BỘ GIÁO DỤC VÀ ĐÀO TẠO BAN CƠ YẾU CHÍNH PHỦ

HỌC VIỆN KỸ THUẬT MẬT MÃ

VŨ THỊ VÂN

PHÁT TRIỂN GIAO THỨC TÍNH TOÁN BẢO MẬT VÀ ỨNG DỤNG CHO HỆ TƯ VẤN NGƯỜI DÙNG CÓ ĐẢM BẢO TÍNH RIÊNG TƯ

LUẬN ÁN TIẾN SĨ NGÀNH AN TOÀN THÔNG TIN

Hà Nội – 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO BAN CƠ YẾU CHÍNH PHỦ

HỌC VIỆN KỸ THUẬT MẬT MÃ

VŨ THỊ VÂN

PHÁT TRIỂN GIAO THỨC TÍNH TOÁN BẢO MẬT VÀ ỨNG DỤNG CHO HỆ TƯ VẤN NGƯỜI DÙNG CÓ ĐẢM BẢO TÍNH RIÊNG TƯ

Ngành: An toàn thông tin
Mã số: 9480202

LUẬN ÁN TIẾN SĨ NGÀNH AN TOÀN THÔNG TIN

CÁN BỘ HƯỚNG DẪN KHOA HỌC **NGHIÊN CỨU SINH**

PGS.TS. Lương Thế Dũng **TS. Hoàng Văn Quân** **Vũ Thị Vân**

XÁC NHẬN CỦA ĐƠN VỊ ĐÀO TẠO

Hà Nội – 2025

LỜI CẢM ƠN

Về phía cá nhân, Nghiên cứu sinh xin bày tỏ lòng biết ơn sâu sắc tới các nhà khoa học: PGS.TS. Lương Thế Dũng và TS. Hoàng Văn Quân, các Thầy đã tận tình giúp đỡ, hướng dẫn, trang bị phương pháp nghiên cứu, kinh nghiệm, kiến thức khoa học và kiểm tra, đánh giá các kết quả trong suốt quá trình thực hiện Luận án.

Về phía tập thể, Nghiên cứu sinh xin trân trọng cảm ơn Học viện Kỹ thuật mật mã là cơ sở đào tạo và đơn vị quản lý chuyên môn, các Đồng chí lãnh đạo Học viện Kỹ thuật mật mã - Ban Cơ yếu Chính phủ, nơi nghiên cứu sinh đang công tác đã tạo mọi điều kiện thuận lợi, hỗ trợ và giúp đỡ nghiên cứu sinh trong suốt quá trình học tập, nghiên cứu và thực hiện Luận án. Xin chân thành cảm ơn các nhà giáo, các nhà khoa học, các đồng chí và đồng nghiệp thuộc Khoa An toàn thông tin, Phòng Đại học - Học viện Kỹ thuật mật mã đã giúp đỡ, hỗ trợ nghiên cứu sinh trong quá trình thực hiện Luận án.

Cuối cùng, tác giả vô cùng biết ơn các bạn bè và người thân trong gia đình vì sự cổ vũ to lớn của họ trong suốt thời gian hoàn thành Luận án này.

Hà Nội, 2025

Nghiên cứu sinh

Vũ Thị Vân

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của đồng tác giả trước khi đưa vào Luận án. Các kết quả nêu trong Luận án là trung thực và chưa từng được ai công bố trong các công trình nào khác.

Hà Nội, 2025

Nghiên cứu sinh

Vũ Thị Vân

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU

Ký hiệu	**Mô tả**
$(\{0,1\}^*)^n$	n phần tử, trong đó mỗi phần tử là một chuỗi hữu hạn 0 và 1
$\|a\|$	Kích thước của a
$\in_R$	Phép chọn một phần tử ngẫu nhiêu, đồng nhất trong một miền dữ liệu
p	Số nguyên tố xác định trường cơ sở F_p cho hệ mật mã Elgamal
p_e	Số nguyên tố xác định trường cơ sở F_{p_e} cho hệ mật mã trên đường cong Elliptic
g	Phần tử sinh trên F_p
q	Số nguyên tố, bậc của phần tử sinh g, sao cho $q = \frac{p-1}{2}$
$\mathbb{E}(F_{p_e})$	Đường cong Elliptic trên trường F_{p_e}
G	Điểm cơ sở trên $\mathbb{E}(F_{p_e})$
d	Bậc của điểm cơ sở G trên $\mathbb{E}(F_{p_e})$
m	Số lượng các mục tin trong hệ tư vấn
n	Số lượng người dùng trong hệ thống
i_j	Mục tin thứ j
U_i	Người dùng thứ i thuộc miền dữ liệu thứ nhất
V_i	Người dùng thứ i thuộc miền dữ liệu thứ hai
CC	Trung tâm tính toán
Rs	Máy chủ tư vấn
Tu	Người dùng cần tư vấn
$r_{i,j}$	Xếp hạng của người dùng thứ i đối với mục tin i_j
$f_{i,j}$	Cờ đánh dấu trạng thái xếp hạng của người dùng thứ i đối với mục tin i_j
$u_{i,j}$	Giá trị tấn suất thứ j của người dùng thứ i trong miền dữ liệu thứ nhất
$v_{i,j}$	Giá trị tấn suất thứ j của người dùng thứ i trong miền dữ liệu thứ hai
su_j	Giá trị tần suất của thứ j của tất cả người dùng trong hai miền dữ liệu
$S_{j,k}$	Độ tương tự giữa hai mục tin i_j và i_k
$P_{i,k}$	Dự đoán xếp hạng của người dùng thứ i đối với mục tin i_k
$Dlog(p,g,B)$	Tính logarit rời rạc của B trên trường Z_p với phần tử sinh g

$Dlog(p, g, max, B)$	Tính logarit rời rạc của B trên trên trường Z_p với phần tử sinh g với giới hạn cận trên của đầu ra biết trước là max
$Dlog_{EC}(\mathbb{E}, G, B)$	Tính logarit rời rạc của điểm B trên $\mathbb{E}(F_p)$ với điểm cơ sở G
$Dlog_{EC}(\mathbb{E}, G, max, B)$	Tính logarit rời rạc của điểm B trên $\mathbb{E}(F_p)$ với điểm cơ sở G với giới hạn cận trên của đầu ra biết trước là max
T_m	Thời gian thực hiện phép nhân modulo trên F_p
T_{ap}	Thời gian thực hiện phép cộng điểm modulo trên $\mathbb{E}(F_{p_e}) \approx 0,12T_m$ [13]
T_mp	Thời gian thực hiện phép nhân điểm modulo trên $\mathbb{E}(F_{p_e}) \approx 29T_m$ [13]
T_e	Thời gian thực hiện phép toán lũy thừa modulo trên $F_p \approx 240T_m$ [13]
T_i	Thời gian thực hiện phép nghịch đảo modulo trên $F_p \approx 11,6T_m$ [31]
T_H	Thời gian thực hiện phép băm
T_{DL}	Thời gian thực hiện phép tính logarit rời rạc
$\|H\|$	Kích thước thuật toán băm

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	**Mô tả**
2PFD	2 Party fully distributed setting – Phân tán đầy đủ 2 bên
CSDL	Cơ sở dữ liệu
CCA	Chosen-ciphertext attack– Tấn công lựa chọn bản mã
CPA	Chosen-plaintext attack – Tấn công lựa chọn bản rõ
DP	Differential privacy – Riêng tư khác biệt
EM	Expectation-maximization
FL	Federated learning – Học kết hợp
FHE	Full Homomorphic encryption – Mã hoá đồng cấu đầy đủ
HE	Homomorphic encryption – Mã hoá đồng cấu
MPC	Multi-party computation – Tính toán nhiều thành viên
OT	Oblivious transfer – Chuyển giao không tiết lộ thông tin
PHE	Partial Homomorphic encryption – Mã hoá đồng cấu một phần
PPDM	Privacy-Preserving Data Mining – Khai phá dữ liệu có đảm bảo tính riêng tư
PPDA	Privacy-Preserving Data Analysis – Phân tích dữ liệu có đảm bảo tính riêng tư
PPRS	Privacy Preserving Recommender System – Hệ tư vấn người dùng có đảm bảo tính riêng tư
RS	Recommender System – Hệ tư vấn người dùng
SM	Secure Computation – Tính toán bảo mật
SMAC	Secure Multiparty Average Computation – Tính trung bình có bảo mật
SMASC	Secure multiparty Multi-Average and Similarity Computation – Tính nhiều giá trị trung bình và độ tương tự có bảo mật
SMC	Secure Multiparty Computation – Tính toán bảo mật nhiều thành viên
SMFC	Secure multiparty Multi- Frequency Computation – Tính nhiều giá trị tần suất có bảo mật
SMSC	Secure Multiparty Similarity Computation – Tính độ tương tự có bảo mật
SWHE	Somewhat homomorphic encryption – Mã hoá đồng cấu một số phần

DANH SÁCH BẢNG

DANH SÁCH HÌNH VẼ

DANH MỤC GIAO THỨC

DANH MỤC GIẢI PHÁP

DANH MỤC THUẬT TOÁN

PHẦN MỞ ĐẦU

1. Lí do chọn đề tài

Sự phát triển và phổ biến của Internet, thiết bị thông minh và dịch vụ mạng xã hội đã dẫn đến việc mở rộng các dịch vụ ứng dụng web khác nhau. Điều này có thể là một thách thức đối với người dùng đối với việc lọc tất cả thông tin đó để thu được thông tin thiết yếu. Nhiều công ty thương mại điện tử trực tuyến giới thiệu sản phẩm cho người dùng, bán hàng triệu sản phẩm trên một nền tảng. Một người dùng hàng ngày phải duyệt qua tất cả các khả năng có thể là quá sức, điều này có thể gây quá tải thông tin. Hệ tư vấn người dùng (RS) nhằm giải quyết vấn đề quá tải thông tin đó đồng thời cá nhân hóa trải nghiệm người dùng bằng cách cung cấp các đề xuất chính xác, được cá nhân hóa về mặt hàng/sản phẩm cho người dùng theo sở thích của họ [68].

Các RS trở thành một phần không thể thiếu trong thế giới hiện đại ngày nay. Sự phổ biến và triển khai rộng rãi của chúng trong nhiều lĩnh vực khác nhau, từ thương mại điện tử đến phương tiện truyền thông xã hội [63]. Hệ tư vấn là hệ thống lọc thông tin cung cấp đề xuất mục tin (item) được cá nhân hóa cho người dùng trong môi trường dịch vụ có thể lưu giữ hoặc thu thập nhiều dữ liệu khác nhau. Lọc thông tin chủ yếu được sử dụng trong các hệ tư vấn, được điều chỉnh theo sở thích của người dùng hoặc chỉ đề xuất các mục tin được đánh giá là hữu ích cho người dùng [39].

Tuy nhiên, các hệ thống này yêu cầu người dùng chia sẻ thông tin của mình với máy chủ tư vấn trung tâm. Máy chủ này thu thập thông tin tùy chọn từ hàng triệu người dùng và sử dụng mối tương quan giữa những thông tin mà người dùng cung cấp để tạo các đề xuất được cá nhân hóa. Điều này gây lo ngại về tính riêng tư vì máy chủ có thông tin chi tiết về sở thích của người dùng. Máy chủ có thể sử dụng thông tin này để dự đoán chính xác tình trạng sức khỏe của người dùng và lập trường về các chủ đề nhạy cảm như tôn giáo, đảng phái chính trị, v.v. Ngoài ra, dữ liệu do người dùng tạo ra thuộc về người dùng và do đó họ phải có quyền kiểm soát dữ liệu đó đang được sử dụng.

Phần lớn người dùng hệ thống tư vấn không biết về những lo ngại về tính riêng tư như vậy. Nhưng đối với những người dùng đã biết, những lo ngại về tính riêng tư này sẽ ngăn cản họ cung cấp các thông tin chính xác cho hệ thống. Điều này dẫn đến việc giảm chất lượng của các đề xuất do có sự đánh đổi cố hữu giữa độ chính xác và tính riêng tư trong các hệ tư vấn. Càng nhiều phản hồi mà người dùng chia sẻ với hệ thống thì các tư vấn của người dùng tốt hơn nhưng sẽ dẫn đến tính riêng tư của người dùng thấp hơn do hệ thống có quyền truy cập vào nhiều thông tin hơn về người dùng đó. Tương tự, khi người dùng chia sẻ ít hơn với hệ thống, chất lượng đề xuất giảm xuống nhưng tính riêng

tư tốt hơn vì họ chia sẻ ít thông tin hơn [53]. Hơn nữa, những thông tin cá nhân này có thể bị lạm dụng hoặc thậm chí bán lại cho các bên trái phép để kiếm lợi nhuận, khiến người dùng gặp rủi ro [16].

Do đó, trong những năm gần đây, rất nhiều nhà nghiên cứu đã cố gắng tìm cách tạo ra một hệ thống tốt hơn có thể duy trì độ chính xác của tư vấn trong khi vẫn bảo vệ tính riêng tư của người dùng [29] như trong các công trình [2, 7, 19, 55, 72, 75]. Tuy nhiên khi dữ liệu nhạy cảm được lưu trữ một cách tập trung dù ở dạng thức nào đi chăng nữa cũng mang lại những nguy cơ về vi phạm tính riêng tư. Do đó, các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho các mô hình phân tán đang nhận được sự quan tâm lớn. Vì thế Luận án tập trung giải quyết bài toán đảm bảo tính riêng tư cho hệ tư vấn người dùng trong mô hình dữ liệu phân tán.

Các phương pháp được sử dụng để bảo vệ tính riêng tư của hệ tư vấn người dùng có thể được chia thành hai loại [29]: dựa trên thống kê, dựa trên mật mã. Phương pháp thống kê là một phương pháp hiệu quả, đảm bảo tính riêng tư của người dùng bằng cách loại bỏ các thuộc tính nhạy cảm trong dữ liệu. Tuy nhiên phương pháp dựa trên thống kê gây ra thêm nhiễu, vì thế phải đánh đổi giữa tính riêng tư và độ chính xác. Phương pháp dựa trên mật mã là một phương pháp an toàn hơn nhưng tiêu tốn tài nguyên tính toán. Bằng cách sử dụng mã hóa đồng cấu, dữ liệu có thể được bảo vệ hoàn toàn mà không bị mất thông tin, điều này có thể đảm bảo chất lượng khuyến nghị của hệ tư vấn người dùng. Mà chất lượng tư vấn là một yếu tố vô cùng quan trọng đối với một hệ tư vấn người dùng. Vì thế nghiên cứu sinh lựa chọn xây dựng giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư dựa trên tính toán bảo mật (SMC). Tính toán bảo mật đề cập đến cách thức cho phép nhiều bên tham gia cùng nhau tính toán một hàm xác định theo một cách an toàn dựa trên các đầu vào riêng tư của các bên tham gia mà không tiết lộ chúng cho bất kỳ ai.

Trong luận án, tính riêng tư dữ liệu của người dùng được đảm bảo bằng cách mỗi người dùng sẽ lưu trữ các giá trị riêng tư của mình, cụ thể là xếp hạng của họ. Khi thực hiện tính toán, các dữ liệu riêng tư này được bảo vệ thông qua sử dụng kỹ thuật dựa trên mã hóa đồng cấu và tính toán bảo mật để đưa ra các tư vấn dựa trên dữ liệu được mã hóa. Ưu điểm chính của kỹ thuật này là bảo toàn được tính đúng đắn của các tư vấn và dễ dàng đáp ứng các ràng buộc an toàn cao. Sử dụng các công cụ và kỹ thuật mã hóa đồng cấu, chúng tôi xây dựng các giao thức đảm bảo tính riêng tư khác nhau cho các mô hình phân tán khác nhau bằng cách phân tích các yêu cầu và thách thức về tính riêng tư của chúng.

Cụ thể hơn, luận án tập trung vào giải pháp hệ tư vấn có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ và phân tán đầy đủ hai bên. Đây là các mô hình dữ liệu

của các bài toán hệ tư vấn hiện có, ví dụ như hệ tư vấn sản phẩm của các trang thương mại điện tử như Amazon, Flip kart, hệ tư vấn bạn bè, địa điểm tham quan và thậm chí cả tư vấn phim dựa trên sở thích của Facebook,...

Mặc dù đã có các giải pháp PPRS khác nhau dựa trên mật mã được đề xuất cho mô hình dữ liệu phân tán đầy đủ, nhưng hoặc là chúng yêu cầu có sự hiện diện của một bên thứ ba tin cậy, hoặc chúng yêu cầu các bên không được thông đồng với nhau như trong các công trình [2, 18, 19, 47, 48, 58, 71, 72] hoặc chỉ bảo vệ tính riêng tư trong giai đoạn mô hình hoá và giai đoạn tính toán như trong các công trình [2, 19, 72, 83] hoặc việc triển khai phức tạp, yêu cầu khả năng tính toán lớn [9, 12, 38, 43, 59, 61] hoặc độ chính xác có thể bị giảm [3, 4, 28, 30, 34, 41, 62].

Trong đó, kỹ thuật tư vấn sử dụng trong các tài liệu [7, 8, 55, 75] rất gần với thực tế được áp dụng cho hệ thống tư vấn sản phẩm của các sàn thương mại điện tử như Amazon, eBay, Alibaba [39]. Đây là các giải pháp sử dụng bộ lọc dựa trên bộ nhớ sử dụng hành vi trong quá khứ của người dùng để đưa ra tư vấn. Lọc dựa trên bộ nhớ rất dễ thực hiện và yêu cầu tính toán tối thiểu, có khả năng diễn giải, có thể xử lý sự thưa thớt của dữ liệu tốt hơn so với lọc dựa trên mô hình. Hơn nữa các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư này không yêu cầu một bên thứ ba tin cậy, đồng thời dữ liệu nhạy cảm của người dùng luôn chỉ tồn tại dạng mã hoá trong quá trình xử lý và truyền thông. Các giải pháp này đồng thời bảo toàn tính riêng tư trong cả ba giai đoạn với mô hình dữ liệu phân tán đầy đủ, quan trọng hơn, tất cả dữ liệu phản hồi chỉ được lưu giữ ở phía máy khách, nghĩa là nó sẽ không có bất kỳ cơ hội nào được phát hành hoặc truyền đến máy chủ không tin cậy hoặc bất kỳ máy khách độc hại nào khác. Tuy nhiên giải pháp trong tài liệu [7, 8] mặc dù có hiệu suất tốt hơn nhưng mức độ an toàn thấp hơn các giải pháp [55, 75]. Vì thế, luận án sẽ tập trung đề xuất giải pháp PPRS có mức độ an toàn tương đường với các giải pháp hiện có có mức độ an toàn cao [55, 75] và hiệu suất tương đương giải pháp có hiệu suất tốt [7, 8].

Ngoài ra, trong thực tế sẽ tồn tại những trường hợp mà các dữ liệu của một người dùng có thể được phân bố ở nhiều vị trí khác nhau (được gọi là các miền dữ liệu khác nhau), ví dụ như dữ liệu của một bệnh nhân có thể một phần do bệnh nhân nắm giữ, các phần khác do các bệnh viện khác nhau nắm giữ hay dữ liệu mua bán, đánh giá của một người dùng có thể nằm trên các nền tảng mua bán trực tuyến khác nhau,... Việc kết hợp thông tin từ các bên có thể giúp cho chất lượng tư vấn được tốt hơn. Hầu hết các phương pháp hiện tại đều dựa vào giả định lưu trữ tập trung dữ liệu người dùng, điều này chắc chắn gây ra rủi ro đáng kể về rò rỉ tính riêng tư của người dùng. Tuy nhiên, các nghiên cứu về giải pháp PPRS cho các kịch bản này là chưa nhiều và còn tồn tại một số vấn đề: các giải pháp được đề xuất trong tài liệu [20, 22, 26] sử dụng phương pháp học tập

trung và dữ liệu vẫn có nguy cơ bị tấn công nếu được lưu trữ ở phía máy chủ [12], các giải pháp trong tài liệu [21, 84] phải đánh đổi giữa độ chính xác và tính riêng tư, các giải pháp được đề xuất trong [26, 57, 58] lại yêu cầu có một bên đáng tin cậy. Đặc biệt chưa có giải pháp nào được đề xuất cho kịch bản dữ liệu được phân tán đầy đủ hai bên, trong đó tập dữ liệu được phân phối trên một số lượng lớn người dùng, mỗi bản ghi thuộc sở hữu của hai người dùng khác nhau, người thứ nhất chỉ biết các giá trị cho một tập hợp con các thuộc tính, trong khi người còn lại biết các giá trị cho các thuộc tính còn lại. Vì vậy trong luận án này sẽ đề xuất giải pháp PPRS mới cho mô hình dữ liệu phân tán đầy đủ hai bên.

2. Mục tiêu nghiên cứu

Luận án tập trung vào giải quyết một số vấn đề chủ yếu sau đây:

1. Phát triển giao thức tính toán bảo mật hiệu quả dựa vào kỹ thuật mật mã: duy trì được tính đúng đắn, tính riêng tư và nâng cao tính hiệu quả so với giao thức điển hình hiện có.

2. Đề xuất ứng dụng giao thức đã được phát triển để đảm bảo tính riêng tư cho hệ tư vấn người dùng. Giải pháp đề xuất duy trì được tính đúng đắn, tính riêng tư và nâng cao hiệu suất so với giải pháp điển hình hiện có.

3. Đối tượng nghiên cứu, phạm vi nghiên cứu

+ Đối tượng nghiên cứu: Giao thức tính toán bảo mật và ứng dụng cho hệ tư vấn người dùng có đảm bảo tính riêng tư.

+ Phạm vi nghiên cứu: Giao thức tính toán bảo mật và ứng dụng tính toán bảo mật trong đảm bảo tính riêng tư trong phân tích dữ liệu cho hệ tư vấn người dùng.

4. Phương pháp nghiên cứu

Luận án sử dụng đồng thời nhiều phương pháp nghiên cứu khoa học như:

- Phương pháp nghiên cứu dựa trên tài liệu: Thu thập, phân tích, xử lý thông tin dựa trên các tài liệu như sách, báo, tạp chí,... đã in ấn hoặc công bố trên internet liên quan đến đề tài.

- Phương pháp nghiên cứu chứng minh giả thuyết dựa trên các luận cứ khoa học: Chứng minh tính đúng đắn, mức độ an toàn của giải pháp đề xuất thông qua các luận cứ khoa học.

- Phương pháp nghiên cứu dựa trên thực nghiệm: Thông qua việc thử nghiệm và

đối sánh với các kết quả đã công bố để chứng minh tính hiệu quả của giải pháp đề xuất.

- Phương pháp nghiên cứu phi thực nghiệm: Tham khảo ý kiến của các chuyên gia thông qua các hội thảo trong nước và quốc tế.

5. Nội dung nghiên cứu

Nội dung nghiên cứu cơ bản được trình bày trong luận án bao gồm:

- Nghiên cứu về hệ tư vấn người dùng, những rủi ro vi phạm tính riêng tư, các kỹ thuật đảm bảo tính riêng tư đối với hệ tư vấn người dùng, tính toán bảo mật. - Phân tích, đánh giá các giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng hiện có để thấy được ưu điểm, nhược điểm của mỗi giải pháp. Trên cơ sở đó đề xuất các giải pháp mới.

- Phát triển giao thức tính toán bảo mật phục vụ cho việc giải quyết bài toán đã được xác định, chứng minh tính đúng đắn, mức độ an toàn và hiệu suất của giao thức.

- Đề xuất ứng dụng các giao thức đã được phát triển để đảm bảo tính riêng tư cho hệ tư vấn người dùng, phân tích tính đúng đắn, độ an toàn và hiệu suất của giải pháp.

6. Tính mới trong khoa học của Luận án

Những đóng góp khoa học mới của Luận án bao gồm:

- Phát triển 02 giao thức tính toán bảo mật hiệu quả:

+ Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự bảo mật trong mô hình dữ liệu phân tán đầy đủ.

+ Giao thức tính đồng thời nhiều giá trị tần suất bảo mật trong mô hình dữ liệu phân tán đầy đủ hai bên.

- Đề xuất 02 giải pháp đảm bảo tính riêng tư trong bài toán phân tích dữ liệu cho hệ tư vấn người dùng sử dụng các giao thức đã được đề xuất:

+ Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình dữ liệu phân tán đầy đủ.

+ Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình dữ liệu phân tán đầy đủ hai bên.

7. Ý nghĩa khoa học và thực tiễn của Luận án

Về phương diện lý thuyết:

- Nghiên cứu, phân tích, đánh giá các ưu điểm và hạn chế của một số giao thức tính toán bảo mật, đặc biệt là các giao thức trong mô hình dữ liệu phân tán đầy đủ và phân tán đầy đủ hai bên ứng dụng trong quá trình đảm bảo tính riêng tư cho hệ tư vấn người dùng.

- Nghiên cứu phát triển các giao thức tính toán bảo mật để đảm bảo tính riêng tư trong quá trình phân tích dữ liệu của hệ tư vấn người dùng với mô hình dữ liệu phân tán đầy đủ và phân tán đầy đủ hai bên.

- Nghiên cứu và đề xuất các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư dựa trên các giao thức đã đề xuất.

Về phương diện thực tiễn:

- Luận án có thể được sử dụng làm tài liệu giảng dạy cho học phần Mật mã ứng dụng trong an toàn thông tin, giao thức an toàn ở bậc đại học, cao học ngành An toàn thông tại Khoa An toàn thông tin, Học viện Kỹ thuật mật mã.

- Luận án có thể được sử dụng là tài liệu tham khảo cho các sinh viên đại học và cao học ngành an toàn thông tin làm đề tài về đảm bảo tính riêng tư. - Nếu được đầu tư về tài chính và nhân lực, kết quả của luận án có thể được hoàn thiện và áp dụng cho các hệ thống thực tế.

8. Cấu trúc của Luận án

Ngoài phần mở đầu, kết luận và phụ lục, nội dung của luận án được bố cục thành 3 chương như sau:

CHƯƠNG 1. TỔNG QUAN VỀ HỆ TƯ VẤN NGƯỜI DÙNG CÓ ĐẢM BẢO TÍNH RIÊNG TƯ VÀ TÍNH TOÁN BẢO MẬT

Trình bày vắn tắt về hệ tư vấn người dùng, những rủi ro về vi phạm tính riêng tư trong hệ tư vấn người dùng, các kỹ thuật đảm bảo tính riêng tư, tính toán bảo mật, phân tích, đánh giá một số công cụ cho việc phát triển, từ đó xác định bài toán đảm bảo tính riêng tư cho hệ tư vấn người dùng của luận án.

CHƯƠNG 2. PHÁT TRIỂN MỘT SỐ GIAO THỨC TÍNH TOÁN BẢO MẬT HIỆU QUẢ

Trình bày về các giao thức bảo mật đề xuất và cải tiến, chứng minh tính đúng đắn và tính riêng tư của giao thức, kết quả thực nghiệm để đánh giá về tính hiệu quả của giao thức.

CHƯƠNG 3. GIẢI PHÁP ĐẢM BẢO TÍNH RIÊNG TƯ CHO HỆ TƯ VẤN NGƯỜI

DÙNG DỰA TRÊN CÁC GIAO THỨC TÍNH TOÁN BẢO MẬT

Trình bày về giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư sử dụng các giao thức đã đề xuất ở Chương 2, chứng minh tính đúng đắn và tính riêng tư của giải pháp, kết quả thực nghiệm để đánh giá về tính hiệu quả của giải pháp mới so với giải pháp ban đầu.

CHƯƠNG 1. TỔNG QUAN VỀ HỆ TƯ VẤN NGƯỜI DÙNG CÓ ĐẢM BẢO TÍNH RIÊNG TƯ VÀ TÍNH TOÁN BẢO MẬT

1.1. Mở đầu chương

Hệ tư vấn người dùng có lịch sử phát triển lâu đời với nhiều loại kỹ thuật tư vấn khác nhau và đã được ứng dụng trong nhiều hệ thống thực tế. Tuy nhiên hệ tư vấn hoạt động dựa trên việc phân tích dữ liệu thu thập được từ các các nhân, tổ chức, trong đó có chứa những dữ liệu nhạy cảm. Điều đó dẫn đến mối lo ngại về vi phạm tính riêng tư của các các nhân, tổ chức. Vì thế, chương này sẽ trình bày bức tranh tổng quan về hệ tư vấn người dùng, bài toán đảm bảo tính riêng tư cho hệ tư vấn người dùng và từ đó xác định những vấn đề cần được giải quyết.

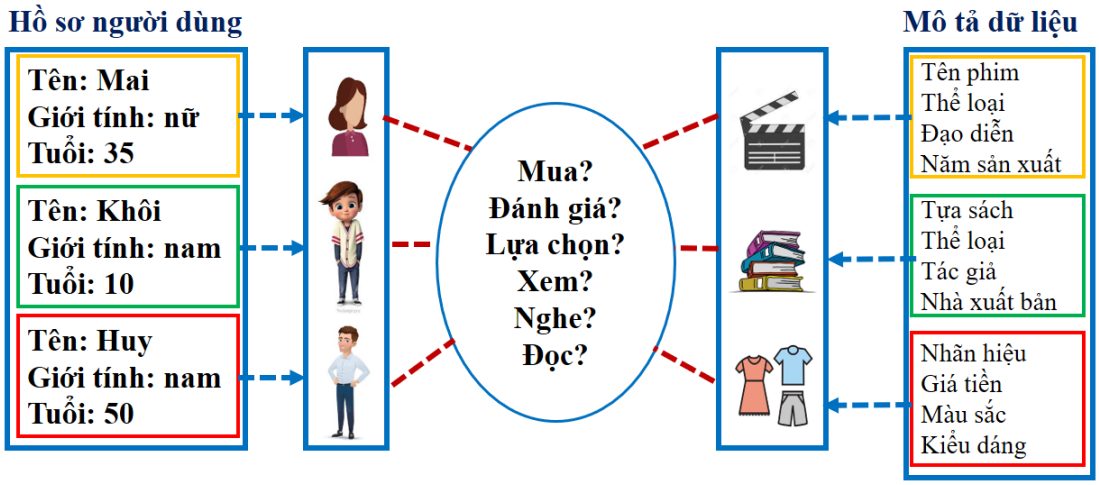
Bố cục của chương được sắp xếp như sau: mục 1.2 sẽ tóm lược về hệ tư vấn người dùng, cấu trúc, phân loại và hoạt động của hệ tư vấn người dùng. Dựa trên đó, những rủi ro về vi phạm tính riêng tư trong quá trình hoạt động của hệ tư vấn người dùng sẽ được trình bày trong mục 1.3. Từ đó, mục 1.4 sẽ đưa ra các kỹ thuật đảm bảo tính riêng tư cho hệ tư vấn người dùng, phân tích các ưu nhược điểm của các kỹ thuật và lựa chọn kỹ thuật dựa tính toán bảo mật để đảm bảo tính riêng tư cho hệ tư vấn. Từ lựa chọn đó, tính toán bảo mật được hệ thống hoá trong mục 1.5. Sau đó, các nghiên cứu liên quan đến giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư sử dụng SMC hiện có được trình bày trong mục 1.6. Từ các phân tích đó, mục 1.7 sẽ đưa ra các vấn đề mở cần nghiên cứu để đảm bảo tính riêng tư cho hệ tư vấn người dùng.

1.2. Hệ tư vấn người dùng

Hệ tư vấn người dùng là hệ thống hỗ trợ ra quyết định nhằm tư vấn các thông tin liên quan đến người dùng một cách dễ dàng và nhanh chóng, phù hợp với từng người dùng [1]. Ví dụ với trang web Amazon, một trong những trang web thương mại điện tử nổi tiếng nhất, khi người dùng truy cập vào trang web này họ sẽ được tư vấn những sản phẩm tiềm năng nhất từ hàng triệu sản phẩm trong hệ thống. RS như một công cụ cung cấp những thông tin hữu ích và riêng biệt theo từng cá nhân trên một hệ thống chứa đựng một lượng lớn thông tin. Các RS được thiết kế nhằm cung cấp cho người dùng những tư vấn liên quan, những tư vấn hiệu quả nhất có thể từ thông tin của các mục dữ liệu, từ hồ sơ người sử dụng và từ mối liên hệ giữa những đối tượng này.

Cấu trúc của một RS gồm có ba thành phần chính [1]: tập hợp các người dùng $U = \{u_1, u_2, ..., u_n\}$ bao gồm các thông tin của người dùng được lưu trên hệ thống; tập hợp các mục dữ liệu $I = \{i_1, i_2, ..., i_m\}$ bao gồm định danh và các thuộc tính của mục dữ

liệu; tập hợp các "mối quan hệ" $R = (u_i, i_j)$ giữa "người dùng" và "mục dữ liệu", đây là tập hợp các giao dịch liên kết giữa tập hợp người dùng U và tập hợp mục dữ liệu I và những mô tả của mối liên kết này [69]. Một RS có thể được miêu tả như trong Hình 1.1.



Hình 1.1: Sơ đồ tổng quát của hệ tư vấn người dùng

RS thông thường được xây dựng dựa trên 3 bước tuần tự:

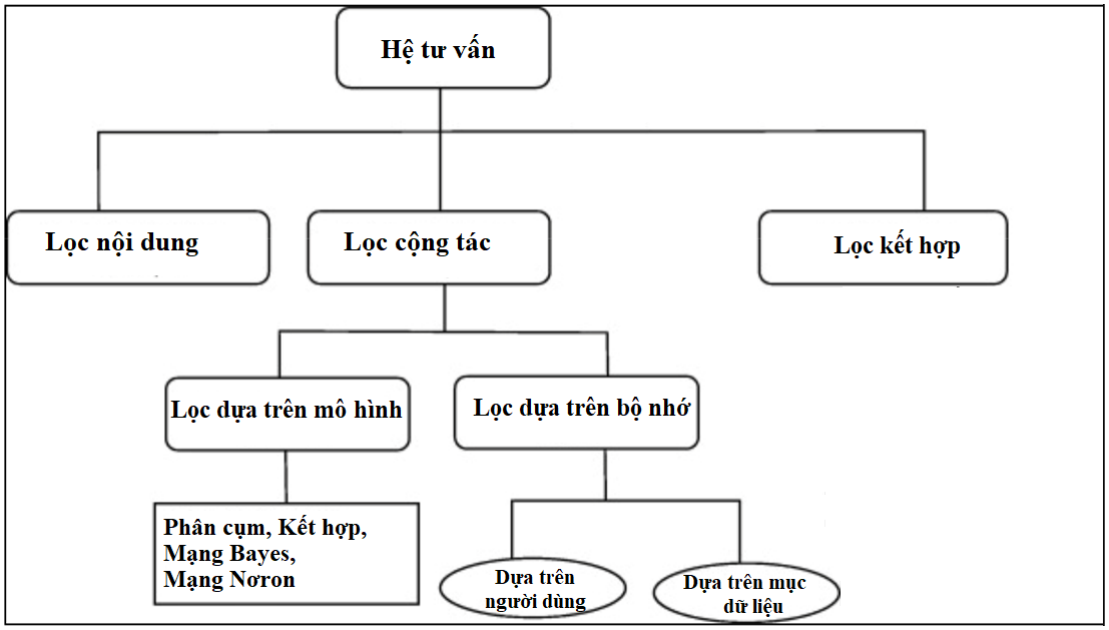
Bước thứ nhất: tiền xử lý dữ liệu sẽ sử dụng để xây dựng hệ thống dựa trên dữ liệu có sẵn trong hệ thống. Những dữ liệu có sẵn này (các mục dữ liệu, danh sách người dùng, mối liên hệ giữa các người dùng) được biểu diễn lại dưới nhiều dạng khác nhau như véc tơ các từ khóa hay các mối quan hệ,...

Bước thứ hai: tính sự tương quan giữa những người dùng hoặc mục dữ liệu. Các chỉ số Pearson, độ tương tự cosin, các phương pháp mô hình,... được sử dụng ở bước này.

Cuối cùng, hệ thống sẽ đưa ra một danh sách các mục dữ liệu khuyến nghị hoặc giá trị đánh giá dự đoán của mục dữ liệu (ví dụ như giá trị đánh giá dự đoán của một bộ phim hay một quyển sách).

Dựa trên cách thức xây dựng, RS được chia thành 3 loại chính [1]: RS dựa trên nội dung (CBF - Content-based filtering), RS dựa trên lọc cộng tác (CF - Collaborative Filtering), và RS lai – kết hợp phương pháp lọc cộng tác và phương pháp dựa trên nội dung như Hình 1.2. RS dựa trên nội dung sử dụng sự tương đồng giữa các mục dữ liệu, các mục được mô tả bởi một số lượng lớn các từ khóa. Hệ thống coi các mục có cùng từ khóa là có sự tương đồng. Số lượng từ khóa được chia sẻ cao hơn, tư vấn mức độ quan hệ chặt chẽ hơn. Nếu người dùng đánh giá một sản phẩm, RS sẽ đề xuất những sản phẩm tương tự nhất tiếp theo.

RS dựa trên lọc cộng tác đưa ra quyết định dựa trên thông tin cộng tác được thu thập bởi nhiều người dùng tương tự. Hệ thống so sánh người dùng dựa trên các quyết



Hình 1.2: Các loại hệ tư vấn người dùng

định hoặc xếp hạng của họ trong quá khứ. Sau đó, hệ thống giả định rằng những người dùng đã đưa ra lựa chọn tương tự sẽ tiếp tục làm như vậy trong tương lai. Những sản phẩm được đánh giá tốt bởi những người dùng tương tự sẽ được tư vấn nhiều hơn.

RS lai là sự kết hợp của hai phương pháp lọc CF và CBF. Việc kết hợp các thuộc tính của các loại lọc khác nhau giúp tăng hiệu suất hoặc giảm các nhược điểm của RS. Tuy nhiên việc kết hợp này có thể dẫn đến các mô hình phức tạp hơn và các vấn đề tương ứng của sự phức tạp này. [63]

Thông thường, ma trận đánh giá rất thưa thớt do thiếu các giá trị vì tất cả người dùng trong hệ thống không thể đánh giá tất cả các mục tin. Đánh giá còn thiếu được biểu thị $r_{i,j} = 0$. Do đó, mục tiêu chính của hệ thống tư vấn là dự đoán đánh giá cho người dùng u_i đối với mục tin i_j mà trước đó người dùng chưa đánh giá. Tùy thuộc vào các thuật toán tư vấn cơ bản trong các tình huống khác nhau, các loại đề xuất cho người dùng có thể khác nhau [7]. Phương pháp CBF và CF để dự đoán các xếp hạng còn thiếu trong hệ thống tư vấn dựa trên mục tin được sử dụng trong luận án.

1.3. Những rủi ro về vi phạm tính riêng tư trong hệ tư vấn người dùng

Nghiên cứu về RS có truyền thống lâu đời và việc bảo vệ tính riêng tư trong hệ thống tư vấn (Privacy Preserving Recommender System - PPRS) đã được cộng đồng nghiên cứu quan tâm đáng kể trong những năm gần đây. "Quyền riêng tư" có nhiều định nghĩa khác nhau, một trong những định nghĩa được chấp nhận rộng rãi nhất là "Quyền của một cá nhân để kiểm soát các điều khoản/giới hạn của thông tin cá nhân có thể nhận dạng được đối với cá nhân đó được thu thập, tiết lộ hoặc sử dụng" [29]. Định nghĩa trên về quyền riêng tư nhấn mạnh sự kiểm soát của chủ sở hữu thông tin đối với nó, thay vì

khả năng không sử dụng những thông tin đã được niêm phong hoàn chỉnh. Do đó, khi xem xét việc sử dụng thông tin trong mạng, đảm bảo tính riêng tư là đảm bảo cho các thông tin của các cá nhân không bị thu thập, lưu trữ hoặc công khai bởi người khác một cách trái phép. Và các thông tin đó có mức độ công khai như thế nào, vào thời điểm nào, theo cách thức nào sẽ do chủ sở hữu thông tin đó quyết định.

Điều này cũng thể hiện rõ nét và cụ thể trong Điều 38 Bộ luật Dân sự của nước Việt Nam năm 2015: "Đời sống riêng tư, bí mật cá nhân, bí mật gia đình là bất khả xâm phạm và được pháp luật bảo vệ. Việc thu thập, lưu giữ, sử dụng, công khai thông tin liên quan đến đời sống riêng tư, bí mật cá nhân phải được người đó đồng ý, việc thu thập, lưu giữ, sử dụng có công khai thông tin liên quan đến bí mật gia đình phải được các thành viên gia đình đồng ý, trừ trường hợp luật có quy định khác; Thư tín, điện thoại, điện tín, cơ sở dữ liệu điện tử và các hình thức trao đổi thông tin riêng khác của cá nhân được bảo đảm an toàn và bí mật. Việc bóc mở, kiểm soát, thu giữ thư tín, điện thoại, điện tín, cơ sở dữ liệu điện tử và các hình thức trao đổi thông tin riêng tư khác của người khác chỉ được hiện trong trường hợp luật quy định; Các bên trong hợp đồng không được tiết lộ thông tin về đời sống riêng tư, bí mật cá nhân, bí mật gia đình của nhau mà mình đã biết được trong quá trình xác lập, thực hiện hợp đồng, trừ trường hợp có thoả thuận khác".

Hệ tư vấn người dùng yêu cầu bộ dữ liệu về sở thích cá nhân của người dùng, điều này có nguy cơ bị tiết lộ bởi những kẻ xâm nhập độc hại. Những vi phạm quyền riêng tư này có thể liên quan đến các hành động có chủ ý như tấn công hoặc rình mò bởi các loại bên khác nhau như người dùng bên ngoài, người dùng hệ thống hoặc nhà cung cấp dịch vụ hoặc dịch vụ bên ngoài. Những sự cố như vậy có thể gây ra hậu quả nghiêm trọng tùy thuộc vào mức độ nhạy cảm của thông tin cá nhân.

Trong hệ tư vấn, máy chủ càng cung cấp nhiều thông tin cá nhân thì các đề xuất càng chính xác hơn. Thông tin của người dùng bao gồm nhận dạng, thông tin nhân khẩu học, hành vi, mô hình hoặc lịch sử mua hàng, xếp hạng, v.v. Những thông tin như vậy rất nhạy cảm về quyền riêng tư và người dùng thường miễn cưỡng chia sẻ những thông tin này với bất kỳ nhà cung cấp dịch vụ bên thứ ba nào. Việc không thể có được hồ sơ xác thực của người dùng hầu hết dẫn đến các đề xuất không đủ tiêu chuẩn, đặc biệt là về độ chính xác dự đoán, vì những dữ liệu dùng được dùng để xây dựng hệ tư vấn không đáng tin cậy [85]. Do đó, các chuyên gia đã có sự quan tâm đáng kể đến việc đảm bảo tính riêng tư của hệ tư vấn để đưa ra dự đoán chính xác ở mức thỏa đáng mà không vi phạm những lo ngại về tính riêng tư của cá nhân [65].

Các rủi ro về vi phạm tính riêng tư của người dùng có thể xảy ra trong cả ba giai đoạn của RS [29]:

Giai đoạn đầu tiên là mô hình hóa người dùng. Trong giai đoạn này, thông tin

nhu cầu và sở thích được cá nhân hóa của người dùng được thu thập và các mô hình người dùng được thiết lập. Các phương pháp để có được sở thích và nhu cầu được cá nhân hóa bao gồm các cách rõ ràng do người dùng chủ động cung cấp, chẳng hạn như các thuộc tính cơ bản như tên người dùng, từ khóa người dùng nhập, phản hồi tùy chọn của người dùng, v.v. Và các cách ngầm như theo dõi hành vi duyệt web của người dùng (chẳng hạn như số lần nhấp chuột, thời gian duyệt, đánh dấu trang, v.v.) bằng cách suy luận sau khi theo dõi hành vi của người dùng. Sau khi có được thông tin được cá nhân hóa, sẽ thu được mô hình người dùng. Các thông tin được cá nhân hoá của người dùng được cấu trúc lại theo cách thích hợp để tạo điều kiện thuận lợi cho việc xử lý và sử dụng trong hệ thống, sau đó kết nối giữa tính toán với mô hình. Tiếp theo, mô hình người dùng sẽ được cập nhật theo sự thay đổi của nhu cầu cá nhân hóa của người dùng. Các lo ngại về tính riêng tư phát sinh trong giai đoạn này bao gồm truy cập, thu thập, giám sát, phân tích, hợp nhất, truyền tải, lưu trữ dữ liệu không đúng cách, v.v.

Giai đoạn thứ hai là giai đoạn tính toán. Ở giai đoạn này, cần tiến hành tính toán các tham số của hệ tư vấn người dùng. Những lo ngại về tính riêng tư bao gồm phân tích, tổng hợp dữ liệu, truyền dữ liệu không đúng cách, gian lận danh tính, ...

Giai đoạn cuối cùng là tạo ra kết quả tư vấn. Các lo ngại về tính riêng tư phát sinh từ giai đoạn này bao gồm phân tích, kết hợp dữ liệu, truyền dữ liệu không đúng cách, tiết lộ thông tin tư vấn, thông tin tư vấn sai lệch.

Việc cung cấp hoặc tiết lộ thông tin đó cho máy chủ tư vấn thường là bất kỳ nhà cung cấp dịch vụ bên thứ ba nào có thể gây ra những rủi ro không mong muốn về quyền riêng tư. Dữ liệu có thể được bán cho bất kỳ bên nào khác mà không có sự đồng ý của chủ sở hữu hoặc kẻ tấn công có thể tấn công và đưa ra tư vấn sai. Vì vậy, điều rất quan trọng là phải bảo vệ những thông tin đó đồng thời đưa ra các tư vấn chính xác.

Hơn thế nữa việc lưu trữ tập trung dữ liệu của nhưng người dùng hệ thống dù ở dạng này hay dạng khác cũng đem đến các mối lo ngại về quyền riêng tư của dữ liệu. Do đó, các giải pháp hệ tư vấn có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán là rất cần thiết và đang được các nhà nghiên cứu đặc biệt quan tâm. Phương pháp mã hoá thường được sử dụng để bảo vệ tính riêng tư của cá nhân trong các hệ tư vấn người dùng phân tán [32]. Vì vậy, bài toán đảm bảo tính riêng tư hệ tư vấn cho mô hình dữ liệu phân tán sẽ được tập trung giải quyết trong luận án này.

1.4. Các kỹ thuật đảm bảo tính riêng tư cho hệ tư vấn

Để giải quyết những lo ngại về vấn đề vi phạm tính riêng tư, các kỹ thuật đảm bảo tính riêng tư đã được phát triển để bảo vệ tính riêng tư dữ liệu của người dùng trong khi vẫn cho phép hệ tư vấn hoạt động hiệu quả gồm có [5]: kỹ thuật dựa trên mật mã

(tính toán bảo mật, mã hoá đồng cấu - HE), kỹ thuật dựa trên nhiễu (riêng tư khác biệt - Differential privacy - DP), Mã thông báo (Tokenization) và kỹ thuật dựa trên mặt nạ (ẩn danh (Anonymization), biệt danh (Pseudonymization)). Trong phần này trình bày các kỹ thuật đảm bảo tính riêng tư cho hệ tư vấn và so sánh các ưu điểm cũng như hạn chế của chúng.

