CONTENTS

ABSTRACT

1. Reason for Choosing the Topic

Recommender Systems (RS) aims to address the problem of information overload while personalizing user experiences by providing accurate and tailored recommendations. However, these systems require users to share their information with a centralized recommendation server, raising concerns about user data privacy. Therefore, solutions to ensure privacy in RS are essential. Methods used to protect the privacy of RS can be categorized into two types [29]: statistical-based and cryptographic-based approaches. Statistical-based methods introduce noise, leading to a trade-off between privacy and accuracy. Cryptographic-based methods are more secure but consume significant computational resources. The quality of recommendations is a critical factor for the effectiveness of a recommender system. Hence, this dissertation focuses on developing a privacy-preserving data analysis solution based on Secure Multi-Party Computation (SMC).).

2. Main contributions

The new scientific contributions of the dissertation include:

- Development of two efficient secure computation protocols.
- Proposal of two privacy-preserving recommender system solutions utilizing the proposed protocols.

3. Thesis organization

The rest of this thesis is organized as follows:

CHAPTER 1: Overview of Privacy-Preserving Recommendation Systems and Secure Computation.

CHAPTER 2: Development of Efficient Secure Computation Protocols.

CHAPTER 3: Privacy-Preserving Solutions for Recommendation Systems Based on Secure Computation Protocols.

Chapter 1. OVERVIEW OF PRIVACY-PRESERVING RECOMMENDATION SYSTEMS AND SECURE COMPUTATION

1.1. Chapter Introduction

This chapter provides an overview of recommender systems, the problem of ensuring privacy for recommender systems, and identifies the issues that need to be addressed.

1.2. Recommendation Systems

1.3. Privacy Violation Risks in Recommender Systems

The risks of privacy violations can occur at all three stages of a recommender system (RS) [29]: the user modeling stage, the computation stage, and the recommendation generation stage.

Centralized storage of user data, whether in one form or another, raises concerns about data privacy. Therefore, this dissertation will focus on addressing the problem of recommender systems for distributed data models.

1.4. Privacy-Preserving Techniques for Recommender Systems

Privacy-preserving techniques include [5]: cryptographic-based techniques (SMC, HE), noise-based techniques (DP), tokenization, and masking techniques (anonymization, pseudonymization). In this section, the dissertation presents privacy-preserving techniques for recommender systems and compares their advantages and limitations.

For the RS problem, the quality of the recommender system (accuracy) is the most important factor, as it plays a decisive role in whether the system will be adopted for use. Therefore, the author has chosen cryptographic-based techniques to ensure privacy for the problem addressed in this dissertation, as these techniques can ensure accuracy and easily meet high-security constraints.

1.5. Secure Computation

First, the dissertation introduces several concepts related to secure multi-party computation (SMC) and secure multi-party computation protocols.

1.5.1. Entities in the SMC Protocol

In the SMC computation model, there are three types of entities [44]: (1) H*onest parties*, (2) C*orrupted parties* and (3) *External adversary*.

1.5.2. Definition of Security

1.5.2.1. Definition Models

There are two approaches to define security models in SMC: game-based and simulation-based. The simulation-based approach is more advantageous and complete [44].

1.5.2.2. Adversary Model

This dissertation focuses on two key parameters to define the adversary model for an SMC protocol [44]: the allowed adversarial behavior and the adversary's strategy for corruption.

1.5.2.2.1. Allowed adversarial behavior

This is the most important parameter for defining the adversary model for an SMC protocol [44]. There are two main types of adversaries: semi-honest adversaries and malicious adversaries. This dissertation will focus on the semi-honest adversary model.

1.5.2.2.2. Corruption strategy

There are three main models [44]: the static corruption model, the adaptive corruption model, and the proactive security model. This dissertation will focus on addressing problems within the adaptive corruption model, where the number of dishonest parties is limited depending on the specific problem at hand.

1.5.2.2.3. Other Parameters

In addition to the two main parameters defining the adversary model, there are several other parameters such as: communication channels, the computational power of the adversary, and the upper limit on the number of dishonest parties. In the adversary model considered in this dissertation, the following assumptions are made: the parties communicate with each other through a secure channel, the adversary is limited in computational power, meaning it operates within polynomial time, and there is an upper bound on the number of colluding parties, determined by the specific size of a subset of colluding parties.

1.5.2.3. Computational Indistinguishability

This section introduces an important concept used to define the security of SMC protocols: computational indistinguishability.

1.5.2.4. Standard Definition of Security for SMC Protocols

In this section, the dissertation presents the standard definition of security for SMC protocols in the semi-honest model (Definition 1.5 [24]) and an important theorem related to proving the privacy of the proposed protocol (Theorem 1.1 [24]).

Definition 1.5 [31]. *(Privacy in the semi-honest model) Let f be a multi-party computation function as defined in Definition 1.1.*

• If f is a deterministic function: the protocol π securely computes function f with t corupted participants if $\forall I \subseteq \{1,2,\dots,n\}$ such that $|I| = t$, there exists a polynomial-time probabilistic algorithm M such that:

$$\{M(I, \bar{x}_I,\ f_I(\bar{x}))\}_{\bar{x}\in(\{0,1\}^*)^n} \overset{c}{\equiv} \{VIEW^{\pi}_{A,I}(\bar{x})\}_{\bar{x}\in(\{0,1\}^*)^n} \quad (1.1)$$

• In the general case: the protocol π securely computes function f with t corupted participants if $\forall I \subseteq \{1,2,\dots,n\}$ such that $|I| = t$, there exists a polynomial-time probabilistic algorithm M such that:

$$\{M(I, \bar{x}_I,\ f_I(\bar{x}), f(\bar{x}))\}_{\bar{x}\in(\{0,1\}^*)^n} \overset{c}{\equiv} \{VIEW^{\pi}_{A,I}(\bar{x}), OUTPUT^{\pi}(\bar{x})\}_{\bar{x}\in(\{0,1\}^*)^n} \quad (1.2)$$

where:

• $VIEW^{\pi}_{A,I}(\bar{x})$: is the views of t corrupted parties and all messages (transferred among the honest parties) that the adversary A eavesdrops during the execution protocol π on the input $\bar{x} = (x_1, x_2, \dots, x_n)$.

