is rating vector of item l held by P_i. It is assumed that there are no overlapped ratings among different parties *i.e.,* $\vec{K_{P_1}} \cap \vec{K_{P_2}} \cap \vec{K_{P_3}}...\cap \vec{K_{P_t}} = \emptyset$ and $\vec{L_{P_1}} \cap \vec{L_{P_2}} \cap \vec{L_{P_3}}...\vec{L_{P_t}} = \emptyset$. For any rating r_k, likelihood probability is computed as:

$$P(r_k|C_y) = \frac{\mu}{\delta} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{t}(\mu_{P_iP_j})}{\sum_{i=1}^{t}\sum_{j=1}^{t}(\delta_{P_iP_j})} \tag{4.3}$$

where $\mu_{P_iP_j}$ is the number of users who rated items k and l which are held by party P_i and P_j respectively, and the ratings for item k are the same as r_k of user u; having class label C_y. Let $\delta_{P_iP_j}$ be the number of users who rated both k and l; have a class label C_y and ratings of k and l are held by party P_i and P_j respectively. Since the data is arbitrarily distributed among multiple parties, four possible cases are considered for calculating numerator and denominator. Further, since binary ratings are considered, parties need to evaluate $P(r_k|C_y)$ for four cases ($r_k = 0$, $r_k = 1$, $C_{dislike} = 0$ and $C_{like} = 1$). In Eq(4.3) for $i = j$, the components which are owned by same party $\mu_{P_1P_1}, \mu_{P_2P_2}...\mu_{P_tP_t}$ and $\delta_{P_1P_1}, \delta_{P_2P_2}...\delta_{P_tP_t}$ are calculated by respective parties $(P_1, P_2...P_t)$ independently. However, for $i \neq j$ the cross-wide components μ_{P_i,P_j} and $\delta_{P_iP_j}$ are calculated if different parties collaborate with each other. For calculating

μ and δ securely, two protocols, namely Protocol I and Protocol II have been proposed. The parties are divided into two sets MP and CP, where MP contains only one party and CP contains other parties who, by collaborating with MP, helps in calculating the value of μ.

Protocol I : Calculation of μ privately on distributed data

For each possible case ($r_k = 0, r_k = 1, C_{dislike} = 0$ and $C_{like} = 1$)

For each item pair (k,l | k=1,2,3...m; l=1,2,3...m; $k \neq l$)

Step 1: $P_1 \in MP$ and all other parties $(P_2...P_t) \in CP$

Step 2: If $r_k = 0$ then party P_1 inverts vector $\vec{K_{P_1}}$.

Step 3: If $C_{dislike} = 0$ then party P_1 inverts their vector $\vec{L_{P_1}}$ and all other collaborating parties $P_j \in CP$ invert their vector $\vec{L_{P_j}}$ respectively.

Step 4: Every party $P_i \in MP$ and $P_j \in CP$ fill empty cells of their sets $\vec{K_{P_i}}$, $\vec{L_{P_i}}$ and $\vec{L_{P_j}}$ with zero.

Step 5: Compute $\mu_{P_1} = \mu_{P_1P_1} + \mu_{P_1P_2} + \mu_{P_1P_3} + ...\mu_{P_1P_t}$ using procedure Privacy Preserving Conditional Probability Protocol (PPCPP(MP, CP)).

Step 6: For computing μ_{P_2} parties P_1 and P_2 interchange their roles and carry out the steps from 1 to 5. Similarly, the value of $\mu_{P_3}, \mu_{P_4}...\mu_{P_t}$ is calculated.

Step 7: Compute $\mu = \mu_{P_1} + \mu_{P_2} + ...\mu_{P_t} = (\mu_{P_1P_1} + \mu_{P_1P_2} + \mu_{P_1P_3} + ...\mu_{P_1P_t}) + (\mu_{P_2P_1} + \mu_{P_2P_2} + \mu_{P_2P_3} + ...\mu_{P_2P_t}) + ...(\mu_{P_tP_1} + \mu_{P_nP_2} + \mu_{P_nP_3} + ...\mu_{P_tP_t}) = \sum_{i=1}^{t}\sum_{j=1}^{t}(\mu_{P_iP_j})$.

By using Protocol I, the numerator μ of Eq. (4.3) is calculated securely. In step 1, the parties are assigned to sets MP and CP. In step 2 and step 3, depending upon the value of r_k and C_y, the related vectors are inverted. Then non-rated cells are filled with zero in respective vectors and the value of μ_{P_1} is calculated by using procedure PPCPP. In procedure PPCPP,

Algorithm 4.1 Privacy Preserving Conditional Probability Protocol

procedure PPCPP(MP,CP)
Step-1: Operations performed by $P_i \in MP$
Compute $\mu_{P_iP_i} = \vec{K_{P_i}}.\vec{L_{P_i}}$
Choose two large prime numbers pn_1 and pn_2 such that $|pn_1| = len_1$, $|pn_2| = len_2$ and $len_2 > (m+1)pn_1^2$.
Set $S = 0$ and choose m positive random numbers $\left(rn_1, rn_2, rn_3 ... rn_m\right)$ such that $\sum_{z=1}^{m} rn_z < pn_1 - m$.
    for each element $k_{P_{i_z}} \in \vec{K_{P_i}}$ **do**
        choose a random number ran_z such that $|ran_z pn_2| = len_3$
        calculate $s_z = ran_z pn_2 - rn_z$
        if $k_{P_{i_z}} = 1$ **then**
            $Q_z = pn_1 + rn_z + ran_z pn_2$
            $S = S + s_z$
        else
            $Q_z = rn_z + ran_z pn_2$
            $S = S + s_z$
        end if
    end for
Keep pn_2 and S confidential, and send ($pn_1, Q_1, Q_2, Q_3 ... Q_z ... Q_m$) to all parties $P_j \in CP$
Step-2: Operations performed by all parties $P_j \in CP$
    for each element $P_j \in CP$ **do**
        for each element $l_{P_{j_z}} \in \vec{L_{P_j}}$ **do**
            if $l_{P_{j_z}} = 1$ **then**
                $D_z = pn_1 Q_z$
            else
                $D_z = Q_z$
            end if
        end for
        Compute D_{P_j} where $D_{P_j} = \sum_{z=1}^{m} D_z$
        Send D_{P_j} to party P_i
    end for
Step-3: Operations performed by party P_i
    for each value of D_{P_j} received from collaborating parties **do**
        $E = \left(D_{P_j} + S\right) \mod pn_2$
        $\mu_{P_iP_j} = \frac{E - E \mod pn_1^2}{pn_1^2}$

    end for
Compute $\mu_{P_i} = \sum_j \mu_{P_iP_j} + \mu_{P_iP_i}$
return μ_{P_i}
end procedure

the dot product between the vectors $\vec{K_{P_1}}$ and $\vec{L_{P_1}}$ is computed as shown in line 3. After that the party P_i masks its vector $\vec{K_{P_1}}$ with the random numbers and send those masked values to other collaborating parties CP. These parties perform the computations on these masked values using their item vector $\vec{L_{P_j}}$ and compute the value of D_{P_j}. Next, they send the value of D_{P_j} to party P_1. Using these values the party P_1 calculate the μ_{P_1} as shown in step 3. Similarly, the parties calculate $\mu_{P_2}, \mu_{P_3} ... \mu_{P_t}$ by interchanging their roles and in the end μ is computed. For calculating conditional probabilities given in Eq(4.3) parties need to calculate the denominator too. The denominator of Eq(4.3) denotes count of users who have rated item k and l and have class label C_y. Protocol II is used to evaluate the denominator δ securely.

Protocol II: Calculation of δ privately on distributed data

For each possible case ($r_k = 0, r_k = 1, C_{dislike} = 0$ and $C_{like} = 1$)

For each item pair (k,l | k=1,2,3...m; l=1,2,3...m;$k \neq l$)

Step 1: $P_1 \in MP$ and all other parties $(P_2 ... P_t) \in CP$.

Step 2: P_1 replaces each rating in vector $\vec{K_{P_1}}$ by 1.

Step 3: If $C_{dislike} = 0$, then party P_1 inverts its vector $\vec{L_{P_1}}$ and all other collaborating parties $P_j \in CP$ invert their vectors $\vec{L_{P_j}}$ respectively, *i.e.,* 1s to 0s and 0s to 1s.

Step 4: Every party $P_j \in CP$ and $P_i \in MP$ fill empty cells of their vectors $\vec{K_{P_i}}$, $\vec{L_{P_i}}$ and $\vec{L_{P_j}}$ with zero.

Step 5: Compute $\delta_{P_1} = \delta_{P_1 P_1} + \delta_{P_1 P_2} + \delta_{P_1 P_3} + ... \delta_{P_1 P_t}$ using procedure PPCPP. (*Note*: Replace μ with δ in PPCPP).

Step 6: For computing δ_{P_2} parties P_1 and P_2 interchange their roles and carry out the same steps from 1 to 5. Similarly the value of $\delta_{P_3}, \delta_{P_4} ... \delta_{P_t}$ will be calculated.

Step 7: Compute $\delta = \delta_{P_1} + \delta_{P_2} + ...\delta_{P_t} = (\delta_{P_1P_1} + \delta_{P_1P_2} + \delta_{P_1P_3} + ...\delta_{P_1P_t}) + (\delta_{P_2P_1} + \delta_{P_2P_2} + \delta_{P_2P_3} + ...\delta_{P_2P_t}) + ...(\delta_{P_tP_1} + \delta_{P_tP_2} + \delta_{P_tP_3} + ...\delta_{P_tP_t}) = \sum_{i=1}^{t}\sum_{j=1}^{t}(\delta_{P_iP_j})$.

Protocol II is almost similar to Protocol I. In first step parties are assigned to sets MP and CP. In the next step, all ratings in $\vec{K_{P_1}}$ are converted to 1. Next, ratings of $\vec{L_{P_1}}$ and for all $\vec{L_{P_j}}$ where $P_j \in CP$ are inverted if $C_{dislike} = 0$. Then similar steps are followed as in Protocol I to calculate δ. Since all parties have the value of numerator μ and denominator δ, they calculate the conditional probabilities by using Eq(4.3). The flow diagrams for calculating the conditional probabilities using protocols I and II are provided in Figure 4.2.

4.2.1.2 Priori probabilities

Priori probability is calculated by finding the probability of selecting a rating having class C_y from all ratings of item vector $\vec{L}$. Since the data is arbitrarily distributed among multiple parties, it can be calculated as follows:

$$P(C_y) = \frac{C_{y,\vec{L}}}{|\vec{L}|} = \sum_{i=1}^{t} \frac{C_{y,\vec{L_{P_i}}}}{|\vec{L_{P_i}}|} \tag{4.4}$$

where $C_{y,\vec{L}}$ is the count of ratings in item vector $\vec{L}$ having class C_y and $C_{y,\vec{L_{P_i}}}$ is the count of ratings in item vector $\vec{L_{P_i}}$ of party P_i having class C_y. For completing the off-line model, parties need the prior probabilities. Protocol III is used for calculating prior probabilities in a privacy preserving manner by using Paillier homomorphic encryption [183]. The additive property of the homomorphic encryption system has been used in the proposed protocol. Let $E_{pk}(.)$ is an encryption function with public key pk and $D_{sk}(.)$ is decryption function with secret key sk of Paillier homomorphic encryption. Then, according to the additive property of homomorphic encryption, if $E_{pk}(b_1)$ and $E_{sk}(b_2)$ are encryption of messages b_1 and b_2, encryption of sum of b_1 and b_2 is: $E_{pk}(b_1 + b_2) = E_{pk}(b_1)E_{pk}(b_2)$ and sum of b_1 and b_2 is:

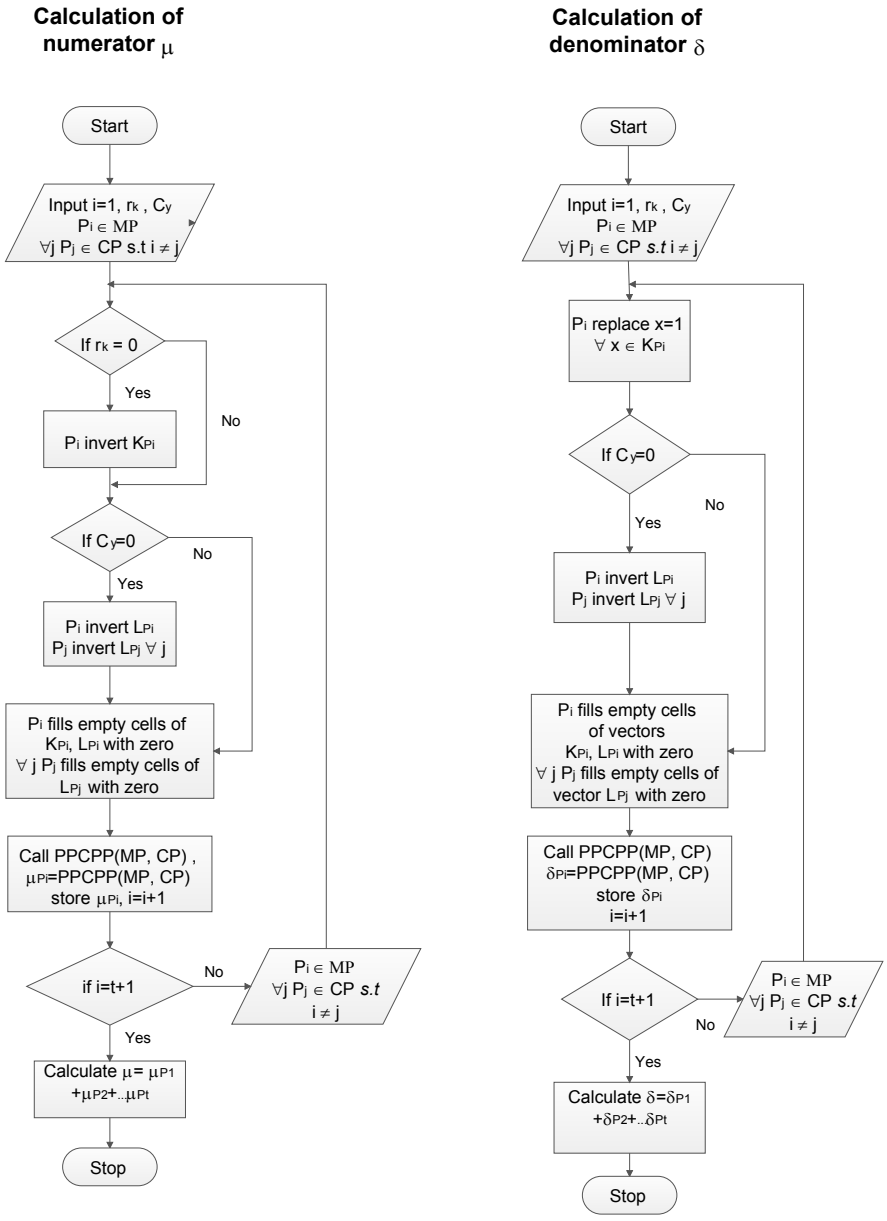


Figure 4.2: Flowchart for calculation of μ and δ

$b_1 + b_2 = D_{sk}(E_{pk}(b_1)E_{pk}(b_2))$. The flow diagram for the calculation of prior probabilities is depicted in Figure 4.3.

Protocol III: Calculation of prior probabilities

Step : 1 The party P_1 generates homomorphic encryption public and secret keys *pk* and *sk* respectively, and share the public key *pk* with all parties.

Step : 2 Party P_1 compute $sumL_1 = E_{pk}(\vec{L_{P_1}})$ and $sumC_{y_1} = E_{pk}(C_y, \vec{L_{P_1}})$

Step : 3 Each party P_i where $n \geq i > 1$, receive $sumL_{i-1}$ and $sumC_{y_{i-1}}$ from party P_{i-1} and compute

$sumL_i = sumL_{i-1} E_{pk}(\vec{L_{P_{i-1}}})$

$sumC_{y_i} = sumC_{y_{i-1}} E_{pk}(C_y, \vec{L_{P_{i-1}}})$

Step : 4 Party P_n sends $sumL_n$ and $sumC_{y_n}$ to party P_1.

