PRIVACY-PRESERVING TWO-PARTY COLLABORATIVE FILTERING ON OVERLAPPED RATINGS

Master of Science Thesis

Burak MEMİŞ

Eskişehir, 2016

PRIVACY-PRESERVING TWO-PARTY COLLABORATIVE FILTERING ON OVERLAPPED RATINGS

Burak MEMİŞ

MASTER OF SCIENCE THESIS

Computer Engineering Program
Supervisor: Asst. Prof. Dr. İbrahim YAKUT

Eskişehir
Anadolu University
Graduate School of Sciences
February, 2016

JÜRİ VE ENSTİTÜ ONAYI

(APPROVAL OF JURY AND INSTITUTE)

Burak MEMİŞ'in **"Privacy-Preserving Two-Party Collaborative Filtering on Overlapped Ratings"** başlıklı tezi 12/02/2016 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, **Bilgisayar Mühendisliği** Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

		Ünvanı-Adı Soyadı	**İmza**
Üye (Tez Danışmanı)	**:**	**Yard. Doç. Dr. İbrahim YAKUT**	**.....**
Üye	**:**	**Doç. Dr. Hüseyin POLAT**	**.....**
Üye	**:**	**Yard. Doç. Dr. M. Müjdat ATANAK**	**.....**

.....

Enstitü Müdürü

ABSTRACT

PRIVACY-PRESERVING
TWO-PARTY COLLABORATIVE FILTERING
ON OVERLAPPED RATINGS

Burak MEMİŞ

Department of Computer Engineering
Anadolu University, Graduate School of Sciences, February, 2016

Supervisor: Asst. Prof. Dr. İbrahim YAKUT

To promote recommendation services through prediction quality, some privacy-preserving collaborative filtering solutions are proposed to make e-commerce parties collaborate on partitioned data. It is almost probable that two parties hold ratings for the same users and items simultaneously; however, existing two-party privacy-preserving collaborative filtering solutions do not cover such overlaps. Since rating values and rated items are confidential, overlapping ratings make privacy-preservation more challenging. In this dissertation, firstly, the subject of how the personal data distribution occurs in information systems will be handled and personal data preserving solutions will be elucidated. Then, how to estimate predictions privately based on partitioned data with overlapped entries between two e-commerce companies is examined. It is considered both user-based and item-based collaborative filtering approaches and proposes novel privacy-preserving collaborative filtering schemes in this sense. It is also evaluated schemes using real movie dataset, and the empirical outcomes show that the parties can promote collaborative services using our schemes.

Keywords: Collaborative filtering, Arbitrarily partitioned data, Overlapped ratings, Pearson similarity, Slope-one predictor, Privacy

ÖZET

ÇAKIŞMALI OYLAR ÜZERİNDEN GİZLİLİK KORUMALI
İKİ PARTİLİ ORTAK FİLTRELEME

Burak MEMİŞ

Bilgisayar Mühendisliği Anabilim Dalı
Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Şubat, 2016

Danışman: Yard. Doç. Dr. İbrahim YAKUT

Tavsiye hizmetlerini öneri kalitesini artırarak geliştirmek için önerilen gizlilik koruyucu ortak filtreleme çözümleri e-ticaret şirketlerinin paylaşılmış veri üzerinden işbirliği yapmalarına imkân sağlar. İki tarafın aynı anda aynı kullanıcıların aynı ürünler için beğeni değerleri tutması muhtemeldir; ancak var olan iki taraflı gizlilik koruyucu ortak filtreleme çözümleri, bu tür çakışmaları ele almamıştır. Kullanıcı oyları ve oylanan öğeler gizli olduğundan çakışan oylamalar gizlilik korumayı daha da güçleştirecektir. Bu çalışmada ilk olarak, bilişim uygulamalarında kişisel veri paylaşımının nasıl gerçekleştiği ele alınıp çeşitli yaklaşımlar ile kişisel verilerin korunmasına yönelik çözümler ifade edilecektir. Daha sonra, iki e-ticaret firması arasında paylaşılmış verilerde çakışan girdiler ile nasıl tahmin yapılacağı araştırılacaktır. Bu bağlamda kullanıcı ve ürün tabanlı ortak filtreleme yöntemleri ele alındı ve özgün gizlilik korumalı ortak filtreleme yöntemleri önerildi. Önerilen yöntemler gerçek veri setleri kullanılarak değerlendirildi ve deneysel sonuçlar şirketlerin bu yöntemleri kullanarak tavsiye servislerini iyileştirebileceğini göstermektedir.

Anahtar Kelimeler: İşbirlikçi filtreleme, Rastgele bölünmüş veri, Çakışan Oylar, Pearson benzerliği, Eğim-bir öngörücüsü, Gizlilik

ACKNOWLEDGEMENTS

I would like to thank my advisor Asst. Prof. Dr. İbrahim Yakut for his excellent guidance, support, patience, and motivation to my research. It was a chance and pleasure to work with him during my thesis. I would also like to thank Asst. Prof. Dr. Gökhan Güneysu for helping and giving his best suggestions. I thank to Assoc. Prof. Dr. Hüseyin Polat, Asst. Prof. Dr. Mustafa Atanak, and Assoc. Prof. Dr. Alpaslan Duysak to participate in my defense committee.

In addition, a thank you to my colleagues, Emrah Demir and Latif Sağlam for their support and helping me while I was busy.

Special thanks go to my sister to cook delicious food while I was working for my dissertation.

Last but not least, I would like to thank my parents for their patience, support, and love.

Burak Memiş

February, 2016

02/03/2016

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ
(DECLARATION OF CONFORMITY FOR ETHIC RULES AND PRINCIPLES)

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz, ve bilgilerin sunumu olmak üzere tüm aşamalardan bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilemeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bu durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

…….

……

CONTENTS

Page

TITLE PAGE ……... i

APPROVAL OF JURY AND INSTITUTE ii

ABSTRACT …….. iii

ÖZET .....…… iv

ACKNOWLEDGEMENTS …….……... v

DECLARATION OF CONFORMITY FOR ETHIC RULES AND PRINCIPLES vi

CONTENTS ……..…..... vii

LIST OF TABLES ……..……. ix

LIST OF FIGURES ……..……... x

ABBREVIATIONS ……..……... xi

1. INTRODUCTION …….….. 1

1.1. Data Mining …….…. 2

1.2. Recommender Systems …….…... 3

1.3. Collaborative Filtering ……..... 4

1.4. Privacy-Preserving Collaborative Filtering …….. 6

1.5. Methods for Preserving Privacy ……..….. 9

1.6. Outline of the Thesis ……..... 10

2. PROTECTING PERSONAL DATA IN INFORMATION SYSTEMS …..... 11

2.1. Introduction ……. 11

2.2. Definition and Historical Development ……. 12

2.2.1. Definition …… 12

2.2.2. Historical development ……. 13

2.3. International Regulations and Our Country ……..….. 14

2.3.1. Organization for economic cooperation and development (OECD) 15

2.3.2. United nations …… 15

2.3.3. Council of europe …….. 15

2.3.4. European union ……. 16

2.3.5. Asia-pacific economic cooperation (APEC) …… 16

2.3.6. Regulations in Turkey ……... 16

2.4. Data Sharing in Information Systems ……..……... 17
2.4.1. User-agency data sharing ……..……. 18
2.4.2. Agency-agency data sharing ……. 20
2.5. Chapter Summary ……...…… 21
3. PRIVACY-PRESERVING USER-BASED COLLABORATIVE FILTERING ON OVERLAPPED RATINGS ……. 23
3.1. User-based Collaborative Filtering with Pearson Similarity …….. 23
3.2. Arbitrarily Partitioning and Overlapped Ratings ……... 23
3.3. Privacy Problem …….. 24
3.4. Privacy-Preserving User-based CF on Overlapped Ratings ……... 25
3.4.1. Preprocessing ……. 25
3.4.2. Similarity computation ……. 26
3.4.3. Prediction computation …… 27
3.4.4. Removing overlaps ……..... 28
3.5. Analysis of the Scheme …… 29
3.6. Experimental Results …….. 32
3.7. Chapter Summary ……..…. 34
4. PRIVACY-PRESERVING ITEM-BASED COLLABORATIVE FILTERING ON OVERLAPPED RATINGS 36
4.1. Item-based Collaborative Filtering with Slope-one Predictor ……. 36
4.2. Privacy-Preserving Item-based CF on Overlapped Ratings ……..... 37
4.2.1. Deviation computation …… 37
4.2.2. Prediction computation …….. 38
4.3. Analysis of the Schemes …… 39
4.4. Experimental Results …… 40
4.5. Chapter Summary …… 42
5. CONCLUSIONS ……... 43
REFERENCES ……. 44
CIRCULUM VITAE

LIST OF TABLES

Table 2.1. Approaches for Protecting Personal Data 22

Table 3.1. Ratio of Overlaps (%) vs. Density and Filling Level 33

Table 3.2. Overall performance with varying density 34

Table 4.1. Overall performance with varying density 41

LIST OF FIGURES

Figure 1.1. Typical CF Process 5

Figure 2.1. Historical Development of Protection of Personal Data 14

Figure 2.2. Stages of Data Sharing 18

Figure 3.1. Partitioned data with sample overlapped ratings 24

Figure 3.2. Privacy-preserving user-based CF on overlapped ratings 26

Figure 3.3. Accuracy with respect to varying level of filling 33

Figure 4.1. Privacy-preserving item-based CF on overlapped ratings 37

Figure 4.2. Accuracy with respect to varying level of filling 41

ABBREVIATIONS

a	: Active User
ADD	: Arbitrarily Distributed Data
APD	: Arbitrarily Partitioned Data
APEC	: Asia-Pacific Economic Cooperation
$card_{jk}$	: Cardinality of Set of Users who have rated both Items j and k
CF	: Collaborative Filtering
DM	: Data Mining
dev_{jk}	: Deviation between Items j and k
EU	: European Union
HCs	: Homomorphic Cryptosystems
KA	: Public Key of A
m	: Number of Items
MAE	: Mean Absolute Error
MLP	: MovieLens Public
MP	: Master Party
n	: Number of Users
NBC	: Naive Bayesian Classifier
OECD	: Organization for Economic Cooperation and Development
p_{aq}	: Prediction on Item *q* for User *a*
PPCF	: Privacy-Preserving Collaborative Filtering
PrivateDevs	: Private Deviation Computation Protocol
PrivateSims	: Private Similarity Computation Protocol
PrivatePreds	: Privately Prediction Computation Protocol
PS	: Plain Scheme
P2P	: Peer to Peer
q	: Target Item
SSL	: Secure Sockets Layer
SVD	: Singular Value Decomposition
TOR	: The Onion Router
u	: Train User
US	: Ultimate Scheme

$v_d s$	: Default Vote
VPN	: Virtual Private Network
w_{au}	: Similarity between a and u
$\xi_K(x)$	: Encrypted Value of x with Public Key K
δ	: Density
θ	: Level of Filling
τ	: Threshold

1. INTRODUCTION

Recommender systems are techniques, which use users' ratings or preferences to produce predictions about target items or top-*n* lists. These systems are used in several applications such as movies, books, film, search queries, and so on. Moreover, a variety of e-commerce companies are used these systems to make a high profit. Recommender systems are designed based on some approaches such as collaborative filtering, content-based filtering and hybrid recommender systems. Collaborative filtering (CF) as a recommender system is useful in the sense that it does not require content analysis for items and provides the ability to recommend items on taste information [1]. Furthermore, recommender system mechanisms have some problems: such as data scarcity, scalability, privacy, and so on.

E-commerce companies such as being newly established or expanding product categories suffer from scarcity of ratings and protection of users' data. To prevent users' data from privacy threats some technical and legal regulations are made. Consequently, such companies are unable to offer quality CF services. One solution for such problems is to collaborate with another data company for featured recommendation services. However, rating data can be subject to privacy risks [2] and e-commerce companies are responsible for the confidentiality of data held by these companies [3,4]. In order to encourage such parties for cooperation, privacy metrics need to be provided. For this reason, a range of privacy-preserving collaborative filtering (PPCF) schemes is proposed considering partitioned data [3,5]. By means of privacy-preserving contribution of bonus data, data scarcity problem can be tackled and companies can provide recommendations having satisfactory quality and quantity.

This thesis focuses on the following problems: *how can personal data be protected in information systems*? And *how can two parties end up with partitioned data having overlapped ratings promote recommendation services ensuring corporate data privacy*?

In Section 1.1, Data Mining is explained while recommender systems are introduced in Section 1.2. While definition collaborative filtering is defined in Section 1.3, privacy-preserving collaborative filtering (PPCF) is introduced in Section 1.4. After explaining privacy-preserving methods in Section 1.5, outline of the thesis is presented is Section 1.6.

1.1. Data Mining

Data mining is a computer science technique which is to get the information or patterns from huge amount of data in databases. This process also known as knowledge discovery in databases. Data analysis techniques are used to build models for exploring these knowledge or patterns. In general, two types of models are being used in data mining. One of them is a predictive model and the other one is a descriptive model. Predictive modeling uses a known data to build a model, which is used to predict results. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place [6]. Descriptive modeling tries to descript patterns or knowledge in existing data.

Data Mining was introduced in 1990s as a term, but the evaluation of data mining has a long history. In 1960s, collection of data has appeared with using computers, disks, and tape recorders. Researchers have started to answer some decision problems with using these data. For example, the question of making a total profit in last year is answered. DBMS software products were programmed, and designed in the 1960s and 1970s. The relational DBMS products were developed during the 1970s and came to prominence during the 1980s and 1990s [7].