1.4.1. Kỹ thuật dựa trên mật mã

Kỹ thuật dựa trên mật mã gồm có: tính toán bảo mật, mã hoá đồng cấu. [5]

Tính toán bảo mật (SMC) là một kỹ thuật mã hóa cho phép hai hoặc nhiều bên tính toán một cách an toàn một chức năng trên dữ liệu riêng tư của họ mà không bên nào có thể truy cập dữ liệu của bên kia. Nói cách khác, SMC đảm bảo rằng dữ liệu riêng tư của mỗi bên vẫn được bảo mật trong khi vẫn cho phép các bên nhận được kết quả tính toán trên tất cả dữ liệu mà không cần phải tự chia sẻ dữ liệu. SMC đã được sử dụng trong RS để bảo vệ dữ liệu người dùng và cho phép các tư vấn hợp tác giữa nhiều bên. Lợi ích của việc sử dụng SMC trong RS bao gồm khả năng bảo vệ tính riêng tư dữ liệu của người dùng, khả năng cộng tác tạo ra các tư vấn được cá nhân hóa từ nhiều bên và khả năng tạo hệ thống tư vấn phi tập trung. Tuy nhiên, việc sử dụng SMC cũng có một số hạn chế, chẳng hạn như chi phí tính toán cao liên quan đến tính toán an toàn, cũng như sự phức tạp trong việc thiết kế và triển khai các giao thức an toàn đa bên.

Mã hóa đồng cấu là một loại mã hoá cho phép bên tính toán thực hiện các phép toán trên dữ liệu ở dạng mã mà cần truy cập đến dữ liệu ở dạng rõ. Trong RS, mã hóa đồng cấu cho phép dữ liệu được mã hóa và chia sẻ giữa nhiều bên đồng thời cho phép tính toán các đề xuất dựa trên dữ liệu được mã hóa. Có hai loại mã hóa đồng cấu chính được sử dụng trong RS: mã hóa đồng cấu đầy đủ (FHE) và mã hóa đồng cấu một phần (PHE). Lợi ích chính của HE trong RS là cung cấp mức độ riêng tư và bảo mật cao cho người dùng, vì dữ liệu của họ được mã hóa và được giữ bí mật. Tuy nhiên, hạn chế chính của HE là tốn kém về mặt tính toán, có thể làm giảm hiệu suất tổng thể của hệ thống.

Tính toán bảo mật có thể tính toán được mọi hàm tổng quát tuy nhiên một nhược điểm chính của SMC là có độ phức tạp tương tác (communication/interaction) cao của các bên tham gia. Ngược lại, mã hoá đồng cấu có độ phức tạp tương tác rất thấp (không tương tác) nhưng lại có nhược điểm là có độ phức tạp tính toán phụ thuộc vào kích thước hàm tính toán (thường là độ phức tạp tính toán cao).

Bên cạnh, chúng ta có thể sử dụng kết hợp hai kỹ thuật này nhằm tận dụng ưu điểm của chúng. Việc này sẽ cho phép thực hiện các phép tính trực tiếp trên dữ liệu đã mã hóa mà không cần giải mã, giúp bảo vệ dữ liệu trong suốt quá trình xử lý. Đồng thời, việc kết hợp này cũng đảm bảo rằng không có bên nào nắm giữ toàn bộ thông tin, mà chỉ

có các phần nhỏ của dữ liệu, giảm nguy cơ rò rỉ dữ liệu. Dữ liệu có thể được phân chia và mã hóa, làm cho cả bên tính toán lẫn các bên liên quan không thể truy cập thông tin gốc. Khi kết hợp hai kỹ thuật này sẽ giúp giải quyết vấn đề chi phí tính toán và lưu trữ rất lớn ở một bên. Tuy nhiên, để áp dụng hiệu quả, cần đánh giá cẩn thận các yêu cầu về hiệu suất, chi phí và tính khả thi của hệ thống. Việc tối ưu hóa thuật toán và phân bổ tài nguyên hợp lý sẽ giúp giảm thiểu nhược điểm và khai thác tối đa ưu điểm của sự kết hợp này. Để thực hiện được điều này đòi hỏi phải có kiến thức chuyên sâu về mật mã, phân tán và các giao thức bảo mật.

1.4.2. Kỹ thuật dựa trên nhiễu

Riêng tư khác biệt là một kỹ thuật bảo vệ tính riêng tư bằng cách thêm nhiễu ngẫu nhiên vào dữ liệu, gây khó khăn cho việc xác định dữ liệu riêng lẻ trong tập dữ liệu [5]. Kỹ thuật này đã được sử dụng trong nhiều bối cảnh khác nhau, bao gồm cả hệ tư vấn người dùng.

Lợi ích của việc sử dụng DP trong RS bao gồm khả năng bảo vệ tính riêng tư dữ liệu của người dùng trong khi vẫn cho phép máy chủ lấy thông tin hữu ích từ dữ liệu. Hơn nữa, nó đảm bảo rằng dữ liệu thô không bao giờ bị lộ ra máy chủ, điều này rất quan trọng để bảo vệ quyền riêng tư của người dùng.

Những hạn chế của việc sử dụng DP trong RS là nhiễu được thêm vào dữ liệu có thể dẫn đến kết quả không chính xác, cũng như tăng thời gian tính toán do cần phải thêm nhiễu [5]. Ngoài ra, rất khó để xác định mức nhiễu tối ưu cho một tập dữ liệu nhất định, điều này có thể dẫn đến kết quả không chính xác. Do đó, điều cần thiết là phải thiết kế một hệ thống có thể cân bằng hiệu quả giữa độ chính xác của đề xuất và tính riêng tư. [5]

1.4.3. Mã thông báo (Tokenization)

Mã thông báo trong RS giúp chia dữ liệu thành các thành phần nhỏ hơn (mã thông báo) để giúp phân tích và xử lý chúng dễ dàng hơn. Điều này gồm việc chia nhỏ văn bản thành các từ, câu và cụm từ cũng như chia nhỏ dữ liệu số thành các số và các giá trị khác. Mã thông báo cũng giúp đảm bảo tính riêng tư và bí mật dữ liệu bằng cách ngăn chặn rò rỉ dữ liệu nhạy cảm của người dùng, chẳng hạn như thông tin cá nhân [5].

Lợi ích của việc sử dụng mã thông báo trong RS là cải thiện độ chính xác, tính riêng tư và khả năng mở rộng. Đặc biệt, tokenization cho phép hệ thống xử lý dữ liệu nhanh chóng và chính xác hơn mà vẫn đảm bảo tính riêng tư dữ liệu của người dùng. Ngoài ra, nó cho phép hệ thống mở rộng quy mô sang các bộ dữ liệu lớn hơn mà không ảnh hưởng đến hiệu suất.

Tuy nhiên, hạn chế của việc sử dụng mã thông báo trong RS bao gồm sự phức tạp của quy trình mã thông báo và khả năng mất dữ liệu. Quá trình mã thông báo có thể tốn nhiều công sức tính toán và tốn thời gian, đồng thời có thể dẫn đến mất dữ liệu nếu mã thông báo không được tạo chính xác. Ngoài ra, một số mã thông báo có thể khó xử lý hơn các mã thông báo khác, dẫn đến các vấn đề tiềm ẩn về độ chính xác. [5]

1.4.4. Kỹ thuật dựa trên mặt nạ

Kỹ thuật dựa trên mặt nạ gồm: ẩn danh và biệt danh [5].

Ẩn danh là một kỹ thuật bảo vệ tính riêng tư dữ liệu của người dùng trong RS bằng cách che giấu danh tính của họ. Ý tưởng chính của kỹ thuật này là che dấu thông tin nhận dạng của người dùng, chẳng hạn như ID người dùng, tên người dùng và các thông tin khác liên quan đến người dùng, trong khi vẫn duy trì tiện ích dữ liệu của họ. Kỹ thuật này có thể bảo vệ thông tin nhận dạng của người dùng trong RS đồng thời cho phép hệ thống đưa ra các tư vấn hữu ích và chính xác. Kỹ thuật ẩn danh rất dễ thực hiện vì nó chỉ yêu cầu che giấu thông tin nhận dạng của người dùng. Những hạn chế của việc sử dụng tính năng ẩn danh trong RS bao gồm khả năng rò rỉ dữ liệu cũng như nguy cơ biến dạng dữ liệu do che giấu thông tin nhận dạng người dùng. [5]

Biệt danh là một kỹ thuật bảo vệ tính riêng tư dữ liệu của người dùng được sử dụng trong RS hoạt động bằng cách thay thế các thuộc tính có thể nhận dạng của họ bằng biệt danh. Đó là một phương pháp che giấu danh tính thực sự của người dùng đồng thời cho phép người đó được xác định bằng biệt danh của mình. Việc đặt biệt danh có thể bảo vệ các thuộc tính nhạy cảm của người dùng, chẳng hạn như tuổi tác, giới tính và vị trí cũng như các sở thích rõ ràng và ngầm định của người đó. Lợi ích của việc sử dụng kỹ thuật biệt danh trong RS bao gồm cải thiện tính riêng tư cho người dùng vì dữ liệu của họ không được lưu trữ và chỉ biệt danh của họ được sử dụng để nhận dạng họ. Hơn nữa, việc đặt biệt danh cũng cho phép khả năng mở rộng RS tốt hơn khi lượng dữ liệu được lưu trữ giảm xuống và dữ liệu có thể được chia thành các phần nhỏ hơn. Tuy nhiên, cũng có một số hạn chế khi sử dụng biệt danh trong RS. Một trong những hạn chế chính là khó có thể liên kết chính xác biệt danh của người dùng với danh tính và sở thích thực sự của họ. Hơn nữa, dữ liệu có biệt danh cũng có thể dễ bị tấn công từ điển và tấn công nhận dạng lại. [5]

1.4.5. So sánh các kỹ thuật đảm bảo tính riêng tư cho hệ tư vấn

Các kỹ thuật ở trên đã được sử dụng rộng rãi trong RS để cung cấp khả năng bảo vệ tính riêng tư mạnh mẽ hơn. Bảng 1.1 trình bày chi tiết sự so sánh ưu, nhược điểm giữa các kỹ thuật này.

Bảng 1.1: So sánh các kỹ thuật đảm bảo tính riêng tư

Kỹ thuật	Ưu điểm	Nhược điểm
Tính toán bảo mật	- Duy trì độ chính xác - Đảm bảo tính riêng tư ở mức cao - Có khả năng cộng tác tạo ra các tư vấn được cá nhân hóa từ nhiều bên và khả năng tạo hệ tư vấn phi tập trung	- Chi phí tính toán cao - Phức tạp trong việc thiết kế và triển khai
Mã hoá đồng cấu	- Duy trì độ chính xác - Đảm bảo tính riêng tư ở mức cao	- Chi phí tính toán rất cao
Riêng tư khác biệt	- Dễ thực hiện	- Tăng thời gian tính toán do cần phải thêm nhiễu - Rất khó để xác định mức nhiễu tối ưu cho một tập dữ liệu nhất định, điều này có thể dẫn đến kết quả không chính xác - Đánh đổi giữa tính riêng tư và tính đúng đắn
Mã thông báo	- Cải thiện độ chính xác, tính riêng tư và khả năng mở rộng	- Tốn nhiều chi phí và thời gian tính toán - Có thể mất dữ liệu - Tồn tại vấn đề tiềm ẩn về độ chính xác
Ẩn danh	- Dễ thực hiện	- Rò rỉ dữ liệu - Biến đạng dữ liệu
Biệt danh	- Cải thiện tính riêng tư dữ liệu của người dùng - Hiệu suất tốt	- Khó có thể liên kết chính xác biệt danh với danh tính và sở thích thực sự - Bị tấn công từ điển và tấn công nhận dạng lại

Mã hóa đồng cấu và SMC đều là các cơ chế dựa trên mã hóa, giúp đảm bảo tính riêng tư ở mức cao trong khi vẫn duy trì độ chính xác, nhưng chi phí tính toán cao. SMC tuy làm giảm chi phí tính toán nhưng lại làm tăng chi phí liên lạc.

Mặt khác, riêng tư khác biệt đảm bảo tính riêng tư bằng cách thêm nhiễu ngẫu nhiên, có chi phí tính toán thấp và không thêm bất kỳ chi phí liên lạc nào. Tuy nhiên, việc thêm nhiễu ngẫu nhiên chắc chắn sẽ ảnh hưởng đến hiệu suất của mô hình ở một mức độ nhất định và làm ảnh hưởng đến tính đúng đắn.

Kỹ thuật mã thông báo mặc dù cải thiện về độ chính xác, tính riêng tư, khả năng mở rộng, tuy nhiên lại có thể mất dữ liệu và tồn tại vấn đề tiềm ẩn về độ chính xác. Ẩn danh và biệt danh đều là các cơ chế dựa trên mặt nạ, đây là kỹ thuật đơn giản, dễ thực hiện, có hiệu suất tốt. Tuy nhiên kỹ thuật này vẫn tồn tại các hiểm hoạ về tính riêng tư (rò rỉ dữ liệu, biến dạng dữ liệu, nhận dạng lại,. . .).

Đối với bài toán xây dựng hệ tư vấn thì chất lượng của hệ tư vấn (tính đúng đắn) là quan trọng nhất, nó có vai trò quyết định trong việc có đưa hệ tư vấn này vào sử dụng hay không. Do đó, kỹ thuật dựa trên mật mã là kỹ thuật được lựa chọn trong luận án để đảm bảo tính riêng tư cho hệ tư vấn người dùng phân tán vì kỹ thuật này có thể đảm bảo được tính đúng đắn và dễ dàng đáp ứng các ràng buộc an toàn cao.

1.5. Tính toán bảo mật

Bài toán tính toán nhiều bên được thực hiện bằng cách xác định một quy trình ngẫu nhiên ánh xạ chuỗi đầu vào (một đầu vào cho mỗi bên) thành chuỗi đầu ra tương ứng (mỗi bên một đầu ra). Như đã đề cập trước đây, tính toán nhiều bên là một mô hình tính toán cụ thể của lĩnh vực tính toán phân tán.

Bài toán tính toán nhiều bên tổng quát được phát biểu như sau [24]:

Định nghĩa 1.1. *Cho $n(n \geq 2)$ là số người tham gia vào quá trình tính toán phân tán, trong đó bên thứ i nắm giữ một đầu vào x_i ($i = 1,\ldots,n$) và tất cả các đầu vào có cùng độ dài ($|x_i| = |x_j|$ với $\forall i, j$). Hàm tính toán nhiều bên f được định nghĩa như sau:*

$$f : (\{0,1\}^*)^n \rightarrow (\{0,1\}^*)^n \tag{1.5.1}$$

là một chuỗi ánh xạ quá trình ngẫu nhiên có dạng:

$$\bar{x} = (x_1,\ldots,x_n) \rightarrow f(\bar{x}) = (f_1(\bar{x}),\ldots,f_n(\bar{x})) \tag{1.5.2}$$

Điều này có nghĩa là $\forall i$, bên thứ i ban đầu giữ một đầu vào x_i và mong muốn lấy phần tử thứ i trong $f(\bar{x})$, ký hiệu là $f_i(\bar{x})$.

Hàm tính toán nhiều thành viên f có thể thuộc một trong các loại sau:

• Các hàm tất định (deterministic functionality): trả về một đầu ra duy nhất có cùng giá trị đầu vào.

• Các hàm tổng quát (bao gồm cả hàm tất định và hàm không tất định): có thể trả về các giá trị đầu ra khác nhau với cùng một giá trị đầu vào trong các lần thực thi khác nhau.

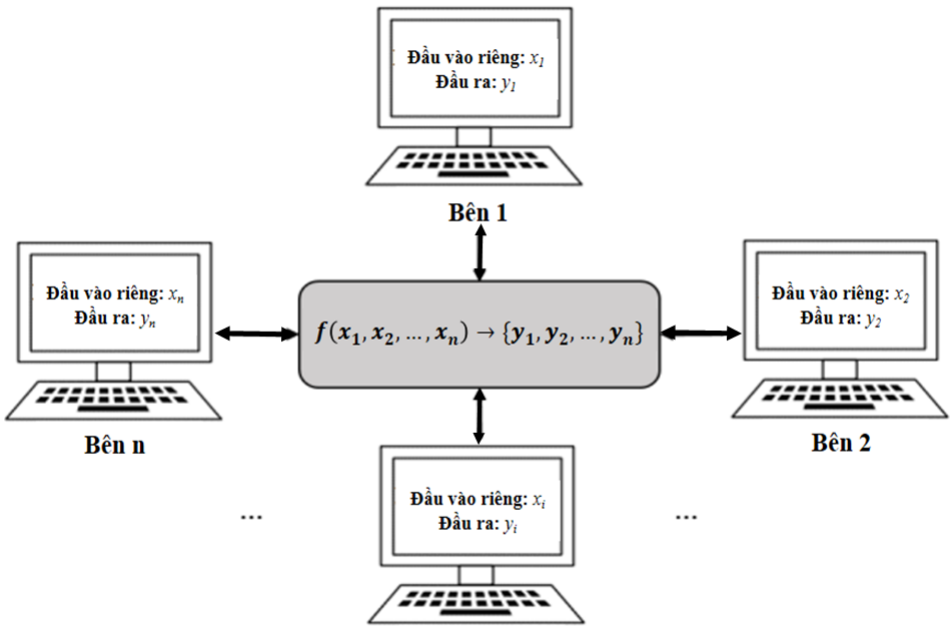
Giao thức tính toán nhiều bên được định nghĩa như sau:

Định nghĩa 1.2. *[24] Giao thức π để tính toán hàm f là một quy trình ngẫu nhiên ánh xạ chuỗi các đầu vào (mỗi bên một đầu vào) với các chuỗi đầu ra tương ứng, sao cho phân phối đầu ra của giao thức (khi được thực hiện đúng theo quy tắc của giao thức) trên chuỗi đầu vào $(x_1,x_2,...,x_n)$ có phân phối giống $f(x_1,x_2,...,x_n)$.*

Tính toán bảo mật (Secure Computation - SC) hay tính toán nhiều thành viên (Multiparty Computation - MPC) hay tính toán bảo mật nhiều thành viên (Secure Multiparty Computation - SMC) là một lĩnh vực của mật mã với mục tiêu tạo ra các phương thức cho phép các bên cùng tính toán một hàm dựa trên các giá trị đầu vào của họ mà vẫn đảm bảo tính riêng tư của những giá trị đầu vào này.

Để thực hiện giao thức SMC, chỉ cần mỗi bên tham gia có một máy tính đáng tin cậy để chạy phần giao thức của mình và cách (có thể không an toàn) để giao tiếp với các bên tham gia khác. Giao thức tạo ra một loạt các thông điệp được trao đổi giữa những bên tham gia, và cuối cùng mỗi bên tham gia có được đầu ra của giao thức. Bản thân giao thức là công khai, cho phép mỗi bên tham gia xác minh độc lập rằng phần mềm chạy trên máy của chính họ là hợp lệ. Các công cụ mật mã khác có thể được tích hợp vào giao thức để ngăn các bên tham gia đi chệch khỏi giao thức quy định. Vì mỗi người dùng chỉ cần một máy tính (được tin tưởng bởi chính mình) và kết nối với người dùng khác, các giao thức SMC có thể dễ dàng được thực hiện qua Internet. [24]

Về cơ bản, lĩnh vực SMC đề cập đến các phương pháp cho phép những người tham gia tính toán một hàm f một cách an toàn dựa trên thông tin đầu vào riêng tư của họ trong khi bất kỳ ai cũng không biết gì về thông tin đầu vào của mỗi bên.



Hình 1.3: Mô hình tính toán bảo mật

Như mô tả trong Hình 1.3, mỗi bên thứ i sẽ sở hữu một giá trị đầu vào bí mật x_i và muốn có được phần tử thứ i trong $f(x_1, x_2, \ldots, x_n)$ là $f_i\ (x_1, x_2, \ldots, x_n)$ (ký hiệu y_i).

1.5.1. Các thực thể trong giao thức SMC

Giao thức SMC thường chịu sự đe dọa do các hành vi cố ý gây hại của một số thực thể đối nghịch. Nghĩa là, việc thực thi giao thức có thể bị "tấn công" bởi một thực thể bên ngoài hoặc thậm chí bởi một nhóm nhỏ các bên tham gia [44]. Mục đích của cuộc tấn công này có thể là để tìm hiểu thông tin riêng tư hoặc làm cho kết quả tính toán không chính xác. Chính vì thế trong mô hình tính toán của SMC gồm ba loại thực thể: (1) các bên trung thực (honest parties) tuân theo quy tắc của giao thức và họ không thông đồng với bất kỳ ai để thực hiện các hành vi xấu, (2) các bên thông đồng (corrupted parties) sẵn sàng thông đồng với bên khác hoặc có thể bị kiểm soát bởi một thực thể đối nghịch bên ngoài để thực hiện các hành vi nguy hiểm chống lại các bên trung thực và (3) kẻ tấn công bên ngoài (external adversary) kiểm soát các bên bán thông đồng để thực hiện các hành vi nguy hiểm.

1.5.2. Định nghĩa an toàn

Trước khi trình bày định nghĩa an toàn của giao thức SMC, Luận án sẽ mô tả về mô hình định nghĩa được lựa chọn.

1.5.2.1. Mô hình định nghĩa

Các giao thức SMC phải chống lại được mọi tấn công được định nghĩa trong mô hình an toàn gắn với giao thức đó. Để chứng minh rằng một giao thức là an toàn, cần phải có một định nghĩa về an toàn cho SMC. Có nhiều định nghĩa khác nhau đã được đề xuất và những định nghĩa này nhằm mục đích đảm bảo một số thuộc tính an toàn quan trọng đủ tổng quát để bao hầu hết các tác vụ tính toán nhiều bên [44]. Có hai cách tiếp cận để định nghĩa mô hình an toàn trong SMC: dựa trên lý thuyết trò chơi (game-based) và dựa trên mô phỏng (simulation).

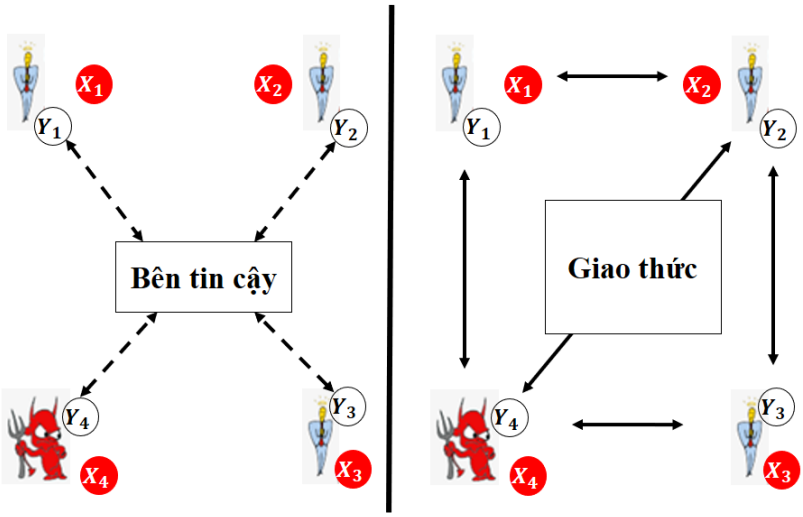
Với cách tiếp cận đầu tiên để định nghĩa an toàn cho một giao thức SMC là chỉ cần tạo một danh sách các yêu cầu riêng biệt và sau đó kết luận một giao thức là an toàn nếu tất cả các yêu cầu này được đáp ứng. Tuy nhiên, cách tiếp cận này không khả quan vì những lý do sau. Đầu tiên, có thể bỏ sót một yêu cầu quan trọng nào đó. Điều này đặc biệt đúng vì các ứng dụng khác nhau có các yêu cầu khác nhau và mong muốn là có một định nghĩa đủ chung để bao quát tất cả các ứng dụng. Thứ hai, định nghĩa phải đủ đơn giản để có thể thấy rằng tất cả các tấn công thù địch có thể xảy ra đều được ngăn chặn bởi định nghĩa được đưa ra. [44]

Do đó, an toàn của một giao thức SMC được định nghĩa một cách khái quát hóa thành "mô hình mô phỏng thực tế/lý tưởng"(ideal/real simulation paradigm) như Hình 1.4. Trong "Mô hình lý tưởng" tồn tại một bên đáng tin cậy bên ngoài sẵn sàng giúp các

bên thực hiện tính toán của họ. Vì thế, các bên có thể chỉ cần gửi thông tin đầu vào của họ cho bên đáng tin cậy, bên này sau đó sẽ tính toán hàm mong muốn và chuyển cho mỗi bên đầu ra theo quy định. Vì hành động duy nhất được thực hiện bởi mỗi bên là gửi thông tin đầu vào của mình cho bên đáng tin cậy, nên khả năng duy nhất mà kẻ tấn công có là lựa chọn thông tin đầu vào của bên trung thực.

Tất nhiên, trong "Mô hình thực tế", một kẻ tấn công không thể mua chuộc được tất cả các bên tham gia giao thức và trong mô hình thực tế không còn bên đáng tin cậy giống như trong mô hình lý tưởng. Thay vào đó, các bên thực hiện một số giao thức mà không có bất kỳ sự trợ giúp nào, và một số trong số họ là bán trung thực và thông đồng với nhau. Mặc dù vậy, một giao thức an toàn nên mô phỏng "Mô hình lý tưởng". Nghĩa là, một giao thức là an toàn, nếu bất cứ điều gì mà kẻ tấn công có thể có được trong giao thức thực thì nó cũng có thể có được trong giao thức lý tưởng.

Một cách tổng quát, tính an toàn của một giao thức được thiết lập bằng cách so sánh kết quả của việc thực thi giao thức thực với kết quả của việc thực thi giao thức lý tưởng [44].



Hình 1.4: Mô hình lý tưởng và thực tế trong tính toán phân tán

Tiếp theo, mô hình kẻ tấn công được lựa chọn trong luận án sẽ được trình bày.

1.5.2.2. Mô hình kẻ tấn công

Luận án sẽ tập trung vào hai tham số chính để xác định mô hình kẻ tấn công cho một giao thức SMC [44]: hành vi đối nghịch được cho phép và chiến lược thông đồng (corruption strategy) của kẻ tấn công.

• *Hành vi của kẻ tấn công*

Đây là tham số quan trọng nhất để xác định mô hình kẻ tấn công cho một giao thức SMC [44]. Có hai loại kẻ tấn công chính: kẻ tấn công bán trung thực (honest-but-curious/semi-honest/passive adversaries) và kẻ tấn công độc hại (malicious/active adversaries).

Trong mô hình bán trung thực, tất cả các bên tham gia (kể cả các bên thông đồng) được giả định tuân theo tất cả các giao thức một cách chính xác. Mặt khác, các bên tham gia có thể thông đồng với nhau hoặc bị kẻ tấn công kiểm soát để thực hiện các hành vi có hại như cố gắng lấy dữ liệu riêng tư của những người tham gia khác thông qua việc phân tích bất kỳ thông tin hoặc thông điệp nào họ nhận được. Kẻ tấn công bán trung thực còn được gọi là "trung thực nhưng tò mò" và "thụ động".

Ngược lại, trong các mô hình độc hại, các bên tham gia ác ý có thể tự ý thực hiện các hành vi của mình mà không tuân theo quy định của giao thức, thậm chí có thể hủy bỏ giao thức bất cứ lúc nào. Một giao thức an toàn trong mô hình độc hại (đôi khi được gọi là an toàn hoàn hảo) về cơ bản sẽ luôn luôn phát hiện ra một bên tham gia đi chệch khỏi giao thức.

Mô hình bán trung thực đóng vai trò là bước đệm và cho phép các nhà thiết kế giao thức thực hiện một cách tiếp cận mô đun hơn đối với thiết kế an toàn. Các giao thức an toàn trong mô hình độc hại cung cấp các đảm bảo an toàn mạnh nhất, nhưng khó thiết kế nhất và do đó chúng là các giao thức kém hiệu quả nhất trong thực tế [24].

Các giao thức SMC theo mô hình bán trung thực khá hiệu quả, vì vậy mô hình này phù hợp với các ứng dụng yêu cầu hiệu suất cao, chẳng hạn như phân tích và khai thác dữ liệu phân tán bảo vệ tính riêng tư [20, 39, 103]. Đây cũng là mô hình gần với thực tế, bởi vì khi một bên sẵn sàng tham gia thực hiện giao thức SMC với thiện chí và danh tiếng của mình, thì bên đó sẽ tuân theo quy tắc của giao thức. Ví dụ, có một số bệnh viện muốn cùng nghiên cứu về hồ sơ bệnh nhân thống nhất của họ. Do những hạn chế về tính riêng tư, mỗi bệnh viện không được phép biết dữ liệu của bệnh viện khác. Rõ ràng, mô hình bán trung thực phù hợp với kịch bản như vậy. Nếu các giao thức hiệu quả trong mô hình bán trung thực thì các giao thức này có thể được điều chỉnh để đạt được các mức an toàn cần thiết cho việc sử dụng trong thực tế.

Vì thế mô hình bán trung thực là mô hình an toàn được lựa chọn cho bài toán đảm bảo tính riêng tư của luận án.

● *Chiến lược mua chuộc*

Tham số này sẽ xác định vấn đề các bên bị mua chuộc khi nào và như thế nào. Có ba mô hình chính [44]:

Mô hình mua chuộc tĩnh (Static corruption model): Trong mô hình này, kẻ tấn công có thể chọn trước/mua chuộc một số bên tham gia trước khi giao thức được thực hiện và các bên bị mua chuộc không thay đổi trong suối quá trình thực hiện giao thức.

Mô hình mua chuộc thích nghi (Adaptive corruption model): Thay vì có một nhóm cố định các bên bị mua chuộc, các kẻ tấn công thích nghi được cung cấp khả năng của các bên bị mua chuộc trong quá trình tính toán. Việc lựa chọn ai để mua chuộc và khi nào có thể được quyết định tùy ý bởi kẻ tấn công và có thể phụ thuộc vào quan sát của họ về việc thực thi (vì lý do này, nó được gọi là thích nghi). Chiến lược này mô phỏng mối đe dọa của một "tin tặc" bên ngoài đột nhập vào máy trong khi thực thi hoặc một bên ban đầu trung thực và sau đó thay đổi hành vi của mình. Trong mô hình này, một khi một bên bị mua chuộc thì bên đó sẽ là bên bị mua chuộc kể từ thời điểm đó.

Mô hình an toàn chủ động (Proactive security model): Mô hình này chỉ xem xét khả năng các bên bị mua chuộc trong một khoảng thời gian nhất định. Do đó, các bên có thể là bị mua chuộc trong suốt quá trình tính toán (chẳng hạn như trong mô hình mua chuộc thích nghi), nhưng các bên bị mua chuộc cũng có thể trở nên trung thực. Mô hình chủ động có ý nghĩa trong trường hợp mối đe dọa là kẻ thù bên ngoài có thể xâm nhập mạng, các dịch vụ và thiết bị khi quá trình tính toán an toàn đang diễn ra. Khi các vi phạm được phát hiện, hệ thống sẽ được làm sạch và kẻ tấn công sẽ mất quyền kiểm soát một số máy, khiến các bên bị mua chuộc trở lại thành các bên trung thực. An toàn trong trường hợp này là kẻ tấn công chỉ có thể học được những gì có được từ trạng thái cục bộ của các bên bị mua chuộc. Kẻ tấn công như vậy đôi khi được gọi là kẻ tấn công di động.

Trong đó, chiến lược mua chuộc được luận án lựa chọn là mô hình mua chuộc thích ứng với số lượng của các bên thông đồng sẽ bị giới hạn tuỳ thuộc vào từng bài toán cụ thể.

Ngoài ra, còn một số tham số khác sẽ được trình bày trong phần tiếp theo.

● *Một số tham số khác*

Bên cạnh hai tham số chính xác định mô hình kẻ tấn công, còn có một số tham số khác như: Kênh liên lạc, Sức mạnh tính toán của kẻ tấn công, Giới hạn trên đối với số lượng các bên bị mua chuộc. Trong đó:

- Kênh liên lạc: Kênh công khai hoặc Kênh riêng (Kênh an toàn).

- Sức mạnh tính toán của kẻ tấn công: Tính toán trong thời gian đa thức hoặc Tính toán không giới hạn, kẻ tấn công có thể sở hữa máy tính lượng tử cỡ lớn. Đối với giao thức SMC, sức mạnh tính toán của kẻ tấn công thường là tính toán trong thời gian đa thức.

- Giới hạn trên đối với số lượng các bên bị mua chuộc t (trong trường hợp không xác định chính xác các bên bị mua chuộc) [14]:

+ Không có đa số trung thực (No honest majority): tính toán hai bên

+ Đa số trung thực $t < \frac{n}{2}$ (Honest majority).

+ Ít hơn một phần ba số bên có thể bị mua chuộc $t < \frac{n}{3}$ (Two-thirds honest majority) với n là tổng số các bên tham gia.

+ Kích thước cụ thể của một tập con các bên bị mua chuộc.

Trong mô hình kẻ tấn công được xem xét trong luận án như sau: các bên liên lạc với nhau qua kênh an toàn, kẻ tấn công bị giới hạn về mặt tính toán, nghĩa là nó chạy trong thời gian đa thức và giới hạn trên đối số lượng các bên bị mua chuộc thông qua thiết lập kích thước cụ thể của một tập con các bên bị mua chuộc.

Phần tiếp theo sẽ trình bày sơ bộ về một số khái niệm cần thiết được sử dụng cho định nghĩa về an toàn của giao thức SMC.

1.5.2.3. Khái niệm không thể phân biệt về mặt tính toán

Định nghĩa 1.3. *[24] Hàm không đáng kể (negligible) Gọi* $\mathbb{N}$ *là tập hợp các số tự nhiên, hàm* $\mu(\cdot) : \mathbb{N} \to [0,1]$ *là không đáng kể theo n, nếu với mọi đa thức dương* $p(\cdot)$ *tồn tại một số nguyên không âm* $n_0 > 0$ *sao cho tất cả* $n > n_0$*,*

$$\mu(n) < \frac{1}{p(n)} \tag{1.5.3}$$

Hàm không đáng kể quan trọng thường được sử dụng là: 2^{-n} [33].

Không thể phân biệt về mặt tính toán (Computational indistinguishability) là một khái niệm quan trọng khi chứng minh thuộc tính an toàn của các giao thức SMC [24]. Đặt $X = \{X_n\}_{n\in\mathbb{N}}$ là một tập hợp được lập chỉ mục bởi tham số an toàn n (thường đề cập đến kích thước của đầu vào), trong đó các X_i là biến ngẫu nhiên.

Định nghĩa 1.4. *[78]. Cho* $X = \{X_n\}_{n\in\mathbb{N}}$ *và* $Y = \{Y_n\}_{n\in\mathbb{N}}$ *là hai phân phối xác suất. Chúng ta nói X và Y là không thể phân biệt tính toán (ký hiệu* $X \overset{c}{\equiv} Y$*) nếu với mọi bộ phân biệt D có dạng là thuật toán xác suất chạy trong thời gian đa thức và* $\forall n$ *đủ lớn, ta*

có:

$$|Pr(t' \leftarrow X_n, D(t') = 1) - Pr(t' \leftarrow Y_n, D(t') = 1)| < \mu(|n|) \tag{1.5.4}$$

Trong đó:

+ $|n|$: kích thước của n.

+ $\mu(|n|)$ là hàm không đáng kể theo $|n|$.

+ $X = \{X_n\}_{n\in\mathbb{N}}$ hoặc $Y = \{Y_n\}_{n\in\mathbb{N}}$: phân phối xác suất hoặc tập các biến ngẫu nhiên được đánh chỉ số.

+ t' là một đại lượng ngẫu nhiên, được lấy mẫu đầu vào một cách ngẫu nhiên và có phân phối đều, làm đầu vào của $D(t)$.

+ $D(t')$ là một bộ phân biệt, có dạng là một thuật toán xác suất chạy trong thời gian đa thức, thuật toán cho ra kết quả đầu ra là 0 hoặc 1. Trong SMC thì $D(t')$ thường là một mô phỏng cách thức mà đối phương tấn công vào giao thức với đầu vào t cho trước, và khi mô phỏng thành công thì đầu ra của $D(t')$ bằng 1.

+ $Pr(t' \leftarrow X_n, D(t') = 1)$ là xác suất xảy ra khi $t' \leftarrow X_n$ ($t' \in X_n$, được lấy mẫu ngẫu nhiên và đồng nhất) và bộ phân biệt $D(t')$ có đầu ra bằng 1.

+ $Pr(t' \leftarrow Y_n, D(t') = 1)$ là xác suất xảy ra khi $t' \leftarrow Y_n$ ($t' \in Y_n$, được lấy mẫu ngẫu nhiên và đồng nhất) và bộ phân biệt $D(t')$ có đầu ra bằng 1.

Như vậy là hai phân phối xác suất thỏa mãn tính chất không phân biệt được về mặt tính toán nếu không có thuật toán hiệu quả chạy trong thời gian đa thức có thể phân biệt được hai phân phối này với nhau, tức là không thể phân biệt được một mẫu ngẫu nhiên đầu vào t' thuộc phân phối nào.

1.5.2.4. Định nghĩa về an toàn của giao thức SMC

Phần này trình bày định nghĩa về an toàn của giao thức SMC trong mô hình bán trung thực được tham chiếu từ khung SMC [24]. Trong mô hình này các bên tuân theo đúng giao thức, ngoại trừ việc các bên đó lưu giữ hồ sơ về tất cả các tính toán trung gian của mình. Nói một cách dễ hiểu, một giao thức SMC an toàn nếu bất kỳ thông tin gì mà một tập hợp (hoặc một liên minh) các bên bị mua chuộc có thể thu được sau khi tham gia vào giao thức về cơ bản có thể thu được từ đầu vào và đầu ra của chính các bên này.

Định nghĩa 1.5. *[24]. (Tính riêng tư đối với mô hình bán trung thực) Cho f là một hàm tính toán nhiều thành viên như được định nghĩa trong Định nghĩa 1.1. Giao thức π tính toán an toàn hàm f với t bên tham gia bị mua chuộc nếu $\forall I \subseteq \{1,2,\ldots,n\}$ sao cho $|I| = t$, tồn tại thuật toán xác suất thời gian đa thức M sao cho:*

• *Trong trường hợp f là hàm tất định:*

$$\{M(I,\overline{x_I},f_I(\overline{x}))\}_{\overline{x}\in(\{0,1\}^*)^n} \stackrel{c}{\equiv} \{VIEW^{\pi}_{A,I}(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \quad (1.5.5)$$

• *Trong trường hợp tổng quát:*

$$\{M(I,\overline{x_I},f_I(\overline{x}),f(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \stackrel{c}{\equiv} \{VIEW^{\pi}_{A,I}(\overline{x}),OUTPUT^{\pi}(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \quad (1.5.6)$$

Trong đó:

• $VIEW^{\pi}_{A,I}(\overline{x})$*: những gì mà t bên thông đồng quan sát được và tất cả các thông điệp (được truyền giữa các bên trung thực) mà kẻ tấn công A nghe lén trong quá trình thực thi giao thức* 4π *với đầu vào* $\overline{x} = (x_1,x_2,...,x_n)$.

• $\overline{x_I}$*: đầu vào của tất cả các bên tham gia bị mua chuộc trong giao thức* π.

• $f_I(\overline{x})$*: chuỗi đầu ra của tất cả các bên tham gia bị mua chuộc trong giao thức* π.

• $OUTPUT^{\pi}(\overline{x})$*: chuỗi đầu ra của tất cả các thành viên tham gia giao thức* π. *Trong trường hợp đầu tiên,* $OUTPUT^{\pi}(\overline{x}) \equiv f(\overline{x})$.

• $\stackrel{c}{\equiv}$ *là không thể phân biệt được về mặt tính toán (mô hình tính toán (computational model)).*

Bên cạnh đó, có một định lý thành phần thường được sử dụng để xây dựng các giao thức SMC trong mô hình bán trung thực (Định lí 1.1). Trước khi trình bày nội dung cụ thể của định lý, luận án sẽ trình bày về khái niệm quan trọng trong định lý này.

Định nghĩa 1.6. *[24]. (Phép quy dẫn an toàn, trường hợp nhiều bên):*

- Một giao thức tiên tri được cho là sử dụng hàm tiên tri f nếu như các câu trả lời từ giao thức tiên tri này đều tuân theo f. Điều này có nghĩa là khi hàm tiên tri được gọi, bên i viết câu hỏi q_i ($i \in [1,n]$), các câu trả lời được phân phối theo $f(q_1,q_2,\ldots,q_n)$, trong đó bên i nhận phần i (tức là $f_i(q_1,q_2,\ldots,q_n)$), độ dài của mỗi câu hỏi phải có mối quan hệ đa thức với độ dài của đầu vào ban đầu.

- Một giao thức tiên tri sử dụng hàm tiên tri f được cho là sẽ tính toán hàm g một cách riêng tư nếu tồn tại thuật toán thời gian đa thức, được ký hiệu là M, thỏa mãn phương trình (1.5.6), trong đó khung nhìn của việc thực thi giao thức tiên tri được xác định theo cách tự nhiên.

- Một giao thức tiên tri được cho là quy dẫn an toàn của hàm g tới hàm f nếu nó tính toán g một cách riêng tư khi sử dụng hàm tiên tri f. Trong trường hợp như vậy, ta nói rằng hàm g có thể quy dẫn an toàn tới hàm f.

Tiếp theo, luận án sẽ trình bày định lý thành phần cho mô hình bán trung thực. Định lý quan trọng này phát biểu rằng nếu một giao thức có thể được phân tách thành các giao thức con, thì tính an toàn của giao thức đó sẽ được chứng minh nếu có thể chỉ ra rằng các giao thức con của nó là an toàn [24].

Định lí 1.1. *[24]. (Định lý thành phần cho mô hình bán trung thực, trường hợp nhiều bên) Giả sử rằng hàm g có thể quy dẫn an toàn tới hàm f và tồn tại một giao thức tính f an toàn, thì tồn tại một giao thức tính g an toàn.*

Chứng minh của định lí này đã được trình bày cụ thể trong tài liệu [24] nên không được trình bày ở trong Luận án này.

1.5.2.5. Một số nguyên thuỷ mật mã

Phần này trình bày một số nguyên thuỷ mật mã được sử dụng cho quá trình phát triển giao thức SMC trong luận án.

• *Hệ mật mã Elgamal*

Chọn nhóm tuần hoàn $\mathbb{G}$ có phần tử sinh g trong $\mathbb{Z}_p$, sao cho $q|(p-1)$, thoả mãn $g\#1$, $g^q mod p = 1$ với p và q là hai số nguyên tố lớn. Tất cả các tính toán đều được thực hiện trong $\mathbb{Z}_p$.

Hệ mật mã Elgamal an toàn dựa trên độ khó của việc giải bài toán logarit rời rạc, với $x_1 \in_R \mathbb{Z}_q$ ::

$$|Pr(D(g^{x_1}) = 1)| < \mu(|q|) \tag{1.5.7}$$

Theo giả thuyết Diffie-Hellman quyết định [11], Hệ mật mã Elgamal là an toàn về mặt ngữ nghĩa, với $x_1, x_2, x_3 \in_R \mathbb{Z}_q$ ta có:

$$|Pr(t \leftarrow \{g^{x_1}, g^{x_2}, g^{x_3}\}, D(t) = 1) - Pr(t \leftarrow \{g^{x_1}, g^{x_2}, g^{x_1 x_2}\}, D(t) = 1)| < \mu(|q|) \tag{1.5.8}$$

Hơn nữa, hệ mật mã Elgamal có tính chất đồng cấu và là một hệ mật mã ngưỡng [10], đây là các tính chất quan trọng được sử dụng trong thiết kế các giao thức SMC. Hệ mật mã Elgamal bao gồm ba thuật toán cụ thể như Thuật toán **1**. Trong đó: KG là thuật toán sinh khoá, thuật toán mã hóa $E(m,X)$ với m là bản rõ và X là khóa công khai, và thuật toán giải mã $D(C,x)$ trong đó C là bản mã và x là khóa bí mật.

Mã hóa Elgamal được chứng minh là an toàn trước CPA, không an toàn trước CCA2 và được phỏng đoán là an toàn trước CCA1, nhưng chưa có bằng chứng chính thức. [81]

Thuật toán 1: Lược đồ mã hoá Elgamal

KG: Chọn một khoá bí mật $x \in_R \mathbb{Z}_q$ và tính khoá công khai $X = g^x$.
$E(m)$: Giả sử $m \in \mathbb{G}$, chọn $r \in_R \mathbb{Z}_q$, kết quả bản mã là

$$C = (C_1, C_2) = (g^r, m.X^r)$$

$D(c)$: Bản rõ được tính như sau:

$$m = C_2.C_1^{-x}$$

• *Hệ mật mã đường cong Elliptic*

Chọn một đường cong Elliptic $\mathbb{E}(\mathbb{Z}_d)$ với điểm vô cực O và d là một số nguyên tố lớn, G là một điểm cơ sở của đường cong $\mathbb{E}$ với bậc d ($d.G = O$). Tất cả các tính toán được thực hiện trong $\mathbb{Z}_{p_e}$, với p_e là một số nguyên tố lớn. Các tham số $(\mathbb{E}, G, d, p_e)$ được biết bởi tất cả các bên tham gia.

Giả sử $P, Q, R \in \mathbb{E}, k \in \mathbb{Z}_d$, các phép toán trên đường cong Elliptic bao gồm:

- Phép cộng điểm: $P + Q = R$.

- Phép nhân một số nguyên với một điểm: $k \cdot P = \underbrace{P + P + \cdots + P}_{k \text{ lần } P} = R$.

Một số tính chất của đường cong Elliptic:

- Cộng với điểm vô cực: $P + O = P$.

- Phần tử nghịch đảo: $P + (-P) = O$ với $P = (x_P, y_P)$, $-P = (x_P, -y_P) = P = (x_P, p - y_P \, mod p)$.

Hệ mật mã đường cong Elliptic (Elliptic curve cryptosystem, viết tắt ECC) dựa trên bài toán logarit rời rạc tổng quát. Sự an toàn của mật mã đường cong elliptic dựa trên độ khó của bài toán logarit rời rạc của đường cong elliptic [40], với $x_1 \in_R \mathbb{Z}_d$:

$$|Pr(D(x_1.G) = 1)| < \mu(|d|) \tag{1.5.9}$$

Theo giả thuyết Diffie-Hellman quyết định [11], Hệ mật mã đường cong Elliptic là an toàn về mặt ngữ nghĩa, với $x_1, x_2, x_3 \in_R \mathbb{Z}_d$ ta có:

$$|Pr(t \leftarrow \{x_1.G, x_2.G, x_3.G\}, D(t) = 1) - Pr(t \leftarrow \{x_1.G, x_2.G, x_1.x_2.G\}, D(t) = 1)| < \mu(|d|) \tag{1.5.10}$$

Hơn nữa, hệ mật mã đường cong Elliptic cũng có tính chất đồng cấu và là một hệ mật mã ngưỡng, đây là các tính chất quan trọng được sử dụng trong thiết kế các giao thức SMC. Hệ mật mã đường cong Elliptic bao gồm ba thuật toán cụ thể được thể hiện trong Thuật toán **2**. Trong đó: KG là thuật toán sinh khoá, thuật toán mã hóa $E(P_m, Y)$

với P_m là điểm của bản rõ m và Y là khóa công khai, và thuật toán giải mã $D(C,x)$ trong đó C là bản mã và x là khóa bí mật.