• $\bar{x}_I$: is input of all the parties involved in the protocol π that are corrupted by the adversary.

• $f_I(\bar{x})$: is output sequence of all the parties involved in the protocol π that are corrupted by the adversary.

• $OUTPUT^{\pi}(\bar{x})$: is the output sequence of all parties involving the protocol π. In the first case, $OUTPUT^{\pi}(\bar{x}) \equiv f(\bar{x})$.

• $\overset{c}{\equiv}$ is computational indistinguishability ((computational model).

1.5.2.5. Some Cryptographic Primitives

1.6. Limitations of Existing PPRS Based on SMC

In this section, the dissertation conducts an analysis and evaluation of existing PPRS solutions based on cryptography, assessing their accuracy, privacy, and performance.

1.7. Issues to Be Researched in the Thesis

1.7.1. Issues to Be Researched in PPRS with Fully Distributed Setting

1.7.1.1. Fully Distributed Setting

1.7.1.2. Research Issues to Be Addressed

Based on these evaluations, the dissertation focuses on improving the performance of PPRS for a full distributed data model by enhancing the efficiency of the ranking aggregation phase, similarity between item pairs, and the generation of recommendations for users. In the system, user ratings for products are used to evaluate the degree of similarity between products, for generating recommendations in two ways: content-based filtering (CBF) and collaborative filtering (CF), using formulas (1.7.11) and (1.7.12), respectively.

$$P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j}.S(i_j, i_k)}{\sum_{j=1}^{m} S(i_j, i_k)} \quad (1.7.11)$$

$$P_{i,k} = \frac{R_k.\sum_{j=1}^{m}.S(i_k, i_j) + \sum_{j=1}^{m}(r_{i,j} - R_j).S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)} \quad (1.7.12)$$

$$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} e_{i,j}} \quad (1.7.13)$$

$$S(i_j, i_k) = \frac{\sum_{i=1}^{n} r_{i,j}.r_{i,k}}{\sqrt{\sum_{i=1}^{n} {r_{i,j}}^2} \cdot \sqrt{\sum_{i=1}^{n} {r_{i,k}}^2}} \quad (1.7.14)$$

1.7.2. Issues to Be Researched in PPRS with 2PFD

1.7.2.1. 2-Part Fully Distributed Setting

1.7.2.2. Research Issues to Be Addressed

Another issue highlighted through the evaluation of existing solutions is the lack of any PPRS solutions for the 2PFD. Therefore, the dissertation will investigate and propose a privacy-preserving recommender system solution for the 2PFD using SMC.

Conclusion of the Chapter

Chapter 2. DEVELOPMENT OF EFFECTIVE SECURE COMPUTATION PROTOCOLS

2.1. Chapter Introduction

This chapter presents the protocols developed for computing secure average values and similarity metrics in the fully distributed data model and the 2PFD (Two-Party Fully Distributed) model. These proposals are associated with Works [CT1, CT3, CT4, CT5, CT6]. The proposed protocols focus on ensuring privacy under the semi-honest adversarial model, where parties are assumed to follow the protocol but may collude with each other to infer sensitive information about other participants. All users are assumed to have an authenticated private communication channel with the computation center. The cryptographic parameters used throughout this dissertation are assumed to satisfy hardness requirements, such that the discrete logarithm problem over finite fields and elliptic curves remains computationally intractable.

2.2. Secure multi-average and similarity computation protocol

2.2.1. The idea

The problem model consists of n participants $U = \{U_1, U_2, \ldots, U_n\}$, where each U_i holds m non-negative integers, either small or medium-sized private values $r_{i,j}$ $(1 \leq j \leq m)$, if $r_{i,j} \neq 0$ then $f_{i,j} = 1$, otherwise $f_{i,j} = 0$. A computing center (CC) needs to calculate the average of the non-zero private values $r_i \neq 0$ held by these participants using formula (1.7.13) without revealing $r_{i,j}$, $e_{i,j}$. Additionally, it computes the similarity values of item pairs based on cosine similarity using formula (1.7.14) without disclosing $r_{i,j}, r_{i,k}$.

2.2.1.1. Secure Average Computation

2.2.1.2. Secure Similarity Computation

Sections 2.2.1.1 and 2.2.1.2 present a comprehensive analysis of the most popular SMAC and SMSC protocols related to the dissertation. Based on this analysis, it is observed that effective multi-average and similarity protocols can be proposed using the following two main steps:

Step 1: Redesign the Protocols in [7] by:

- Initialization Phase: Require each U_i to use $n_k = \left\lceil \frac{1}{2} + \sqrt{m(m+5) + \frac{1}{4}} \right\rceil$ key pairs, and the CC computes only n_k shared public keys for encrypting the users' private values instead of $\frac{m(m+5)}{2} + 1$ key pairs.

- Phase 1: Require each U_i to encrypt their private values using the common public key and to decrypt components using their corresponding private keys.

