A Trust-Based Recommender System Over Arbitrarily Partitioned Data With Privacy

Ibrahim Yakut and Huseyin Polat

SIRF TRADE, 38050 Melikgazi, Kayseri, Turkiye

E-mail:

iyakutcs@gmail.com

hpolat@gmail.com

Corresponding Author

Ibrahim Yakut

SIRF TRADE

38050 Melikgazi, Kayseri, Turkey

E-mail: iyakutcs@gmail.com

A Trust-based Recommender System over Arbitrarily Partitioned Data with Privacy

Ibrahim Yakut and Huseyin Polat

ABSTRACT

Recommender systems are effective mechanisms for recommendations about what to watch, read, or taste based on user ratings about experienced products or services. To achieve higher quality recommendations, e-commerce parties may prefer to collaborate over partitioned data. Due to privacy issues, they might hesitate to work in pairs and some solutions motivate them to collaborate. This study examines how to estimate trust-based predictions on arbitrarily partitioned data in which two parties have ratings for similar sets of customers and items. A privacy-preserving scheme is proposed, and it is justified that it efficiently offers trust-based predictions on partitioned data while preserving privacy.

Keywords: arbitrarily partitioned data, privacy, trust, collaborative filtering, sparsity, accuracy

1. INTRODUCTION

Recommender systems are substantial for online shopping or digital content services, which provides personalized predictions to particular user's tastes [1]. Collaborative filtering (CF) algorithms effectively and serendipitously handle rating profiles of a set of users to provide personalized recommender system services. These systems take inspiration from *word of mouth*, which spreads knowledge by exchanging information between people having similar tastes [2]. Conventionally speaking, a CF system consists of n users' preferences about m products. The main functionality of the system is to estimate a prediction for an active user (a) about a target item (q), referred to as p_{aq}. Traditional prediction estimation process includes (*i*) calculating similarities between a and each user u in the database, (*ii*) finding a's neighbors, and (*iii*) computing a prediction based on such neighbors' data using an algorithm.

Trust is complex term with multiple facets [3, 4]. In general, trust is hard to be built but easy to be collapsed. In terms of artificial intelligence, the trust and trustworthiness are the first requirement rather than performance issues such as efficiency and accuracy. Concerns related to risk, trust, and security are emerging with the rising prevalence of AI systems. One of the most beneficial solutions for ensuring the reliability and trustworthiness of AI systems is AI trust, risk, and security management (AI TRiSM) framework, which plays a vital role for organizations in ensuring the proper regulation to deploy AI models and effective management [5]. According to Nusrat and Vassileva [6], trust can be defined as one person's belief in another person's capabilities, which is needed to differentiate good members of society from bad ones. With the proliferation of online social networks, trust-based relationships and recommendation systems have come into prominence [3, 7]. Trust concept has been already discussed in terms of the applicability in recommender systems [8-10]. Massa and Bhattacharjee [8] show that trust-based metrics can be applied to determine relation between users to compute prediction about items. Hence, trust metrics are applied to CF algorithms and satisfactory results are obtained. Massa and Avesani [9] offer propagated trust metric to improve the coverage of trust-based CF systems. Hwang and Chen [10] present a CF method deriving both direct and propagated trust values from traditional rating profile data. They experimentally show that rating-based trust approach offers better referrals over correlation-based CF methods.

Data sparsity is crucial problem to drive recommender systems effectively [1, 11]. Such problem faced by some e-companies brings about privacy-preserving partitioned data mining (P3DM) solutions. In such solutions, the goal is to promote quality and coverage of data mining services via contribution of additional data of other party while ensuring data confidentiality. During such mining processes, the foremost challenge is to preserve privacy of each party. Unless their confidentiality is ensured, such companies are expected to go through serious legal and financial deadlocks in managerial operations. Online vendors are responsible for protecting customer profile data [12]. Moreover, such data are valuable asset due to enhancing web personalization facilities. Hence, privacy concerns must be satisfied so that e-companies are able to cooperate for better mining purposes.
Bilge et al. [13] investigate various privacy-preserving recommendation methods according to the data configurations and the utilized techniques for preserving confidentiality. They address limitations and practical

implementation challenges of state privacy-preserving recommendation systems in literature. Corporate privacy is an issue, where two or more parties want to collaborate over rating profiles data to drive higher quality recommendation services. Privacy-preserving collaboration over data is desirable for many parties to benefit from each other's data without harming their privacy [14]. Data holders might end up with alternatively three kinds of data configurations; namely, horizontal, vertical, or arbitrary. Arbitrary partitioning can be considered as the combination of both horizontal and vertical partitioning. There are numerous studies on horizontally partitioned data (HPD) and vertically partitioned data (VPD) covering broad range of data mining tasks [14-18]. Besides horizontal and vertical partitioning cases, there are some solutions for arbitrarily partitioned data (APD), which is more probable and convenient than others in real life scenarios [19, 20]. APD can be formed when two parties have similar sets of customers and product portfolios. In APD, while some part of user rating profile arbitrarily exists in either of the parties' data, the remaining part of user profile is also arbitrarily exists in the other party, as shown in **Figure 1**, where *A* and *B* show two online vendors.

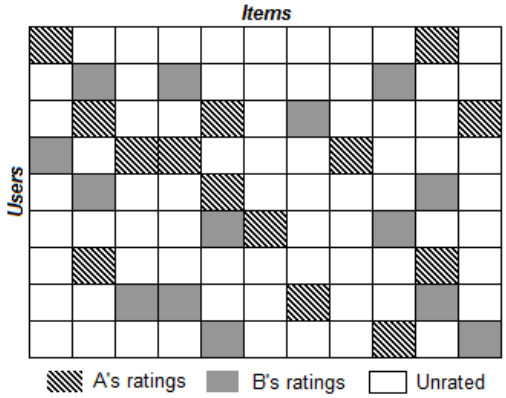

Figure 1: Arbitrarily partitioned rating profile data

In this work, we investigate how to provide trust-based predictions on APD without violating the data holders' confidentiality. In e-commerce, trust is directly related factor to customer privacy and loyalty as depicted in [21], as well. We focus on trust-based CF mechanism while preserving privacy of each parties' data. We propose a privacy-preserving scheme to provide trust-based referrals efficiently on APD with privacy. Since privacy-preserving is the main factor help e-commerce sites cooperate, accuracy is the key parameter to measure CF output quality, and efficiency is the core requirement for online responding information systems, the proposed scheme should provide privacy, accuracy, and efficiency. However, they are conflicting goals. Despite of such conflicting goals, we propose a computationally achievable solution providing privacy requirements and responding quality predictions. We also justify such solution in terms of privacy, supplementary overheads, and prediction quality via analyzing it theoretically and empirically.