Thuật toán 2: Lược đồ mã hoá đường cong Elliptic

KG: Chọn một khoá bí mật $x \in_R \mathbb{Z}_d$ và tính khoá công khai $Y = x.G$.
$E(P_m)$: Giả sử $P_m \in \mathbb{E}$, chọn $r \in_R \mathbb{Z}_d$, kết quả bản mã là

$$C = (C_1, C_2) = (r.G, P_m + r.Y)$$

$D(c)$: Bản rõ được tính như sau:

$$P_m = C_2 - x.C_1$$

Từ đó tính được m.

1.6. Các nghiên cứu liên quan

Các nghiên cứu về hệ gợi ý trong nước đã tiến xa, với sự kết hợp của các công nghệ mới như học sâu và xử lý đồ thị, cùng với việc chú trọng đến trải nghiệm người dùng và tính minh bạch. Những tiến bộ này không chỉ nâng cao hiệu suất của hệ thống mà còn mở rộng phạm vi ứng dụng, đáp ứng nhu cầu ngày càng đa dạng của người dùng [10, 51, 56, 73]. Tuy nhiên theo hiểu biết của tác giả, hiện nay trong nước chưa có công trình nghiên cứu về đảm bảo tính riêng tư cho RS, nếu không kể tới các công bố của bản thân và nhóm nghiên cứu. Trong khi đó đã có khá nhiều công trình nghiên cứu trên thế giới trong lĩnh vực này. Phần này sẽ phân tích, đánh giá những ưu nhược điểm của các giải pháp hệ tư vấn có đảm bảo tính riêng tư sử dụng phương pháp dựa trên mật mã hiện có.

Đầu tiên, luận án xem xét giải pháp dựa trên đồng cấu mà trong đó các tác giả đã sử dụng hệ mật mã Paillier để đảm bảo tính riêng tư của người dùng trong thuật toán CF dựa trên mục tin trong tài liệu [71]. Trong giải pháp này, hệ tư vấn hoạt động dựa trên dữ liệu xếp hạng của người dùng, dữ liệu này được đảm bảo tính riêng tư và được lưu trữ bởi mỗi người dùng. Những gửi dùng sẽ mã hoá các giá trị riêng tư của mình bằng khoá công khai của máy chủ tin cậy và tương tác với những người dùng khác để tính bản mã tổ hợp, sau đó sẽ tương tác với máy chủ tin cậy để giải mã các bản mã đó.

Tiếp theo, giải pháp PPRS được đề xuất trong các tài liệu [2, 19, 72, 83] cũng hoạt động dựa trên mô hình dữ liệu phân tán đầy đủ, trong đó mỗi người dùng sẽ chịu trách nhiệm lưu trữ dữ liệu xếp hạng riêng tư của mình. Tuy nhiên, những giải pháp này đòi hỏi sự hiện diện của một bên trung thực và các bên không thông đồng.

Tác giả Badsha [7] đã đề xuất một hệ thống PPRS cho phép tất cả tính toán được thực hiện theo cách phân phối riêng bằng cách sử dụng mã hóa đồng cấu Elgamal, mà

không ảnh hưởng đến độ chính xác và hiệu quả của đề xuất. Tuy nhiên, nhóm tác giả Mu và các cộng sự [55] đã chỉ ra một nhược điểm tiềm ẩn của giải pháp trong tài liệu [7] và đề xuất một giao thức được sửa đổi lọc cộng tác bảo vệ tính riêng tư nhằm giải quyết lỗ hổng được xác định này. Họ sử dụng một sổ cái công khai lưu trữ các giá trị băm của tất cả người dùng và hiển thị cho tất cả người dùng. Vì vậy, giao thức này rất tốn kém về mặt tính toán.

Trong một trong những nghiên cứu gần đây hơn trong tài liệu [75], các tác giả đưa ra PPRS nâng cao, hệ thống này bảo vệ tính riêng tư của các xếp hạng trước một máy chủ độc hại như được chỉ ra trong tài liệu [55]. Họ thêm khóa tạm để che giá trị xếp hạng. Mặc dù giải pháp sửa đổi này đã nâng cao tính riêng tư mà không ảnh hưởng đến độ chính xác của đề xuất của lược đồ trong tài liệu [7], nhưng chi phí truyền thông tương đương với lược đồ trước đó và đi kèm với độ phức tạp tính toán.

Giải pháp PPRS được đề xuất trong [38] sử dụng mã hóa đồng cấu đầy đủ (FHE) để cải thiện hiệu suất so với các phương pháp trước đó. Họ đã khai thác các thuộc tính SIMD của các mô hình FHE để thực hiện các phép tính hàng loạt trong miền văn bản mã hóa, mang lại những cải tiến hiệu suất đáng kể. Nhóm tác giả Chai và các cộng sự [12] đã đề xuất một hệ tư vấn phân tán đảm bảo quyền riêng tư với HE để tổng hợp dữ liệu. Tương tự, nhóm tác giả Du và các cộng sự [16] cũng cho biết khả năng rò rỉ quyền riêng tư khi sử dụng đào tạo phân tán, tác giả đã nâng cao hơn nữa quyền riêng tư bằng cách kết hợp HE và ngẫu nhiên với nhau.

Tiếp theo đó, các giải pháp PPRS sử dụng phương pháp học kết hợp và mã hoá đồng cấu để đảm bảo tính riêng tư [34, 61]. Hạn chế chính của mã hóa đồng cấu là chi phí tính toán cao, có thể làm giảm hiệu suất tổng thể của hệ thống.

Nhóm tác giả Ammad-Ud-Din [3] sử dụng học kết hợp (Federated learning - FL) cho hệ tư vấn. Mặc dù học kết hợp cho phép đào tạo mô hình mà không cần công khai dữ liệu cục bộ, việc truyền tải các bản cập nhật của mô hình học máy cục bộ sẽ gây ra thêm các vấn đề về quyền riêng tư [62]. Để giải quyết vấn đề này, nhiều nghiên cứu đã tích hợp thêm các kỹ thuật khác [4, 28, 30, 62]. Bên cạnh đó, học kết hợp còn có một nhược điểm đó là chi phí truyền thông cao và độ chính xác của mô hình toàn cục có thể bị giảm nếu dữ liệu phân bố không định danh và độc lập (Non-IID) [79].

Một xu hướng nghiên cứu gần đây cũng đã được thực hiện về việc sử dụng môi trường phân tán để bảo vệ tính riêng tư dữ liệu [9, 43, 59]. Các giải pháp này cung cấp một môi trường an toàn cho người dùng bằng cách sử dụng các tính năng phân tán trong đó dữ liệu của người dùng được sử dụng với quyền của họ. Tuy nhiên, công nghệ blockchain đòi hỏi sức mạnh tính toán cực lớn, điều này vẫn khiến hiệu suất của giải pháp giảm đi đáng kể.

Một nghiên cứu gần đây hơn trong các tài liệu [47, 48] đề xuất một hệ tư vấn có đảm bảo tính riêng tư cũng hoạt động dựa trên xếp hạng của người dùng, trong đó tính đúng đắn của kết quả đề xuất và hiệu quả cao. Tuy nhiên, giải pháp này cần có một bên hỗ trợ gọi là nhà cung cấp dịch vụ an toàn (security provider, Security Server), tất cả các thông tin về xếp hạng và độ tương tự cục bộ của người dùng đều được mã hoá bằng khoá công khai của nhà cung cấp dịch vụ an toàn. Bên này sẽ làm nhiệm vụ hỗ trợ giải mã thông điệp bằng khoá bí mật của mình. Các thực thể đều được giả định không thông đồng với nhau.

Mặc dù đã có các giải pháp PPRS khác nhau dựa trên mật mã được đề xuất, nhưng hoặc là chúng yêu cầu có sự hiện diện của một bên thứ ba tin cậy, hoặc chúng yêu cầu các bên không được thông đồng với nhau như trong các công trình [2,18,19,47,48,58,71,72] hoặc chỉ bảo vệ tính riêng tư trong giai đoạn mô hình hoá và giai đoạn tính toán như trong các công trình [2, 19, 72, 83] hoặc việc triển khai phức tạp, yêu cầu khả năng tính toán lớn [9, 12, 38, 43, 59, 61] hoặc độ chính xác có thể bị giảm [3, 4, 28, 30, 34, 41, 62].

Kỹ thuật tư vấn trong các giải pháp [2, 7, 18, 19, 47, 48, 55, 71, 72, 75, 83] rất gần với thực tế được áp dụng cho hệ thống tư vấn sản phẩm của các sàn thương mại điện tử như Amazon, eBay, Alibaba [39]. Đây là các giải pháp sử dụng bộ lọc dựa trên bộ nhớ sử dụng hành vi trong quá khứ của người dùng để đưa ra tư vấn. Lọc dựa trên bộ nhớ rất dễ thực hiện và yêu cầu tính toán tối thiểu, có khả năng diễn giải, có thể xử lý sự thưa thớt của dữ liệu tốt hơn so với lọc dựa trên mô hình.

Trong đó, các giải pháp trong tài liệu [7, 55, 75] không yêu cầu một bên thứ ba tin cậy, đồng thời dữ liệu nhạy cảm của người dùng luôn chỉ tồn tại dạng mã hoá trong quá trình xử lý và truyền thông. Các giải pháp này đồng thời bảo toàn tính riêng tư trong cả ba giai đoạn với mô hình dữ liệu phân tán đầy đủ, quan trọng hơn, tất cả dữ liệu phản hồi chỉ được lưu giữ ở phía máy khách, nghĩa là nó sẽ không có bất kỳ cơ hội nào được phát hành hoặc truyền đến máy chủ không tin cậy hoặc bất kỳ máy khách độc hại nào khác.

Tuy nhiên trong các giải pháp này vẫn còn tồn tại một số vấn đề: với giải pháp trong tài liệu [7] mặc dù có hiệu suất khá tốt nhưng lại tồn tại các vấn đề an toàn như đã được chỉ ra trong nghiên cứu [55]. Giải pháp trong tài liệu [55] đã giải quyết được vấn đề an toàn của giải pháp trong tài liệu [7] nhưng lại gặp trở ngại về vấn đề hiệu suất khá thấp. Giải pháp gần đầy trong tài liệu [75] cũng giải quyết được vấn đề an toàn của giải pháp trong tài liệu [7] và phần nào giải quyết được vấn đề hiệu suất của giải pháp trong tài liệu [55]. Tuy nhiên để có thể đưa giải pháp vào áp dụng thực tế thì hiệu suất của giải pháp này vẫn cần phải cải tiến hơn nữa. Đánh giá chi tiết về các giải pháp này được trình bày cụ thể trong Bảng 1.2.

Ngoài ra, trong thực tế sẽ tồn tại những trường hợp mà các dữ liệu của một người

Bảng 1.2: So sánh về hiệu suất và an toàn của các giải pháp PPRS điển hình cho mô hình dữ liệu phân tán đầy đủ

	Chi phí tính toán		Chi phí truyền thông (bit)		Tính riêng tư
	Mỗi người dùng	Máy chủ	Mỗi người dùng	Máy chủ	
Giải pháp trong tài liệu [7]	$\frac{m(m+5)}{2}(4T_e + T_m) \approx (480,5m^2 + 2402,5m)T_m$	$\frac{m(m+5)}{2}((3n+1)T_m + T_i + T_D L)$	$\frac{3m(m+5)\|p\|}{2}$	$\frac{nm(m+5)\|p\|}{2}$	- Mô hình bán trung thực
Giải pháp trong tài liệu [55]	$\frac{m(m+5)}{2}(4T_e + nT_m + 2T_H) \approx (482m^2 + 0,5nm^2 + 2410m + 2,5nm)T_m$		$\frac{3m(m+5)\|p\|}{2}(l) + 2\|H\|$	$\frac{3nm(m+5)\|p\|}{2} + 2n\|H\|$	- Mô hình bán trung thực - Máy chủ độc hại như được chỉ ra trong tài liệu [55]
Giải pháp trong tài liệu [75]	$\frac{m(m+5)}{2}(5T_e + 2T_m) \approx (601m^2 + 3005m)T_m$		$\frac{3m(m+5)\|p\|}{2}$	$\frac{nm(m+5)\|p\|}{2}$	- Mô hình bán trung thực - Máy chủ độc hại như được chỉ ra trong tài liệu [55]

dùng có thể được phân bố ở nhiều vị trí khác nhau (được gọi là các miền dữ liệu khác nhau), ví dụ như dữ liệu của một bệnh nhân có thể một phần do bệnh nhân nắm giữ, các phần khác do các bệnh viện khác nhau nắm giữ hay dữ liệu mua bán, đánh giá của một người dùng có thể nằm trên các nền tảng mua bán trực tuyến khác nhau,... Việc kết hợp thông tin từ các bên có thể giúp cho chất lượng tư vấn được tốt hơn. Hầu hết các phương pháp hiện tại đều dựa vào giả định lưu trữ tập trung dữ liệu người dùng, điều này chắc chắn gây ra rủi ro đáng kể về rò rỉ tính riêng tư của người dùng. Tuy nhiên, các nghiên cứu về giải pháp PPRS cho các kịch bản này là chưa nhiều:

Các công trình [20, 22] đề xuất chỉ chia sẻ các thông tin của mục dữ liệu thay vì thông tin của người dùng giữa miền nguồn và miền đích để tránh rò rỉ thông tin liên quan đến người dùng, phương pháp này trích xuất thông tin hữu ích từ các mục dữ liệu bằng cách tận dụng khả năng biểu diễn mạnh mẽ của mạng nơron. Lưu ý rằng mặc dù các phương pháp này bảo vệ tính riêng tư của người dùng ở một mức độ nào đó, nhưng tất cả chúng đều dựa trên phương pháp học tập trung và dữ liệu vẫn có nguy cơ bị tấn công nếu được lưu trữ ở phía máy chủ [12].

Nhóm tác giả Gao và các cộng sự [21] đề xuất ẩn dữ liệu thực trong miền nguồn với tính riêng tư khác biệt trong giai đoạn đầu tiên bằng cách sử dụng ma trận độ tin cậy để

làm cho dữ liệu nhiễu hiệu quả hơn và sau đó chia sẻ dữ liệu được bảo vệ với miền đích. Một khung học kết hợp bảo vệ tính riêng tư cho RS miền chéo, được gọi là FedCDR được đề xuất trong tài liệu [84], trong đó mô hình tư vấn chung được huấn luyện trên thiết bị cá nhân của mỗi người dùng để lấy các bản nhúng của người dùng và các mục, đồng thời mỗi ứng dụng máy trạm tải các trọng số lên máy chủ trung tâm. Trước khi gửi các trọng số này, ứng dụng máy trạm sẽ sử dụng tính riêng tư khác biệt để bảo vệ tính riêng tư của các trọng số. Sau đó, máy chủ trung tâm sẽ tổng hợp các trọng số và phân phối chúng cho từng máy trạm để cập nhật. Tuy nhiên các giải pháp này phải đánh đổi giữa độ chính xác và tính riêng tư.

Nhóm tác giả trong các tài liệu [57,58] đã đề xuất một framework bảo vệ tính riêng tư cho các RS có sự hợp tác của các miền dữ liệu khác nhau với mục tiêu có thể tạo ra đề xuất tốt hơn. Giải pháp này cũng sử dụng mã hoá đồng cấu một số phần (SWHE) để đảm bảo tính riêng tư với khoá bí mật do máy chủ bảo mật chịu trách nhiệm giữ an toàn. Đây là bên thứ ba đáng tin cậy.

Một nghiên cứu gần đây trong tài liệu [26] đã đề xuất một lược đồ RS giữa các miền có đảm bảo tính riêng tư bằng cách sử dụng mã hoá đồng cấu đối xứng và giao thức tính độ tương tự Jaccard an toàn để tính điểm tương đồng về xã hội và vị trí địa lý giữa những người dùng theo cách được mã hóa. Tuy nhiên giải pháp này vẫn lưu trữ tập trung dữ liệu của người dùng tại cục bộ từng miền.

1.7. Những vấn đề cần được nghiên cứu của Luận án

Trong luận án này hướng tới giải quyết những thách thức chính trong các lĩnh vực khác nhau của hệ tư vấn người dùng trong khi vẫn đảm bảo tính riêng tư của người dùng cho mô hình dữ liệu phân tán.

Mục tiêu về đảm bảo tính riêng tư trong giải pháp hệ tư vấn người dùng đề xuất giống với các giải pháp đã có [7, 55, 75] và bao gồm:

– Máy chủ tư vấn và người dùng khác không thể biết được về sự đánh giá của người dùng (kể cả mức độ đánh giá và trạng thái đánh giá).

– Máy chủ tư vấn không thể biết được liệu người dùng có đánh giá một mục cụ thể hay không.

– Giá trị tư vấn chỉ nên được biết bởi người dùng mục tiêu.

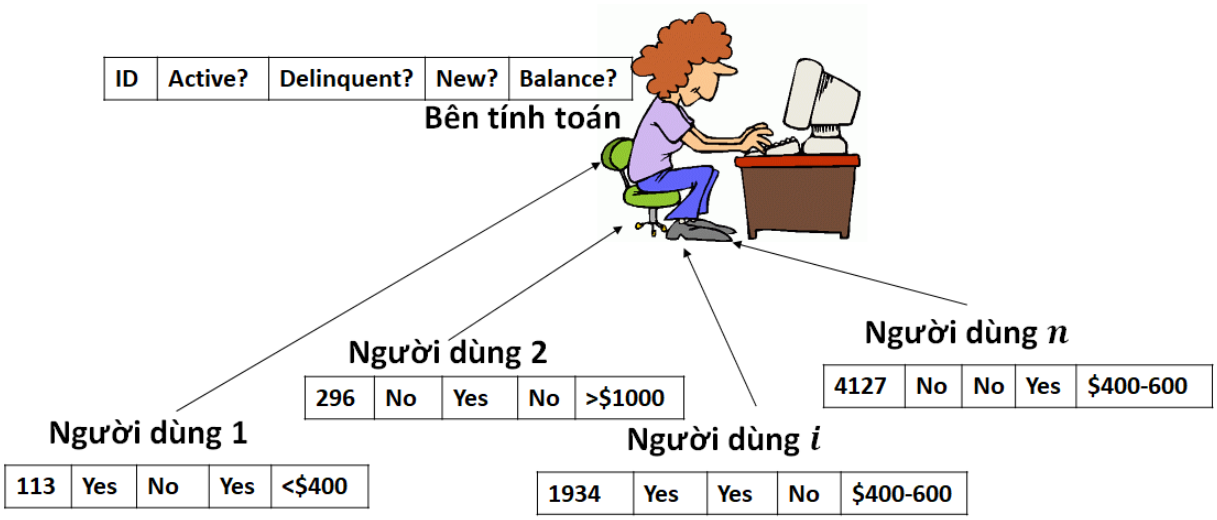
– Máy chủ tư vấn không thể biết được kết quả tư vấn của bất kỳ người dùng nào. – Người dùng không thể có được bất kỳ thông tin nào về dữ liệu của máy chủ tư vấn (các tham số của hệ tư vấn).

Phần tiếp theo sẽ trình bày cụ thể về những vấn đề cần được nghiên cứu trong một số trường hợp cụ thể.

1.7.1. Những vấn đề cần được nghiên cứu trong hệ tư vấn người dùng có đảm bảo tính riêng tư với mô hình dữ liệu phân tán đầy đủ

1.7.1.1. Mô hình dữ liệu phân tán đầy đủ

Mô hình dữ liệu phân tán đầy đủ (Fully Distributed Setting) là phân tán một quan hệ tổng thể n-bộ R thành các quan hệ con $R_1, R_2, ..., R_n$ (mỗi quan hệ con $R_i, i \in [1,n]$ gồm 1 bộ) sao cho quan hệ R có thể được khôi phục lại từ các quan hệ con này bằng phép hợp: $R = R_1 \cup R_2 \cup ... \cup R_n$. Ví dụ về dữ liệu phân tán đầy đủ như Hình 1.5, quan hệ n-bộ R được tách thành n quan hệ con, mỗi quan hệ con này chỉ chứa một bản ghi và được sở hữu bởi mỗi người dùng của hệ thống.



Hình 1.5: Ví dụ về mô hình phân tán đầy đủ

1.7.1.2. Vấn đề cần được nghiên cứu

Từ những đánh giá ở trên có thể thấy mặc dù giải pháp PPRS của nhóm tác giả Badsha và các cộng sự [7] là một giải pháp rất gần với mô hình RS thực tế, có thể dễ dàng triển khai, bảo toàn được tính đúng đắn. Giải pháp này cũng có hiệu suất tốt, dữ liệu riêng tư của các bên tham gia được lưu trữ cục bộ phía các bên tham gia mà không cần phải lưu trữ tập trung, trong suốt quá trình tính toán để xây dựng hệ tư vấn người dùng cũng như quá trình sinh tư vấn dữ liệu riêng tư này đều tồn tại dưới dạng bản mã vì thế đảm bảo tính riêng tư của các bên tham gia trong mô hình bán trung thực.

Tuy nhiên, giải pháp này tồn tại vấn đề an toàn trong trường hợp máy chủ tư vấn có hành vi độc hại như được chỉ ra trong tài liệu [55]. Các giải pháp cải tiến của [7], các tác

giả trong công trình [55] đã giải quyết được vấn đề an toàn này nhưng có hiệu suất thấp hơn so với giải pháp ban đầu. Để có thể áp dụng thực tế thì cần tìm ra giải pháp vừa giải quyết được vấn đề an toàn như được chỉ ra trong tài liệu [55] vừa có hiệu suất tốt hơn giải pháp cải tiến trong tài liệu [75].

Trong hệ thống sẽ sử dụng các xếp hạng của người dùng đối với các sản phẩm để đánh giá mức độ tương tự giữa những sản phẩm cho mục tạo ra các khuyến nghị cho người dùng theo hai cách: dựa trên nội dung (CBF) và dựa trên lọc cộng tác (CF) theo hai công thức (1.7.11) và (1.7.12) tương ứng.

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

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

Trong đó:

m là số lượng mục tin trong hệ thống.

$P_{i,k}$ biểu thị dự đoán xếp hạng cho mục tin thứ k của người dùng U_i, $k = \{1, 2, ..., m\}$ là thứ tự mục tin mà người dùng mục tiêu đã yêu cầu sinh tư vấn.

$r_{i,j}$ là xếp hạng do người dùng U_i cung cấp trên mục tin i_j.

$f_{i,j}$ là trạng thái xếp hạng do người dùng U_i trên mục tin i_j, nếu người dùng đã xếp hạng cho mục tin thì giá trị trạng thái này nhận giá trị là 1, ngược lại sẽ nhận giá trị là 0.

R_k, R_j là trung bình xếp hạng của những người dùng đã xếp hạng đối với mục tin i_k và i_j tương ứng, được xác định theo công thức sau:

$$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}} \tag{1.7.13}$$

$S(i_j, i_k)$ là độ tương tự cosin giữa hai mục tin i_j và i_k được xác định theo công thức sau:

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

Với n là số lượng người dùng trong hệ thống.

Như vậy, giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng sẽ được chia thành hai giai đoạn: giai đoạn đầu tiên là tính trung bình xếp hạng của những người dùng đã xếp hạng đối với các mục tin và tính độ tương tự giữa các cặp mục tin an toàn, giai đoạn thứ hai là sinh ra các tư vấn cho người dùng an toàn.

Chính vì vậy, luận án tập trung vào vấn đề nâng cao hiệu suất của giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ thông qua việc cải tiến hiệu suất của giai đoạn tính trung bình xếp hạng của các mục tin, độ tương tự giữa các cặp mục tin và giai đoạn sinh tư vấn cho người dùng.

1.7.2. Những vấn đề cần được nghiên cứu trong hệ tư vấn người dùng có đảm bảo tính riêng tư với mô hình dữ liệu phân tán đầy đủ hai bên

Một vấn đề nữa có thể thấy được qua phần đánh giá về các giải pháp hiện có là chưa hề có giải pháp PPRS nào cho kịch bản dữ liệu nhạy cảm của người dùng được phân tán đầy đủ hai bên (2 party fully distributed setting - 2PFD).

1.7.2.1. Mô hình dữ liệu phân tán đầy đủ hai bên

Mô hình 2PFD là một trường hợp đặc biệt của mô hình kết hợp giữa phân tán dọc và phân tán ngang.

Mô hình dữ liệu phân tán ngang là phân tán một quan hệ tổng thể n-bộ R là tách R thành các quan hệ con n-bộ $R_1, R_2, \ldots, R_k$ sao cho quan hệ R có thể được khôi phục lại từ các quan hệ con này bằng phép hợp: $R = R_1 \cup R_2 \cup \cdots \cup R_k$.

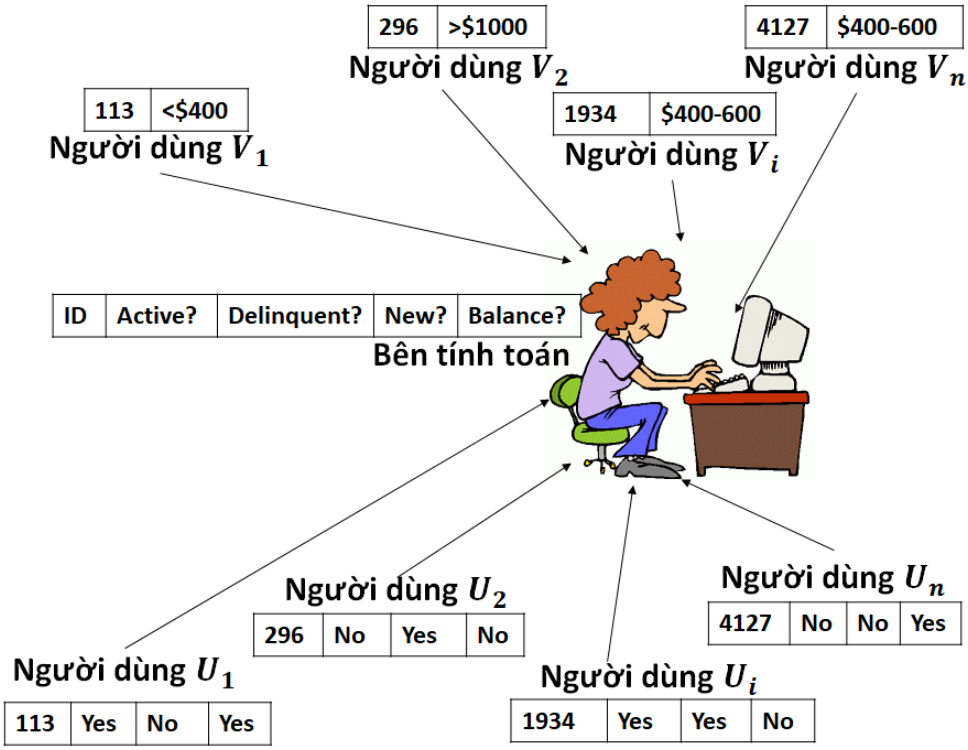
Mô hình dữ liệu phân tán dọc là phân tán một quan hệ tổng thể n-bộ R là tách R thành các quan hệ con $R_1, R_2, \ldots, R_k$ sao cho quan hệ R có thể được khôi phục lại từ các quan hệ con này bằng phép nối: $R = R_1 \bowtie R_2 \bowtie \cdots \bowtie R_k$.

Mô hình 2PFD được xác định lần đầu tiên trong [27], trong đó tập dữ liệu được phân phối trên một số lượng lớn người dùng, mỗi bản ghi thuộc sở hữu của hai người dùng khác nhau, người thứ nhất chỉ biết các giá trị cho một tập hợp con các thuộc tính, trong khi người còn lại biết các giá trị cho các thuộc tính còn lại.

Mô hình 2PFD là phân tán một quan hệ tổng thể n-bộ R thành các quan hệ con $R_{11}, R_{21}, R_{12}, R_{22} ..., R_{1n}, R_{2n}$ (mỗi quan hệ con $R_{1j} \bowtie R_{2j}$, $j \in [1,n]$ gồm 1 bộ) sao cho quan hệ R có thể được khôi phục lại từ các quan hệ con này bằng phép hợp và phép nối: $R = (R_{11} \bowtie R_{21}) \cup (R_{12} \bowtie R_{22}) \cup ... \cup (R_{1n} \bowtie R_{2n})$.

Ví dụ về dữ liệu phân tán đầy đủ hai bên như Hình 1.6, quan hệ n-bộ R được tách thành $2n$ quan hệ con, mỗi quan hệ con $R_{ij} (i \in [1,2], j \in [1,n])$ này chỉ chứa một phần bản ghi và được sở hữu bởi mỗi người dùng của hệ thống.

Các ví dụ về mô hình này có thể thấy: Nhà xã hội học muốn tìm hiểu mối quan hệ giữa hành vi nhân cách của trẻ em và phương pháp nuôi dạy của cha mẹ như trong tài liệu [60, 80]. Nhà xã hội học cung cấp mẫu khảo sát để thu thập thông tin về phương pháp nuôi dạy con cái của cha mẹ và hành vi của con cái họ. Rõ ràng, đây là những thông



Hình 1.6: Ví dụ về mô hình phân tán đầy đủ hai bên

tin khá nhạy cảm, cha mẹ không muốn bộc lộ một cách khách quan những hạn chế của mình trong việc nuôi dạy con cái, đồng thời cũng khó yêu cầu con cái trả lời thẳng thắn, trung thực về hành vi nhân cách của mình. Vì vậy, để có được thông tin chính xác, Nhà nghiên cứu phải đảm bảo nguyên tắc bảo mật thông tin cho từng đối tượng. Trong trường hợp này, mỗi bản ghi dữ liệu thuộc sở hữu riêng của cả cha mẹ và con cái của họ.

Một ví dụ khác là tình huống một nhà nghiên cứu y học cần nghiên cứu mối quan hệ giữa thói quen sinh hoạt, thông tin lâm sàng và một căn bệnh nào đó [6,82]. Một bệnh viện có bộ dữ liệu lâm sàng của bệnh nhân có thể được sử dụng cho mục đích nghiên cứu và thông tin về thói quen sinh hoạt có thể được thu thập bằng một cuộc khảo sát bệnh nhân, tuy nhiên, cả bệnh viện và bệnh nhân đều không sẵn sàng chia sẻ dữ liệu của họ với trung tâm tính toán vì lo ngại vấn đề riêng tư. Trong trường hợp này, mỗi đối tượng dữ liệu bao gồm hai phần: một phần bao gồm thói quen sinh hoạt của bệnh nhân, phần còn lại bao gồm dữ liệu lâm sàng của bệnh nhân này do bệnh viện lưu giữ.

Trong trường hợp một trường học muốn tìm mối liên quan giữa thói quen ăn uống và chỉ số cơ thể của học sinh, sinh viên [25,52]. Trường học có các thông tin dữ liệu nhân khẩu học của học sinh, sinh viên: bao gồm thông tin về các đặc điểm nhân khẩu học của đối tượng nghiên cứu như tuổi, giới tính và các phép đo nhân trắc học (cân nặng, chiều cao và chỉ số cơ thể (body mass index – BMI),. . .). Thông tin về thói quen ăn uống có thể được thu thập bằng một cuộc khảo sát học sinh, sinh viên, tuy nhiên, cả trường học và học sinh, sinh viên đều không sẵn sàng chia sẻ dữ liệu của họ với bên phân tích vì lo

ngại vấn đề riêng tư.

1.7.2.2. Vấn đề cần được nghiên cứu

PPRS trong mô hình 2PFD phức tạp hơn so với các mô hình dữ liệu khác, bởi vì một bộ giá trị ở đây có thể thuộc về hai người dùng khác nhau [27]. Giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng trong mô hình 2PFD về cơ bản là tương tự với mô hình phân tán đầy đủ, tức là sẽ được chia thành hai giai đoạn: Giai đoạn đầu tiên là tính trung bình xếp hạng của những người dùng đã xếp hạng đối với các mục tin và tính độ tương tự giữa các cặp mục tin có bảo mật, giai đoạn thứ hai là sinh ra các tư vấn cho người dùng có bảo mật.

Tuy nhiên trong giai đoạn thứ nhất vì một bộ dữ liệu thuộc về hai người dùng thuộc hai miền dữ liệu khác nhau nên quá trình tính độ tương tự giữa các cặp mục tin thuộc hai miền dữ liệu khác nhau đòi hỏi phải có giải pháp phức tạp hơn so với mô hình phân tán đầy đủ. ***Vì thế, luận án cần đề xuất giao thức tính tần suất có đảm bảo tính riêng tư trong mô hình 2PFD để tính tổng các tích vô hướng của các xếp hạng của người dùng đối với các mục tin thuộc hai miền dữ liệu khác nhau theo công thức 1.7.15. Từ đó ứng dụng giao thức để xây dựng giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình 2PFD sử dụng SMC.***

$$su_{j,k} = \sum_{i=1}^{n} r_{i,j}.r_{i,k} \tag{1.7.15}$$

1.8. Kết luận chương

Kết quả đạt được trong chương này gồm:

- Trình bày tổng quan về hệ tư vấn người dùng, những rủi ro về vi phạm tính riêng tư, các kỹ thuật đảm bảo tính riêng tư cho hệ tư vấn người dùng, tính toán bảo mật.

- Trình bày, tổng hợp và khái quát một số công trình nghiên cứu trước luận án về giải pháp PPRS, trong đó thực hiện việc phân tích rõ về những hạn chế của các giải pháp PPRS sử dụng SMC hiện có.

- Phân tích và đưa ra những vấn đề mở cần giải quyết trong luận án.

CHƯƠNG 2. PHÁT TRIỂN MỘT SỐ GIAO THỨC TÍNH TOÁN BẢO MẬT HIỆU QUẢ

2.1. Mở đầu chương

Trong các bài toán PPRS cho mô hình dữ liệu phân tán đầy đủ và mô hình dữ liệu 2PFD mà luận án cần giải quyết, các giá trị đầu vào thường là các giá trị xếp hạng thể hiện ý kiến đánh giá của người dùng đối với một mục tin, trong đó các giá trị này là các số nguyên không âm khá nhỏ, ví dụ trong một số ứng dụng đánh giá thực tế hiện nay thì thang đánh giá thường có giá trị từ 0 đến 5 như: các hệ thống của amazon, ebay, google, shopee,... . Trong bài toán xem xét của luận án các giá trị bí mật này có giá trị cũng từ 0 đến 5.

Để xây dựng PPRS thì cần phải tính được giá trị trung bình xếp hạng của các mục dữ liệu và độ tương tự của các cặp mục dữ liệu theo công thức 1.7.13 và 1.7.14. Tuy nhiên với việc xây dựng RS cho mô hình dữ liệu 2PFD, một cặp mục dữ liệu thuộc sở hữu bởi hai miền dữ liệu khác nhau, vì thế việc tính độ tương tự của các cặp mục dữ liệu loại này cần phải có phương pháp tính khác biệt so với việc tính đối với các cặp mục cùng thuộc một miền dữ liệu. Do đó luận án cần phát triển giao thức tính các giá trị tần suất cho mô hình 2PFD theo công thức 1.7.15.

Vì thế, chương này trình bày 2 giao thức đề xuất, giao thức đề xuất 1 về cải tiến giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật cho mô hình dữ liệu phân tán đầy đủ và giao thức đề xuất 2 về cải tiến giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trên mô hình 2PFD. Những đề xuất này có liên quan đến Công trình [CT1, CT3, CT4, CT5, CT6]. Trong đó: Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật sẽ được trình bày trong mục 2.2; Giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD sẽ được trình bày trong mục 2.3.

Các giao thức đề xuất trong luận án tập trung đảm bảo tính riêng tư trong mô hình bán trung thực, trong đó các bên tham gia sẽ tuân thủ giao thức nhưng các bên cũng có thể thông đồng với nhau để tìm hiểu các thông tin nhạy cảm của các bên tham gia khác. Tất cả người dùng có một kênh được xác thực riêng tư với trung tâm tính toán. Trung tâm tính toán làm nhiệm vụ tổng hợp các thông điệp nhận được từ các bên tham gia và tính toán để thu được giá trị cần tính. Trung tâm tính toán sẽ không thông đồng với các bên tham gia để tìm hiểu các thông tin nhạy cảm của các bên tham gia.

Các tham số của các hệ mật được sử dụng trong luận án được giả định là thoả mãn yêu cầu sao cho bài toán logarit rời rạc trên trường hữu hạn và trên đường cong Elliptic

là bài toán khó.

2.2. Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật

2.2.1. Ý tưởng

Mô hình bài toán gồm n bên tham gia $U = U_1, U_2, ..., U_n$, mỗi bên U_i nắm giữ m giá trị bí mật là số nguyên không âm nhỏ hoặc vừa $r_{i,j}$ $(0 \leqslant j < m)$, nếu $r_{i,j}$#0 thì $f_{i,j} = 1$, ngược lại $f_{i,j} = 0$. Một trung tâm tính toán cần tính các giá trị trung bình của các giá trị bí mật $r_{i,j}$#0 mà các bên này nắm giữ theo công thức (1.7.13) mà không tiết lộ $r_{i,j}$, $f_{i,j}$ và tính các giá trị độ tương tự của cặp mục tin dựa trên độ tương tự cosin như công thức (1.7.14) mà không tiết lộ $r_{i,j}$, $r_{i,k}$.

2.2.1.1. Tính trung bình có bảo mật

Bài toán tính trung bình có bảo mật (Secure multiparty Average computation - SMAC) đã thu hút được rất nhiều sự quan tâm của giới nghiên cứu. Có hai cách để xây dựng một giao thức SMAC: phương pháp mã hóa và phương pháp thống kê. Trong phần này, các giao thức SMAC điển hình theo phương pháp mã hoá được xem xét toàn diện về cả thuộc tính bảo mật và hiệu quả.

Phần này phân tích về các giao thức SMAC phổ biến nhất liên quan đến luận án, bao gồm: giao thức SMAC hai bên [32] được sử dụng trong đảm bảo tính riêng tư cho bài toán phân cụm, giao thức SMAC trong mô hình phi tập trung [86] áp dụng cho bài toán đồng thuận trung bình có đảm bảo tính riêng tư có các ứng dụng quan trọng trong cân bằng tải động và điều khiển hợp tác các nhóm phương tiện, giao thức SMAC cho bài toán hệ tư vấn người dùng có đảm bảo tính riêng tư [7], giao thức SMAC [55] cải tiến giao thức đề xuất trong tài liệu [7] nhằm loại bỏ rủi ro an toàn khi máy chủ có hành vi độc hại được mô tả trong tài liệu [55], và đặc biệt là giao thức cải tiến do nhóm tác giả Verma và các cộng sự [75] đề xuất nhằm nâng cao hiệu suất hoạt động của giao thức ban đầu [55] mà vẫn duy trì được mức độ an toàn.

Đầu tiên, trong tài liệu [32] đề xuất hai giao thức bảo vệ tính riêng tư cho tính trung bình có trọng số (Weighted Average Problem viết tắt WAP) hiệu quả hơn các giao thức tiêu chuẩn. Giao thức đầu tiên dựa trên đánh giá đa thức không tiết lộ thông tin và giao thức thứ hai dựa trên mã hóa đồng cấu. Cả hai giao thức này đều cho kết quả đầu ra chính xác và an toàn trong mô hình bán trung thực. Hiệu năng của các giao thức này còn thấp, hơn nữa mới dừng lại ở việc tính toán giữa hai bên, việc mở rộng tính toán bảo mật nhiều bên sẽ dẫn tới hiệu suất rất thấp và có thể không thực tế.

Giao thức SMAC thứ hai [86] được áp dụng cho mô hình phi tập trung. Mục tiêu

của giao thức là tính giá trị trung bình trạng thái ban đầu của các nút trong mạng như trong công thức đồng thời bảo vệ tính riêng tư của các nút tuân theo giao thức. Trong giao thức sử dụng hệ mật mã Paillier để mã hóa các thông điệp được trao đổi giữa các nút. Trong giao thức các bên phải trao đổi nhiều lần và trực tiếp với nhau, do đó với những ứng dụng có sự hiện diện của một bên đóng vai trò là bên hỗ trợ tính toán thì không hiệu quả. Mà trong thực tế có khá nhiều các ứng dụng có sự hiện diện của trung tâm tính toán như vậy, ví dụ như hệ tư vấn người dùng, học máy, khai phá dữ liệu,...

Các giao thức còn lại [7, 55, 75] được sử dụng cho mô hình dữ liệu phân tán nhưng có sự hiện diện của bên hỗ trợ tính toán (máy chủ, trung tâm tính toán,...) trong bài toán bảo vệ tính riêng tư cho hệ tư vấn người dùng. Giả sử có n người dùng được ký hiệu $U = U_1, U_2, ..., U_n$ đã tham gia với máy chủ và có xếp hạng là r_i tương ứng với giá trị cờ f_i (1 nếu có xếp hạng, ngược lại là 0). Mục tiêu của máy chủ là tính giá trị trung bình như công thức (2.2.1) mà vẫn đảm bảo tính riêng tư của các giá trị r_i và trạng thái xếp hạng của người dùng f_i.

$$R = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} f_i} \tag{2.2.1}$$

Trong giao thức của nhóm tác giả Badsha và các cộng sự [7] sử dụng hệ mật Elgamal để che dấu những dữ liệu riêng tư. Giao thức này đảm bảo được tính đúng đắn của kết quả đầu ra, an toàn trong mô hình bán trung thực. Giao thức được so sánh với ba giao thức nổi tiếng: Giao thức dựa trên nhiễu loạn ngẫu nhiên được đề xuất trong tài liệu [64], Giao thức dựa trên đồng cấu mà trong đó sử dụng hệ mật mã Paillier để đảm bảo tính riêng tư của người dùng [35], Giao thức cuối cùng [71] cũng dựa trên hệ mật mã Paillier của cùng một tác giả trong công trình [35], được cải thiện về mặt tính toán và chi phí truyền thông so với giao thức [35]. Giao thức [7] cũng hoạt động tốt hơn giải pháp không dựa trên mật mã, mang lại hiệu quả cao bằng cách giảm chi phí tính toán.

Tuy nhiên, các tác giả trong tài liệu [55] đã chỉ ra hai lỗ hổng an toàn của giao thức trong tài liệu [7] và đề xuất giao thức cải tiến để giảm tăng cường mức độ riêng tư cho người dùng trước lỗ hổng tiềm ẩn này. Hai lỗ hổng tiềm ẩn được chỉ ra như sau: Lỗ hổng tiềm ẩn thứ nhất, máy chủ có thể thực hiện một cuộc tấn công đơn giản bằng cách đi chệch khỏi giao thức ở bước 2 khi tính giá trị trung bình. Máy chủ được yêu cầu truyền phát tích của tất cả các bản mã riêng lẻ cho người dùng và mỗi người dùng giải mã tích này bằng khóa bí mật của họ. Nếu, thay vì gửi tích của các bản mã riêng lẻ, máy chủ gửi cho tất cả người dùng phần đầu tiên của bản mã xếp hạng và cờ của một người dùng cụ thể (ví dụ g^{α_1}), thì giá trị mà người dùng trả về sẽ là $(g^{\alpha_1})^{x_i}$, trong đó α_1 là giá trị được chọn ngẫu nhiên được sử dụng để mã hóa các xếp hạng. Sau đó, máy chủ có thể dễ dàng

giải mã xếp hạng của người dùng bằng cách chia xếp hạng được mã hóa cho:

$$\prod_{i=1}^{n}(g^{\alpha_1})^{x_i} \tag{2.2.2}$$

Lỗ hổng tiềm ẩn thứ hai, máy chủ có thể thêm một người dùng "giả"khác và chọn xếp hạng được mã hóa của người dùng một cách thích hợp để đánh lừa một người dùng cụ thể cung cấp xếp hạng của họ. Để đảm bảo giao thức có thể loại bỏ được hai lỗ hổng này, các tác giả đã sử dụng thêm sổ cái công khai để lưu trữ các bản mã hoá và giá trị băm của người dùng, yêu cầu người dùng thực hiện băm các bản mã và gửi giá trị băm này cho máy chủ trước khi gửi các bản mã, đồng thời máy chủ cũng phải gửi tất cả các bản mã và giá trị băm lên sổ cái công khai để người dùng có thể kiểm tra xem liệu máy chủ có hành vi độc hại không. Các bước bổ sung do người dùng thực hiện trong đề xuất mới rõ ràng đã thêm chi phí tính toán đáng kể. Mỗi thông điệp M_i bao gồm mã hóa đánh giá và gắn cờ cho từng mục. Do đó, việc xác minh $H(M_i)$ cho mỗi người dùng i mất $O(n)$ thời gian, trong đó n là số lượng người dùng. Ngoài ra, mỗi người dùng phải nhân các bản mã riêng lẻ để xác minh giá trị do máy chủ cung cấp. Nhân n bản mã người dùng mất $O(n)$ thời gian. Trong giao thức ban đầu, chi phí tính toán cho người dùng chỉ đơn giản là $O(1)$, vì người dùng chỉ cần mã hóa từng xếp hạng của họ. Chi phí tính toán cho một người dùng không phụ thuộc vào tổng số người dùng trong giao thức gốc. Trong giao thức đã sửa đổi, số lượng người dùng cũng là một yếu tố trong chi phí tính toán của một người dùng. Với một kịch bản có số lượng lớn người dùng, rõ ràng là chi phí tính toán của đề xuất sửa đổi sẽ cao hơn đáng kể so với đề xuất ban đầu. Do đó, tồn tại một sự đánh đổi rõ ràng giữa an toàn và hiệu suất, việc đạt được an toàn cao hơn phải trả giá bằng việc tăng chi phí tính toán cho người dùng, đặc biệt là khi số lượng người dùng tăng lên.

Một giao thức điển hình gần đây do nhóm tác giả Verma và các cộng sự đề xuất trong tài liệu [75], trong đó các tác giả đã phân tích lại các lỗ hổng của giao thức đề xuất trong công trình [7] và giải pháp khắc phục được đề xuất trong công trình [55]. Từ đó nhóm tác giả đã đề xuất giải pháp khắc phục khác dựa trên việc cộng thêm các giá trị mặt nạ ngẫu nhiên vào các giá trị nhạy cảm trước khi thực hiện mã hoá các giá trị này như trong công trình [7], khi những người dùng tham gia vào quá trình giải mã thành phần cũng sẽ thực hiện việc loại bỏ giá trị mặt nạ này để máy chủ có thể thu lại giá trị trung bình ban đầu. Giao thức này không những duy trì tính đúng đắn và an toàn trong mô hình bán trung thực như giao thức gốc [7] mà còn loại bỏ được các lỗ hổng được xác định trong [55]. Hơn nữa giao thức cũng không yêu cầu người dùng hay máy chủ phải thực hiện thêm quá nhiều tính toán như giải pháp trong [55]. Giao thức này có chi phí truyền thông bằng với giao thức gốc [7]. So với giao thức gốc thì giao thức mới này chỉ

yêu cầu mỗi người dùng tính thêm 2 phép nhân modulo, vì thế chi phí tính toán của giao thức mới cao hơn giao thức cũ một chút.