Step : 5 Party P_1 compute $D_{sk}(sumL_n)$ and $D_{sk}(sumC_{y_n})$ and send these values to all parties.

The item l can be any item from m items, so prior probabilities should be calculated for all m items. At the end of the protocol, all parties have prior probabilities for all items.

4.2.2 On-line model: Prediction generation

The recommendation for user u for item l is generated by using the off-line model discussed in the previous section. In this process, one party acts as an Active Party (AP) which interacts with user u. Let P_1 is AP and u's ratings are distributed among different parties and m_{P_i} denotes the number of ratings held by party P_i. Thus, AP has to calculate probabilities for both classes with the collaboration of other parties and assign it to class having higher probability. OP, the collection of other parties is generated in a privacy preserving manner by Protocol IV. The flow diagram for the same is provided in Figure 4.3.

Protocol IV: Prediction Generation

Step:1 Party AP informs other parties $P_j \in OP \forall j$ through cloud about l.

Step:2 Party AP computes $\prod_{i=1}^{m_{AP}} p(r_i|C_j)$ based on u's ratings held by it.

Step:3 The other parties $\forall j P_j \in OP$ compute $p(C_j)\prod_{i=1}^{m_{P_j}} p(r_i|C_y)$ based on user u's ratings held by them respectively.

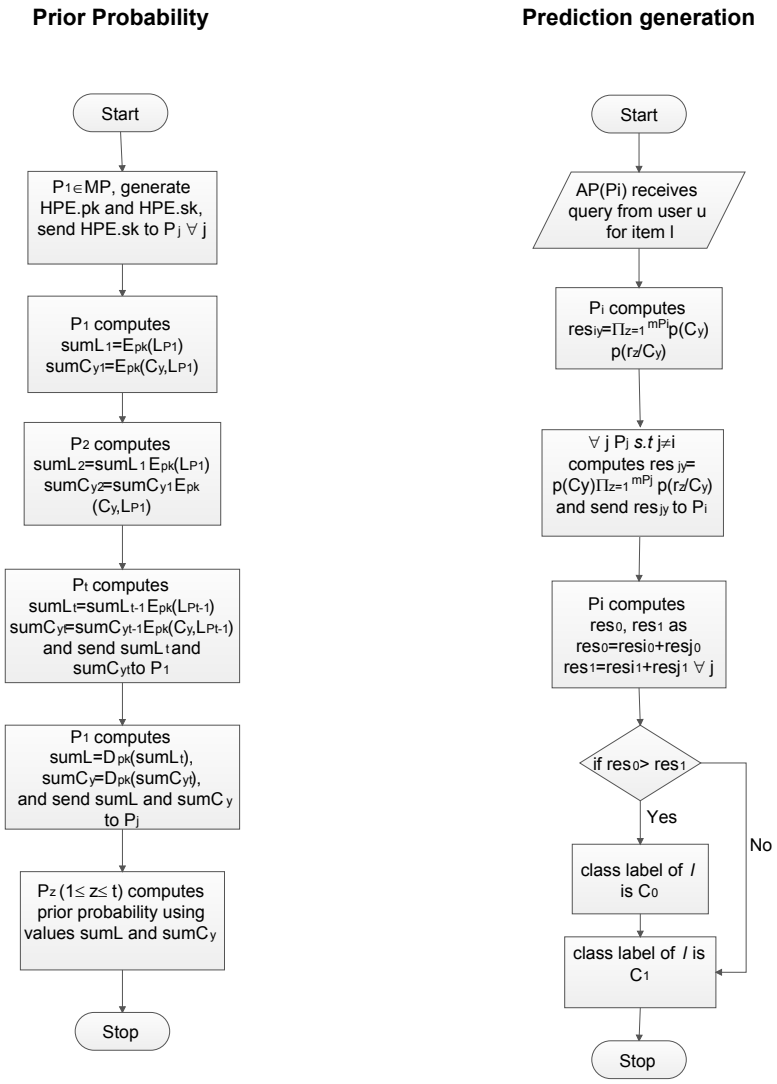


Figure 4.3: Flowchart for calculation of prior probabilities and prediction generation

Step:4 All other parties send calculated values of both classes to *AP* through the cloud.

Step:5 *AP* determines the final probabilities for both classes and compares them to find the class label of item l.

Step:6 It assigns the item l to class having higher probability.

4.3 Privacy Analysis

In the ClaMPP, data transfer is performed in off-line and on-line phases. Protocol I and Protocol II of the off-line phase are used to calculate μ and δ securely to determine conditional probabilities as described in Eq.(4.3). The security of Protocol I and Protocol II depends on PPCPP algorithm, information leakage from self-crosswise components and final computed value of numerator μ and denominator δ. The prior probabilities are calculated using Protocol III, whose security depends on homomorphic encryption system used. Protocol IV is used for generating the predictions by using the offline model. Its security depends on the information leakage while generating predictions from calculated values received from other parties. Since the semi-honest model is considered, each party will follow the defined protocol and try to guess information from intermediate and final results it received from other parties. The privacy analysis of each protocol has been described below.

i) Protocol I and II

PPCPP : As described in algorithm 4.1, for every element $k_{P_{i_z}} \in \vec{K_{P_i}}$, $Q_z = pn_1 + rn_z + ran_z pn_2$ or $Q_z = rn_z + ran_z pn_2$ depending upon whether $k_{P_{i_z}} = 1$ or $k_{P_{i_z}} = 0$ respectively. Since every value of $k_{P_{i_z}}$ is masked by random numbers rn_z and ran_z, and large prime number pn_2, it is almost impossible to find whether Q_z is formed by $pn_1 + rn_z + ran_z pn_2$ or $rn_z + ran_z pn_2$. Different random numbers (rn_z, ran_z) are used in every iteration so Q_z and Q'_z are not linked. Therefore, each $k_{P_{i_z}} \in \vec{K_{P_1}}$ is secure during computation of μ. For other collaborating parties P_j such that $P_j \in CP$, for each $l_{P_{j_z}} \in \vec{L_{P_j}}$, we have $D_z = pn_1 Q_z$ or $D_z = Q_z$ when $l_{P_{j_z}} = 1$ or $l_{P_{j_z}} = 0$. This operation will not hide the value of $l_{P_{j_z}}$. However, since summation is performed on all values of D_z i.e $D_{P_j} = \sum_{z=1}^{m} D_z$, it will hide the operation on each D_z and hence the value of each $l_{P_{j_z}} \in \vec{L_{P_j}}$. By using the value of D_{P_j} and pn_2, μ_{P_i,P_j} is computed. So the computation of μ_{P_i,P_j} is secured. The

same computations are done to calculate δ_{P_i,P_j}, hence, it is secured.

Self-Crosswise Component: The terms $\mu_{P_iP_j}$ and $\delta_{P_iP_j}$ such that $i \neq j \forall i \forall j$ are called self-crosswise component. Let us examine the security of self-crosswise component in terms of party P_1 and value of $\mu_{P_1P_2}$. If P_1 knows the value of $\mu_{P_1P_2}$, then the probability of guessing the users who rated l as C_y held by P_2 is 1 out of $\binom{|C_{r_{k_{P_1}},\vec{K_{P_1}}}|}{\mu_{P_1P_2}}$. Similarly, if P_1 knows $\delta_{P_1P_2}$, it can guess those users who rated l as C_y held by P_2 with probability of 1 out of $\binom{\vec{K_{P_1}}|}{\delta_{P_1P_2}}$. Since the value of $\binom{|C_{r_{k_{P_1}},\vec{K_{P_1}}}|}{\mu_{P_1P_2}} < \binom{|\vec{K_{P_1}}|}{\delta_{P_1P_2}}$ the security of Protocol I and II depends on value of $|C_{r_{k_{P_1}},\vec{K_{P_1}}}|$. The inference probability value for both protocols is $1/\binom{|C_{r_{k_{P_1}},\vec{K_{P_1}}}|}{\mu_{P_1P_2}}$. Since protocol is symmetric, this inference probability value applies to other parties as well.

Final μ and δ values: Let us examine the security in terms of party P_1. As given in Eq.4.3, for calculating μ, the other parties $P_j \in CP \forall j$ need to exchange the value of μ_{P_j}. It is difficult to derive information from exchanging aggregates, since the self-computable values are known by other parties P_j only, *i.e.* $\mu_{P_jP_j}$ are added to collaborative-computable values *i.e.* $\mu_{P_jP_i}$. Similarly, P_1 cannot infer useful information from δ. Same applies to other parties too. Hence the proposed PPCPP algorithm is secured.

ii) Protocol III

Security of prior probability calculation depends on Paillier homomorphic encryption scheme used for calculating the secure sum. In this protocol, party P_1 shares the public key with all other parties. All parties encrypt their values with the public key. Party P_1 sends its encrypted value to P_2, which performs the addition on encrypted values and sends it to the next party P_3 and so on. Except for the party P_1, none of the parties can decrypt the value because they do not have the secret key. In the end party P_1 gets the encrypted value from party P_n. Party P_1 decrypts the value, calculates the prior

Table 4.1: Datasets

Dataset	Category	Size	Ratings	Density	Range	Type
MovieLens 100K	Movie	943 x 1682	1 L	6.30%	[1,5]	Discrete
MovieLens 20M	Movie	138000 x 27000	20 M		[1,5]	Discrete
Jester	Joke	59132 x 140	1.8 M	21.28%	[-10,+10]	Real

probability and shares it with other parties.

iii) Protocol IV

During the online prediction generation process, the AP gets two aggregate values for each class from other collaborating parties. Through these aggregate values, *AP* cannot determine the conditional probabilities used for calculating these aggregated values. These aggregate values are multiples of the unknown number of conditional probabilities. So, it is difficult to determine the conditional probabilities from these aggregate values.

4.4 Experimental Results and Analysis

ClaMPP is analyzed experimentally to evaluate its prediction quality using three publicly available datasets: Movielens Public (MLP), Movielens 20M (MLM) [184] and Jester [188] datasets. Since there is no accumulated binary-ratings based dataset for recommendation generation purposes, so the dataset containing numeric ratings has been used. Table 4.1 provides the properties of these datasets.

These datasets are converted into the binary form using the method suggested by [190] and as done in [156], [191] and [192]. Each dataset is divided into training and testing datasets in the percentage of 70 and 30 respectively, and 10-fold cross-validation scheme is used for evaluating the prediction quality. Two metrics are considered: Classification Accuracy (CA) and F-measure (F1). CA is computed by finding the percentage of correctly

predicted items to total predictions generated. Let P_u be a set of x recommendations generated for the user u then CA is:

$$CA = \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{x} \tag{4.5}$$

where $p_{u,l}$ is prediction generated for user u, for item l and $r_{u,l}$ is an original prediction for user u of item l.

F1 is computed as:

$$F1 = \frac{2\psi\omega}{\psi + \omega} \tag{4.6}$$

where ψ represents precision, which is the proportion of relevant recommended items to the total number of recommendations generated and ω represents the recall which is the proportion of relevant recommended items to the total number of relevant items. ψ and ω are calculated as:

$$\psi = \frac{1}{U} \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{x} \tag{4.7}$$

$$\omega = \frac{1}{U} \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{(l \in P_u | r_{u,l} = 1) + (l \in P_u | r_{u,l} = 0)} \tag{4.8}$$

In our experimental work, initially analysis is done to verify that ADD-collaboration improves the prediction quality of different parties. The CA and F1 values are calculated for different parties (t) and different datasets as shown in Figure 4.4. The recommendation data are distributed arbitrarily among 7 parties. For $t = 1$, CA is calculated when a single party generates recommendations as shown in Table 4.2. All parties generate predictions individually and the average of their respective values of CA are considered. For $t = 3$, CA is depicted when three parties generate recommendations collaboratively. All possible combinations of 3 parties ($t = 3$) from a group of 7 parties have been considered for generating recommendations collaboratively. The average of their respective CA values are considered

Table 4.2: CA values for $t = 1$

S.no	Party	CA		
		MLP	MLM	Jester
1	1	0.5911	0.6122	0.6534
2	2	0.586	0.5987	0.6211
3	3	0.5521	0.5993	0.5989
4	4	0.6129	0.6212	0.6434
5	5	0.5744	0.6116	0.6061
6	6	0.5779	0.5829	0.6326
7	7	0.5801	0.6129	0.5991
Average		0.5821	0.6055	0.6221

(shown in Table 4.3) and similarly the CA for $t = 5$ and $t = 7$ is calculated, shown in Table 4.4 and Table 4.5, respectively. In the similar manner, values of F1 are calculated for $t = 1$, $t = 3$, $t = 5$ and $t = 7$. As shown in Figure 4.4, when the number of collaborating parties (t) increases, the CA and F1 values increase; indicating that prediction quality is increased due to collaboration. It is required that there should be negligible or less accuracy loss due to the introduction of privacy measures while calculating the results on integrated data. In ClaMPP, no operations are performed during an on-line phase which affects the prediction quality. In the off-line phase, while calculating likelihood probabilities, PPCPP algorithm is used which does not affect the output accuracy and during prior probability computation, homomorphic encryption based operations are used, which also do not affect the accuracy. To verify that there is no accuracy loss in ClaMPP, experimental trials are carried out using the same MLP data set. The outcomes obtained from originally combined data without any privacy measure and with ClaMPP are shown in Figure 4.5, which shows that there is no accuracy loss due to privacy measure.

Table 4.3: CA values for $t = 3$

S.no	Parties	CA		
		MLP	MLM	Jester
1	123	0.5921	0.6689	0.6713
2	124	0.6125	0.6965	0.7015
3	125	0.5945	0.6741	0.6739
4	126	0.6013	0.6669	0.6589
5	127	0.6112	0.6681	0.6845
6	134	0.6243	0.6764	0.6775
7	135	0.5986	0.6697	0.6718
8	136	0.6012	0.6724	0.6771
9	137	0.6039	0.6636	0.6961
10	145	0.6231	0.6939	0.7039
11	146	0.6235	0.6976	0.7041
12	147	0.6289	0.70977	0.7089
13	156	0.5932	0.6523	0.6431
14	157	0.5987	0.6631	0.6751
15	167	0.5976	0.6634	0.6671
16	234	0.6189	0.6801	0.6983
17	235	0.5963	0.6634	0.6657
18	236	0.5879	0.6751	0.6845
19	237	0.5865	0.6631	0.6683
20	245	0.6231	0.6821	0.6883
21	246	0.6236	0.6829	0.6884
22	247	0.6138	0.6793	0.6779
23	256	0.5971	0.6611	0.6673
24	257	0.5901	0.6701	0.6775
25	267	0.5863	0.6687	0.6771
26	345	0.6213	0.6742	0.6885
27	346	0.6143	0.6624	0.6971
28	347	0.6362	0.6719	0.6898
29	356	0.5901	0.6625	0.6641
30	357	0.5824	0.6631	0.6671
31	367	0.5901	0.6601	0.6682
32	456	0.6139	0.6945	0.6985
33	457	0.6431	0.7058	0.6979
34	467	0.6275	0.7018	0.6921
35	567	0.5949	0.6941	0.7012
Average		0.6069	0.6758	0.6821

Table 4.4: CA values for $t = 5$

S.no	Parties	CA		
		MLP	MLM	Jester
1	12345	0.6709	0.7556	0.7541
2	12346	0.6758	0.6893	0.7706
3	12347	0.6824	0.6971	0.7428
4	12356	0.6721	0.6884	0.6941
5	12357	0.6818	0.6979	0.7248
6	12367	0.6763	0.6983	0.6895
7	12456	0.6793	0.7158	0.7686
8	12457	0.6877	0.7444	0.7503
9	12467	0.7057	0.7275	0.7816
10	12567	0.6619	0.7171	0.6884
11	13456	0.6639	0.7189	0.7731
12	13457	0.6927	0.7289	0.7482
13	13467	0.6934	0.7098	0.7764
14	13567	0.6825	0.7085	0.6865
15	14567	0.6849	0.7167	0.7753
16	23456	0.6713	0.7082	0.6959
17	23457	0.6983	0.7387	0.7039
18	23467	0.6815	0.6993	0.7089
19	23567	0.6701	0.6891	0.7015
20	24567	0.6955	0.7187	0.7298
21	34567	0.6764	0.6895	0.7104
Average		0.6812	0.7123	0.7321

Table 4.5: CA values for $t = 7$

S.no	Parties	CA		
		MLP	MLM	Jester
1	1234567	0.7231	0.7435	0.7437

4.4.1 Performance

The recommendation generation over distributed data with privacy measures can have various overhead. Table 4.6 highlights overheads with respect to storage, communication, and computation costs.

