On Applications of Privacy-Preserving Collaborative Filtering Schemes[1]

A. Bilge, C. Kaleli, I. Yakut, H. Polat[2]

Abstract— **Privacy-preserving collaborative filtering (PPCF) has been receiving increasing attention lately. Several PPCF schemes have been proposed for performing collaborative filtering services without deeply jeopardizing data owners' privacy. The proposed methods have been investigated in terms of privacy, performance, and accuracy. For each method, it has been shown that they are secure, able to provide accurate recommendations, and their online performance is comparable. However, they have not been investigated in terms of possible challenges and problems that might occur while utilizing them in real applications.**

In this study, we scrutinize different PPCF methods in terms of usability; and discuss the challenges that might be faced while utilizing them. We first classify the PPCF schemes according to various dimensions. We then list possible challenges and/or problems that might occur while applying them in recommender systems. We finally recommend some suggestions to eliminate them at all if possible or to lower their effects. Our study can give an idea about the applicability of the proposed PPCF approaches to those users and/or companies that plan to utilize them.

Index Terms — **Application, privacy, collaborative filtering, recommendation, usability.**

I. Introduction

With increasing popularity of e-commerce, number of online vendors utilizing recommender systems has been rapidly growing. Collaborative filtering (CF) techniques are widely used by recommender systems for estimating recommendations on customers' past experiences. The idea behind CF is that if two customers share similar interests in the past, they tend to share similar preferences in the future, as well [1]. Customers get predictions about the products they plan to buy, which help them choose the right products without wasting money and time. Online vendors are able to increase their sales and profits by providing recommendations to their customers because such services help them increase number of loyal customers and recruit new ones.

Performing CF services on plain data might cause some privacy, financial, and legal concerns [2, 3]. First, data collected for CF purposes are considered confidential data. In case of bankruptcy, such data are considered valuable asset. E-commerce sites can derive useful information about their customers from the ratings they provided. Online vendors can learn how much their customers like or dislike an item; or what products they bought before or showed interest. Moreover, since they are used to increase sales and profits, disclosing them might cause financial losses. Finally, legal regulations prevent data collectors from sharing or transferring such data. Revealing them might cause legal concerns. Due to such issues, it becomes important to provide recommendations without divulging private data. Hence, various schemes have been proposed to achieve CF services while preserving data owners' confidentiality. On one hand, privacy-preserving schemes should protect private data; on the other hand, they should still be able to offer predictions with decent accuracy. Moreover, they should be comparable with the traditional methods in terms of online efficiency. But, due to conflicting nature, it is a challenge to achieve three goals simultaneously.

In addition to privacy, accuracy, and performance, usability should be another dimension of which the proposed schemes should be analyzed. Although various PPCF schemes have been proposed; and they have been analyzed in terms of privacy, accuracy, and online efficiency, they have not been scrutinized in terms of usability. In other words, there exist no studies concerning the applications of such methods. Practical applications of PPCF schemes might face with various challenges. It is critical for the success of PPCF schemes to determine the possible problems and propose related solutions to eliminate them at all or reduce their effects. In addition to investigating such methods in terms of privacy, accuracy, and performance, their usability should be analyzed, as well.

In this study, we investigate PPCF schemes in terms of challenges that might be faced while using them in recommender systems. There are various PPCF methods having different properties besides the common ones. Since different schemes might face different challenges, we first classify PPCF schemes according to some dimensions that we determined. We then study various challenges of each class of the methods. Besides conflicting goals of privacy, accuracy, and performance, we scrutinize the PPCF approaches in terms of usability. The methods might achieve the expected goals theoretically. However, applications of them in recommender systems might face with some challenges. Thus, our major intend is to unveil such problems. After determining possible challenges, we recommend some possible solutions to eliminate them at all if it is possible or reduce their effects as much as possible. Our work guides those online vendors that plan to apply PPCF schemes in their recommender systems.

[1] This work was supported by TUBITAK under Grant 108E221.
[2] Corresponding Author: Department of Computer Engineering, Anadolu University, 26470 Eskisehir, Turkey (e-mail: polath@anadolu.edu.tr)

II. Related Works

Since privacy is an important challenge for CF schemes, researchers have great intention on this subject and propose several solutions for different privacy challenges in recommender systems. We can partition those studies into two groups, first one covers the studies related protecting individuals' privacy while the second group considers data holders' privacy.