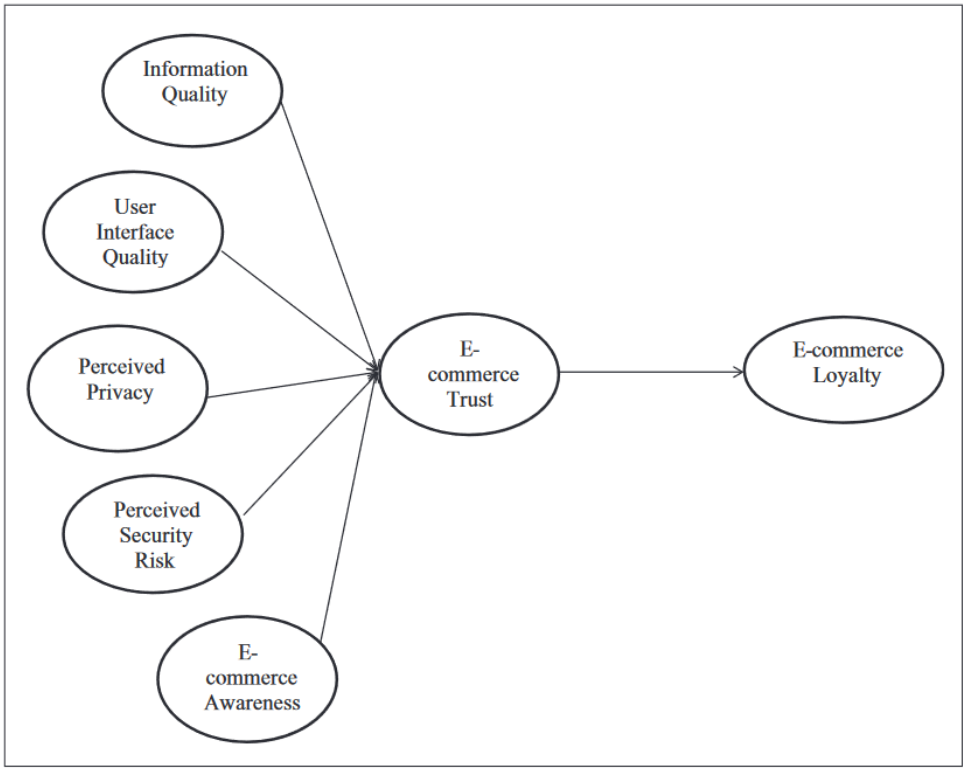

Figure 2. Factors that affect e-commerce trust and loyalty [21]

The paper is structured, as follows. Section 2 presents the related work highlighting the studies among the state-of-the-art. After introducing trust-based CF mechanism and the related preliminaries in Section 3, we present our privacy-preserving trust-based CF solution on APD in Section 4. Then, the proposal is theoretically examined in terms of privacy and supplementary costs in Section 5 and Section 6, respectively. While experimental analysis is presented and results are discussed in Section 8, allied conclusions are drawn, and future directions are remarked in the last section.

2. RELATED WORK

There are many privacy-preserving collaborative filtering (PPCF) schemes concentrating on user privacy protection when interacting with data collectors [22, 23]. Rather than such individual privacy driven CF studies, there are also other studies focused on partitioned data-based CF with privacy. Polat and Du [24, 25] introduce privacy-preserving partitioned collaborative filtering (P3CF) problem. They examine how to realize recommendations using conventional CF algorithm on VPD [24] and propose a top-*N* recommendation solution for HPD [25]. Yakut and Vaidya [26] investigate how to provide CF services on partitioned data between two parties using item-based algorithms while considering overlaps. They propose novel P3CF solutions when partitioned data is overlapped. Yakut and Polat [27] examine model-based CF in the context of P3CF and offer an SVD-based solution for two party suffering from data sparsity. Hsieh et al. [28] focus on correlation-based CF on HPD and propose a P3CF framework utilizing El Gamal-based homomorphic encryption. In the study [29], there are two party: one client company and one cloud computations service provider. They focus on integrity of recommendation computation services results, as a security service, on the other hand our security goal is to achieve confidentiality of data collaboration. Zhan et al. [30] empirically investigate efficiency issues in P3CF on HPD by comparing computation and transportation time costs of El Gamal-, commodity-, and their revised commodity-based approach; and their experimental findings show that the revised approach outperforms the others. Comparing to the above-mentioned P3CF solutions, we focus on more complex data partitioning scheme. Moreover, our scheme focuses on trust-based CF mechanism rather than correlation-based ones.

There are some PPCF schemes focusing on trust metrics [31, 32]. In [31], the authors investigate optimal privacy in trust-aware social networks using randomized disguising techniques as a preprocessing step. While their scheme

is a solution for such networks in peer-to-peer manner, in this proposal, there are two parties whose data constitute APD. In [32], the authors examine how to provide trust-based recommendations on VPD among multiple parties. Unlike their study, this work focuses on estimating trust-based predictions from APD between two parties only.

Performing various data mining functionalities on APD with privacy has been receiving increasing attention. Prasad and Rangan [33] offer a privacy-preserving clustering solution on APD. One clustering solution on APD with privacy is offered by Upmanyu et al. [34]. Their proposal is based on cloud computing using the paradigm of secret sharing. Han and Ng [35] propose a privacy-preserving decision tree induction algorithm on APD among multiple parties. They present an efficient method performing the secure scalar product operation on APD among multiple parties. Li et al. [36] offer distance-based outlier detection protocol on APD with confidentiality. In another study, Yunhong et al. [37] propose a privacy-preserving support vector machine (SVM) classifier scheme on APD. While their scheme bids SVM classifier as public, it does not divulge any privately held data. Bansal et al. [38] present a privacy-oriented scheme for neural network learning on APD. Yuan and Yu [39] study back-propagation learning with privacy via cloud computing on APD. Shi et al. [40] present a privacy preserving protocol for the multiparty training of growing neural gas while the data are arbitrarily partitioned over different parties.

Yakut and Polat [41] examine APD for PPCF systems. Their work initiates how to realize PPCF services on APD. Our work differs from their study due to the utilized CF algorithm. While their work is based on item-based CF algorithms, we focus on trust-based mechanisms used in CF systems. Yakut and Polat [42] also investigate how to realize naïve Bayesian classifier-based CF services on arbitrarily partitioned binary data and propose a novel P3CF solution. There are two main differences between the problem definition of ours and their study: type of rating data and exploited CF method. In another work [43], the same authors propose a privacy-preserving hybrid CF solution on cross partitioned data. Their CF algorithm and data partitioning are different than the ones we utilize here. Basu et al. [44] discuss the feasibility of multi-party HPD- or VPD-based Slope One predictor schemes on APD among multiple parties rather than proposing a complete APD-based solution. This work also differs from theirs in terms of CF mechanism and we offer a complete trust-based CF scheme on APD across two parties.