Dựa trên những phân tích ở trên, có thể thấy rằng các giao thức của nhóm tác giả Verma và các cộng sự [75] là giao thức SMAC có mức độ an toàn cao tuy nhiên hiệu suất lại thấp hơn giao thức gốc [7]. So sánh về độ phức tạp tính toán và chi phí truyền thông của ba giao thức điển hình cho mô hình dữ liệu phân tán với sự có mặt của bên hỗ trợ tính toán được thể hiện trong Bảng 2.1.

Bảng 2.1: So sánh về hiệu suất và bảo mật của các giao thức SMAC điển hình

	Chi phí tính toán		Chi phí truyền thông (bit)		Tính riêng tư
	Mỗi người dùng	Máy chủ	Mỗi người dùng	Máy chủ	
Giải pháp trong tài liệu [7]	$(1+8m)T_e + 2mT_m \approx (1922m + 240)T_m$	$n(1+6m)T_m + 2mT_i + 2mT_{DL}$	$(1+6m)\|p\|$	$(1+2m)n\|p$	- Mô hình bán trung thực
Giải pháp trong tài liệu [55]	$(1+8m)T_e + 4mT_H + 2nmT_m \approx (1924m + 2nm + 240)T_m$		$(1+6m)\|p\| + 2m\|H\|$	$(1+6m)n\|p\| + 2mn\|H\|$	- Mô hình bán trung thực - Máy chủ độc hại như được chỉ ra trong tài liệu [55]
Giải pháp trong tài liệu [75]	$(1+10m)T_e + 4mT_m \approx (2404m + 240)T_m$		$(1+6m)\|p\|$	$(1+2m)n\|p\|$	- Mô hình bán trung thực - Máy chủ độc hại như được chỉ ra trong tài liệu [55]

Từ kết quả so sánh về độ phức tạp tính toán giữa các giao thức SMAC điển hình được trình bày ở trên có thể thấy rằng, các giao thức được đề xuất trong [68, 93] là các giao thức SMAC điển hình cho mô hình có sự hiện diện của một bên hỗ trợ tính toán, đảm bảo được tính riêng tư của các bên tham gia ở mức cao, tuy nhiên hiệu suất lại thấp hơn giao thức ban đầu [7], đặc biệt là giao thức trong công trình [55].

Vì vậy, giao thức ban đầu [7] được luận án phân tích để tìm ra hướng cải tiến đảm bảo mức độ an toàn của giao thức [75] và có thể cải tiến hiệu suất của giao thức này. Trước khi thực hiện giao thức này, các bên sẽ thống nhất các tham số của hệ mật mã Elgamal trên trường hữu hạn như trong mục 1.5.2.5.1. Chi tiết về giao thức gốc [7] được trình bày trong Giao thức **2.1**.

Đầu vào:
n: Số lượng người dùng.
U_i: Người dùng thứ i, với $i \in [1,n]$.
m: Số lượng các giá trị trung bình cần tính.
$r_{i,j}$: Các số nguyên không âm nhỏ hoặc vừa, là các giá trị bí mật mà mỗi U_i nắm giữ $(0 \leq j < m)$.
max: Giá trị lớn nhất trong các giá trị bí mật mà các U_i nắm giữ.
Đầu ra: Tại trung tâm tính toán (CC):
$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$

Pha khởi tạo:
- Mỗi U_i thực hiện:
1. $x_i \in_R \mathbb{Z}_q^*; X_i = g^{x_i}$;
2. Gửi X_i cho CC.

- CC thực hiện:
3. $X = \prod_{i=1}^{n} X_i$
4. Gửi X cho tất cả U_i.
5. **Pha 1: Mỗi U_i thực hiện**
6. For $(0 \leq j < m)$
7. $\alpha_{i,j}, \beta_{i,j} \in_R \mathbb{Z}_q^*$
8. $E(g^{r_{i,j}}) = (g^{\alpha_{i,j}}, g^{r_{i,j}} X^{\alpha_{i,j}})$;
9. $f_{i,j} = 0$;
10. if $r_{i,j} \neq 0$ $f_{i,j} = 1$;
11. $E(g^{f_{i,j}}) = (g^{\beta_{i,j}}, g^{f_{i,j}} X^{\beta_{i,j}})$;
12. Gửi $\{E(g^{r_{i,j}}), E(g^{e_{i,j}})\}_{j=0}^{m-1}$ cho CC;

Pha 2: CC thực hiện
13. for $(0 \leq j < m)$
14. $A_{1,j} = \prod_{i=1}^{n} g^{\alpha_{i,j}}$; $A_{2,j} = \prod_{i=1}^{n} g^{\beta_{i,j}}$;
15. $B_{1,j} = \prod_{i=1}^{n} g^{r_{i,j}} X^{\alpha_{i,j}}$;
16. $B_{2,j} = \prod_{i=1}^{n} g^{f_{i,j}} X^{\beta_{i,j}}$;
17. Gửi $\{A_{1,j}, A_{2,j}\}_{j=0}^{m-1}$ cho U_i.

Pha 3: Mỗi U_i thực hiện
18. for $(0 \leq j < m)$
19. $(A_{1,j})^{x_i}$; $(A_{2,j})^{x_i}$;
20. Gửi $\{(A_{1,j})^{x_i}, (A_{2,j})^{x_i}\}_{j=0}^{m-1}$ cho CC.

Pha 4: CC thực hiện
21. for $(0 \leq j < m)$
22. $C_{1,j} = \frac{B_{1,j}}{\prod_{i=1}^{n} (A_{1,j})^{x_i}} = g^{\sum_{i=1}^{n} r_{i,j}}$
23. $C_{2,j} = \frac{B_{2,j}}{\prod_{i=1}^{n} (A_{2,j})^{x_i}} = g^{\sum_{i=1}^{n} e_{i,j}}$
24. $d^{(1)} = \text{Dlog}(p,g,C_{1,j}) = \sum_{i=1}^{n} r_{i,j}$
25. $d^{(2)} = \text{Dlog}(p,g,C_{2,j}) = \sum_{i=1}^{n} f_{i,j}$
26. $R_j = \frac{d^{(1)}}{d^{(2)}} = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$

Giao thức 2.1: Giao thức SMAC được đề xuất trong tài liệu [7]

Giao thức trên có một số vấn đề có thể cải tiến để nâng cao hiệu suất cũng như mức độ an toàn như sau:

(1) Pha 1 trong giao thức của nhóm tác giả Badsha và các cộng sự [7] mặc dù mỗi U_i gửi cho máy chủ đầy đủ cả 2 bản mã $E(g^{r_{i,j}}) = (g^{\alpha_{i,j}}, g^{r_{i,j}}.X^{\alpha_{i,j}}), E(g^{f_{i,j}}) = (g^{\beta_{i,j}}, g^{f_{i,j}}.X^{\beta_{i,j}})$ nhưng ở pha 2 máy chủ chỉ gửi kết quả tính toán phần đầu tiên của các bản mã cho mỗi U_i. Vì thế, có thể cải tiến giao thức ở pha 1 bằng cách mỗi U_i chỉ tính và gửi thành phần đầu tiên của các bản mã cho *CC*.

Ở pha 3 trong giao thức gốc mỗi U_i sẽ thực hiện tính thành phần giải mã tương ứng với người dùng thứ i $\{(A_{1,j})^{x_i}, (A_{2,j})^{x_i}\}$ của các bản mã tổng hợp mà *CC* nắm giữ. Điều này dẫn đến việc *CC* có thể có hành vi độc hại [55]: *CC* không gửi phần đầu tiên của bản mã tổng hợp $(A_{1,j})^{x_i}$ theo qui định của giao thức mà gửi phần đầu tiên

của bản mã thành phần $g^{\alpha_{i,j}}$ của người dùng mục tiêu thứ i nào đó và có thể thu được thông tin về giá trị riêng tư của người dùng này. Để khắc phục lỗ hổng này, giao thức có thể cải tiến như sau: thay vì các bên chỉ tham gia giải mã các bản mã tổng hợp như giao thức gốc thì ở pha 3, mỗi U_i sẽ thực hiện đồng thời quá trình mã hoá và giải mã thành phần tương ứng với người dùng thứ i $\{\frac{g^{r_{i,j}}.X^{\alpha_{i,j}}}{(A_{1,j})^{x_i}}, \frac{g^{f_{i,j}}.X^{\beta_{i,j}}}{(A_{2,j})^{x_i}}\}$ của các bản mã tổng hợp $\{B_{1,j} = \prod_{i=1}^{n} g^{r_{i,j}}.X^{\alpha_{i,j}}, B_{2,j} = \prod_{i=1}^{n} g^{f_{i,j}}.X^{\beta_{i,j}}\}$. Vì thế mỗi U_i sẽ không phải gửi riêng thành phần thứ hai của các bản mã tại pha 1 nữa.

(2) Trong trường hợp các tác vụ tính toán yêu cầu thực thi giao thức SMAC để tính đồng thời m ($m > 2$) giá trị trung bình, biến thể cải tiến ở trên chỉ sử dụng $n_k = \left\lceil \frac{1}{2} + \sqrt{2m + \frac{1}{4}} \right\rceil$ cặp khóa riêng và công khai mà không lo ngại về sự an toàn, trong khi những lượng khóa này của các giao thức điển hình [7, 55, 75] đều là $1 + 2m$. Đặc biệt là đối với các ứng dụng chuyên biệt yêu cầu tính toán đồng thời nhiều giá trị trung bình chỉ trong một vòng tính toán (ví dụ: bài toán PPRS).

(3) Bên cạnh đó, hiệu năng của giao thức gốc còn có thể tăng lên bằng cách ứng dụng các hệ mã trên đường cong elliptic. Điều này sẽ làm cho dung lượng cần trao đổi giữa các bên và chi phí tính toán giảm đi đáng kể.

Chi tiết ý tưởng cải tiến được trình bày trong Bảng 2.2.

Bảng 2.2: Chi tiết cải tiến so với giao thức gốc

	Giao thức gốc	**Đề xuất cải tiến**
Pha khởi tạo	$X_i = g^{x_i}$ $X = \prod_{i=1}^{n} X_i$	$\{KPU_{i,j} = g^{ksu_{i,j}}\}_{j=1}^{n_k}$
Pha 1	$\{E(g^{r_{i,j}}) = (g^{\alpha_{i,j}}, g^{r_{i,j}}.X^{\alpha_{i,j}})\}_{j=1}^{m}$ $\{E(g^{f_{i,j}}) = (g^{\beta_{i,j}}, g^{f_{i,j}}.X^{\beta_{i,j}})\}_{j=1}^{m}$	
Pha 2	$\{A_{1,j} = \prod_{i=1}^{n} g^{\alpha_{i,j}}\}_{j=1}^{m}$ $\{A_{2,j} = \prod_{i=1}^{n} g^{\beta_{i,j}}\}_{j=1}^{m}$ $\{B_{1,j} = \prod_{i=1}^{n} g^{r_{i,j}}.X^{\alpha_{i,j}}\}_{j=1}^{m}$ $\{B_{2,j} = \prod_{i=1}^{n} g^{f_{i,j}}.X^{\beta_{i,j}}\}_{j=1}^{m}$	$\{KP_{i,j} = \prod_{i=1}^{n} KPU_{i,j}\}_{j=1}^{n_k}$
Pha 3	$A_{1,j}^{x_i}, A_{2,j}^{x_i}$	$B_{1,j} = \frac{g^{r_{i,j}}.X^{\alpha_{i,j}}}{(A_{1,j})^{x_i}}$, $B_{2,j} = \frac{g^{f_{i,j}}.X^{\beta_{i,j}}}{(A_{2,j})^{x_i}}$
Pha 4	$C_{1,j} = \frac{B_{1,j}}{\prod_{i=1}^{n} A_{1,j}}$ $C_{2,j} = \frac{B_{2,j}}{\prod_{i=1}^{n} A_{2,j}}$	$C_{1,j} = \prod_{i=1}^{n} B_{1,j}$ $C_{2,j} = \prod_{i=1}^{n} B_{2,j}$

2.2.1.2. Tính độ tương tự có bảo mật

Bài toán tính độ tương tự cosin có đảm bảo tính riêng tư (Secure multiparty similarity computation - SMSC) đã thu hút được rất nhiều sự quan tâm trong giới nghiên cứu. Có hai cách để xây dựng một giao thức SMSC: phương pháp mã hóa và phương pháp thống kê. Trong phần này, các giao thức SMSC điển hình theo phương pháp mã hoá sẽ được xem xét toàn diện về cả thuộc tính bảo mật và hiệu quả.

Có thể kể đến các giao thức trong các công trình [36, 76] được ứng dụng cho bài toán xác thực có đảm bảo tính riêng tư. Tuy nhiên đây là giao thức tính độ tương tự giữa hai vector riêng của hai bên. Trong các tài liệu [37, 66] đã đề xuất giao thức tính độ tương tự giữa hai bên sử dụng hệ mật đồng cấu toàn phần Cheon-Kim-Kim-Song áp dụng cho bài toán xác thực giọng nói và so khớp văn bản có đảm bảo tính riêng tư. Giao thức SMSC cho bài toán khai phá dữ liệu dữ liệu có đảm bảo tính riêng tư được đề xuất trong tài liệu [46] đối với ngữ cảnh hai bên sử dụng hệ mật Pallier. Các giải pháp này đều đề xuất cho mô hình hai bên, mỗi bên sở hữu một vector bí mật vì thế không thể áp dụng vào bài toán của luận án.

Các giao thức SMSC được đề xuất trong các tài liệu [7, 55, 75] cho bài toán hệ tư vấn người dùng có đảm bảo tính riêng tư phù hợp với mô hình bài toán của Luận án tuy nhiên: mặc dù giao thức trong tài liệu [7] có hiệu suất tốt tuy nhiên lại tồn tại lỗ hổng an toàn tiềm ẩn khi máy chủ có hành vi độc hại như đã được chỉ ra trong tài liệu [55]. Giao thức trong tài liệu [55] đã loại bỏ được các lỗ hổng này bằng cách sử dụng sổ cái công khai và giải pháp băm các giá trị bản mã được gửi bởi người dùng, điều đó làm giảm đáng kể hiệu suất. Một giao thức gần đây trong tài liệu [75] được đề xuất nhằm nâng cao hiệu suất hoạt động của giao thức [55] mà vẫn đảm bảo loại bỏ được các lỗ hổng của giao thức trong công trình [7]. Tuy nhiên hiệu suất của giao thức này vẫn chưa cao.

Vì thế, Luận án thiết kế giao thức mới nhằm loại bỏ được các lỗ hổng của giao thức trong công trình [7], đồng thời nâng cao hiệu suất của giao thức trong công trình [75]. Trước khi thực hiện giao thức, các bên sẽ thống nhất các tham số của hệ mật mã Elgamal trên trường hữu hạn như trong mục 1.5.2.5.1.

Trong pha khởi tạo, mỗi U_i sẽ phải chuẩn bị một cặp khoá $(x_i, X_i = g^{x_i})$ và gửi cho CC để thực hiện tính khoá công khai chung $X = \prod_{i=1}^{n} X_i$. Khoá công khai chung này được gửi cho tất cả U_i để dùng cho quá trình mã hoá các giá trị riêng tư của mình. Sau đó ở pha 1, mỗi U_i mã hoá giá trị bình phương, giá trị tích của từng cặp giá trị riêng tư sử dụng khoá công khai chung X và gửi cho CC. CC sẽ thực hiện tính các bản mã tổng hợp và gửi thành phần thứ nhất của các bản mã này cho tất cả U_i ở pha 2. Trong pha 3, mỗi U_i sẽ thực hiện tính các thành phần giải mã tương ứng của mình cho bản

Đầu vào:
n: Số lượng người dùng.
U_i: Người dùng thứ i, với $i \in [1,n]$.
m: Số lượng các giá trị trung bình cần tính.
$r_{i,j}, r_{i,k}$: Các số nguyên không âm nhỏ hoặc vừa, là các giá trị bí mật mà mỗi U_i nắm giữ.
max: Giá trị lớn nhất trong các giá trị bí mật mà các U_i nắm giữ.
Đầu ra: Tại trung tâm tính toán (CC):
$S_{j,k} = \frac{\sum_{i=1}^{n} r_{i,j} r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$

Pha khởi tạo:
1. - *Mỗi U_i thực hiện:*
2. $x_i \in_R \mathbb{Z}_q^*$;
3. $X_i = g^{x_i}$;
4. Gửi X_i cho CC.
- *CC thực hiện:*
5. $X = \prod_{i=1}^{n} X_i$
6. Gửi X cho tất cả U_i.

Pha 1: Mỗi U_i thực hiện
7. $c_{i,1}, c_{i,2}, c_{i,3} \in_R \mathbb{Z}_q^*$;
8. $E(g^{r_{i,j}^2}) = (g^{c_{i,1}}, g^{r_{i,j}^2} \cdot X^{c_{i,1}})$;
9. $E(g^{r_{i,k}^2}) = (g^{c_{i,2}}, g^{r_{i,j}^2} \cdot X^{c_{i,2}})$;
10. $E(g^{r_{i,j} \cdot r_{i,k}}) = (g^{c_{i,3}}, g^{r_{i,j} \cdot r_{i,k}} \cdot X^{c_{i,3}})$;
11. Gửi $\{E(g^{r_{i,j} \cdot r_{i,k}}), E(g^{r_{i,j}^2}), E(g^{r_{i,k}^2})\}$ cho CC;

Pha 2: CC thực hiện
12. $A_1 = \prod_{i=1}^{n} g^{c_{i,1}}$;
13. $A_2 = \prod_{i=1}^{n} g^{c_{i,2}}$;
14. $A_1 = \prod_{i=1}^{n} g^{c_{i,3}}$;
15. $B_1 = \prod_{i=1}^{n} g^{r_{i,j}^2} \cdot X^{c_{i,1}}$
16. $B_2 = \prod_{i=1}^{n} g^{r_{i,k}^2} \cdot X^{c_{i,2}}$
17. $B_3 = \prod_{i=1}^{n} r_{i,j} \cdot r_{i,k} \cdot X^{c_{i,3}}$
18. Gửi $\{A_1, A_2, A_3, B_1, B_2, B_3\}$ cho U_i.

Pha 3: Mỗi U_i thực hiện
19. $(A_1)^{x_i}$;
20. $(A_2)^{x_i}$;
21. $(A_3)^{x_i}$;
22. Gửi $\{(A_1)^{x_i}, (A_2)^{x_i}, (A_3)^{x_i}\}$ cho CC.

Pha 4: CC thực hiện
23. $C_1 = \frac{B_1}{\prod_{i=1}^{n} (A_1)^{x_i}} = g^{\sum_{i=1}^{n} r_{i,j}^2}$;
24. $C_2 = \frac{B_2}{\prod_{i=1}^{n} (A_2)^{x_i}} = g^{\sum_{i=1}^{n} r_{i,k}^2}$;
25. $C_3 = \frac{B_3}{\prod_{i=1}^{n} (A_3)^{x_i}} = g^{\sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}}$;
26. $d_1 = \mathrm{Dlog}(p,g,C_1) = \sum_{i=1}^{n} r_{i,j}^2$;
27. $d_2 = \mathrm{Dlog}(p,g,C_2) = \sum_{i=1}^{n} r_{i,k}^2$;
28. $d_3 = \mathrm{Dlog}(p,g,C_3) = \sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}$;
29. $S_{j,k} = \frac{d_3}{\sqrt{d_1} \cdot \sqrt{d_2}} = \frac{\sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$;

Giao thức 2.2: Giao thức SMSC ban đầu [7]

mã tổng hợp và gửi chúng cho CC. CC sẽ thực hiện giải mã các bản mã tổng hợp, thực hiện tính logarit rời rạc để thu được các giá trị tổng hợp và tính các giá trị độ tương tự $S_{j,k} = \frac{\sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$.

Phân tích giao thức trên và nhận thấy ý tưởng phát triển của giao thức này tương tự với giao thức SMAC [7] chỉ khác về số lượng các thông tin cần tính toán và trao đổi. Ví thế, các vấn đề tồn tại với giao thức này sẽ giống với giao thức SMAC [7].

2.2.1.3. Đề xuất ý tưởng cải tiến

Xem xét trường hợp cần thực hiện tính nhiều giá trị trung bình và độ tương tự của cặp mục tin với tập m ($m \geq 2$) mục tin. Trong các giải pháp trước đây [7, 55, 75] để có thể tính giá trị trung bình xếp hạng và độ tương tự giữa các cặp mục tin của tập mục tin (tức là cần tính tập các giá trị trung bình A như công thức (2.2.3) và tập các giá trị độ tương tự S như công thức (2.2.4)) cần thực thi 01 lần giao thức tính trung bình xếp hạng mục tin và 01 lần giao thức tính độ tương tự giữa các cặp mục tin.

$$A = \{a_1 = \frac{\sum_{i=1}^{n} r_{i,1}}{\sum_{i=1}^{n} f_{i,1}}, \ldots, a_m = \frac{\sum_{i=1}^{n} r_{i,m}}{\sum_{i=1}^{n} f_{i,m}}\} \tag{2.2.3}$$

$$S = \{S_1 = \frac{\sum_{i=1}^{n} r_{i,1} . r_{i,2}}{\sqrt{\sum_{i=1}^{n} r_{i,1}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,2}^2}}, \ldots, S_{\frac{m(m-1)}{2}} = \frac{\sum_{i=1}^{n} r_{i,m-1} . r_{i,m}}{\sqrt{\sum_{i=1}^{n} r_{i,m-1}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,m}^2}}\} \tag{2.2.4}$$

- Có thể thấy trong các giao thức trước đây mỗi giá trị bí mật của người dùng sẽ sử dụng hai cặp khoá riêng và khoá công khai để mã hoá. Do đó, nếu mỗi người dùng sử dụng n_k cặp khoá riêng và khoá công khai thì người dùng đó có thể mã hoá lên tới $C_{n_k}^2 = \frac{n_k(n_k-1)}{2}$ giá trị riêng tư. Vì thế mỗi người dùng chỉ cần chuẩn bị n_k cặp khoá riêng và khoá công khai thay vì $\frac{m(m+5)}{2} + 1$ cặp như trong các giao thức trước đây, trong đó n_k được tính theo công thức sau:

$$n_k = \left\lceil \frac{1}{2} + \sqrt{m(m+5) + \frac{1}{4}} \right\rceil \tag{2.2.5}$$

$$n_s = \frac{m(m+5)}{2} \tag{2.2.6}$$

- CC tổng hợp n_s bản mã của các giá trị tổng từ các bản mã của các bên, sau đó tính bản rõ của n_s bản mã tổng hợp này bằng cách chỉ thực hiện thuật toán vét cạn một lần vì các bài toán logarit rời rạc có cùng không gian nghiệm (tức là $\{0, 1, \ldots, n.max^2\}$, trong đó max là giá trị lớn nhất trong của giá trị xếp hạng mà mỗi người dùng có thể nắm giữ).

Từ các phân tích này, phần tiếp theo giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự hiệu quả được đề xuất dựa trên ý tưởng gồm hai bước chính như sau:

Bước 1: Thiết kế lại giao thức trong [7] bằng cách:

- Ở pha khởi tạo: chỉ yêu cầu mỗi U_i sử dụng n_k cặp khoá, tương ứng CC chỉ tính n_k khoá công khai chung dùng cho quá trình mã hoá các giá trị riêng tư của người dùng thay vì $n_s + 1$ cặp khoá.

- Ở pha 1: yêu cầu mỗi U_i thực hiện mã hoá các giá trị riêng tư của mình bằng khoá

công khai chung và giải mã thành phần sử dụng khoá bí mật tương ứng của mình.

- Ở pha 2: CC chỉ tính các bản mã tổng hợp của các giá trị trung bình, các giá trị độ tương tự và thực hiện tính logarit rời rạc một lần tất cả các giá trị này.

Bước 2: Chuyển đổi giao thức được thiết kế lại ở bước 1 sang một biến thể sử dụng hệ mật mã trên đường cong Elliptic.

Từ các phân tích này, phần tiếp theo trình bày giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật hiệu quả được đề xuất. Đề xuất này được công bố trong các Công trình [CT3], [CT4], [CT6].

2.2.2. Giao thức đề xuất

Giao thức đề xuất cũng giả thiết CC bán tin cậy giống như giao thức gốc [7] và vì thế CC sẽ không thông đồng để tiết lộ xếp hạng của người dùng. Giao thức trong công trình [7] được thiết kế lại theo ý tưởng cải tiến số 1 đã được đề xuất trong mục 2.2.1 và được giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật cải tiến được mô tả như trong Giao thức **2.3**. Giao thức cải tiến cũng sử dụng hệ mật Elgamal trên trường hữu hạn giống với giao thức ban đầu [7], các tham số được mô tả trong mục 1.5.2.5.1.

Các giá trị $\sum_{i=1}^{n} a_{i,j}$ không quá lớn nên việc tính logarit rời rạc không khó.

2.2.3. Phân tích tính đúng đắn

Để cho thấy tính đúng đắn của giao thức đề xuất, luận án chứng minh mệnh đề sau:

Mệnh đề 2.1. *Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật tính đúng các giá trị trung bình xếp hạng của các mục tin* $R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$ *và các giá trị độ tương tự giữa cặp mục tin* $S_{j,k} = \frac{\sum_{i=1}^{n} r_{i,j}.r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2}.\sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$ *với* $0 \leq j < k < m$.

Chứng minh. Ta có $sa_j = Dlog(p, g, max^2, A_j)$ nên:

$$g^{sa_j} = A_j = \prod_{i=1}^{n} (g^{a_{i,j}}.KP_k^{-ksu_{i,t}}.KP_t^{ksu_{i,k}})$$

$$= \frac{\prod_{i=1}^{n} g^{a_{i,j}}.\prod_{i=1}^{n} KP_k^{ksu_{i,k}}}{\prod_{i=1}^{n} KP_t^{ksu_{i,t}}}$$

$$= \frac{\prod_{i=1}^{n} g^{a_{i,j}}.\prod_{i=1}^{n} g^{ksu_{i,k}}.\prod_{i=1}^{n} g^{ksu_{i,t}}}{\prod_{i=1}^{n} g^{ksu_{i,t}}.\prod_{i=1}^{n} g^{ksu_{i,k}}}$$

$$= \prod_{i=1}^{n} g^{a_{i,j}} = g^{\sum_{i=1}^{n} a_{i,j}}$$

Do $0 \leq j < m$ nên $a_{i,j} = r_{i,j}$, vậy:

Đầu vào:
n: Số lượng người dùng.
U_i: Người dùng thứ i, với $i \in [1,n]$.
m: Số lượng các giá trị trung bình cần tính.
RM: Ma trận xếp hạng của những người dùng đối với các mục tin, trong đó:$r_{i,j}$: Các số nguyên không âm nhỏ hoặc vừa, là các giá trị bí mật mà mỗi U_i nắm giữ ($0 \leq j < m$).
max: Giá trị lớn nhất trong các giá trị bí mật mà các U_i nắm giữ.
Đầu ra: Tại trung tâm tính toán (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 $0 \leq j < k < m$.

*** Pha khởi tạo:**
- Mỗi U_i thực hiện:
1. for $(0 \leq j < n_k)$
2. $ksu_{i,j} \in_R \mathbb{Z}_q^*; KPU_{i,j} = g^{ksu_{i,j}}$;
3. Gửi $\{KPU_{i,j}\}_{j=0}^{n_k-1}$ cho CC.
- CC thực hiện:
4. for $(0 \leq j < n_k)$
5. $KP_j = \prod_{i=1}^{n} KPU_{i,j}$;
6. Gửi $\{KP_j\}_{j=0}^{n_k-1}$ cho $\{U_i\}_{i=1}^{n}$;

*** Pha 1: Mỗi U_i thực hiện**
7. for $(0 \leq j < m)$
8. $a_{i,j} = r_{i,j}; f_{i,j} = 0$;
9. if $(r_{i,j} \neq 0)$ $f_{i,j} = 1$;
10. for $(m \leq j < 2m)$
11. $a_{i,j} = f_{i,j-m}$;
12. for $(2m \leq j < 3m)$
13. $a_{i,j} = r_{i,j-2m} \cdot r_{i,j-2m}$;
14. for $(0 \leq t < m-1)$
15. for $(t+1 \leq k < m)$
16. $a_{i,j} = r_{i,t} \cdot r_{i,k}$;
17. j++;
18. j = 0;
19. for $(0 \leq t < n_k - 1)$
20. for $(t+1 \leq k < n_k)$
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.

*** Pha 2: CC thực hiện**
26. for $(0 \leq j < n_s)$
27. $A_j = \prod_{i=1}^{n} AU_{i,j}$;
28. $sa = \mathrm{Dlog}(p, g, \max^2, A)$;
/*Sử dụng thuật toán vét cạn một lần với $A = \{A_j, j \in [1,n_s]\}$, $sa = \{sa_j, j \in [1,n_s]\}$*/.
29. for $(0 \leq j < m)$
30. $R_j = \frac{sa_j}{sa_{j+m}}$;
31. $l = 0$;
32. for $(0 \leq j < m-1)$
33. for $(j+1 \leq k < m)$
34. $S_{j,k} = \frac{sa_{3m+l}}{\sqrt{sa_{2m+j}} \cdot \sqrt{sa_{2m+k}}}$;
35. l++;

Giao thức 2.3: Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật

$g^{sa_j} = g^{\sum_{i=1}^{n} a_{i,j}} = g^{\sum_{i=1}^{n} r_{i,j}}$

Vì thế $sa_j = \sum_{i=1}^{n} r_{i,j}$.

Theo cách tương tự ta có $sa_{j+m} = \sum_{i=1}^{n} f_{i,j}$, do đó

$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$.

Do $0 \leq j < k < m$ nên $a_{2m+j} = r_{i,j}^2$, $a_{2m+k} = r_{i,k}^2$, $a_{3m+l} = r_{i,j}.r_{i,k}$. Vậy $g^{sa_{2m+j}} = g^{\sum_{i=1}^{n} a_{i,2m+j}} = g^{\sum_{i=1}^{n} r_{i,j}^2}$, $g^{sa_{2m+k}} = g^{\sum_{i=1}^{n} a_{i,2m+k}} = g^{\sum_{i=1}^{n} r_{i,k}^2}$, $g^{sa_{3m+l}} = g^{\sum_{i=1}^{n} a_{i,2m+l}} = g^{\sum_{i=1}^{n} r_{i,j} \cdot r_{i,k}}$.
Do đó, ta có:

$$S_{j,k} = \frac{sa_{3m+l}}{\sqrt{sa_{2m+j}} \cdot \sqrt{sa_{2m+k}}} = \frac{\sum_{i=1}^{n} r_{i,j} . r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n} r_{i,k}^2}}.$$

Mệnh đề được chứng minh. □

2.2.4. Phân tích tính riêng tư

Để cho thấy mức độ an toàn của giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật được đề xuất, cần chứng minh mệnh đề sau:

Mệnh đề 2.2. *Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật đảm bảo tính riêng tư của mỗi người dùng trong mô hình bán trung thực trong trường hợp không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng lên đến tối đa $n-2$ người dùng, giao thức vẫn bảo vệ tính riêng tư của những người dùng trung thực.*

Chứng minh. Đầu tiên, luận án cần chứng minh rằng giao thức đề xuất đảm bảo tính riêng tư trong mô hình bán trung thực trong trường hợp không có sự thông đồng giữa các bên. Trong giao thức, mỗi người dùng U_i chỉ gửi các giá trị $AU_{i,j}$ ($j \in [0, n_s)$) theo công thức (2.2.7) trong một luồng giao tiếp tới CC. Chúng ta có thể thấy $AU_{i,j}$ chính là phần đầu tiên của bản mã $(z_{i,j}.KP_t^{ksu_{i,k}}, g^{ksu_{i,k}})$ trong hệ mật Elgamal. Trong đó: $z_{i,j} = g^{a_{i,j}}.KP_k^{-ksu_{i,t}}$ là bản rõ, $ksu_{i,k} \in_R \mathbb{Z}_q^*$ là giá trị ngẫu nhiên được chọn và khoá công khai chung KP_t được sử dụng để mã hoá và không ai có khoá bí mật $\sum_{i=1}^{n} ksu_{i,t}$ tương ứng.

$$AU_{i,j} = g^{a_{i,j}}.KP_k^{-ksu_{i,t}}.KP_t^{ksu_{i,k}} \tag{2.2.7}$$

Vì $ksu_{i,k}$, $ksu_{i,t} \in_R \mathbb{Z}_q^*$, g là phần tử sinh của một nhóm con nguyên tố lớn $\mathbb{Z}_q$ nên với mọi phần tử có dạng $y = g^x \ mod \ p$ với $x \in_R \mathbb{Z}_q$ là duy nhất, thuộc tập $\mathbb{Z}_q$ và đều là phần tử sinh của nhóm này, do đó:

$$Pr(t \leftarrow \{z_{i,j}, AU_{i,j}\}_{i\in[1,n], j\in[0,n_s)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.8}$$

Vì thế bộ mô phỏng có thể chọn bất kỳ tập các giá trị bản mã ngẫu nhiên của hệ mật mã Elgamal ta đều có tập hợp thông điệp đo bộ mô phỏng sinh ra và tập hợp thông điệp do giao thức thực sinh ra đều là các tập phân phối đều rời rạc trên $\mathbb{Z}$ nên:

$$\{M(I, \overline{x_I}, f_I(\bar{x}))\}_{\bar{x}\in(\{0,1\}^*)^n} \stackrel{c}{\equiv} \{VIEW_{A,I}^{\pi}(\bar{x})\}_{\bar{x}\in(\{0,1\}^*)^n} \tag{2.2.9}$$

Theo Định nghĩa 1.5 giao thức đề xuất đảm bảo tính riêng tư của các bên tham gia trong mô hình bán trung thực trường hợp không có sự thông đồng giữa các bên.

Tiếp theo, luận án cần chứng minh giao thức đề xuất bảo vệ tính riêng tư của mỗi người dùng trung thực trước sự thông đồng của lên đến tối đa $n-2$ người dùng thông đồng. Để chứng minh, cần chỉ ra một trình mô phỏng M (còn được gọi là thuật toán thời gian đa thức) mô phỏng chế độ xem chung của những người dùng thông đồng và sau đó trình mô phỏng này được kết hợp với trình mô phỏng cho mã hóa Elgamal để thu được một trình mô phỏng hoàn chỉnh. Để làm như vậy, về cơ bản, cần chỉ ra một thuật toán thời gian đa thức để tính toán chế độ xem chung của những người dùng thông đồng. Mô phỏng dựa trên những gì những người dùng thông đồng đã quan sát được trong quá trình thực hiện giao thức chỉ sử dụng thông tin của người dùng thông đồng và khóa công khai. Thuật toán cho đầu ra là các giá trị mô phỏng cho các bản mã hóa được tạo bởi một trình mô phỏng mã hóa Elgamal.

Không mất tính tổng quát, giả sử U_1 và U_2 không thông đồng và những người dùng U_i là những người dùng thông đồng ($i \in I = \{3,4,\ldots,n\}$). Do đó, thuật toán M chỉ cần mô phỏng tính toán cho các giá trị $AU_{1,j}$ và $AU_{2,j}$ ($j \in [0,n_s)$). Dưới đây là trình bày chi tiết về trình mô phỏng M.

$$AU'_{1,j} = \frac{u_{12}.\prod_{i\in I} KPU_{1,k}^{ksu_{i,t}}}{u_{21}.\prod_{i\in I} KPU_{1,t}^{ksu_{i,k}}} \tag{2.2.10}$$

$$AU'_{2,j} = \frac{u_{21}.\prod_{i\in I} KPU_{2,k}^{ksu_{i,t}}}{u_{12}.\prod_{i\in I} KPU_{2,t}^{ksu_{i,k}}} \tag{2.2.11}$$

Trong đó:

- $(u_{12}, v_{12}) = (g^{a_{1,j}}.g^{ksu_{1,t}.ksu_{2,k}}, g^{ksu_{1,t}})$
- $(u_{21}, v_{21}) = (g^{a_{2,j}}.g^{ksu_{2,t}.ksu_{1,k}}, g^{ksu_{2,t}})$

Vì các giá trị $ksu_{1,t}$, $ksu_{2,t}$, $ksu_{1,k}$, $ksu_{2,k} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(t \leftarrow \{AU'_{1,j}, AU'_{2,j}\}_{j\in[0,n_s)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.12}$$

Kết hợp với công thức (2.2.8) ta có:

$$\{M(I, \overline{x_I}, f_I(\overline{x}))\}_{\overline{x}\in(\{0,1\}^*)^n} \overset{c}{\equiv} \{VIEW_{A,I}^{\pi}(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \tag{2.2.13}$$

Trình mô phỏng M ở trên đã thoả mãn Định nghĩa 1.5, do đó giao thức bảo vệ tính riêng tư của các bên trung thực trước sự thông đồng của lên tới $n-2$ bên thông đồng.

Đối với kịch bản hành vi độc hại của trung tâm tính toán như được đề cập trong công trình [55], giao thức đề xuất vẫn đảm bảo tính riêng tư cho tất cả người dùng trung thực. Không mất tính tổng quát, giả sử rằng CC muốn biết giá trị $r_{1,0}$.

Ở pha khởi tạo, CC sẽ gửi cho $\{U_i\}_{i\in[1,n]}$: $KP_1' = KPU_{1,1} = g^{ksu_{1,1}}$ thay vì $KP_1 = \prod_{i=1}^{n} KPU_{i,1}$ như quy định của giao thức.

Vì $\{ksu_{i,1}\}_{i\in[1,n]} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(D(KP_1) = 1) = Pr(D(KP_1') = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.14}$$

Trong pha 1, mỗi U_i sẽ tính giá trị:

$$AU_{i,0}' = g^{r_{i,0}}.KPU_{1,1}^{-ksu_{i,0}}.KP_0^{ksu_{i,1}} \tag{2.2.15}$$

và gửi cho CC thay vì:

$$AU_{i,0} = g^{r_{i,0}}.KPU_{1}^{-ksu_{i,0}}.KP_0^{ksu_{i,1}} \tag{2.2.16}$$

Vì $\{ksu_{i,0}, ksu_{i,1}\}_{i\in[1,n]} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(t \leftarrow \{AU_{i,0}^{'}\}_{i\in[1,n]}, D(t) = 1) = Pr(t \leftarrow \{AU_{i,0}\}_{i\in[1,n]}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.17}$$

Pha 2: CC thực hiện tính:

$$A_0' = \sum_{i=1}^{n} g^{r_{i,0}}.KPU_{1,1}^{-ksu_{i,0}}.KP_0^{ksu_{i,1}} = g^{\sum_{i=1}^{n} r_{i,0}}.KP_0^{\sum_{i=2}^{n} ksu_{i,1}} \tag{2.2.18}$$

Do $\{ksu_{i,1}\}_{i\in[2,n]} \in_R \mathbb{Z}_q^*$ ta có:

$$Pr(D(A_0') = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.19}$$

Từ công thức (2.2.14), (2.2.17) và (2.2.19) ta có:

$$\{KP_1', AU_{i,0}', A_0'\}_{i\in[1,n]} \overset{c}{\equiv} \{KP_1, AU_{i,0}, A_0\}_{i\in[1,n]} \tag{2.2.20}$$

Do đó giao thức đề xuất bảo vệ được tính riêng tư của các bên tham gia trước hành vi độc hại của trung tâm tính toán như được chỉ ra trong [55].

Theo Định nghĩa 1.5, mệnh đề được chứng minh. □

Từ đó cho thấy giao thức đề xuất có mức độ an toàn cao hơn giao thức gốc và duy trì mức độ an toàn so với giao thức cải tiến [75].

Cuối cùng, mức độ an toàn của giao thức đề xuất được đánh giá trước những tấn công phổ biến.

Trong giao thức đề xuất, các bên tham gia được giả định là tuân thủ giao thức và giao tiếp thông qua một kênh an toàn vì thế các tấn công bên ngoài được loại bỏ.

Đối với các sơ đồ mã hóa khóa công khai, ba mô hình tấn công thường được sử dụng để phân tích tính an toàn là: tấn công bằng văn bản gốc được chọn (CPA), tấn công bằng văn bản mã hóa được chọn không thích ứng (CCA1) và tấn công bằng văn bản mã hóa được chọn thích ứng (CCA2). CCA2 mạnh hơn CCA1 và CCA1 mạnh hơn CPA. Mã hóa Elgamal được chứng minh là an toàn trước CPA, không an toàn trước CCA2 và được phỏng đoán là an toàn trước CCA1, nhưng chưa có bằng chứng chính thức. [81]

Như đã được chứng minh trong Mệnh đề 2.2, giao thức đề xuất là an toàn ngữ nghĩa, tức là giao thức chống lại được tấn công lựa chọn bản rõ (CPA).

Thêm vào đó, trong giao thức đề xuất:

- Các cặp khoá được sử dụng đều là cặp khoá sử dụng một lần duy nhất trong mỗi lần gọi giao thức.

- Không có hoạt động giải mã diễn ra trong giao thức, trung tâm tính toán chỉ đơn giản là tổ hợp các bản mã để thu được kết quả cần tính.

- Giá trị bản rõ trước khi mã hoá đã được bổ sung thêm một lớp mặt nạ là một giá trị được lựa chọn đồng nhất, ngẫu nhiên trong $\mathbb{Z}_q^*$, vì thế cho dù kẻ tấn công có toàn quyền truy cập bộ giải mã với một khoá nhất định (CCA2) thì tính riêng tư của dữ liệu đầu vào của người dùng vẫn được đảm bảo:

Giả sử kẻ tấn công thực hiện CCA2 thành công và tìm được bản rõ cần tìm

$$z = D_{ks_t}(AU_{i,j}) = g^{a_{i,j}}.KP_k^{-ksu_{i,t}} \tag{2.2.21}$$

Do $ksu_{i,t}{}_{i\in[1,n])} \in_R \mathbb{Z}_q^*$ nên ta có:

$$Pr(D(z) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.2.22}$$

Vì vậy giao thức đề xuất an toàn CCA2 với các giả thuyết trên.

2.2.5. Phân tích tính hiệu quả

Trong phần này trình bày về việc so sánh giao thức đề xuất với giao thức gốc [7] và giao thức [75] (được gọi là giải pháp 1 và giải pháp 2 tương ứng). Các giao thức được đánh giá về chi phí truyền thông, chi phí tính. Đối với mỗi giải pháp xem xét độ phức tạp tính toán và chi phí truyền thông của từng giai đoạn. Trong đó n lần lượt là số lượng các bên tham gia tính toán.

Các giao thức sẽ sử dụng hệ mật Elgamal kích thước khoá bí mật 160 bit và khoá công khai 1024 bit [42], vì thế mỗi thông điệp sẽ có kích thước là $|p| = 1024$ bít.

2.2.5.1. Đánh giá về chi phí truyền thông

Chi phí truyền thông được tính toán dựa trên số lượng mục tin và người dùng trong hệ thống.

Trước tiên quá trình thực thi của giải pháp 1 và giải pháp 2 được luận án xem xét. Trong pha khởi tạo, mỗi người dùng sẽ phải gửi 1 thông điệp tới CC và CC cũng phải gửi 1 thông điệp đến tất cả các bên tham gia. Để tính các m giá trị xếp hạng trung bình, mỗi người dùng sẽ phải gửi $4m$ thông điệp ở pha 1 và $2m$ thông điệp ở pha 3, vì thế số lượng thông điệp mà mỗi người dùng cần gửi cho CC là $6m$; CC phải gửi $2m$ thông điệp cho tất cả những người dùng trong hệ thống ở pha 2, vì vậy số lượng thông điệp mà CC phải gửi là $2nm$. Do đó, để tính m giá trị xếp hạng trung bình, mỗi người dùng sẽ phải gửi $6m$ thông điệp và CC phải gửi $2nm$ thông điệp. Tương tự đối với quá trình tính độ tương tự, để tính $\frac{m(m-1)}{2}$ giá trị độ tương tự, mỗi người dùng sẽ phải gửi $\frac{3m^2+3m}{2}$ thông điệp và CC phải gửi $\frac{nm(m+1)}{2}$ thông điệp. Do đó, để tính tất các các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi người dùng phải gửi $6m + \frac{3m^2+3m}{2} + 1 = \frac{3m^2+15m+2}{2}$ thông điệp và CC cần gửi $2nm + \frac{nm(m+1)}{2} + n = \frac{n(m^2+5m+2)}{2}$.

Đối với giải pháp đề xuất, để tính các giá trị các các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi người dùng phải gửi n_k (n_k được tính theo công thức (2.2.5)) thông điệp ở pha 1 và $\frac{m(m+5)}{2}$ thông điệp ở pha 3, CC sẽ phải gửi $n_k n$ thông điệp ở pha 2. Như vậy, mỗi người dùng sẽ phải gửi $n_k + \frac{m(m+5)}{2}$ thông điệp, CC sẽ phải gửi $n_k.n$ thông điệp. Ta có:

$$n_k = \left\lceil \frac{1}{2} + \sqrt{m(m+5) + \frac{1}{4}} \right\rceil < \left\lceil \frac{1}{2} + m + \sqrt{5m} + \frac{1}{2} \right\rceil = m + \sqrt{5m} + 1 \leq 1,5m + 3,5 \tag{2.2.23}$$

Suy ra:

$$n_k + \frac{m(m+5)}{2} < 0,5m^2 + 4m + 3,5 \tag{2.2.24}$$

Bảng 2.3 tóm tắt các chi phí truyền thông để thực hiện tính các giá trị trung bình, tính độ tương tự của người dùng và CC.

Bảng 2.3: Bảng so sánh chi phí truyền thông giữa các giao thức

	Người dùng	Trung tâm tính toán
Giao thức [7,75]	$\|p\|(1,5m^2+7,5m+1)$	$n\|p\|(0,5m^2+2,5m+1)$
Giao thức đề xuất	$\|p\|(n_k+\frac{m(m+5)}{2}) < \|p\|(0,5m^2+4m+3,5)$	$\|p\|.n.n_k < n\|p\|(1,5m+3,5)$

Từ kết quả trên có thể thấy rõ giao thức đề xuất có chi phí truyền thông hiệu quả hơn hai giải pháp còn lại. Nhờ việc thiết kế lại giao thức gốc giúp giảm đáng kể số lượng thông điệp cần trao đổi của các bên: mỗi U_i chỉ cần gửi $n_k+\frac{m(m+5)}{2}$ thông điệp và CC chỉ phải gửi $n.n_k$ thông điệp thay cho $\frac{3m^2+15m+2}{2}$ và $\frac{n(m^2+5m+2)}{2}$ thông điệp của các giao thức điển hình.

2.2.5.2. Đánh giá về chi phí tính toán

Bởi vì chi phí tính toán của các giao thức chủ yếu liên quan đến thời gian thực thi tính toán phép nhân, phép luỹ thừa, phép nghịch đảo modulo của các số nguyên lớn, phép tính logarit rời rạc với số nguyên lớn. Do đó chi phí tính toán của các giao thức được đánh giá thông qua các phép toán này.