Data mining techniques have several benefits. For example, companies hold on the market with using these techniques. Because they can specify a new marketing strategies while customers' behaviors change day by day. Besides, these techniques will help to find new customers, who will make a high profit margin. At the same time, data mining techniques have some issues. One of them is performance issue. Data mining techniques and algorithms must be efficient in order to extract the knowledge from huge amount of data in databases. The other one is mixed data types issue. It is a challenge to enable data mining approaches to deal with mixed data types because there are difficulties in finding a measure of similarity between objects with mixed data type attributes [8]. The third one is trying to extract different kinds of knowledge in database. To overcome these issues Han et al. [9] study techniques for the discovery of various kinds of knowledge, including generalization, characterization, discrimination, association, classification, clustering, and so on.

Data mining techniques are used in many research areas, including marketing, banking, genetics, communication, and criminology. For example, credit card fraud might be stopped with using data mining model, which tracks and expects personal

credit card habits. Thus, the using of lost or stolen credit card can be blocked. Furthermore, data mining techniques are frequently used to predict the ratio of catching a disease in genetics. Lee and Stolfo [10] propose a systematic framework that uses data mining techniques for intrusion detection. Jourdan, Dhaenens, and Talbi [11] study a genetic algorithm dedicated for a particular feature selection problem encountered in genetic analysis of different diseases.

1.2. Recommender Systems

Recommender Systems have recently become very important and popular in the context of e-business applications [12,13]. Schafer J. B. et al. study how recommendation systems help e-business sites increase sales [14]. Such systems not only facilitate decision process of users having limited time for consuming on the web but also inform Internet users about music, film, and books which they intend to taste. These systems create a user model with using a collection of personal data, such as what a user clicks on in the online website, time spent looking at a page, etc. to inform users.

Several approaches are used to the design of recommender systems. One of them is content-based filtering. It may be used several applications, such as recommending web pages, news articles, restaurants, television programs, and items for sale [15]. In a content-based recommender system, the user is recommended items similar to the user has liked in the past. Last et al. study to develop an innovative methodology for abnormal activity detection on the web content [16]. Moreover, Bogdanov et al. [17] propose a content-based user modeling technique for music recommendation and visualization of the user's musical preferences.

Another common approach is collaborative filtering. Collaborative filtering (CF) is a technique, which is used in recommender systems. CF is the process of filtering or evaluating items using the opinions of other people [18]. CF as a recommender systems in useful in the sense that it does not require content analysis for items and provides the ability to recommend items on taste information [1]. The main purpose of using CF algorithm is to give the best recommendations to people with respect to given ratings of similar set of users or items.

Nevertheless, some recommender algorithms develop a model to provide item recommendation to improve the prediction and find a solution for scalability problem. These algorithms use machine learning algorithms such as Bayesian network CF,

clustering CF, and rule-based CF approaches to build a model. Bayesian network CF model formulates a probabilistic model with a node corresponding to each item in the domain [19]. The clustering CF model uses data partitioning and clustering algorithms to partition the set of items based on user rating data and computing predictions independently within each partition [19,20]. Pham et al. study a clustering approach to CF recommendation technique to apply on the social network of users to propose the recommendations [21]. The rule-based CF model uses association rules to find association between co-purchased items for making a prediction [22,23].

Furthermore, hybrid approach combines memory-based algorithms such as user-based or item-based algorithms and model based algorithms such as Bayesian network [24]. This approach tries to overcome some limitations of traditional collaborative filtering such as accuracy, scalability, scarcity, and so on. Ghazanfar et al. [25] study kernel mapping recommender to make reliable recommendations under sparse. Similarly, to improve accuracy of recommender systems and scarcity of data Badaro et al. [26] introduce a hybrid approach based on simultaneous combination of user-based and item-based collaborative filtering. Besides, Google news recommender system algorithm uses three approaches such as collaborative filtering using MinHash clustering, Probabilistic Latent Semantic Indexing (PLSI), and covisitation counts and combines recommendations from these approaches using a linear model [27].

1.3. Collaborative Filtering

Collaborative filtering (CF) is commonly used for recommender systems. CF makes recommendations based on other people's tastes. Because of using other people's tastes, firstly, CF algorithm collects active user's data, which is called rates of active user's and these rates show the preferences of a user. Active user is a customer who wants a prediction for a target item *q*. In general, CF algorithm uses implicit and explicit ratings to find preferences of active user. In implicit rating, active user doesn't give a rate directly and CF algorithm observes of behaviors of active user. An example of implicit rating is keeping track of what kind of music is listened by active user in the online website. Besides, active user rates the item explicitly in explicit rating. For example, active user can give a rate to the item on a 1-5 scale.

Figure 1.1 shows that user ratings are taken as input for CF algorithm, while prediction about q and top-*N* lists of recommended items. CF algorithm mainly consists

of three steps [1]: At first, CF algorithm calculates similarity of each user respect to an active user. To calculate similarity, CF algorithm uses some formulas. These are *Pearson Correlation*, *Spearman Correlation*, and so on. Then, CF algorithm selects neighborhoods for an active user with using weight thresholding or best-n neighbors methods. Finally, CF process will perform tasks. CF process has mainly two types of tasks. One of them is how much active user will like q. Moreover, CF process will create for an active user a top-N list of recommended items.

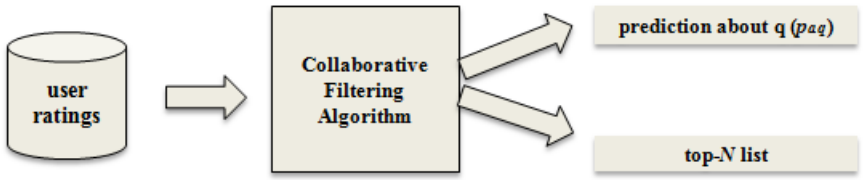

Figure 1.1 Typical CF Process

Figure 1.1. *Typical CF Process*

There are some challenges for collaborative filtering algorithms. One of them is the cold start problem. In the cold start problem, new user must rate enough number of rates to get a reliable recommendation from CF algorithm. The other one is the data scarcity. Sparse data can reduce the quality of the recommendation. Besides, growth of user and item causes a scalability problem. If data set is too large, to make a prediction will be slow. Furthermore, CF algorithms have to overcome privacy protection problems. To handle this problem privacy-preserving collaborative filtering is proposed.

There are two different CF approaches with respect to reference entities; these are user-based and item-based methods. In user-based CF methods, user-to-user relations based on similarity and proximity metrics are key elements to drive recommendation mechanisms. Typically, similarities are computed between users, and for each user, neighbor users are determined from the most similar users. Output predictions and recommendations are computed over neighbor users' similarities and ratings. Since, Pearson similarity is representative and widely utilized in user-based recommendation algorithms [28], we are going to examine such similarity metrics through our user-based CF investigation. User-based CF was appeared in the GroupLens Research firstly [29]. GroupLens is a system for collaborative filtering of netnews, to help people find articles

they will like in the huge stream of available articles. Hill et al. [30] present BellCore video recommender system. This recommender system also uses user-based CF. Besides, Shardanand and Maes [31] introduce The Ringo music recommender, which makes personalized recommendations for music albums and artists. Ringo calculates similarities between the interest profile of that user and those of other users.

In order to achieve more accurate CF results in more scalable ways, item-to-item relations are considered, and succeeding CF studies show that item-based *CF* approaches give satisfactory results and even outperform user-based CF in terms of performance and prediction quality [22]. Since item relations are more static than user relations, item similarities can be computed off-line to achieve faster online response with more throughputs. Since item-based CF notion introduced by Sarwar et al. [22], many item-based solutions are proposed [32,33]. In this sense, Lemire and Machlan [32] proposed Slope-one algorithms for recommender systems based on the *popularity differential* intuition. Ratings differences for two item vectors are the key issue to evaluate item-to-item deviations and this makes the method simple but effective to produce predictions. They have been shown to be accurate even with sparse datasets while being updatable on the fly [34]. Amazon.com recommendation system generates recommendations based on customers who are most similar to the user and uses a cosine measure to calculate similarity between each item pairs [35]. In this work with respect to such prominent features, we investigate slope-one predictor that was proposed by Lemire and Mahlachlan [32].

1.4. Privacy-Preserving Collaborative Filtering

Data mining (DM) is a science to extract information or knowledge from a huge data. While DM has some advantages, it has also number of problems including privacy concerns. To achieve DM tasks considering privacy concerns, there are several studies. One of them is randomized approach. In this approach noise can be added to values of data and so privacy of users' data can be preserved. Agrawal and Srikant [36] propose a privacy-preserving data mining technique, which is used randomizing function to perturb user's record with sensitive values. Cryptographic Approach is also used in privacy-preserving data mining. Pinkas [37] studies cryptographic research on secure distributed computation, and their applications to data mining. Furthermore, anonymization techniques are used to preserve privacy in data mining. Anbazhagan K.

et al. [38] propose statistical anonymization methods for privacy-preserving data mining.

Collaborative Filtering algorithms have also privacy issues similar to data mining algorithms. First studies about privacy issues were proposed by Canny [39]. He studies an algorithm in which users can compute a public "aggregate" data with preserving individual users' data. Furthermore, he uses a homomorphic encryption to calculate sum of encrypted vectors without disclosing users' individual data. His study also can be implemented to peer-to-peer (P2P) system with untrusted servers. Kaleli and Polat [40] also study caring about privacy-preserving collaborative filtering (PPCF) for P2P networks. In this study, they focus on to produce naïve Bayesian classifier (NBC)-based recommendations with preserving users' privacy. Moreover, a fully distributed collaborative filtering method, which is self-organizing and operates in a distributed way, is proposed by Wang et al. [41]. This method is a promising technique to facilitate filtering for relevant data in P2P networks.

Mainly there are two problems in PPCF. These problems are occurred between users-data holder(s) or two or more data holders. Data holder is a company or agency that holds users gathered from many customers and performs filtering services with other companies by sharing data. The most important problem is that users' don't want to share their information without preserving privacy. Firstly, PPCF algorithms are encountered while providing recommendations to masked data with preserving users' privacy. Calandrino et al. [42] propose an algorithm to ensure privacy metrics while providing CF services because of collecting and processing user profiles could be threat to privacy. Furthermore, Xiong et al. [43] propose a comprehensive approach, called Privacy pReserving Identity and Access Management scheme, referred to as PRIAM, which is able to satisfy security requirements in cloud computing because of each cloud service has numerous users and it is important to preserve privacy mechanism for each users.

As a privacy-preserving memory-based collaborative filtering scheme with respect to shilling attacks, Gunes et al. [44] study the modified versions of two low-knowledge shilling attacks models and integrate them in masked databases by employing random perturbation protocol. Moreover, without violating users' privacy, Lathia et al. [45] propose a new measure of similarity, which achieves prediction accuracy successfully. Zhang et al. [46] introduce a two-way communication privacy-

preserving scheme in which users perturb their ratings for each item based on the server's guidance instead of using an item-variant perturbation with using a modified perturbation techniques in PPCF. To obfuscate parts of users' profile, Berkovsky [47] propose a decentralized CF model with storing users' profiles only on the client side. In this approach users' profiles is stored several different locations and thus the risk of having the users' data exposed to a malicious attacker is being reduced.

Because of the data scarcity problem, data holders want to share users' data between each other. Sharing users' data may cause privacy problems nevertheless increasing prediction quality. Still data holders want to collaborate to achieve correct predictions. Yakut and Polat [48] study item-based predictions on arbitrarily distributed data (ADD) between two e-commerce sites with preserve their privacy. As a privacy-preserving collaborative filtering over distributed data, Basu et al. [34] propose a solution based on the weighted slope one predictor and uses homomorphic encryption. Moreover, preserving privacy on partitioned data with using naïve Bayesian classifier (NBC)-based CF tasks is proposed by Kaleli and Polat [49].

As a privacy-preserving scheme to estimate naïve Bayesian classifier-based predictions on arbitrarily partitioned data between two parties, Yakut and Polat [3] propose a method to provide binary ratings-based predictions on partitioned data with preserving online vendors' confidentiality requirements. Another study is to produce recommendations privately using Singular value decomposition (SVD) [50]. In this paper, they show that how to provide SVD-based referrals on partitioned data with ensure data holders' privacy. Another study is about preserving privacy in merging recommender system databases using a novel algorithm based on ElGamal scheme of homomorphic encryption [51]. Moreover, Polat and Du [5] present a scheme for binary ratings-based-top N recommendation on horizontally partitioned data, in which two parties own disjoint sets of users' for the same items while preserving their privacy. They also study a privacy-preserving protocol for CF grounded on vertically partitioned data [52].

Kaleli and Polat [53] study how to provide predictions based on vertically distributed data (VDD) among multiply parties with preserving their confidentiality. In this paper, users are first grouped into various clusters off-line using self-organizing map clustering while protecting the online vendors' privacy. The same authors [54] examined how to provide recommendations using rating-derived trust metrics on

vertically distributed data with privacy. Rather than investigating symmetrically behaving parties, Zhao et al. [55] introduced shared collaborative filtering approach in which parties have asymmetric roles, i.e., while contributor party's data improves the beneficiary party's CF performance, privacy of contributed data cannot be compromised. In all work examining horizontal and vertical partitioned data, no overlaps are expected since authors concentrate on perfectly disjoint set of users or items. Bilge et al. [56] reviewed the state-of-start techniques, from the viewpoint of privacy basics of PPCF, and recently developed mechanisms with the emphasis on the partitioning data.