3. PRELIMINARIES

Hwang and Chen [10] define trust between users *a* and *u*, $t_{a\to u}$, which means how much *a* trusts *u*, or vice versa. The trust can be computed, as follows:

$$t_{a\to u}=\frac{1}{|I_a\cap I_u|}\sum_{j\in J}\left(1-\frac{|p_{aj}^u-v_{aj}|}{\rho}\right), \tag{1}$$

where I_a and I_u stand for the rated item sets of users *a* and *u*, respectively, *J* is the set of items rated by both users, and ρ is the range of the operated ratings, p_{aj}^u is prediction for trust computation and it can also be derived, as follows:

$$p_{aj}^u=\overline{v_a}+(v_{uj}-\overline{v_u}), \tag{2}$$

where v_{uj} is the rating of item *j* given by *u*, $\overline{v_a}$ and $\overline{v_u}$ are mean ratings of users *a* and *u*, respectively. Hwang and Chen [10] also introduce trust propagation metric in order to evaluate trust values between users who have no commonly rated items, as shown in Eq. (3):

$$t_{s\to\underset{v}{h}}=t_{s\to v}\oplus t_{v\to h}=\frac{|I_s\cap I_v|\times t_{s\to v}+|I_v\cap I_h|\times t_{v\to h}}{|I_s\cap I_v|+|I_v\cap I_h|}, \tag{3}$$

where users *s* and *h* are non-commonly rated users, however, *v* has co-rated items with both of them. The final inferred trust $t_{s\to h}$ is the average of the values for each user *v* computed by Eq. (3). After computing trusts between users, p_{aq} can be computed, as follows:

$$p_{aq}=\overline{v_a}+\frac{\sum_{u\in S}(v_{uq}-\overline{v_u})\times t_{a\to u}}{\sum_{u\in S}t_{a\to u}}, \tag{4}$$

where *S* stands for the users who have rated *q* and in trust neighborhood of *a*.

In order to ensure privacy constraints, we employ some cryptosystems such as homomorphic encryption (HE) and 1-out-of *n* oblivious transfer (OT) protocol. Since Paillier cryptosystem [45] avoids many of the drawbacks of the earlier homomorphic cryptosystems and provides faster encryption and decryption comparing to its alternatives [46], it is preferred to be utilized in this study. Paillier's HE scheme [45] allows to perform addition and multiplication operation *on ciphertexts* such as $\xi_K(X + Y) = \xi_K(X) \times \xi_K(Y)$ and $\xi_K(X \times Y) = \xi_K(X)^Y$, where X and Y are private data items while K is public key. This scheme also supports *self-blinding* property, which allows *ciphertexts* of the same plaintext be distinct to each other. OT protocol provides the secure exchange opportunity of one message over *n* values between two parties; one having *n* values and the other needs one of those values. It must satisfy three key requirements such as correctness of the essential value, confidentiality of *n* values, privacy of which one is needed. There are many OT studies done by cryptographic community and an example of efficient solution is given by Naor and Pinkas [47].

4. TRUST-BASED RECOMMENDATIONS ON APD WITH PRIVACY

Our privacy perspective envisages that the rating values and the rated and/or unrated items are confidential. Online vendors are obliged to keep them private. In case of collaboration, each party prevents the other from deriving useful information about such confidential data. It is assumed that the collaborating sites are semi-honest obeying the protocol while being ready to evaluate any obtained data in order to infer as much sensitive information as possible. *Principal privacy constraints* are violated by direct exchange of such ratings and/or the rated items. Additionally, *auxiliary privacy constraints* also prohibit any transaction conflicting principal privacy constraint between parties. Hence, the proposed protocol cannot allow any information leakage, which causes inference of any confidential values. In addition to assuming semi-honest parties, we also assume that any user can rate any item only one time in rating collection interface of any of the parties so that non-overlapping APD occurs between parties. In other words, it is assumed that the APD configuration is non-overlapping.

In the following, we describe our privacy-preserving trust-based CF on APD scheme. The proposal consists of four different sub-processes. At first, preprocessing is performed to determine user means and normalize data held by each party. Secondly, secure trust computation process is done covering users having commonly rated items. Then, for user pairs having no commonly rated item, trust propagation computation is taken place. Finally, how predictions are estimated online over the constructed models is explained. To estimate trust-based similarities, the method proposed by Hwang and Chen [10] is utilized due to the ease of availability of rating profile data.

4.1 Preprocessing

Preprocessing step includes estimation of user mean ratings from filled data. As seen from Eq. (2) and Eq. (4), normalized ratings are needed. User ratings are normalized using the deviation from mean approaches. Thus, to normalize user ratings using deviation from mean normalization and compute prediction for trust values, the parties need user mean ratings. User (row) mean can be expressed as *sum*/*count* of user ratings. To calculate user mean ratings, each party should find *sum* and *count* values based on its available data for each user and exchange them. However, if they directly exchange them, they can infer useful information about each other's confidential data. Rather than directly sharing of such values for each user, it is more convenient to compute and exchange them after filling some unrated cells with default votes (v_ds) so that the original *sum* and *count* values are kept private. Our proposed filling scheme can be briefly described, as follows:

1. Each company uniformly randomly chooses some of the unrated item cells to be filled.
2. They estimate v_ds based on their available data.
3. They fill chosen unrated cells with the related v_ds.

In the proposed filling scheme, v_ds can be determined alternatively, as follows:

1. For each item, each party computes the related v_d as the average from the ratings available for that item using POP algorithm [48].
2. As row-variant of the previous method, for each user, each party computes v_d as average from the ratings available for that user.
3. Overall mean of the available data can be utilized.

The other issue that should be addressed for the proposed filling scheme is the number of unrated cells to be filled. Such number can be determined, as follows:

1. Each party j uniformly randomly selects θ_j from the range $(0, \beta_j)$, where β_j is upper bound parameter of filling amount. The value of β_j should be selected in such a way so that the parties achieve the accuracy and privacy levels that they want.
2. Each party then uniformly randomly selects θ_j percent of their unrated cells to be filled with v_ds.

Now, the parties can calculate *sum* and *count* values for each user based on the filled data sets. They then exchange such values and compute user mean ratings. They finally normalize their ratings with the related user mean ratings using the deviation from mean approach.