Trước tiên quá trình thực thi của giải pháp 1 được xem xét. Trong pha khởi tạo, mỗi U_i sẽ phải thực hiện 1 phép tính luỹ thừa modulo và CC phải thực hiện n phép nhân modulo. Để tính m giá trị xếp hạng trung bình, mỗi U_i sẽ phải thực hiện $8m$ phép luỹ thừa modulo và $2m$ phép nhân modulo và CC phải thực hiện $6mn$ phép nhân modulo, $2m$ phép nghịch đảo modulo, $2m$ phép tính logarit rời rạc. Tương tự đối với quá trình tính độ tương tự, để tính $\frac{m(m-1)}{2}$ giá trị độ tương tự, mỗi U_i sẽ phải thực hiện $2m(m+1)$ phép luỹ thừa modulo, $\frac{m(m+1)}{2}$ phép nhân modulo và CC phải thực hiện $\frac{3nm(m+1)}{2}$ phép nhân modulo, $\frac{m(m+1)}{2}$ phép nghịch đảo modulo, $\frac{m(m+1)}{2}$ phép tính logarit rời rạc. Do đó, để tính tất cả các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi U_i phải thực hiện $1+8m+2m(m+1)=2m^2+10m+1$ phép luỹ thừa modulo, $2m+\frac{m(m+1)}{2}=\frac{m^2+5m}{2}$ phép nhân modulo và CC phải thực hiện $n+6mn+\frac{3nm(m+1)}{2}=n.\frac{3m^2+15m+2}{2}$ phép nhân modulo, $2m+\frac{m(m+1)}{2}=\frac{m(m+5)}{2}$ phép nghịch đảo modulo, $\frac{m^2+5m}{2}$ phép tính logarit rời rạc.

Tiếp theo, quá trình thực thi của giải pháp 2 được xem xét. Giải pháp 2 và giải pháp 1 chỉ khác nhau về chi phí tính toán của người dùng. Để tính m giá trị trung bình, mỗi U_i sẽ phải thực hiện 10m phép luỹ thừa modulo, $4m$ phép nhân modulo. Để tính $\frac{m(m-1)}{2}$ giá

tri độ tương tự, mỗi U_i sẽ phải thực hiện $\frac{5m(m+1)}{2}$ phép luỹ thừa modulo, $m(m+1)$ phép nhân modulo. Vì vậy, để tính tất cả các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi người dùng phải thực hiện $(1+10m+\frac{5m(m+1)}{2} = \frac{5m^2+25m+2}{2}$ phép luỹ thừa modulo, $(4m+m(m+1) = m(m+5))$ phép nhân modulo.

Đối với giải pháp đề xuất, để tính tất cả các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi người dùng phải thực hiện $n_k + \frac{3m(m+5)}{2}$ phép luỹ thừa modulo, $\frac{m(m+5)}{2}$ phép nhân modulo, $\frac{m(m+5)}{2}$ phép nghịch đảo modulo và CC phải thực hiện $n.(n_k + \frac{m(m+5)}{2})$ phép nhân modulo, 1 phép tính logarit rời rạc. So sánh chi tiết về chi phí tính toán giữa các giao thức được trình bày trong Bảng 2.4.

Bảng 2.4: So sánh chi phí tính toán giữa các giao thức

	Người dùng	Trung tâm tính toán
Giao thức [7]	$(2m^2+10m+1)T_e + (0,5m^2 + 2,5m)T_m$ $\approx (480,5m^2 + 2402,5m + 240)T_m$	$n(1,5m^2 + 7,5m + 1)T_m + (0,5m^2 + 2,5m)T_i + (0,5m^2 + 2,5m)T_{DL}$
Giao thức [75]	$(2,5m^2+12,5m+1)T_e + m(m+5)T_m$ $\approx (601m^2+3005m+240)T_m$	$\approx n(7,3m^2 + 36,5m + 1)T_m + (0,5m^2+2,5m)T_{DL}$
Giao thức đề xuất	$(n_k + 1,5m^2 + 7,5m)T_e + (0,5m^2 + 2,5m)T_m + (0,5m^2 + 2,5m)T_i$ $< (1,5m^2 + 9m + 3,5)T_e + (0,5m^2 + 2,5m)T_m + (0,5m^2 + 2,5m)T_i$ $\approx (336,3m^2 + 2191,5m + 840)T_m$	$n(n_k+0,5m^2+2,5m)T_m + T_{DL}$ $< n(0,5m^2+4m+3,5)T_m + T_{DL}$

Từ bảng kết quả này ta có thể thấy chi phí tính toán của giải pháp mới thấp hơn các giao thức còn lại. Kết quả này là do việc thiết kế lại giao thức [7] đã giúp giảm số lượng các giá trị cần tính toán của mỗi U_i trong pha 1 và pha khởi tạo từ $\frac{m^2+5m+2}{2}$ xuống n_k, giảm số lượng các giá trị cần tính toán của CC trong pha 2 và pha khởi tạo từ $\frac{m^2+5m+2}{2}$ xuống n_k, giảm số lượng các phép tính logarit rời rạc từ $\frac{m^2+m}{2}$ xuống còn 1.

2.2.6. Chuyển đổi giao thức đề xuất sử dụng hệ mật đường cong Elliptic

Để tiếp tục nâng cao hiệu suất, giao thức đề xuất được chuyển đổi sang một biến thể sử dụng hệ mật mã trên đường cong Elliptic theo ý tưởng cải tiến số 2 đã được đề

xuất trong Mục 2.2.1. Vì hệ mật mã Elgamal trên trường hữu hạn và trên đường cong Elliptic là tương đồng nên với cùng mức độ an toàn thì tính đúng đắn và tính riêng tư của giao thức khi chuyển đổi sang sử dụng hệ mật đường cong Elliptic sẽ tương đương với giao thức ban đầu sử dụng hệ mật mã Elgamal trên trường hữu hạn. Do đó trong phần này chỉ phân tích về tính hiệu suất của giao thức chuyển đổi này so với giao thức ban đầu sử dụng hệ mật mã Elgamal trên trường hữu hạn được mô tả trong Mục 2.2.2.

Trong đó các tham số hệ thống được thiết lập như mô tả trong Mục 1.5.2.5.2. Chi tiết giao thức tính đồng thời nhiều giá trị trung bình xếp hạng và độ tương tự giữa các mục tin bảo mật sử dụng hệ mật đường cong Elliptic được mô tả như trong Giao thức **2.4**.

Phần tiếp theo sẽ đánh giá mức độ hiệu quả của giao thức này so với giao thức sử dụng hệ mật Elgamal trên trường hữu hạn về chi phí truyền thông, chi phí tính toán. Giao thức dựa trên hệ mật mã đường cong Elliptic sẽ sử dụng đường cong Elliptic BrainpoolP160r1 với kích thước khoá bí mật là 160 bit, khoá công khai là 320 bit [45] (hai hệ mật này có cùng mức an toàn là 80 bit).

Khi xem xét về chi phí truyền thông ta có thể thấy: giao thức đề xuất sử dụng hệ mật Elgamal trên trường hữu hạn và trên đường cong Elliptic có số lượng các thông điệp cần trao đổi giữa các bên tham gia là như nhau. Tuy nhiên thì kích thước của mỗi thông điệp của các hệ mật này là khác nhau: mỗi thông điệp của hệ mật Elgamal trên trường hữu hạn là $|p| = 1024$ bít, mỗi thông điệp của hệ mật Elgamal trên đường cong Elliptic tương ứng với 1 điểm và có kích thước là $2.160 = 320 = \frac{5|p|}{16}$ bít. Tương ứng với chi phí truyền thông của giao thức đề xuất sử dụng hệ mật trên đường cong Elliptic giảm đi $\frac{5|p|}{16}$ lần so với sử dụng hệ mật mã Elgamal trên trường hữu hạn. Chi tiết so sánh được trình bày trong Bảng 2.5.

Tiếp theo, ta có thể thấy để tính các giá trị các các giá trị trung bình xếp hạng và giá trị độ tương tự cho m mục tin thì mỗi U_i phải thực hiện $n_k + \frac{3m(m+5)}{2}$ phép nhân trên đường cong elliptic, $m(m+5)$ phép cộng điểm trên đường cong elliptic và CC phải thực hiện $n.(n_k + \frac{m(m+5)}{2})$ phép cộng điểm trên đường cong elliptic, 1 phép tính logarit rời rạc trên đường cong elliptic. So sánh chi tiết về chi phí tính toán giữa các giao thức được trình bày trong Bảng 2.5.

Từ kết quả so sánh trên cho thấy việc chuyển đổi giao thức đề xuất sang sử dụng hệ mật mã Elgamal trên đường cong Elliptic giúp cải thiện đáng kể về cả chi phí truyền thông và chi phí tính toán của giao thức đối với cả phía người dùng và phía trung tâm tính toán. Ví dụ: chi phí của truyền thông của mỗi người dùng trong giao thức sử dụng hệ mật đường cong Elliptic chỉ bằng khoảng $\frac{1}{3}$ so với giao thức sử dụng hệ mật mã Elgamal.

Đầu vào:
n: Số lượng người dùng.
U_i: Người dùng thứ i, $i \in [1,n]$.
m: Số lượng các mục tin trong hệ thống.
RM: Ma trận xếp hạng của những người dùng đối với các mục tin, trong đó: $r_{i,j}$: số nguyên không âm nhỏ hoặc vừa, là giá trị bí mật thứ j mà mỗi người dùng U_i nắm giữ $(0 \le j < m)$.
max: Giá trị xếp hạng lớn nhất trong thang xếp hạng.
Đầu ra (đối với trung tâm tính toán CC):
$R_j = \frac{\sum_{i=1}^{n} r_{i,j}}{\sum_{i=1}^{n} f_{i,j}}, S_{j,k} = \frac{\sum_{i=1}^{n} r_{i,j}.r_{i,k}}{\sqrt{\sum_{i=1}^{n} r_{i,j}^2}.\sqrt{\sum_{i=1}^{n} r_{i,k}^2}}$ với $0 \le j < k < m$.

*** Pha khởi tạo:**
- Mỗi U_i thực hiện:
1. for $(0 \le j < n_k)$
2. $ksu_{i,j} \in_R \mathbb{Z}_d^*$; $KPU_{i,j} = ksu_{i,j}.G$;
3. Gửi $\{KPU_{i,j}\}_{j=0}^{n_k-1}$ cho CC;
- CC thực hiện:
4. for $(0 \le j < n_k)$
5. $KP_j = \sum_{i=1}^{n} KPU_{i,j}$;
6. Gửi $\{KP_j\}_{j=0}^{n_k-1}$ cho $\{U_i\}_{i=1}^{n}$;

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

*** Pha 2: CC thực hiện:**
27. for $(0 \le j < n_s)$
28. $A_j = \sum_{i=1}^{n} AU_{i,j}$;
29. $sa = Dlog_{EC}(\mathbb{E}, G, max^2, A)$; /*Sử dụng thuật toán vét cạn 1 lần $A = \{A_j, j \in [0,n_s)\}, sa = \{sa_j, j \in [0,n_s)\}$*/
30. for $(0 \le j < m)$
31. $R_j = \frac{sa_j}{sa_{j+m}}$;
32. l=1;
33. for $(0 \le j < m-1)$
34. for $(j+1 \le k < m)$
35. $S_{j,k} = \frac{sa_{3m+l}}{\sqrt{sa_{2m+j}}.\sqrt{sa_{2m+k}}}$;
36. l++;

Giao thức 2.4: Giao thức SMASC() tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật sử dụng hệ mật đường cong Elliptic

2.2.7. Đánh giá về mặt thực nghiệm

Để so sánh hiệu suất, giao thức đề xuất, giao thức gốc [7] và giao thức [75] được thực thi trong cùng một môi trường. Để thuận tiện, giả sử rằng mỗi giá trị đầu vào bí mật của các bên tham gia là 1 và không gian của nghiệm của các bài toán logarit rời rạc là từ 0 đến n. Trong kết quả này đã tính toán thời gian tính toán thực tế cần thiết trong từng pha. Các thực nghiệm được thực thi với số lượng các bên tham gia thay đổi, tức là 2000, 4000, 6000, 8000, 10000 và số lượng các giá trị cần tính là 100.

Bảng 2.5: Bảng so sánh chi phí truyền thông và chi phí tính toán

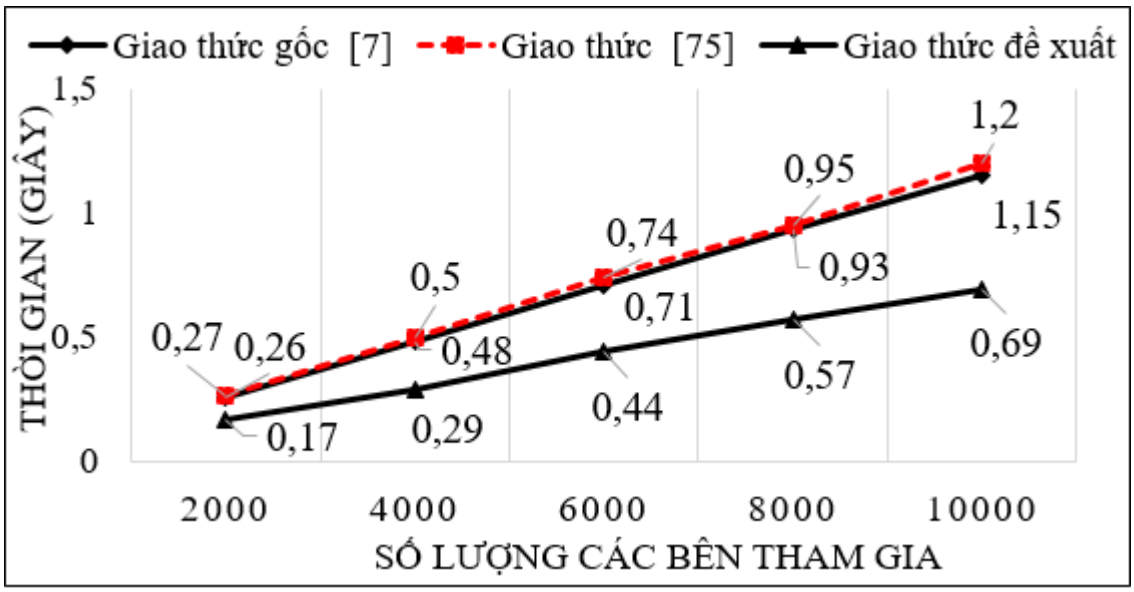
Tiêu chí	Đối tượng	Giao thức đề xuất sử dụng hệ mật Elgamal trên trường hữu hạn	Giao thức đề xuất sử dụng hệ mật Elgamal trên đường cong Elliptic
Chi phí truyền thông	U_i	$\|p\|(n_k+0,5m^2+2,5m)$ $<\|p\|(m^2+4m+3,5)$	$\frac{5\|p\|}{16}(n_k+0,5m^2+2,5m)$ $<\|p\|(0,32m^2+1,25m+1,12)$
	CC	$\|p\|.n.n_k < n\|p\|(1,5m+3,5)$	$\frac{5\|p\|}{16}n.n_k$ $<n\|p\|(0,48m+1,12)$
Chi phí tính toán	U_i	$(n_k+1,5m^2+7,5m)T_e+(0,5m^2+2,5m)T_m+(0,5m^2+2,5m)T_i$ $<(336,3m^2+2191,5m+840)T_m$	$(n_k+1,5m^2+7,5m)T_{mp}+(m^2+5m)T_{ap}$ $<(1,5m^2+9m+3,5)T_{mp}+m(m+5)T_{ap}$ $\approx(43,62m^2+261,6m+101,5)T_m$
	CC	$n(0,5m^2+4m+3,5)T_m+T_{DL}$	$n(n_k+0,5m^2+2,5m)T_{ap}+T_{DL}$ $<n(0,06m^2+0,48m+0,42)T_m+T_{DL}$

Các giao thức cần so sánh sẽ được thực thi bằng ngôn ngữ C# trong môi trường Visual Studio 2019, với sự hỗ trợ của thư viện System.Numerics để so sánh hiệu suất giữa các giao thức này. Các thực nghiệm cùng chạy trên một máy tính xách tay Intel Core i5-4210M CPU 2.60GHz, 8GB RAM. Để giảm thiểu sự sai khác về môi trường thực nghiệm, mỗi giao thức sẽ được thực thi độc lập trong điều kiện không có ứng dụng nào khác được chạy trên máy tính. Thời gian thực thi của các giao thức sẽ được đo bằng cách sử dụng lớp Stopwatch trong thư viện System.Diagnostics.

Để so sánh thời gian chạy của các giao thức, mỗi giao thức được thực thi 10 lần cho mỗi bộ tham số và lấy giá trị trung bình. Luận án giả định rằng tất cả người dùng thực hiện tác vụ của họ cùng một lúc và độ trễ mạng không được tính vào tổng thời gian thực thi. Do đó, chi phí tính toán ở phía người dùng có thể được giảm xuống bằng chi phí tính toán cho một người, chi phí tính toán và giao tiếp ở phía máy chủ tư vấn được tính cho tất cả người dùng tham gia vào hệ thống vì chúng phụ thuộc vào sự cộng tác của tất cả người dùng với CC.

Hình 2.1 thể hiện thời gian tính toán cho từng U_i và CC với số lượng các bên tham gia thay đổi từ 2000 đến 10000 và số lượng các giá trị cần tính là 100.

Từ các kết quả trên có thể thấy giao thức mới có hiệu suất tốt hơn hai giao thức còn lại. Với số lượng người dùng là 10000, giao thức đề xuất chỉ mất thời gian tính toán là 0,69 giây, giao thức gốc là 1,15 giây và giao thức [75] là 1,2 giây. Điều này là do việc thiết kế lại giao thức giúp chi phí tính toán cho việc sinh khoá và tạo khoá công khai chung giảm, chi phí giải bản mã tổ hợp của CC giảm, hơn nữa việc chuyển sang sử dụng



Hình 2.1: So sánh thời gian thực thi giữa các giao thức

hệ mật trên đường cong Elliptic cũng làm giảm chi phí tính toán của giao thức.

Từ Mệnh đề 2.1, Mệnh đề 2.2, đánh giá về mặt lí thuyết và thực nghiệm về hiệu suất giữa các giao thức có thể kết luận: Giao thức đề xuất không những bảo toàn được tính đúng đắn mà còn có mức độ an toàn cao tương đương giao thức trong công trình [75] và hiệu suất cao hơn giao thức gốc [7].

2.2.8. Ví dụ minh hoạ

Để mô tả rõ ràng quá trình tính các giá trị trung bình và độ tương tự có đảm bảo tính riêng tư, luận án sử dụng ma trận xếp hạng kích thước 3×3 như trong Bảng 2.6. Trong đó, người dùng và mục được biểu diễn dưới dạng $U_i = \{U_1, U_2, U_3\}$ và $i_j = \{i_1, i_2, i_3, i_4\}$ tương ứng. Trước khi bắt đầu, các bên tham gia thống nhất các tham số đường cong Elliptic $(\mathbb{E}, G, d, p_e)$.

Bảng 2.6: Ví dụ về ma trận xếp hạng

U_i/i_j	i_1	i_2	i_3
U_1	3	5	0
U_2	0	1	5
U_3	2	3	2
R_j	2,5	3	3,5

Người dùng U_i trong đó $i = \{1,2,3\}$, chọn ngẫu nhiên khóa bí mật của họ $ksu_{i,j}$, trong đó $ksu_{i,j} \in_R \mathbb{Z}^*$, $1 \leq j \leq \left\lceil \frac{1}{2} + \sqrt{3(3+5)+\frac{1}{4}} \right\rceil = 6$ và tính khóa công khai là $KPU_{i,j} = ksu_{i,j}.G$ và gửi cho CC.

CC tính các khoá công khai dùng chung như sau:

$$KP_1 = KPU_{1,1} + KPU_{2,1} + KPU_{3,1}$$

$$KP_2 = KPU_{1,2} + KPU_{2,2} + KPU_{3,2}$$

$$KP_3 = KPU_{1,3} + KPU_{2,3} + KPU_{3,3}$$

$$KP_4 = KPU_{1,4} + KPU_{2,4} + KPU_{3,4}$$

$$KP_5 = KPU_{1,5} + KPU_{2,5} + KPU_{3,5}$$

$$KP_6 = KPU_{1,6} + KPU_{2,6} + KPU_{3,6}.$$

Sau đó, người dùng U_1, U_2 và U_3 gửi các giá trị xếp hạng của họ như sau:

$$AU_1 = \{3.G - ksu_{1,1}.KP_2 + ksu_{1,2}.KP_1, 5.G - ksu_{1,1}.KP_3 + ksu_{1,3}.KP_1,$$
$$0.G - ksu_{1,1}.KP_4 + ksu_{1,4}.KP_1, 1.G - ksu_{1,1}.KP_5 + ksu_{1,5}.KP_1,$$
$$1.G - ksu_{1,1}.KP_6 + ksu_{1,6}.KP_1, 0.G - ksu_{1,2}.KP_3 + ksu_{1,3}.KP_2,$$
$$9.G - ksu_{1,2}.KP_4 + ksu_{1,4}.KP_2, 25.G - ksu_{1,2}.KP_5 + ksu_{1,5}.KP_2,$$
$$0.G - ksu_{1,2}.KP_6 + ksu_{1,6}.KP_2, 15.G - ksu_{1,3}.KP_4 + ksu_{1,4}.KP_3,$$
$$0.G - ksu_{1,3}.KP_5 + ksu_{1,5}.KP_3, 0.G - ksu_{1,3}.KP_6 + ksu_{1,6}.KP_3\}$$

$$AU_2 = \{0.G - ksu_{2,1}.KP_2 + ksu_{2,2}.KP_1, 1.G - ksu_{2,1}.KP_3 + ksu_{2,3}.KP_1,$$
$$5.G - ksu_{2,1}.KP_4 + ksu_{2,4}.KP_1, 0.G - ksu_{2,1}.KP_5 + ksu_{2,5}.KP_1,$$
$$1.G - ksu_{2,1}.KP_6 + ks_2, 6.KP_1, 1.G - ksu_{2,2}.KP_3 + ksu_{2,3}.KP_2,$$
$$0.G - ksu_{2,2}.KP_4 + ksu_{2,4}.KP_2, 1.G - ksu_{2,2}.KP_5 + ksu_{2,5}.KP_2,$$
$$25.G - ksu_{2,2}.KP_6 + ksu_{2,6}.KP_2, 0.G - ksu_{2,3}.KP_4 + ksu_{2,4}.KP_3,$$
$$0.G - ksu_{2,3}.KP_5 + ksu_{2,5}.KP_3, 5.G - ksu_{2,3}.KP_6 + ksu_{2,6}.KP_3\}$$

$$AU_3 = \{2.G - ksu_{3,1}.KP_2 + ksu_{3,2}.KP_1, 3.G - ksu_{3,1}.KP_3 + ksu_{3,3}.KP_1,$$
$$2.G - ksu_{3,1}.KP_4 + ksu_{3,4}.KP_1, 1.G - ksu_{3,1}.KP_5 + ksu_{3,5}.KP_1,$$
$$1.G - ksu_{3,1}.KP_6 + ksu_{3,6}.KP_1, 1.G - ksu_{3,2}.KP_3 + ksu_{3,3}.KP_2,$$
$$4.G - ksu_{3,2}.KP_4 + ksu_{3,4}.KP_2, 9.G - ks_3, 2.KP_5 + ksu_{3,5}.KP_2,$$
$$4.G - ksu_{3,2}.KP_6 + ksu_{3,6}.KP_2, 6.G - ksu_{3,3}.KP_4 + ksu_{3,4}.KP_3,$$
$$4.G - ksu_{3,3}.KP_5 + ksu_{3,5}.KP_3, 6.G - ksu_{3,3}.KP_6 + ksu_{3,6}.KP_3\}$$

Khi CC nhận các thông điệp này, CC bắt đầu tổng hợp các bản mã:

$$A_1 = AU_{1,1} + AU_{2,1} + AU_{3,1} = 3.G + 0.G + 2.G = 5.G$$

$$A_2 = AU_{1,2} + AU_{2,2} + AU_{3,2} = 5.G + 1.G + 3.G = 9.G$$

$$A_3 = AU_{1,3} + AU_{2,3} + AU_{3,3} = 0.G + 5.G + 2.G = 7.G$$

$$A_4 = AU_{1,4} + AU_{2,4} + AU_{3,4} = 1.G + 0.G + 1.G = 2.G$$

$$A_5 = AU_{1,5} + AU_{2,5} + AU_{3,5} = 1.G + 1.G + 1.G = 3.G$$

$A_6 = AU_{1,6} + AU_{2,6} + AU_{3,6} = 0.G + 1.G + 1.G = 2.G$

$A_7 = AU_{1,7} + AU_{2,7} + AU_{3,7} = 9.G + 0.G + 4.G = 13.G$

$A_8 = AU_{1,8} + AU_{2,8} + AU_{3,8} = 25.G + 1.G + 9.G = 35.G$

$A_9 = AU_{1,9} + AU_{2,9} + AU_{3,9} = 0.G + 25.G + 4.G = 29.G$

$A_{10} = AU_{1,10} + AU_{2,10} + AU_{3,10} = 15.G + 0.G + 6.G = 21.G$

$A_{11} = AU_{1,11} + AU_{2,11} + AU_{3,11} = 0.G + 0.G + 4.G = 4.G$

$A_{12} = AU_{1,12} + AU_{2,12} + AU_{3,12} = 0.G + 5.G + 6.G = 11.G$

Sử dụng logarit rời rạc để xác định các bản rõ tương ứng là $sa = \{5,9,7,2,3,2,13, 35,29,21,4,11\}$. Do đó, trung bình được tính:

$R_1 = \frac{sa_1}{sa_4} = \frac{5}{2} = 2,5$, $R_2 = \frac{sa_2}{sa_5} = \frac{9}{3} = 3$, $R_3 = \frac{sa_3}{sa_6} = \frac{7}{2} = 3,5$.

Xếp hạng trung bình của tất cả các mục tin được hiển thị trong Bảng 2.6. Các giá trị độ tương tự được tính bởi CC như sau:

$s(i_1, i_2) = \frac{sa_{10}}{\sqrt{sa_7} \cdot \sqrt{sa_8}} = \frac{21}{\sqrt{13} . \sqrt{35}} = 0,99$

$s(i_1, i_3) = \frac{sa_{11}}{\sqrt{sa_7} \cdot \sqrt{sa_9}} = \frac{4}{\sqrt{13} . \sqrt{29}} = 0,21$

$s(i_2, i_3) = \frac{sa_{12}}{\sqrt{sa_8} \cdot \sqrt{sa_9}} = \frac{11}{\sqrt{35} . \sqrt{29}} = 0,35.$

2.3. Giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD

2.3.1. Ý tưởng

Mô hình 2PFD khá phổ biến trong thực tế và các giao tính tần suất có bảo mật (Secure Multiparty Frequency Computation viết tắt SMFC) trong 2PFD rất quan trọng và có thể được áp dụng cho nhiều tình huống dữ liệu phân tán tương tự khác [27]. Tính tần suất an toàn đóng vai trò vô cùng quan trọng trong các giải pháp đảm bảo tính riêng tư cho khai phá dữ liệu, phân tích dữ liệu và các tính toán trong mô hình phân tán đầy đủ hai bên cũng vậy.

Bài toán tính đồng thời nhiều giá trị tần suất an toàn được định nghĩa là một giao thức có khả năng tính toán đồng thời nhiều giá trị tần suất (được gọi là giao thức SMFC). Trên thực tế, vấn đề này khá phổ biến trong lĩnh vực SMC. Ví dụ: bài toán phân lớp Naive Bayes có đảm bảo tính riêng tư cho cài đặt dữ liệu được phân tán đầy đủ 2 bên [17], hay cho mô hình dữ liệu phân tán dọc [74] yêu cầu trung tâm tính toán hợp tác với các bên tham gia để tính toán an toàn nhiều giá trị tần số được sử dụng để tính toán các xác suất cần thiết. Ngoài ra, có rất nhiều bài toán bảo vệ tính riêng tư thực tế khác liên quan đến

như khai phá cây quyết định, luật kết hợp, phân cụm, phân tích tương quan, v.v., có đảm bảo tính riêng tư.

Trong mô hình 2PFD, một tập dữ liệu (một bảng dữ liệu) chứa n_1 bản ghi, và mỗi bản ghi được biểu diễn bằng các giá trị của các thuộc tính. Tập dữ liệu được phân tán trên $n = 2n_1$ người dùng thuộc hai nhóm $U = \{U_1, U_2, \ldots, U_{n_1}\}$ và $V = \{V_1, V_2, \ldots, V_{n_1}\}$, trong đó mỗi người dùng U_i, V_i giữ các giá trị nhị phân riêng của mình $\{u_{i,0}, u_{i,1}, \ldots, u_{i,m_u-1}\}$, $\{v_{i,0}, v_{i,1}, \ldots, v_{i,m_v-1}\} \in \{0,1\}$. Mỗi cặp người dùng (U_i, V_i) sẽ nắm giữ một bản ghi trong đó người dùng U_i biết các giá trị của một tập con các thuộc tính thích hợp, và người dùng V_i biết các giá trị của các thuộc tính còn lại. Trung tâm tính toán cần tính tập hợp $m_u.m_v$ giá trị tần suất $SU = \{su_j = \sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}\}_{j\in[0,m_u.m_v-1)}$.

Để tính giá trị tần suất đối với các mô hình dữ liệu phân tán khác nhau sẽ có các giao thức tính tần suất có bảo mật tương ứng, có thể kể đến như: Các giao thức đề xuất trong các công trình [15, 23, 70, 72] cho mô hình dữ liệu phân tán hai bên, giao thức đề xuất trong các công trình [54, 70, 78] cho mô hình dữ liệu phân tán đầy đủ, giao thức [77] cho mô hình dữ liệu phân tán bán đầy đủ (semi-fully distributed) và giao thức trong các công trình [27, 49, 50] cho mô hình 2PDF.

Đối với bài toán đặt ra trong luận án có thể được giải quyết thông qua việc thực thi kết hợp nhiều lần giữa các giao thức cho mô hình phân tán hai bên [15, 23, 70, 72] và các giao thức cho mô hình phân tán đầy đủ [54, 70, 78]. Tuy nhiên, cách tiếp cận này đòi hỏi chi phí tính toán và truyền thông cao. Hơn nữa việc kết hợp các giao thức khác nhau nhiều lần sẽ làm cho thiết kế phức tạp và có thể phát sinh các vấn đề liên quan đến tính riêng tư trong quá trình kết hợp. Cách đơn giản để tính các giá trị tần suất này là thực hiện nhiều lần giao thức tính tần suất có đảm bảo tính riêng tư [27, 49, 50] trong mô hình 2PFD. So sánh, đánh giá tổng quan về các giải pháp này được trình bày trong Bảng 2.7.

Bảng 2.7: So sánh các giải pháp tính tần suất cho mô hình 2PFD

Giải pháp	**Mô hình dữ liệu**	**Tính khả thi**	**Mô tả tóm lược giải pháp**	**Số lượng giao tiếp giữa người dùng với CC**	**Số lượng giao tiếp giữa những người dùng**	**Mức độ an toàn**

Giao thức [15, 23, 70, 72]	Mô hình phân tán hai bên	Không	-	-	-	-
Giao thức [54, 70, 78]	Mô hình phân tán đầy đủ	Không	-	-	-	-
Giao thức [77]	Mô hình phân tán bán đầy đủ	Không	-	-	-	-
Kết hợp nhiều lần một trong các giao thức [15, 23, 70, 72] cho mô hình dữ liệu phân tán hai bên và một trong các giao thức [54, 70, 78] cho mô hình dữ liệu phân tán đầy đủ	Mô hình phân tán hai bên + Mô hình phân tán đầy đủ	Có	Để tính 1 giá trị tần suất giữa n_1 cặp người dùng: - Từng cặp người dùng và trung tâm tính toán gọi thực thi giao thức tính tích vô hướng có đảm bảo tính riêng tư. - Sau đó n_1 người dùng trong cùng một miền dữ liệu cùng với trung tâm tính toán thực thi giao thức tính tổng có đảm bảo tính riêng tư.	$4m_u.m_v$	$4m_u.m_v$	- Mô hình bán trung thực - Không có sự thông đồng

Giao thức [27]	Mô hình 2PFD	Có	Thực thi $m_u.m_v$ lần giao thức	$5m_u.m_v$	0	- Mô hình bán trung thực - Tối đa $2n_1 - 2$ người dùng thông đồng với CC
Giao thức [49]						
Giao thức [50]						

Từ bảng so sánh trên có thể thấy trong các giải pháp khả thi, việc thực thi giải pháp đầu thứ ba có độ phức tạp cao và có mức độ an toàn thấp hơn các giải pháp khả thi còn lại [27, 49, 50]. Bên cạnh đó giải pháp này còn yêu cầu từng cặp người dùng giao tiếp trực tiếp với nhau vì thế tính khả thi áp dụng trong thực tế là không cao. Vì thế các giải pháp còn lại tiếp tục được phân tích và đánh giá chi tiết hơn.

Theo hiểu biết của tác giả, giao thức được giới thiệu trong [27] là giao thức đầu tiên cho SMFC trong mô hình 2PFD. Giải pháp này không cần kênh liên lạc giữa những người dùng với nhau. Ngoài ra, giao thức này cung cấp sự riêng tư mạnh mẽ cho mỗi người dùng mà không làm giảm độ chính xác.

Sau đó, nhóm tác giả tiếp tục đề xuất một phương pháp tính toán bảo vệ tính riêng tư mới đối với dữ liệu được phân phối đầy đủ 2 bên [49] dựa trên hệ mật mã Elgamal và chia sẻ bí mật Shamir. Giải pháp này cho phép người khai thác có thể thu được tần số mà không yêu cầu sự tham gia đầy đủ của n cặp người dùng nhưng đổi lại sự phức tạp tính toán gia tăng.

Trong công trình [50], các tác giả đã đề xuất giao thức tính tần suất có đảm bảo tính riêng tư cho mô hình dữ liệu trong đó bản ghi được phân cùng hai phần (Two-Part Partitioned Record Model - TPR), áp dụng giao thức này trong quá trình khai phá luật kết hợp và phân lớp Naïve bayes.

Tuy nhiên, các giao thức này hoặc là có hiệu suất chưa tốt, yêu cầu các bên tương tác qua lại với trung tâm tính toán nhiều lần nên khó áp dụng thực tế. Hơn nữa các giao thức trên đều sử dụng hệ mật Elgamal dẫn đến hiệu suất không cao. Chi tiết so sánh về chi phí tính toán và chi phí truyền thông của các giải pháp điển hình được trình bày cụ thể trong Bảng 2.8.

Từ so sánh trên có thể thấy, giao thức [27] có ưu thế về mặt chi phí truyền thông so với các giao thức còn lại và có chi phí tính toán thấp hơn giao thức [49], cao hơn giao thức [50]. Với mục tiêu duy trì mức độ an toàn cao và cải tiến về mặt hiệu suất, các giao thức [27, 50] được chọn để phân tích sâu hơn nhằm tìm ra cách thức cải tiến. Sau khi phân tích, tác giả nhận thấy khả năng cải tiến của giao thức [27]. Trước khi thực hiện

Bảng 2.8: So sánh về chi phí tính toán và truyền thông giữa các giải pháp tính tần suất trong mô hình 2PFD điển hình

Giao thức	**Chi phí truyền thông**	**Chi phí tính toán**
Giao thức [27]	$20n_1\|p\|$	$2T_i + T_{DL} + (6n+7)T_m + 15T_e$ $\approx (6n_1 + 3630,2)T_m + T_{DL}$
Giao thức [49]	$22n_1\|p\|$	$2T_i + T_{DL} + (6n_1+7)T_m + 18T_e$ $\approx (6n_1 + 4350,2)T_m + T_{DL}$
Giao thức [50]	$24n_1\|p\|$	$2T_i + T_{DL} + (4n+11)T_m + 15T_e$ $\approx (4n_1 + 3634,2)T_m + T_{DL}$

giao thức, các bên sẽ thống nhất các tham số cho hệ mật mã Elgamal như mô tả trong mục 1.5.2.5.1. Chi tiết giao thức này được mô tả trong Giao thức **2.5**

Trong giao thức này, ở pha 3 mỗi U_i sẽ phải gửi 2 thông điệp $K_1^{(i)}$, $K_2^{(i)}$ cho CC, sau đó CC tính giá trị $\frac{\prod_{i=1}^{n} K_1^{(i)}}{\prod_{i=1}^{n} K_2^{(i)}}$ trong pha 4. Vì thế có thể thiết kế lại giao thức bằng cách để mỗi U_i tính $\frac{K_1^{(i)}}{K_2^{(i)}}$ và chỉ gửi giá trị này cho CC thay vì phải gửi 2 thông điệp riêng lẻ như giao thức gốc. Tương ứng ở pha 4, CC cũng chỉ cần tính giá trị $\prod_{i=1}^{n_1} \frac{K_1^{(i)}}{K_2^{(i)}}$. Do đó sẽ giảm tổng chi phí tính toán của giao thức đi $(n_1 - 1)$ phép nhân modulo và chi phí truyền thông giảm n_1 thông điệp.

Một vấn đề nữa đối với giao thức đã đươc thiết kế lại ở pha thứ 3 mỗi U_i sẽ tính giá trị $\frac{K_1^{(i)}}{K_2^{(i)}} = \frac{Q_1^{(i)} (Q_3^{(i)})^{c_i^{(1)}} KP_1{}^{ksu_{i,2}}}{Q_2{}^{(i)}.KP_2{}^{ksu_{i,1}}} = \frac{Q_1^{(i)}}{Q_2^{(i)}}.\frac{(Q_3^{(i)})^{c_i^{(1)}} (KP_1)^{ksu_{i,2}}}{(KP_2)^{ksu_{i,1}}}$, trong đó $\frac{Q_1^{(i)}}{Q_2^{(i)}}$ được tính từ hai giá trị $Q_1^{(i)}, Q_2^{(i)}$ mà mỗi V_i gửi. Vì thế có thể để mỗi V_i tính $\frac{Q_1^{(i)}}{Q_2^{(i)}}$, $Q_3^{(i)}$ và chỉ gửi 2 thông điệp này cho CC thay vì gửi 3 thông điệp như ban đầu.

Bên cạnh đó, trong một số tác vụ yêu cầu phải tính đồng thời nhiều giá trị tần suất thì phải thực hiện gọi nhiều lần giao thức. Để tính mỗi giá trị riêng tư $su_j = \sum_{i=1}^{n_1} u_{i,j}.v_{i,l}$, mỗi U_i, V_i sử dụng hai cặp khoá bí mật và khoá công khai chung. Vì vậy để tính m giá trị tần suất thì sẽ cần sử dụng $2m_u.m_v$ cặp khoá bí mật và khoá công khai chung, tương ứng mỗi người dùng sẽ phải chuẩn bị $2m$ cặp khoá bí mật và khoá công khai. Giả sử các bên sử dụng n_k cặp khoá bí mật và khoá công khai chung thì có thể mã hoá lên tới $C_{n_k}^2$ giá trị bí mật. Vì thế chỉ cần chuẩn bị n_k cặp khoá bí mật và khoá công khai chung, trong đó n_k được tính theo công thức sau:

$$n_k = \left\lceil \frac{1}{2} + \sqrt{2m_u.m_v + \frac{1}{4}} \right\rceil \tag{2.3.25}$$

Như giao thức gốc có thể thấy mỗi cặp khoá bí mật và khoá công khai chung lại

Đầu vào:
n: Số lượng người dùng trong hệ thống, mỗi miền dữ liệu có n_1 người dùng ($n = 2n_1$)
U_i, V_i, với $i \in [1, n_1]$.
u_i: Giá trị nhị phân bí mật mà U_i nắm giữ.
v_i: Giá trị nhị phân bí mật mà V_i nắm giữ.
Đầu ra (đối với CC):
$su = \sum_{i=1}^{n_1} u_i v_i$

*** Pha khởi tạo:**
- Mỗi U_i thực hiện:
1. $x_i \in_R \mathbb{Z}_q^*$;
2. $X_i = g^{x_i}$;
3. for $(1 \le j \le 2)$
4. $\quad ksu_{i,j} \in_R \mathbb{Z}_q^*$;
5. $\quad KPU_{i,j} = g^{ksu_{i,j}}$;
6. Gửi $\{X_i, KPU_{i,1}, KPU_{i,2}\}$ cho CC.
- Mỗi V_i thực hiện:
7. $y_i \in_R \mathbb{Z}_q^*$;
8. $Y_i = g^{y_i}$;
9. for $(1 \le j \le 2)$
10. $\quad ksv_{i,j} \in_R \mathbb{Z}_q^*$;
11. $\quad KPV_{i,j} = g^{ksv_{i,j}}$;
12. Gửi $\{KPV_{i,1}, KPV_{i,2}\}$ cho CC.
- CC thực hiện:
13. for $(1 \le j \le 2)$
14. $\quad KP_j = \prod_{i=1}^{n_1}(KPU_{i,j} \cdot KPV_{i,j})$;
15. Gửi $\{KP_j\}_{j=1}^{2}$ cho $\{U_i, V_i\}_{i\in[1,n]}$.

*** Pha 1: Mỗi U_i thực hiện**
16. $c_i^{(1)} \in_R \mathbb{Z}_q^*$;
17. $C_1^{(i)} = g^{u_i} \cdot X_i^{c_i^{(1)}}$;
18. $C_2^{(i)} = g^{c_i^{(1)}}$;
19. Gửi $\{C_1^{(i)}, C_2^{(i)}\}$ cho CC.

20. *** Pha 2: Mỗi V_i thực hiện**
21. Nhận $\{C_1^{(i)}, C_2^{(i)}, X_i\}$ từ CC.
22. $c_i^{(2)} \in_R \mathbb{Z}_q^*$;
23. $Q_1^{(i)} = (C_1^{(i)})^{v_i} \cdot (KP_1)^{ksv_{i,2}}$;
24. $Q_2^{(i)} = (C_2^{(i)})^{-c_i^{(2)} y_i} \cdot (KP_2)^{ksv_{i,1}}$;
25. $Q_3^{(i)} = Y_i^{c_i^{(2)}} \cdot X_i^{-v_i}$;
26. Gửi $\{Q_1^{(i)}, Q_2^{(i)}, Q_3^{(i)}\}$ cho CC.

*** Pha 3: Mỗi U_i thực hiện**
27. Nhận $\{Q_1^{(i)}, Q_2^{(i)}, Q_3^{(i)}\}$ từ CC.
28. $K_1^{(i)} = Q_1^{(i)} \cdot (Q_3^{(i)})^{c_i^{(1)}} \cdot (KP_1)^{ksu_{i,2}}$;
29. $K_2^{(i)} = Q_2^{(i)} \cdot (KP_2)^{ksu_{i,1}}$;
30. Gửi $\{K_1^{(i)}, K_2^{(i)}\}$ cho CC.

*** Pha 4: CC thực hiện**
31. $A = \frac{\prod_{i=1}^{n_1} K_1^{(i)}}{\prod_{i=1}^{n_1} K_2^{(i)}}$;
32. $su = Dlog(p, g, n_1, A)$;

Giao thức 2.5: Giao thức tính tần suất có bảo mật trong mô hình 2PFD gốc

được hình thành từ hai cặp khoá bí mật và khoá công khai của mỗi U_i, V_i. Do đó, nếu mỗi bên sử dụng n_{k1} cặp khoá bí mật và khoá công khai thì có tạo lên tới $F_{n_{k1}}^2$ cặp khoá bí mật và công khai chung. Vì thế mỗi bên tham gia chỉ cần chuẩn bị n_{k1} cặp khoá bí mật và khoá công khai, trong đó n_{k1} được tính theo công thức sau:

$$n_{k1} = \lceil \sqrt{n_k} \rceil \tag{2.3.26}$$

Hơn nữa, giao thức gốc đang sử dụng hệ mật mã Elgamal sẽ khiến cho hiệu suất của giao thức không cao. Vì vậy ta có thể chuyển đổi giao thức đã thiết kế lại ở trên sang một biến thể mới sử dụng hệ mật mã Elgamal trên đường cong Elliptic.

Phần này đề xuất một giao thức tính đồng thời nhiều giá trị tần suất có đảm bảo tính riêng tư mới trong mô hình 2PFD thông qua việc cải tiến hiệu suất của giao thức ban đầu [27] dựa trên ý tưởng các bước như sau:

Bước 1: Thiết kế lại các bước trong giao thức gốc [27] bằng cách:

- Ở pha khởi tạo: chỉ yêu cầu mỗi U_i, V_i sử dụng $n_{k1}+1$ cặp khoá (n_{k1} được tính theo công thức (2.3.26)), tương ứng CC chỉ tính n_k khoá công khai chung (n_k được tính theo công thức (2.3.25)) dùng cho quá trình mã hoá các giá trị riêng tư của người dùng thay vì $2m+1$ cặp khoá như giao thức gốc.

- Ở pha 1: mỗi U_i mã hoá các giá trị riêng tư sử dụng khoá công khai của mình và gửi các bản mã cho CC.

- Ở pha 2: chỉ yêu cầu mỗi V_i tính $Q_{11}^{(i,j)} = \frac{Q_1^{(i,j)}}{Q_2^{(i,j)}}$, $Q_{21}^{(i,j)} = Q_3^{(i,j)}$, gửi các giá trị này đến CC thay vì tính $Q_1^{(i,j)}$, $Q_2^{(i,j)}$, $Q_3^{(i,j)}$ và gửi các giá trị này đến CC như giao thức gốc.

- Ở pha 3: yêu cầu mỗi U_i tính $\frac{K_1^{(i,j)}}{K_2^{(i,j)}}$ và gửi giá trị duy nhất này đến CC thay vì tính riêng lẻ $K_1^{(i,j)}$, $K_2^{(i,j)}$ và gửi cho CC như giao thức gốc.

- Ở pha 4: CC chỉ thực hiện tổng hợp các bản mã nhận được và thực hiện tính các giá trị logarit rời rạc một lần duy nhất thay vì phải giải mã bản mã tổng và thực hiện tính nhiều lần các giá trị logarit rời rạc như giao thức gốc.

Bước 2: Chuyển đổi giao thức được thiết kế lại sang một biến thể dựa vào hệ mã hóa Elgamal trên đường cong elliptic.

Đề xuất này liên quan đến các Công trình [CT1], [CT5].

2.3.2. Giao thức đề xuất

Giao thức **2.6** mô tả chi tiết giao thức đề xuất thông quan việc thiết kế lại giao thức trong công trình [27] theo ý tưởng cải tiến số 1 được đề xuất trong mục 2.3.1. Giao thức cải tiến cũng sử dụng hệ mật Elgamal trên trường hữu hạn giống với giao thức ban đầu [27], các tham số được mô tả trong mục 1.5.2.5.1.

Các giá trị $\sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}$ không quá lớn nên việc tính logarit rời rạc không khó.