- Phase 2: The CC calculates the aggregated ciphertexts for the averages and similarity values and performs discrete logarithm calculations for all these values just once.

Step 2: Convert the redesigned protocol from Step 1 into a variant that utilizes cryptographic systems based on elliptic curves.

This proposal has been published in the Publications [CT3, CT4, CT6].

2.2.2. Proposed Protocol

The dissertation redesigns the protocol from works [7] based on the improved idea 1 proposed in Section 2.2.1. The redesigned protocol simultaneously computes multiple secure averages and similarity values, as described in Protocol 2.3. This protocol employs the ElGamal cryptosystem with parameters as described in Section 1.5.2.5.1.

2.2.3. Correctness Analysis

The dissertation has proven the correctness of the proposed protocol.

2.2.4. Privacy Analysis

The thesis demonstrates that the improved protocol offers higher security compared to the original protocol in [7] and is equivalent to the protocol by Verma et al. [75], as stated in the following proposition:

Proposition 2.1 *The secure multi-average and similarity computation protocol ensures the privacy of each user under a semi-honest model in the absence of collusion among parties. Even in cases*

where up to n−2 users collude, the protocol still preserves the privacy of the honest users.

Input:
n: The number of users.
U_i: the i^{th} user, where $i \in [1,n]$.
m: The number of items.
RM: The rating matrix for users on items, where: $r_{i,j}$: Non-negative small or medium-sized integers are secret values held by each U_i $(0 \le j < m)$.
max: The maximum value among the secret values held by the users U_i.

Output: At the computing center"(CC):

$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$, $S_{j,k} = \frac{\sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$, với $0 \le j < k < m$.

*** Initialization Phase:**
- U_i does:
1. for $(0 \le j < n_k)$
2. $ksu_{i,j} \in_R \mathbb{Z}_q^*; KPU_{i,j} = g^{ksu_{i,j}}$;
3. Sends $\{KPU_{i,j}\}_{j=0}^{n_k-1}$ to CC.
- CC does:
4. for $(0 \le j < n_k)$
5. $KP_j = \prod_{i=1}^{n} KPU_{i,j}$;
6. Sends $\{KP_j\}_{j=0}^{n_k-1}$ to $\{U_i\}_{i=1}^{n}$;

*** Phase 1: U_i does**
7. for $(0 \le j < m)$
8. $a_{i,j} = r_{i,j}; f_{i,j} = 0$;
9. if $(r_{i,j} \ne 0)$ $f_{i,j} = 1$;
10. for $(m \le j < 2m)$
11. $a_{i,j} = f_{i,j-m}$;
12. for $(2m \le j < 3m)$
13. $a_{i,j} = r_{i,j-2m} \cdot r_{i,j-2m}$;
14. for $(0 \le t < m-1)$
15. for $(t+1 \le k < m)$
16. $a_{i,j} = r_{i,t} \cdot r_{i,k}$;
17. j++;
18. j = 0;
19. for $(0 \le t < n_k - 1)$
20. for $(t+1 \le k < n_k)$
21. $AU_{i,j} = g^{a_{i,j}} \cdot KP_k^{-ksu_{i,t}} \cdot KP_t^{ksu_{i,k}}$;
22. j++;
23. if $(j == n_s - 1)$ break;
24. if $(j == n_s - 1)$ break;
25. Gửi $\{AU_{i,j}\}_{j=0}^{n_s-1}$ cho CC.

*** Phase 2: CC does**
26. for $(0 \le j < n_s)$
27. $A_j = \prod_{i=1}^{n} AU_{i,j}$;
28. $sa = \mathrm{Dlog}(p, g, \max^2, A)$;
/*Using the brute-force algorithm once time with $A = \{A_j, j \in [1,n_s]\}$, $sa = \{sa_j, j \in [1,n_s]\}$*/.
29. for $(0 \le j < m)$
30. $R_j = \frac{sa_j}{sa_{j+m}}$;
31. $l = 0$;
32. for $(0 \le j < m-1)$
33. for $(j+1 \le k < m)$
34. $S_{j,k} = \frac{sa_{3m+l}}{\sqrt{sa_{2m+j}} \cdot \sqrt{sa_{2m+k}}}$;
35. l++;

Protocol 2.3. *Secure multi-average and similarity computation protocol*

2.2.5. Efficiency Analysis

The dissertation evaluates and compares the communication and computational costs of the original protocols [7], the protocol by Verma et al. [75], and the proposed protocol. The results show that the proposed protocol achieves lower communication and computational costs than the other protocols.

2.2.5.1. Communication Costs

2.2.5.2. Computational Costs

2.2.6. Conversion of the Proposed Protocol Using Elliptic Curve Cryptography

To further enhance the efficiency of the proposed protocol, the dissertation transforms it into a variant that employs elliptic curve cryptography (ECC) based on the improved idea 2 discussed in Section 2.2.1. The dissertation provides a detailed comparison of computational costs between the proposed protocol before and after adopting ECC. The results demonstrate that converting the proposed protocol to use the ElGamal cryptosystem on elliptic curves significantly reduces both communication and computational costs, leading to notable performance improvements.

2.2.7. Experimental Evaluation

The dissertation implements the proposed protocol, the original protocol considered the most efficient [7], and the protocol with relatively high efficiency but stronger security [75] to compare execution times. The experimental results show that the proposed protocol achieves significantly lower execution times compared to the other protocols, underscoring its superior performance in practical scenarios..