1.5. Methods for Preserving Privacy

To ensure users' privacy, some techniques such as randomization-based techniques and homomorphic cryptosystems (HCs) are used. General idea for randomization-based techniques, to add random values to users' ratings and send these ratings to recommender systems. In this study, randomized vote filling procedure where default votes can be row mean, column mean, or overall mean from available ratings of a party P is used to preserve users' privacy. Also Gong [57] presents a collaborative filtering algorithm based on randomized perturbation techniques and secures multiparty computation. Besides, Polat and Du [58] propose a randomized perturbation technique to preserve privacy while still producing accurate recommendations results.

Based on homomorphic cryptosystem, Paillier HC [59] can perform addition of two numbers as ciphertext and obtain encrypted version of the actual sum. Suppose that a and b are two numbers and ξ_K is encryption function with public key (K). Then, the ciphertexts of the numbers are $\xi_K(a)$ and $\xi_K(b)$ and their multiplication is $\xi_K(a) \times \xi_K(b) = \xi_K(a + b)$. Additionally in an analogous manner multiplication of plaintext can be performed as $\xi_K(a)^b = \xi_K(ab)$. Paillier HC has self-blinding property permitting public modification of ciphertexts by multiplying with R^N without affecting the plaintext, where R is a random integer value and N is modulus of the operated public cryptosystem.

1.6. Outline of the Thesis

In the following chapter, protecting of personal data in information systems is studied. While user-based collaborative filtering on overlapped ratings is proposed in Chapter 3, item-based collaborative filtering on overlapped ratings is presented in Chapter 4. Finally, in Chapter 5, conclusions are drawn and future research directions are introduced.

2. PROTECTING PERSONAL DATA IN INFORMATION SYSTEMS

Along with the proliferation of the information technologies, the protection of personal data becomes a severe problem. As information systems take much more place in our life day by day, the requirements to protect the personal data becomes crucial issue and requirements in this context come up with the solution approaches for them. In this study, after giving the definition and the historical development about the protection of the personal data, we will examine the legal regulations in the international scope and in our country. In the light of the concerning regulations, the subject of how the personal data distribution occurs in information systems will be handled and personal data protecting solutions will be elucidated.

2.1. Introduction

Along with the development of technology, communication instruments, especially computers, started to take part in all fields of our lives. These technologies make our lives easier. Besides, they come with many problems too. Sharing data easily via these technologies and lack of knowledge about data privacy leave users vulnerable. 20. Significant legal regulations are made in the field of fundamental rights and freedoms around the world with the century [60, 61]. Concordantly, some regulations become necessary on communication technologies, especially in computer usage. After 1950s some essential regulations about protection of personal data actualized consecutively especially in Europe. Although there is no specific law on this issue, it's guaranteed with several legal regulations to protect personal data. Also, protection of personal data is assured by adding provision to The Constitution of the Republic of Turkey article no: 20 regarding to law no: 5982 Making Changes in Some Articles of The Constitution of the Republic of Turkey under the date of 7/5/2010 [62]. The right of personal data protection also protects individuals' freedom together with itself. However, protection of data itself is not about protection of personal data but protection of data. In this respect, protection of personal data serves for personal data protecting within fundamental rights and freedoms [63].

As a result of developing technologies, the idea that without legal regulations, individuals won't be able to develop their personality before data processing actions of public body and won't be able to take place in democracy has emerged [63]. It can be said that there are three factors basically in arising of law of personal data protection:

i. Need for personal data by various organizations
ii. Developments in technology
iii. Arising consideration as a result of developments in surveillance technologies [64].

A movie, "Person of Interest" which is broadcasted nowadays, shows how convenient it is to make law of protection of personal data. In this movie, cell phone conversations, e-mails and daily actions of people are recorded via a designed machine. And in this way, person whose data are recorded should feel himself/herself secure. And this will be provided by law of personal data protection. Law of personal data protection gives the right to find out which personal data of real person are collected and treated by whom and for whom [65].

2.2. Definition and Historical Development

In this section, some definitions about information systems' are given and history of protecting personal data is explained in detail.

2.2.1. Definition

We frequently encounter the word confidentiality as "mahremiyet" in Turkish and privacy as "gizlilik" in informatics applications. Confidentiality is defined as a term not known by everybody and may harm the individual if it is known by everyone. And according to Warren and Brandeis, it is defined as a right that guarantees common civil privileges [66]. Privacy, as one of the most basic objectives of information security, means protecting content of the knowledge from anyone except authorities.

Beings that are competent with rights and obligations are called individual in juridical literature. Individuals are divided to two articles in our civil law as real person and legal person. Real people are individuals. Legal people are merchandise and human communities which are established to accomplish a specific purpose.

Personal data is defined as a term which states every kind of data regarding to a specific or specifiable person [62]. And is stated in a sentence of constitutional court that personal data is "all the data of a person who is or can be identified." [67]. As one can understand from these definitions, main factor of personal data is, it's being belong to an identifiable person even if it is not identifiable itself. Other factors that rise from these definitions are the terms data and information. Data can be defined as raw

information. And information is defined as significant data which is acquired by searching and using senses.

Informatics applications mean collecting data by hand or by automatic methods and to turn them into useful information. And increasing data quantity enabled development of some technologies which will help them to be accessible. And so, some factors such as data base systems, data warehouse, data mining occurred. Data base is defined as updatable, erasable, portable regular associative information in computer literature. Data warehouse is a storage in which related information are stored, interrogated and required transactions can be done. And data mining is to attain useful information from database and data warehouse.

Privacy in personal data protecting process means, usage of this data by relevant-authorized people or performing the required process. In addition to this, it's possible share data with other people or agencies in some exceptional circumstances. For instance, after September 11 attack, information about people who traveled to America was shared with USA by air carriers [64].

2.2.2. Historical Development

Even though technological developments make our lives easier in any field, they bring some problems with them. And one of these problems is, as it is mentioned on previous section, the problem of protecting these data which occurred as a result of shearing personal data. Along with these problems, legal regulations regarding to protecting of data started to arise in 2nd half of 1900s. Protecting data, basically aims to protect not "data" but related people [64]. For that reason, it's understood that fountain head of law of protecting data is protecting people [68]. Initiative legal regulations regarding to protecting personal data were seen especially in Europe and USA. And after that, some national and international arrangements arouse.

Legal regulations since the term of protection of personal data had existed until today are given in Figure 2.1 chronologically. Some of the terms that are used for protection of data which are used in our day, started to be subject for academic studies 100 years ago. And with the study of Right of Privacy in 1890, which is accepted as one of the most important studies in this field, the terms privacy and confidentiality came to light [66]. As protection of personal data is a rapidly spreading term, some clauses existed about this subject in EU Universal Declaration of Human Rights and European

Convention of Human Rights. Together with 1950s, when computer usage became a part of our lives, protection of personal data started to take place in legal arrangements. First legal regulation regarding protection of data was seen in Hessen, in Germany. And this regulation is followed by first national one which is performed to protect data in Switzerland in 1973. Later, some legal arrangements were made for protection of personal data in USA in 1974, in Portuguese in 1976, in Germany in 1977, and in Spain in 1978. And in 1980, Organization for Economic Cooperation and Development (OECD) released directory principles regarding to this subject. In 1981, first development in protection of personal data was realized by Council of Europe. In 1990, European Union released Directory Principles Regarding to Computerized Personal Data Files. In 1995, low no. 95/46/EC "European Parliament and Council of Europe Directive on European Parliament and Council of Europe Directive on protection of individuals with regard to the processing of personal data and on the free movement of such data" which is obligatory for member states, was released by European Union. With the legal regulations in our country, after the referendum which was arranged in 2010, related provision was added to 20th article of Turkish Constitution.

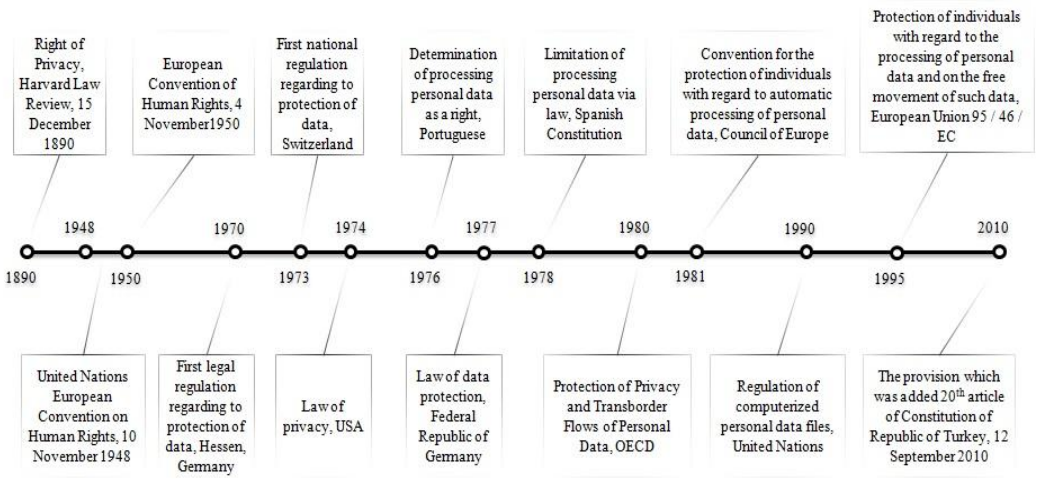

Figure 2.1. *Historical Development of Protection of Personal Data*

2.3. International Regulations and Our Country

In this section, firstly, international organizations, which they are responsible to release some regulations to protect personal data. These regulations specify to gather, process, and hide personal data for countries. Some regulations are compulsivity and

countries must obey them. Secondly, regulations of protecting personal data in Turkey will be proposed.

2.3.1. Organization for economic cooperation and development (OECD)

The first organization, which has brought up protection of personal data subject to agenda internationally, is OECD [69]. OECD has released directory principles regarding to protection of personal data [70]. With the Directory Principles of OECD, personal data are guaranteed internationally. Directory principles of OECD serves as a recommendation and these principles are not binding for member states.

2.3.2. United nations

One of the most important developments of United Nations for protection of personal data is "Directory Principles for Computerized Personal Data Files" which was confirmed in 1990. According to Directory Principles of United Nations, assurances which are to be guaranteed in national law system are related to articles given below:

i. Principle of collecting and processing with legal and fair procedures
ii. Principle of data accuracy
iii. Principle of purpose certainty
iv. Principle of access of relative person
v. Non-discrimination principle
vi. Principle of data security
vii. Inspection and suction
viii. Extraterritorial data flow [64].

2.3.3. Council of europe

Although European Convention of Human Rights which was accepted in 4th November 1950 by the Council doesn't include direct regulations about protection of personal data, there are provisions in article no: 8 "Respect for privacy and family life". First development about protection of personal data in Europe is article no: 108 "Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data" under the date of 28 January 1981. With this convention, only automatically processed data are guaranteed. Fundamental principles of this convention are given below:

i. Acquirement of data with rightful and legal ways, using them objectively

ii. Texture like being appropriate and updated
iii. Keeping sensitive personal data even more secure
iv. Keeping data secure
v. Right of demanding for accessing, correcting or erasing [12].

2.3.4. European union

The most differential feature of European Union data protection model is its "compulsivity". There are units in each member state of the Union which protect [64] and lead the application to rules personal data protection law no. 95/46/EC " European Parliament and Council of Europe Directive on protection of individuals with regard to the processing of personal data and on the free movement of such data" dated 1985 is the most effective regulation about protection of personal data [72]. One of the most important features of this directive is it not only assures automatically processed data but also manually processed ones.

2.3.5. Asia-pacific economic cooperation (APEC)

"APEC Privacy Framework" which was accepted by the organization, shows its own approach regarding to protection of APEC countries' personal data [64].

2.3.6. Regulations in Turkey

Recent improvements in information technologies are investigated closely in our country as well. According to a research of Turkish Statistical Institute, computer usage rate of the population between the ages of 16 and 74 had been %49.9 in 2013, while the rate have risen to %53,5 in 2014 [73]. In respect to the same research, Internet usage rate in 2013 was %48,9 and in 2014 it increased to %53,8. In Turkey, one of two people uses computer and also the Internet access rate increases in direct proportion to computer usage rate according to these results. However the increase of these technologies and the usage of them brought new problems along. The need for protection of data which are personal information shared by users is at the top of these problems. International regulations about this topic are explained in the previous section.

Unfortunately, Turkey falls behind other countries, especially European Union countries, on the subject of regulations on the area of protection elements called personal data although the technological developments are followed closely. There

aren’t any special legal arrangements on the subject yet. European Commission Data Protection Act is signed by Turkey but confirmation is yet to be completed. Although there isn’t any legal regulation for personal data protection a new provision is added to 20th article of Republic Constitution by referendum in 2010 [74].

There are some regulations in Turkish Criminal Code No. 5237 under the date of 26.09.2004. With the related law of article 135(1), users who keep personal data unlawfully will be sentenced for a term of six months to three years. The provision in the same article states “Any person who illegally obtains, disseminates or gives to another person someone's personal data shall be sentenced to a penalty of imprisonment for a term of one to four years”. The article 137 of the same law also penalizes people who don’t dispose the acquired data for a length of time. As it is evident from the articles personal data needs to be gathered, used and disposed within boundaries of law. Personal data are also attempted to be secured by Electronic Communication Law. Besides many regulations are provided in areas of private law, administrative law, etc.