2.3.3. Phân tích tính đúng đắn

Để cho thấy tính đúng đắn của giao thức đề xuất, cần chứng minh mệnh đề sau:

Đầu vào:
n: Số lượng người dùng trong hệ thống, trong đó mỗi miền dữ liệu U_i, V_i có n_1 người dùng, $i \in [1, n_1], n = 2n_1$.
$u_{i,j}$: Các giá trị nhị phân bí mật mà U_i nắm giữ ($0 \le j < m_u$).
$v_{i,j}$: Các giá trị nhị phân bí mật mà V_i nắm giữ ($0 \le j < m_v$), $m_u \le m_v$.
Output (đối với CC):
$SU = \{su_j = \sum_{i=1}^{n_1} u_{i,j\%m_v}.v_{i,\lfloor \frac{j}{m_v} \rfloor}\}_{j \in [0, m_u.m_v - 1)}$

*** Pha khởi tạo:**
- Mỗi U_i thực hiện:
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. Gửi $\{KPU_{i,j}\}_{0 \le j < n_{k1}}, X_i$ cho CC;
- Mỗi V_i thực hiện:
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. Gửi $\{KPV_{i,j}\}_{0 \le j < n_{k1}}$ cho CC;
- CC thực hiện:
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. Gửi $\{KP_j\}_{0 \le j < n_k}$ cho $\{U_i, V_i\}_{1 \le i \le n}$;

*** Pha 1: Mỗi U_i thực hiện**
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. Gửi $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}$ cho CC;

*** Pha 2: Mỗi V_i thực hiện**
20. Nhận $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}, X_i$ từ 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. Gửi $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u.m_v}$ cho CC;

*** Pha 3: Mỗi U_i thực hiện:**
29. Lấy $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u.m_v}$ từ 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. Gửi $\{A_{i,j}\}_{0 \le j < m_u.m_v}$ cho CC;

*** Pha 4: CC thực hiện:**
38. for $(0 \le j < m_u.m_v)$
39. $A_j = \prod_{i=1}^{n_1} A_{i,j}$;
40. $SU = Dlog(p, g, n_1, A)$;
//Sử dụng thuật toán vét cạn 1 lần

Giao thức 2.6: Giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD

Mệnh đề 2.3. *Giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD tính đúng các giá trị tần suất* $su_j = \sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}$.

Chứng minh. Ta có $su_j = Dlog(p, g, n_1, A_j)$, nên:

$$g^{su_j} = A_j = \prod_{i=1}^{n_1}(Q_1^{(i,j)}.(Q_2^{(i,j)})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}.KP_k^{-ksu_{i,\lfloor t/n_{k1} \rfloor}}.KP_t^{ksu_{i,\lfloor k/n_{k1} \rfloor}})$$

$$= \prod_{i=1}^{n_1}((C_1^{(i,\lfloor j/m_v \rfloor)})^{v_{i,j\%m_v}}.KP_t^{ksv_{i,k\%n_{k1}}}.(C_2^{(i,\lfloor j/m_v \rfloor)})^{-c_{i,j}^{(2)}.y_i}.KP_k^{-ksv_{i,t\%n_{k1}}})$$
$$.\prod_{i=1}^{n_1}(Y_i^{c_{i,j}^{(2)}}.X_i^{-v_{i,j\%m_v}})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}.\prod_{i=1}^{n_1}(KP_t)^{ksu_{i,\lfloor k/n_{k1} \rfloor}}.\prod_{i=1}^{n_1}(KP_k)^{-ksu_{i,\lfloor t/n_{k1} \rfloor}}$$

$$= \prod_{i=1}^{n_1}(C_1^{(i,\lfloor j/m_v \rfloor)})^{v_{i,j\%m_v}}.\prod_{i=1}^{n_1}(KP_t)^{ksv_{i,k\%n_{k1}}}.\prod_{i=1}^{n_1}(C_2^{(i,\lfloor j/m_v \rfloor)})^{-c_{i,j}^{(2)}.y_i}.\prod_{i=1}^{n_1}(KP_k)^{-ksv_{i,t\%n_{k1}}}$$
$$.\prod_{i=1}^{n_1}(Y_i^{c_{i,j}^{(2)}}.X_i^{-v_{i,j\%m_v}})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}.\prod_{i=1}^{n_1}(KP_t)^{ksu_{i,\lfloor k/n_{k1} \rfloor}}.\prod_{i=1}^{n_1}(KP_k)^{-ksu_{i,\lfloor t/n_{k1} \rfloor}}$$

$$= \prod_{i=1}^{n_1}(g^{u_{i,\lfloor j/m_v \rfloor}})^{v_{i,j\%m_v}}.\prod_{i=1}^{n_1}(X_i^{c_{i,\lfloor j/m_v \rfloor}^{(1)}})^{v_{i,j\%m_v}}.\prod_{i=1}^{n_1}(KP_t)^{ksv_{i,k\%n_{k1}}+ksu_{i,\lfloor k/n_{k1} \rfloor}}$$
$$.\prod_{i=1}^{n_1}(g^{c_{i,\lfloor j/m_v \rfloor}^{(1)}})^{-c_{i,j}^{(2)}.y_i}.\prod_{i=1}^{n_1}(KP_k)^{-(ksu_{i,t\%n_{k1}}+ksv_{i,\lfloor t/n_{k1} \rfloor})}.\prod_{i=1}^{n_1}(Y_i^{c_{i,j}^{(2)}})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}$$
$$.\prod_{i=1}^{n_1}(X_i^{-v_{i,j\%m_v}})^{c_{i,\lfloor j/m_v \rfloor}^{(1)}}$$

$$= \prod_{i=1}^{n_1} g^{u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}} = g^{\sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}}$$

Vì $g^{su_j} = g^{\sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}}$ nên $su_j = \sum_{i=1}^{n_1} u_{i,\lfloor j/m_v \rfloor}.v_{i,j\%m_v}$.

Chú ý rằng với giá trị của su_j không quá lớn, ta có thể sử dụng thuật toán vét cạn để tìm su_j. Giao thức này vẫn đúng trong trường hợp các giá trị riêng tư mà những người dùng nắm giữ là những số nguyên dương chứ không phải chỉ là giá trị {0,1} miễn là giá trị su_j không quá lớn và nó có thể được giới hạn (ví dụ 10^6). □

2.3.4. Phân tích tính riêng tư

Để cho thấy mức độ an toàn của giao thức đề xuất, cần chứng minh mệnh đề sau:

Mệnh đề 2.4. *Giao thức tính đồng thời nhiều giá trị tần suất có đảm bảo tính riêng tư trong mô hình 2PFD bảo vệ tính riêng tư của mỗi người dùng trung thực trong mô hình bán trung thực. Trong trường hợp có sự thông đồng của các bên tham gia, giao thức bảo vệ tính riêng tư của người dùng trung thực chống lại trung tâm tính toán và tối đa $n-2$ người dùng thông đồng. Trong trường hợp chỉ có hai người dùng trung thực, điều này vẫn chính xác miễn là hai người dùng trung thực không sở hữu các giá trị thuộc tính của cùng một bản ghi.*

Chứng minh. Đầu tiên, giao thức tính đồng thời nhiều giá trị tần suất có bảo mật được chứng minh là đảm bảo tính riêng tư trong mô hình bán trung thực.

Trong pha khởi tạo các bên tham gia chỉ gửi các giá trị $\{KPU_{i,j}, X_i, KPV_{i,j}\}_{1\le i\le n_1, 0\le j<n_{k1}}$ cho CC và CC gửi các giá trị $\{KP_j\}_{0\le j<n_k}$ cho tất cả các bên tham gia. Vì $\{x_i, ksu_{i,j}, ksv_{i,j}\}_{1\le i\le n_1, 0\le j<n_{k1}} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(t \leftarrow \{KPU_{i,j}, X_i, KPV_{i,j}, KP_l\}_{i\in[1,n_1], j\in[0,n_{k1}), l\in[0,n_{k1})}, D(t)=1) = \frac{1}{|\mathbb{Z}_q^*|} \qquad (2.3.27)$$

Trong Pha 1, CC nhận các thông điệp $C_1^{(i,j)}$ và $C_2^{(i,j)}$ của mỗi U_i. Ta có $c_{i,j}^{(1)} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(t \leftarrow \{C_1^{(i,j)}, C_2^{(i,j)}\}_{i\in[1,n_1],j\in[0,m_u)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.3.28}$$

Trong pha 2, các thông điệp $Q_1^{(i,j)}$ và $Q_2^{(i,j)}$ được gửi bởi mỗi V_i, trong đó $\{c_{i,j}^{(2)}, ksu_{i,j}, ksv_{i,j}\}_{1\leq i\leq n_1, 0\leq j<n_k} \in_R \mathbb{Z}_q^*$ nên ta có:

$$Pr(t \leftarrow \{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{i\in[1,n_1],j\in[0,m_u.m_v)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.3.29}$$

Tương tự, trong pha 3, thông điệp A_i được gửi bởi mỗi U_i ta cũng có:

$$Pr(t \leftarrow \{A_{i,j}\}_{i\in[1,n_1],j\in[0,m_u.m_v)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.3.30}$$

Từ công thức 2.3.27-2.3.30 có thể thấy với bất kỳ tập các giá trị bản mã ngẫu nhiên của hệ mật mã Elgamal mà bộ mô phỏng chọn ta đều có tập hợp thông điệp đo bộ mô phỏng sinh ra và tập hợp thông điệp do giao thức thực sinh ra đều là các tập phân phối đều rời rạc trên $\mathbb{Z}_p$ nên:

$$\{M(I, \overline{x_I}, f_I(\overline{x}))\}_{\overline{x}\in(\{0,1\}^*)^n} \stackrel{c}{\equiv} \{VIEW_{A,I}^{\pi}(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \tag{2.3.31}$$

Do đó, giao thức đề xuất bảo vệ tính riêng tư của mỗi người dùng trung thực trong mô hình bán trung thực.

Tiếp theo, giao thức được chứng minh là bảo vệ tính riêng tư của các bên trung thực trước sự thông đồng của CC và tối đa $n-2$ bên thông đồng, trong đó hai người dùng trung thực không sở hữu các giá trị thuộc tính của cùng một bản ghi. Để chứng minh, cần chỉ ra một trình mô phỏng M (còn được gọi là thuật toán thời gian đa thức) mô phỏng chế độ xem chung của CC và những người dùng thông đồng và sau đó trình mô phỏng này được kết hợp với trình mô phỏng cho mã hóa Elgamal để thu được một trình mô phỏng hoàn chỉnh. Để làm như vậy, về cơ bản, luận án chỉ ra một thuật toán thời gian đa thức để tính toán chế độ xem chung của CC và những người dùng thông đồng. Mô phỏng dựa trên những gì CC và những người dùng thông đồng đã quan sát được trong quá trình thực hiện giao thức chỉ sử dụng kết quả su_j, thông tin của người dùng thông đồng và khóa công khai. Thuật toán cho đầu ra là các giá trị mô phỏng cho các bản mã hóa được tạo bởi một trình mô phỏng mã hóa Elgamal. Không mất tính tổng quát, giả sử U_1 và V_2 không thông đồng, những U_i thông đồng ($i \in I_1 = \{2,3,...,n_1\}$) và những V_l thông đồng ($l \in I_2 = \{1,3,4,...,n_1\}$). Do đó, thuật toán M sẽ thực hiện như sau với

$j \in \{0,\ldots,m-1\}$:

- M mô phỏng $C_1^{(1,j)}, C_2^{(1,j)}$ sử dụng các bản mã Elgamal ngẫu nhiên.
- M lấy các bản mã sau làm đầu vào của nó

$$(a_{1,j}, a_{2,j}) = (\alpha_j . g^{-ksv_{2,t\%n_{k1}} . (ksu_{1,\lfloor k/n_{k1}\rfloor} + ksv_{2,k\%n_{k1}})} . (g^{(ksu_{1,\lfloor t/n_{k1}\rfloor} + ksv_{2,t\%n_{k1}})})^{ksv_{2,k\%n_{k1}}} , g^{ksv_{2,k\%n_{k1}}}) \tag{2.3.32}$$

Trong đó $\alpha_j = (C_1^{(2,\lfloor j/m_v\rfloor)})^{v_{2,j\%m_v}} . (C_2^{(2,\lfloor j/m_v\rfloor)})^{-y_2 . c_{2,j}^{(2)}}$ và tính các giá trị sau:

$$(Q_1^{'(2,j)} = a_{1,j} . KPU_{2,t\%n_{k1}}^{-(\sum_{i\in I_1} ksu_{i,\lfloor k/n_{k1}\rfloor} + \sum_{l\in I_2} ksv_{l,k\%n_{k1}})} . KPV_{2,k\%n_{k1}}^{(\sum_{i\in I_1} ksu_{i,\lfloor t/n_{k1}\rfloor} + \sum_{l\in I_2} ksv_{l,t\%n_{k1}})} . g^{-(su_j - \sum_{l=3}^{n_1} u_{l,\lfloor j/m_v\rfloor} v_{l,j\%m_v} - \varepsilon_j v_{1,j\%m_v} - \theta_j u_{2,\lfloor j/m_v\rfloor})} \tag{2.3.33}$$

Trong đó: $\varepsilon_j, \theta_j \in \{0,1\}$. Tiếp theo, M mô phỏng $Q_2^{(2,j)}$ sử dụng một bản mã Ellgamal ngẫu nhiên.

- M mô phỏng $A'_{1,j}$:

$$(A'_{1,j} = b_{1,j} . KPU_{1,\lfloor k/n_{k1}\rfloor}^{(\sum_{i\in I_1} ksu_{i,\lfloor t/n_{k1}\rfloor} + \sum_{l\in I_2} ksu_{l,t\%n_{k1}})} . KPU_{1,\lfloor t/n_{k1}\rfloor}^{-(\sum_{i\in I_1} ksu_{i,\lfloor k/n_{k1}\rfloor} + \sum_{l\in I_2} ksu_{l,k\%n_{k1}})} \tag{2.3.34}$$

Trong đó:

$$(b_{1,j}, b_{2,j}) = (Q_1^{(1,j)} . (Q_2^{(1,j)})^{c_{1,\lfloor j/m_v\rfloor}^{(1)}} . g^{-ksu_{1,\lfloor t/n_{k1}\rfloor} . (ksu_{1,\lfloor k/n_{k1}\rfloor} + ksu_{2,k\%n_{k1}})} . (g^{(ksu_{1,\lfloor t/n_{k1}\rfloor} + ksu_{2,t\%n_{k1}})})^{ksu_{1,\lfloor k/n_{k1}\rfloor}} , g^{ksu_{1,\lfloor k/n_{k1}\rfloor}}) \tag{2.3.35}$$

Vì các giá trị $\{ksu_{i,\lfloor t/n_{k1}\rfloor}, ksu_{i,\lfloor k/n_{k1}\rfloor}, ksu_{l,t\%n_{k1}}, ksu_{l,k\%n_{k1}}\}_{i\in I_1, l\in I_2} \in_R \mathbb{Z}_q^*$ nên:

$$Pr(t \leftarrow \{Q_1^{'(2,j)}, A'_{1,j}\}_{j\in[0,m_u.m_v)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_q^*|} \tag{2.3.36}$$

Kết hợp với công thức 2.3.29 và 2.3.29 suy ra:

$$\{M(I, \overline{x_I}, f_I(\overline{x}))\}_{\overline{x}\in(\{0,1\}^*)^n} \stackrel{c}{\equiv} \{VIEW_{A,I}^{\pi}(\overline{x})\}_{\overline{x}\in(\{0,1\}^*)^n} \tag{2.3.37}$$

Trình mô phỏng M thoả mãn Định nghĩa 1.5, do đó Mệnh đề 2.4 được chứng minh. □

Cuối cùng, mức độ an toàn của giao thức đề xuất được đánh giá trước những tấn công phổ biến. Tương tự như mục 2.2.4, có thể thấy giao thức đề xuất an toàn trước các tấn công từ bên ngoài và các tấn công phổ biến đối với sơ đồ mã hoá khoá công khai bởi vì các bên tham gia được giả định là tuân thủ giao thức và giao tiếp thông qua một kênh an toàn, các cặp khoá được sử dụng đều là cặp khoá sử dụng một lần duy nhất trong mỗi lần gọi giao thức, không có hoạt động giải mã diễn ra trong giao thức, trung tâm tính toán chỉ đơn giản là tổ hợp các bản mã để thu được kết quả cần tính.

2.3.5. Phân tích tính hiệu quả

Trong phần này, giao thức đề xuất được được phân tích và so sánh tính hiệu quả của về chi phí giao tiếp, chi phí tính toán với các giao thức trong cùng mô hình tính toán, có cùng mức độ an toàn giao [27, 50]. Trong đó n, m lần lượt là số lượng người dùng tham gia tính toán và số giá trị tần suất cần phải tính. Các giao thức sử dụng hệ mật mã Elgamal với các tham số được thiết lập như trong mục 1.5.2.5.1.

Các giao thức sẽ sử dụng hệ mật Elgamal kích thước khoá bí mật 160 bit và khoá công khai 1024 bit [42], vì thế mỗi thông điệp sẽ có kích thước là $|p| = 1024$ bít.

2.3.5.1. Đánh giá về chi phí truyền thông

Xem xét giao thức trong [27], trong pha khởi tạo, mỗi người dùng cần gửi hai khóa công khai cho CC. Sau khi CC tính hai khóa công khai chung thì sẽ gửi các khóa này cho tất cả người dùng. Trong pha 1, mỗi U_i cũng cần gửi hai giá trị $C_1^{(i)}, C_2^{(i)}$ cho CC. Trong pha 2, mỗi V_i cũng cần nhận hai giá trị $C_1^{(i)}, C_2^{(i)}$ từ CC và gửi ba giá trị $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ cho CC. Trong pha 3, mỗi U_i cần nhận ba giá trị $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ từ CC và gửi hai giá trị $K_1^{(i)}, K_2^{(i)}$ cho CC. Vậy giao thức trong [27] trao đổi $20n_1$ thông điệp tương ứng $20n_1|p|$ bit. Để tính m giá trị tần suất, sẽ cần trao đổi $20n_1m_um_v$ thông điệp tương ứng $20n_1m_um_v|p|$ bit.

Đối với giao thức trong [50], trước khi giao thức này bắt đầu, mỗi người dùng cần gửi hai khóa công khai cho CC. Trong pha đầu tiên của [50], mỗi U_i cần nhận hai giá trị P_i, Q_i từ CC và gửi bốn giá trị $C_1^{(i)}, C_2^{(i)}, C_3^{(i)}, C_4^{(i)}$ cho CC. CC sẽ gửi bốn giá trị $C_1^{(i)}, C_2^{(i)}, X, Y$ cho tất cả V_i, hai giá trị X, Y cho tất cả U_i. Trong giai đoạn 2, mỗi V_i cũng cần nhận bốn giá trị $C_1^{(i)}, C_2^{(i)}, C_3^{(i)}, C_4^{(i)}$ từ CC và gửi ba giá trị $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ cho CC. Trong pha 3, mỗi U_i cần nhận ba giá trị $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ từ CC và gửi hai giá trị $K_1^{(i)}, K_2^{(i)}$ cho bên này. Vậy giao thức trong [50] trao đổi $24n_1$ thông điệp sử dụng $24n_1|p|$ bit. Để tính m giá trị tần suất, sẽ cần trao đổi $24n_1m_um_v$ thông điệp tương ứng $24n_1m_um_v|p|$ bit.

Đối với giao thức đề xuất, trong pha khởi tạo, mỗi U_i chỉ cần gửi $n_{k1} + 1$ thông điệp cho CC, mỗi V_i chỉ cần gửi n_{k1} thông điệp cho CC, sau khi CC tính n_k khóa công

khai chung thì sẽ gửi các khóa này cho tất cả người dùng. Trong pha 1, mỗi U_i cũng cần gửi các giá trị $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0\leq j<m_u}$ cho CC. Trong pha 2, mỗi V_i cũng cần nhận các giá trị $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0\leq j<m_u}, X_i$ từ CC và gửi các giá trị $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0\leq j<m_u m_v}$ cho CC. Trong pha 3, mỗi U_i cần nhận các giá trị $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0\leq j<m_u m_v}$ từ CC và gửi các giá trị $\{A_{i,j}\}_{0\leq j<m_u m_v}$ cho CC. Vì thế để tính $m_u m_v$ giá trị tần suất, sẽ cần trao đổi $n_1(2n_{k1}+1+2n_k+4m_u+5m_u m_v)$ thông điệp tương ứng $n_1(2n_{k1}+1+2n_k+4m_u+5m_u m_v)|p|$ bit. Từ công thức (2.2.23), (2.3.25) và (2.3.26) ta có:

$$n_k < 0.5m_u m_v + 2$$

$$n_{k1} < \lceil \sqrt{0.5m_u m_v + 2} \rceil$$

Bảng 2.9 cho thấy sự so sánh chi phí truyền thông giữa các giải pháp. Từ bảng này, Luận án có thể thấy rằng giải pháp đề xuất trao đổi số lượng bit thấp hơn so với các giao thức còn lại [27, 50]. Do đó có thể thấy rõ giao thức đề xuất có chi phí truyền thông hiệu quả

Bảng 2.9: So sánh về chi phí truyền thông giữa các giao thức

Giao thức	Số lần giao tiếp với CC	Chi phí truyền thông
Giao thức [27]	$5m_u m_v$	$20n_1 m_u m_v\|p\|$
Giao thức [50]		$24n_1 m_u m_v\|p\|$
Giao thức đề xuất	5	$n_1(2n_k 1+1+2n_k+4m_u+5m_u m_v)\|p\|$

hơn các giải pháp còn lại, hơn nữa số lượng các lần giao tiếp giữa mỗi người dùng với trung tâm tính toán thấp hơn nhiều so với các giao thức [27, 50]. Điều này là do việc giảm số lượng thông điệp cần trao đổi giữa các bên: trong pha khởi tạo, số lượng cặp khóa cần tính toán và trao đổi giảm từ $2m_u m_v$ (trong giao thức gốc) xuống còn n_k, trong pha 2 và pha 3 đối với mỗi giá trị tần suất cần tính đã được thiết kế lại để giảm bớt đi 01 thông điệp, vì thế số lượng thông điệp giảm đi so với giao thức gốc trong hai pha này là $4m_u m_v$. Hơn nữa, trong pha 1 mỗi U_i chỉ phải gửi $2m_u$ thông điệp thay vì tổng số $2m_u m_v$ thông điệp như trong các giao thức còn lại.

2.3.5.2. Đánh giá về chi phí tính toán

Bởi vì chi phí tính toán của các giao thức chủ yếu liên quan đến thời gian thực thi tính toán phép nhân, phép luỹ thừa, phép nghịch đảo modulo của các số nguyên lớn. Do đó chi phí tính toán của các giao thức được đánh giá thông qua các phép toán này.

Đầu tiên, quá trình tính một giá trị tần suất theo giao thức gốc [27] được xem xét. Mỗi người dùng phải dùng $3T_e$ để tính toán ba khóa công khai. Thời gian để mỗi U_i tính

toán các thông điệp $C_1^{(i)}, C_2^{(i)}, K_1^{(i)}, K_2^{(i)}$ của mình là $6T_e + 4T_m$. Thời gian để mỗi V_i chuẩn bị thông điệp $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ của mình là $6T_e + 3T_m + T_i$. Trong pha khởi tạo, thời gian cần thiết để CC tính toán các giá trị công khai KP_1 và KP_2 là $4n_1T_m$. Hơn nữa, CC phải sử dụng $T_i + 2n_1T_m + T_{DL}$ để tính toán 1 giá trị tần suất.

Tiếp theo, quá trình tính một giá trị tần suất của giao thức trong công trình [50] được xem xét. Mỗi người dùng phải dùng $2T_e$ để tính toán ba khóa công khai. Thời gian để mỗi U_i tính toán các thông điệp $C_1^{(i)}, C_2^{(i)}, C_3^{(i)}, C_4^{(i)}, K_1^{(i)}, K_2^{(i)}$ của mình là $8T_e + 8T_m$. Thời gian để mỗi V_i chuẩn bị thông điệp $Q_1^{(i)}, Q_2^{(i)}$ và $Q_3^{(i)}$ của mình là $6T_e + 3T_m + T_i$. Chi phí tính toán mà CC cần thực hiện để tính các giá trị công khai KP_1, KP_2 là $2n_1T_m$. Hơn nữa, CC phải sử dụng $T_i + 2n_1T_m + T_{DL}$ để tính toán 1 giá trị tần suất.

Đối với giao thức đề xuất, để tính m giá trị tần suất, mỗi U_i sử dụng $(n_k + 1)T_e$ để tính toán trong pha khởi tạo, sau đó dành $m_u(3T_e + T_m) + m_u m_v(2T_e + 2T_m + T_i)$ để chuẩn bị các thông điệp của mình. Mỗi V_i sử dụng $(n_k + 1)T_e$ để tính toán trong pha khởi tạo, sau đó dành $m_u m_v(6T_e + 2T_m + 2T_i)$ để chuẩn bị các thông điệp của mình. CC sử dụng $2n_1n_kT_m$ trong pha khởi tạo và sử dụng $m_u m_v n_1 T_m + T_{DL}$ để tính kết quả trong pha 4. Bảng 2.10 thể hiện sự so sánh độ phức tạp tính toán lý thuyết giữa các giao thức. Từ bảng trên có thể thấy chi phí tính toán cho mỗi người dùng và CC trong hai giao

Bảng 2.10: So sánh độ phức tạp tính toán giữa các giao thức

	Giao thức gốc [27]	Giao thức [50]	Giao thức đề xuất
Mỗi U_i	$m_u m_v(9T_e + 4T_m)$ $\approx 2164 m_u m_v T_m$	$m_u m_v(10T_e + 8T_m)$ $\approx 2408 m_u m_v T_m$	$(3m_u + 2m_u m_v + n_k + 1)T_e +$ $(m_u + 2m_u m_v)T_m + m_u m_v T_i$ $< (493,6 m_u m_v + 721 m_u +$ $240\lceil\sqrt{0,5 m_u m_v + 2}\rceil + 240)T_m$
Mỗi V_i	$m_u m_v(9T_e + 3T_m +$ $T_i)$ $\approx 2174,6 m_u m_v T_m$	$m_u m_v(8T_e + 3T_m +$ $T_i)$ $\approx 1934,6 m_u m_v T_m$	$(n_k + 1)T_e + m_u m_v(6T_e + 2T_m +$ $2T_i)$ $< (1453,6 m_u m_v + 240 +$ $240\lceil\sqrt{0,5 m_u m_v + 2}\rceil)T_m$
CC	$m_u m_v(6n_1 T_m +$ $T_i + T_{DL})$ $\approx (6n_1 +$ $11,6)m_u m_v T_m +$ $m_u m_v T_{DL}$	$m_u m_v(4n_1 T_m +$ $T_i + T_{DL})$ $\approx (4n_1 +$ $11,6)m_u m_v T_m +$ $m_u m_v T_{DL}$	$n_1(2n_k + m_u m_v)T_m + T_{DL}$ $< 2n_1(m_u m_v + 2)T_m + T_{DL}$

thức [27, 50] lớn hơn nhiều so với giao thức đề xuất. Ngay ở bước thiết kế lại giao thức gốc khi sử dụng cùng một hệ mật với hai giao thức điển hình, do số lượng khoá sử dụng của giao thức đề xuất giảm đi đáng kể và số lượng tính toán giữa các bên được thiết kế

và tối ưu hơn nên chi phí tính toán của mỗi người dùng và của CC đều thấp hơn so với các giao thức còn lại [27, 50]. Hơn thế nữa, thay vì yêu cầu mỗi người dùng U_i phải tính toán và gửi $2m_u \cdot m_v$ thông điệp thì trong giao thức mới chỉ cần tính và gửi $2m_u$ thông điệp. Điều đó làm cho giao thức đề xuất có hiệu suất cao.

2.3.6. Chuyển đổi giao thức đề xuất sử dụng hệ mật đường cong Elliptic

Để tiếp tục nâng cao hiệu suất, giao thức đề xuất được chuyển đổi sang một biến thể sử dụng hệ mật mã trên đường cong Elliptic theo ý tưởng cải tiến số 2 đã được đề xuất trong Mục 2.3.1. Vì hệ mật mã Elgamal trên trường hữu hạn và trên đường cong Elliptic là tương đồng nên với cùng mức độ an toàn thì tính đúng đắn và tính riêng tư của giao thức khi chuyển đổi sang sử dụng hệ mật đường cong Elliptic sẽ tương đương với giao thức ban đầu sử dụng hệ mật mã Elgamal trên trường hữu hạn. Do đó, trong phần này chỉ phân tích về tính hiệu suất của giao thức chuyển đổi này so với giao thức ban đầu sử dụng hệ mật mã Elgamal trên trường hữu hạn.

Trong đó, các tham số hệ thống được thiết lập như mô tả trong Mục 1.5.2.5.2. Chi tiết giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD sử dụng hệ mật đường cong Elliptic được mô tả như trong Giao thức **2.7**.

Phần tiếp theo sẽ đánh giá mức độ hiệu quả của giao thức này so với giao thức sử dụng hệ mật Elgamal trên trường hữu hạn về chi phí truyền thông và chi phí tính toán. Giao thức dựa trên hệ mật mã đường cong Elliptic sẽ sử dụng đường cong Elliptic BrainpoolP160r1 với kích thước khoá bí mật là 160 bit, khoá công khai là 320 bit [45] (hai hệ mật này có cùng mức an toàn là 80 bit).

Khi xem xét về chi phí truyền thông ta có thể thấy: giao thức đề xuất sử dụng hệ mật Elgamal trên trường hữu hạn và trên đường cong Elliptic có số lượng các thông điệp cần trao đổi giữa các bên tham gia là như nhau. Tuy nhiên thì kích thước của mỗi thông điệp của các hệ mật này là khác nhau: mỗi thông điệp của hệ mật Elgamal trên trường hữu hạn là $|p| = 1024$ bít, mỗi thông điệp của hệ mật Elgamal trên đường cong Elliptic tương ứng với 1 điểm và có kích thước là $2.160 = 320 = \frac{5|p|}{16}$ bít. Tương ứng với chi phí truyền thông của giao thức đề xuất sử dụng hệ mật trên đường cong Elliptic giảm đi $\frac{5|p|}{16}$ lần so với sử dụng hệ mật mã Elgamal trên trường hữu hạn. Bảng 2.11 cho thấy sự so sánh chi phí truyền thông và chi phí tính toán giữa hai giao thức này.

Từ kết quả so sánh trên cho thấy việc chuyển đổi giao thức đề xuất sang sử dụng hệ mật mã Elgamal trên đường cong Elliptic giúp cải thiện đáng kể cả chi phí truyền thông và chi phí tính toán của giao thức. Chi phí truyền thông giảm còn khoảng $\frac{1}{3}$, chi phí tính toán của mỗi người dùng U_i giảm khoảng 5 lần, chi phí tính toán của mỗi người dùng V_i và CC giảm khoảng 8 lần

Đầu vào:
n: Số lượng người dùng trong hệ thống, trong đó mỗi miền dữ liệu U_i, V_i có n_1 người dùng, $i \in [1, n_1], n = 2n_1$.
$u_{i,j}$: Các giá trị nhị phân bí mật mà U_i nắm giữ $(0 \le j < m_u)$.
$v_{i,j}$: Các giá trị nhị phân bí mật mà V_i nắm giữ $(0 \le j < m_v), m_u \le m_v$.
Output (đối với CC):
$SU = \left\{ su_j = \sum_{i=1}^{n_1} u_{i,j\%m_v} \cdot v_{i,\lfloor j/m_v \rfloor} \right\}_{j \in [0, m_u m_v - 1)}$

Pha khởi tạo
- Mỗi U_i thực hiện:
1. $x_i \in_R \mathbb{Z}_d^*$; $X_i = x_i G$;
2. **for** $(0 \le j < n_{k1})$
3. $ksu_{i,j} \in_R \mathbb{Z}_d^*$; $KPU_{i,j} = ksu_{i,j} G$;
4. Gửi $\{KPU_{i,j}\}_{0 \le j < n_{k1}}, X_i$ cho CC;
- Mỗi V_i thực hiện:
5. $y_i \in_R \mathbb{Z}_d^*$; $Y_i = y_i G$;
6. **for** $(0 \le j < n_{k1})$
7. $ksv_{i,j} \in_R \mathbb{Z}_d^*$; $KPV_{i,j} = ksv_{i,j} G$;
8. Gửi $\{KPV_{i,j}\}_{0 \le j < n_{k1}}$ cho CC;
- CC thực hiện:
9. $j = 0$;
10. **for** $(0 \le t < n_{k1})$
11. **for** $(0 \le k < n_{k1})$
12. $KP_j = \sum_{i=1}^{n_1} (KPU_{i,t} + KPV_{i,k})$;
13. $j++$;
14. **if** $(j == n_k - 1)$ **break**;
15. **if** $(j == n_k - 1)$ **break**;
16. Gửi $\{KP_j\}_{0 \le j < n_k}$ cho $\{U_i, V_i\}_{1 \le i \le n_1}$;
Pha 1: Mỗi U_i thực hiện
17. **For** $0 \le j < m_u$
18. $c_{i,j}^{(1)} \in_R \mathbb{Z}_d^*$;
19. $C_1^{(i,j)} = u_{i,j} G + c_{i,j}^{(1)} X_i$; $C_2^{(i,j)} = c_{i,j}^{(1)} G$;
20. Gửi $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}$ cho CC.
Pha 2: Mỗi V_i thực hiện:
21. Nhận $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \le j < m_u}, X_i$ từ CC;
22. $j = 0$;
23. **for** $(0 \le t < n_k - 1)$
24. **for** $(t + 1 \le k < n_k)$
25. $c_{i,j}^{(2)} \in_R \mathbb{Z}_d^*$;
26. $Q_1^{(i,j)} = v_{i,j\%m_v} C_1^{(i,\lfloor j/m_v \rfloor)} + ksv_{i,k\%n_{k1}} KP_t - c_{i,j}^{(2)} y_i C_2^{(i,\lfloor j/m_v \rfloor)} - ksv_{i,t\%n_{k1}} KP_k$;
27. $Q_2^{(i,j)} = c_{i,j}^{(2)} Y_i - v_{i,j\%m_v} X_i$;
28. $j++$;
29. **if** $(j == m_u \cdot m_v - 1)$ **break**;
30. **if** $(j == m_u \cdot m_v - 1)$ **break**;
31. Gửi $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u \cdot m_v}$ cho CC;
Pha 3: Mỗi U_i thực hiện:
32. Lấy $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < m_u \cdot m_v}$ từ CC;
33. $j = 0$;
34. **for** $(0 \le t < n_k - 1)$
35. **for** $(t + 1 \le k < n_k)$
36. $A_{i,j} = Q_1^{(i,j)} + c_{i,\lfloor j/m_v \rfloor}^{(1)} Q_2^{(i,j)} - ksu_{i,\lfloor t/n_{k1} \rfloor} KP_k + ksu_{i,\lfloor k/n_{k1} \rfloor} KP_t$;
37. $j++$;
38. **if** $(j == m_u \cdot m_v - 1)$ **break**;
39. **if** $(j == m_u \cdot m_v - 1)$ **break**;
40. Gửi $\{A_{i,j}\}_{0 \le j < m}$ cho CC;
Pha 4: CC thực hiện
41. **For** $0 \le j < m_u m_v$
42. $A_j = \sum_{i=1}^{n_1} A_{i,j}$;
43. $SU = Dlog_{EC}(\mathbb{E}, G, n_1, A)$
//Sử dụng thuật toán vét cạn 1 lần.

Giao thức 2.7: Giao thức PPMFC() tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD sử dụng hệ mật đường cong Elliptic

2.3.7. Đánh giá về mặt thực nghiệm

Trong phần này, các giao thức [27, 50] có cùng mức độ an toàn, cùng số lần giao tiếp giữa các bên tham gia với trung tâm tính toán và giao thức đề xuất được cài đặt bằng ngôn ngữ C# trong môi trường Visual Studio 2019, với sự hỗ trợ của thư viện System.Numerics để so sánh hiệu suất giữa các giao thức này. Các thực nghiệm cùng

Bảng 2.11: So sánh về chi phí truyền thông và chi phí tính toán giữa các giao thức

Tiêu chí	**Đối tượng**	**Giao thức Elgamal trên trường hữu hạn**	**Giao thức trên đường cong Elliptic**
Chi phí truyền thông		$n_1(2n_{k1}+1+2n_k+9m)\|p\|$ $< n_1(4m_u + 6m_um_v + \lceil\sqrt{2m_um_v+8}\rceil+5)\|p\|$	$0,32\|p\|n_1(2n_{k1} + 1 + 2n_k + 4m_u+5m_um_v)$ $< 0,32n_1(4m_u + 6m_um_v + \lceil\sqrt{2m_um_v+8}\rceil+5)\|p\|$
Chi phí tính toán	Mỗi U_i	$(493,6m_um_v + 721m_u + 240\lceil\sqrt{0,5m_um_v+2}\rceil+240)T_m$	$(3m_u+3m_um_v+n_{k1}+1)T_{mp} + (3m_um_v+m_u)T_{ap}$ $< (3m_u + 3m_um_v + \lceil\sqrt{0,5m_um_v+2}\rceil + 1)T_{mp} + (3m_um_v+m_u)T_{ap}$ $\approx (87,36m_um_v + 29 + 87,12m_u + 29\lceil\sqrt{0,5m_um_v+2}\rceil)T_m$
	Mỗi V_i	$(1453,6m_um_v + 240 + 240\lceil\sqrt{0.5m_um_v+2}\rceil)T_m$	$(6m_um_v + n_{k1} + 1)T_{mp} + 4m_um_vT_{ap}$ $< (174,48m_um_v + 29 + 29\lceil\sqrt{0,5m_um_v+2}\rceil)T_m$
	CC	$2n_1(m_um_v+2)T_m+T_{DL}$	$n_1(2n_k+m_um_v)T_{ap}+T_{DL}$ $<(4+2m_um_v)T_{ap}+T_{DL}$ $\approx 0,24n_1(m_um_v+2)T_m+T_{DL}$

chạy trên một máy tính xách tay Intel Core i5-4210M CPU 2.60GHz, 8GB RAM. Để giảm thiểu sự sai khác về môi trường thực nghiệm, mỗi giao thức sẽ được thực thi độc lập trong điều kiện không có ứng dụng nào khác được chạy trên máy tính. Thời gian thực thi của các giao thức sẽ được đo bằng cách sử dụng lớp Stopwatch trong thư viện System.Diagnostics.

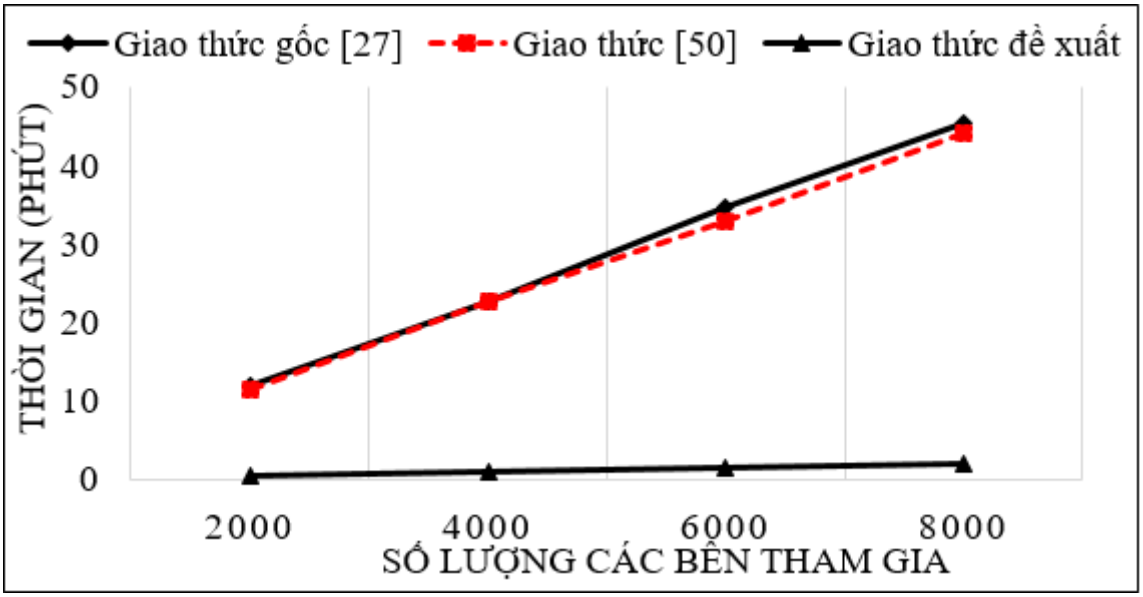
Để so sánh thời gian chạy của các giao thức, mỗi giao thức được thực thi 10 lần cho mỗi bộ tham số và lấy giá trị trung bình. Luận án giả định rằng tất cả người dùng thực hiện tác vụ của họ cùng một lúc và độ trễ mạng không được tính vào tổng thời gian thực thi. Do đó, chi phí tính toán ở phía người dùng có thể được giảm xuống bằng chi phí tính toán cho một người, chi phí tính toán và giao tiếp ở CC được tính cho tất cả người dùng tham gia vào hệ thống vì chúng phụ thuộc vào sự cộng tác của tất cả người dùng với CC.

Để thuận tiện, giả sử rằng mỗi giá trị đầu vào bí mật của các bên tham gia là 1 và không gian của nghiệm của các bài toán logarit rời rạc là từ 0 đến n_1. Trong kết quả này đã tính toán thời gian tính toán thực tế cần thiết trong từng pha. Các thực nghiệm được thực thi với số lượng người dùng trong hệ thống thay đổi, tức là 2000, 4000, 6000, 8000 và số lượng giá trị tần suất cần tính 8000.

Như đã đề cập, giao thức mới được nâng cấp từ giao thức trong tài liệu [27]. Đặc

biệt, trong giao thức gốc, mỗi U_i phải tính toán hai giá trị $K_1^{(i)}$, $K_2^{(i)}$ để gửi cho CC. Dựa trên bộ giá trị của hai giá trị này, CC thực hiện phép nhân:$K_1^{(i)} \cdot (K_2^{(i)})^{-1}$. Do đó, độ phức tạp tính toán tại CC là cao. Không giống như giao thức [27], trong giao thức đề xuất, đối với mỗi giá trị tần suất, mỗi U_i chỉ cần tính một điểm duy nhất $A_{i,j}$ và CC chỉ cần thực hiện phép cộng các điểm $A_{i,j}$. Điều này chỉ làm tăng không đáng kể độ phức tạp tính toán của mỗi U_i, nhưng lại giảm đáng kể độ phức tạp tính toán tại CC.

Thêm vào đó, việc giảm số lượng cặp khóa cần sử dụng giúp giảm đáng kể chi phí tính toán cả ở phía người dùng và phía CC. Hơn nữa, khi chuyển sang sử dụng hệ mật mã trên đường cong Elliptic, chi phí tính toán của giao thức đề xuất giảm đi đáng kể. Do đó, tổng thời gian thực thi của giao thức đề xuất thấp hơn nhiều so với các giao thức còn lại, như thể hiện trong Hình 2.2.



Hình 2.2: So sánh thời gian thực thi giữa các giao thức

Từ các kết quả trên cho thấy giao thức đề xuất hiệu quả hơn so với các giao thức còn lại.

2.3.8. Ví dụ minh hoạ

Để mô tả rõ ràng quá trình tính các giá trị tần suất có đảm bảo tính riêng tư, luận án sử dụng ma trận các giá trị tần suất kích thước 3×8 như trong Bảng 2.12. Trong đó, người dùng và mục được biểu diễn dưới dạng $U_i = \{U_1, U_2, U_3\}$, $V_i = \{V_1, V_2, V_3\}$ và $i_j = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8\}$ tương ứng. Trước khi bắt đầu, các bên tham gia thống nhất các tham số đường cong Elliptic $(\mathbb{E}, G, d, p)$.

Bảng 2.12: Ví dụ về ma trận giá trị tần suất

U_i/i_j	i_1	i_2	V_i/i_j	i_3	i_4	i_5
U_1	1	1	V_1	0	1	1
U_2	0	1	V_2	1	0	1
U_3	1	1	V_3	1	1	0

Người dùng U_i trong đó $i = \{1,2,3\}$, chọn ngẫu nhiên khóa bí mật của họ $x_i, ksu_{i,j}$, trong đó $x_i, ksu_{i,j} \in_R \mathbb{Z}_d^*$, $0 \leq j < \left\lceil \sqrt{\left\lceil \frac{1}{2} + \sqrt{2.2.3 + \frac{1}{4}} \right\rceil} \right\rceil = 2$ và tính khóa công khai là: $X_i = x_i G$, $KPU_{i,j} = ksu_{i,j} G$. Sau đó gửi cho CC.

Người dùng V_i trong đó $i = \{1,2,3\}$, chọn ngẫu nhiên khóa bí mật của họ $y_i, ksv_{i,j}$, trong đó $y_i, ksv_{i,j} \in_R \mathbb{Z}_d^*$, $0 \leq j < 2$ và tính khóa công khai là:$Y_i = y_i G$, $KPV_{i,j} = ksv_{i,j} G$. Sau đó gửi cho CC.

CC tính $n_k = \left\lceil \frac{1}{2} + \sqrt{2.2.3 + \frac{1}{4}} \right\rceil = 4$ khóa công khai dùng chung như sau:

$$
\begin{aligned}
KP_0 &= KPU_{1,0} + KPU_{2,0} + KPU_{3,0} + KPV_{1,0} + KPV_{2,0} + KPV_{3,0} \\
KP_1 &= KPU_{1,0} + KPU_{2,0} + KPU_{3,0} + KPV_{1,1} + KPV_{2,1} + KPV_{3,1} \\
KP_2 &= KPU_{1,1} + KPU_{2,1} + KPU_{3,1} + KPV_{1,0} + KPV_{2,0} + KPV_{3,0} \\
KP_3 &= KPU_{1,1} + KPU_{2,1} + KPU_{3,1} + KPV_{1,1} + KPV_{2,1} + KPV_{3,1}
\end{aligned}
$$

CC gửi $\{KP_j\}_{0 \leq j < 4}$ cho tất cả các bên tham gia.

Người dùng U_1, U_2 và U_3 gửi các giá trị xếp hạng của họ tới CC như sau:

$$
\begin{aligned}
(C_1^{(1)}, C_2^{(1)}) &= \{(1.G + c_{1,0}^{(1)} X_1, c_{1,0}^{(1)} G), (1.G + c_{1,1}^{(1)} X_1, c_{1,1}^{(1)} G)\} \\
(C_1^{(2)}, C_2^{(2)}) &= \{(0.G + c_{2,0}^{(1)} X_2, c_{2,0}^{(1)} G), (1.G + c_{2,1}^{(1)} X_2, c_{2,1}^{(1)} G)\} \\
(C_1^{(3)}, C_2^{(3)}) &= \{(1.G + c_{3,0}^{(1)} X_3, c_{3,0}^{(1)} G), (1.G + c_{3,1}^{(1)} X_3, c_{3,1}^{(1)} G)\}
\end{aligned}
$$

Sau đó, người dùng V_1, V_2 và V_3 nhận các giá trị $\{C_1^{(i,j)}, C_2^{(i,j)}\}_{0 \leq j < 2}, X_i$ từ CC và

thực hiện các tính toán sau:

$$\begin{aligned}(Q_1^{(1)}, Q_2^{(1)}) = \{&(0.C_1^{(1,0)} + ksv_{1,1}KP_0 - c_{1,0}^{(2)} y_1 C_2^{(1,0)} - ksv_{1,0}KP_1, c_{1,0}^{(2)} Y_1 - 0.X_1),\\ &(1.C_1^{(1,0)} + ksv_{1,2}KP_0 - c_{1,1}^{(2)} y_1 C_2^{(1,0)} - ksv_{1,0}KP_2, c_{1,1}^{(2)} Y_1 - 1.X_1),\\ &(1.C_1^{(1,0)} + ksv_{1,3}KP_0 - c_{1,2}^{(2)} y_1 C_2^{(1,0)} - ksv_{1,0}KP_3, c_{1,2}^{(2)} Y_1 - 1.X_1),\\ &(0.C_1^{(1,1)} + ksv_{1,2}KP_1 - c_{1,3}^{(2)} y_1 C_2^{(1,1)} - ksv_{1,1}KP_2, c_{1,3}^{(2)} Y_1 - 0.X_1),\\ &(1.C_1^{(1,1)} + ksv_{1,3}KP_1 - c_{1,4}^{(2)} y_1 C_2^{(1,1)} - ksv_{1,1}KP_3, c_{1,4}^{(2)} Y_1 - 1.X_1),\\ &(1.C_1^{(1,1)} + ksv_{1,3}KP_2 - c_{1,5}^{(2)} y_1 C_2^{(1,1)} - ksv_{1,2}KP_3, c_{1,5}^{(2)} Y_1 - 1.X_1)\}\end{aligned}$$

Tương tự cho $(Q_1^{(2)}, Q_2^{(2)})$ và $(Q_1^{(3)}, Q_2^{(3)})$. Gửi $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < 6}$ cho CC.