Canny [4] proposes a homomorphic encryption-based solution to protect users' privacy in a fully P2P system. Polat and Du [5] discuss privacy protection for correlation-based and singular-value decomposition (SVD)-based CF algorithms. They employ randomized perturbation techniques (RPT) to protect users' confidentiality. To achieve private recommendations with Eigentaste-based CF, Yakut and Polat [6] introduce an RPT-based solution. In addition to RPT, data obfuscation and randomized response techniques (RRT) are utilized in CF to protect users' privacy. Parameswaran and Blough [7] propose a framework for obfuscating confidential information of individuals in CF. Kaleli and Polat [8] introduce a privacy-preserving scheme based on RRT for naïve Bayesian Classifier (NBC)- based CF algorithm. The same authors also propose a solution for protection individuals privacy on a P2P network while performing NBC-based CF algorithm [9]. Lathia et al. [10] introduce a new measure to compute correlation between users while protecting their privacy. This measure is employed in P2P systems-based CF applications. Shokri et al. [11] enable users to control their private profile and they disguise their profile by partly merging the profiles with other similar users' profile.

Besides protecting individuals' privacy, data holders need to consider their privacy if they collaborate with other parties. Since users do shopping from different online vendors, data collected for CF purposes might be distributed among parties. Thus, researchers investigate solutions providing collaboration of parties while producing recommendations. Kaleli and Polat [12] introduce private protocols for parties aiming to collaborate while producing NBC-based recommendations on partitioned (horizontally or vertically) data. In the same way, Yakut and Polat [13] propose solutions for providing SVD-based referrals on partitioned data without greatly jeopardizing data holders' privacy. In addition to collaboration of two parties, data might be distributed among multiple parties. Kaleli and Polat [14] study on private protocols to enable data holders providing SOM-based recommendations on vertically distributed data. Besides horizontally and vertically distributed data configuration, data might be partitioned arbitrarily between two parties. To make possible collaboration of parties when data arbitrarily partitioned, Yakut and Polat [15, 16] introduce different solutions.

III. Classification of PPCF Schemes

As explained previously, there are various PPCF approaches proposed in the literature. It is not appropriate to discuss the challenges that each method might face. Better alternative is to group such schemes into some classes according to some dimensions and attributes. Hence, in this section, we first determine the conceivable dimensions. Then, possible attributes for each dimension are decided. We finally create classes based on them. Notice that there are various CF algorithms. Namely, there are memory- or model-based algorithms and hybrid schemes. Different approaches might apply different soft computing techniques like Bayesian classification, neural network approaches, self-organizing map (SOM) clustering, and so on. Moreover, some utilize binary ratings while others use numeric ratings (discrete or continuous). Each algorithm might have different weaknesses or disadvantages. Similarly, they might have various challenges related to their applications. Our focus here is not CF algorithms and their challenges. We want to focus on the challenges of PPCF schemes.

To better scrutinize the possible challenges introduced during applications of PPCF methods, we first determine dimensions, which can be used for classification. After investigating various PPCF schemes, we conclude the following three major dimensions for classification:

A. Data Partitioning-DP
B. Data Partitioning Configuration-DPC
C. Privacy-preserving Measure-2PM

The dimensions have different attributes that can be used for classifying various PPCF approaches. In the following, we explain each dimension and their related attributes.

A. Data Partitioning (DP)

Data collected for CF purposes might be held by different entities. Such data collected as ratings are saved in an $n \times m$ matrix, called user-item matrix **D**, where n and m represent number of users and items, respectively. The way in which data are held is called data partitioning (DP). DP is an important dimension that might affect the applicability of various PPCF schemes. Different partitioning cases may pose different challenges. In terms of DP, we determine the following attributes:

I. Central Server-based DP-1P: In some applications like client-server (the systems run by Yahoo, Amazon, etc.), collected data are held by a single vendor, referred to as central server-based DP (**1P**). The server gathers ratings from n users for m items; and then saves them in **D**. Notice that **D** is held by a single party only. The data collector or the server controls all collected data. When an active user (a) who is looking for a prediction for a target item (q) wants a prediction (p_{aq}), she communicates with the server. After estimating p_{aq}, the server returns it to a.