2.4. Data Sharing in Information Systems

Improving rapidly, information practice has turned out to be one of the indispensable elements of life. While applications like shopping sites meet our needs social networks like Facebook and Twitter helps us have a good time. As well as these beneficial services information practice might cause several disadvantages like the privacy of personal information. Gathering, processing and delivering users’ data can be realized in two stages as shown in Figure 2.2. The first stage takes place between the user and information service agency and at this level users’ personal data is gathered, processed and delivered by the agency [75]. The agency supplying information service cannot only be a shopping site but it can only be a site on art and literature where the users share comments on films and books. Social networking sites, which provide social and personal sharing and building social networks, can be also regarded as examples for agencies supplying information services. In addition to this, nowadays governments also present their citizens some on-line services so they can also be regarded as information service agencies. In Figure 2.2, the data sharing between user and agency is shown via vertical arrows and users usually enter the information that online process necessitates and send it to the agency’s database. In the next stage the information

service agencies are in the position of sharing users' data between each other [76]. Agency might need this sharing model, shown in Figure 2.2, while they process the data. In this section these data sharing stages will take place under different headings.

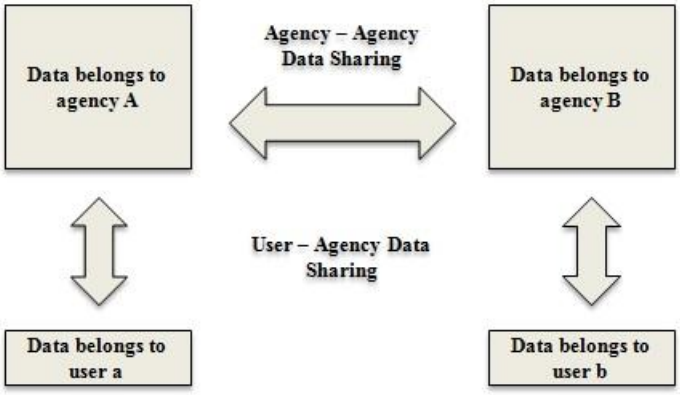

Figure 2.2. *Stages of Data Sharing*

2.4.1. User-agency data sharing

In the use of most information practice we first face the first step that includes receiving some sensitive data like user's obligatory demographical information or address. The data called ID number, via which one can acquire all kind of data about one person, precedes this sensitive data. In this step, where this kind of sensitive data is received, we face attacks in order to gain these sensitive data. The attack technique called social engineering, which is defined as the art and science of learning what they want from people, precedes these attacks. The aim here is to enter the system without permission and acquire the user's sensitive data.

After the user ends membership process s/he starts using the application. At this stage the user involuntarily or voluntarily lets the system gather his personal data by likes, comments or advices. In addition to this, the information about how much time the user spent in a site or when the user logged in the site is also kept in the system. By using these kind of data the user's profiling is easily done and user data is processed. While some information practice have in their membership contract the warning that these kind of data will be gathered and used, most don't have any warning about it. This situation causes the problem of illegally collecting and using the user data.

In the field of computer sciences different studies were done about data sensitivity by using cryptographic, random fault, anonymization techniques. Yakut and Polat [75]

have proposed a method to carry out some main filtering services like assumption and suggestion by protecting the privacy via stable time common filtering algorithm that uses random fault technique. Pinkas [37] realized safe distributed calculating and their application in data mining basing on cryptographic techniques. Sweeney [77] suggested the concept of k-anonymity and proposed the method that will maintain data privacy in a way to anonymize it in the data like this: one etiquette will meet at least *k* number of entries. The use of networking technologies for protecting private data, as well as academic studies, has become widespread. To exemplify these technologies The Onion Router (TOR), Virtual Private Network (VPN), proxy servers can be mentioned. Private data protection is carried out in TOR technology via several tunnels either serialized or imaginary; as for VPN it is carried out by connecting into a cipher network called imaginary private network. Thanks to the Proxy servers that are used to maintain anonymity, instead of connecting a network directly the user can hide his/her identity by connecting via inter-servers.

One of the risks for private data during online communication comes up during communication between people via social media. Actually, the user shares data with the agency technically but the structure of these sites bring along user-user data sharing, too. In this kind of communication users can easily share their private data. Especially the sensitive data shared in this kind of environment might cause serious problems to show up. Dangers like being subject to dishonesty and fraud precede the problems. In order to prevent and decrease these kind of problems users should be made conscious of whether to share their private data or not and how to share how much of it.

Besides, attacks, which aim to expose private data, might be organized by third batch people during the communication between the person and the agency. Foremost among these are the security gaps during sharing credit card information at online shopping. Users should especially be careful about the site they do the shopping having a valid SSL certificate. Through SSL certification sensitive data like credit card information are ciphered before they are sent and only the correct receiver can decipher it. In the SSL certificate, there are 40 bytes or 128 bytes ciphering methods. Occasionally, in online shopping applications 128 bytes ciphering method is used and with this method acquiring the cipher costs much time. Another attack aiming at gaining credit card information is fishing. Fishing attacks is one of the most popular attack methods online, and with this method e- mails that pretend to be from banks personal

credit card or bank account information can be acquired. In addition to this, by using key logger software private data can be acquired. By utilizing spywares, which make a record of keyboard key entries, the user's sensitive data can be recorded. As precautions against these kind of attacks pop-up keyboard and 3D security measures are taken.

2.4.2. Agency-agency data sharing

The firms, which develop information practice, might try the way of sharing costumer data with other firms for the aim of increasing quality of service and customer satisfaction; to gain moral and material income over the data. In Facebook users' contract it is obviously seen that the user data can be shared with other agencies [78]. Although sometimes this is done by informing users in the membership contract, sometimes it comes true beyond their knowledge. As mentioned above sharing private data with other agencies is only possible with one's consent. In addition it might sometimes be possible to share the data when the law requires. Sharing private data except for these conditions will be illegal.

Data mining over inter-institutional data sharing has been handled academically in many articles and several methods have been suggested [76,79]. For example in Memiş and Yakut's study about suggestion systems [76], in order to increase the quality of the suggestion to be offered to the customer sharing his data between two companies operating in the same field is resorted. The starting point of this study is companies that provide suggestion services don't have enough data of real use. In this study [76], authors developed privacy protection method against limited data problem. In their study [79] Vaidya and Clifton argued how to realize k-means clustering algorithm by using data owner organizations' data as entries in a shared way and they solved this problem by using protocols in which cryptographic techniques are used.

A significant event on inter-institutional data sharing has been experienced in Spain. In this event a Spanish Peugeot vendor's transferring customer data to another Peugeot vendor in Spain again has been evaluated by Data Protection Authority. The Data Protection Committee found the firm's act of informing its customers about transferring their data to another vendor firm in the same group with general expressions, insufficient [64]. The case above is important in two aspects. The first is; it is required that informing people about their data being shared must be put into direct words but shouldn't be in general expressions. Another situation is the necessity of

direct informing not only between firms in the same group but also between intragroup firms.

2.5. Chapter Summary

Especially in the twentieth century the development the concept of private data and the need for protecting this data has brought along some legal and technical approaches. As well as these approaches making people aware of protecting their private data is also important. In Table 2.1 these approaches, the methods they use and their profits are summarized. The needs that increase with the aim of protecting private data laid the groundwork for several legal regulations both internationally and in our country. The data, which was guaranteed through legal regulations, was also protected with different privacy protection solutions in the field of computer sciences. Besides, though not being very common in application field, the methods of data collection, procession and sharing present different solutions for data protection. Together with the changing and increasing needs for data protection the solutions will take their place in engineering and data processing.

In information practice another approach to protect private data is to create awareness in this issue. Both official positions and agencies providing information services carry on studies with face to face education and seminar works, advertising videos on mass media tools, public spots, and leaflets to increase public awareness. Moreover these agencies raise awareness campaigns through using social media effectively. For example Security General Directorate of our country informs the citizens by texting as a precaution against engineering and fishing threat. Another advantage of protecting private data guaranteed by technical and awareness raising approaches is that it will decrease the load of work of courts. These approaches on protecting private data will also lessen the unjust treatment being experienced or having already been experienced by individuals.

Table 2.1. *Approaches for Protecting Personal Data*

Scope	Methods	Profits
Legal Approaches	• International Regulations • Constitutional Regulations • Related Legal Regulations	Private data is protected through international regulations aiming at private data protection, and legal regulations like constitution laws, related laws and written regulations
Technical Approaches	• Cryptographic Algorithms • Randomization-based Techniques • Anonymization Techniques • Network Technologies	Improving computer sciences approaches provide the chance of collecting, processing and conveying the data.
Awareness Approaches	• Education and Seminars • Mass Media Tools • Social Media	It is necessary to make people aware of legal and technical approaches about private data protection. In this way awareness about which legal methods and techniques to use in a matter of any illegal situation before or after data sharing.

3. PRIVACY-PRESERVING USER-BASED COLLABORATIVE FILTERING ON OVERLAPPED RATINGS

In this section, how to perform privacy-preserving of user-based CF over arbitrarily partitioned data with overlapping ratings is examined. To achieve privacy-preservation through schemes, default votes and homomorphic cryptosystems (HCs) are exploited. The proposed schemes will be introduced in detail in the following subsections.

3.1. User-based Collaborative Filtering with Pearson Similarity

In user-based CF, similarities are calculated based on users similarity and neighbors are found from the most similar users. One main task of CF systems is to produce a prediction p_{aq} for an *active user* (*a*), about *the target item* (*q*) using $n \times m$ user-item rating matrix where *n* and *m* are the number of users and items, respectively. There are mainly three steps in a typical CF process: similarity computations, neighborhood determination, and prediction generation based on the similarity-weighted average of neighbor's ratings on *q*. According to Herlocker et al. [1], similarity between *a* and train user *u* can be computed using Pearson correlation coefficient:

$$w_{au} = \frac{\sum_{j \in C} (r_{aj} - \bar{r}_a) \times (r_{uj} - \bar{r}_u)}{s_a \times s_u} \tag{3.1}$$

where *C*, w_{au}, r_{uj}, $\bar{r}_u$ and $\sigma_u$represent commonly rated items, similarity between *a* and train user *u*, the given rating value by *u* on item *j*, user *u*'s mean and user *u*'s standard deviation, respectively [2]. After calculating similarity between *a* and each train user *u*, *a*'s neighborhood is determined from the best similar users. Then, the final prediction p_{aq} equals to the similarity weighted average of ratings given by the neighbors for *q*:

$$p_{aq} = \bar{r}_a + \frac{\sum_{u \in N} w_{au} \times (r_{uq} - \bar{r}_u)}{\sum_{u \in N} w_{au}} \tag{3.2}$$

where, *N* stands for *a*'s neighbors [1].

3.2. Arbitrarily Partitioning and Overlapped Ratings

Two parties, say *A* and *B*, want to provide CF services on partitioned data with overlapped ratings. They have similar sets of customer and item portfolios. According to Figure 3.1, with respect to rating belongings there are three subsets of ratings: R_A, R_B

and R_ϕ. While R_A and R_B hold ratings only belong to A and B, respectively, R_ϕ includes overlapped ratings given by the same user for the same item to the both parties. If R_ϕ is empty, there is no rating overlap and the partitioning case becomes arbitrarily partitioned data (APD) as examined in [48]. However, such overlaps make this study more challenging through prediction quality and privacy-preservation compared to APD. Figure 3.1 also demonstrates the scarcity of CF rating data which have many unrated items shown with empty cells. In this configuration, for the sake of simplicity, overlapped ratings are assumed to be consistent, thus, users have already given the same rating value for the same item in both parties' data.

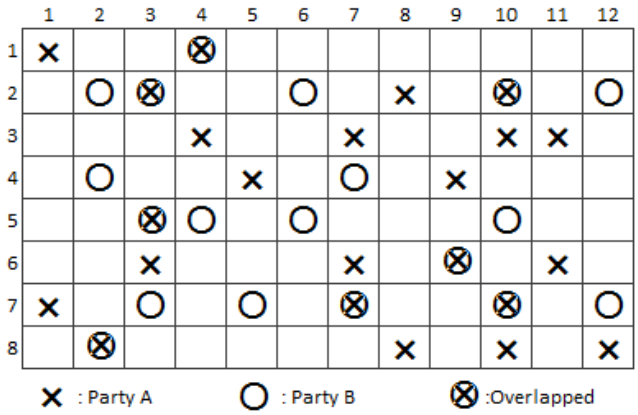

Figure 3.1. *Arbitrarily partitioned data with sample overlapped ratings*

3.3. Privacy Problem

In the context of PPCF [3], *the private* denotes each rating values and also denotes which items are rated by which user. To achieve privacy-preservation, there should be no direct exchange of each individual rating values and rated items without sharing any intermediate and aggregate values that may reveal individual private information. This is necessary as parties are *semi-honest* and greedy about gathering as much private data as possible, while obeying the predefined procedure. Note that there is no problem for parties to learn which ratings are overlapped, and the information about which ratings are overlapped can be considered public information. Since the value of overlapped ratings for the same user-item pair is equal, any party's awareness of whether such overlapped item is rated to be a nonissue regarding privacy is considered. After introducing all the related preliminaries, the concentrated *problem* can be described to be in the junction of two viewpoints:

i. From the prediction quality viewpoint, proposed schemes should promote user-based CF services of the two parties suffering from data scarcity.
ii. From the privacy viewpoint, privacy is preserved when proposed protocols executed by semi-honest parties ending up with arbitrarily partitioned data with overlaps.