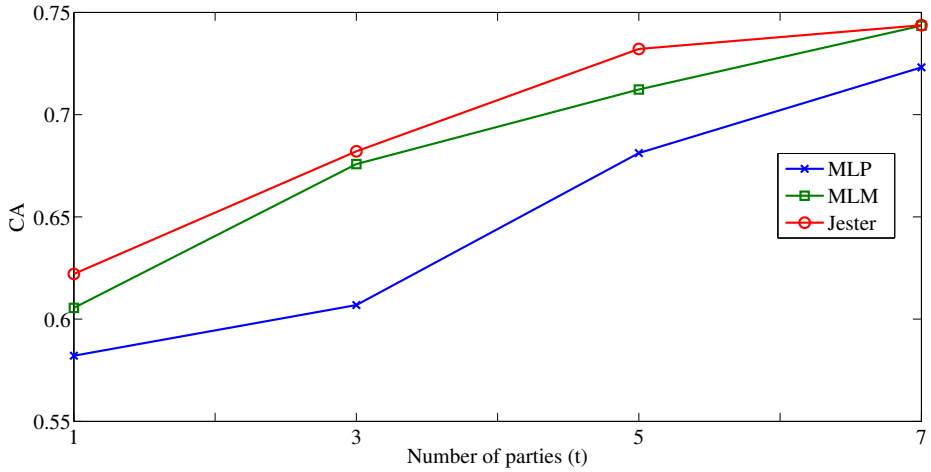
The computation, communication, and storage costs of the scheme proposed by Yakut and Polat [164] is given in Table 4.7. The off-line model needs to be updated periodically when the new data is inserted. Therefore, its computation time should not be very high, which hinders the update process. As shown in Table 4.7, theoretically computation cost of the off-line phase of ClaMPP is less than other related schemes. Our proposed method is for multi-party, whereas another related scheme is for only two-party. Hence to compare the computation costs of the off-line phase of ClaMPP experimentally with another related scheme, only two parties have been considered. When $t = 2$ the computation overhead of ClaMPP for off-line phase is $O(m)HPE.enc$, $O(m)HPE.mul$ and $O(m)HPE.dec$. The benchmarks available at CRYPTO++ tool kit (https://www.cryptopp.com/) are used to find total running time of cryptographic functions. Based on that, Figure 4.6 shows the computation time of model generation of different schemes. It is observed that the model generation time of ClaMPP is less than the other related scheme.

4.4.2 Limitations of proposed scheme

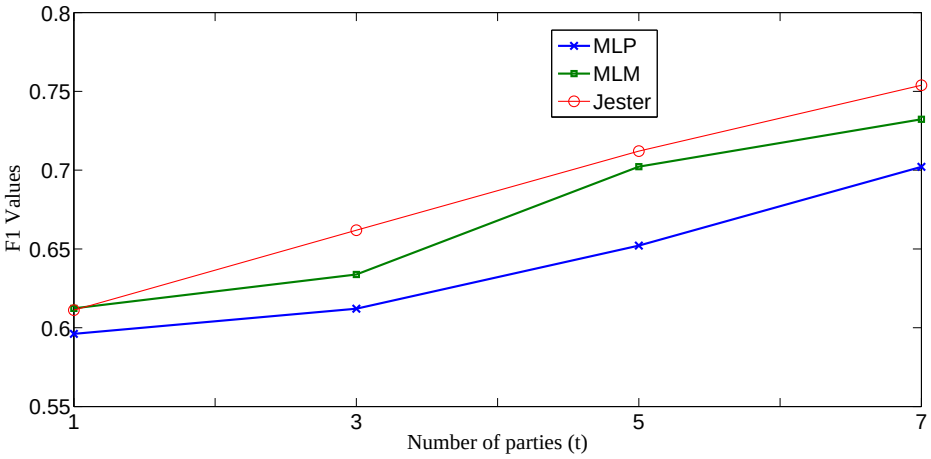
The proposed scheme is devised for the homogeneous network domain and needs few modification before executing it on the heterogeneous network domain. The limitations of the

Table 4.6: Storage, communication and computation complexity of ClaMPP

Phase	Storage complexity	Communication complexity	Computation complexity
Off-line phase	The four m x m matrices are required for storing likelihood probabilities and one 2 x m matrix is needed for storing priori probabilities. So, the storage complexity is $O(m^2)$	In protocol I, step 5 requires $2t-2$ communications, and as mentioned in step 6, t parties will execute step 5. Thus, total communication rounds of protocol I are $t(2t-2) = 2t^2-2t$, *i.e*, $O(t^2)$. Communication complexity of protocol II is also $O(t^2)$. In protocol III, step 1 requires $t-1$ communications to share the public key pk generated by party P_1 to $t-1$ collaborating parties. The total t number of communications are required from step 2 to step 4. Further, in step 5, $t-1$ communications are required to send the decrypted values to $t-1$ collaborating parties. So, the total communication rounds of protocol III are $t-1+t+t-1 = 3t-2$, *i.e*, $O(t)$.	The number of multiplications performed in PPCPP algorithm for party P_i are $2\|K_{P_i}\|+\|L_{P_1}\|+\|L_{P_2}\|+...\|L_{P_t}\| = 2\|K_{P_i}\|+m$. As mentioned in step 6 of protocol I, PPCPP algorithm is run by all t parties. Therefore total number of multiplications performed in protocol I and protocol II are $2\|K_{P_1}\|+2\|K_{P_2}\|+...2\|K_{P_t}\|+mt = 2m+mt$. Thus, computational complexity is $O(mt)$. If $(HPE.key, HPE.enc, HPE.dec, HPE.mul)$ homomorphic encryption system is used for calculation of prior probabilities for single item vector, $2tHPE.enc$, $2(t-1)HPE.mul$ and $2HPE.dec$ are required. Therefore, computational cost for calculation of prior probabilities of m item vectors is $2mtHPE.enc$, $2m(t-1)HPE.mul$ and $2mHPE.dec$, and computational complexity is $O(mt)HPE.enc$, $O(mt)HPE.mul$ and $O(m)HPE.dec$.
On-line phase	The storage complexity is same as conventional NBC based scheme without any privacy measure.	In ClaMPP, $t-1$ communications take place when AP sends label of item l to $t-1$ collaborating parties, and $t-1$ communications take place when $t-1$ collaborating parties send results to AP. So, the total communication rounds are $t-1+t-1 = 2t-2$, *i.e.*, $O(t)$.	The computation complexity is same as conventional NBC based scheme without any privacy measure.



(a) CA Values



(b) F1 values

Figure 4.4: CA and F1 values for MLP, MLM and Jester dataset

proposed scheme with respect to the heterogeneous network domain is that single cryptographic primitives cannot run on the heterogeneous network domains and synchronization of the operations executed by multiple parties may become difficult. The datasets MLP and MLM have the same type of interaction: user-item. The Jester dataset too has one type of interaction: user-joke. So, datasets also depict the homogeneous information network as discussed by Pham *et al.* [193].

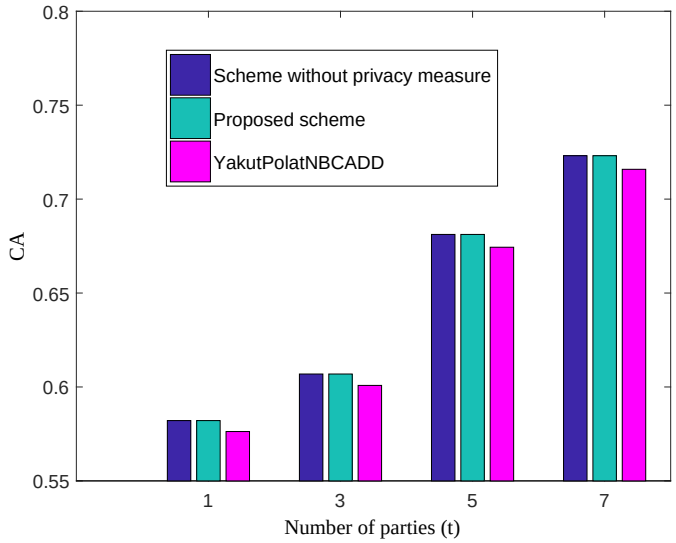


Figure 4.5: CA values for MLP for with and without privacy measure

4.5 Discussion

The distributed data mining techniques should use security models to preserve the sensitive information of individual parties. The cloud computing environment can be used to store the massive datasets and perform the high-intensive computing on it. Existing privacy preserving techniques for prediction generation on distributed data use data perturbation and homomorphic encryption. Data perturbation lead to accuracy losses whereas

Table 4.7: Comparison of proposed and related schemes

Paper	Data Distribution	Multi-party	Privacy preserving method	Accuracy loss	Storage Overhead	Off-line Communication overhead	On-line communication overhead	Off-line Computation overhead	On-line computation overhead
Yakut and Polat [164]	Arbitrary	No	Homomorphic encryption, addition of default and fake votes	Less than 1%	$O(m^2)$	$O(2)$	$O(m)$	$O(mn)HPE.enc$, $O(m^2)HPE.dec$, $O(m^2n)HPE.mul$	$O(nm)$
ClaMPP	Arbitrary	Yes	Random masking and polynomial aggregation techniques	No accuracy loss	$O(m^2)$	$O(t)$	$O(t^2m)$	$O(mt)HPE.enc$, $O(mt)HPE.mul$, $O(m)HPE.dec$	$O(nm)$

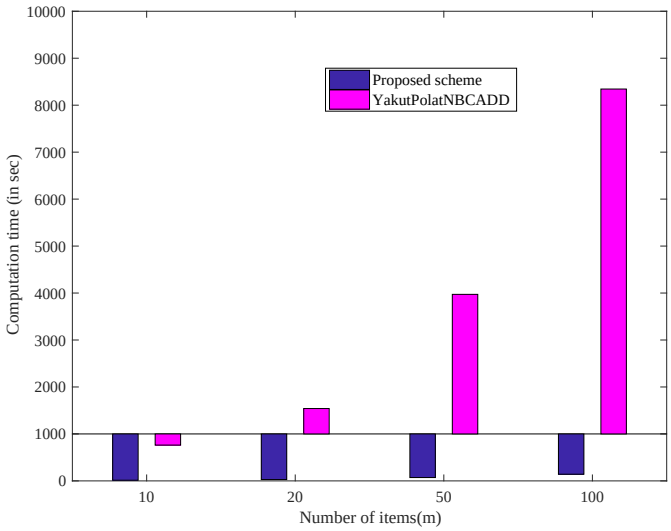


Figure 4.6: Off-line model computation cost

homomorphic encryption lead to high computing cost. In this work, a privacy preserving NBC based prediction generation scheme on a cloud environment for ADD among multiple parties is proposed. It is theoretically and empirically proved in the above sections that ClaMPP is secure and efficient; prediction quality is improved due to collaboration and there is no accuracy loss due to the proposed privacy measure. The computation overhead in the off-line phase is less as compared to another related scheme since its computation cost is $O(m)HPE.enc$, $O(m)HPE.mul$,$O(m)HPE.dec$ whereas another related scheme has $O(mn)HPE.enc$, $O(m^2)HPE.dec$, $O(m^2n)HPE.mul$ for two parties.

The main contributions of the ClaMPP are:

i) The PPCF scheme based on NBC classifier on the cloud environment is proposed for ADD among multiple parties such that the parties from different regions can collaborate with each other efficiently.

ii) ClaMPP does not use data perturbation for privacy, so it has negligible accuracy loss.

iii) Four protocols are proposed in this work. Conditional probability between items is

calculated using protocols I and II, while protocol III calculates the prior probability of item vectors and protocol IV generates the predictions.

iv) The proposed scheme is analyzed using different evaluation parameters such as: - security, accuracy and coverage. It has been experimentally demonstrated that accuracy and coverage of the proposed scheme are better in comparison to the other existing schemes.

In this chapter and previous chapter, the privacy preserving data mining techniques are proposed for distributed datasets. For some applications, there can be a centralized dataset. In the next chapter, the privacy preserving technique is proposed for centralized social network dataset. The optimized k-anonymization technique is proposed for privacy preservation of social network dataset.

Chapter 5

K-Neuro SVM based Privacy Preservation technique for Social Network

With the coming up of social network sites like Twitter, Facebook, LinkedIn, *etc.* information sharing has seen tremendous growth in the past few years. Although people wish to interact with like-minded persons through such platforms, there is a lot of personal information about individuals that need to be secured. To maintain the privacy of sensitive information in the social network, efficient privacy preserving techniques like k-anonymization, randomization, generalization, *etc.*, are used. The majority of the proposed techniques apply edge editing and clustering-based methods for creating k-anonymized social network graphs. Since these techniques distort the original properties of the graph to a great extent, various enhancements have been proposed which consider the addition of noisy nodes as one of the alternatives. In this chapter, an improved version of k-degree-anonymization on social network graph has been provided which uses the hybridization of NN and SVM, called as NeuroSVM where average path length of graph is preserved and addition of noisy nodes and noisy edges is

reduced significantly.

5.1 Overview

Social network sites such as Facebook, Twitter and LinkedIn are internet-based applications used by people to build relations with persons having similar hobbies, activities, career interest or real-life connections. People share a lot of information over the social network which includes photos, videos, political and religious views, creative ideas, real-world activities, and events, *etc.*; considering it safe and secure. Social networks can be easily modeled as a graph representing the relations and processes between entities (*e.g.,* person, community). The social networks can be categorized into three main types [35]:

- ◇ *Socializing social networks* used by people to create and maintain social relations with their friends and relatives.
- ◇ *Networking social networks* used by people for developing business or profession oriented relations.
- ◇ *Social navigation social networks* used by people to discover particular information or resources.

The organizations and government bodies perform data mining on social network data to extract the information, which they can use for their business growth, law enforcement, prevent epidemics, identify terrorist networks and organized crime and street gangs, *etc* [194]. Different types of network patterns (divided, unified, fragmented, clustered, broadcast) can be identified, which can be further used to find the target audience [195]. Although social networks provide a lot of benefits to social community as a whole, it is bundled with few issues which include spamming, unauthorized access, influence on employability, child safety, cyber-bullying, trolling and privacy.

5.1.1 Anonymized social network graph

To extract useful information, the organizations might provide their social network data to the third party. If the data is provided to the third party after naive anonymization *i.e.,* unique identifiers are removed before providing the data, it can reveal a lot of information. If third party has some background knowledge about the data like graph metrics, attributes of vertices, vertex degree, neighborhood and relationships between vertices and embedded subgraphs [196], actual information can be easily extracted. Such serious privacy concerns on graph data have been demonstrated in [197], [198]. So, social network data must be properly anonymized before giving it to the third party. The anonymization of the social network graph is more challenging than relational data because, in relational databases, change in one tuple does not affect the remaining tuples, whereas in graphs, change in one vertex or edge can affect the entire network of graphs.