II. Distributed Data Partitioning-2DP: Unlike **1P** partitioning case, users' preferences might be distributed between various online vendors, referred to as distributed DP (**2DP**). Some online vendors, especially recently established companies, might have problems collecting enough ratings for CF purposes. Those companies holding scarce data may want to integrate their data for better services. In **2DP**, **D** is distributed among various number of e-commerce sites, even competing ones. When data are distributed, there are three possible cases, as follows:

a. Partitioned Data-2P: In this case, **D** is partitioned between two companies only. Thus, we call this partitioning as partitioned data or **2P**. In this scenario, a asks a prediction from one of the parties. The party estimates the predictions based on integrated data.

b. Multi-party Distributed Data Partitioning-MP: Unlike **2P**, data are distributed among multiple parties. In other words, **D** is partitioned among more than two parties. We call this data partitioning case as multi-party or **MP**. Notice that if M shows number of parties, then M is usually a small constant and $M << n$ and $M << m$. In the **MP**, one of the collaborating companies acts as a leading party that communicates with both a and other helping vendors.

c. Fully Distributed Data Partitioning-P2P: Although data are distributed in this case, each user actually acts as an involving company. This type of partitioning is common in social network applications like one-to-one (chat) communication applications. When data are distributed among n parties or users, we call this scheme as peer-to-peer (**P2P**).

As expected, abovementioned partitioning cases might pose different challenges and application issues because each scenario has its own specific properties. Therefore, we investigate application challenges of PPCF schemes after grouping them into different classes using data partitioning attributes. Fig. 1 shows these data partitioning cases. On one extreme point, data are held by a single party (**1P**), on the other extreme point; **D** is fully distributed among n users (**P2P**). Between these two extreme points, we have partitioned and multi-distributed cases (**2P** and **MP**, respectively).

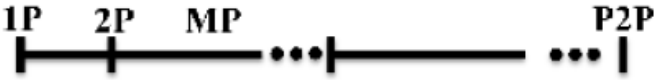

Fig 1. Data partitioning cases.

B. Data Partitioning Configuration (DPC)

Another dimension that can be used for classification is called data partitioning configuration (DPC). DPC represents how data are distributed between different online vendors. Data might be partitioned horizontally, vertically, or arbitrary. Hence, we determine the following attributes:

I. Horizontal Partitioning (HP): In horizontal partitioning, each party holds the ratings of the disjoint sets of users for the same items. Suppose that data are partitioned among C vendors, where C can be 2 (representing **2P**), $2 < C << n$ (representing **MP**), or $C = n$ (representing **P2P**). Each party c holding ratings of disjoint sets of n_c users for the same m items, where $c = 1, 2, \ldots, C$. Thus, the integrated data has the size $(n_1 + n_2 + \ldots + n_C) \times m$, while split data held by each vendor c, called $\mathbf{D}_c$, has the size $n_c \times m$.

II. Vertical Partitioning (VP): Unlike horizontal partitioning, in vertical partitioning, each party has the same users' ratings for the disjoint sets of items. Assume again that data are partitioned among C vendors, where C can be 2 (representing **2P**), $2 < C << m$ (representing **MP**), or $C = m$ (representing **P2P**). Each party c holding ratings for disjoint sets of m_c items of the same n users, where $c = 1, 2, \ldots, C$. Thus, the integrated data has the size $n \times (m_1 + m_2 + \ldots + m_C)$, while split data held by each vendor c, called $\mathbf{D}_c$, has the size $n \times m_c$.

III. Arbitrary Partitioning (AP): Arbitrary partitioning is the combination of both horizontal and vertical partitioning. Given the **D** including the ratings of n users for m items, each party holds some of the ratings in **D**. Suppose that v_{ij} represents a rating of user i for item j in **D**, then each company owns v_{ij} values for some i and j. As an example, we show AP for two parties only in Fig. 2. We assume that **D** is partitioned between two companies, c_1 and c_2. Given m products, users usually rate some of them leading to very sparse data. Hence, some of the cells are unrated in Fig. 2.