4.2 Privacy-preserving trust computation

Eq. (1) can be written, as follows, after replacing p_{aj}^{u} by its equivalent given in using Eq. (2):

$$t_{a\to u} = \frac{1}{c_{au}}\left(c_{au} - \sum_{j\in J} \frac{\left|(\overline{v_a} - v_{aj}) + (v_{uj} - \overline{v_u})\right|}{\rho} \right), \tag{5}$$

in which $c_{au} = \left|I_a \cap I_u\right|$. Considering Eq. (5) and arbitrary data partitioning, the followings are observed:

1. Since the ratings range is publicly known, ρ is obviously public.
2. The value of c_{au} can also be made public (the parties can share the related partial values and the final aggregate) because they are estimated from filled (disguised) data.
3. Recall that to compute the trust value between users *a* and *u*, they must rate at least one item commonly. If this condition is satisfied, there are two cases of availability of the ratings for the commonly rated item *j*. In the first case, *full availability* occurs as both ratings are held by either *A* or *B*. Second case is *cross-wise availability* in which one of the ratings is held by *A* and the other is held by *B*. There is no problem for determination of commonly rated ones and computation of this expression in one-side full availability case. However, cross-wise available ratings make trust computation task challenging.
4. Trust values between any two users are symmetric, i.e., $t_{a\to u} = t_{u\to a}$.
5. *Absolute difference* - $\left|(\overline{v_a} - v_{aj}) + (v_{uj} - \overline{v_u})\right|$ makes the computation more challenging.

To compute trust values privately on APD, we propose privacy-preserving trust computation protocol (P2TCP), which is described in the following. Due to symmetric trust values, the matrix created for storing trust values is triangular matrix. It is proposed that such triangular trust matrix would be shared between two parties. In other words, half of the trust values are held by *A* and the remaining ones are held by *B*.

Protocol I: Privacy-Preserving Trust Computation Protocol

For the first half of the trust values, perform the followings:

For each distinct user pair (*a*, *u*), do the followings:

1. *A* and *B* compute and store absolute differences for fully available ratings in their own databases.
2. *A* and *B* ignore the set of ratings found in *Step* 1.
3. *A* encrypts all available $(\overline{v_a} - v_{aj})$ values using HE with its public key *KA* and obtains $\xi_{KA}(v_{aA})$. It then generates *N*-1 random vectors and hides the vector holding the rated item indices of v_{aA} into such random vectors. It finally send $\xi_{KA}(v_{aA})$ and all *N* vectors to *B*.
4. *B* encrypts all available $(v_{uj} - \overline{v_u})$ values using HE with *KA* and obtains $\xi_{KA}(v_{uB})$ values.
5. *B* also performs $\xi_{KA}(v_{aA} + v_{uB}) = \xi_{KA}(v_{aA}) \times \xi_{KA}(v_{uB})$ for only commonly corresponding item indices for each of *N* vectors.
6. For each different vector, *B* permutes each obtained values using its private permutation function f_B. Then, it sends all permuted values to *A*.
7. Using OT protocol, *A* takes the permuted set holding actual rated ones.
8. *A* decrypts them, takes absolute values, and accumulates them. *A* also adds initially found absolute differences for fully available ones. *A* now has the *half-trust* values.
9. Switching their roles, applying *Step* 3-8, *B* also has the *complementary half-trust* values.
10. *B* sends such complementary values to *A* that obtains the final trust values.

For the remaining half, *A* and *B* switch their roles, perform *Step* 1-10; and *B* obtains final trust values.

4.3 Trust propagation computation

After performing the P2TCP, each party ends up with its corresponding trust values. However, some of them are null because of the absence of co-rated items among any two users. Thus, the parties must utilize trust propagation metric given in Eq. (3) to determine trust values for such users. For any users *s* and *h*, if the required trust values and the numbers of commonly rated items are held by the same company, that party can easily estimate the trust propagation using Eq. (3). However, if they are held by different parties (for example, $\left|I_s \cap I_v\right|$ and $t_{s\to v}$ are held

by *A* and similarly, $|I_v \cap I_h|$ and $t_{v \to h}$ are held by *B*), they then compute partial aggregates for numerator and exchange the required data for numerator and denominator. They finally compute propagated trust values using Eq. (3).

4.4 Recommendation estimation

Eq. (4) is used to estimate a prediction for *a* on *q*. Suppose that the ratings of *q* held by *A* and *B* are labeled as v_{uqA} and v_{uqB}, respectively. Remember that the ratings (due to arbitrary partitioning) and the trust values (half of them are held by *A* and the remaining ones are held by *B*) are partitioned between *A* and *B*. In order to provide predictions without violating confidentiality, we propose privacy-preserving prediction computation protocol (P3CP) explained in the following. Note that, as in the P2TCP, *A* and *B* perform computations with fully available components and store such *fully available sub-aggregates* just after being informed about *a* and *q*.

Protocol II: Privacy-Preserving Prediction Computation Protocol

0. Active user *a* asks a prediction about *q* from *A*, which acts as a master party.
1. *A* zeroes all trust values below the threshold τ. Note that those users whose trust value with *a* satisfies a predefined threshold τ are selected as neighbors.
2. Using its own public key *KA* and self-blinding property of HE, *A* encrypts all available trust values of *a* using HE; and sends such encrypted values to *B*.
3. *B* multiplies the rated ones of v_{uqB} only using homomorphic property. It then accumulates the results for numerator and the corresponding trust values for denominator. It obtains $\xi_{KA}\left(\sum_{u \in S} (v_{uqB} - \overline{v_u}) \times [t_{a \to u}]_A\right)$ and $\xi_{KA}\left(\sum_{u \in S} [t_{a \to u}]_A\right)$, respectively, where $[t_{a \to u}]_A$ is trust value of *a* held by *A*.
4. Switching their roles, they perform *Step* 1-3 for v_{uqA} and trust values of *a* held by *B*. *A* obtains $\xi_{KB}\left(\sum_{u \in S} (v_{uqA} - \overline{v_u}) \times [t_{a \to u}]_B\right)$ and $\xi_{KB}\left(\sum_{u \in S} [t_{a \to u}]_B\right)$. After encrypting fully available sub-aggregates with *KB*, *A* adds corresponding parts of numerator and denominator and sends them to *B*.
5. *B* decrypts the obtained values in *Step* 4 and encrypts them with *KA*. Then, it also encrypts its fully available sub-aggregates similarly.
6. *B* adds correspondent obtained values in *Step* 3 and *Step* 5.
7. *B* returns $\xi_{KA}\left(\sum_{u \in S} (v_{uq} - \overline{v_u}) \times t_{a \to u}\right)$ and $\xi_{KA}\left(\sum_{u \in S} t_{a \to u}\right)$ to *A*.
8. *A* determines p_{aq} using Eq. (4) and returns it to active user *a*.

5. PRIVACY ANALYSIS

In our proposed scheme, there are no direct exchanges of information about the individual rating values and the rated items. Thus, it can be said that the principal privacy constraints are satisfied. However, there are protocols used to achieve the exchange of aggregate values in a private manner. They should be examined whether they conflict with the auxiliary privacy constraints or not. The proposed protocols' privacy protection is based on random filling with default votes and cryptographic tools. Since Paillier [45] justifies that his HE schemes are semantically secure and Naor and Pinkas [47] examine the security of their OT protocols, the proposed protocols are secure in their anticipated framework. However, in privacy perspective, it is still interesting to investigate disclosed intermediary values, aggregates, and default votes in addition to actual rating values. Considering such values, the proposed scheme is going to be analyzed in terms of inference probability rates and privacy enhancement.