Tiếp theo, người dùng U_1, U_2 và U_3 nhận các giá trị $\{Q_1^{(i,j)}, Q_2^{(i,j)}\}_{0 \le j < 6}$ từ CC và thực hiện các tính toán tiếp theo.

$$A_{1,j} = \begin{Bmatrix} 0.1G + (ksv_{1,1} + ksu_{1,0})KP_0 - (ksv_{1,0} + ksu_{1,0})KP_1, \\ 1.1G + (ksv_{1,0} + ksu_{1,1})KP_0 - (ksv_{1,0} + ksu_{1,0})KP_2, \\ 1.1G + (ksv_{1,1} + ksu_{1,1})KP_0 - (ksv_{1,0} + ksu_{1,0})KP_3, \\ 0.1G + (ksv_{1,0} + ksu_{1,1})KP_1 - (ksv_{1,1} + ksu_{1,0})KP_2, \\ 1.1G + (ksv_{1,1} + ksu_{1,1})KP_1 - (ksv_{1,1} + ksu_{1,0})KP_3, \\ 1.1G + (ksv_{1,1} + ksu_{1,1})KP_2 - (ksv_{1,0} + ksu_{1,1})KP_3 \end{Bmatrix}$$

$$A_{2,j} = \begin{Bmatrix} 1.0G + (ksv_{2,1} + ksu_{2,0})KP_0 - (ksv_{2,0} + ksu_{2,0})KP_1, \\ 0.0G + (ksv_{2,0} + ksu_{2,1})KP_0 - (ksv_{2,0} + ksu_{2,0})KP_2, \\ 1.0G + (ksv_{2,1} + ksu_{2,1})KP_0 - (ksv_{2,0} + ksu_{2,0})KP_3, \\ 1.1G + (ksv_{2,0} + ksu_{2,1})KP_1 - (ksv_{2,1} + ksu_{2,0})KP_2, \\ 0.1G + (ksv_{2,1} + ksu_{2,1})KP_1 - (ksv_{2,1} + ksu_{2,0})KP_3, \\ 1.1G + (ksv_{2,1} + ksu_{2,1})KP_2 - (ksv_{2,0} + ksu_{2,1})KP_3 \end{Bmatrix}$$

$$A_{3,j} = \begin{Bmatrix} 1.1G + (ksv_{3,1} + ksu_{3,0})KP_0 - (ksv_{3,0} + ksu_{3,0})KP_1, \\ 1.1G + (ksv_{3,0} + ksu_{3,1})KP_0 - (ksv_{3,0} + ksu_{3,0})KP_2, \\ 0.1G + (ksv_{3,1} + ksu_{3,1})KP_0 - (ksv_{3,0} + ksu_{3,0})KP_3, \\ 1.1G + (ksv_{3,0} + ksu_{3,1})KP_1 - (ksv_{3,1} + ksu_{3,0})KP_2, \\ 1.1G + (ksv_{3,1} + ksu_{3,1})KP_1 - (ksv_{3,1} + ksu_{3,0})KP_3, \\ 0.1G + (ksv_{3,1} + ksu_{3,1})KP_2 - (ksv_{3,0} + ksu_{3,1})KP_3 \end{Bmatrix}$$

CC thực hiện tính các bản mã tổng hợp:

$$A_0 = 0.1G + 1.0G + 1.1G + \sum_{i=1}^{3} \left((ksv_{i,1} + ksu_{i,0})KP_0 - (ksv_{i,0} + ksu_{i,0})KP_1 \right) = 1G$$

$$A_1 = 1.1G + 0.0G + 1.1G + \sum_{i=1}^{3} \left((ksv_{i,0} + ksu_{i,1})KP_0 - (ksv_{i,0} + ksu_{i,0})KP_2 \right) = 2G$$

$$A_2 = 1.1G + 1.0G + 0.1G + \sum_{i=1}^{3} \left((ksv_{i,1} + ksu_{i,1})KP_0 - (ksv_{i,0} + ksu_{i,0})KP_3 \right) = 1G$$

$$A_3 = 0.1G + 1.1G + 1.1G + \sum_{i=1}^{3} \left((ksv_{i,0} + ksu_{i,1})KP_1 - (ksv_{i,1} + ksu_{i,0})KP_2 \right) = 2G$$

$$A_4 = 1.1G + 0.1G + 1.1G + \sum_{i=1}^{3} \left((ksv_{i,1} + ksu_{i,1})KP_1 - (ksv_{i,1} + ksu_{i,0})KP_3 \right) = 2G$$

$$A_5 = 1.1G + 1.1G + 0.1G + \sum_{i=1}^{3} \left((ksv_{i,1} + ksu_{i,1})KP_2 - (ksv_{i,0} + ksu_{i,1})KP_3 \right) = 2G$$

Sử dụng logarit rời rạc để xác định các bản rõ tương ứng là: $F = \{1,2,1,2,2,2\}$.

2.4. Kết luận chương

Trong chương này, luận án đã phát triển 02 giao thức tính toán bảo mật hiệu quả: Giao thức tính đồng thời nhiều giá trị trung bình, độ tương tự có bảo mật và Giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD.

Đồng thời cũng đã chứng minh rằng các giao thức đề xuất vừa đảm bảo được tính đúng đắn của kết quả đầu ra, vừa đảm bảo được độ an toàn trong mô hình bán trung thực. Các đánh giá về mặt lí thuyết và thực nghiệm cũng cho thấy tính hiệu quả cao của các giao thức này so với các giao thức tính trung bình, giao thức tính độ tương tự được trình bày trong tài liệu [7] và giao thức tính tần suất trong mô hình 2PFD được trình bày trong tài liệu [27], tương ứng.

Do đó, các giao thức đề xuất có khả năng áp dụng vào các bài toán thực tế. Cụ thể như: giao thức tính trung bình và độ tương tự có bảo mật được đề xuất đã được áp dụng vào trong bài toán đảm bảo tính riêng tư cho hệ tư vấn người dùng đã được công bố trong công trình [CT3, CT4, CT5], giao thức tính tần suất có bảo mật trong mô hình 2PFD [CT1] đã được áp dụng vào trong bài toán đảm bảo tính riêng tư trong quá trình huấn luyện cây quyết định đã được công bố trong [CT2].

Nội dung của chương này liên quan đến các công trình [CT1, CT3, CT4, CT5, CT6] trong Danh mục các công trình khoa học sử dụng trong Luận án. Tiếp tục hướng nghiên cứu phát triển giao thức tính toán bảo mật, luận án nghiên cứu đề xuất các giải pháp ứng dụng các giao thức đề xuất để đảm bảo tính riêng tư hệ tư vấn người dùng trong chương tiếp theo.

CHƯƠNG 3. GIẢI PHÁP ĐẢM BẢO TÍNH RIÊNG TƯ CHO HỆ TƯ VẤN NGƯỜI DÙNG DỰA TRÊN CÁC GIAO THỨC TÍNH TOÁN BẢO MẬT

3.1. Mở đầu chương

Trong chương này, các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ và phân tán đầy đủ hai bên được đề xuất dựa trên việc sử dụng các giao thức tính toán bảo mật phục vụ cho việc tính các giá trị trung bình, độ tương tự và giao thức tính tần suất trong mô hình dữ liệu 2PDF được phát triển ở chương trước. Trong đó, các giá trị đầu vào là các giá trị xếp hạng thể hiện ý kiến đánh giá của người dùng đối với một mục tin, trong đó các giá trị này là các số nguyên không âm khá nhỏ, ví dụ trong một số ứng dụng đánh giá thực tế hiện nay thì thang đánh giá thường có giá trị từ 0 đến 5 như: các hệ thống của amazon, ebay, google, shopee,. . . . Các giao thức được phát triển này sẽ được ứng dụng trong giai đoạn đầu tiên của giải pháp Hệ tư vấn người dùng có đảm bảo tính riêng tư: giai đoạn tính giá trị trung bình xếp hạng của mục tin và độ tương tự của cặp mục tin. Những đề xuất này có liên quan đến các Công trình [CT3, CT4, CT6].

Ý tưởng chung của bài toán hệ tư vấn người dùng có đảm bảo tính riêng tư sẽ được trình bày trong mục 3.2. Sau đó các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư đề xuất cho mô hình dữ liệu phân tán đầy đủ và phân tán đầy đủ hai bên được trình bày trong mục 3.3 và 3.4 tương ứng.

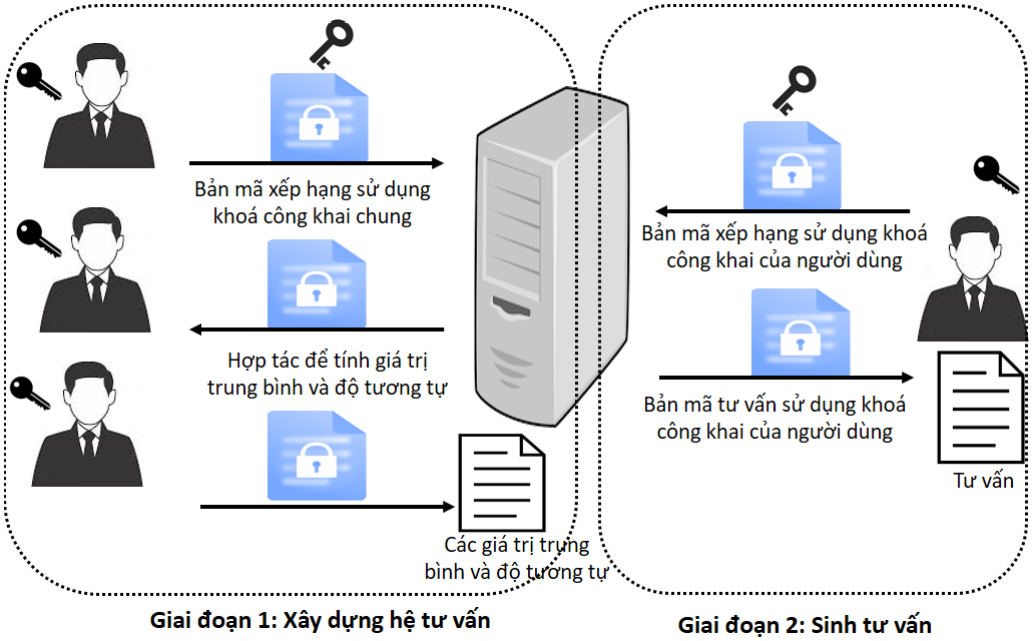
Các giải pháp này có thể được triển khai thực tế dưới dạng mô hình máy chủ - máy trạm, đây là một mô hình phổ biến hiện nay. Trong đó phía người dùng sẽ sử dụng máy trạm có ứng dụng để thực thi các tác vụ của mình và kết nối với máy chủ tư vấn thông qua hệ thống mạng để trao đổi các thông điệp. Trên máy chủ tư vấn cũng sẽ cài đặt ứng dụng để thực hiện các tác vụ tương ứng của mình.

Để tiến hành thực nghiệm đánh giá tính hiệu quả của các giải pháp, luận án sử dụng dữ liệu có sẵn công khai do GroupLens [67] cung cấp, bao gồm 100.000 xếp hạng, 1982 mục được cung cấp bởi 943 người dùng trên thang điểm đánh giá từ 1–5. Đồng thời, luận án cũng giả định rằng tất cả người dùng thực hiện tác vụ của họ cùng một lúc và độ trễ mạng không được tính vào tổng thời gian thực thi. Do đó, chi phí tính toán ở phía người dùng có thể được giảm xuống bằng chi phí tính toán cho một người, chi phí tính toán và giao tiếp ở phía máy chủ tư vấn được tính cho tất cả người dùng tham gia vào hệ thống vì chúng phụ thuộc vào sự cộng tác của tất cả người dùng với máy chủ tư vấn.

Trong thực tế quá trình trao đổi giữa các bên tham gia sẽ có độ trễ và điều này sẽ ảnh hưởng lớn tới tính hiệu quả của giải pháp. Khi đó thời gian thực thi phía người dùng trong một pha nào đó sẽ tương đương với thời gian từ khi pha bắt đầu cho tới khi người dùng cuối cùng thực thi xong. Vì thế để hạn chế độ trễ này thì cần có biện pháp liên quan đến việc đảm bảo băng thông của mạng kết nối và có những chính sách phù hợp để các bên tham gia thực thi các tác vụ của mình một cách nhanh nhất.

3.2. Bài toán tư vấn người dùng có đảm bảo tính riêng tư

Từ những phân tích, đánh giá trong Mục 1.7.1 có thể thấy mặc dù các giải pháp PPRS được đề xuất trong các công trình [7,55,75] phù hợp với bài toán của luận án. Giải pháp trong công trình [7] đảm bảo tính đúng đắn của tư vấn, hiệu suất tốt tuy nhiên mức độ an toàn chưa cao. Giải pháp trong công trình [55,75] cũng đảm bảo tính đúng đắn, có mức độ an toàn cao tuy nhiên hiệu suất thấp hơn so với giải pháp trong công trình [7]. Đối với ngữ cảnh tính toán phân tán, tính hiệu suất đóng một vai trò quan trọng trong việc quyết định ứng dụng giải pháp trong thực tế.



Hình 3.1: Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư [7]

Vì vậy, giải pháp trong công trình [7] sẽ tiếp tục được phân tích chi tiết hơn. Giải pháp này được chia thành hai giai đoạn như Hình 3.1, cụ thể như sau:

- **Giai đoạn thứ nhất: Xây dựng hệ tư vấn**. Trong giai đoạn này, các bên thực hiện tính trung bình xếp hạng của các mục tin và tính độ tương tự giữa các các cặp mục tin có đảm bảo tính riêng tư. Người dùng mã hóa xếp hạng của họ và gửi các bản mã đến máy chủ. Máy chủ tính các giá trị trung bình xếp hạng của từng mục tin cũng như độ

tương tự giữa các mục sử dụng thuộc tính đồng cấu và sự hợp tác của tất cả người dùng. Sau đó, các giá trị độ tương tự và giá trị trung bình xếp hạng của các mục tin được lưu trữ trong cơ sở dữ liệu của máy chủ. Tuy nhiên giao thức tính các giá trị trung bình và độ tương tự giữa các mục tin được sử dụng trong giai đoạn này mặc dù đảm bảo được tính đúng đắn nhưng tính an toàn và hiệu suất chưa cao như được phân tích trong Mục 2.2.1.

- **Giai đoạn thứ hai: Sinh tư vấn**. Trong giai đoạn này, chỉ một người dùng tham gia nhận các đề xuất (được gọi là "người dùng mục tiêu") và gửi các bản mã của các đánh giá mục tin của người dùng đó tới máy chủ. Máy chủ tính điểm tư vấn theo cách đồng cấu và gửi bản mã kết quả cho người dùng. Người dùng mục tiêu cuối cùng giải mã được bản mã bằng khóa riêng của mình. Giải pháp cung cấp hai loại phương pháp để sinh tư vấn có đảm bảo tính riêng tư bằng cách sử dụng Lọc dựa trên nội dung (CBF) và Lọc cộng tác (CF).

Hơn nữa, giải pháp này sử dụng hệ mật mã Elgamal vì thế hiệu suất chưa cao. Vì thế luận án thực hiện thiết kế lại cả hai giai đoạn của giải pháp [7] nhằm nâng cao hiệu suất và mức độ an toàn của giải pháp. Cụ thể như sau:

- Ở giai đoạn thứ nhất của giải pháp đề xuất sẽ sử dụng các giao thức được phát triển trong Chương 2 để tính các giá trị trung bình xếp hạng của các mục tin và các giá trị độ tương tự giữa các cặp mục tin.

- Ở giai đoạn sinh tư vấn, luận án sẽ tiếp tục phát triển các giải pháp sinh tư vấn tương ứng với hai phương pháp CBF, CF trên cơ sở cải tiến giải pháp đề xuất trong tài liệu [7] thông qua việc sử dụng hệ mật mã trên đường cong Elliptic thay cho hệ mật mã Elgamal. Sau đó, các giải pháp này sẽ được đánh giá về tính đúng đắn, mức độ an toàn và hiệu suất.

Giải pháp đề xuất này sẽ được trình bày cụ thể trong Mục 3.3.

Từ những phân tích, đánh giá trong Mục 1.7.2, mặc dù hiện nay có rất nhiều giải pháp PPRS nhưng vẫn chưa có giải pháp nào cho kịch bản dữ liệu nhạy cảm của người dùng được phân tán đầy đủ hai bên. Giải pháp PPRS trong mô hình 2PFD phức tạp hơn so với các cài đặt khác, bởi vì một bộ giá trị ở đây có thể thuộc về hai người dùng thuộc hai miền dữ liệu khác nhau. Do đó cần phải có giải pháp mới cho giai đoạn đầu tiên của giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình 2PFD. Vì vậy trong Mục 3.4, luận án đề xuất giải pháp PPRS mới cho mô hình 2PFD.

Các giải pháp đề xuất tập trung đảm bảo tính riêng tư trong mô hình bán trung thực, trong đó các bên tham gia sẽ tuân thủ giao thức nhưng các bên cũng có thể thông đồng với nhau để tìm hiểu các thông tin nhạy cảm của các bên tham gia khác.

Cả hai giải pháp đề xuất đều sử dụng hệ mật mã trên đường cong Elliptic nên giai đoạn khởi tạo hệ thống sẽ thống nhất như Mục 1.5.2.5.2.

Các tham số đường cong elliptic cần được lựa chọn cẩn thận để chống lại tất cả các cuộc tấn công đã biết vào bài toán logarit rời rạc trên đường cong elliptic (ECDLP). Ta có thể sử dụng các bộ tham số đường cong elliptic mẫu theo các chuẩn đưa ra (ví dụ: FIPS 186-4, NIST SP 800-186, SEC 2,...) hoặc tự lựa chọn ngẫu nhiên bộ tham số và sau đó kiểm chứng lại các rằng buộc an toàn của bộ tham số này theo cách thức được đề cập chi tiết trong tài liệu [35]. Luận án giả định các tham số này được lựa chọn đúng đắn để đảm bảo an toàn cho hệ mật mã trên đường cong Elliptic.

Việc lựa chọn bộ tham số đường cong elliptic đóng vai trò quan trọng trong việc đảm bảo an toàn cho giải pháp. Tùy theo từng yêu cầu cụ thể về mức an toàn của lĩnh vực tư vấn mà có thể chọn các tham số của đường cong Elliptic cho phù hợp. Yêu cầu về kích thước khóa hay chính là kích thước của bậc đường cong elliptic cho các mức an toàn như được thể hiện trong Bảng 3.1.

Bảng 3.1: Yêu cầu về kích thước khóa của EC cho các mức an toàn

Mức an toàn (bit)	**Kích thước khóa (bit)**
80	160
112	224
128	256
192	384
256	512

Phần tiếp theo sẽ trình bày chi tiết về hai giải pháp đề xuất: Hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình phân tán đầy đủ và Hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình 2PFD.

3.3. Giải pháp đảm bảo tính riêng tư cho Hệ tư vấn người dùng trong mô hình dữ liệu phân tán đầy đủ

3.3.1. Định nghĩa bài toán

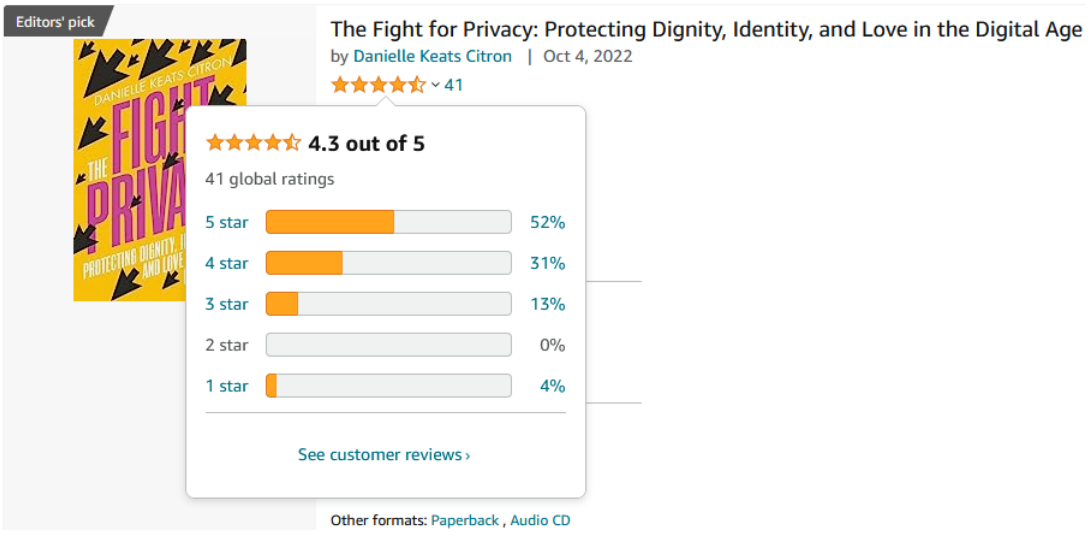
Hệ tư vấn người dùng gồm có hai thực thể cơ bản:

Người dùng: sử dụng hệ thống tư vấn (RS) để đưa ra ý kiến đánh giá cũng như nhận các tư vấn và các mục dữ liệu được người dùng đánh giá. Giả sử trong hệ thống có n người dùng $U = \{U_1, U_2, ..., U_n\}$ và m mục dữ liệu $I = \{i_1, i_2, ..., i_m\}$.

Máy chủ tư vấn: hợp tác với người dùng để tính toán giá trị trung bình xếp hạng của những người dùng đối với mục tin, độ tương tự giữa cặp mục tin, lưu trữ các giá trị này trong cơ sở dữ liệu của mình và hỗ trợ người dùng mục tiêu trong việc sinh tư vấn. Máy chủ tư vấn sẽ không tiết lộ các tham số của hệ tư vấn (trung bình xếp hạng và độ tương tự giữa các mục tin khác nhau) cho bất kỳ ai.

Điều này xuất phát từ thực tế hệ tư vấn chính là bí mật kinh doanh nhằm cải tiến hiệu quả hoạt động của tổ chức. Hơn nữa, để có được một hệ tư vấn có chất lượng tốt thì thông thường các tổ chức sẽ phải đầu tư chi phí đáng kể cho quá trình người dùng đưa ra các đánh giá cho các mục tin của mình.

Mỗi người dùng U_i sẽ đưa ra ý kiến đánh giá đối với các mục dữ liệu thông qua việc chọn câu trả lời cho các câu hỏi liên quan đến việc đánh giá các mục dữ liệu. Các câu trả lời này sẽ tuân theo một thang số cụ thể (ví dụ: 1: kém đến 5: xuất sắc) như ví dụ trong Hình 3.2.



Hình 3.2: Ví dụ về thanh đánh giá đối với mục dữ liệu trong hệ thống thương mại điện tử của Amazon

Đầu vào cho hệ thống tư vấn (RS) là các giá trị xếp hạng số học, tức các giá trị $r_{i,j}$, với $j \in [0,m)$ là giá trị xếp hạng của người dùng U_i cho mục dữ liệu i_j. Nếu người dùng chưa đánh giá mục dữ liệu, thì $r_{i,j} = 0$. Dữ liệu đầu vào gốc cho quá trình xây dựng hệ tư vấn được biểu diễn dưới dạng ma trận R như mô tả trong Bảng 3.2.

Mục đích của giải pháp là cho phép máy chủ sử dụng dữ liệu từ tất cả người dùng để xây dựng hệ tư vấn người dùng trong khi vẫn bảo vệ tính riêng tư của mỗi người dùng. Máy chủ sẽ không tiết lộ các tham số của hệ tư vấn đã được xây dựng. Mỗi người dùng

Bảng 3.2: Bảng ma trận Người dùng - Xếp hạng mục tin

Users/items	i_1	...	i_j	...	i_k	...	i_m
U_1	$r_{1,0}$	...	$r_{1,j}$	...	$r_{1,k}$	...	$r_{1,m-1}$
...	...	...	...	...	...	...	...
U_i	$r_{i,0}$	...	$r_{i,j}$	...	$r_{i,k}$	...	$r_{i,m-1}$
...	...	...	...	...	...	...	...
U_n	$r_{n,0}$	...	$r_{n,j}$	...	$r_{n,k}$	...	$r_{n,m-1}$

u_i nắm giữ các giá trị bí mật $r_{i,j}$. Không ai biết về các giá trị này trừ người dùng đó.

Khi người dùng muốn được tư vấn, họ sẽ gửi các ý kiến đã đánh giá của mình đối với các mục tin (cũng thông qua việc đã trả lời các câu hỏi liên quan) tới máy chủ tư vấn. Đánh giá của người dùng cũng được chuyển đổi sang các giá trị số học cụ thể theo thang số của các câu trả lời đã được định nghĩa trước $r_{i,j}$. Mục tiêu là cho phép máy chủ sử dụng các giá trị này để sinh tư vấn cho người dùng trong khi vẫn đảm bảo tính riêng tư của người dùng, cụ thể là:

- Tính riêng tư về các giá trị bí mật $r_{i,j}$.

- Kết quả tư vấn về các mục tin mà người dùng chưa đánh giá.

3.3.2. Giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng

Trong giải pháp mới này, luận án thiết kế lại cả hai pha của hệ tư vấn người dùng có đảm bảo tính riêng tư trong [7] nhằm tăng hiệu quả của giải pháp. Giải pháp đề xuất cụ thể như sau:

Giai đoạn 1: Để tính trung bình xếp hạng của các mục tin và độ tương tự giữa các cặp mục tin, luận án sử dụng giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có đảm bảo tính riêng tư được đề xuất trong Mục 2.2 (được gọi là giao thức **SMASC**).

Giai đoạn 2: Để sinh tư vấn có đảm bảo tính riêng tư cho người dùng mục tiêu, luận án đề xuất hai phương thức:

- Sinh tư vấn dựa trên CBF có đảm bảo tính riêng tư.

- Sinh tư vấn dựa trên CF có đảm bảo tính riêng tư.

Có ba bước chính để tạo các đề xuất: (1) Người dùng mục tiêu bắt đầu bằng cách mã hóa vectơ xếp hạng của mình và gửi các bản mã đến máy chủ.(2) Máy chủ sử dụng các phép toán đồng cấu và mã hóa để tạo ra các bản mã của tử số và mẫu số của phương trình (1.7.11) hoặc (1.7.12) (tương ứng với phương pháp sinh tư vấn dựa trên CBF hoặc CF). (3) Người dùng lấy các bản mã này từ máy chủ và giải mã chúng để tính điểm đề

xuất.

Trong đó, $P_{i,k}$ là điểm đề xuất được dự đoán cho mục tin i_k với $k \in [1,m]$, và mục tin có điểm số cao nhất cuối cùng được đề xuất cho người dùng.

Gọi max là giá trị lớn nhất trong thang đánh giá mục tin, khi đó: $su_j \in \{0,1,\ldots,n \cdot \texttt{max}^2\}$

Giải pháp **3.1** sẽ trình bày chi tiết về giải pháp này.

3.3.3. Đánh giá về tính đúng đắn

Giải pháp PPRS được đề xuất gồm hai giai đoạn: Giai đoạn thứ nhất là xây dựng hệ tư vấn người dùng thông qua việc tính các giá trị xếp hạng trung bình của mục tin, các giá trị độ tương tự giữa cặp mục tin và Giai đoạn thứ hai là tạo các tư vấn cho người dùng.

Trong đó, các giá trị xếp hạng trung bình của mục tin và độ tương tự giữa các cặp mục tin được tính thông qua việc thực thi giao thức tính nhiều giá trị trung bình và độ tương tự có đảm bảo tính riêng tư đã được đề xuất trong Mục 2.2. Do đó, quá trình này đã được chứng minh là đảm bảo tính đúng đắn của các giá trị tham số của hệ tư vấn người dùng.

Ở giai đoạn thứ hai, giải pháp cung cấp hai cách thức để tạo tư vấn cho người dùng: tư vấn dựa trên **CBF** và **CF**.

Ta có: $d^{(3)} = Dlog_{EC}(\mathbb{E}, G, \max^2, C^{(3)})$

Vì thế, với $0 \leq k < m$, ta có: $d_k^{(3)} \cdot G = C_k^{(3)} = F_{1,k} - x_i \cdot F_{2,k}$

• Nếu tạo tư vấn dựa trên CBF:

$d_k^{(3)} \cdot G = \sum_{j=1}^{m} S(i_k, i_j) \cdot (r_{i,j} \cdot G + c_j^{(1)} \cdot X_i) - x_i \cdot \sum_{j=1}^{m} S(i_k, i_j) \cdot c_j^{(1)} \cdot G$

$= \sum_{j=1}^{m} S(i_k, i_j) \cdot r_{i,j} \cdot G$

Do đó: $d_k^{(3)} = \sum_{j=1}^{m} r_{i,j} \cdot S(i_k, i_j)$

Tương tự, ta có: $d_{k+m}^{(3)} = \sum_{j=1}^{m} S(i_k, i_j)$

Vậy: $P_{i,k} = \frac{d_k^{(3)}}{d_{k+m}^{(3)}} = \frac{\sum_{j=1}^{m} r_{i,j} \cdot S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$

• Nếu tạo tư vấn dựa trên CF:

$d_k^{(3)} \cdot G = F_{5,k} + \sum_{j=1}^{m} (C_j^{(3)} - F_{7,k}) \cdot S(i_k, i_j) - x_i \cdot (F_{6,k} + \sum_{j=1}^{m} (C_j^{(4)} - F_{8,k}) \cdot S(i_k, i_j))$

$= R_k \cdot \sum_{j=1}^{m} S(i_k, i_j) \cdot G + \sum_{j=1}^{m} (r_{i,j} - R_j) \cdot G \cdot S(i_k, i_j)$

Giải pháp 3.1: Giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng trong mô hình dữ liệu phân tán đầy đủ

Đầu vào:
n: Số lượng người dùng.
m: Số lượng các giá trị trung bình cần tính.
RM: Ma trận xếp hạng của những người dùng đối với các mục tin, trong đó mỗi $r_{i,j}$ là giá trị xếp hạng riêng tư của người dùng thứ i đối với mục tin thứ j.
max: Giá trị lớn nhất trong các giá trị bí mật mà các U_i nắm giữ.
k: Chỉ số của mục tin mà người dùng U_i yêu cầu sinh tư vấn.
Đầu ra: (đối với U_i):
$P_{i,k}$: Điểm đề xuất được dự đoán xếp hạng cho mục tin thứ k của người dùng U_i, với $k = \{0,2,\ldots,m-1\}$ là thứ tự mục tin mà U_i đã yêu cầu sinh tư vấn.
Nếu sinh tư vấn dựa trên nội dung – CBF thì: $P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j} S(i_k,i_j)}{\sum_{j=1}^{m} S(i_k,i_j)}$
Nếu sinh tư vấn dựa trên lọc cộng tác – CF thì: $P_{i,k} = \frac{R_k \sum_{j=1}^{m} S(i_k,i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) S(i_k,i_j)}{\sum_{j=1}^{m} S(i_k,i_j)}$
R_k, R_j được tính theo công thức (1.7.13) và $S(i_k,i_j)$ được tính theo công thức (1.7.14).

Giai đoạn 1: Tính trung bình xếp hạng của các mục tin và độ tương tự giữa cặp mục tin

1. $\{R_j, S(i_j,i_k)\}_{0 \le j,k < m; j<k} =$ SMASC$(n,m,RM,n.\max^2)$;
/* Thực hiện giao thức SMASC
2. **for** $(0 \le j < m-1)$
3. **for** $(j+1 \le k < m)$
4. $S(i_k,i_j) = S(i_j,i_k)$;

Giai đoạn 2: Sinh tư vấn cho người dùng mục tiêu U_i

*** Pha 1. U_i thực hiện**

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 là thuật toán mã hoá của hệ mật mã đường cong Elliptic.*/
8. Gửi $\{E(r_{i,j} \cdot G)\}_{0 \le j < m}$ cho Rs;

*** Pha 2. Rs thực hiện**

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** (Phương thức sinh tư vấn = 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. Gửi cho U_i: $\{F_{1,k}, F_{2,k}, F_{3,k}, F_{4,k}\}_{0 \le k < m}$;

*** Pha 3. U_i thực hiện** 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. $P_{i,k} = \frac{d_k^{(3)}}{d_{k+m}^{(3)}}$;

$= \left(R_k \cdot \sum_{j=1}^{m} S(i_k, i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) \cdot S(i_k, i_j)\right) \cdot G$

Do đó: $d_k^{(3)} = R_k \cdot \sum_{j=1}^{m} S(i_k, i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) \cdot S(i_k, i_j)$

Tương tự, ta có: $d_{k+m}^{(3)} = \sum_{j=1}^{m} S(i_k, i_j)$

Vậy: $P_{i,k} = \frac{d_k^{(3)}}{d_{k+m}^{(3)}} = \frac{R_k \cdot \sum_{j=1}^{m} S(i_k,i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) \cdot S(i_k,i_j)}{\sum_{j=1}^{m} S(i_k,i_j)}$

Kết luận 3.1. *Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư được đề xuất cho mô hình dữ liệu phân tán đầy đủ tính đúng đắn của giá trị điểm tư vấn:*

- *Với tư vấn dựa trên CBF:* $P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j} \cdot S(i_k,i_j)}{\sum_{j=1}^{m} S(i_k,i_j)}$
- *Với tư vấn dựa trên CF:* $P_{i,k} = \frac{R_k \cdot \sum_{j=1}^{m} S(i_k,i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) \cdot S(i_k,i_j)}{\sum_{j=1}^{m} S(i_k,i_j)}$ ■

Về mặt lý thuyết luận án đã chứng minh giải pháp đề xuất bảo toàn tính đúng đắn của kết quả tư vấn. Tuy nhiên do hệ mật mã trên đường cong Elliptic chỉ làm việc với bản rõ là số nguyên nên trong quá trình sinh tư vấn đôi khi phải thực hiện làm tròn các giá trị trung bình và độ tương tự, vì thế sẽ có sự sai lệch nhỏ trong kết quả điểm tư vấn $P_{i,k}$.

Tiếp theo, tính riêng tư của giải pháp sẽ được phân tích, đánh giá.

3.3.4. Đánh giá về tính riêng tư

Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ được đề xuất ở trên được tạo nên từ việc thực thi tuần tự hai giai đoạn: xây dựng hệ tư vấn và sinh tư vấn.

Đầu tiên, giai đoạn xây dựng hệ tư vấn thông qua việc sử dụng giao thức tính nhiều giá trị trung bình và độ tương tự có đảm bảo tính riêng tư như được mô tả trong Mục 2.2. Giai đoạn này đã được chứng minh là: *"đảm bảo tính riêng tư của mỗi người dùng trong mô hình bán trung thực trong khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng của lên đến tối đa $n-2$ người dùng, giao thức vẫn bảo vệ tính riêng tư của các bên trung thực. Đồng thời giao thức cũng bảo vệ tính riêng tư của các bên trung thực trước hành vi độc hại của trung tâm tính toán nhằm tìm hiểu các giá trị riêng tư của người dùng theo kịch bản cụ thể đã được chỉ ra trong tài liệu [55]."*

Tiếp theo, mức độ an toàn trong giai đoạn sinh tư vấn với cả hai phương pháp CBF và CF được đánh giá.

Trong pha 1, người dùng mục tiêu sẽ mã hóa giá trị xếp hạng cho các mục tin bằng

khóa công khai của mình:

$$(C_j^{(1)} = r_{i,j}G + c_j^{(1)}X_i, \quad C_j^{(2)} = c_j^{(1)}G) \tag{3.3.1}$$

và gửi chúng đến máy chủ ở pha 1. Vì thế không ai (kể cả máy chủ tư vấn) ngoài người dùng mục tiêu có thể giải mã để biết được giá trị bí mật. Hơn nữa, vì $x_i, c_j^{(1)} \in_R \mathbb{Z}_d^*$ nên theo giả thuyết Diffie-Hellman quyết định [11] ta có:

$$Pr(t \leftarrow \{C_j^{(1)}, C_j^{(2)}\}_{j\in[0,m)}, D(t) = 1) = \frac{1}{|\mathbb{Z}_d^*|} \tag{3.3.2}$$

Trong pha 2, máy chủ sử dụng các giá trị độ tương tự của mục tin-mục tin, vốn đã có sẵn ở máy chủ để tính các giá trị $(F_{1,k}, F_{2,k}, F_{3,k}, F_{4,k})_{0\le j,k<m, j\neq k}$ và gửi các giá trị này đến người dùng mục tiêu. Trong đó:

- $(F_{3,k} = \sum_{j=1}^{m} s(i_k, i_j) \cdot G + c_k^{(2)}X_i, F_{4,k} = c_k^{(2)}G)$ chính là bản mã của giá trị bản rõ $\sum_{j=1}^{m} s(i_k, i_j) \cdot G$ sử dụng khóa công khai của người dùng mục tiêu X_i. Vì thế cũng chỉ có người dùng mục tiêu mới có thể giải mã để thu được giá trị bản rõ $\sum_{j=1}^{m} s(i_k, i_j) \cdot G$.

Do $c_k^{(2)} \in_R \mathbb{Z}_d^*$ nên theo giả thuyết Diffie-Hellman quyết định [11] ta có:

$$\Pr\left(t \leftarrow \{F_{3,k}, F_{4,k}\}_{k\in[0,m)}, D(t) = 1\right) = \frac{1}{|\mathbb{Z}_d^*|} \tag{3.3.3}$$

- Để tạo các đề xuất dựa trên CBF: $(F_{1,k} = \sum_{j=1}^{m} s(i_k, i_j) \cdot C_j^{(1)}, F_{2,k} = \sum_{j=1}^{m} s(i_k, i_j) \cdot C_j^{(2)})$ chính là bản mã của bản rõ $\sum_{j=1}^{m} s(i_k, i_j) \cdot r_{i,j}G$ sử dụng khóa công khai của người dùng mục tiêu X_i. Vì thế cũng chỉ có người dùng mục tiêu mới có thể giải mã để thu được giá trị bản rõ $\sum_{j=1}^{m} s(i_k, i_j) \cdot r_{i,j}G$. Kết hợp với công thức (3.3.3) ta có:

$$\Pr\left(t \leftarrow \{F_{1,k}, F_{2,k}\}_{k\in[0,m)}, D(t) = 1\right) = \frac{1}{|\mathbb{Z}_d^*|} \tag{3.3.4}$$

- Để tạo các đề xuất dựa trên CF: $(F_{1,k} = F_{5,k} + \sum_{j=1}^{m}(C_j^{(3)} - F_{7,k})s(i_k, i_j)$
$= R_k \sum_{j=1}^{m} s(i_k, i_j)G + c_k^{(4)}X_i + \sum_{j=1}^{m}(r_{i,j}G + c_j^{(3)}X_i - R_jG - c_k^{(5)}X_i)s(i_k, i_j) =$
$\left(R_k \sum_{j=1}^{m} s(i_k, i_j) + \sum_{j=1}^{m}(r_{i,j}G - R_jG)s(i_k, i_j)\right)G + \left(c_k^{(4)} + \sum_{j=1}^{m}(c_j^{(3)} - c_k^{(5)})s(i_k, i_j)\right)X_i,$
$F_{2,k} = F_{6,k} + \sum_{j=1}^{m}(C_j^{(4)} - F_{8,k})s(i_k, i_j) = \left(c_k^{(4)} + \sum_{j=1}^{m}(c_j^{(3)} - c_k^{(5)})s(i_k, i_j)\right)G)$ chính là bản mã của bản rõ $\left(R_k \sum_{j=1}^{m} s(i_k, i_j) + \sum_{j=1}^{m}(r_{i,j}G - R_jG)s(i_k, i_j)\right)G$ sử dụng khóa công khai của người dùng mục tiêu X_i, với giá trị ngẫu nhiên là $\left(c_k^{(4)} + \sum_{j=1}^{m}(c_j^{(3)} - c_k^{(5)})s(i_k, i_j)\right)$. Vì $c_j^{(2)}, c_k^{(4)}$, và $c_k^{(5)} \in_R \mathbb{Z}_d^*$, do đó:

$$\Pr\left(t \leftarrow \{F_{1,k}, F_{2,k}\}_{k\in[0,m)}, D(t) = 1\right) = \frac{1}{|\mathbb{Z}_d^*|} \tag{3.3.5}$$

Trong pha 3, các bản mã được giải mã ở phía người dùng mục tiêu sử dụng khóa bí mật x_i của người dùng mục tiêu. Vì thế chỉ người dùng mục tiêu mới biết được kết quả tư vấn ở dạng rõ.

Từ công thức (3.3.2)-(3.3.5) có thể thấy trong quá trình sinh tư vấn, kết quả đề xuất và xếp hạng cá nhân của người dùng mục tiêu không bị tiết lộ và việc sinh tư vấn mục tin được tiến hành ở phía người dùng mục tiêu.

Kết hợp với Định lý 1.1, ta có kết luận sau:

Kết luận 3.2. *Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư đề xuất cho mô hình dữ liệu phân tán đầy đủ ở trên đảm bảo tính riêng tư của mỗi người dùng trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng giữa các bên, giao thức này đảm bảo tính riêng tư của các bên trung thực trước sự thông đồng của tối đa $n-2$ người dùng thông đồng. Hơn nữa, giải pháp còn bảo vệ tính riêng tư của mỗi người dùng trung thực trước hành vi độc hại của máy chủ nhằm tìm hiểu các giá trị riêng tư của người dùng theo kịch bản cụ thể đã được chỉ ra trong tài liệu [55].* ■

3.3.5. Đánh giá tính hiệu quả

Trong phần này, giải pháp đề xuất được so sánh với giải pháp ban đầu [7] và giải pháp [75] về chi phí truyền thông và thời gian thực thi. Tất cả người dùng được giả định cùng tham gia để tính các giá trị trung bình xếp hạng và độ tương tự giữa các mục tin trong hệ thống và chỉ một người dùng (người dùng mục tiêu) tham gia tạo tư vấn.

Các giải pháp [7, 75] sẽ sử dụng hệ mật mã Elgamal kích thước khoá bí mật 160 bit và khoá công khai 1024 bit [42], vì thế mỗi thông điệp sẽ có kích thước là $|p| = 1024$ bít. Giải pháp đề xuất sử dụng hệ mật mã trên đường cong Elliptic sẽ sử dụng đường cong Elliptic BrainpoolP160r1 với kích thước khoá bí mật là 160 bit, khoá công khai là 320 bit [45] (hai hệ mật này có cùng mức an toàn là 80 bit).

3.3.5.1. Đánh giá về chi phí truyền thông

Chi phí truyền thông được tính toán dựa trên số lượng mục tin và người dùng trong hệ thống. Trong giai đoạn sinh tư vấn, cả ba phương pháp đều có số lượng các thông điệp được trao đổi giữa các bên là như nhau: người dùng gửi $2m$ thông điệp và máy chủ gửi $4m$ thông điệp. Tuy nhiên, mỗi thông điệp trong giải pháp đề xuất là một điểm trên

đường cong Elliptic (gồm hai giá trị tọa độ) nên có kích thước là $2 \times 160 = 320$ bit, chỉ bằng $\frac{5}{16}$ kích thước của một thông điệp trong các giải pháp còn lại.

Bảng 3.4: So sánh chi phí truyền thông giữa các giải pháp

		Tính trung bình, độ tương tự	Sinh tư vấn CBF	Sinh tư vấn CF
Giải pháp [7, 75]	Người dùng	$\|p\|(1,5m^2+7,5m+1)$	$2m\|p\|$	$2m\|p\|$
	Máy chủ	$n\|p\|(0,5m^2+2,5m+1)$	$4m\|p\|$	$4m\|p\|$
Giải pháp đề xuất	Người dùng	$\|p\|(0,32m^2 + 1,25m + 1,1)$	$0,63m\|p\|$	$0,63m\|p\|$
	Máy chủ	$n\|p\|(3,2m+1,1)$	$1,25m\|p\|$	$1,25m\|p\|$

Từ bảng trên cho thấy giải pháp đề xuất có chi phí truyền thông hiệu quả hơn các giải pháp còn lại [7, 75]. Ví dụ chi phí truyền thông của mỗi người dùng trong các giải pháp này cao hơn 3 lần so với giải pháp đề xuất trong tất cả các giai đoạn của hệ tư vấn.

3.3.5.2. Đánh giá về chi phí tính toán

Chi phí tính toán của các giao thức chủ yếu liên quan đến thời gian thực thi các phép toán như: phép nhân, phép lũy thừa, phép nghịch đảo modulo của số nguyên lớn, phép tính logarit rời rạc, phép cộng điểm và phép nhân điểm trên đường cong Elliptic.

Từ Bảng 3.5, có thể thấy chi phí tính toán của giải pháp mới thấp hơn so với các giải pháp [7, 75], mang lại hiệu quả cao hơn. Để tính các giá trị trung bình và độ tương tự giữa các mục tin, chi phí tính toán của mỗi người dùng đối với giải pháp gốc khoảng $480,5m^2$, với giải pháp cải tiến [75] khoảng $601m^2$, trong khi đó theo giải pháp của luận án khoảng $43,62m^2$ phép nhân modulo. Tương tự ta có thể thấy chi phí tính toán của người dùng và máy chủ trong giải pháp đề xuất thấp hơn nhiều so với các giải pháp còn lại trong tất cả các giai đoạn của hệ tư vấn.

3.3.5.3. Đánh giá về mặt thực nghiệm

Trong các thử nghiệm này, 500 mục tin và 943 người dùng đã xếp hạng cho các mục tin đó (tương ứng với 30.442 xếp hạng) được chọn. Do đó, việc phân tích hiệu suất của giải pháp gồm 943 người dùng và 500 mục tin. Chi phí thời gian tính trung bình xếp hạng và độ tương tự giữa các mục tin. Mỗi giải pháp được thực thi với số lượng mục tin thay đổi từ 100 đến 500.