In order to solve this problem, the proposed solutions should cater to both the aforementioned viewpoints. Since efficiency is the conflicting goal with respect to prediction quality and privacy-preservation, the solution should promise agreeable computational performance as well.

3.4. Privacy-Preserving User-based CF on Overlapped Ratings

In order to solve this problem, the proposed solutions should cater to both the aforementioned viewpoints. Since efficiency is the conflicting goal with respect to prediction quality and privacy-preservation, the solution should promise agreeable computational performance as well.

3.4.1. Preprocessing

Regarding Equation 3.1, it can be said that each party needs to normalize its own data. To perform such normalization, each party needs user means. In order to determine the denominator in the same equation they need the standard deviation of each user. Mean and standard deviation are statistically algebraic measures which are composed of distributed measures. Distributed measures can be easily calculated in distributed manner. For example, arithmetic mean equals *sum* of numbers in an array divided by the *count* of this array. If the array is partitioned among the two parties then by exchanging partial sum and partial size each party can obtain mean of the elements in the array. However in this thesis, direct exchange of such statistical measures may cause some privacy breaches especially if there are a small amount of available ratings from a user.

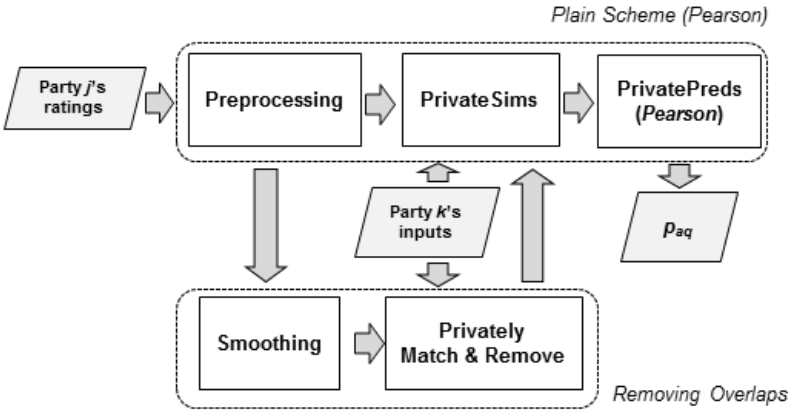

Figure 3.2. *Privacy-preserving user-based CF on overlapped ratings*

To ensure privacy, randomized default vote filling procedure where default votes can be row mean, column mean, or overall mean from available ratings of a party P is offered. After parties agree on *level of filling* (θ) in percentage of density, party P can enhance its own data with v_ds as given below:

1. Randomly or selectively determine β_P from the range [0, θ].
2. Randomly select $\beta_P \cdot \delta_P$% of unrated cells where δ_P is the number of available ratings.
3. Fill such selected cells with v_ds.

After filling its own data, parties can exchange partial sum and count values and estimate user mean. Then, they normalize their data using deviation from user mean approach and estimate user standard deviation similar to mean estimation. After preprocessing, each party ends up with estimates of user mean and standard deviation.

3.4.2. Similarity computation

To compute similarities, two complete user profiles are needed. However, such profiles are arbitrarily distributed among two parties. Hence, there are two parties and two users then the similarity between users a and u can be considered as follows:

$$w_{au} = XY = X_A Y_A + X_A Y_B + X_B Y_A + X_B Y_B \tag{3.3}$$

where X and Y represent the normalized rating profiles of a and u, respectively; X_P and Y_P stand for available part of such profiles in party P. Overlaps affect the accuracy of the recommender, however, it can be hypothesized that explainable results can be obtained despite of overlapped ratings. In plain approach, private similarity computation protocol

(*PrivateSims*) is given, which does not consider overlaps. Moreover, how to tackle with overlaps with preserving privacy is provided in the following subsections.

PrivateSims: Private similarity computation protocol

For each user with *a* the following is performed:

1. Each party assigns zero to all unrated cells.
2. Each party *P* computes $X_P Y_P$.
3. For train user *u* being 1 to *n*/2
   3.1. *A* encrypts each element *i* of X_A and Y_A with its public key *KA.*
   3.2. *A* sends all $\xi_{KA}(X_{Ai})$ and $\xi_{KA}(Y_{Ai})$ to B.
   3.3. *B* computes all $\xi_{KA}(X_{Ai})^{Y_{Bi}}$ then finds $\xi_{KA}(X_A Y_B)$.
   3.4. *B* computes all $\xi_{KA}(Y_{Ai})^{X_{Bi}}$ then finds $\xi_{KA}(X_B Y_A)$.
   3.5. *B* encrypts $X_B Y_B$ with *KA.*
   3.6. Using Paillier's addition, *B* finds $\xi_{KA}(X_A Y_B + X_B Y_A + X_B Y_B)$.
   3.7. *B* sends resultant ciphertext to *A.*
   3.8. *A* decrypts it, adds $X_A Y_A$ to it and divide proper $\sigma_a \cdot \sigma_u$ and obtains w_{au}.
4. For the remaining train users
   4.1. By switching roles, repeat steps 3.1–3.8.
5. Finally, each party has *n*/2 pieces of *n* similarities.

PrivateSims protocol's privacy mechanism is based on Paillier HC. In the initial step, unrated cells are set to zero since it is intended to utilize absorbing element property of zero during multiplication. In step 2, each party performs partial similarity calculation over only available ratings. With steps 3–4, each party privately computes components of w_{au} and end up with half of the total similarity values between *a* and each train user *u*. Note that self-blinding property of Pailler HC is exploited for all encryptions in this scheme in order to discriminate similar plaintexts from each other.

3.4.3. Prediction computation

Now, it is necessary to compute Equation 3.2. Considering that similarities and ratings are distributed among the parties, Equation 3.2 can be rearranged as follows:

$$p_{aq} = \bar{r}_a + \frac{\sum_{u \in N} (w_{auA} \times \tilde{r}_{uqA} + w_{auA} \times \tilde{r}_{uqB} + w_{auB} \times \tilde{r}_{uqA} + w_{auB} \times \tilde{r}_{uqB})}{\sum_{u \in N} (w_{auA} + w_{auB})} \tag{3.4}$$

where w_{auP} and $\tilde{r}_{uqP}$ stands for similarity values and normalized rating of *u* on *q* held by party *P*, respectively. Private prediction computation protocol (PrivatePreds) for distributed Pearson similarities and ratings is proposed. First of all, such protocol is demonstrated for the case where *A* is master party (*MP*) queried for p_{aq}. If *MP* is *B* then they must switch the roles and move further. *A*'s neighbors based on threshold (τ) is determined and select neighbors comprised of users having similarities greater than τ in step 1. In step 3, each party generates binary clone rating vector whose entries having value of one if *q* is rated by *u* otherwise it is zero. Since one is an identity element for multiplication, The binary clones to add up proper similarity values in the denominator is used. In steps 4–8, *B* computes for the numerator while in step 9 computations are performed for the denominator. In step 11, $(w_{auA})_P$ stands for similarity values available in *A* exploited in numerator calculation by party *P*. At the end of PrivatePreds, *MP* returns prediction p_{aq} to *a*.

PrivatePreds: Privately prediction computation protocol for Pearson similarity

1. Each party assigns zero to all its similarity values less than τ.
2. Each party assigns zero to all unrated cells for *q*.
3. Each party *P* generates binary clone rating vector (c_{uqP}).
4. *A* encrypts each element *i* of w_{auA}, $\tilde{r}_{uqA}$, and c_{uqA} with *KA*.
5. *A* sends all $\xi_{KA}(w_{auA})$ and $\xi_{KA}(\tilde{r}_{uqA})$ values to *B*.
6. *B* computes $\xi_{KA}(w_{auA})^{\tilde{r}_{uqB}}$ then obtains $\xi_{KA}(w_{auA}\tilde{r}_{uqB})$.
7. *B* computes $\xi_{KA}(\tilde{r}_{uqA})^{w_{auB}}$ then obtains $\xi_{KA}(w_{auB}\tilde{r}_{uqA})$.
8. *B* computes $w_{auB}\tilde{r}_{uqB}$ and encrypts it with *KA*.
9. *B* repeats steps 6–8 replacing $\tilde{r}_{uqP}$ with proper c_{uqP}.
10. *B* adds up and finds $\xi_{KA}(w_{auA}\tilde{r}_{uqB} + w_{auB}\tilde{r}_{uqA} + w_{auB}\tilde{r}_{uqB})$ and $\xi_{KA}((w_{auA})_B + w_{auB})$ sends to *A*.
11. *A* decrypts them and adds $w_{auA}\tilde{r}_{uqA}$ to the former and $(w_{auA})_A$ to the latter.
12. *A* divides numerator by the denominator, adds *a*'s mean, finds prediction p_{aq}.

3.4.4. Removing overlaps

As seen from Figure 3.2, in order to remove overlaps, there are two processes: eliminating initially filled votes (*smoothing*) and privately determining and removing

overlaps (*privately match & remove)*. In the first step, each party deletes v_ds after preprocessing. Note that such v_ds are avoided to cause additional overlaps. In the second step, the problem is how to privately determine which ratings are overlapped. Such a problem can be deliberated as two parties having two sets and want to find commonly existing items. In privacy-preserving data mining, such problems are paid so much attention and some privacy-preserving set intersection protocols are proposed for parties having confidential data. In this context, Freedman et al. [80] presented some efficient schemes and in order to find overlaps, applying one of them is prefered, namely *private matching for semi-honest parties(PM-Semi-Honest)*. PM-Semi-Honest scheme is a two-party protocol between *chooser* and *sender* both having different size of sets having numbers from the same domain. At the end of the protocol, chooser learns which of inputs are shared by both of them.

Privately matching and removing overlaps protocol (*Privately Match & Remove*) is proposed in order to tackle with overlaps. Initially, each party *P* finds indices of rated cells and computes cutting index point (λ_{ci}) where $\lambda_{ci} = (nm)/2$. Finally, each party *P* ends up with knowledge of approximately half of the total overlaps and deletes ratings held by *P* having indices corresponding such overlaps. After removing overlaps, parties move on to the next process *PrivateSims*. This solution is named as *ultimate scheme* (*US*). If the parties do not need or prefer to remove overlaps, *plain scheme* (*PS*), which does not involve overlap removing process, can be applied.

Privately Match & Remove: Privately matching and removing overlaps protocol

1. Each party *P* finds indices of rated cells and computes λ_{ci}
2. For rating index from the first to λ_{ci}
   2.1. Set *A* as chooser and B *as* sender
   2.2. Apply *PM-Semi-Honest*
   2.3. *A* learns about half of the overlaps and removes corresponding rating values
3. For rating index from λ_{ci} to the end
   3.1. Switch parties' roles in steps 2.1–2.3, *B* removes remaining of the overlaps

3.5. Analysis of the Scheme

First of all, proposed scheme meets the privacy requirements mentioned in Subsection 3.3. Via randomized filling with default votes and homomorphic encryption, confidentiality of rated items and rating values are ensured. In the preprocessing step,

v_ds is exploited to avoid share of actual sum, count, and sum of squares. Such v_ds improve privacy-preservation especially when there are a few number ratings for in a row (user). For instance to compute user mean values, each party P share disguised numbers such as $count + 0.01\beta_P\delta_P$ rather than sharing actual *count* values. From the side of other party Q, before trying to infer which items are rated, he should guess *count* first. The probability of correctly guessing β_P is $1 / \theta$ if β_P is considered integer. If β_P is considered rational number, this probability reduces with increasing precision of selection interval of $[0,\theta]$. However, at the same time, Q still has no certain information about density (δ_P) of P. One way to estimate *count* values approximately, Q can analyze shared count values for the same users over number of trials where β_P is expected to be $\theta/2$. To avoid such kind of inferences, parties should scramble labels of users in a particular frequency of sharing. Note also that inference of individual rating values using disguised *sum* values is much more difficult than correctly guessing of which items are rated.

Default votes enhance privacy-preservation along the remaining procedures of plain schemes of user-based scheme as well. How about the proper values of default votes? v_ds can be row mean or column mean of held data. In particular, for this user-based CF scheme, column mean can be considered as more privacy enhancing solution since sum and count values of each row are shared among parties. In addition to randomization provided by v_ds, cryptographic mechanisms is exploited as well in order to accomplish privacy-preservation. Paillier [59] proved that his homomorphic cryptosystem achieves semantic security for any probabilistic polynomial time adversary. Privacy-preservation of these protocols *PrivateSims* and *PrivatePreds* is directly based on such evidence. The privacy of *Privately Match & Remove* is fulfilled by Freedman et al's *PM-Semi-Honest* [81]. Their private matching protocol can be implemented based on Paillier's scheme or its subsequent versions hence privacy-preservation is based on the same proof. Also, self-blinding property of Paillier's homomorphic cryptosystem makes much more sense for a typical user-item data. There are numerous unrated cells and there are many cells expected to have the same value from a particular integer range, and such property effectively camouflages unrated and same-rated cells.