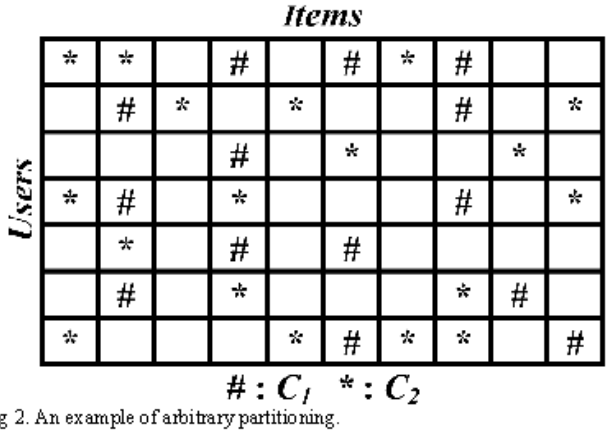

Fig 2. An example of arbitrary partitioning.

C. Privacy-preserving Measure (2PM)

In PPCF schemes, users' preferences as ratings and the rated and/or unrated items are considered confidential. For any data owner, either a customer or a company, performing CF services without divulging private data is imperative. With increasing popularity of privacy-preserving data mining, researchers have been recommended different privacy-preserving measures. Since CF can be considered as a subset of data mining, those techniques invented for preserving privacy in data mining functionalities can be applied to and have been applied to CF schemes, as well.

In traditional CF algorithms without privacy concerns, various challenges have been identified and efficient solutions have also been proposed. When privacy is a concern, offering CF services while achieving confidentiality brings some issues, too. Two major constraints that each CF scheme should satisfy are accuracy and efficiency (online performance). However, as expected, applying privacy-preserving measures might make accuracy and/or online performance worse. The reason for this phenomenon is that privacy, accuracy, and efficiency are conflicting goals. Different privacy-preserving technique might bring different issues. Although effects of privacy measures on accuracy and online performance have been investigated, there are still open questions about how such measures affect usability of PPCF schemes. Moreover, it is likely that the effects of different privacy-preserving methods cause different issues or challenges related to applications of PPCF schemes as a part of recommender systems used by various e-commerce sites. Therefore, we consider privacy-preserving methods as another dimension for grouping PPCF schemes; and scrutinize challenges of applications of PPCF schemes in terms of such methods. We determine the following techniques used as privacy-preserving methods in PPCF schemes:

I. Randomization Techniques (RT): Randomized methods have been widely used in PPCF schemes for data masking.

Randomization-based approaches are useful when we are interested in aggregate data. There are three randomization schemes applied to CF algorithms for privacy-preserving:

a. Randomized Perturbation Techniques (RPT): Random numbers, drawn from over a specified range with mean being 0 using some distribution like uniform or Gaussian, are added to confidential data. Instead of actual data items, perturbed data values are sent to data collectors. Such techniques are suitable for numeric ratings-based schemes and widely used in central server-based PPCF methods.

b. Randomized Response Techniques (RRT): When users' preferences are represented using binary ratings, RRT are utilized for data masking. According to the result of a comparison between a specified value (θ) and a random number (r), data owners decide whether to send true data or false data (exact opposite of actual ratings) to data collectors. RRT are useful in central server-based PPCF schemes. There are two variants of RRT. They are one-group scheme (**1G**) and multi-group scheme (**MG**).

1. In **1G**, all ratings are placed in a vector. Entire ratings are disguised together.
2. In **MG**, ratings are divided into groups and each group is masked independently.

c. Random Filling (RF): In addition to perturbing real ratings, it is also crucial to mask rated and/or unrated item cells. To do so, PPCF schemes also utilize random filling (**RF**) in which some uniformly randomly chosen cells are filled with some default votes or noise data.

II. Cryptographic Techniques (CT): In addition to randomized methods, cryptographic methods have been also widely used for achieving privacy in PPCF schemes. Most commonly utilized CT are given in the following:

a. Homomorphic Encryption (HE): Encryption functions with homomorphic property allow performing certain operations on encrypted data without decrypting them [17]. HE methods can be used in various systems for ensuring privacy of processed data. For example, assuming that ξ is an encryption function, K is a public key, and x and y are private data values, HE property allows an addition operation to be conducted based on the encrypted data without decrypting them, as follows: $\xi_K(x) \times \xi_K(y) = \xi_K(x + y)$. HE methods happen to be secure schemes. However, they have efficiency and implementation problems.

b. 1-out-n Oblivious Transfer Protocol (OT): OT refers to a protocol, where at the beginning of the protocol one party, Bob has n inputs $X_1, \ldots, X_n$ and at the end of the protocol the other party, Alice, learns one of the inputs X_i for some $1 \leq i \leq n$ of her choice, without learning anything about the other inputs and without allowing Bob to learn anything about i.

c. Secure Multi-party Computation (SMC): SMC enable two or more parties to evaluate a specified function of their inputs collaboratively without disclosing their inputs to each other. When multiple companies want to perform some computations jointly while keeping their private data secret, they might decide to use SMC protocols.