In the existing studies, two models have been proposed to anonymize the graph: edge-editing based and clustering-based. In an edge-editing based model, addition or deletion of edges is done to anonymize the graph, whereas in clustering-based model, similar nodes are clustered into supernodes and supernodes are connected with each other using super edges. The supernode represents multiple nodes of the original graph and similarly super edge represents multiple edges of the original graph. One of the major shortcomings of such models is that the original properties of graph like edge intersection, diameter, Average Path Length (APL) are distorted. Yuan *et al.* [150] presented a technique of noisy node addition to anonymize the graph while preserving the APL property of the graph, but there is the scope of further improvement in the proposed technique. This work proposes a novel technique to generate a k-degree anonymous graph based on the noisy node addition method and neutralized cluster behavior through the use of two well-known classification algorithms: NN and

SVM.

5.2 Preliminaries

Definition 1. *A social network is made up of 4 tuples represented as G(V, E, CW, ψ) where V is the vertex set, E is the edge set, CW is the connection weight and f is the function* $f : V \times V \mapsto CW$ *which maps pair of vertices (u,v) to its corresponding connection weight (CW).*

The number of structural attacks that can be performed on graph are: degree attack, subgraph attack, neighborhood attack and hub fingerprint attack. Among them, degree (count of incoming and outgoing connections form the node) attack is common because collecting the background knowledge about the degree of the node is easy. Figure 1 represents a social

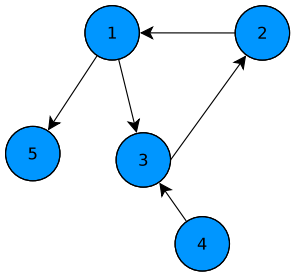


Figure 5.1: Original graph G

Vertex id	In degree	Out degree	Total degree
1	1	2	3
2	1	1	2
3	2	1	3
4	0	1	1
5	1	0	1

Table 5.1: Degree of each node for graph G

Vertex id	In degree	Out degree	Total degree	
1	1	2	3	
3	2	1	3	Group 1
2	1	1	2	
4	0	1	2	Group 2
5	1	0	2	

Figure 5.2: Grouping as per k-anonymity

network graph where V=5 and E=5. Vertex 1 and 3 possess the same degree and 4 and 5

possess same degree but vertex 2 has a degree which is not similar to any vertex degree, so it is an attack prone as shown in Table 5.1. Considering Figure 5.1, if an attacker knows that one person has two friends, then he can re-identify node 2 from the published social network graph. To prevent such attacks it is desired that each node in a graph has at least $k-1$ other nodes with the same degree.

Definition 2 : *A social network graph is called a k-degree anonymous graph G'(V', E', CW', f') if for each vertex (v') there are k-1 other vertices with same degree.*

If an edge is added between nodes 4 and 5 then the graph in Figure 5.1 will become 2-anonymous graph such that, nodes 2, 4 and 5 will have degree 2 and remaining nodes *i.e.* 1 and 3 will have degree 3 as shown in Figure 5.2.

Figure 5.3 represents an unbalanced graph in which the degree varies a lot as shown in Table 5.2. It is difficult to make this graph k-degree anonymous by the simple addition of

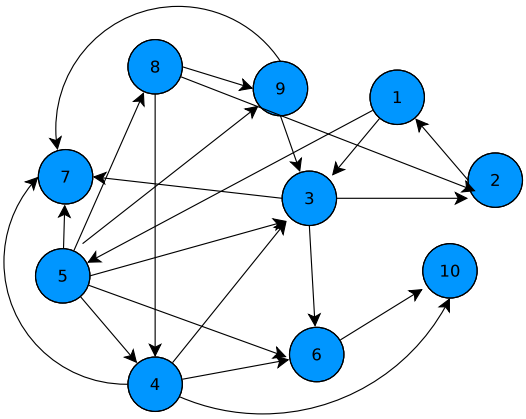


Figure 5.3: Unstructured graph G(V, E)

Vertex id	In degree	Out degree	Total degree
1	1	2	3
2	2	1	3
3	4	3	7
4	2	4	6
5	1	6	7
6	3	1	4
7	4	0	4
8	1	2	3
9	2	2	4
10	2	0	2

Table 5.2: Degree of each node for graph G

noisy edges. To overcome such issues, a solution was proposed by Yuan *et al.* [150] where they added noisy nodes to the graph and connected it to different nodes who were lacking in the degree value. The downside of this approach is that adding noisy nodes to graph solves

the problem of degree, but adds the problem of integrity *i.e.,* the original properties of the graph like edges, nodes, degree, APL, *etc.* will change.

5.3 Proposed Solution

The proposed work tries to find an optimal k-degree anonymous graph such that distortion in the original properties of graph (APL) is reduced to a great extent.

Mathematically, the objective function can be defined as :

$$Grp\_data^{[V,E,CW]}[x_i] \xrightarrow[input]{} [Cl] \xrightarrow[NeuroSVM(C,\gamma)]{min(\delta,IL,Nsy_v,Nsy_e)} [Op_{Cl}] \begin{cases} x_i'.c_i \\ \\ x_i'.deg \end{cases} \tag{5.1}$$

$$where \quad \delta = \frac{2\sum_i x_i'.deg}{n'(n'-1)} - \frac{2\sum_i x_i.deg}{n(n-1)} \tag{5.2}$$

$$IL = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i.deg - x_i'.deg)^2} \tag{5.3}$$

$$Nsy_v = \left|(V' \cap V)\right| \tag{5.4}$$

$$Nsy_e = \left|(E' \cap E)\right| \tag{5.5}$$

$$subject \quad to \quad \frac{\sum_i x_i'.deg}{n'} \geq \frac{\sum_i x_i.deg}{n} \tag{5.6}$$

$$x_i.deg, x_i'.deg', Nsy_e, Nsy_v, IL, n, n' \geq 0 \tag{5.7}$$

where V,E, CW and x_i represents the set of vertices, set of edges , set of connection weights and node of graph G respectively. Cl represents the clusters of nodes of graph G and Op_{Cl} represents the optimized cluster obtained from cluster Cl having minimum value of δ, IL, Nsy_v and Nsy_e. As represented by Eq.5.2, the δ represents the distortion in APL property of the original graph. IL represents the information loss occurred in newly constructed graph G', Nsy_v represents the number of noisy nodes added in new graph G' and Nsy_e represents the number of noisy edges added in new graph G'.

For achieving k-anonymization in graph, the proposed solution is divided into three protocols named as:

Protocol I (Clus_Creation): This protocol organized the nodes into clusters having minimum size k such that minimum number of noisy nodes and noisy edges are added for achieving the k-degree anonymous graph.

Protocol II (Degree_Shift): The main role of this protocol is to shift the degree of the vertex.

Protocol III (Class_NeuroSVM): This protocol classifies the elements of cluster created after Protocol II, through NeuroSVM.

Protocol I (Clus_Creation)

In this protocol, clusters of nodes $c_1, c_2, c_3 ... c_s$ are formed such that each cluster c_i contains at least k elements, where k is the degree of anonymity. The vertices of the graph are arranged in decreasing order according to their degree represented as $P =< p_1, p_2 ... p_n >$.

For obtaining the clusters having at least k nodes, algorithm 5.1 is followed. The degrees of nodes are used to calculate the mean μ and median md of cluster c_i. Based on the value of μ and md, the number of edges, E_μ and E_{md}, required to be added or deleted to achieve k-anonymity in the cluster are calculated. If $E_\mu < E_{md}$, then each node in cluster c_i has degree μ, else it has degree md. After completion of algorithm 5.1, there are s clusters such that each cluster c_i is associated with its cluster degree $c_i.deg$. Each node in cluster c_i should have a degree equal to cluster degree $c_i.deg$ to achieve k-anonymity. The graph is reconstructed since there are uncertainties in the clusters. The graph reconstruction phase is completed in two steps: The first step evaluates which vertex has a higher degree reference and which has a lower degree reference in the cluster. The lower reference is shifted to higher reference. A direct removal of the link will produce disorder in the graph and the graph will become attack prone. Hence, to reduce the possibility of being attack prone, a security table is maintained which contains the shifted link.

Algorithm 5.1 Cluster Formation

1: Input P, k
2: **for** $i = 1$ to k **do**
3: $c_1[i] = P[i]$
4: $m = k, p = k$
5: **end for**
6: Mean $\mu = \frac{\sum_{i=1}^{k} c_1[i].deg}{k}$
7: Median $md = \frac{c_1[i].deg + c_1[k].deg}{2}$
8: Calculate E_μ and E_{md}.
9: **if** $E_\mu < E_{md}$ **then**
10: $c_1.deg = \mu$
11: **else**
12: $c_1.deg = md$
13: **end if**
14: $c_1[m+1] = P[m+1]$
15: Calculate μ, md, E_μ and E_{md} as in Steps 6,7 and 8
16: **if** $E_\mu < E_{md}$ **then**
17: $Min_{E_{next}} = E_\mu$
18: $Min_{deg_{next}} = \mu$
19: **else**
20: $Min_{E_{next}} = E_{md}$
21: $Min_{deg_{next}} = md$
22: **end if**
23: **for** $i = p$ to $p + k - 1$ **do**
24: $l = 1$
25: $c_2[l] = P[i]$
26: **end for**
27: Calculate μ, md, E_μ and E_{md} as in Steps 6,7 and 8
28: **if** $E_\mu < E_{md}$ **then**
29: $Min_{E_{c_2}} = E_\mu$
30: $Min_{deg_{c_2}} = \mu$
31: **else**
32: $Min_{E_{c_2}} = E_{md}$
33: $Min_{deg_{c_2}} = md$
34: **end if**

35: **if** $Min_{E_{next}} < Min_{E_{c_2}}$ **then**
36: discard cluster c_2
37: $c_1.deg = Min_{deg_{next}}$
38: $p = m = m + 1$
39: **else**
40: remove last element from c_1
41: $c_2.deg = Min_{deg_{c_2}}$
42: $p = p + m - 1$
43: $m = p$
44: **end if**
45: Repeat the steps from 14 to 44, until left with less than k nodes in last and then add these last nodes in cluster c_s

Protocol 2 (Degree_Shift)

The s clusters created in algorithm 5.1 are added into cluster list C_{list}. Each cluster c_i is further divided into two groups *less* and *more* based on the cluster degree $c_i.deg$. The graph reconstruction algorithm 5.2 results into a much more sophisticated architecture which achieves k-anonymity and maintains the integrity of the network too. Figure 5.4 depicts the working of algorithm 5.2. Figure 5.4a, 5.4b, 5.4c and 5.4d represent different stages of group formation of a graph G (V, E).

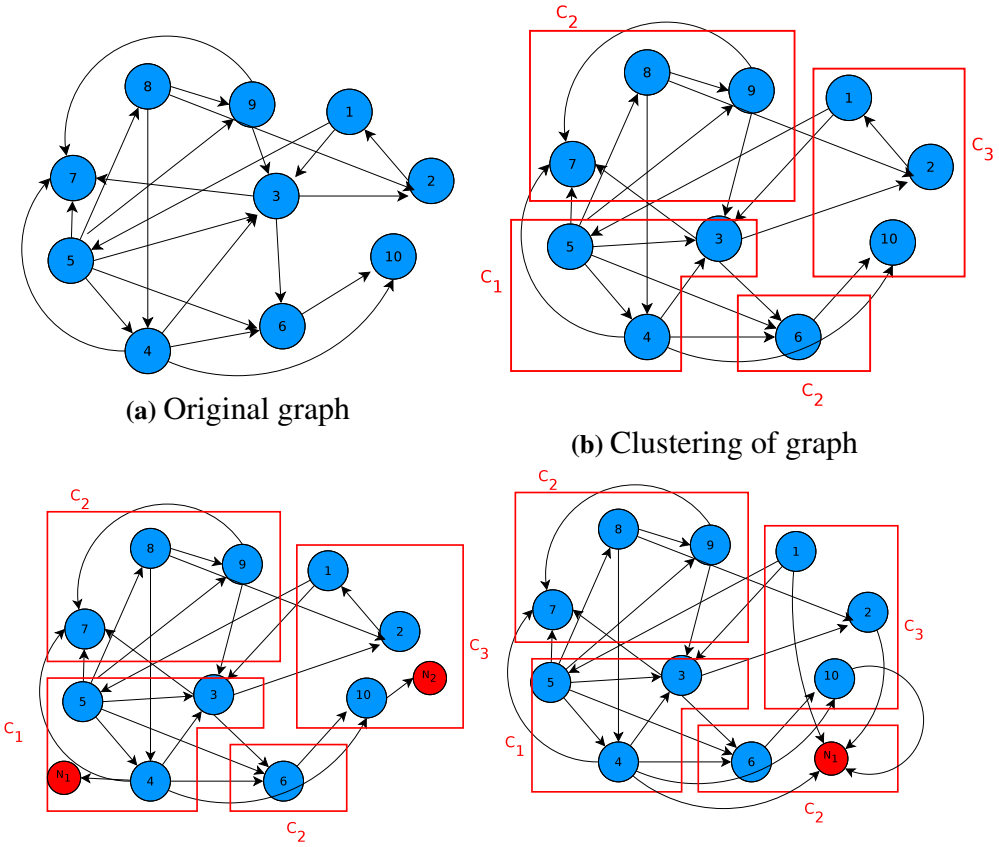
Figure 5.4a is the raw data that is passed to the algorithm 5.1 and Figure 5.4b represents the clusters created from it. The edge replacement is done which results into a graph shown in Figure 5.4c. The algorithm 5.2 also includes constraints for the situation when the graph is re-arranged but the condition is not satisfied. In such situations, algorithm 5.2 adds noisy nodes, which generates Figure 5.4d. The rule architecture generated requires strict monitoring and hence some of the deep learning algorithms have been applied and tested.

Protocol 3 (Class_NeuroSVM)

The classification of elements in clusters is executed by the hybridization of NN and SVM as shown in Figure 5.5. NN is a deep learning algorithm which consists of 3 layers. The first layer, called the input layer is used for providing input to the processing layer. The proposed layer utilizes *CW* as the input layer. The second layer is the hidden layer termed as the brain

Algorithm 5.2 Graph Reconstruction

for each cluster c in C_{list} **do**
  $D = c_i.deg$
  less=[]; // elements with less degree
  l=0; // counter for less
  more=[]; // elements with high degree
  m=0; // counter for More
  for each node el in c **do**
    $curr_{deg} = el_{deg}$
    if $curr_{deg} > D$ **then**
      $more[m] = el$
      $m = m + 1$
    else
      $less[l] = el$
      $l = l + 1$
    end if
  end for
  $min_v = min(V(more[]))$// find node having minimum degree
  $min_{Cw} = find(min(Cw_v))$// find connection of node v having minimum weight
  $J_{node_1} = V(min_{CW})$
  $node_{conn} = min_{Cw}$
  Remove $node_{conn}$ and attributes from cluster c and add to security table
  $min_v = min(V(less[]))$
  $J_{node_2} = min_v$
  Add $node_{(node_{conn})}$ to J_{node_2}
  Repeat until all nodes are not neutralized or each node is traversed at least once.
  if el_{deg} is not neutralized **then**
    Add $noisy_{nodes}$
  end if
end for



(a) Original graph

(b) Clustering of graph

(c) Cluster after shifting vertex group

(d) Regrouped clusters with noisy nodes

Figure 5.4: Example of Graph Reconstruction algorithm

of the NN. In the second layer, NN changes the input data with some threshold value as per the layering. The normal transformation for this layer is represented as:

$$w_1 = ax + b \tag{5.8}$$

where w_1 is the changed weight of the input data, a and b are arbitrary constants.

The trained architecture propagates with N number of hidden neurons. The SVM is usually used for binary classification which can distinguish between two elements.