In normalization, default votes hide the total number of ratings of each user has already rated and avoid directly sharing actual local mean of each user. In the P2TCP, *B* can guess a subset of *A*'s rated items. Let the size of this subset *f*, over random vectors, the probability of guessing such subset is $1/N$ in *Step* 3. Similarly, after switching their roles, *A* can also guess it with same probability. The value of *N* should be set to proper value depending on sensitivity of items and privacy requirements. Again, in the same protocol, *A* obtains individual aggregates of commonly rated items for the subset of cross-wise components in *Step* 8. Let *A* obtains *g* pieces of such aggregates. Then, *A* can infer the subset of the rated items with probability C_g^f . By switching their roles, *B* may also infer with the same possibility for the complementary cross-wise components.

In trust propagation, each party learns which trust value is null in the other party's trust sub-matrix and two sub-aggregate values for each of such values; one for numerator and the other for denominator part of Eq. (3). In order

to deduce trust values owned by the other party, how many values are included to compute such sub-aggregates should be known. However, it is unknown in the proposed scheme. After guessing such value, they are still conundrums that which trust values are included and what are such trust values. In prediction generation, *cooperator*, who is not master party, learns only sub-aggregates of final prediction value and the master party learns only final prediction value. For both parties, the same applies as in trust propagation because there are sub-aggregates in similar computational manner.

Filling with default votes and removing processed fully available components enhances privacy. Both operations decrease the rate of original rating components over totally contributed ones to generate aggregate values. Default votes also give denial of possession of the rated items in case of inferring the other's ratings. To compare privacy preservation with respect to type of default votes, POP algorithm can be considered to be the best over the others. Since the computation process is realized user by user and user-based aggregates are shared through the proposed protocols, each user rating vector is expected to have different default vote values if item means for default votes are used. However, in user mean usage, filled default votes are the same for each user and this may facilitate the inference manner of the other party. Also, row-variant POP preference is not suitable for applications, where local user means are sensitive to privacy because the actual mean is disclosed to each party. Overall mean's handicap about privacy is that it is the same for all local data. This fact is also advantage to the party intending to deduce some extra sensitive information from the other's data.

Conducted privacy analysis indicates that there is no conflict of both principal and auxiliary privacy constraints in the proposed approach. There are no direct or indirect leakage of the parties' individual rating values and the rated items. However, the inference possibilities are scrutinized over the shared intermediate values. One additional issue is related to trust updates. To enhance privacy and complicate inference possibilities, the parties prefer to re-fill their original data with defaults for each update phase because default vote values and filled unrated cells are changed and different input data are obtained.

6. SUPPLEMENTARY COST ANALYSIS

The proposed scheme brings about some overheads of computation, communication, and storage. In this section, the proposal is examined in terms of such extra costs. First of all, considering computational resources, this scheme can be contemplated for implementation as two phases: off-line and online. Preprocessing and computations of direct and propagated trust values can be computed off-line. However, prediction generation needs online interactions, and it should be considered for running online. Since off-line costs are not critical, P3CP must be evaluated in terms of computational efficiency. While the party j totally realizes $n/2$ encryptions, rf_j homomorphic multiplications, rf_j homomorphic additions, and two decryptions, the cooperator party performs additional two decryptions and two homomorphic additions, where rf_j stands for number of rating and filled ratings in a's vector held by the party j. Since τ can be determined previously, the parties can encrypt trust values after comparing them with τ and store the encrypted trust values off-line. However, this brings $n^2/4$ storage requirement for each party. To benchmark cryptographic operations, CRYPTO++ [49] can be referred. Therewithal there are some hardware-based and systematic solutions to overcome computational cumbersome due to homomorphic encryption operations [14, 50].

Secondly, the proposed method bids parties to communicate for trust computation and prediction generation. In the P2TCP, there must be at least two communications consisting of significant sizes of data exchange in bi-directional way while trust propagation requires two mutual communications; one for informing which of the held trusts are null and the other for sharing the sub-aggregates for propagated trusts. During the P2TCP, there must be at least three communications. Recall that if two parties collaborate on APD with off-line generated trust values, two online communications are needed to provide prediction services.

Thirdly, there are also storage overheads with respect to off-line model generation. During off-line phase, the parties temporarily need spaces to keep default votes, user mean values, and c_{au} values. However, trust values computed off-line require $n^2/4$ spaces from each party in order to utilize the constructed prediction model online. Note that, depending on data entry traffic and recommended product profiles, trust model must be updated in a particular period.

7. PREDICTION QUALITY ANALYSIS: EXPERIMENTS

Using MovieLens Public (MLP) data [51], we empirically evaluated our proposal in terms of accuracy and coverage. MLP is widely used data set by CF community. The results based on this data set can be generalized. This data set consists of 100,000 ratings collected from 943 users on 1,682 movies. While ratings are integers from

1 (*dislike*) to 5 (*like*), each user has rated at least 20 movies. We divided available ratings into two disjoint subsets. 80% and 20% of them were uniformly randomly selected for training and testing, respectively. While training subset was used as input data for specified CF process, test ratings were queried for prediction. To reach more dependable results, we performed each experiment 100 times and presented the overall averages. Returned prediction values were compared based on the accuracy metric, mean absolute error (MAE). It can be obtained through averaging the absolute values of difference between generated predictions and original ratings [41]. Another metric to evaluate the CF recommender system is coverage, which is the percentage of number of prediction-responded queries over total number of queries [41].

Hwang and Chen [10] evaluated experiments in which the scheme determines a's neighborhood selecting the best k similar users. However, rather than such determination, it is preferred to use threshold-based scheme in order to simplify prediction generation process. Herlocker et al. [48] empirically demonstrated that such process can be performed either of both methods. To determine the optimum value of the threshold τ, various experiments were conducted using MLP. According to the outcomes, it is concluded that 0.7 produces satisfactory accuracy and coverage values. Thus, τ was set at 0.7 in the following trials.

In the first experiment, how collaboration on APD affects accuracy and coverage of trust-based CF system was investigated. For this reason, an experiment was conducted comparing split and combined data without any privacy considerations. The number of users in input data was varied and MAE and coverage values were computed. The results in terms of accuracy and coverage are given in **Table 1** and **Table 2**, respectively. Both accuracy and coverage gains show similar manner with respect to increasing number of users. Such gains are initially higher; however, with joining more users into CF process, they decrease. APD generally contributes more significant to accuracy rather than coverage according to results in **Table 1** and **Table 2**. This experiment shows that APD contributes more to the prediction quality of CF system when amount of available rating profile is lower.