III. Data Obfuscation (DO): Basic idea behind DO methods is to substitute private data values with some other data whose disclosure does not violate privacy constraint. DO techniques used in PPCF schemes substitute real ratings with predefined values or their neighbors in the metric space. Berkovsky et al. [18] list three general policies for obfuscating the ratings in the user profiles in CF systems for accomplishing confidentiality. *Default obfuscation* of real rating votes, v_{ij} values, means that substituting the real ratings in the user profile with a fixed predefined value. *Uniform random obfuscation* allows substituting the real ratings in the user profile with random values chosen uniformly in the range of ratings in the dataset. Finally, in *bell-curved random obfuscation*, real ratings in the user profile are substituted with values chosen using a bell-curve distribution reflecting the distribution of ratings in the data set. Paremeswaran and Blough [7] propose to use *Nearest Neighbor Data Substitution* (NeNDS) approach to CF systems for protecting privacy. NeNDS substitutes individual data elements with one of their neighbors in the metric space.

IV. Others: There are some other techniques used in PPCF schemes as privacy-preserving measures. Anonymity and k-anonymity are among such methods. They are not widely used like randomization and cryptographic techniques. Anonymity prevents data collectors from learning the origin or the owner of the ratings. k-anonymity demands that every tuple in the data table released be indistinguishably related to no fewer than k respondents.

In Fig. 3, we show the dimensions we determined, which can be used to classify PPCF schemes. Also, we display the related attributes for each dimension. The figure shows the big picture of PPCF schemes in terms of various aspects.

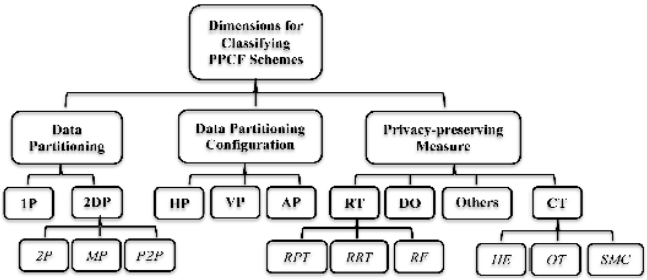

Fig 3. Dimensions and their related attributes for classifying PPCF schemes.

IV. Applications of PPCF Schemes: Challenges

In theory, theory and practice are the same; but in practice, they never are[3]. Although everything is clearly depicted and organized during design stages of any system, unexpected problems might occur due to avoidable and/or inevitable challenges. We describe such challenges in the following.

A. Data Losses during Communication

CF is designed to automate the "word-of-mouth" habit in web environment and hence, all PPCF applications run over

[3] Albert Einstein.

such applications; e.g., typically web browsers. Users of PPCF applications interact with the system through an unreliable packet-switched network infrastructure, which is vulnerable to packet losses during transmission. Since PPCF is an interactive and online process, users require quick and accurate responses to their queries. If any query or part of preference data is lost on the way through, the server might not be able to produce a recommendation. Even worse, after successful transmission of preference data and query, produced recommendation might not be delivered to the client side. Both can reduce credibility of PPCF systems and lead customers to refrain from using such systems. Similar problem can be observed in distributed PPCF schemes. Notice that the collaborating parties exchange data during off-line and online stages. Although data losses during off-line computations can be overcome due to unrestricted requirements, data losses during online process might hinder the quality of predictions. When some of the data required for estimating recommendations get lost during online phase, accuracy losses are expected because smaller amount of data is used for generating predictions reducing the gains due to collaboration.