NeuroSVM: The proposed architecture utilizes the advantages of both the algorithm. The main aim of applying a hybrid algorithm is to classify the elements of clustered groups and

to check whether there is any need to update the groupset. If the classification accuracy of elements in the clustered group is more than 75 % then the node editing is said to be successful. The classification algorithm 5.3 takes the connection weight of the grouped

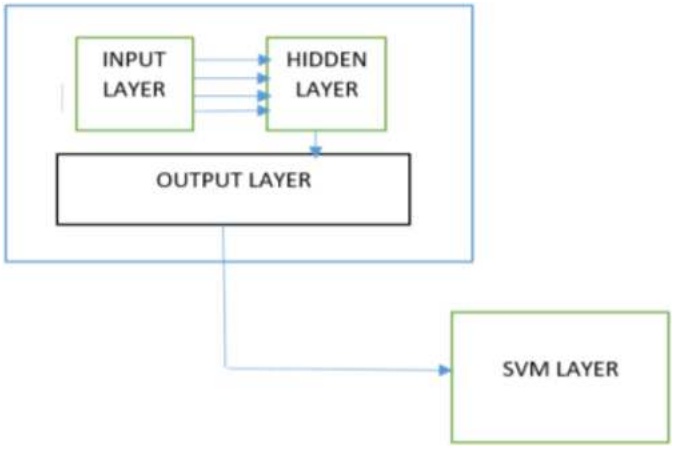


Figure 5.5: Hybridization of NN and SVM

elements as the input. The concerned groups are also part of the input to algorithm 5.3. The input set is initially analyzed with the neural network. The hidden layer of the neural network is divided into 10 hidden neurons and it is trained on the basis of Mean Square Error according to Levenberg-Marquardt algorithm as shown in Figure 5.6a. The algorithm 5.3 takes the processed trained data and the raw data passed at the time of training. It provides the group set at the training layer if the data is structured well. If not, it might provide distorted group architecture. In the case of distorted group architecture, there is more than one group for one element in the result set. In such cases, the binary classifier is quite effective. Algorithm 5.3 has used SVM in such situations and finds out the best-fitted class. The basic kernels used in SVM are [199].

1. Linear Kernel $x^T y + b$
2. Polynomial Kernel $(\gamma x^T y + b)^n$
3. Radial Kernel $exp(-\gamma(x-y)^2$

Algorithm 5.3 Classify NeuroSVM

$Input_{set} = Grouped_{Connection_{Weight}}()$
$Group_{set} = Group_{Cluster}(Input_{set})$
$NeuralTrainSet = newff(Input_{set}, Group_{set}, 20)$ // The NN is initialized with 20 Hidden Neurons at one layer
$nTrain = Train(NeuralTrainSet)$ // nTrained is the trained Neural Set
$TestSet = NeuralTrainSet.$
$nResult = Classify(nTrain, TestSet$ // Employed architecture is for supervised learning and hence the trainset will act as the test set itself
$Anl = Analyze(Group_{set}, nResult)$
if $Anl.Contains.Groups > 1$ **then**
    $Initialize(SVM)$
    $SVML = Input(ntrain(Anl), Group_{set}(Anl))$
    $SvmOutput = Classify(SVML, ntrain(Anl))$
    $Output = SvmOutput$
else
    $Output = nResult$
    $l = l + 1$
end if
$FindErrorRate(Output)$
$Accuracy = 100 - ErrorRate$
if $Accuracy \geq 75$ **then**
    $No\_need\_to\_reconstruct$;
    $m = m + 1$
else
    $Reconstruct\ \ Group\ \ and\ \ Call\ \ NeuroSVM()$
end if

4. Sigmoid Kernel $tanh(\gamma x^T y + b)$

where b, γ and n are kernel parameters. The Gaussian Radial Base Fucntion Kernel (RBF) is most commonly used kernel in SVM. Since datasets are non-linear in nature with fewer features sets, Gaussian RBF is used for classification. It is represented as:

$$K(x,y) = exp(-\frac{1}{2\sigma^2}||x-y||^2) \tag{5.9}$$

where $\gamma = \frac{1}{2\sigma^2}$. The RBF kernel defines the decision boundary which separates the training data into two classes. The decision boundary function $g(x)$ of RBF kernel can be represented as [200] :

$$g(x) = w.K(x,y) + b \qquad w \in x, b \in R^m \tag{5.10}$$

where w is the normal vector to decision boundary and b is regularization parameter in m-dimensional space R^m. The decision function $g(x)$ can be represented as optimization problem:

$$min(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i) \tag{5.11}$$

where $\frac{1}{2}||w||^2$ represenst maximum margin and $C\sum_{i=1}^{n}\xi_i$ represents the minimum error. The ξ are slack variables added to counter the occurrences of inequalities during classification. The Lagrange multipliers $(\alpha_i, \beta_i > 0)$ are used to solve Eq. (5.11). By using Lagrange multiplier the Eq. (5.10) can be represented as:

$$g(x) = \sum_{i=1}^{n}(\alpha_i - \beta_i)K(x,y) + b \qquad \alpha_i, \beta_i \in [0, C] \tag{5.12}$$

After choosing the kernel, its modeling parameters need to be chosen for getting a better quality of classification. In RBF kernel, C and γ have a direct effect on the accuracy of classification. In order to find the right value of C and γ, the grid search approach is used as

defined in [199].

5.4 Results

For the evaluation of proposed work, metrics considered are APL, Information loss, Precision, Recall and F-measure. The results obtained after the simulation of proposed work are shown in this section.

i) APL: The APL is a ratio of the total distance between two vertices to the total number of vertices in the given graph set. Evaluation of distance can be done through various methods or algorithms. Mathematically, it can be represented as:

$$APL = \frac{2\sum_{i,j \in G} d_{i,j}}{n(n-1)} \tag{5.13}$$

where $d_{i,j}$ is the distance between two vertices.

ii) Information Loss (IL): Information loss is the total loss in the integrity of the data when a graph is rearranged to achieve the k-anonymization. Different graph metrics can be used to measure information loss. In this paper, the degree of vertices is used for measuring information loss based on metric defined in [201].

$$IL = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(deg_{v_i} - deg'_{v'_i})^2} \tag{5.14}$$

where deg_{v_i} is the degree of vertex v_i of original graph and $deg'_{v'_i}$ is the degree of vertex v'_i of the k-degree anonymous graph.

iii) Precision (P): It is defined as the ratio of relevant data (T_p) with respect to total retrieve

data ($T_p + F_p$).

$$P = \frac{T_p}{T_p + F_p} \tag{5.15}$$

iv) Recall (R): It is the ratio of relevant data (T_p) with respect to the total relevant data ($T_p + F_n$) during training.

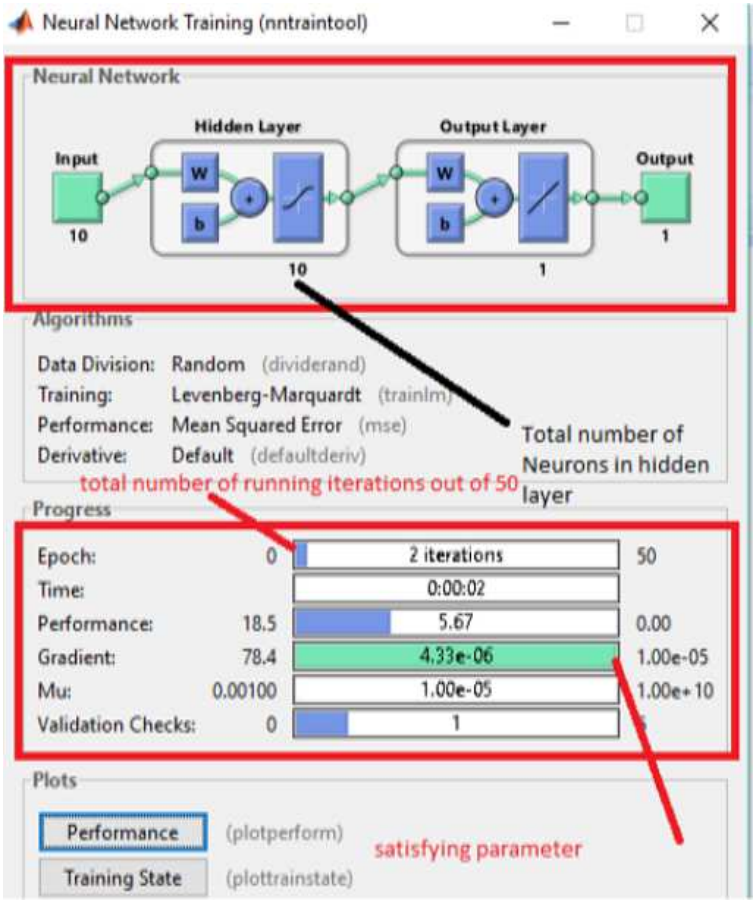
$$R = \frac{T_p}{T_p + F_n} \tag{5.16}$$

v) F-measure (ψ): It is the harmonic mean of precision and recall. it is approximately the average value of precision and recall.

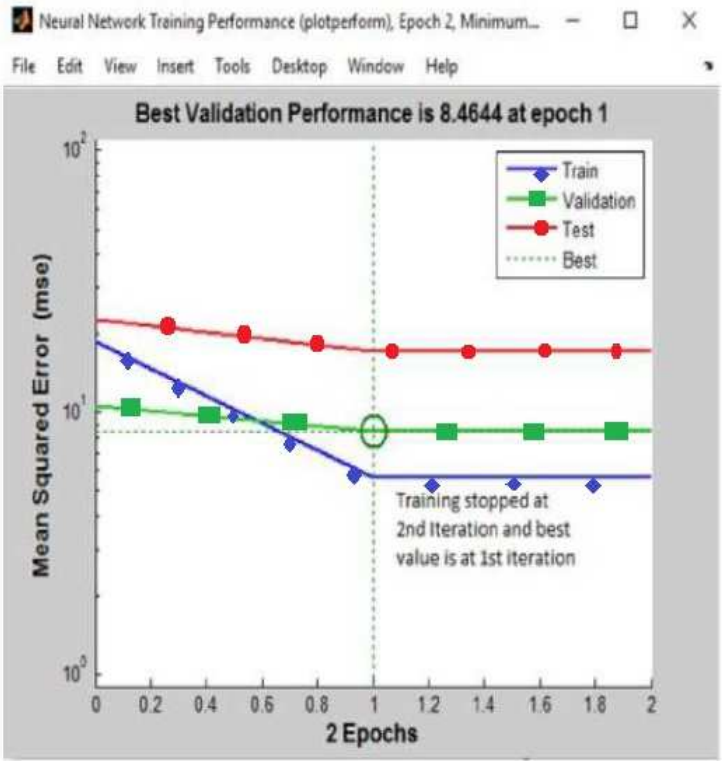
$$\psi = \frac{2PR}{P + R} \tag{5.17}$$

In the first place, the classification of the re-grouped vertices takes place. The neural network discussed earlier initiates training mechanism with certain iterations. The network understands the input data if any of the provided validation is complete. The architecture has been implemented in the MATLAB platform and there are 5 validation parameters other than the total number of provided iteration as shown in Figure 5.6a. If any of the validation parameter is satisfied then the network backtracks and find the cross point of the training weight and the iterative best measurement as shown in Figure 5.6b.

The NN varies the training data with fragmentation value of data along with the variation of threshold for segmentation. NN also performs a regression analysis over the train set to assure that the classification process gets the best out of the training data. The results in Figure 5.7 represent the regression analysis. High regression value results in good training mechanism $(0 \leq R \leq 1)$ where R represents regression. The support vector machine has also been utilized with the Gaussian RBF kernel. The best value of C and γ parameters of Gaussian RBF kernel for maximum accuracy is calculated through grid search methodology.



(a) Neural network training description



(b) Neural network training performance

Figure 5.6: Neural network training

For loose grid search, C value varies from $(2^{-10}, 2^{-7}, 2^{-4}, 2^{-1}, 2^{1}, 2^{4}, 2^{7}, 2^{10})$ and γ value varies from $(2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7})$. The best value of C and γ is found as $(2^{1}, 2^{-1})$ as shown in Figure 5.8. To further refine the value of C and γ, fine grid search is performed in the neighborhood of the best value $(2^{1}, 2^{-1})$. After fine grid search, the best value of C and γ obtained for maximum accuracy is $(1.9, 0.7)$. The SVM is trained using Gaussian RBF kernel having values for C and γ as $(1.9, 0.7)$. Next, the trained SVM is used on the test data for performing the classification.

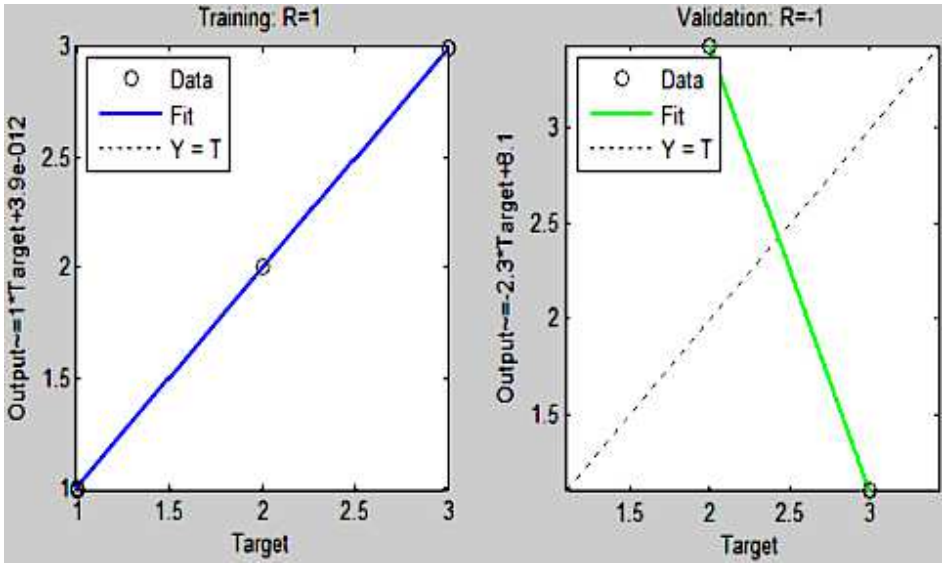
To test the effectiveness of the proposed algorithm, two datasets Arnet [202] and Cora [203] have been used. For regression analysis, the value of k has been varied from 5 to 30. Figure 5.9, demonstrates that APL value of the proposed technique is less than Yuan *et al.* [150] technique. The proposed architecture also utilizes the data mining concepts and hence the classification parameters have also been evaluated as shown in Figure 5.10. High precision and recall values result in a high F-measure which leads to a better classification mechanism. The average F-measure is approximately 0.75 which indicates an approximate accuracy of 75% as shown in Table 5.3.

As shown in Figure 5.11 and Table 5.4 information loss of the proposed technique is less compared to Yuan *et al.* technique. Less modification in original degree leads to less information loss as explained in Eq. (5.14).

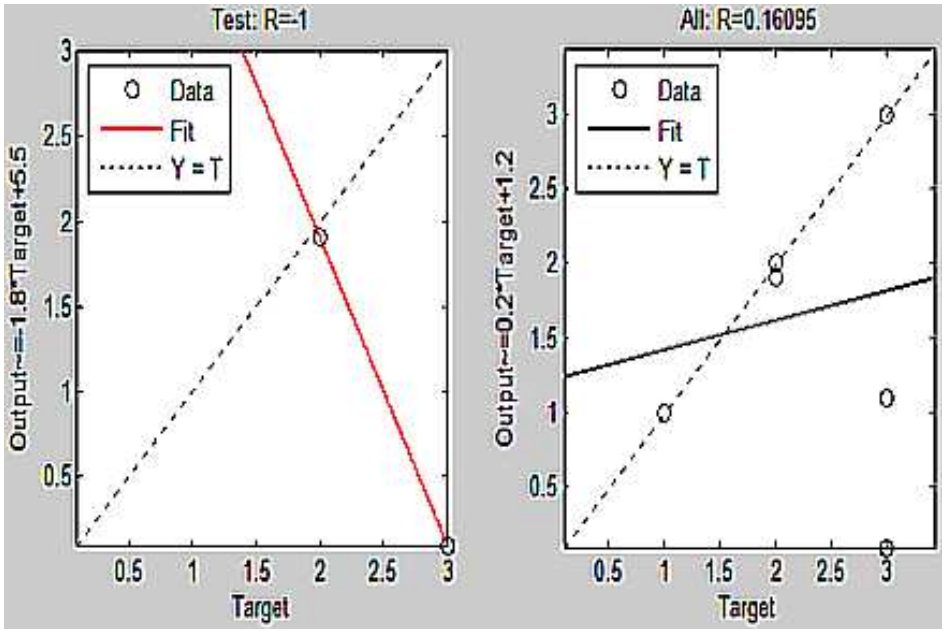
5.5 Discussion

In this work, a novel algorithm is proposed for optimal k-degree anonymization by using an artificial neural network along with the support vector machine. The main idea of the proposed work is to anonymize the graph to prevent degree attacks in such a way that there is less distortion of the properties of the graph. It has been experimentally demonstrated in the

above section that distortion in APL of the anonymized graph is less as compared to other related technique. The information loss is also reduced and the accuracy is more than 75% with better precision and recall rate.