2.2.8. Illustrative Example

2.3. Secure Multi-Frequency Computation in the 2PFD

2.3.1. The idea

he problem model consists of two groups of users $U = \{U_1, U_2, \ldots, U_{n_1}\}$ and $V = \{V_1, V_2, \ldots, V_{n_1}\}$, where $n = 2n_1$, each user U_i, V_i holds their own binary values: $\{u_{i,0}, u_{i,1}, \ldots, u_{i,m_u-1}\}$, $\{v_{i,0}, v_{i,1}, \ldots, v_{i,m_v-1}\} \in \{0,1\}$. Each pair of users (U_i, V_i) holds a record where user U_i knows the values of a subset of relevant attributes, and user V_i knows the values of the remaining attributes. The computing center (CC) needs to compute the set of $m_u.m_v$ frequency values $SU = \{su_j = \sum_{i=1}^{n} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}\}_{j\in[0,m_u.m_v-1)}$.

The thesis evaluates existing multi-member secure frequency calculation protocols in the 2PFD model and suggests a new privacy-preserving protocol for multi-frequency calculations by improving the

performance of the original protocol [27]. The proposed protocol follows these steps:

Step 1: Redesigning the Steps in the Original Protocol [27]:

- Initialization Phase: Each U_i, V_i are required to use $n_k + 1$ key pairs, where ($n_k = \lceil \sqrt{s_k} \rceil$), and the CC computes s_k common public keys ($s_k = \left\lfloor \frac{1}{2} + \sqrt{2m + \frac{1}{4}} \right\rfloor$) for encrypting users' private values. This is a reduction compared to the original protocol, which used $2m + 1$ key pairs.

- Phase 1: Each U_i encrypts their private values using their public key and sends the ciphertexts to the CC.

- Phase 2: Each V_i computes $Q_{11}^{(i,j)} = \frac{Q_1^{(i,j)}}{Q_2^{(i,j)}}$, $Q_{21}^{(i,j)} = Q_3^{(i,j)}$, and sends these values to the CC, rather than calculating and sending all three values $Q_1^{(i,j)}$, $Q_2^{(i,j)}$, $Q_3^{(i,j)}$ as in the original protocol.

- Phase 3: Each U_i computes $\frac{K_1{}^{(i,j)}}{K_2{}^{(i,j)}}$ and sends this single value to the CC instead of separately sending $K_1{}^{(i,j)}$, $K_2{}^{(i,j)}$ as in the original protocol.

- Phase 4: CC only performs the decryption of the aggregated ciphertexts of the frequency values and computes the discrete logarithm values just once, instead of decrypting the ciphertext and calculating discrete logarithms multiple times as in the original protocol.

Bước 2: Conversion to an Elliptic Curve Cryptography Variant.

This proposal is related to the Publications [CT1, CT5].

2.3.2. Proposed Protocol

The thesis redesigns the protocol from [27] based on the improved idea 1 proposed in Section 2.3.1, and the enhanced secure multi-frequency calculation protocol for multiple members is described in Protocol 2.6. The protocol uses the ElGamal cryptosystem with parameters as described in Section 1.5.2.5.1.

Input:

n: The number of users in the system, where each data domain U_i, V_i has n_1 users, $i \in [1, n_1], n = 2n_1$.

$u_{i,j}$: The secret binary values held by U_i $(0 \le j < m_u)$.

$v_{i,j}$: The secret binary values held by V_i $(0 \le j < m_v)$, $m_u \le m_v$.

Output (for CC):

$SU = \{su_j = \sum_{i=1}^{n_1} u_{i,j\%m_v} . v_{i,\lfloor \frac{j}{m_v} \rfloor}\}_{j \in [0, m_u.m_v - 1)}$

*** Initialization Phase:**

- U_i does:

1. $x_i \in_R \mathbb{Z}_q^*; X_i = g^{x_i}$;
2. for $(0 \le j < n_{k1})$
3. $ksu_{i,j} \in_R \mathbb{Z}_q^*; KPU_{i,j} = g^{ksu_{i,j}}$;
4. Sends $\{KPU_{i,j}\}_{0 \le j < n_{k1}}, X_i$ to CC;

- V_i does:

5. $y_i \in_R \mathbb{Z}_q^*; Y_i = g^{y_i}$;
6. for $(0 \le j < n_{k1})$
7. $ksv_{i,j} \in_R \mathbb{Z}_q^*; KPV_{i,j} = g^{ksv_{i,j}}$;
8. Sends $\{KPV_{i,j}\}_{0 \le j < n_{k1}}$ to CC;

- CC does:

9. j=0;
10. for $(0 \le t < n_{k1})$
11. for $(0 \le k < n_{k1})$
12. $KP_j = \prod_{i=1}^{n_1} (KPU_{i,t}.KPV_{i,k})$; j++;
13. if $(j == n_k)$ break;
14. if $(j == n_k)$ break;
15. Sends $\{KP_j\}_{0 \le j < n_k}$ to $\{U_i, V_i\}_{1 \le i \le n}$;

*** Phase 1: U_i does**

16. for $(0 \le j < m_u)$
17. $c_{i,j}^{(1)} \in_R \mathbb{Z}_q^*$;
18. $C_1^{(i,j)} = g^{u_{i,j}}.X_i^{c_{i,j}}; C_2^{(i,j)} = g^{c_{i,j}}$;
19. Send $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}$ to CC;