B. *User Interface Design*

According to Stone et al. [19], "good user interface design encourages an easy, natural, and engaging interaction between a user and a system, and it allows users to carry out their required tasks." However, PPCF schemes increase the visual design and algorithmic complexities. Without privacy-preserving measures, it is easy to create a user-friendly interface. Privacy concerns make it difficult for designers to design intelligible interfaces. Moreover, e-commerce sites must serve prediction services based on the privately collected data by means of web sites having a clear visual design.

C. *Unconscious Users*

One goal of PPCF systems is to protect individual privacy. However, the term "privacy" is getting interest through developments in the information systems and most of the average users of such systems are unaware of this phenomenon. Privacy, in the context of PPCF systems, includes both hiding true preferences and rated items. Such aspects are controlled via different parameters and selecting their values optimally due to personal requirements is vital because privacy protection and accuracy of recommendations are inversely correlated. If users are unconscious of those aspects and/or uninformed on how to choose privacy parameters optimally, all effort might become void for designing PPCF applications.

Cranor et al. [20] carried out a survey on the Internet users concluding that privacy concerns differs from one user to another. There are PPCF algorithms, which are sensitive to users' different privacy characteristics [21, 22]. To realize such schemes, PPCF web designers must consider various privacy concern levels of the Internet users. One of the application challenge of PPCF is that designed web site should serve from unconscious users to sophisticated users with all concerning about privacy in different levels. For those users living in developed countries, preserving privacy is vital. Customers are aware of their privacy and situations causing privacy violations. They act in such a way so that their privacy is not jeopardized. However, those users living in developing countries are not that aware of their privacy. Confidentiality might not be a big concern for them. Thus, they might hesitate to use PPCF schemes.

D. *Overheads due to Cryptographic Techniques*

Cryptographic techniques such as HE and OT protocol are effectively utilized in PPCF schemes. Due to their computational complexity, such techniques are heavily placed in off-line phase. However, there are also schemes offering them in online transactions [15]. To respond for user prediction request in decent time, such cryptographic overheads are cumbersome.

Although some works are performed off-line in PPCF applications, in essence, it is an online process; and like all online applications, the bottleneck of PPCF is time constrained. However, while encryption schemes provide a secure and privacy-preserving framework, they lead much computation complexity. Such side effects of encryption schemes must be overcome to make PPCF systems operate online and respond within reasonable amount of time.

E. *Practical Problems due to Encryptions*

To overcome privacy challenges, it is inevitable employing HE schemes. However, HE has no natural way of encrypting and decrypting floating point numbers and negative integers. In CF algorithms, similarities between users are operated and they are from the range [-1, 1]. When data are distributed, the parties perform sub-computations with their private data to use them in online recommendation process [14]. To protect their privacy, they encrypt those interim results, which are generally floating point numbers. Besides distributed data-based CF schemes, some solutions for protecting individuals' privacy in central server-based CF systems employ HE solutions. Thus, encryption of floating point numbers and negative integers are practical challenges that might be overcome for PPCF schemes.

F. *Synonymy*

In distributed data-based PPCF schemes, especially in HDD-based schemes, it is assumed that all collaborating parties own the same item sets. In other words, they provide predictions for the same products. However, different names might refer to the same item. Different parties might use different names for representing the same items leading synonymy problem.

V. SMOOTHING CHALLENGES: PROPOSED SOLUTIONS

Described challenges on deploying PPCF systems might be eliminated at all or their effects can be lowered by means of some application level measures. We suggest some procedures to handle referred problems on applications of PPCF schemes.

To reduce packet losses during communications between client applications and the server, or between multiple parties, secure and bandwidth allocated channels might be used and such applications need to be run over TCP to guarantee packet transmissions. In addition to TCP's 3-way handshaking reliable packet transfer protocol, also application level implementations can be deployed to assure successful data exchange and inform users and the server on this. Therefore,

an extra module responsible for reliable data transmission issues can be very useful to avoid such inconveniences.