(a) Phase 1



(b) Phase 2

Figure 5.7: Regression analysis

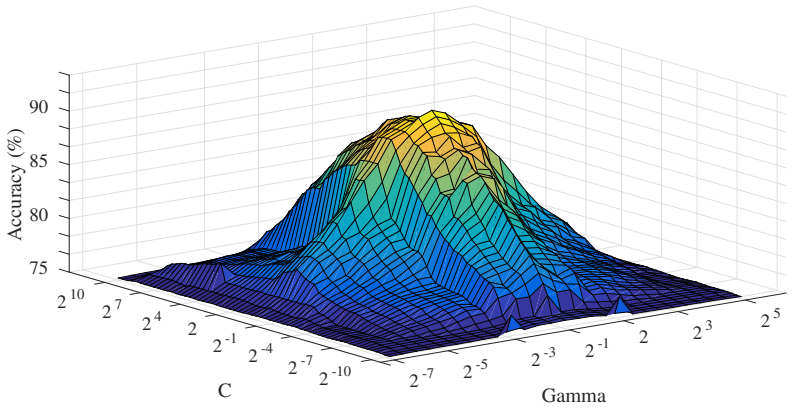
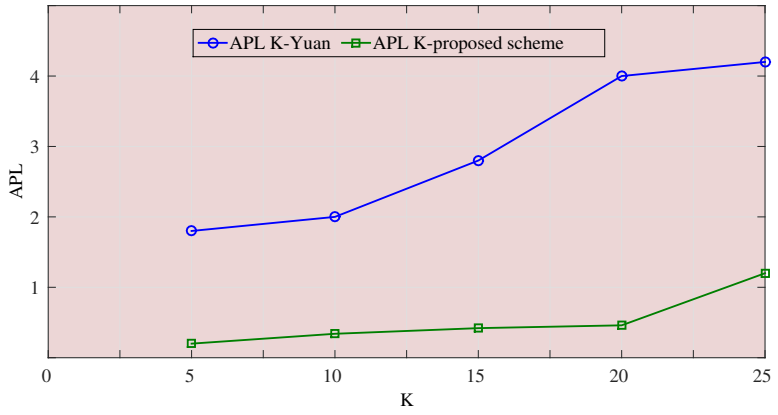
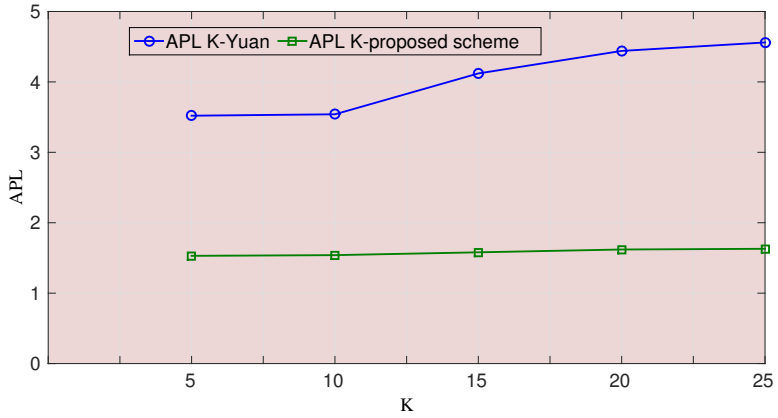


Figure 5.8: Grid search cross-validation of NeuroSVM

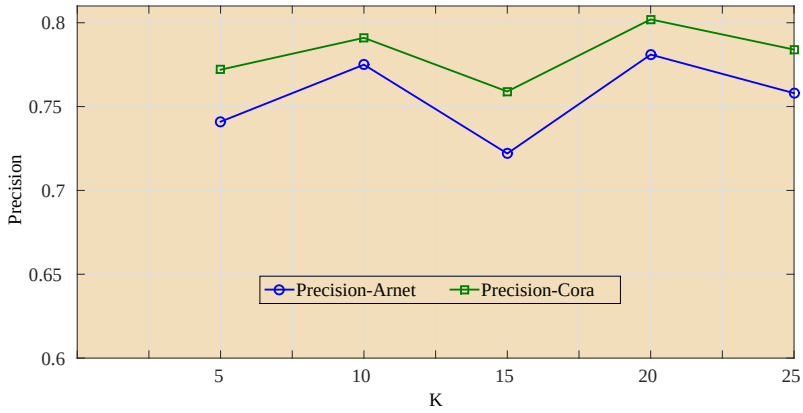


(a) Arnet dataset

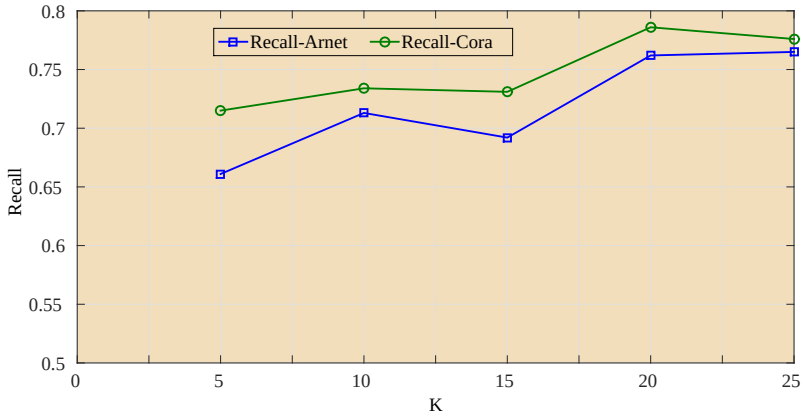


(b) Cora dataset

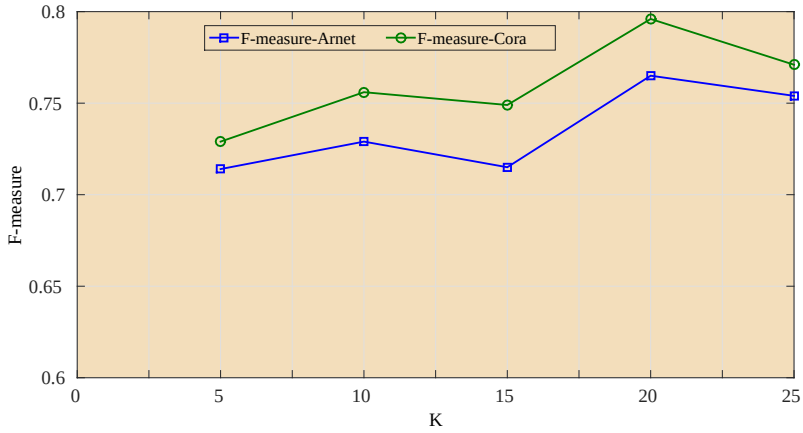
Figure 5.9: APL for Arnet and Cora dataset



(a) Precision



(b) Recall



(c) F-measure

Figure 5.10: Classification parameters

K	5	10	15	20	25
Precision	0.76	0.78	0.74	0.79	0.77
Recall	0.69	0.72	0.71	0.77	0.77
F-measure	0.72	0.74	0.73	0.78	0.76

Table 5.3: Average Precision, Recall and F-measure for Arnet and Cora dataset

K	5	10	15	20	25
Info-Loss-Yuan	1.25	1.45	1.46	1.42	1.47
Info-Loss-Proposed	1.21	1.35	1.42	1.58	1.59

Table 5.4: Average Information Loss for Arnet and Cora dataset

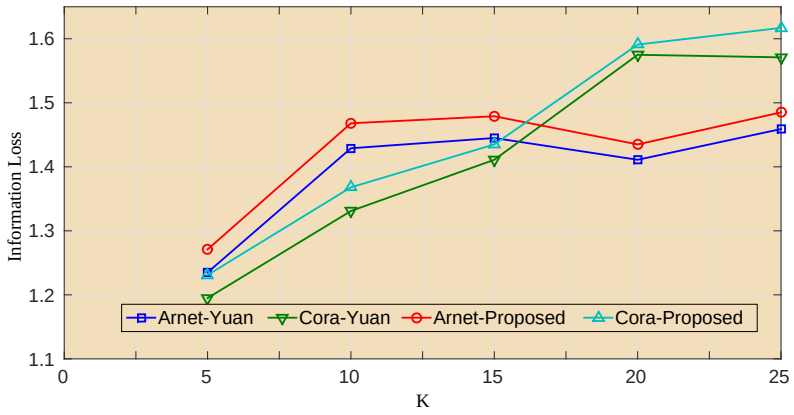


Figure 5.11: Information loss

The main contributions of this work are:

i) A novel technique is proposed for generating a k-degree anonymous graph based on noisy node addition and classification algorithm NeuroSVM.

ii) The proposed technique is evaluated on different parameters like APL, precision, recall, F-measure and information loss.

iii) The proposed technique has accuracy more than 75%. The information loss and distortion in APL have also been reduced as compared to other related technique.

Chapter 6

Conclusion and Future scope

The focus of this thesis has been on efficient PPDM techniques for big data applications like healthcare, social network and recommender systems, *etc*.

The tremendous growth in Information technology has revolutionized healthcare industry in a big way. Patients use the Internet for getting information about diseases, diagnoses, treatments, physicians, and hospitals. Patient-oriented decision-making can enhance the efficiency of the healthcare recommender system provided the data scattered across different geographical regions is collected, mined and analyzed efficiently. In this thesis, a PPCF technique based upon randomization, masking and homomorphic encryption technique, is proposed for healthcare recommendation generation on ADD among multiple parties. Three protocols have been proposed in our work. The similarity between items is generated by using Protocol I and PPSDPC algorithm, Protocol II calculates the length of item vectors using homomorphic encryption technique and Protocol III generates the predictions. The proposed scheme is efficient in terms of computation cost and maintains the patient's privacy. For an e-commerce based recommender system, a multiparty privacy preserving naive Bayesian classification scheme based on the cloud environment, named as ClaMPP, is proposed. Four protocols have been proposed. Protocol I and Protocol II are used for calculating conditional

probabilities securely using PPCPP algorithm, Protocol III is used for calculating the prior probabilities using homomorphic encryption and Protocol IV is used for generating the predictions online. The proposed scheme is secure, and the accuracy of prediction generation is improved due to the collaboration of multiple parties. Also, the accuracy loss due to the introduction of privacy measures in the proposed approach is negligible.

With the coming up of social network sites like Twitter, Facebook, LinkedIn, etc. information sharing has seen tremendous growth in the past few years, but these sites also contain the personal information of individuals that need to be secured. To maintain the privacy of sensitive information in the social networks, an improved version of k-degree-anonymization on social network graph has been provided which uses the hybridization of NN and SVM, called as k-NeuroSVM. The proposed technique is divided into three protocols: Protocol I (Clus Creation), Protocol II (Degree Shift) and Protocol III (Class NeuroSVM). Protocol I is used for dividing the nodes of graph into cluster. Protocol II is used for shifting the degree of nodes to achieve k-anonymization and Protocol III is used for classifying the elements of the cluster using NeuroSVM. In the proposed technique, the average path length of the graph is preserved and the addition of noisy nodes and noisy edges is reduced significantly.

6.1 Contributions

Major contributions of this work are:

- The work proposes an efficient framework PPDMB for privacy preserving distributed data mining on Big data. Based upon the proposed framework two novel efficient PPCF techniques: EMS-PPCF and ClaMPP are proposed for privacy preserving recommendation generation on ADD among multiple parties.

  i) EMS-PPCF is efficient model-based privacy preserving collaborative filtering

technique for recommendation generation in the healthcare recommender system where data is arbitrarily distributed among multiple parties. It is analyzed using different parameters: security, accuracy, coverage, and performance. With respect to these parameters, it has been experimentally demonstrated that accuracy and coverage of the EMS-PPCF scheme are better in comparison to the other competing schemes when different parties collaborate with each other. Further, the computation cost of the EMS-PPCF scheme is less in comparison to other schemes.

ii) ClaMPP, cloud-based privacy preserving collaborative filtering scheme for e-commerce based recommender system use NBC for recommendation generation. It does not use data perturbation for privacy, so it has negligible accuracy loss and also the computational overhead for off-line model generation phase is low as compared to other related techniques.

- EMS-PPCF and ClaMPP schemes are implemented in Java with major performance improvements compared to other schemes.

- For a centralized social network dataset, the K-neuroSVM anonymization technique is proposed for the privacy preservation of the the social network graph. It is based on noisy node addition and classification algorithm NeuroSVM. It is implemented in MATLAB and it has accuracy more than 75%, and the information loss and distortion in APL has also been reduced as compared to other related technique.

6.2 Future Scope

The proposed PPCF scheme for healthcare recommender system provides the recommendations based on the item-item similarity model on ADD among various parties. As future

work, parallel computing techniques can be considered to further reduce the computation time of the off-line model generation process. The proposed scheme depends on item-based CF technique. In the future, research can be carried out on privacy preserving schemes for demographic, social or content-based recommendation generation on ADD among multiple parties.

NBC based PPCF scheme on the cloud environment, has been used for generating the predictions on arbitrarily distributed binary ratings among multiple parties. In the future, the analysis of secure NBC based prediction generation on arbitrarily distributed numerical data sets can be considered.

The proposed k-NeuroSVM technique is used for the k-anonymization of the social network graph to prevent degree attack. This work can be extended for dynamic social network analysis. Another direction is to extend the proposed algorithm for dealing with other structural attacks like neighborhood, embedded subgraph, and hub fingerprinting for social network graphs.

Bibliography

[1] Domo. Data Never Sleeps 6.0. Press release.

[2] Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[3] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, 47:98–115, 2015.

[4] Bernard Marr. How much data do we create every day? the mind-blowing stats everyone should read. Blog, May 2018.

[5] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[6] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

[7] John Gantz and David Reinsel. Extracting value from chaos. *IDC iview*, 1142(2011):1–12, 2011.

[8] Jonathan Stuart Ward and Adam Barker. Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*, 2013.

[9] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.

[10] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[11] Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. Sliq: A fast scalable classifier for data mining. In *International Conference on Extending Database Technology*, pages 18–32. Springer, 1996.

[12] Rong Chen, Krishnamoorthy Sivakumar, and Hillol Kargupta. Collective mining of bayesian networks from distributed heterogeneous data. *Knowledge and Information Systems*, 6(2):164–187, 2004.

[13] Ioannis Kopanas, Nikolaos M Avouris, and Sophia Daskalaki. The role of domain knowledge in a large scale data mining project. In *Hellenic Conference on Artificial Intelligence*, pages 288–299. Springer, 2002.

[14] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, and Dino Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(1):10, 2014.

[15] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Annual International Cryptology Conference*, pages 36–54. Springer, 2000.

[16] Pinaki Mitra, Rinku Das, and Girish Sundaram. Privatizing user credential information of web services in a shared user environment. *arXiv preprint arXiv:1308.3482*, 2013.

[17] Robert Gibbons. *A primer in game theory*. Harvester Wheatsheaf, 1992.