Since privacy and efficiency are two clashing goals, privacy-preservation mechanisms require additional communicational, computational and storage

requirements. Using *PrivateSims*, for each similarity values, parties need to exchange $O(n)$ vectors with each other in two different communications. They can exchange such values of all similarities over just two communications: one from P to Q and one from vice versa. Similarly, for the case of PrivatePreds, $O(m)$ vectors are exchanged between two parties and they can also be performed over two communications**.** A distributed model is proposed in which similarity and deviation values are distributed between two parties. To compute each p_{aq}, parties need each other and one communication is needed from each party to other. To avoid prediction computation on distributed model, such similarity and/or deviation values can be entirely on each party depending on application.

Computational overheads are dominated by homomorphic operations. For *PrivateSims*, to compute each similarity value, there are totally $3m$ encryptions, $2m+1$ homomorphic multiplications and 1 decryption performed collaboratively by two party. For *PrivatePreds* based on user similarities, to compute each prediction value, MP needs to perform $5n/2$ encryptions and 1 decryption while the other party performs $2(n+1)$ homomorphic multiplications. To compute prediction based on item deviations by *PrivatePreds*, assuming that each party holds $m/2$ of deviations related to item q, each party is expected to perform m encryptions, m/2 homomorphic multiplications, m homomorphic additions and 2 decryptions. Considering large dataset where n and m are greater values, cryptographic operations may be bulky in computation, however recent research on implementation of efficient homomorphic encryption [82] shows that homomorphic encryption takes 24 ms, decryption takes at least 15 ms, addition is instantenous as taking as 1 ms whereas multiplication takes 41 ms on ordinary computer with 2.1 GHz Intel Core 2 duo processor with 1 GB of memory. With utilization of more powerful hardware infrastructures and parallel computation techniques, more satisfactory performance can be obtained. Also, to increase efficiency, some improvements such as pre-computation of normalization, similarity values and predictions before a's request may be possible. However, the parties must be ready for additional storage overheads in this case. For example, there will be requirement of $n^2/2$ of floating point number space

3.6. Experimental Results

In the experimental analysis of the proposed schemes, MLP datasets having ratings from 943 users for 1682 movies is used. It is collected by GroupLens research community and publicly available at their web site *www.grouplens.org*. There are in all 100.000 integer ratings from the domain of [1,5]. In these experiments, available ratings are divided into train and test subsets having 90% and 10% of available ratings randomly assigned to corresponding subsets, respectively. Ratings in the train subsets are utilized to achieve CF algorithm and generate prediction while actual rating values in test subset are compared with predicted values to observe prediction quality in terms of accuracy. To evaluate accuracy, mean absolute error (MAE) is used, which is popularly exploited in CF researches [1,22]. MAE equals average of absolute differences between predicted values and actual test ratings. To reach dependable results, 100 trials for each experiment is performed and in each trial, train and test ratings are randomly determined. Each displayed MAE value is the average of MAEs obtained from all trials for each experiment.

First of all, how ratio of overlaps changes with varying density of rating data and the level of filling is observed. Trials by increasing δ from 10 to 100 and θ from 0 to 100 are performed and demonstrate the percentages of overlaps in Table 3.1. Such percentage values reflect number of overlaps over the cardinality of union of ratings between both parties. When the data type is whole, the all available 100.000 ratings are taken into account and then the ratings are randomly selected. Else, such ratings are determined from train data consisting of 90.000 ratings. Note that when θ is 0 there is no filling, and when θ is 100 there may be default votes as much as actual ratings. According to Table 3.1, with increasing density overlapping ratio increases for all of the rows. However, such ratio is inversely proportional to θ since rating values can only be from fixed 90.000 cells while v_ds can be assigned to remaining cells, i.e., 1.496.126 cells.

Table 3.1. *Ratio of Overlaps (%) vs. Density and Filling Level*

Data Type	Filling (θ)	Density (δ)					
		10	**20**	**40**	**60**	**80**	**100**
Whole	0	5.28	11.11	24.99	42.86	66.67	100.00
Train	0	4.70	9.93	21.95	36.97	56.26	81.82
Train	10	4.47	9.40	20.67	34.72	52.24	75.29
Train	20	4.29	8.93	19.71	32.29	48.32	71.34
Train	40	3.94	8.11	18.08	29.69	43.22	61.87
Train	60	3.64	7.57	16.65	26.83	40.37	55.70
Train	80	3.49	7.13	15.50	24.88	36.22	49.89
Train	100	3.18	6.84	14.65	22.86	34.64	48.59

In the second experiment, how accuracy changes with different levels of filling is examined. For this reason, θ is varied from 10 to 100 and MAE values are computed for PS and US for such θ values. Regarding the analysis in subsection 3.5, column mean as v_d is selected for user-based CF scheme. δ is set as 60 then each party holds 60% of ratings randomly selected from train subset and 36.97% of them are expected to be overlapped according to Table 3.1. For user-based CF algorithm, MAEs of PS and US with respect to varying θ are given in Figure 3.3. As seen from Figure 3.3, two schemes show different accuracy characteristics against increasing θ. While accuracy of PS worsen with the large level of filling, that of US gets better insignificantly, and US has the lowest MAEs for all θ values. The figure shows that θ does not affect accuracy of US as much as PS since US eliminates v_ds by the smoothing process. For each scheme, the best MAEs are 0.7513 and 0.7442 achieved at PS (θ = 20) and US (θ = 60), respectively.

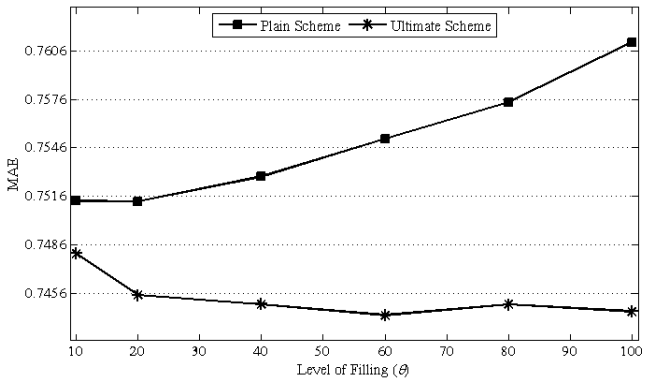

Figure 3.3. *Accuracy with respect to varying level of filling*

In the context of this study, any party can produce prediction using three different methods: singly without collaboration and schemes PS and US. In this set of experiments, these three methods were considered and MAEs are computed for varying densities from 10 to 80. From Figure 3.3, θ is set to optimum values 20 and 60 for PS and US, for user-based schemes, respectively. In Table 3.1 the outcomes are displayed related to user-based schemes and corresponding gains obtained by PPCF schemes with respect to single evaluation of CF in percentages where $Gain(X) = 100 \times (MAE_{Single} - MAE_X)/MAE_{Single}$ and MAE_X stands for the obtained MAE value from method X. According to Table 3.2, observed gains due to PPCF schemes get higher with lower densities. Hence, proposed user-based schemes work well for the parties having fewer amounts of ratings. This complies with motivation which promotes the prediction quality of the parties that suffer from data scarcity. Statistical significance of the results is also checked. For example, *t*-values of the results from PS and US are 47.60 and 31.14, respectively, for δ = 20. For both *t*-values, the two-tailed P value is less than 0.0001, and by conventional criteria the differences between single party and each of the user-based PPCF schemes are considered to be extremely statistically significant. The other *t*-values provide the same confidence level for promised accuracies by schemes, except PS (δ = 60) and US (δ = 80). For PS (δ = 60), *t*-value is less than 1 and it can be said that it is not statistically significant. For US (δ = 80), *t*-value equals 2.96 and this means that the two-tailed P value is 0.0035 and by the way the difference caused by US can be said to be statistically very significant according to conventional criteria.

Table 3.2. *Overall performance with varying density*

Method	**δ = 10**	**20**	**40**	**60**	**80**
Single Party	0.9265	0.8381	0.7729	0.7517	0.7416
Plain S.	0.8627	0.7936	0.7624	0.7513	0.7457
Ultimate S.	0.8798	0.8003	0.7562	0.7443	0.7391
Gain (PS)	6.88	5.31	1.35	0.06	-0.54
Gain (US)	5.04	4.51	2.16	0.99	0.34

3.7. Chapter Summary

In this chapter, two-fold solution framework privacy-preserving user-based CF on overlapped ratings is presented. The name of the first solution is the plain scheme which investigates the problem without eliminating overlaps. The name of the other solution is

ultimate scheme which determines overlaps privately and then eliminates them. The proposed method makes it possible to produce predictions on partitioned data between two parties. The experimental analyses show that proposed method produces satisfactory predictions while protecting privacy.

4. PRIVACY-PRESERVING ITEM-BASED COLLABORATIVE FILTERING ON OVERLAPPED RATINGS

In this section, how to perform privacy-preserving of item-based CF over arbitrarily partitioned data with overlapping ratings is examined. To achieve privacy-preservation through schemes, default votes and homomorphic cryptosystems (HCs) are exploited. The proposed schemes will be introduced in detail in the following subsections.

4.1. Item-based Collaborative Filtering with Slope-one Predictor

In item-based CF, the similarities between different items are calculated by using items which have been rated by all the users. Slope-one predictor algorithms [32] evaluate how much an item is likely to be compared to another one using predictors of the form $f(x) = x + b$. One way to measure this differential is by simply subtracting the average rating of the two items. Deviation dev_{jk} between items j and k can be computed by the following:

$$dev_{jk} = \frac{\sum_i (r_{ij} - r_{ik})}{card_{jk}} \tag{4.1}$$

where $card_{jk}$ is the cardinality of the set of users i who have rated both items j and k. In order to take the number of ratings observed into consideration, a weighted Slope-one prediction formula is introduced in [32]. Hence, prediction p_{aq} can be computed through the following:

$$p_{aq} = \frac{\sum_j (dev_{qj} + r_{aj}) \times card_{qj}}{\sum_j card_{qj}} \tag{4.2}$$

where j is each of the available items except q.

Similar to subsection 3.2 and 3.3, arbitrarily partitioning and overlapped ratings are used. Partitioned data with overlapped ratings holds ratings only belong to party A and B and overlapped ratings which are given by the same user for the same item to the both parties. Because of these overlapped ratings this study is more challenging through prediction quality and privacy-preservation than APD. Besides, to achieve privacy-preservation, default votes and homomorphic cryptosystems (HCs) are exploited.

4.2. Privacy-Preserving Item-based CF on Overlapped Ratings

Similar to subsection 3.4, two different schemes, as with the case of plain and ultimate ones in terms of privacy-preserving items-based CF are proposed. Such solutions are schematized as in Figure 4.1 there are some common blocks with user-based solution which are "Preprocessing" and "Privately Match & Remove". Such common blocks are the same as those presented in subsections 3.4.1 and 3.4.4. However, the remaining ones are going to be mentioned in the following texts. In contrast to user-based scheme, the ultimate scheme does not include preprocessing step in the item-based CF since preprocessing makes no sense for non-overlapping case of Slope-one algorithm.

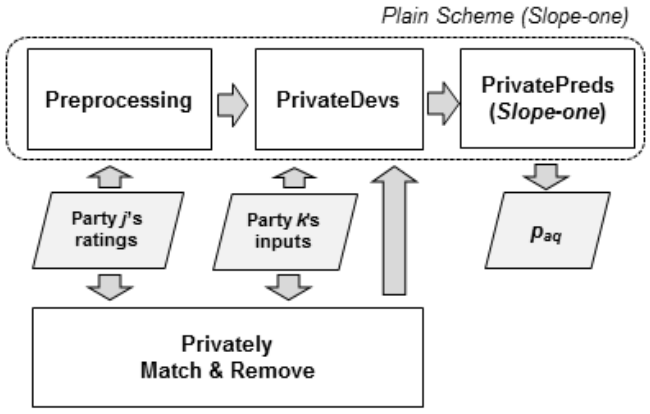

Figure 4.1. *Privacy-preserving item-based CF on overlapped ratings*

4.2.1. Deviation computation

Deviation computation given in Equation 4.1 can be considered as in Equation 3.3 and, similarly, it can be rewritten as:

$$dev_{jk} = \frac{\sum_i (r_{ij} - r_{ik})}{card_{jk}} = \frac{\sum\sum_{i \in Z_{PQ}} (X_{Pi} - Y_{Qi})}{\sum_{\forall (P,Q)} |Z_{PQ}|} = \frac{\sum_{\forall (P,Q)} (R_{PQ\,X} - R_{PQ\,Y})}{\sum_{\forall (P,Q)} |Z_{PQ}|} = \frac{\sum_{\forall (P,Q)} (R_{PQ})}{\sum_{\forall (P,Q)} |Z_{PQ}|} \quad (4.3)$$

where X_P and Y_Q are column vectors consisting of r_{ij} and r_{ik} values held by party P and Q, respectively; Z_{PQ} stands for commonly rated users through vectors X_P and Y_Q. Hence, there are 4 different sub-components as a combination of $P = A$, $Q = A$, $P = A$, $Q = B$, etc. If $P = Q$, then numerator and dominator parts can be locally computed by each party. However, similar to user-based scheme, the computation for cross sub-components is still challenging. Such challenge can be solved via private deviation

computation protocol (PrivateDevs) as given below. In the PrivateDevs protocol, each part ends up with half of all $dev_{j,k}$ and $card_{j,k}$ values.