*** Phase 2: V_i does**

20. Receives $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}, X_i$ from CC;
29. j=0;
21. for $(0 \le t < n_k - 1)$
22. for $(t + 1 \le k < n_k)$
23. $c_{i,j}^{(2)} \in_R \mathbb{Z}_q^*$;
24. $Q_1^{(i,j)} = (C_1^{(i,\lfloor j/m_v \rfloor)})^{v_{i,j\%m_v}}.$ $KP_t^{ksv_{i,k\%n_{k1}}}.(C_2^{(i,\lfloor j/m_v \rfloor)})^{-c_{i,j}^{(2)}.y_i}.KP_k^{-ksv_{i,t\%n_{k1}}}$
25. $Q_2^{(i,j)} = Y_i^{c_{i,j}^{(2)}}.X_i^{-v_{i,j\%m_v}}$; j++;
26. if $(j == m_u.m_v - 1)$ break;
27. if $(j == m_u.m_v - 1)$ break;
28. Sends $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u.m_v}$ to CC;

*** Phase 3: U_i does:**

29. Receives $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u.m_v}$ from CC;
30. j=0;
31. for$(0 \le t < n_k - 1)$
32. for $(t + 1 \le k < n_k)$
33. $A_{i,j} = Q_1^{(i,j)}.(Q_2^{(i,j)})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}.$ $KP_k^{-ksu_{i,\lfloor t/n_{k1} \rfloor}}.KP_t^{ksu_{i,\lfloor k/n_{k1} \rfloor}}$,
34. j++;
35. if $(j == m_u.m_v - 1)$ break;
36. if $(j == m_u.m_v - 1)$ break;
37. Sends $\{A_{i,j}\}_{0 \le j < m_u.m_v}$ to CC;

*** Phase 4: CC does:**

38. for $(0 \le j < m_u.m_v)$
39. $A_j = \prod_{i=1}^{n_1} A_{i,j}$;
40. $SU = Dlog(p, g, n_1, A)$;

//Using the brute-force algorithm once only

Protocol 2.6. *.Secure Multi-Frequency Computation Protocol in the 2PFD*

$\sum_{i=1}^{n} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}$ are not too large, so computing the discrete logarithm is not difficult.

2.3.3. Correctness Analysis

The thesis has proven the correctness of the proposed protocol.

2.3.4. Privacy Analysis

This section proves that the proposed protocol achieves the same level of security as the original protocol [27], as stated in the following proposition:

Proposition 2.4. *The protocol for computing multiple frequency values with privacy guarantees in the 2PFD preserves the privacy of each honest user in the semi-honest model. In the case of collusion among participants, the protocol ensures the privacy of honest users against the computation center and up to $n-2$ dishonest users. In the scenario where there are only two honest users, this remains accurate as long as the two honest users do not possess attribute values from the same record.*

2.3.5. Efficiency Analysis

In this section, the thesis analyzes and compares the efficiency of the proposed protocol with the original protocol [27] and the recent protocol [50] in terms of communication costs and computation costs. The results show that the proposed protocol is more efficient than the others.

2.3.5.1. Communication Costs

2.3.5.2. Computational Costs

2.3.6. Conversion of the Proposed Protocol Using Elliptic Curve Cryptography

To further enhance the performance of the proposed protocol, the thesis transforms the redesigned protocols into a variant utilizing elliptic curve cryptography based on the improvement idea number 2 presented in Section 2.3.1. The thesis also provides a detailed comparison of computation costs between the proposed protocol before and after adopting elliptic curve cryptography. The comparison results show that transitioning the proposed protocol to use the Elliptic Curve ElGamal cryptosystem significantly improves both communication and computation costs.

2.3.7. Experimental Evaluation

The thesis implements the proposed protocol, the original protocol [27], and protocol [50] to compare execution times and observes that

the proposed protocol has a lower execution time compared to the other protocols.

2.3.8. Illustrative Example

Conclusion of the Chapter

In this chapter, the dissertation has developed two efficient secure computation protocols: a protocol for simultaneously computing multiple secure average and similarity values, and a protocol for simultaneously computing multiple secure frequency values in the 2PFD.

It has also been proven that the proposed protocols ensure both the correctness of the output and the security under the semi-honest adversarial model. Theoretical and experimental evaluations demonstrate the high efficiency of these protocols compared to the secure average and similarity computation protocols presented in [7], and the frequency computation protocol in the 2PFD model presented in [27], respectively.

Therefore, the proposed protocols are applicable to real-world problems. Specifically, the proposed secure average and similarity computation protocol has been applied to privacy-preserving user recommendation systems, as published in [CT3, CT4, CT5]. The secure frequency computation protocol in the 2PFD model [CT1] has been applied to privacy-preserving decision tree training, as published in [CT2].

The content of this chapter is related to the works [CT1, CT3, CT4, CT5, CT6] listed in the List of Scientific Publications Used in the Dissertation.

Chapter 3. PRIVACY-PRESERVING SOLUTIONS FOR RECOMMENDATION SYSTEMS BASED ON SECURE COMPUTATION PROTOCOLS

3.1. Chapter Introduction

Building on the development of secure multi-party computation protocols for calculating averages, similarities, and frequency values in the 2PFD discussed in the previous chapter, these proposed protocols are applied in the first phase of the Privacy-Preserving Recommender System solution. This phase involves calculating the average ratings of items and the similarity between item pairs. These proposals are related to works [CT3, CT4, CT6].

3.2. Privacy-Preserving Recommendation Problem

This section analyzes existing PPRS solutions that address the problem under consideration in the thesis and identifies the general concept of the proposed solutions. These solutions are divided into two phases, as illustrated in Figure 3.1.