[18] Wikipedia contributors. Sockpuppet (internet) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Sockpuppet_(Internet)&oldid=849953293, 2018. [Online; accessed 18-July-2018].

[19] Sarah Downey. Introducing maskme: why ever give away your real personal info online? Blog, July 2013.

[20] Stephen R Carter. Techniques to pollute electronic profiling, November 29 2011. US Patent 8,069,485.

[21] Rui Li, Denise de Vries, and John Roddick. Bands of privacy preserving objectives: Classification of ppdm strategies. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 137–152. Australian Computer Society, Inc., 2011.

[22] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[23] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.

[24] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, page 51. American Medical Informatics Association, 1997.

[25] Shivaramakrishnan Narayan, Martin Gagné, and Reihaneh Safavi-Naini. Privacy preserving ehr system using attribute-based infrastructure. In *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*, pages 47–52. ACM, 2010.

[26] Mohammad Jafari, Reihaneh Safavi-Naini, and Nicholas Paul Sheppard. A rights management approach to protection of privacy in a cloud of electronic health records. In *Proceedings of the 11th annual ACM workshop on Digital rights management*, pages 23–30. ACM, 2011.

[27] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, December 1992.

[28] Huseyin Polat and Wenliang Du. Privacy-preserving collaborative filtering on vertically partitioned data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 651–658. Springer, 2005.

[29] Bradley Malin and Latanya Sweeney. Determining the identifiability of dna database entries. In *Proceedings of the AMIA Symposium*, page 537. American Medical Informatics Association, 2000.

[30] Bradley Malin and Latanya Sweeney. Re-identification of dna through an automated linkage process. In *Proceedings of the AMIA Symposium*, page 423. American Medical Informatics Association, 2001.

[31] Bradley Malin. Why methods for genomic data privacy fail and what we can do to fix it. In *AAAS Annual Meeting*, 2004.

[32] Bradley Malin. *Protecting dna sequence anonymity with generalization lattices*. Carnegie Mellon University, School of Computer Science [Institute for Software Research International], 2004.

[33] Latanya Sweeney. Privacy-preserving bio-terrorism surveillance. In *AAAI spring symposium, AI Technologies for Homeland Security*, 2005.

[34] Yair Amichai-Hamburger and Tsahi Hayat. *Social Networking*, pages 1–12. American Cancer Society, 2017.

[35] Mike Thelwall. Social network sites: Users and uses. *Advances in computers*, 76:19–73, 2009.

[36] Saul Hansell. Social network launches worldwide spam campaign. *New York Times daily paper*, 27, 2007.

[37] Teenage Social Life. Why youth (heart) social network sites. *Gender, Race, and Class in Media: A Critical Reader*, page 409, 2011.

[38] Alessandro Acquisti and Ralph Gross. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, pages pnas–0904891106, 2009.

[39] Laura Van Eperen and Francesco M Marincola. How scientists use social media to communicate their research, 2011.

[40] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[41] Ljiljana Brankovic and Vladimir Estivill-Castro. Privacy issues in knowledge discovery and data mining. In *Australian institute of computer ethics conference*, pages 89–99, 1999.

[42] Charu C Aggarwal and S Yu Philip. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*, pages 11–52. Springer, 2008.

[43] Mukesh Kumar Jha and TP Sharma. A new approach to secure data aggregation protocol for wireless sensor network. *Int. J. Comput. Sci. Eng.(IJCSE)*, 2(05):1539–1543, 2010.

[44] Majid Bashir Malik, M Asger Ghazi, and Rashid Ali. Privacy preserving data mining techniques: current scenario and future prospects. In *Computer and Communication Technology (ICCCT), 2012 Third International Conference on*, pages 26–32. IEEE, 2012.

[45] Stan Matwin. Privacy-preserving data mining techniques: survey and challenges. *Discrimination and Privacy in the Information Society*, pages 209–221, 2013.

[46] Geetha Jagannathan and Rebecca N Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 593–599. ACM, 2005.

[47] Elena Dasseni, Vassilios S Verykios, Ahmed K Elmagarmid, and Elisa Bertino. Hiding association rules by using confidence and support. In *International Workshop on Information Hiding*, pages 369–383. Springer, 2001.

[48] Yücel Saygin, Vassilios S Verykios, and Ahmed K Elmagarmid. Privacy preserving association rule mining. In *Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, 2002. RIDE-2EC 2002. Proceedings. Twelfth International Workshop on*, pages 151–158. IEEE, 2002.

[49] Stanley RM Oliveira and Osmar R Zaiane. Algorithms for balancing privacy and knowledge discovery in association rule mining. In *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International*, pages 54–63. IEEE, 2003.

[50] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Information Systems*, 29(4):343–364, 2004.

[51] Jaideep Vaidya and Chris Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 639–644. ACM, 2002.

[52] Murat Kantarcıoglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge & Data Engineering*, 16(9):1026–1037, 2004.

[53] Andrew Chi-Chih Yao. How to generate and exchange secrets. In *Foundations of Computer Science, 1986., 27th Annual Symposium on*, pages 162–167. IEEE, 1986.

[54] Justin Zhan, Stan Matwin, and LiWu Chang. Privacy-preserving collaborative association rule mining. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 153–165. Springer, 2005.

[55] Gopalan Kesavaraj and Sreekumar Sukumaran. A study on classification techniques in data mining. In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pages 1–7. IEEE, 2013.

[56] Ming-Jun Xiao, Liu-Sheng Huang, Yong-Long Luo, and Hong Shen. Privacy preserving id3 algorithm over horizontally partitioned data. In *Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on*, pages 239–243. IEEE, 2005.

[57] Charu C Aggarwal. Privacy-preserving data mining. In *Data Mining*, pages 663–693. Springer, 2015.

[58] Wenliang Du and Zhijun Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 505–510. ACM, 2003.

[59] Kshitij Pathak, Narendra S Chaudhari, and Aruna Tiwari. Privacy preserving association rule mining by introducing concept of impact factor. In *Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on*, pages 1458–1461. IEEE, 2012.

[60] Nikunj H Domadiya and Udai Pratap Rao. Hiding sensitive association rules to maintain privacy and data quality in database. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pages 1306–1310. IEEE, 2013.

[61] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[62] Hai Quoc Le and Somjit Arch-int. A conceptual framework for privacy preserving of association rule mining in e-commerce. In *Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on*, pages 1999–2003. IEEE, 2012.

[63] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22 of *2*, pages 207–216. ACM, 1993.

[64] George V Moustakides and Vassilios S Verykios. A maxmin approach for hiding frequent itemsets. *Data & Knowledge Engineering*, 65(1):75–89, 2008.

[65] Komal Shah, Amit Thakkar, and Amit Ganatra. Association rule hiding by heuristic approach to reduce side effects & hide multiple rhs items. *International Journal of Computer Applications*, 45(1):1–7, 2012.

[66] Dhyanendra Jain, Pallavi Khatri, Rishi Soni, and Brijesh Kumar Chaurasia. Hiding sensitive association rules without altering the support of sensitive item (s). In *International Conference on Computer Science and Information Technology*, pages 500–509. Springer, 2012.

[67] Marzena Kryszkiewicz. Representative association rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 198–209. Springer, 1998.

[68] Haoliang Lou, Yunlong Ma, Feng Zhang, Min Liu, and Weiming Shen. Data mining for privacy preserving association rules based on improved mask algorithm. In *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on*, pages 265–270. IEEE, 2014.

[69] Tzung-Pei Hong, Chun-Wei Lin, Kuo-Tung Yang, and Shyue-Liang Wang. Using tf-idf to hide sensitive itemsets. *Applied Intelligence*, 38(4):502–510, 2013.

[70] Chieh-Ming Wu, Yin-Fu Huang, and Jian-Ying Chen. Privacy preserving association rules by using greedy approach. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, volume 4, pages 61–65. IEEE, 2009.

[71] Hai Quoc Le, Somjit Arch-Int, Huy Xuan Nguyen, and Ngamnij Arch-Int. Association rule hiding in risk management for retail supply chain collaboration. *Computers in Industry*, 64(7):776–784, 2013.

[72] Chun-Wei Lin, Tzung-Pei Hong, and Hung-Chuan Hsu. Reducing side effects of hiding sensitive itemsets in privacy preserving data mining. *The Scientific World Journal*, 2014, 2014.

[73] Rakesh Agrawal and Ramakrishnan Srikant. Method and system for building a decision-tree classifier from privacy-preserving data, April 8 2003. US Patent 6,546,389.

[74] Wenliang Du and Zhijun Zhan. Building decision tree classifier on private data. In *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*, pages 1–8. Australian Computer Society, Inc., 2002.

[75] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. Privacy-preserving svm using nonlinear kernels on horizontally partitioned data. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 603–610. ACM, 2006.

[76] Hwanjo Yu, Jaideep Vaidya, and Xiaoqian Jiang. Privacy-preserving svm classification on vertically partitioned data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 647–656. Springer, 2006.

[77] Keng-Pei Lin and Ming-Syan Chen. On the design and analysis of the privacy-preserving svm classifier. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1704–1717, 2011.

[78] Yogachandran Rahulamathavan, Raphael C-W Phan, Suresh Veluru, Kanapathippillai Cumanan, and Muttukrishnan Rajarajan. Privacy-preserving multi-class support vector machine for outsourcing the data classification in cloud. *IEEE Transactions on Dependable and Secure Computing*, 11(5):467–479, 2014.

[79] Li Sun, Wei-Song Mu, Biao Qi, and Zhi-Jian Zhou. A new privacy-preserving proximal support vector machine for classification of vertically partitioned data. *International Journal of Machine Learning and Cybernetics*, 6(1):109–118, 2015.

[80] J Hyma, P Sanjay Varma, SVS Nitish Kumar Gupta, and R Salini. Heterogeneous data distortion for privacy-preserving svm classification. In *Smart Intelligent Computing and Applications*, pages 459–468. Springer, 2019.

[81] Bettahally N Keshavamurthy, Asad M Khan, and Durga Toshniwal. Privacy preserving association rule mining over distributed databases using genetic algorithm. *Neural Computing and Applications*, 22(1):351–364, 2013.

[82] NV Muthu Lakshmi and K Sandhya Rani. Privacy preserving association rule mining without trusted party for horizontally partitioned databases. *International Journal of Data Mining & Knowledge Management Process*, 2(2):17, 2012.

[83] NV Muthu Lakshmi and K Sandhya Rani. Privacy preserving association rule mining in horizontally partitioned databases using cryptography techniques, 2012.

[84] Mohammed Golam Kaosar, Russell Paulet, and Xun Yi. Fully homomorphic encryption based two-party association rule mining. *Data & Knowledge Engineering*, 76:1–15, 2012.

[85] Andrew C Yao. Protocols for secure computations. In *Foundations of Computer Science, 1982. SFCS'08. 23rd Annual Symposium on*, pages 160–164. IEEE, 1982.

[86] Xuan Canh Nguyen, Hoai Bac Le, and Tung Anh Cao. An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases. In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on*, pages 1–4. IEEE, 2012.

[87] Mahmoud Hussein, Ashraf El-Sisi, and Nabil Ismail. Fast cryptographic privacy preserving association rules mining on distributed homogenous data base. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 607–616. Springer, 2008.

[88] Tamir Tassa. Secure mining of association rules in horizontally distributed databases. *IEEE Transactions on Knowledge and Data Engineering*, 26(4):970–983, 2014.

[89] Omar Abdel Wahab, Moulay Omar Hachami, Arslan Zaffari, Mery Vivas, and Gaby G Dagher. Darm: a privacy-preserving approach for distributed association rules mining on horizontally-partitioned data. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, pages 1–8. ACM, 2014.

[90] Nirali R Nanavati and Devesh C Jinwala. Privacy preserving approaches for global cycle detections for cyclic association rules in distributed databases. In *SECRYPT*, pages 368–371, 2012.

[91] Banu Ozden, Sridhar Ramaswamy, and Abraham Silberschatz. Cyclic association rules. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 412–421. IEEE, 1998.

[92] Feng Zhang, Chunming Rong, Gansen Zhao, Jinxia Wu, and Xiangning Wu. Privacy-preserving two-party distributed association rules mining on horizontally partitioned data. In *Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on*, pages 633–640. IEEE, 2013.

[93] David W Cheung, Jiawei Han, Vincent T Ng, Ada W Fu, and Yongjian Fu. A fast distributed algorithm for mining association rules. In *Parallel and Distributed Information Systems, 1996., Fourth International Conference on*, pages 31–42. IEEE, 1996.

[94] NV Muthu Lakshmi and K Sandhya Rani. Privacy preserving association rule mining in vertically partitioned databases. *International Journal of Computer Applications (0975–8887) Volume*, 2012.

[95] S KumaraSwamy, SH Manjula, KR Venugopal, SS Iyengar, and LM Patnaik. Association rule sharing model for privacy preservation and collaborative data mining efficiency. In *Engineering and Computational Sciences (RAECS), 2014 Recent Advances in*, pages 1–6. IEEE, 2014.

[96] K. Gurney. *An Introduction to Neural Networks*. CRC Press, 2018.

[97] Mauro Barni, Claudio Orlandi, and Alessandro Piva. A privacy-preserving protocol for neural-network-based computation. In *Proceedings of the 8th workshop on Multimedia and security*, pages 146–151. ACM, 2006.

[98] Tingting Chen and Sheng Zhong. Privacy-preserving backpropagation neural network learning. *IEEE Transactions on Neural Networks*, 20(10):1554–1564, 2009.

[99] Ankur Bansal, Tingting Chen, and Sheng Zhong. Privacy preserving back-propagation neural network learning over arbitrarily partitioned data. *Neural Computing and Applications*, 20(1):143–150, 2011.

[100] Jiawei Yuan and Shucheng Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 25(1):212–221, 2014.

[101] Hervé Chabanne, Amaury de Wargny, Jonathan Milgram, Constance Morel, and Emmanuel Prouff. Privacy-preserving classification on deep neural network. *IACR Cryptology ePrint Archive*, 2017:35, 2017.

[102] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210, 2016.

[103] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[104] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[105] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997.

[106] Zhijun Zhan, LiWu Chang, and Stan Matwin. Privacy-preserving naive bayesian classification. Technical report, SYRACUSE UNIV NY, 2004.

[107] Rakesh Agrawal and Ramakrishnan Srikant. Method and system for building a naive bayes classifier from privacy-preserving data, February 17 2004. US Patent 6,694,303.

[108] Peng Zhang, Yunhai Tong, Shiwei Tang, and Dongqing Yang. Privacy preserving naive bayes classification. In *International Conference on Advanced Data Mining and Applications*, pages 744–752. Springer, 2005.

[109] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive bayes classification. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 571–576. IEEE Computer Society, 2013.

[110] Ximeng Liu, Rongxing Lu, Jianfeng Ma, Le Chen, and Baodong Qin. Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. *IEEE journal of biomedical and health informatics*, 20(2):655–668, 2016.

[111] Chong-zhi Gao, Qiong Cheng, Pei He, Willy Susilo, and Jin Li. Privacy-preserving naive bayes classifiers secure against the substitution-then-comparison attack. *Information Sciences*, 444:72–88, 2018.

[112] Murat Kantarcıoglu, Jaideep Vaidya, and C Clifton. Privacy preserving naive bayes classifier for horizontally partitioned data. In *IEEE ICDM workshop on privacy preserving data mining*, pages 3–9, 2003.

[113] Jaideep Vaidya, Murat Kantarcıoğlu, and Chris Clifton. Privacy-preserving naive bayes classification. *The VLDB Journal*, 17(4):879–898, 2008.

[114] Justin Zhan, Stan Matwin, and Li Wu Chang. Privacy-preserving naive bayesian classification over horizontally partitioned data. In *Data Mining: Foundations and Practice*, pages 529–538. Springer, 2008.

[115] Xun Yi and Yanchun Zhang. Privacy-preserving naive bayes classification on distributed data via semi-trusted mixers. *Information systems*, 34(3):371–380, 2009.