Table 1: Effects of APD on accuracy

	MAE			
Type	***n* = 125**	**250**	**500**	**943**
Split	0.8196	0.7935	0.7730	0.7631
Integrated	0.7803	0.7621	0.7498	0.7446
Gain (%)	4.80	3.96	3.00	2.42

Table 2: Effects of APD on coverage

	Coverage (%)			
Type	***n* = 125**	**250**	**500**	**943**
Split	91.51	95.91	98.01	98.97
Integrated	95.99	97.74	98.71	99.14
Gain (%)	4.90	1.91	0.71	0.17

After justifying how APD contributes to prediction accuracy and coverage of CF system, how inserting default votes affect accuracy was then analyzed. By setting n to 500, for different levels of filling, i.e., β_j and different types of default votes, accuracy changes were observed on combined data. The obtained results are given in **Figure 3**.

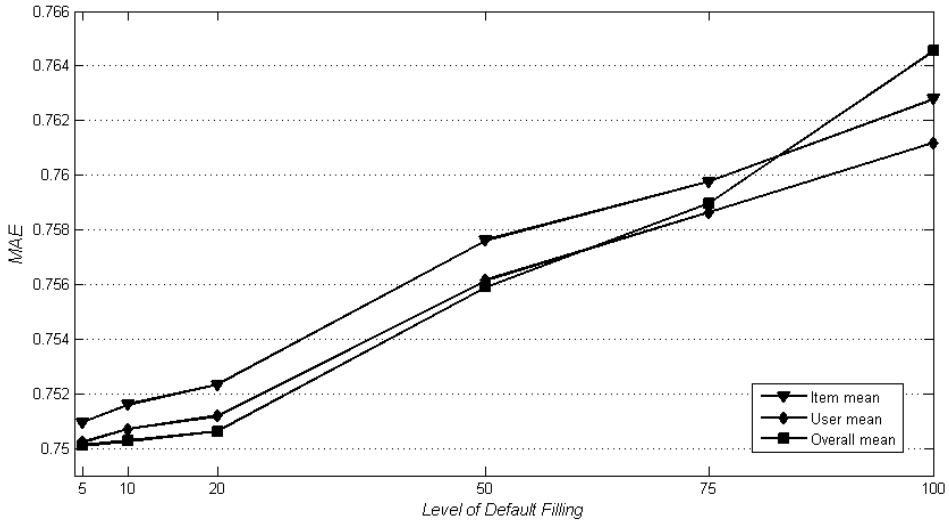

Figure 2: Accuracy vs. level of default filling

According to outcomes given in **Figure 3**, it can be said that accuracy is inversely proportional to filling level for all types of default votes. However, considering MAE value for split data for n = 500 in **Table 1**, which is 0.7730, accuracy for all types and levels in the proposed model is more accurate. To speak about specifically types of default votes, for β_j is 50, the best default vote type is overall mean. However, after such value, for overall mean-based default filling, accuracy significantly becomes worse. There is a parallel relation in terms of accuracy between filling with item and user mean default votes. Although the outcomes become very closer to each other, user mean shows better accuracy, as seen from **Figure 3**.

In the final experiment, the goal was to benchmark the accuracy values obtained by using the split data only and the combined data by the proposed method. For this purpose, trials were conducted for different level of filling with respect to varying number of users. Considering average number of the rated items per user is 106 in MLP, if β_j = 10, then E(θ_j)= 5 and E(|fc|) = 0.05 × 106 = 5.3, where E(x) is expected value of x and |fc| stands for the number of filled cells. Roughly speaking, it is expected that about five of the unrated cells would be filled with default votes. Similarly, for β_j = 20 and 50, E(|fc|) values are 10.6 and 26.5, respectively. Since such listed E(|fc|) values can be considered decent values providing balance between privacy and data originality, for β_j being 10, 20, and 50 with the best filling scheme, i.e., overall mean, trials were performed for such values and the results are displayed in **Figure 4**.

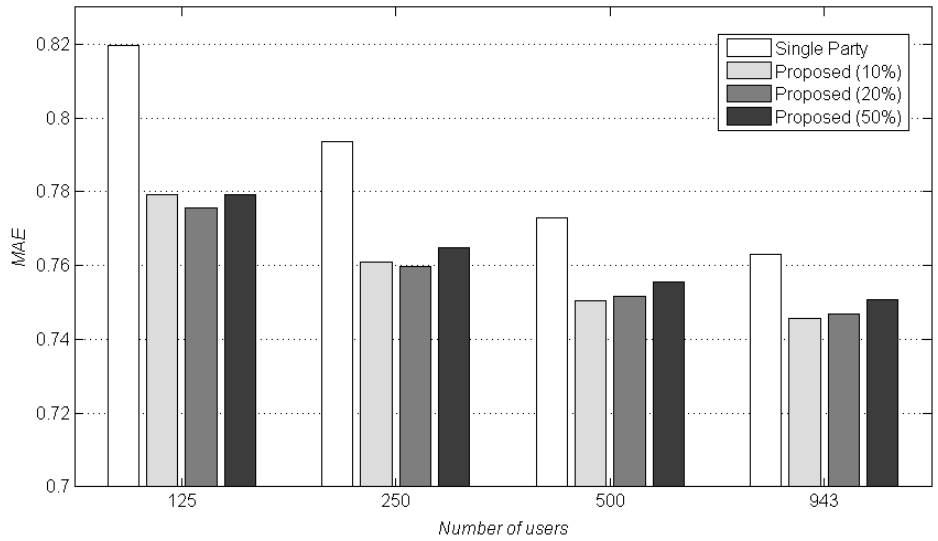

Figure 4: Single party vs. the proposed method

According to outcomes presented in **Figure 4**, it is obvious that for all n and focused β_j values, the proposed methods outperform the results on split data only. Especially for smaller number of users, the proposed scheme

provides more quality referrals due to the insufficient number of ratings in partitioned case. These outcomes show that the proposed scheme is preferable in order to overcome problems caused by split data.

8. CONCLUSIONS AND FUTURE WORK

In this study, we presented a novel solution in order to provide trust-based predictions on arbitrarily partitioned data between two parties while preserving their privacy. Our solution makes it possible for two parties to provide predictions using their joint data without divulging their sensitive data to each other. The proposed scheme gives control of some parameters to the collaborating parties. The solution was justified in terms of efficiency, privacy-preservation, and accuracy through theoretical and experimental analysis. The experimental analysis demonstrate that the solution produces satisfactory results in prediction quality especially in situations, where available data are insufficient. Our theoretical analysis shows that additional costs caused by privacy concerns are not that critical for overall performance. Our privacy analysis confirms that the proposed scheme is able to protect data holders' confidential data against each other. Our scheme can be used by those sites struggling with insufficient data for collaborative and want to provide trust-based recommendations to their customers.