To smooth above-mentioned user design challenges, first of all, rather than entering numerical privacy-adjusting parameters from keyboard, the system should offer user privacy level selection bar as in Fig. 3. Similar bar is utilized in Jester recommender system [23] to collect continuous ratings from users. Visuality facilitates user comprehension about privacy, enhances the user comfort and simplifies the use of interface. Moreover, if user clicks *no privacy* region, system ought to warn user as "you are publicly saving your rating information." Touch screen technologies also contributes visuality and can be considered in use for additional private information collection frameworks. Alternatively, privacy menu can be used by users for selecting the required level of privacy. For generation of such menu, one classification is realized by Aïmeur et al [24]. Considering different levels of concerns, they classify users into four classes: no privacy, soft privacy, hard privacy, and full privacy. Similar classification and labels can be given for particular value of privacy parameters. Finally, PPCF applications must provide a user-friendly GUI to make users feel comfortable on using such systems offering easy-to-use interactive interfaces to adjust level of privacy. Also, a detailed but easy to follow "Help" section must be made available to inform users about the aspects of their individual privacy and how to tune parameters of them.

Fig. 4. Privacy level selection bar.

To handle overhead costs due to cryptographic methods, parallel computation techniques can be used. E-commerce parties can apply parallel and concurrent computation techniques to improve online time. If the selected cryptographic method is not parallelizable satisfactorily, parallelizable algorithms can be preferred. For instance, Kamara and Raykova [25] propose a parallel HE algorithm. Moreover, application-oriented hardware design can be another alternative. If the cryptographic tools must be used in online and parallel computation is unavailable or insufficient, then application-oriented hardware may be a solution. Also, rapidly evolving hardware, software, and information technologies may produce a solution to smooth such problem. Finally, some preprocessing schemes and model-based approaches must be followed to lower their side effects on online performance and accuracy. Extensive experiments must be performed periodically before and during publication of PPCF schemes. Also they must be updated occasionally according to experimental findings and latest improvements in scientific research area.

Encrypting floating point numbers and negative integers is not a challenge only for PPCF schemes. It is a general privacy-preserving data mining problem. Thus, researchers introduce solutions for practical challenges of utilizing HE. Pathak and Raj [26] study computing eigenvectors from distributed data without jeopardizing data holders' privacy. Since eigenvalue computation requires working with floating numbers and negative integers, the authors propose practical solutions for encrypting such values. To encrypt floating point numbers, they propose to floor this numbers by a fixed constant that is power of two for efficient arithmetic. After decryption, they divide the results with the same constant to get the actual result. In PPCF schemes, the same approach can be employed as a solution introduced by the authors to encrypt positive floating point numbers. To encrypt negative floating numbers, we can shift the interval of the numbers into a positive range after we can use the solution. In the case of encrypting negative integers using an HE, Pathak and Raj [26] represent negative numbers in binary two's complement system. It is possible to perform arithmetic operations such as addition, subtraction, and multiplication using two's complements numbers. In PPCF schemes, the proposed solution overcomes challenges caused by encryption of negative integers. In another study, Pathak et al. [27] utilize a similar solution to encrypt floating point numbers.

Collaborating parties make sure that they share or use the same item names for the same products before they start providing recommendations collaboratively to resolve synonymy problems. In order to offer predictions for the same items, the parties must use the same IDs or names referring for the same products. Similar case is true for the user IDs. Collaborating companies also utilize the same IDs representing for the same users. Hence, before they start collaborative work, they should agree on user and item IDs and/or names referring for the same entities.

VI. Conclusions and Future Work

Although there are various studies describing how to offer predictions with privacy, they fail to explain possible challenges that might occur while implementing the proposed solutions in real life applications. Therefore, in this paper, we first determined possible attributes and dimensions with which the proposed privacy-preserving collaborative filtering schemes can be classified. We then classified such schemes in terms of data partitioning, partitioning configuration, and utilized privacy-preserving measures. We also investigated them in terms of the challenges and problems that they might face with while utilizing them in real applications. We finally suggested some solutions and approaches to remove such challenges at all or reduce their effects.

Similar application issues might occur during implementation of such methods. After intensive user requirement analyses, more issues can be determined. More work should be conducted to suggest possible solutions to overcome such practical challenges.

References

[1] M. Grcar, "User profiling: Collaborative filtering," in *7th International Multiconference Information Society IS 2004* Slovenia, 2004, pp. 75–78.

[2] OECD, "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data," 2005.

[3] OECD, "Guidelines for Consumer Protection in the Context of Electronic Commerce," 2000.

[4] J. Canny, "Collaborative Filtering with Privacy," in *Proceedings of the 2002 IEEE Symposium on Security and Privacy*, Oakland, California, USA, 2002, pp. 45-57.