[116] Stanley RM Oliveira and Osmar R Zaïane. Toward standardization in privacy-preserving data mining. In *ACM SIGKDD 3rd Workshop on Data Mining Standards*, pages 7–17, 2004.

[117] Rupa Parameswaran and D Blough. A robust data obfuscation approach for privacy preservation of clustered data. In *Workshop on Privacy and Security Aspects of Data Mining*, pages 18–25. Citeseer, 2005.

[118] Sara Hajian and Mohammad Abdollahi Azgomi. A privacy preserving clustering technique using haar wavelet transform and scaling data perturbation. In *Innovations in Information Technology, 2008. IIT 2008. International Conference on*, pages 218–222. IEEE, 2008.

[119] Liming Li and Qishan Zhang. A privacy preserving clustering technique using hybrid data transformation method. In *Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on*, pages 1502–1506. IEEE, 2009.

[120] RR Rajalaxmi and AM Natarajan. An effective data transformation approach for privacy preserving clustering 1, 2008.

[121] Jie Liu and Yifeng Xu. Privacy preserving clustering by random response method of geometric transformation. In *Internet Computing for Science and Engineering (ICICSE), 2009 Fourth International Conference on*, pages 181–188. IEEE, 2009.

[122] Srujana Merugu and Joydeep Ghosh. Privacy-preserving distributed clustering using generative models. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 211–218. IEEE, 2003.

[123] Jaideep Vaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM, 2003.

[124] Wenliang Du and Mikhail J Atallah. Privacy-preserving cooperative statistical analysis. In *Computer Security Applications Conference, 2001. ACSAC 2001. Proceedings 17th Annual*, pages 102–110. IEEE, 2001.

[125] Saeed Samet, Ali Miri, and Luis Orozco-Barbosa. Privacy preserving k-means clustering in multi-party environment. In *Proceedings of the International Conference on Security and Cryptography*, pages 523–531, Barcelona, Spain, July 2007.

[126] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu. Tools for privacy preserving distributed data mining. *SIGKDD Explor. Newsl.*, 4(2):28–34, December 2002.

[127] Mahir Can Doganay, Thomas B Pedersen, Yücel Saygin, Erkay Savaş, and Albert Levi. Distributed privacy preserving k-means clustering with additive secret sharing. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pages 3–11. ACM, 2008.

[128] Somesh Jha, Luis Kruger, and Patrick McDaniel. Privacy preserving clustering. In *European Symposium on Research in Computer Security*, pages 397–417. Springer, 2005.

[129] Chunhua Su, Feng Bao, Jianying Zhou, Tsuyoshi Takagi, and Kouichi Sakurai. Privacy-preserving two-party k-means clustering via secure approximation. In *null*, pages 385–391. IEEE, 2007.

[130] Paul Bunn and Rafail Ostrovsky. Secure two-party k-means clustering. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 486–497. ACM, 2007.

[131] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[132] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[133] Yan Zhao, Ming Du, Jiajin Le, and Yongcheng Luo. A survey on privacy preserving approaches in data publishing. In *Database Technology and Applications, 2009 First International Workshop on*, pages 128–131. IEEE, 2009.

[134] E Poovammal and M Ponnavaikko. Task independent privacy preserving data mining on medical dataset. In *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*, pages 814–818. IEEE, 2009.

[135] Yan Zhu and Lin Peng. Study on k-anonymity models of sharing medical information. In *Service Systems and Service Management, 2007 International Conference on*, pages 1–8. IEEE, 2007.

[136] Nissim Matatov, Lior Rokach, and Oded Maimon. Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, 180(14):2696–2720, 2010.

[137] Benjamin CM Funga, Ke Wangb, Lingyu Wanga, and Patrick CK Hungc. Privacy-preserving data publishing for cluster analysis. *Data & Knowledge Engineering, Elsevier journal*, 68(6):552–575, 2009.

[138] Dan Zhu, Xiao-Bai Li, and Shuning Wu. Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *Decision Support Systems*, 48(1):133–140, 2009.

[139] Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and Information Systems*, 28(1):47–77, Jul 2011.

[140] James Cheng, Ada Wai-chee Fu, and Jia Liu. K-isomorphism: privacy preserving network publication against structural attacks. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pages 459–470. ACM, 2010.

[141] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 506–515, April 2008.

[142] Benjamin C. M. Fung, Yan'an Jin, and Jiaming Li. Preserving privacy and frequent sharing patterns for social network data publishing. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 479–485, NY, USA, 2013. ACM.

[143] Mingxuan Yuan and Lei Chen. Semi-edge anonymity: graph publication when the protection algorithm is available. In *International Conference on Database Systems for Advanced Applications*, pages 367–381. Springer, 2012.

[144] M Vamsi Krishna and K VVS Narayana Murthy. Method to prevent re-identification of individual nodes by combining k-degree anonymity with l-diversity. *International Journal of Science Engineering and Advance Technology*, 2(11):742–745, 2014.

[145] Rangarajan Vasudevan, Ajith Abraham, and Sugata Sanyal. A novel scheme for secured data transfer over computer networks. *Journal of Universal Computer Science*, 11(1):104–121, 2005.

[146] Lei Xu, Chunxiao Jiang, Nengqiang He, Zhu Han, and Abderrahim Benslimane. Trust-based collaborative privacy management in online social networks. *IEEE Transactions on Information Forensics and Security*, 14(1):48–60, 2019.

[147] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 93–106. ACM, 2008.

[148] Lei Zou, Lei Chen, and M Tamer Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment*, 2(1):946–957, 2009.

[149] Xiaowei Ying and Xintao Wu. Randomizing social networks: a spectrum preserving approach. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 739–750. SIAM, 2008.

[150] M. Yuan, L. Chen, P. S. Yu, and T. Yu. Protecting sensitive labels in social network data anonymization. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):633–647, March 2013.

[151] Huseyin Polat and Wenliang Du. Privacy-preserving collaborative filtering. *International journal of electronic commerce*, 9(4):9–35, 2005.

[152] Alper Bilge and Huseyin Polat. An improved privacy-preserving dwt-based collaborative filtering scheme. *Expert Systems with Applications*, 39(3):3841–3854, 2012.

[153] Alper Bilge and Huseyin Polat. A comparison of clustering-based privacy-preserving collaborative filtering schemes. *Applied Soft Computing*, 13(5):2478–2489, 2013.

[154] Alper Bilge and Huseyin Polat. A scalable privacy-preserving recommendation scheme via bisecting k-means clustering. *Information Processing & Management*, 49(4):912–927, 2013.

[155] Javier Parra-Arnau, David Rebollo-Monedero, and Jordi Forné. A privacy-protecting architecture for recommendation systems via the suppression of ratings. *Int. J. Security, Appl*, 6:61–80, 2012.

[156] Huseyin Polat and Wenliang Du. Privacy-preserving top-n recommendation on distributed data. *Journal of the American Society for Information Science and Technology*, 59(7):1093–1108, 2008.

[157] Cihan Kaleli and Huseyin Polat. Privacy-preserving som-based recommendations on horizontally distributed data. *Knowledge-Based Systems*, 33:124 – 135, 2012.

[158] Cihan Kaleli and Huseyin Polat. *Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007. Proceedings*, chapter Providing Naïve Bayesian Classifier-Based Private Recommendations on Partitioned Data, pages 515–522. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[159] Ibrahim Yakut and Huseyin Polat. Privacy-preserving svd-based collaborative filtering on partitioned data. *International Journal of Information Technology & Decision Making*, 9(03):473–502, 2010.

[160] A. Jeckmans, Q. Tang, and P. Hartel. Privacy-preserving collaborative filtering based on horizontally partitioned dataset. In *2012 International Conference on Collaboration Technologies and Systems (CTS)*, pages 439–446, May 2012.

[161] H. Kikuchi, Y. Aoki, M. Terada, K. Ishii, and K. Sekino. Accuracy of privacy-preserving collaborative filtering based on quasi-homomorphic similarity. In *2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing*, pages 555–562, Sept 2012.

[162] Ibrahim Yakut and Huseyin Polat. Privacy-preserving hybrid collaborative filtering on cross distributed data. *Knowledge and Information Systems*, 30(2):405–433, 2012.

[163] Ibrahim Yakut and Huseyin Polat. Arbitrarily distributed data-based recommendations with privacy. *Data & Knowledge Engineering*, 72:239–256, 2012.

[164] Ibrahim Yakut and Huseyin Polat. Estimating nbc-based recommendations on arbitrarily partitioned data with privacy. *Knowledge-Based Systems*, 36:353 – 362, 2012.

[165] Burak Memis and Ibrahim Yakut. Privacy-preserving two-party collaborative filtering on overlapped ratings. *KSII Transactions on Internet and Information Systems (TIIS)*, 8(8):2948–2966, 2014.

[166] S. Katzenbeisser and M. Petkovic. Privacy-preserving recommendation systems for consumer healthcare services. In *2008 Third International Conference on Availability, Reliability and Security*, pages 889–895, March 2008.

[167] T Ryan Hoens, Marina Blanton, Aaron Steele, and Nitesh V Chawla. Reliable medical recommendation systems with patient privacy. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):67, 2013.

[168] J. Yuan and S. Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 25(1):212–221, Jan 2014.

[169] Xun Yi, Fang-Yu Rao, Elisa Bertino, and Athman Bouguettaya. Privacy-preserving association rule mining in cloud computing. In *Proceedings of the 10th ACM symposium on information, computer and communications security*, pages 439–450. ACM, 2015.

[170] Lifei Wei, Haojin Zhu, Zhenfu Cao, Xiaolei Dong, Weiwei Jia, Yunlu Chen, and Athanasios V Vasilakos. Security and privacy for storage and computation in cloud computing. *Information Sciences*, 258:371–386, 2014.

[171] Kawuu W Lin and Der-Jiunn Deng. A novel parallel algorithm for frequent pattern mining with privacy preserved in cloud computing environments. *International Journal of Ad Hoc and Ubiquitous Computing*, 6(4):205–215, 2010.

[172] Manisha Jindal and Mayank Dave. Data security protocol for cloudlet based architecture. In *Recent Advances and Innovations in Engineering (ICRAIE), 2014*, pages 1–5. IEEE, 2014.

[173] Yuan-Hung Kao, Wei-Bin Lee, Tien-Yu Hsu, Chen-Yi Lin, Hui-Fang Tsai, and Tung-Shou Chen. Data perturbation method based on contrast mapping for reversible privacy-preserving data mining. *Journal of Medical and Biological Engineering*, 35(6):789–794, Dec 2015.

[174] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao. Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security*, 11(8):1847–1861, 2016.

[175] Junzuo Lai, Yingjiu Li, Robert H Deng, Jian Weng, Chaowen Guan, and Qiang Yan. Towards semantically secure outsourcing of association rule mining on categorical data. *Information Sciences*, 267:267–286, 2014.

[176] Rasheed Hussain, Donghyun Kim, Junggab Son, Jooyoung Lee, Chaker Abdelaziz Kerrache, Abderrahim Benslimane, and Heekuck Oh. Secure and privacy-aware incentives-based witness service in social internet of vehicles clouds. *IEEE Internet of Things Journal*, 5(4):2441–2448, 2018.

[177] Fosca Giannotti, Laks VS Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui Wang. Privacy-preserving mining of association rules from outsourced transaction databases. *IEEE Systems Journal*, 7(3):385–395, 2013.

[178] Yehida Lindell. Secure multiparty computation for privacy preserving data mining. In *Encyclopedia of Data Warehousing and Mining*, pages 1005–1009. IGI Global, 2005.

[179] D. He, N. Kumar, H. Wang, L. Wang, K. K. R. Choo, and A. Vinel. A provably-secure cross-domain handshake scheme with symptoms-matching for mobile healthcare social network. *IEEE Transactions on Dependable and Secure Computing*, PP(99):1–1, 2016.

[180] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[181] Max Levchin, Luke Paul Nosek, Peter Thiel, and Scott Alan Banister. System and method for electronically exchanging value among distributed users, August 8 2006. US Patent 7,089,208.

[182] Rongxing Lu, Xiaodong Lin, and Xuemin Shen. Spoc: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency. *IEEE Transactions on Parallel and Distributed Systems*, 24(3):614–624, 2013.

[183] Pascal Paillier. *Public-Key Cryptosystems Based on Composite Degree Residuosity Classes*, pages 223–238. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[184] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015.

[185] Huseyin Polat and Wenliang Du. Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE International Conference on Data Mining*, pages 625–628. IEEE, 2003.

[186] Rupa Parameswaran and Douglas M Blough. Privacy preserving collaborative filtering using data obfuscation. In *2007 IEEE International Conference on Granular Computing (GRC 2007)*, pages 380–380. IEEE, 2007.

[187] Koji Miyahara and Michael J. Pazzani. *PRICAI 2000 Topics in Artificial Intelligence: 6th Pacific Rim International Conference on Artificial Intelligence Melbourne, Australia, August 28 – September 1, 2000 Proceedings*, chapter Collaborative Filtering with the Simple Bayesian Classifier, pages 679–689. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[188] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[189] Elisa Bertino, Latifur R Khan, Ravi Sandhu, and Bhavani Thuraisingham. Secure knowledge management: confidentiality, trust, and privacy. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 36(3):429–438, 2006.

[190] Koji Miyahara and Michael J Pazzani. Improvement of collaborative filtering with the simple bayesian classifier 1. *Information Processing Society of Japan*, 2002.

[191] Cihan Kaleli and Huseyin Polat. Privacy-preserving naïve bayesian classifier–based recommendations on distributed data. *Computational Intelligence*, 31(1):47–68, 2015.

[192] Huseyin Polat and Wenliang Du. Privacy-preserving top-n recommendation on horizontally partitioned data. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 725–731, Washington, DC, USA, 2005. IEEE Computer Society.

[193] T. A. N. Pham, X. Li, G. Cong, and Z. Zhang. A general recommendation model for heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3140–3153, Dec 2016.

[194] Amir Rostami and Hernan Mondani. The complexity of crime network data: A case study of its consequences for crime control and the study of networks. *PloS one*, 10(3):e0119309, 2015.

[195] Sean Evan Wise. *The impact of intragroup social network topology on group performance: understanding intra-organizational knowledge transfer through a social capital framework*. PhD thesis, University of Glasgow, 2013.

[196] Bin Zhou, Jian Pei, and WoShun Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.*, 10(2):12–22, December 2008.

[197] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, pages 181–190. ACM, 2007.

[198] Michael Hay, Gerome Miklau, David Jensen, Philipp Weis, and Siddharth Srivastava. Anonymizing social networks. *Computer Science Department Faculty Publication series*, page 180, 2007.

[199] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science National Taiwan University, Taipei 106, Taiwan, 2003.

[200] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra. Decision tree and svm-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3):1005–1016, June 2016.

[201] Jordi Casas Roma. *Privacy-Preserving and Data Utility in Graph Mining*. PhD thesis, Universitat Autònoma de Barcelona, 2014.

[202] Mingxuan Yuan and Lei Chen. Node protection in weighted social networks. In Jeffrey Xu Yu, Myoung Ho Kim, and Rainer Unland, editors, *Database Systems for Advanced Applications*, pages 123–137, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[203] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

List of Publications

- Published

  1) Harmanjeet Kaur, Neeraj Kumar and Shalini Batra, "An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system," *Future Generation Computer Systems, Elsevier*, Vol. 86, March 2018, pp. 297-307, https://doi.org/10.1016/j.future.2018.03.017, SCIE Indexed (IF: 4.63).
  2) Harmanjeet Kaur, Neeraj Kumar and Shalini Batra, "ClaMPP: A Cloud-based Multi-party Privacy Preserving Classification Scheme for Distributed Applications," *Journal of Supercomputing, Springer*, Vol. 75 (6), 2019, pp. 3046-3075, https://doi.org/10.1007/s11227-018-2691-0, SCIE Indexed (IF: 1.326).