PrivateDevs: Private deviation computation protocol

1. Each party *P* assigns zero to unrated cells in X_P and Y_P.
2. For half of deviation values
3. For each item pairs (*j*, *k*)

   3.1. Each party *P* computes $\sum(X_P - Y_P)$ and $|Z_{PP}|$

   3.2. Each party *P* generates binary clone rating column vector (c_{jP}) and (c_{kP}).

   3.3. Party *A* encrypts all X_A,-Y_A, c_{jA} and c_{kA} with its public key *KA*.

   3.4. *A* sends $\xi_{KA}(X_A)$, $\xi_{KA}(Y_A)$, $\xi_{KA}(c_{jA})$, and $\xi_{KA}(c_{kA})$ to *B*.

   3.5. *B* computes $\xi_{KA}(R_{AB_X}) = \prod_i \xi_{KA}(X_{Ai})^{c_{ikB}}$, $\xi_{KA}(R_{AB_Y}) = \prod_i \xi_{KA}(c_{ijA})^{Y_{Bi}}$,

   $\xi_{KA}(R_{BA_X}) = \prod_i \xi_{KA}(c_{ikA})^{X_{Bi}}$, and $\xi_{KA}(R_{BA_Y}) = \prod_i \xi_{KA}(Y_{Ai})^{c_{ijB}}$.

   3.6. *B* computes $\xi_{KA}(|Z_{AB}|) = \xi_{KA}(c_{jA})^{c_{kB}}$ and $\xi_{KA}(|Z_{BA}|) = \xi_{KA}(c_{jB})^{c_{kA}}$.

   3.7. *B* computes $\xi_{KA}(R_{AB})\xi_{KA}(R_{BA})\xi_{KA}(R_{BB})$, and $\xi_{KA}(|Z_{AB}|)\xi_{KA}(|Z_{BA}|)\xi_{KA}(|Z_{BB}|)$ sends these encrypted sub-aggregates to *A*.

   3.8. *A* decrypts such encrypted sub-aggregates and adds $\sum(X_A - Y_A)$ and $|Z_{AA}|$ values to the corresponding sub-aggregates and obtains dev_{jk} and $card_{jk}$.
4. For the remaining deviation values

   4.1. By switching their roles, *B* obtains such deviations and corresponding cardinalities.

4.2.2. Prediction computation

Prediction computation is triggered with the prediction query "p_{aq}" of active user from *MP* whose rating profile is distributed among the parties. Deviations and cardinalities are also distributed among the parties. The necessity is privately computed Equation 4.2 from the distributed elements. After rearranging Equation 4.2, the new equation is the following:

$$p_{aq} = \frac{\sum_j (dev_{qj} + r_{aj}) \times card_{qj}}{\sum_j card_{qj}} = \frac{\sum_j (num(dev_{qj}) + r_{aj} \times card_{qj})}{\sum_j card_{qj}} \qquad (4.4)$$

where $num(x)$ is the numerator of x. To solve Equation 4.3, parties use the protocol PrivatePreds (Slope-one) which privately computes prediction for the two parties. In this protocol, parties share encrypted version of held deviation values for item j, then the other party computes partial values of numerator and denominator of p_{aq} using homomorphic encryption properties. At the end of PrivatePreds (Slope-one), *MP* ends up with the final value of p_{aq} and inputs it a.

PrivatePreds : Privately prediction computation protocol for Slope-one predictor

1. A informs B about p_{aq}
2. Each party computes partial $num(p_{aq})$ and $den(p_{aq})$ for $dev_{jk,}$ and r_{aj} is held by the party.
3. Each party encrypts all held $num(dev_{jk})$ and $card_{jk}$ values related to item q and send it to the other party with its own public key.
4. A computes $\xi_{KB}(card_{qj_B})^{r_{aj_A}}\ \xi_{KB}(num(dev_{qj_B}))$ and $\xi_{KB}(\sum_j card_{qj_B})$ for the r_{aj_A} values and sends these values to B.
5. B decrypts these partial values.
6. B computes $\xi_{KA}(card_{qj_A})^{r_{aj_B}}\ \xi_{KA}(num(dev_{qj_A}))$ and $\xi_{KA}(\sum_j card_{qj_A})$ for r_{aj_B} and adds other available partial $num(p_{aq})$ and $den(p_{aq})$ values to this ciphertext and sends it to A.
7. A decrypts them and adds available corresponding partial data and obtains $num(p_{aq})$ and $den(p_{aq})$.
8. A divides $num(p_{aq})$ and $den(p_{aq})$. to find p_{aq} and returns it to a.

4.3. Analysis of the Schemes

Similar to user-based CF proposed scheme meets the privacy requirements. *Preprocessing* and *Privately Match & Remove* blocks are the same as user-based scheme. In the preprocessing step, randomized default vote filling procedure is offered. Default votes increase preservation of privacy for plain scheme. Default votes can be row mean, column mean, or overall mean and these mean values are calculated from available ratings of a party P. The proper values of default votes can be considered row and column mean values. Then, parties fill selected unrated cells with v_ds. After preprocessing, each party winds up with estimates of user mean and standard deviation. In the deviation computation and the prediction computation steps *PrivateDevs* and

PrivatePreds protocols are proposed respectively. To provide privacy homomorphic cryptosystem is used in these protocols.

In order to tackle with overlaps, privately matching and removing overlaps protocol is proposed. In this protocol, firstly, each party deletes v_ds. Then, to determine which ratings are overlapped PM-Semi-Honest is applied. PM-Semi-Honest is one of the efficient set intersection protocol which is proposed by Freedman et al. [81]. This private matching protocol is implemented based on Paillier's cryptosystem [59] and subsequent constructions. Furthermore, privacy-preservation mechanisms require additional requirements for communication, computation, and storage. Besides, homomorphic operations are needed for computational overheads. For For *PrivateDevs*, to compute each deviation value there are totally 6*n* encryptions, 6*n* homomorphic multiplications and 4*n* homomorphic additions and 2 decryptions performed in collaboration of parties**.**

4.4. Experimental Results

MLP datasets having ratings from 943 users for 1682 movies is used for each experiment. In these experiments, available ratings are divided into train and test subsets having 90% and 10% of available ratings randomly assigned to corresponding subsets, respectively. Ratings in the train subsets are utilized to achieve CF algorithm and generate prediction while actual rating values in test subset are compared with predicted values to observe prediction quality in terms of accuracy. To evaluate accuracy, mean absolute error (MAE) is used, which is popularly exploited in CF researches [1,22]. MAE equals average of absolute differences between predicted values and actual test ratings. To reach dependable results, 100 trials for each experiment is performed and in each trial, train and test ratings are randomly determined. Each displayed MAE value is the average of MAEs obtained from all trials for each experiment. The fist experiment is similar to subsection 3.6. In the second experiment, how accuracy changes with different levels of filling is examined. For this reason, θ is varied from 10 to 100 and compute MAE values for PS and US for such values. Similar to user-based scheme, some trials to evaluate change of accuracy with respect to varying level of filling for item-based CF is conducted. Since US does not include the preprocessing step, there no need to compare it with PS in terms of level of filling. Row and column means as v_d are used for item-based CF scheme and display corresponding

accuracy outcomes in Figure 4.2. According to Figure 4.2, row mean usage is slightly better than column mean for Slope-one CF, and both types provide worse accuracy with increasing amount of filling. Except for θ = 20, where the best MAE value is observed.

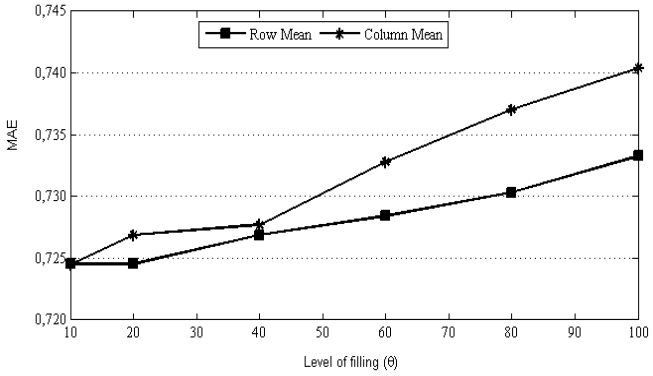

Figure 4.2. *Accuracy with respect to varying level of filling*

To evaluate overall performance of item-based solutions, some experiments are conducted, and displayed obtained MAEs in Table 4.1. For PS, data is filled using optimum settings of θ = 20 and v_d; row mean is set according to Figure 4.2. Comparing to Table 3.1, gain values for item-based schemes are much greater than user-based schemes. Hence, item-based schemes promise substantial contribution to accuracy especially for parties having sparse data. PS gives better accuracy than US especially for $\boldsymbol{\delta}$ values of 20 and 40, and it is said that default votes and rating overlaps can be expected to contribute to the accuracy of CF. Two-tailed t-values for PS as {87.25, 69.47, 32.56, 12.03, 4.71} and US as {87.21, 52.87, 17.25, 1.76, 0.13} for δ values of 10, 20, 40, 60, and 80, respectively can be listed. Such *t*-values show that the results are more statistically significant especially for lower values of δ. Another point is that the statistical significance parameters of PS are greater than that of US despite of randomization-based mechanism in PS.

Table 4.1. *Overall performance with varying density*

Method	**δ = 10**	**20**	**40**	**60**	**80**
Single Party	0.9936	0.8233	0.7613	0.7413	0.7378
Plain S.	0.7957	0.7416	0.7288	0.7292	0.7321
Ultimate S.	0.7955	0.7633	0.7455	0.7400	0.7394
Gain (PS)	19.91	9.93	4.27	1.63	0.78
Gain (US)	19.93	7.29	2.07	1.73	-0.02

4.5. Chapter Summary

In this chapter, similar to section 3, two different schemes are proposed. One of them is plain scheme and the other one is ultimate scheme. To compute prediction privately slope-one predictor is used for the two parties. Default votes and homomorphic encryption is used to ensure privacy protection. The empirical results show that proposed schemes give successful predictions while ensuring privacy.

5. CONCLUSIONS

In this thesis, how to preserve privacy while increase prediction quality using two-party CF on overlapped ratings is proposed. Although several studies [3,48,80] are proposed on arbitrarily partitioned data, in these studies overlapped ratings are not considered. Overlapped ratings make this thesis more challenging through prediction quality and preserving privacy than arbitrarily partitioned data. Besides, because of the fact users give the same rating value to the same item for both parties' data, overlapped ratings will be agreed consistent.

Two different collaborative filtering approaches and proposed novel schemes are investigated for conventional user-based collaborative filtering and slope-one which is an effective item-based collaborative filtering method. Such schemes come up with two alternative schemes such as the plain scheme and ultimate scheme. While the plain scheme gives the de facto solution involving some privacy-preserving collaborative filtering process blocks without considering rating overlaps, ultimate scheme consists of such blocks and an overlap removing process. The empirical results show that these schemes contribute to the prediction quality of the parties while ensuring their privacy. Plain schemes for user-based or item-based collaborative filtering are very effective for lower data density. At the same time, these schemes promise a more practical setup over existing some privacy-preserving collaborative filtering solutions.

Within the scope of this dissertation, two papers [83,84] are presented at international conferences and one SCI-Expanded journal article [76] is published.

As a future study, more complicated scenarios can be considered, as in this dissertation since the problem is simplified by equalizing the overlapping entries; however, in practice, much more complex overlapping cases could be faced. It is worth to examine such cases in the privacy-preserving manner. In this dissertation, just two parties are considered, but there are some e-commerce sites that collaborate with multiple parties. This is also another interesting topic to focus in further research.

REFERENCES

[1] Herlocker, J. L., Konstan, J. A., Borcher, A., and Riedl, J. T., "An algorithmic framework for performing collaborative filtering," *Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, USA, 230-237, 1999.

[2] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., and Shmatikov, V., "You might also like: Privacy risks of collaborative filtering," *Proceedings of 2011 IEEE Symposium on Security and Privacy*, Oakland, California, USA, 231-246, 2011.

[3] Yakut, I. and Polat, H., "Estimating NBC-based recommendations on arbitrarily partitioned data with privacy," *Knowledge-based Systems*, **36**, 353-362, 2012.

[4] Kim, J., Park, C., Hwang J., Kim, H., J., "Privacy Level Indicating Data Leakage Prevention System," *KSII Transaction on Internet and Information System*, **7(3)**, 558-575, 2013.

[5] Polat, H. and Du, W., "Privacy-preserving top-N recommendation on horizontally partitioned data," *Proceedings of International Conference on Web Intelligence*, Compiegne, France, 725-731, 2005.

[6] Finlay, S., *Predictive Analytics, Data Mining and Big Data Myths, Misconceptions and Methods*, Palgrave Macmillan, UK, 2014.

[7] Grad, B., Bergin, T. J., "Guest Editors' Introduction: History of Database Management Systems," *IEEE Annals of the History of Computing*, **31(4)**, 3-5, 2009.

[8] Wu, Q., McGinnity, M., Prasad, G., and Bell, D., "Knowledge Discovery in Databases with Diversity of Data Types," *Encyclopedia of Data Warehousing and Mining*, 1117-1123, 2009.

[9] Han, J., Fu, Y., Koperski, K., Melli, G., Wang, W., Zaiane, O. R., "Knowledge Mining in Databases: An Integration of Machine Learning Methodologies with Database Technologies," *Canadian Artificial Intelligence*, **38**, 4-8, 1996.

[10] Lee, W. and Stolfo, S. J., "Data Mining Approaches for Intrusion Detection," *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, USA, 1998.