[5] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering," *International Journal of Electronic Commerce,* vol. 9, pp. 9-35, 2005.

[6] I. Yakut and H. Polat, "Privacy-Preserving Eigentaste-Based Collaborative Filtering Advances in Information and Computer Security." vol. 4752, A. Miyaji, H. Kikuchi, and K. Rannenberg, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 169-184.

[7] R. Parameswaran and D. M. Blough, "Privacy Preserving Collaborative Filtering Using Data Obfuscation," in *Proceedings of the 2007 IEEE International Conference on Granular Computing*, Silicon Walley, CA, USA, 2007, pp. 380-386.

[8] C. Kaleli and H. Polat, "Providing Private Recommendations Using Naïve Bayesian Classifier," in *Advances in Intelligent Web Mastering*. vol. 43, K. Wegrzyn-Wolska and P. Szczepaniak, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 168-173.

[9] C. Kaleli and H. Polat, "P2P collaborative filtering with privacy," *Turkish Journal of Electric Electrical Engineering and Computer Sciences,* vol. 8, pp. 101-116, 2010.

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

[11] R. Shokri, P. Pedarsani, G. Theodorakopoulos, and J.-P. Hubaux, "Preserving privacy in collaborative filtering through distributed aggregation of offline profiles," presented at the Proceedings of the third ACM conference on Recommender systems, New York, New York, USA, 2009.

[12] C. Kaleli and H. Polat, "Providing Naïve Bayesian Classifier-Based Private Recommendations on Partitioned Data," in *Knowledge Discovery in Databases: PKDD 2007*. vol. 4702, J. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenic, and A. Skowron, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 515-522.

[13] I. Yakut and H. Polat, "Privacy-preserving SVD-based collaborative filtering on partitioned data," *Int. J. Information Technology & Decision Making,* vol. 9, pp. 473–502, 2010.

[14] C. Kaleli and H. Polat, "SOM-based recommendations with privacy on multi-party vertically distributed data," *Journal of Operational Research Society,* 2011.

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

[16] I. Yakut and H. Polat, "Privacy-preserving hybrid collaborative filtering on cross distributed data," *Knowledge and Information Systems,* vol. 30, pp. 405-433, 2012.

[17] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proceedings of the 17th international conference on Theory and application of cryptographic techniques*, Prague, Czech Republic, 1999, pp. 223-238.

[18] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci, "Privacy-Enhanced Collaborative Filtering," in *Workshop on Privacy-Enhanced Personalization, at the International Conference on User Modeling*, Edinburgh, UK, 2005, pp. 75-83.

[19] D. Stone, C. Jarret, M. Woodroffe, and S. Minocha, *User Interface Design and Evaluation*: Morgan Kaufmann 2005.

[20] L. F. Cranor, J. Reagle, and M. S. Ackerman, "Beyond Concern: Understanding Net Users' Attitudes About Online Privacy," AT&T Labs-Research1999.

[21] H. Polat and W. Du, "Effects of inconsistently masked data using RPT on CF with privacy," in *Proceedings of the 2007 ACM symposium on Applied computing*, Seoul, Korea, 2007, pp. 649-653.

[22] I. Yakut and H. Polat, "Achieving Private SVD-based Recommendations on Inconsistently Masked Data," in *Achieving Private SVD-based Recommendations on Inconsistently Masked Data*, 2007.

[23] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A Constant Time Collaborative Filtering Algorithm," *Inf. Retr.,* vol. 4, pp. 133-151, 2001.

[24] E. Aimeur, G. Brassard, J. M. Fernandez, and F. S. M. Onana, "Alambic: a privacy-preserving recommender system for electronic commerce," *International Journal of Information Security,* vol. 7, pp. 307-334, 2008.

[25] S. Kamara and M. Raykova, "Parallel Homomorphic Encryption," Microsoft Research2011.

[26] M. Pathak and B. Raj, "Privacy preserving protocols for eigenvector computation," presented at the Proceedings of the international ECML/PKDD conference on Privacy and security issues in data mining and machine learning, Barcelona, Spain, 2011.

[27] M. Pathak, S. Rane, S. Wei, and B. Raj, "Privacy preserving probabilistic inference with Hidden Markov Models," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 5868-5871.