It is still interesting topic to investigate trust-based collaborative filtering on arbitrarily partitioned data among multiple parties because there are some challenges due to extension from two-party to multi-party scenario. Moreover, our arbitrary data partitioning case assumes that the ratings distinctly exist in each party's databases. In actual case, there may be overlapping ratings. It is a proper research task to scrutinize how such overlapping ratings can occur and how to handle such cases. Also, such future studies should investigate performance changes with different amounts of overlapping data and effects of overlapping in terms of accuracy and privacy.

REFERENCES

1. E. Kannout, M. Grzegorowski, M. Grodzki, and H. S. Nguyen, "Clustering-based Frequent Pattern Mining Framework for Solving Cold-Start Problem in Recommender Systems," *IEEE Access*, vol. 11, pp. 1-21, 2024. https://doi.org/10.1109/ACCESS.2024.3355057
2. R. Baraglia, P. Dazzi, M. Mordacchini, and L. Ricci, "A peer-to-peer recommender system for self-emerging user communities based on gossip overlays," *Journal of Computer and System Sciences*, vol. 79, pp. 291-308, 2013. https://doi.org/10.1016/j.jcss.2012.05.011
3. Z. Yan, P. Zhang, and R. H. Deng, "TruBeRepec: A trust-behavior-based reputation and recommender system for mobile applications," *Personal and Ubiquitous Computing*, vol. 16, pp. 485-506, 2012. https://doi.org/10.1007/s00779-011-0420-2
4. H. Fang, G. Guo, and J. Zhang, "Multi-faceted trust and distrust prediction for recommender systems," *Decision Support Systems*, vol. 71, pp. 37-47, 2015. https://doi.org/10.1016/j.dss.2015.01.005
5. A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions," *Expert Systems with Applications*, vol. 240, 122442, 2024. https://doi.org/10.1016/j.eswa.2023.122442
6. S. Nusrat and J. Vassileva, "Recommending services in a trust-based decentralized user modeling system," *Lecture Notes in Computer Science*, vol. 7138, pp. 230-242, 2011. https://doi.org/10.1007/978-3-642-28509-7_22
7. L. Guo, C. Zhang, and Y. Fang, "A trust-based privacy-preserving friend recommendation scheme for online social networks," *IEEE Transactions on Dependable and Secure Computing*, vol.12, no. 4, pp. 413-427, 2015. https://doi.org/10.1109/TDSC.2014.2355824
8. P. Massa and B. Bhattacharjee, "Using trust in recommender systems: An experimental analysis trust management," *Lecture Notes in Computer Science*, vol. 2995, pp. 221-235, 2004. https://doi.org/10.1007/978-3-540-24747-0_17
9. P. Massa and P. Avesani, "Trust-aware recommender systems," in *Proceedings of the 2007 ACM Conference on Recommender Systems* (pp. 17-24), 2007. https://doi.org/10.1145/1297231.1297235
10. C.-S. Hwang and Y.-P. Chen, "Using trust in collaborative filtering recommendation," in *Proceedings of the 20th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* (pp. 1052-1060), 2007. https://doi.org/10.1007/978-3-540-73325-6_105

11. Y. Wang, L. Li, and G. Liu, "Social context-aware trust inference for trust enhancement in social network based recommendations on service providers," *World Wide Web*, vol. 18, pp. 159-184, 2015. https://doi.org/10.1007/s11280-013-0241-5
12. OECD, "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data," 2005.
13. A. Bilge, C. Kaleli, I. Yakut, I. Gunes, and H. Polat, A Survey of Privacy-Preserving Collaborative Filtering Schemes," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, pp. 1085-1108, 2013. https://doi.org/10.1142/S0218194013500320
14. I. San, N. At, I. Yakut, and H. Polat, "Efficient Paillier Cryptoprocessor for Privacy-Preserving Data Mining," *Security and Communication Networks*, vol. 9, pp. 1535-1546, 2016. https://doi.org/10.1002/sec.1442
15. A. Inan and Y. Saygin, "Privacy preserving spatio-temporal clustering on horizontally partitioned data," *Lecture Notes in Artificial Intelligence*, vol. 6202, pp. 187-198, 2010. https://doi.org/10.1007/978-3-642-16392-0_11
16. N. R. Nanavati and D. C. Jinwala, "A novel privacy-preserving scheme for collaborative frequent itemset mining across vertically partitioned data," *Security and Communication Networks*, vol. 8, pp. 4407-4420, 2015. https://doi.org/10.1002/sec.1377
17. K. Y. Yigzaw, A. Michalas and J. G. Bellika, "Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation," *BMC Medical Informatics and Decision Making*, vol. 17, pp. 1-19, 2017. https://doi.org/10.1186/s12911-016-0389-x
18. Y. Li, Z. L. Jiang, X. Wang, S. M. Yiu and P. Zhang, "Outsourcing privacy-preserving ID3 decision tree over horizontally partitioned data for multiple parties," *International Journal of High Performance Computing and Networking*, vol. 12, pp. 207-215, 2018. https://doi.org/10.1504/IJHPCN.2018.094370
19. B. Y. Yilmazel and C. Kaleli, "Robustness analysis of arbitrarily distributed data-based recommendation methods," *Expert Systems with Applications*, vol. 44, pp. 217-229, 2016. https://doi.org/10.1016/j.eswa.2015.09.012
20. S. Mehnaz and E. Bertino, "Privacy-preserving multi-party analytics over arbitrarily partitioned data," in *Proceedings of the 10th IEEE International Conference on Cloud Computing* (pp. 342-249), 2017. https://doi.org/10.1109/CLOUD.2017.51
21. W. Aslam, A. Hussain, K. Farhat, and I. Arif, "Underlying Factors Influencing Consumers' Trust and Loyalty in E-commerce," *Business Perspectives and Research*, vol. 8, pp. 186-204, 2020. https://doi.org/10.1177/2278533719887451
22. E. Aïmeur, 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. https://doi.org/10.1007/s10207-007-0049-3
23. H. Polat and W. Du, "Privacy-preserving collaborative filtering," *International Journal of Electronic Commerce*, vol. 9, pp. 9-35, 2005. https://doi.org/10.1080/10864415.2003.11044341
24. H. Polat and W. Du, "Privacy-preserving collaborative filtering on vertically partitioned data," *Lecture Notes in Computer Science*, vol. 3721, pp. 651-658, 2005. https://doi.org/10.1007/11564126_69
25. H. Polat and W. Du, "Privacy-preserving top-*N* recommendation on horizontally partitioned data," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 725-731), 2005.
26. I. Yakut and J. Vaidya, "Privacy-preserving item-based recommendations over partitioned data with overlaps," *International Journal of Business Information Systems*, vol. 25, pp. 336-351, 2017. https://doi.org/10.1504/IJBIS.2017.084449
27. I. Yakut and H. Polat, "Privacy-preserving SVD-based collaborative filtering on partitioned data," *International Journal of Information Technology and Decision Making*, vol. 9, pp. 473-502, 2010. https://doi.org/10.1142/S0219622010003919
28. C. L. A. Hsieh, Z. Zhan, D. Zeng, and W. Feiyue, "Preserving privacy in joining recommender systems," in *Proceedings of the International Conference on Information Security and Assurance* (pp. 561-566), 2008.
29. J. Vaidya, I. Yakut, and A. Basu, "Efficient Integrity Verification for Outsourced Collaborative Filtering," in *Proceedings of the 2014 IEEE International Conference on Data Mining* (pp. 560-569), 2014. https://doi.org/10.1109/ICDM.2014.145
30. Z. Zhan, I. C. Wang, C.-L. Hsieh, T.-S. Hsu, C.-J. Liau, and D.-W. Wang, "Towards efficient privacy-preserving collaborative recommender systems," in *Proceedings of the IEEE International Conference on Granular Computing* (pp. 778-783), 2008. https://doi.org/10.1109/GRC.2008.4664769
31. N. Dokoohaki, C. Kaleli, H. Polat, and M. Matskin, "Achieving optimal privacy in trust-aware social recommender systems," in *Proceedings of the 2nd International Conference on Social Informatics* (pp. 62-79), 2010. https://doi.org/10.1007/978-3-642-16567-2_5