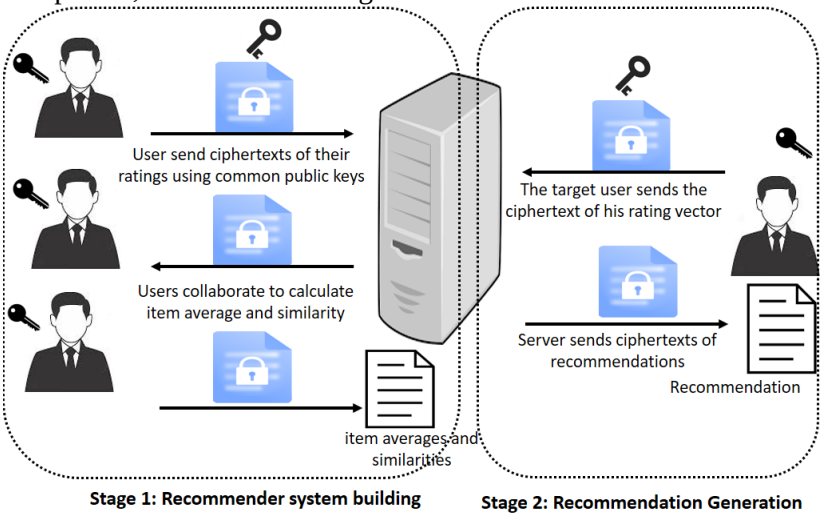

Figure 3.1. *PPRS solution [7]*

3.3. Privacy-Preserving Solution for Recommendation System in the Fully Distributed setting

3.3.1. Problem Definition

This section describes the RS problem in the fully distributed setting.

3.3.2. Privacy-Preserving Recommendation System Solution

In this new solution, the thesis redesigns both phases of the privacy-preserving recommendation system solution from [7] to improve its efficiency. The proposed solution is as follows:

- Stage 1: To compute the average ratings of items and the similarity between item pairs, the thesis uses the privacy-preserving multi-value average and similarity computation protocol proposed in Section 2.2.6 (referred to as the SMASC protocol).

- Stage 2: To generate privacy-preserving recommendations for target users, the dissertation proposes two methods: recommendation generation based on CBF (Content-Based Filtering) and recommendation generation based on CF (Collaborative Filtering) with privacy preservation. However, the proposed solution replaces the original ElGamal cryptosystem with an elliptic curve cryptosystem in order to maintain consistency with Stage 1 and improve computational efficiency.

The solution is detailed in Solution 3.1.

3.3.3. Illustrative Example

3.3.4. Correctness Analysis

This section of the thesis proves the correctness of the proposed solution.

3.3.5. Privacy Analysis

The thesis demonstrates that the proposed solution provides the same level of security as the solution of Verma et al. [75] and a higher level of security compared to the original solution [7].

3.3.6. Efficiency Analysis

The thesis compares the communication costs, computation costs, and execution time between the proposed solution, the original solution [7], and the solution by Verma et al. [75]. The results show that the proposed solution is more efficient than the other solutions.

3.3.6.1. Communication Costs

3.3.6.2. Computational Costs

3.3.6.3. Experimental Evaluation

Solution 3.1. *PPRS in the Fully Distributed setting*

Input:

n: The number of users.

m: The number of items.

RM: The rating matrix for users on items, where $r_{i,j}$ is the private rating value of user i for j.

max: The highest rating value in the rating scale.

k: The index of the item requested by user U_i for generating a recommendation.

Output: (for U_i):

$P_{i,k}$: The predicted ranking score for item k of U_i, where $k = \{0, 2, \ldots, m-1\}$ is the order of the item that the target user has requested for recommendation.

If the recommendation is generated based on content – CBF: $P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j} S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$

If the recommendation is generated based on collaborative filtering – CF: $P_{i,k} = \frac{R_k \sum_{j=1}^{m} S(i_k, i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$ R_k, R_j are computed using formula (1.7.13) và $S(i_k, i_j)$ is computed using formula (1.7.14).

Stage 1: Calculating the Average Ratings of Items and the item-item Similarities

1. $\{R_j, S(i_j, i_k)\}_{0 \le j,k < m; j < k} =$ SMASC$(n, m, RM, n.\max^2)$;

/* Execute protocol SMASC

2. **for** $(0 \le j < m-1)$
3. **for** $(j+1 \le k < m)$
4. $S(i_k, i_j) = S(i_j, i_k)$;

Stage 2: Generating Recommendations for the Target User U_i

*** Phase 1. U_i does**

5. **for** $(0 \le j < m)$
6. $c_j^{(1)} \in_R \mathbb{Z}_d^*$;
7. $E(r_{i,j} \cdot G) = (C_j^{(1)} = r_{i,j} \cdot G + c_j^{(1)} \cdot KPU_{1,1}, C_j^{(2)} = c_j^{(1)} \cdot G)$;