[11] Jourdan, L., Dhaenens, C., and Talbi, E., "A Genetic Algorithm for Feature Selection in Data-Mining for Genetics," *Proceedings of the 4th Metaheuristics International Conference (MIC '2001)*, Porto, Portugal, 29-34, 2001.

[12] Koschmider, A., Hornung, T., and Oberweis, A., "Recommendation-based editor for business process modeling," *Data & Knowledge Engineering*, **70(6)**, 483-503, 2011.

[13] Park, J.H., "A recommender system for device sharing based on context-aware and personalization," *KSII Transaction on Internet and Information Systems*, **4(2)**, 174-190, 2010.

[14] Schafer, J. B., Konstan, J., and Riedl, J., "Recommender Systems in e-commerce," *Proceedings of the 1st ACM Conference on Electronic Commerce (EC '99)*, Denver, CO, USA, 158-166, 1999.

[15] Pazzani, M. J., Billsus, D, "Content-Based Recommendation Systems," *Lecture Notes in Computer Science*, **4321**, 325-341, 2007.

[16] Last, M., Shapira, B., Elovici, Y., Zaafrany, O., and Kandal, A., "Content-based methodology for anomaly detection on the web," *Proceedings of the 1st International Atlantic Web Intelligence Conference on Advances in Web Intelligence (AWIC '03)*, Madrid, Spain, 113-123, 2003.

[17] Bogdanov, D., Haro, M., Fuhrmann, F., Xambo, A., Gomez, E., Herrera, P, "Semantic audio content-based music recommendation and visualization based on user preference examplaes," *Information Processing and Management*, **49(1)**, 13-23, 2013.

[18] Schafer, B., Frankowski, D., Herlocker, J., and Sen, S., "Collaborative Filtering Recommender Systems," *Lecture Notes in Computer Science*, **4321**, 291-327, 2007.

[19] Breese, J. S., Heckerman, D., and Kadie, C., "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of 14th conference on Uncertainty in artificial intelligence*, Madison, Wisconsin, USA, 43-52, 1998.

[20] O'Connor, M. and Herlocker, J., "Clustering Items for Collaborative Filtering," *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, Berkeley, CA, USA, 1999.

[21] Pham, M. C., Cao, Y., Klamma, Y., Jarke, M., "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis," *Journal of Universal Computer Science*, **17(4)**, 583-604, 2011.

[22] Sarwar, B. M., Karypis, G., Konstan, J., and Reidl, J., "Item-based collaborative filtering recommendation algorithms," *Proceedings of the 10th international conference on World Wide Web*, Hong Kong, 285-295, 2001.

[23] Sarwar, B. M., Karypis, G., Konstan, J., and Riedl, J., "Analysis of Recommendation Algorithms for E-Commerce," *Proceeding of the 2nd ACM conference on Electronic commerce (EC '00)*, Minneapolis, 158-167, 2000.

[24] Shih, Y. and Liu, D., "Hybrid recommendation approaches: collaborative filtering via valuable content information", *Proceedings of the 38th Hawaii International Conference on System Sciences*, Hawaii, 217b, 2005.

[25] Ghazanfar, M. A., Prügel-Bennett, A., and Szedmák, S., "Kernel-Mapping Recommender system algorithms," *Information Sciences*, **208**, 81-104, 2012.

[26] Badaro, G., Hajj, H., El-Hajj, W., and Nachman, L., "A Hybrid Approach with Collaborative Filtering for Recommender systems," *Proceedings of the 9th International Wireless Communications and Mobile Computing Conference*, Sardinia, Italy, 349-354, 2013.

[27] Das, S. A., Datar, M., Garg, A., and Rajaram, S., "Google news personalization: scalable online collaborative filtering," *Proceedings of the 16th International Conference on World Wide Web*, Alberta, Canada, 271-280, 2007.

[28] Su, X., Khoshgoftaar, T.M., "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, **2009**, 2-21, 2009.

[29] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: an open architecture for collaborative filtering of netnews," *Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94)*, Chapel Hill, North Carolina, USA, 175-186, 1994.

[30] Hill, W., Stead, L., Rosenstein, M., and Furnas, G., "Recommending and evaluating choices in a virtual community of use," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, Denver, CO, USA, 194-201, 1995.

[31] Shardanand, U., Maes, P., "Social information filtering: algorithms for automating "word of mouth"," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, Denver, CO, USA, 210-217, 1995.

[32] Lemire, D. and Maclachlan, A., "Slope one predictor for online rating-based collaborative filtering," *Proceedings of 2005 SIAM International Conference on Data Mining*, Newport Beach, California, USA, 471-475, 2005.

[33] Papagelis, M. and Plexousakis, D., "Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents," *Engineering Applications of Artificial Intelligence*, **18(7)**, 781-789, 2005.

[34] Basu, A., Vaidya, J., and Kikuchi, H., "Efficient privacy-preserving collaborative filtering based on the weighted Slope One predictor," *Journal of Internet Services and Information Security*, **1(4)**, 26-46, 2011.

[35] Linden, G., Smith, B., and York, J., "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, **7(1)**, 76-80, 2003.

[36] Agrawal, R. and Srikant, R., "Privacy-preserving data mining," *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*, Dallas, Texas, USA, 439-450, 2000.

[37] Pinkas, B., "Cryptographic Techniques for Privacy-Preserving Data Mining," *ACM SIGKDD Explorations Newsletter*, **4(2)**, 12-19, 2002.

[38] Anbazhagan, K., Sugumar, R., Mahendran, M., and Natarajan, R., "An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining," *International Journal of Advanced Research in Computer and Communication Engineering*, **1(7)**, 482-485, 2012.

[39] Canny, J., "Collaborative filtering with privacy," *Proceedings of the 2002 IEEE Symposium on Security and Privacy*, Oakland, California, USA, 45-57, 2002.

[40] Kaleli, C. and Polat, H., "P2P collaborative filtering with privacy," *Turkish Journal of Electrical Engineering and Computer Sciences*, **18(1)**, 101-116, 2010.

[41] Wang, J., Pouwelse, J., Lagendijk, R. L., and Reinders, M. J. T., "Distributed collaborative filtering for peer-to-peer file sharing systems," *Proceedings of the 2006 ACM symposium on Applied computing*, Dijon, France, 1026-1030, 2006.

[42] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., and Shmatikov, V., "You might also like: Privacy risks of collaborative filtering," *Proceedings of 2011 IEEE Symposium on Security and Privacy*, Oakland, California, USA, 231-246, 2011.

[43] Xiong, J., Yao, Z., Ma, J., Liu, X., Li, Q., Ma, J., "PRIAM: Privacy preserving identity and access management scheme in cloud," *KSII Transaction on Internet and Information Systems*, **8(1)**, 282-304, 2014.

[44] Gunes, I., Bilge, A., Kaleli, C., and Polat, H., "Shilling attacks against privacy-preserving collaborative filtering," *Journal of Advanced Management Science*, **1(1)**, 54-60, 2013.

[45] Lathia, N., Hailes, S., and Capra, L., "Private distributed collaborative filtering using estimated concordance measures," *Proceedings of the 2007 ACM conference on Recommender Systems*, Minneapolis, MN, USA, 1-8, 2007.

[46] Zhang, S., Ford, J., and Makedon, F., "A privacy-preserving collaborative filtering scheme with two-way communication," *Proceedings of the 7th ACM conference on Electronic commerce,* Ann Arbor, MI, USA, 316-323, 2006.

[47] Berkovsky, S., Eytani, Y., Kuflik, T., and Ricci, F., "Privacy-enhanced collaborative filtering," *Proceedings of the PEP05, UM05 Workshop on Privacy-Enhanced Personalization*, Edinburgh, UK, 75-84, 2005.

[48] Yakut, I. and Polat, H., "Arbitrarily distributed data-based recommendations with privacy," *Data & Knowledge Engineering*, **72**, 239-256, 2012.

[49] Kaleli, C. and Polat H., "Providing naïve Bayesian classifier-based private recommendations on partitioned data," *Lecture Notes in Computer Science*, **4702**, 515-522, 2007.

[50] Yakut, I. and Polat, H., "Privacy-preserving SVD-based collaborative filtering on partitioned data," *International Journal of Information Technology and Decision Making*, **9(3)**, 473-502, 2010.

[51] Hsieh, C., L, Zhan, J., Zeng, D., and Wang, F., "Preserving privacy in joining recommender systems," *Proceedings of International Conference on Information Security and Assurance*, Busan, Korea, 561-566, 2008.

[52] Polat, H. and Du, W., "Privacy-preserving collaborative filtering on vertically partitioned data," *Lectures Notes in Computer Science*, **3721**, 651-658, 2005.

[53] Kaleli, C. and Polat, H., "SOM-based recommendations with privacy on multi-party vertically distributed data," *Journal of Operational Research*, **63**, 826-838, 2012.

[54] Kaleli, C. and Polat, H., "Privacy-preserving trust-based recommendations on vertically distributed data," *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, Laguna Hills, California, USA, 376-379, 2011.

[55] Zhao, Y., Feng, X., Li, J., and Liu, B., "Shared collaborative filtering," *Proceedings of the fifth ACM conference on Recommender system*, Chicago, IL, USA, 29-36, 2011.

[56] Bilge, A., Kaleli, C., Yakut, I., Gunes, I., and Polat, H., "A survey of privacy-preserving collaborative filtering schemes," *International Journal of Software Engineering and Knowledge Engineering*, **23(8)**, 1085-1108, 2013.

[57] Gong, S., "Privacy-preserving Collaborative Filtering based on Randomized Perturbation Techniques and Secure Multiparty Computation," *International Journal of Advancements in Computing Technology*, **3(4)**, 89-99, 2011.

[58] Polat, H. and Du, W., "Privacy-Preserving Collaborative Filtering," *International Journal of Electronic Commerce*, **9(4)**, 9-35, 2005.

[59] Paillier, P., "Public key cryptosystems based on composite degree residuosity classes," *Lecture Notes in Computer Science*, **1592**, 223-238, 1999.

[60] European Union, *Universal Declaration of Human Rights*, 1948.

[61] Council of Europe, *European Convention of Human Rights*, 1950.

[62] Akgül, A., *Danıştay ve Avrupa İnsan Hakları Mahkemesi Kararları Işığında Kişisel Verilerin Korunması*, Beta Yayınevi, İstanbul, 2014.

[63] Şimşek, O., *Anayasa Hukukunda Kişisel Verilerin Korunması*, Beta Yayınevi, Ankara, 2008.

[64] Küzeci, *E., Kişisel Verilerin Korunması*, Turhan Kitapevi, Ankara, 2010.

[65] Başalp, N., *Kişisel Verilerin Korunması ve Saklanması*, İstanbul Bilgi Üniversitesi Yayınları, İstanbul, 2004.

[66] Warren, S., D. and Brandeis, L., D., "The Right to Privacy," *Harvard Law Review*, **4(5)**, 193, 1890.

[67] Anayasa Mahkemesi 19.01.2012, E:2010/40, K:2012/8, 6.3.2013 T. ve 28579 sayılı R.G.

[68] Claes, E., Duff, A., and Gutwirth, S., *Privacy and the Criminal Law*, Intersentia, Antwerp, Belgium, 2006.

[69] Room, S., "Data Protection and Compliance in Context," *The British Computer Society Publishing and Information Product*, Swimdon, UK, 2007.

[70] OECD, *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 1980.

[71] Council of Europe, *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*, 1981.

[72] European Parliament and Council, *Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data*, 1995.

[73] TÜİK, *Individuals' Computer and Internet Usage Rates According to Last Usage Time*, 2014. http://www.tuik.gov.tr/PreTablo.do?alt_id=1028. (Date accessed: 20.02.2015)

[74] *The Constitution of Republic of Turkey*, 1982. https://global.tbmm.gov.tr/docs/constitution_en.pdf. (Date accessed: 20.02.2015)

[75] Yakut, İ. and Polat, H., "Privacy-preserving Eigentaste-based collaborative filtering," *Lecture Notes in Computer Science*, **4572**, 169-184, 2007.

[76] Memiş, B. and Yakut, İ., "Privacy-Preserving Two-Party Collaborative Filtering on Overlapped Ratings," *KSII Transactions on Internet and Information Systems*, **8(8)**, 2948-2966, 2014.

[77] Sweeney, L., "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **10(5)**, 557-570, 2002.

[78] *Statement of Rights and Responsibilities*, 2015. https://www.facebook.com/legal/terms. (Date accessed: 24.02.2015)

[79] Vaidya, J. and Clifton, C., "Privacy-Preserving k-means clustering over vertically partitioned data," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03)*, Washington, DC, USA, 206-215, 2003.

[80] Jagannathan, G. and Wright, R. N., "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data," *Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, Chicago, IL, USA, 593-599, 2005.

[81] Freedman, M., Nissim, K., and Pinkas, B., "Efficient private matching and set intersection," *Lecture Notes in Computer Science*, **3027**, 1-19, 2004.

[82] Naehrig, M., Lauter, K., and Vaikuntanathan, V., "Can homomorphic encryption be practical?," *Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop*, Chicago, IL, USA, 113-124, 2011.

[83] Memis, B. and Yakut, I., "Privacy-Preserving Collaborative Filtering on Overlapped Ratings," *IEEE 22nd Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2013)*, Hammamet, Tunisia, June 2013.

[84] Memis, B. and Yakut, I., "Bilişim Uygulamalarında Kişisel Verilerin Korunması," *The Third International Symposium on Digital Forensics and Security (ISDFS 2015)*, Ankara, Turkey, 2015.