32. C. Kaleli and H. Polat, "Privacy-preserving trust-based recommendations on vertically distributed data," in *Proceedings of the 5th IEEE International Conference on Semantic Computing* (pp. 376-379), 2011. https://doi.org/10.1109/ICSC.2011.43
33. P. K. Prasad and C. P. Rangan, "Privacy preserving BIRCH algorithm for clustering over arbitrarily partitioned databases," in *Proceedings of the 3rd International Conference on Advanced Data Mining and Applications* (pp. 146-157), 2007. https://doi.org/10.1007/978-3-540-73871-8_15
34. M. Upmanyu, A. Namboodiri, K. Srinathan, and C. Jawahar, "Efficient privacy preserving *k*-means clustering," *Lecture Notes in Computer Science*, vol. 6122, pp. 154-166, 2010. https://doi.org/10.1007/978-3-642-13601-6_17
35. S. Han and W. K. Ng, "Multi-party privacy-preserving decision trees for arbitrarily partitioned data," *International Journal of Intelligent Control and Systems*, vol. 12, pp. 351-358, 2007.
36. L. Li, L. Huang, and W. Yang, "Privacy-preserving outlier detection over arbitrarily partitioned data," in *Proceedings of the 3rd International Symposium on Information Engineering and Electronic Commerce* (pp. 103-103), 2011. https://doi.org/10.1115/1.859759.paper24
37. H. Yunhong, H. Guoping, F. Liang, and T. Jingyong, "Privacy-preserving SVM classification on arbitrarily partitioned data," in *Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing* (pp. 67-71), 2010. https://doi.org/10.1109/PIC.2010.5687397
38. A. Bansal, T. Chen, and S. Zhong, "Privacy preserving Back-propagation neural network learning over arbitrarily partitioned data," *Neural Computing & Applications*, vol. 20, pp. 143-150, 2010. https://doi.org/10.1007/s00521-010-0346-z
39. J. Yuan and S. Yu, "Privacy preserving Back-propagation learning made practical with cloud computing," *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 106, pp. 292-309, 2013. https://doi.org/10.1007/978-3-642-36883-7_18
40. J. Shi, C. Chen, and S. Zhong, "Privacy preserving growing neural gas over arbitrarily partitioned data," *Neurocomputing*, vol. 144, pp. 427-435, 2014. https://doi.org/10.1016/j.neucom.2014.04.033
41. I. Yakut and H. Polat, "Arbitrarily distributed data-based recommendations with privacy," *Data & Knowledge Engineering*, vol. 72, pp. 239-256, 2012. https://doi.org/10.1016/j.datak.2011.11.002
42. I. Yakut and H. Polat, "Estimating NBC-based recommendations on arbitrarily partitioned data with privacy," *Knowledge-Based Systems*, vol. 36, pp. 353-362, 2012. https://doi.org/10.1016/j.knosys.2012.07.015
43. I. Yakut and H. Polat, "Privacy-preserving hybrid collaborative filtering on cross distributed data," *Knowledge and Information Systems*, vol. 30, pp. 405-433, 2012. https://doi.org/10.1007/s10115-011-0395-3
44. A. Basu, J. Vaidya, and H. Kikuchi, "Efficient privacy-preserving collaborative filtering based on the weighted Slope One predictor," *Journal of Internet Services and Information Security*, vol. 1, pp. 26-46, 2011.
45. P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," *Lecture Notes in Computer Science*, vol. 1592, pp. 223-238, 1999. https://doi.org/10.1007/3-540-48910-X_16
46. T. B. Pedersen, Y. Saygin, and E. Savas, "Secret sharing vs. encryption-based techniques for privacy preserving data mining," in *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Disclosure Control* (pp. 17-19), 2007.
47. M. Naor and B. Pinkas, "Efficient oblivious transfer protocols," in *Proceedings of the Symposium on Discrete Algorithms* (pp. 448-457), 2001.
48. J. L. Herlocker, J. A. Konstan, A. Borchers, and J. T. Riedl, "An algorithmic framework for performing collaborative filtering," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 230-237), 1999. https://doi.org/10.1145/312624.312682
49. Crypto++. (20/02/2016). *5.6.0 Benchmarks*. http://www.cryptopp.com/benchmarks.html
50. C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning," in *2020 USENIX Annual Technical Conference* (pp. 493-506), 2020.
51. GroupLens. (20/02/2016). *MovieLens Data Sets*. http://www.grouplens.org/node/73

Yakut, Ibrahim: Dr. Yakut works at SIRF TRADE, Kayseri, Turkey. Before joining the company, Dr. Yakut was an Assistant Professor in Computer Engineering Department at Anadolu University from 2013 to 2016. He received his MSc. degree and PhD from Computer Engineering Department at Anadolu University in 2008 and 2012, respectively. He was a one-year visiting research fellow at MSIS Dept. of Rutgers University, Newark, USA in academic term 2013-2014. His research interests are artificial intelligence, data science, recommender systems and information security.

Polat, Huseyin: Dr. Polat is a consultant at SIRF TRADE, Kayseri, Turkey. Before joining the company, Dr. Polat was an Associate Professor in Computer Engineering Department at Anadolu University, Eskisehir, Turkey. He received his MSc. degree and PhD from Computer Science Department at Syracuse University in 2001 and 2006, respectively. His research interests are primarily collaborative filtering with privacy, private predictions on selected models, and privacy-preserving data mining in general.

Figure Captions

Figure 1: Arbitrarily partitioned rating profile data
Figure 2: Factors that affect e-commerce trust and loyalty [21]
Figure 3: Accuracy vs. level of default filling
Figure 4: Single party vs. the proposed method