/*E is the encryption algorithm of the Elliptic Curve Cryptosystem (ECC).*/

8. Sends $\{E(r_{i,j} \cdot G)\}_{0 \le j < m}$ to Rs;

*** Phase 2. Rs does**

9. **for** $(0 \le j < k < m)$
10. $c_k^{(2)} \in_R \mathbb{Z}_d^*$;
11. $E(\sum_{j=1, j \ne k}^{m} S(i_k, i_j) \cdot G) = (F_{3,k} = \sum_{j=1}^{m} S(i_k, i_j) \cdot G + c_k^{(2)} \cdot KPU_{1,1}, F_{4,k} = c_k^{(2)} \cdot G)$;
12. **if** (Recommendation generation= CBF)
13. $(F_{1,k}, F_{2,k}) = \sum_{j=1, j \ne k}^{m} S(i_k, i_j) \cdot E(r_{i,j} \cdot G)$;
14. **else**
15. $c_k^{(4)} \in_R \mathbb{Z}_d^*; c_k^{(5)} \in_R \mathbb{Z}_d^*$;
16. $E(R_k \sum_{j=1, j \ne k}^{m} S(i_k, i_j) \cdot G) = (R_k \sum_{j=1, j \ne k}^{m} S(i_k, i_j) \cdot G + c_k^{(4)} \cdot KPU_{1,1}, c_k^{(4)} \cdot G)$;
17. $E(R_j \cdot G) = (R_j \cdot G + c_k^{(5)} \cdot X_i, c_k^{(5)} \cdot G)$;
18. $(F_{1,k}, F_{2,k}) = E(R_k \sum_{j=1, j \ne k}^{m} S(i_k, i_j) \cdot G) + \sum_{j=1, j \ne k}^{m} (E(r_{i,j} \cdot G) - E(R_j \cdot G)) \cdot S(i_k, i_j)$;
19. Sends to U_i: $\{F_{1,k}, F_{2,k}, F_{3,k}, F_{4,k}\}_{0 \le k < m}$;

*** Phase 3. U_i does**

20. **for** $(0 \le k < m)$
21. $C_k^{(3)} = F_{1,k} - ksu_{1,1} \cdot F_{2,k}$;
22. $C_{k+m}^{(3)} = F_{3,k} - ksu_{1,1} \cdot F_{4,k}$;
23. $d^{(3)} = Dlog_{EC}(\mathbb{E}, G, \max^2, C^{(3)})$;
24. **for** $(0 \le k < m)$
25. $P_{i,k} = \frac{d_k^{(3)}}{d_{k+m}^{(3)}}$;

3.4. Privacy-Preserving Solution for Recommendation System in the 2PFD

3.4.1. Problem Definition

Suppose we need to build a recommendation system with m items, where the rating data for these items is owned by each pair of users (u_i, v_i) from two different data domains (data from two different organizations), and there are n_1, $n = 2n_1$ pairs of users involved in the ratings. Each u_i will provide feedback on a subset of m_1 items $(m_1 < m)$. Each v_i will provide feedback on the remaining subset of $m - m_1$ items.

3.4.2. Privacy-Preserving Recommendation System Solution

Where: $n_{k1} = \left\lceil \frac{1}{2} + \sqrt{m_1(m_1+5) + \frac{1}{4}} \right\rceil$, $n_{k2} = \left\lceil \frac{1}{2} + \sqrt{(m-m_1)(m-m_1+5) + \frac{1}{4}} \right\rceil$, $n_{k3} = \left\lceil \frac{1}{2} + \sqrt{2m_1(m-m_1) + \frac{1}{4}} \right\rceil$

3.4.3. Illustrative Example

3.4.4. Correctness Analysis

This section of the thesis proves the accuracy of the proposed solution.

3.4.5. Privacy Analysis

The proposed solution ensures the privacy of each user in the semi-honest model. Furthermore, it protects the privacy of each honest user against collusion between the server and up to $n - 2$ corrupted users, provided that the two honest users do not share attribute values from the same record.

Solution 3.2. *PPRS in the 2PFD*

Input:
n: The number of users in the system, where each data domain has n_1 users ($n = 2n_1$).
m: The number of items in both data domains.
m_1: The number of items in the first data domain (U_i).
RM_1: The rating matrix for users on items in the first data domain, where $r_{i,j}$ is the private rating value of user U_i for item j, $j \in [0, m_1)$.
RM_2: The rating matrix for users on items in the second data domain (V_i), where $r_{i,j}$ is the private rating value of user V_i for item j, $j \in [0, m - m_1)$.
max: The highest rating value in the rating scale.
Output (for U_i):
$P_{i,k}$: The predicted ranking score for item k of U_i, $k \in [0, m)$ is the order of the item that the target user has requested for recommendation.

If the recommendation is generated based on content – CBF: $P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j} S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$

If the recommendation is generated based on collaborative filtering – CF: $P_{i,k} = \frac{R_k \sum_{j=1}^{m} S(i_k, i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$

Stage 1: Calculating the Average Ratings of Items and the item-item Similarities

1. $\{R_j, \sum_{i=1}^{\frac{n}{2}} r_{i,j}^2, S(i_j, i_k)\}_{0 \le j,k < m_1; j \le k} = SMASC(\frac{n}{2}, m_1, RM_1, \max)$;

/* Execute protocol SMASC with a set of $\frac{n}{2}$ users U_i.*/

2. $\{R_j, \sum_{i=1}^{\frac{n}{2}} r_{i,j}^2, S(i_j, i_k)\}_{m_1 \le j,k < m; j \le k} = SMASC(\frac{n}{2}, m - m_1, RM_2, \max)$;

/* Execute protocol SMASC with a set of $\frac{n}{2}$ users V_i. where: Each U_i has n_{k1} private keys and public keys $(ksu_{i,j}, KPU_{i,j})$ and has sent the public keys to the server when executing the SMASC protocol for the first time. Each V_i has n_{k2} private keys and public keys $(ksv_{i,j}, KPV_{i,j})$ and has sent the public keys to the server. */

/*Calculate the frequency values between item pairs from two data domains*/

Initialization Phase: Rs does

3. $j = 0$;
4. for $(0 \le t < n_{k1})$
5. for $(0 \le k < n_{k2})$
6. $K_j = KU_t + KV_k$;
7. j++;
8. if $(j == n_{k3} - 1)$ break;
9. if $(j == n_{k3} - 1)$ break;

/* KU_t, KV_k are the shared public keys for the user sets U_i, V_i which have been computed by the server during the execution of the SMASC protocols mentioned above*/

10. Sends to $\{U_i, V_i\}_{i=1}^{\frac{n}{2}}$: $\{K_j\}_{j=0}^{n_{k3}-1}$;

/*Execute the modified PPMFC protocol with a set of n users U_i, V_i*/

/*Where the initialization phase is not required*/

11. $suv = \{\sum_{i=1}^{\frac{n}{2}} r_{i,j} r_{i,k}\}_{0 \le j < m_1 \le k < m} = PPMFC(n, RM_1, RM_2, \max)$
12. $l = 0$;
13. for $(0 \le j < m_1)$
14. for $(m_1 \le k < m)$
15. $S(i_j, i_k) = S(i_k, i_j) = \frac{suv_l}{(\sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,k}^2})}$;
16. l++;

Stage 2: Generating Recommendations for the Target User U_i

// Similar to the proposed solution in section 3.3.2.

3.4.6. Efficiency Analysis

This section provides a detailed evaluation of the computation time for each user and the server during the construction of the recommendation system. The number of participants is fixed at 943, with the number of items varying from 100 to 500. The server's

computation time is also evaluated as the number of participants changes from 100 to 900, while the number of items remains fixed at 500. The results demonstrate the efficiency of the proposed solution.

Conclusion of the Chapter

This chapter presented two proposed privacy-preserving user recommendation system solutions: the privacy-preserving user recommendation system for the fully distributed data model, and the privacy-preserving user recommendation system for the 2PFD data model. These solutions incorporate the protocols proposed in Chapter 2 for the first phase of the recommendation system. In the second phase, two privacy-preserving recommendation generation methods are proposed: content-based filtering (CBF) and collaborative filtering (CF). These methods improve upon the original approach by employing Elliptic Curve Cryptography (ECC), ensuring consistency with the first phase of the proposed solution and enhancing the overall efficiency of the system.

- These solutions not only guarantee correctness but also preserve the privacy of users' profiles and rating histories against any third party or other users.
- Experimental results demonstrate that the proposed solution for the fully distributed model not only meets practical application requirements but also outperforms previous solutions [7, 75]. The newly proposed solution for the 2PFD model achieves good performance and is fully applicable to real-world systems.
- These solutions can be practically implemented using a client-server model, which is a widely adopted architecture today.

The content of this chapter is related to the works [CT3, CT4, CT6] in the List of Scientific Publications used in the Dissertation.

CONCLUSION

In this thesis, the researcher has studied privacy-preserving solutions for recommendation systems. The researcher has found that designing such solutions is not straightforward, even when advanced encryption techniques like homomorphic encryption and secure computation are used, especially in distributed models, such as the fully distributed setting and the 2PFD. Privacy-preserving solutions for recommendation systems that use homomorphic encryption or secure computation can ensure the correctness of the computation results, thus maintaining the quality of the recommendation system equivalent to the original system. Moreover, these solutions also ensure a high level of privacy for user data. However, the performance of these solutions is often low due to encryption operations. Therefore, in this thesis, the researcher proposes a new privacy-preserving solution for recommendation systems that ensures high privacy for user data while maintaining the correctness of recommendations equivalent to existing typical solutions and offering better performance.

I. Results

The results of the thesis can be summarized as follows:

1. Development of Two Secure Computation Protocols:

- The thesis has proposes a methodology for securely computing multiple averages and similarities in a fully distributed situation. This protocol ensures that the output is valid and equivalent to primitive operations without security protections. It also has higher computational and communication efficiency than previously published protocols within the same computational model. Theoretical proofs show that its security level is equivalent to typical high-security protocols for computing averages and similarities.

- The thesis has developed a secure multi-frequency computation protocol in 2PFD. The proposed protocol preserves the correctness of the output, ensuring equivalence to primitive operations without security measures, while improving computational and communication efficiency over previously published protocols within the same computational model. Theoretical proofs demonstrate that its security level is equivalent to typical high-security protocols.

2. Design of a PPRS based on the developed secure computation protocols:

- PPRS in the fully distributed setting
- PPRS in the 2PFD

These solutions can be easily implemented in a client-server model, which is a commonly used model today.

For each proposal:

+ Prove that the correctness of the recommendation results is preserved, equivalent to the original solution without security measures applied.

+ Prove that the proposed solutions ensure the privacy of honest participants in the semi-honest model, even when there is no collusion between parties. In cases where there is collusion among a large number of semi-honest users, the privacy of the honest participants remains guaranteed. This is equivalent to the high security level of existing solutions with strong privacy guarantees.

+ Theoretically demonstrate that the communication cost and computational cost of the proposed solutions are better than those of several existing typical solutions.

Finally, the simulation program of the protocols and solutions was implemented and constructed to experiment and demonstrate the feasibility of the developed protocols and solutions. The results were compared with previously published works, showing that the protocols and solutions of this dissertation outperform the existing solutions in terms of both communication cost and computational cost. Therefore, these solutions are fully feasible for practical implementation.

II. Future Works

Although the developed protocols can be used for several privacy-preserving recommendation system applications, there are still some areas that can be improved in the future:

1. Continue to improve to enhance the security level of the solutions.

2. Further research on other homomorphic encryption systems to develop quantum-resistant SMC protocols that only execute with authenticated data.

3. Continue research and improvements to reduce the number of interactions between users and the computation center in the proposed frequency calculation solution for the 2PFD.

4. Conduct research to expand the practical application scope of the protocols proposed in the dissertation.

5. Investigate solutions to ensure privacy for recommendation systems using more complex recommendation techniques.

6. Based on the requirements of real-world problems, new distributed data models, or different security levels, new secure computation protocols need to be developed.