Bảng 3.5: So sánh chi phí tính toán giữa các giải pháp

		Tính trung bình, độ tương tự	Sinh tư vấn CBF	Sinh tư vấn CF
Giải pháp [7]	Người dùng	$(480,5m^2 + 2402,5m + 240)T_m$	$m(5T_e + T_m + 2T_i) + 2mT_{DL} \approx 1224,2mT_m + 2mT_{DL}$	
	Máy chủ	$n(7,3m^2 + 36,5m + 1)T_m + (0,5m^2 + 2,5m)T_{DL}$	$m((2m+3)T_e + (2m-1)T_m) \approx m(482m+719)T_m$	$m((2m+9)T_e + (2m+3)T_m + 2mT_i) \approx m(504,6m+2163)T_m$
Giải pháp [75]	Người dùng	$(601m^2 + 3005m + 240)T_m$	$m(5T_e + T_m + 2T_i) + 2mT_{DL} \approx 1224,2mT_m + 2mT_{DL}$	
	Máy chủ	$n(7,3m^2 + 36,5m + 1)T_m + (0,5m^2 + 2,5m)T_{DL}$	$m((2m+3)T_e + (2m-1)T_m) \approx (482m+719)T_m$	$m((2m+9)T_e + (2m+3)T_m + 2mT_i) \approx m(504,6m+2163)T_m$
Giải pháp đề xuất	Người dùng	$(43,62m^2 + 261,6m + 101,5)T_m$	$m(5T_{mp} + 3T_{ap}) + T_{DL} \approx 145,36mT_m + T_{DL}$	
	Máy chủ	$n(0,06m^2 + 0,48m + 0,42)T_m + T_{DL}$	$m((2m+3)T_{mp} + (2m-1)T_{ap}) \approx (58,24m+86,88)mT_m$	$m((2m+9)T_{mp} + (4m+3)T_{ap}) \approx (58,48m+261,36)mT_m$

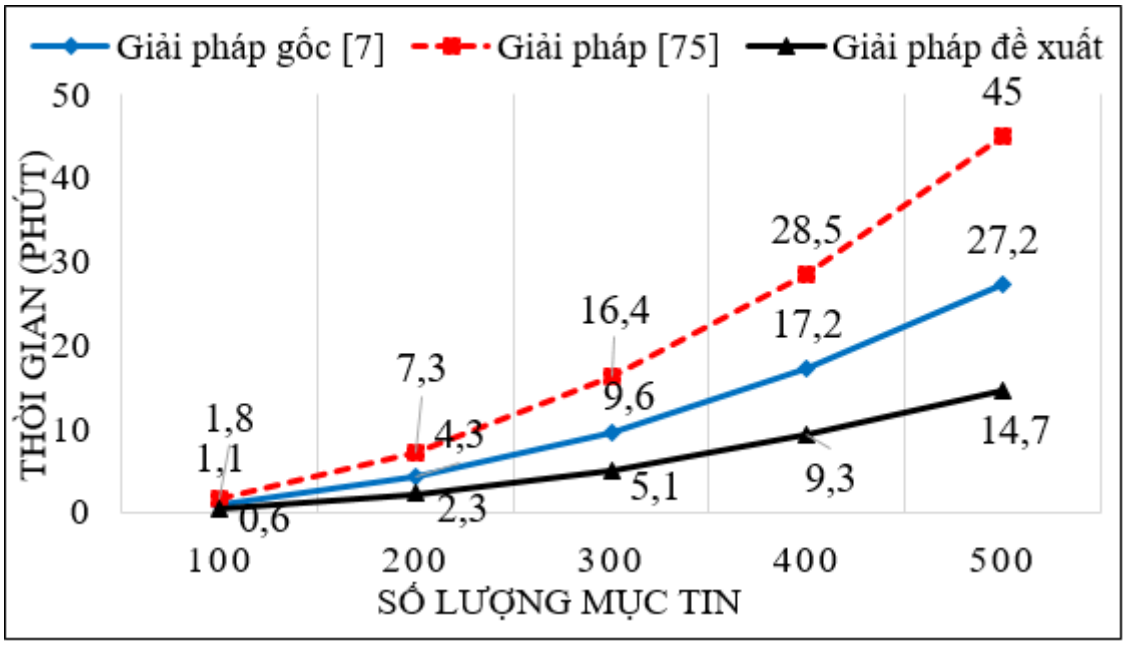
Các giải pháp cần so sánh sẽ được thực thi bằng ngôn ngữ C# trong môi trường Visual Studio 2019, với sự hỗ trợ của thư viện System.Numerics để so sánh hiệu suất giữa các giải pháp này. Các thực nghiệm cùng chạy trên một máy tính xách tay Intel Core i5-4210M CPU 2.60GHz, 8GB RAM. Để giảm thiểu sự sai khác về môi trường thực nghiệm, mỗi giao thức sẽ được thực thi độc lập trong điều kiện không có ứng dụng nào khác được chạy trên máy tính. Thời gian thực thi của các giải pháp sẽ được đo bằng cách sử dụng lớp Stopwatch trong thư viện System.Diagnostics.

Thời gian thực thi

Để so sánh thời gian chạy của các giải pháp, mỗi giải pháp được thực thi 10 lần cho mỗi bộ tham số và lấy giá trị trung bình.

Đầu tiên, thời gian thực thi của mỗi người dùng trong các giải pháp được so sánh,

kết quả chi tiết được trình bày trong Hình 3.3.



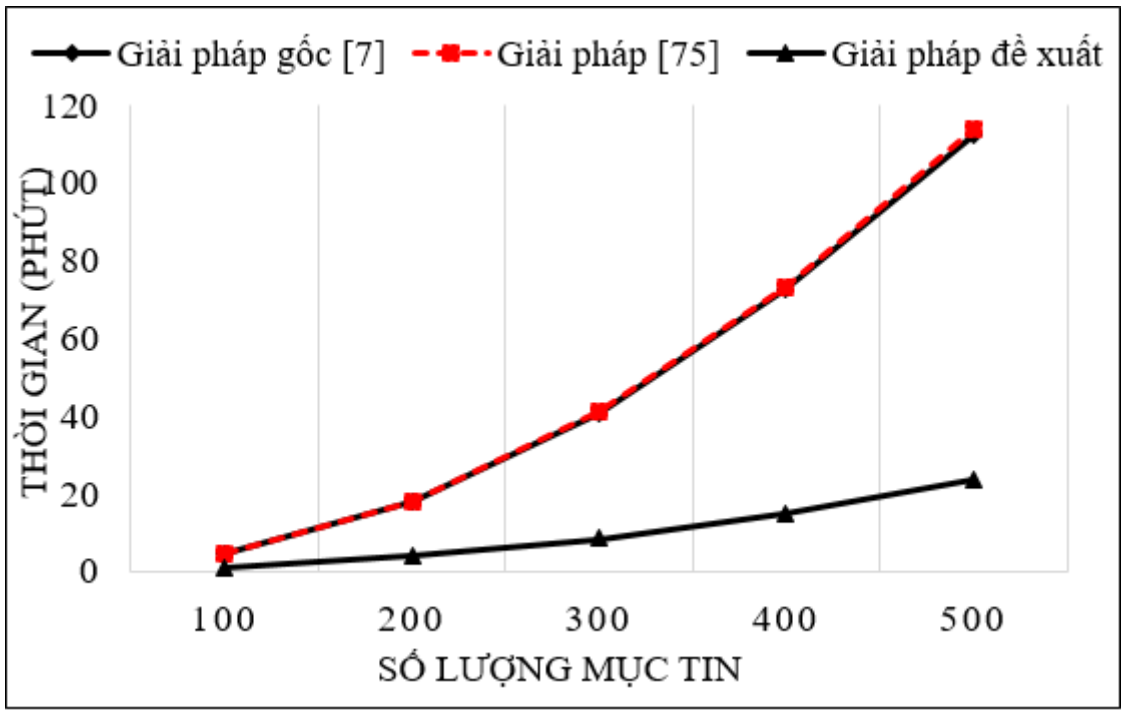
Hình 3.3: Thời gian thực thi của mỗi người dùng trong quá trình tính các giá trị trung bình xếp hạng và độ tương tự giữa các mục tin

Kết quả trong hình này cho thấy với 500 mục tin, mỗi người dùng trong giải pháp đề xuất chỉ mất 14,7 phút để tính giá trung bình xếp hạng và độ tương tự cho tất cả mục tin, trong khi đó giao thức [75] mất 45 phút và giao thức gốc [7] mất 27,2 phút. Điều này khẳng định hiệu quả cao của giải pháp đề xuất. Khi số lượng các mục tin tăng lên, thời gian thực thi của mỗi người dùng sẽ tăng theo hàm số mũ.

Tiếp theo, thời gian thực thi của máy chủ tư vấn trong các giải pháp được so sánh.

Hình 3.4 cho thấy thời gian thực thi của máy chủ trong quá trình tính các giá trị trung bình xếp hạng cho các mục tin và độ tương tự giữa các mục tin. Kết quả trong hình này cho thấy với 500 mục tin và số lượng người dùng tham gia xếp hạng là 943, máy chủ tư vấn trong giải pháp đề xuất chỉ mất 23,5 phút để tính giá trung bình xếp hạng và độ tương tự cho tất cả mục tin, trong khi đó các giao thức còn lại mất 113,9 và 112,5 phút. Điều này khẳng định hiệu quả cao của các giao thức đề xuất.

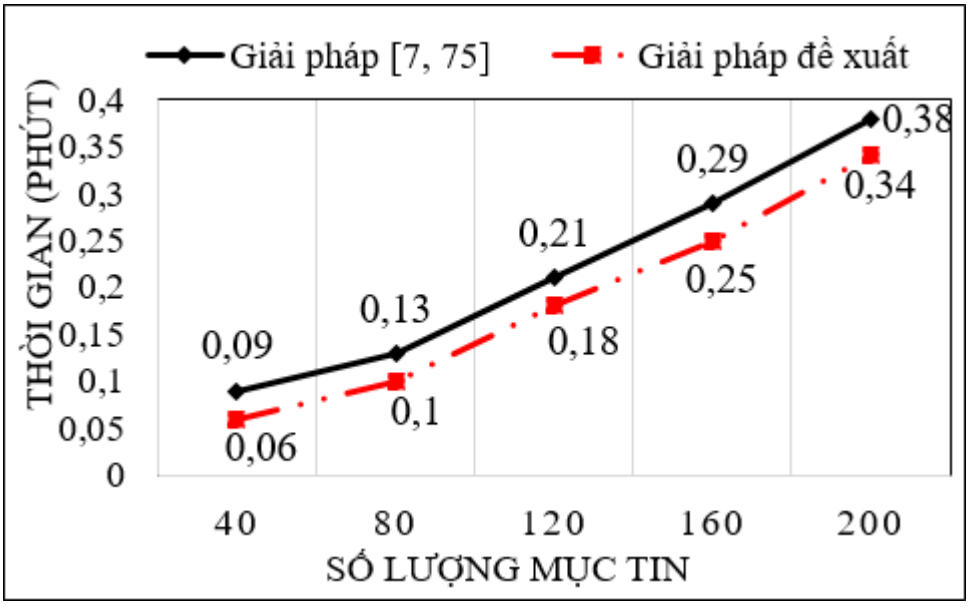
Từ những phân tích này, có thể nhận thấy giải pháp đề xuất hiệu quả hơn giải pháp ban đầu, đặc biệt là khi số lượng mục tin lớn, thời gian thực thi của máy chủ tư vấn trong quá trình tính trung bình xếp hạng và độ tương tự giữa các mục tin sẽ tăng theo hàm mũ. Kết quả này có được là do việc thiết kế lại các bước tính toán trong giao thức pha 1, pha 3 và pha 4 trong giao thức gốc, việc tối ưu hoá số lượng cặp khoá bí mật, công khai cần



Hình 3.4: Thời gian thực thi của máy chủ tư vấn trong quá trình tính trung bình xếp hạng và độ tương tự giữa các mục tin

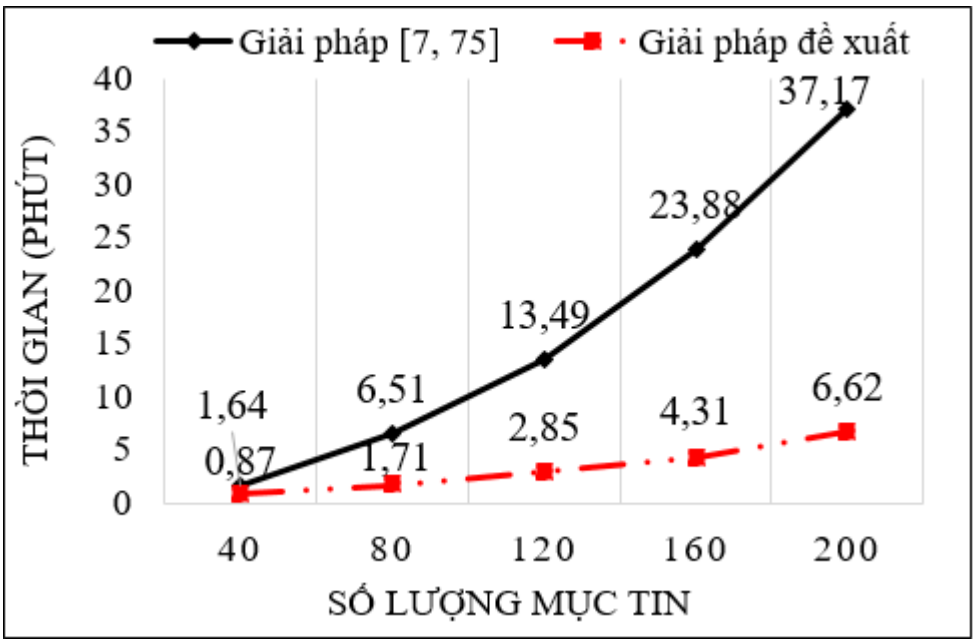
sử dụng và việc chuyển đổi giao thức ban đầu [7] sang hệ mật mã đường cong Elliptic.

Cuối cùng, thời gian tính toán thực tế cần thiết để tạo tư vấn với số lượng mục tin thay đổi từ 40 đến 200 được đo. Các kết quả thu được như trong Hình 3.5 và Hình 3.6.



Hình 3.5: So sánh về thời gian sinh tư vấn CBF

Các số liệu này cho thấy rằng: giải pháp mới có thời gian sinh tư vấn CBF và CF với



Hình 3.6: So sánh về thời gian sinh tư vấn CF

200 mục tin lần lượt là 0,34 và 6,62 phút, trong khi thời gian này trong giải pháp [7, 75] lên tới 0,38 và 37,17 phút. Nguyên nhân của kết quả này là do việc chuyển đổi giải pháp ban đầu [7] sang một biến thể sử dụng hệ mật trên đường cong Elliptic. Điều này cho thấy tính hiệu quả của giải pháp mới trong quá trình sinh tư vấn.

Có thể thấy rằng khi số lượng mục tin càng lớn thì giải pháp mới càng hiệu quả hơn các giải pháp còn lại [7, 75]. Thời gian thực thi sẽ tăng theo hàm luỹ thừa của số lượng mục tin.

Kết quả thực nghiệm cho thấy giải pháp đề xuất hiệu quả hơn các giải pháp điển hình hiện có. Thời gian thực thi của mỗi người dùng và máy chủ sẽ tăng theo hàm số luỹ thừa của số lượng mục tin. Khi số lượng người dùng không đổi, việc tăng số lượng mục sẽ khiến chi phí tính toán cho mỗi người dùng và máy chủ tăng theo tỷ lệ thuận với mức tăng số lượng mục. Trong hệ thống đề xuất, trong đó số lượng mục là cố định, việc tăng số lượng xếp hạng của người dùng sẽ ảnh hưởng đến thời gian thực hiện của máy chủ trong khi vẫn giữ nguyên chi phí thực hiện cho mỗi người dùng. Thời gian thực hiện của máy chủ tăng theo tỷ lệ thuận với số lượng người dùng.

Chất lượng tư vấn

Để đánh giá chất lượng của giải pháp đề xuất, chúng tôi sử dụng Sai số bình phương trung bình (RMSE), căn bậc hai của giá trị trung bình của các lỗi bình phương và Sai số tuyệt đối trung bình (MAE), dùng để đo độ lớn trung bình của các lỗi. Sai số được tính là chênh lệch giữa xếp hạng thực và xếp hạng dự đoán.

Trong thực nghiệm, giải pháp đề xuất đạt được các giá trị MAE lần lượt là 0,997753 và 0,860595 cho các tư vấn dựa trên CBF và dựa trên CF, và các giá trị RMSE lần lượt là 1,276316 và 0,942845 cho các tư vấn dựa trên CBF và dựa trên CF. Các kết quả này bằng với các kết quả của giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư gốc [7] được thử nghiệm trên cùng một tập dữ liệu xếp hạng phim. So với hệ tư vấn người dùng nguyên thuỷ không có ràng buộc về quyền riêng tư (đạt được MAE = 0,997749 và 0,860618, RMSE = 1,276310 và 0,942845 cho các khuyến nghị dựa trên CBF và dựa trên CF, tương ứng), có thể thấy rằng việc làm tròn đến hai chữ số thập phân có tác động không đáng kể đến độ chính xác của giải pháp. Hơn nữa, rõ ràng là khuyến nghị dựa trên CF tạo ra sai số ít hơn (0,860595 và 0,942845) so với khuyến nghị dựa trên CBF (0,997753 và 1,276310).

3.3.6. Ví dụ minh hoạ

Để mô tả rõ ràng giải pháp đề xuất, luận án sử dụng ma trận xếp hạng kích thước 3×3 như ví dụ minh hoạ trong mục 2.2.8. Trong đó, người dùng và mục tin được biểu diễn dưới dạng $U_i = \{U_1, U_2, U_3\}, \quad i_j = \{i_1, i_2, i_3, i_4\}$ tương ứng. Trước khi bắt đầu, các bên tham gia thống nhất các tham số đường cong Elliptic $(\mathbb{E}, G, d, p_e)$.

Giai đoạn 1: Tính trung bình xếp hạng của các mục tin và độ tương tự giữa cặp mục tin

Thực thi giao thức SMASC theo các bước được mô tả như trong mục 2.2.8 và có kết quả các tham số của hệ tư vấn như sau:

$$R_1 = 2,5, \quad R_2 = 3, \quad R_3 = 3,5,$$

$$S(i_1, i_2) = 0,99, \quad S(i_1, i_3) = 0,21, \quad S(i_2, i_3) = 0,35.$$

Vì hệ mật mã đường cong Elliptic không thể xử lý số thập phân nên máy chủ sẽ nhân các giá trị trung bình xếp hạng và độ tương tự giữa các mục tin khác nhau với 100 trước khi lưu vào cơ sở dữ liệu của mình, tương ứng:

$$R_1 = 250, \quad R_2 = 300, \quad R_3 = 350,$$

$$S(i_2, i_1) = S(i_1, i_2) = 99,$$

$$S(i_3, i_1) = S(i_1, i_3) = 21,$$

$$S(i_3, i_2) = S(i_2, i_3) = 35.$$

Giai đoạn 2: Sinh tư vấn cho người dùng mục tiêu

Luận án giả định rằng người dùng mục tiêu đã yêu cầu tư vấn cho tất cả các mục tin và gửi xếp hạng được mã hóa của mình đến máy chủ. Trong cả hai loại đề xuất, máy chủ tạo ra hai bản mã khác nhau (tử số và mẫu số của phương trình (1.7.11) và (1.7.12) và lặp lại quy trình cho tất cả các mục tin. Cuối cùng, người dùng mục tiêu giải mã kết quả và chọn mục tin có điểm tư vấn cao nhất. Quá trình để tạo tư vấn có đảm bảo tính riêng tư được minh hoạ bằng số cụ thể dưới đây.

Để minh hoạ, luận án trình bày quá trình sinh tư vấn có đảm bảo tính riêng tư cho người dùng U_1 trên mục i_1. Đầu tiên, người dùng U_1 gửi thông điệp chứa các bản mã xếp hạng của mình như sau:

$\{E(3\cdot G), E(5\cdot G), E(0\cdot G)\}$

Máy chủ tư vấn mã hoá giá trị tương ứng với phần mẫu trong phương trình (1.7.11) và (1.7.12) sử dụng khoá công khai của người dùng U_1 như sau:

$(F_{3,1}, F_{4,1}) = E((99+21)\cdot G) = E(120\cdot G)$

3.3.6.1. Sinh tư vấn dựa trên CBF

Máy chủ tư vấn sử dụng thuộc tính đồng cấu để tính bản mã tương ứng với phần tử trong phương trình (1.7.11) như sau:

$(F_{1,1}, F_{2,1}) = 99\cdot E(5\cdot G) + 21\cdot E(0\cdot G) = E(495\cdot G)$

Máy chủ gửi các thông điệp $\{F_{1,k}, F_{2,k}, F_{3,k}, F_{4,k}\}_{1\le k\le 3}$ cho người dùng U_1.

Người dùng U_1 giải mã cục bộ các bản mã sử dụng khoá bí mật của mình $ksu_{1,1}$ như sau:

$C_1^{(3)} = F_{1,1} - ksu_{1,1}\cdot F_{2,1} = 495\cdot G$

$C_1^{(4)} = F_{3,1} - ksu_{1,1}\cdot F_{4,1} = 120\cdot G$

Người dùng U_1 tính logarit rời rạc kết quả giải mã để thu được giá trị tử và mẫu thực sự của phương trình (1.7.11) như sau:

$d_1^{(3)} = Dlog_{EC}(E, G, 495\cdot G) = 495$

$d_1^{(4)} = Dlog_{EC}(E, G, 120\cdot G) = 120$

Cuối cùng, người dùng U_1 tính giá trị tư vấn dự đoán cho mục tin i_1:

$P_{1,1} = \frac{495}{120} = 4,125$

Theo cách tương tự, giá trị tư vấn dự đoán cho tất cả các mục tin khác được tính:

$P_{1,2} = 2.22, \quad P_{1,3} = 4,25$

Vì mục tin i_3 có điểm tư vấn cao nhất nên nó được tư vấn cho người dùng U_1.

3.3.6.2. Sinh tư vấn dựa trên CF

Máy chủ tư vấn sử dụng khoá công khai của người dùng U_1 để mã hoá như sau:

$E(250 \cdot (99+21) \cdot G) = E(30000 \cdot G)$

$E(R_1 \cdot G) = E(250 \cdot G)$

Sau đó, máy chủ sử dụng thuộc tính đồng cấu để tính bản mã tương ứng với phần tử trong phương trình (1.7.12) như sau:

$(F_{1,1}, F_{2,1}) = E(30000 \cdot G) + 99 \cdot (100 \cdot E(5 \cdot G) - E(300 \cdot G)) + 21 \cdot (100 \cdot E(0 \cdot G) - E(350 \cdot G)) = E(42450 \cdot G)$

Máy chủ gửi các thông điệp $\{F_{1,k}, F_{2,k}, F_{3,k}, F_{4,k}\}_{1 \le k \le 3}$ cho người dùng U_1.

Người dùng U_1 giải mã cục bộ các bản mã sử dụng khoá bí mật của mình $ksu_{1,1}$ như sau:

$C_1^{(3)} = F_{1,1} - ksu_{1,1} \cdot F_{2,1} = 42450 \cdot G$

$C_1^{(4)} = F_{3,1} - ksu_{1,1} \cdot F_{4,1} = 120 \cdot G$

Người dùng U_1 tính logarit rời rạc kết quả giải mã để thu được giá trị tử và mẫu thực sự của phương trình (1.7.11) như sau:

$d_1^{(3)} = Dlog_{EC}(\mathbb{E}, G, 42450 \cdot G) = 42450$

$d_1^{(4)} = Dlog_{EC}(\mathbb{E}, G, 120 \cdot G) = 120$

Cuối cùng, người dùng U_1 tính giá trị tư vấn dự đoán cho mục tin i_1:

$P_{1,1} = \frac{42450}{120} = 353,75 \approx 3,54$

Theo cách tương tự, giá trị tư vấn dự đoán cho tất cả các mục tin khác được tính:

$P_{1,2} = 2.46, \quad P_{1,3} = 4,94$

Vì mục tin i_3 có điểm tư vấn cao nhất nên nó được tư vấn cho người dùng U_1.

3.4. Giải pháp đảm bảo tính riêng tư cho Hệ tư vấn người dùng trong mô hình dữ liệu 2PFD

3.4.1. Định nghĩa bài toán

Giả sử cần xây dựng hệ tư vấn người dùng có m mục tin, trong đó dữ liệu đánh giá về các mục tin này là sở hữu của từng cặp người dùng (U_i, V_i) thuộc hai miền dữ liệu khác nhau (dữ liệu thuộc hai tổ chức khác nhau) và có n_1 cặp người dùng tham gia đánh giá.

Mỗi người dùng U_i sẽ đưa ra ý kiến đánh giá đối với tập con gồm m_1 mục dữ liệu thông qua việc chọn câu trả lời cho các câu hỏi liên quan đến việc đánh giá các mục dữ liệu ($m_1 < m$), trong đó các câu trả lời này sẽ tuân theo một thang số cụ thể (ví dụ: 1: kém đến 5: xuất sắc). Đầu vào cho hệ tư vấn người dùng là các giá trị xếp hạng số học này, tức các giá trị $r_{i,j}, j \in [0, m_1)$ là giá trị xếp hạng của người dùng U_i cho mục dữ liệu i_j, nếu người dùng chưa đánh giá mục dữ liệu thì $r_{i,j} = 0$.

Tương tự, mỗi người dùng V_i sẽ đưa ra ý kiến đánh giá đối với tập con gồm $m - m_1$ mục dữ liệu còn lại thông qua việc chọn câu trả lời cho các câu hỏi liên quan đến việc đánh giá các mục dữ liệu, trong đó các câu trả lời này sẽ tuân theo một thang số cụ thể (ví dụ: 1: kém đến 5: xuất sắc). Đầu vào cho hệ tư vấn người dùng là các giá trị xếp hạng số học này, tức các giá trị $r_{i,j}, j \in [m_1, m-1]$ là giá trị xếp hạng của người dùng V_i cho mục dữ liệu i_j, nếu người dùng chưa đánh giá mục dữ liệu thì $r_{i,j} = 0$.

Dữ liệu đầu vào gốc cho quá trình xây dựng hệ tư vấn người dùng chính là ma trận R như được mô tả trong Bảng 3.6. Mục đích là cho phép bên máy chủ sử dụng dữ liệu từ tất cả người dùng để xây dựng hệ tư vấn người dùng trong khi vẫn bảo vệ tính riêng tư của mỗi người dùng. Mỗi người dùng U_i nắm giữ các giá trị bí mật $r_{i,j}$. Không ai biết về các giá trị này trừ người dùng đó.

Khi người dùng muốn được tư vấn, người dùng sẽ gửi các ý kiến đã đánh giá của mình đối với các mục tin (cũng thông qua việc đã trả lời các câu hỏi liên quan) tới máy chủ tư vấn. Đánh giá của người dùng cũng được chuyển đổi sang các giá trị số học cụ thể theo thang số của các câu trả lời đã được định nghĩa trước $r_{i,j}$. Mục tiêu là cho phép máy chủ sử dụng các giá trị này để sinh tư vấn cho người dùng trong khi vẫn đảm bảo tính riêng tư của người dùng. Tính riêng tư về các giá trị bí mật $r_{i,j}$ và kết quả tư vấn về các mục tin mà người dùng chưa đánh giá.

Mỗi người dùng U_i nắm giữ các giá trị bí mật $r_{i,j}, j \in [0, m_1)$, mỗi người dùng V_i nắm giữ các giá trị bí mật $r_{i,j}, j \in [m_1, m)$, không ai biết về các giá trị này trừ người dùng đó. Một máy chủ được gọi là máy chủ tư vấn sẽ hợp tác với người dùng để tính toán mức trung bình cũng như điểm tương đồng giữa các mục tin và lưu trữ chúng trong cơ sở dữ

Bảng 3.6: Bảng ma trận xếp hạng của người dùng-mục tin

Người dùng/ mục tin	i_0	$\dots$	i_{m_1-1}	**Người dùng/ mục tin**	i_{m_1}	$\dots$	i_{m-1}
U_1	$r_{1,0}$	$\dots$	r_{1,m_1-1}	V_1	r_{1,m_1}	$\dots$	$r_{1,m-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
U_i	$r_{i,0}$	$\dots$	r_{i,m_1-1}	V_i	r_{i,m_1}	$\dots$	$r_{i,m-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
U_{n_1}	$r_{n_1,0}$	$\dots$	r_{n_1,m_1-1}	V_{n_1}	r_{n_1,m_1}	$\dots$	$r_{n_1,m-1}$

liệu của mình và hỗ trợ người dùng mục tiêu trong việc sinh tư vấn. Máy chủ tư vấn sẽ không tiết lộ các tham số của hệ tư vấn cho bất kỳ ai (các giá trị trung bình xếp hạng và độ tương tự giữa các mục tin khác nhau).

3.4.2. Giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng

Giải pháp đề xuất cũng được chia thành hai giai đoạn giống như giải pháp trong Mục 3.3 và chỉ khác giải pháp ban đầu ở giai đoạn đầu tiên (giai đoạn xây dựng hệ tư vấn người dùng) do mô hình dữ liệu của hai giải pháp khác nhau.

- Giai đoạn 1: Sử dụng giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự có bảo mật được đề xuất trong Mục 2.2 và giao thức tính đồng thời nhiều giá trị tần suất có bảo mật trong mô hình 2PFD được đề xuất trong Mục 2.3 để tính trung bình xếp hạng của các mục tin và độ tương tự giữa các cặp mục tin. Trong đó giao thức trong Mục 2.2 được sử dụng để tính các giá trị trung bình xếp hạng của các mục tin và các giá trị độ tương tự giữa các cặp mục tin thuộc cùng một miền dữ liệu. Giao thức trong Mục 2.3 được sử dụng kết hợp với các giá trị tính được ở giao thức trong Mục 2.2 để tính các giá trị độ tương tự giữa các cặp mục tin thuộc hai miền dữ liệu khác nhau.

Trước khi bắt đầu giải pháp, mỗi U_i chọn một khoá bí mật $x_i \in_R \mathbb{Z}_d^*$, tính khoá công khai $X_i = x_i G$ và gửi X_i cho máy chủ tư vấn. Mỗi V_i chọn một khoá bí mật $y_i \in_R \mathbb{Z}_d^*$, tính khoá công khai $Y_i = y_i G$ và gửi Y_i cho máy chủ tư vấn.

Trong giai đoạn này, những người dùng và máy chủ tư vấn cùng tham gia vào quá trình tính các giá trị trung bình xếp hạng của các mục tin và các giá trị độ tương tự giữa cặp mục tin. Để xây dựng các tham số cho hệ tư vấn người dùng cần phải tính:

$$R_{i,j} = \frac{\sum_{i=1}^{n_1} r_{i,j}}{\sum_{i=1}^{n_1} f_{i,j}}, \quad j \in [0,m) \tag{3.4.6}$$

$$S(i_j, i_k) = \frac{\sum_{i=1}^{n_1} r_{i,j} r_{i,k}}{\sqrt{\sum_{i=1}^{n_1} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{n_1} r_{i,k}^2}}, \quad j,k \in [0,m), j < k \tag{3.4.7}$$

Trong đó:

- Các giá trị nhóm 1:

$$n_{k1} = \left\lceil \frac{1}{2} + \sqrt{2n_{s1} + \frac{1}{4}} \right\rceil \tag{3.4.8}$$

$$n_{s1} = \frac{m_1(m_1+5)}{2} \tag{3.4.9}$$

- Các giá trị nhóm 2:

$$n_{k2} = \left\lceil \frac{1}{2} + \sqrt{2n_{s2} + \frac{1}{4}} \right\rceil \tag{3.4.10}$$

$$n_{s2} = \frac{m_2(m_2+5)}{2} \tag{3.4.11}$$

- Các giá trị nhóm 3:

$$n_{s3} = m_1 \times m_2 \tag{3.4.12}$$

$$n_{k3} = \left\lceil \frac{1}{2} + \sqrt{2m_1m_2 + \frac{1}{4}} \right\rceil \tag{3.4.13}$$

Có thể dễ dàng nhận thấy: $n_{k3} < n_{k1} \times n_{k2}$ nên những người dùng sẽ không cần chuẩn bị thêm khoá.

Giai đoạn 2: Để sinh tư vấn có đảm bảo tính riêng tư cho người dùng mục tiêu, quá trình sinh tư vấn giống với giải pháp hệ tư vấn có đảm bảo tính riêng tư được đề xuất cho mô hình dữ liệu phân tán đầy đủ trong Mục 3.3.2.

Trong giải pháp này, để tính các giá trị trung bình xếp hạng và độ tương tự giữa các cặp mục tin có đảm bảo tính riêng tư này, các giao thức tính toán bảo mật đề xuất được sử dụng cụ thể như sau:

- Các giá trị nhóm 1 được tính bằng cách sử dụng giao thức SMASC được trình bày trong Mục 2.2 đối với tập n_1 bên tham gia U_i. - Các giá trị nhóm 2 được tính bằng cách sử dụng giao thức SMASC được trình bày trong Mục 2.2 đối với tập n_1 bên tham gia V_i. - Các giá trị nhóm 3 được tính bằng cách sử dụng giao thức PPMFC được trình bày trong Mục 2.3 với tập n bên tham gia U_i, V_i.

Như phân tích ở trên, các bên tham gia sẽ không cần chuẩn bị thêm các cặp khoá. Tuy nhiên, các giá trị cần tính này lại có sự lặp lại, vì thế để giảm chi phí tính toán cũng như chi phí truyền thông ở pha 1 của giao thức thì cần sửa đổi giao thức cho phù hợp.

Giải pháp chi tiết sẽ được trình bày cụ thể trong Giải pháp **3.2**.

Giải pháp 3.2: Giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng trong mô hình 2PFD

Đầu vào:
n: Số lượng người dùng trong hệ thống, trong đó mỗi miền dữ liệu có n_1 người dùng ($n = 2n_1$).
m: Số lượng các mục tin trong cả hai miền dữ liệu.
m_1: Số lượng các mục tin thuộc miền dữ liệu đầu tiên (U_i).
RM_1: Ma trận xếp hạng của những người dùng đối với các mục tin thuộc miền dữ liệu đầu tiên, trong đó mỗi $r_{i,j}$ là giá trị xếp hạng riêng tư của người dùng U_i đối với mục tin thứ j, $j \in [0, m_1)$.
RM_2: Ma trận xếp hạng của những người dùng đối với các mục tin thuộc miền dữ liệu thứ hai (V_i), trong đó mỗi $r_{i,j}$ là giá trị xếp hạng riêng tư của người dùng V_i đối với mục tin thứ j, $j \in [0, m - m_1)$.
max: Giá trị xếp hạng lớn nhất trong thang xếp hạng.
Đầu ra (đối với U_i):
$P_{i,k}$: Điểm đề xuất được dự đoán xếp hạng cho mục tin thứ k của người dùng U_i, $k \in [0, m)$ là thứ tự mục tin mà người dùng mục tiêu đã yêu cầu sinh tư vấn.
Nếu sinh tư vấn dựa trên nội dung - CBF thì: $P_{i,k} = \frac{\sum_{j=1}^{m} r_{i,j} S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$
Nếu sinh tư vấn dựa trên lọc cộng tác – CF thì: $P_{i,k} = \frac{R_k \sum_{j=1}^{m} S(i_k, i_j) + \sum_{j=1}^{m} (r_{i,j} - R_j) S(i_k, i_j)}{\sum_{j=1}^{m} S(i_k, i_j)}$

Giai đoạn 1: Tính trung bình xếp hạng của các mục tin và độ tương tự giữa cặp mục tin

1. $\{R_j, \sum_{i=1}^{\frac{n}{2}} r_{i,j}^2, S(i_j, i_k)\}_{0 \le j,k < m_1; j \le k} = SMASC(\frac{n}{2}, m_1, RM_1, \max)$;
/* Thực hiện giao thức SMASC với tập $\frac{n}{2}$ bên tham gia U_i.*/
2. $\{R_j, \sum_{i=1}^{\frac{n}{2}} r_{i,j}^2, S(i_j, i_k)\}_{m_1 \le j,k < m; j \le k} = SMASC(\frac{n}{2}, m - m_1, RM_2, \max)$;
/* Thực hiện giao thức SMASC với tập $\frac{n}{2}$ bên tham gia V_i. Trong đó: Mỗi người dùng U_i có n_{k1} khóa bí mật và khóa công khai $(ksu_{i,j}, KPU_{i,j})$ và đã gửi khóa công khai cho máy chủ khi thực hiện giao thức SMASC lần đầu. Mỗi người dùng V_i có n_{k2} khóa bí mật và khóa công khai $(ksv_{i,j}, KPV_{i,j})$ và đã gửi khóa công khai cho máy chủ. */
/*Thực hiện tính các giá trị tần suất giữa các cặp mục tin thuộc hai miền dữ liệu*/
Pha khởi tạo: Rs thực hiện
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 lần lượt là các khóa công khai dùng chung của tập người dùng U_i, V_i đã được máy chủ tính trong quá trình thực thi các giao thức SMASC bên trên */
10. Gửi cho $\{U_i, V_i\}_{i=1}^{\frac{n}{2}}$: $\{K_j\}_{j=0}^{n_{k3}-1}$;
/*Thực hiện giao thức PPMFC có sửa đổi với tập n bên tham gia U_i, V_i*/
/*Trong đó không cần thực thi pha khởi tạo*/
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++;

Giai đoạn 2: Sinh tư vấn cho U_i
// Giống với giải pháp đề xuất trong mục 3.3.2.

3.4.3. Đánh giá về tính đúng đắn

Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình 2PFD được đề xuất gồm hai giai đoạn: Giai đoạn 1: Xây dựng hệ tư vấn người dùng thông qua việc

tính các giá trị xếp hạng trung bình của mục tin và độ tương tự giữa các cặp mục tin, Giai đoạn 2: Tạo các tư vấn cho người dùng.

Trong giai đoạn 1, các giá trị xếp hạng trung bình của mục tin và các giá trị độ tương tự giữa cặp mục tin được tính thông qua việc thực thi 2 lần giao thức SMASC và 1 lần giao thức PPMFC. Do đó, kết quả đầu ra của các giao thức này là hoàn toàn chính xác. Cụ thể:

Thực thi giao thức SMASC thứ nhất SMASC$(\frac{n}{2}, m_1, RM_1, \max)$ thu được các giá trị: R_j, $S(i_j, i_k)$, $\sum_{i=1}^{\frac{n}{2}} r_{i,j}^2$, với $0 \leq j,k < m_1, j \leq k$.

Thực thi giao thức SMASC thứ hai SMASC$(\frac{n}{2}, m - m_1, RM_2, \max)$ thu được các giá trị: R_j, $S(i_j, i_k)$, $\sum_{i=1}^{\frac{n}{2}} r_{i,j}^2$, với $m_1 \leq j,k < m, j \leq k$.

Thực thi giao thức PPMFC PPMFC$(n, m, m_1, RM_1, RM_2, \max)$ thu được các giá trị: $\sum_{i=1}^{\frac{n}{2}} r_{i,j} r_{i,k}$, với $j < m_1$ và $k \geq m_1$.

Do đó, các giá trị độ tương tự giữa cặp mục tin thuộc các miền dữ liệu khác nhau ($j < m_1$ và $k \geq m_1$) được tính như sau:

$$S(i_j, i_k) = \frac{\sum_{i=1}^{\frac{n}{2}} r_{i,j} r_{i,k}}{\sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,k}^2}} \tag{3.4.14}$$

Do đó, luận án có thể kết luận: giải pháp đề xuất tính đúng đắn các giá trị trung bình xếp hạng của các mục tin và các giá trị độ trương tự giữa cặp mục tin

$$R_j = \frac{\sum_{i=1}^{\frac{n}{2}} r_{i,j}}{\sum_{i=1}^{\frac{n}{2}} f_{i,j}} \tag{3.4.15}$$

$$S(i_j, i_k) = \frac{\sum_{i=1}^{\frac{n}{2}} r_{i,j} r_{i,k}}{\sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{\frac{n}{2}} r_{i,k}^2}}, \quad j,k \in [0,m), j \leq k. \tag{3.4.16}$$

Ở giai đoạn 2, giải pháp cung cấp hai cách thức để tạo tư vấn cho người dùng: tạo tư vấn dựa trên CBF và tạo tư vấn dựa trên CF có đảm bảo tính riêng tư. Quá trình này giống với giải pháp cho mô hình dữ liệu phân tán đầy đủ như đã được chứng minh trong mục 3.3.3 là đảm bảo tính đúng đắn của kết quả tư vấn. Từ đó luận án có thể kết luận:

Kết luận 3.3. *Giải pháp hệ tư vấn có đảm bảo tính riêng tư được đề xuất cho mô hình 2PFD đảm bảo tính đúng đắn của hệ tư vấn người dùng cũng như kết quả tư vấn.*

Về mặt lý thuyết luận án đã chứng minh giải pháp đề xuất bảo toàn tính đúng đắn

của kết quả tư vấn. Tuy nhiên do hệ mật mã trên đường cong Elliptic chỉ làm việc với bản rõ là số nguyên nên trong quá trình sinh tư vấn đôi khi phải thực hiện làm tròn các giá trị trung bình và độ tương tự, vì thế sẽ có sự sai lệch nhỏ trong kết quả điểm tư vấn $P_{i,k}$.

3.4.4. Đánh giá về tính riêng tư

Giải pháp đề xuất gồm hai giai đoạn:

Giai đoạn 1: Xây dựng hệ tư vấn người dùng có đảm bảo tính riêng tư bằng cách kết hợp giao thức SMASC và PPMFC đã được chứng minh trong Chương 2.

- Trong mục 2.2.4 của luận án đã chứng minh được: giao thức SMASC *"đảm bảo tính riêng tư của mỗi người dùng trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng giữa tối đa n-2 bên thông đồng, giao thức vẫn bảo vệ tính riêng tư của các bên trung thực. Đồng thời giao thức cũng bảo vệ tính riêng tư của các bên trung thực trước hành vi độc hại của trung tâm tính toán nhằm tìm hiểu các giá trị riêng tư của người dùng theo kịch bản cụ thể đã được chỉ ra trong tài liệu [55]"*.

- Trong mục 2.3.4 của luận án cũng đã chứng minh rằng: giao thức PPMFC *"bảo vệ tính riêng tư của mỗi người dùng trung thực trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng của các bên tham gia, giao thức bảo vệ tính riêng tư của người dùng trung thực chống lại trung tâm tính toán và tối đa n-2 người dùng thông đồng. Trong trường hợp chỉ có hai người dùng trung thực, điều này vẫn chính xác miễn là hai người dùng trung thực không sở hữu các giá trị thuộc tính của cùng một bản ghi"*.

Kết hợp với Định lí 1.1, luận án có kết luận: Giai đoạn 1 của giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình 2PFD bảo vệ tính riêng tư của mỗi người dùng trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Khi có sự thông đồng giữa các bên, giải pháp bảo vệ tính riêng tư của những người dùng trung thực trước sự thông đồng tối đa $n-2$ người dùng thông đồng của mỗi miền dữ liệu.

Giai đoạn 2, quá trình sinh tư vấn giống với hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình phân tán đầy đủ được đề xuất trong mục 3.3. Quá trình này đã được chứng minh là bảo vệ tính riêng tư của mỗi người dùng trong mô hình bán trung thực, cũng như bảo vệ tính riêng tư trước bất kỳ sự thông đồng nào.

Kết hợp với Định lí 1.1, ta có kết luận:

Kết luận 3.4. *Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư được đề xuất*

cho mô hình 2PFD bảo vệ tính riêng tư của mỗi người dùng trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Khi có sự thông đồng, giải pháp bảo vệ tính riêng tư của người dùng trung thực trước sự thông đồng tối đa $n-2$ người dùng của mỗi miền dữ liệu.

3.4.5. Đánh giá tính hiệu quả

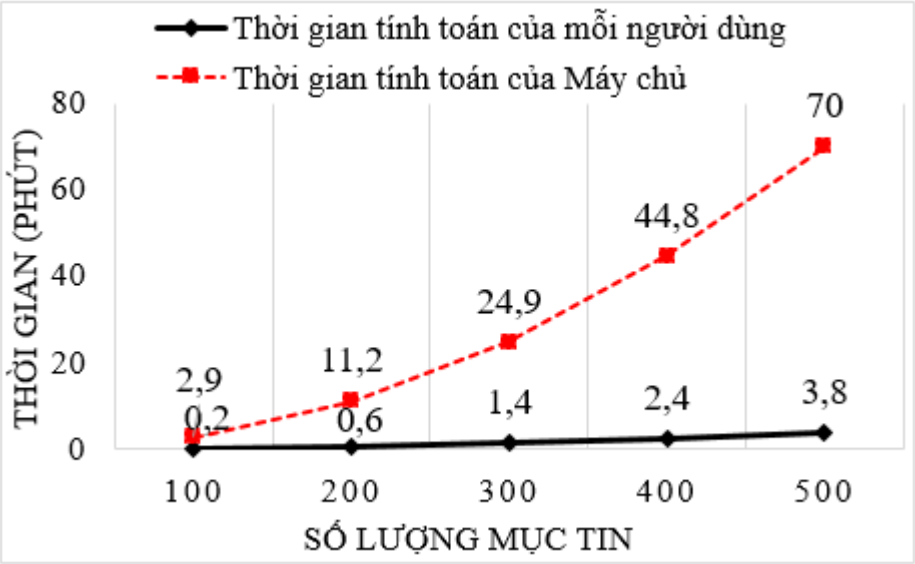
Trong các thực nghiệm, 500 mục tin và 943 người dùng đã xếp hạng cho các mục tin được chọn với giả định một nửa số mục tin thuộc sở hữu của tổ chức thứ nhất và số mục tin còn lại thuộc sở hữu của tổ chức thứ hai. Giải pháp đề xuất sử dụng hệ mật mã trên đường cong Elliptic sẽ sử dụng đường cong Elliptic BrainpoolP160r1 với kích thước khoá bí mật là 160 bit, khoá công khai là 320 bit [45]

Tất cả người dùng được giả định là cùng tham gia để tính các giá trung bình xếp hạng và độ tương tự giữa các mục tin trong hệ thống, và chỉ một người dùng (người dùng mục tiêu) tham gia tạo tư vấn. Đồng thời, người dùng cũng được giả định rằng họ thực thi các tính toán cần thiết và gửi các thông điệp song song tới máy chủ, do đó, chi phí tính toán ở phía người dùng có thể được giảm xuống bằng cách tính toán chỉ cho một người. Về phía máy chủ, chi phí tính toán và giao tiếp được tính cho tất cả người dùng tham gia vào hệ thống vì chúng phụ thuộc vào sự cộng tác của tất cả người dùng với máy chủ.

Giải pháp này chỉ khác giải pháp trong mô hình dữ liệu phân tán đầy đủ về quá trình thực hiện tính các giá trị trung bình xếp hạng và độ tương tự giữa các mục tin, nên trong phần đánh giá này, Luận án chỉ thực hiện đánh giá quá trình này. Tất cả người dùng được giả định thực hiện tác vụ của họ cùng một lúc và độ trễ mạng không được tính vào tổng thời gian thực thi.

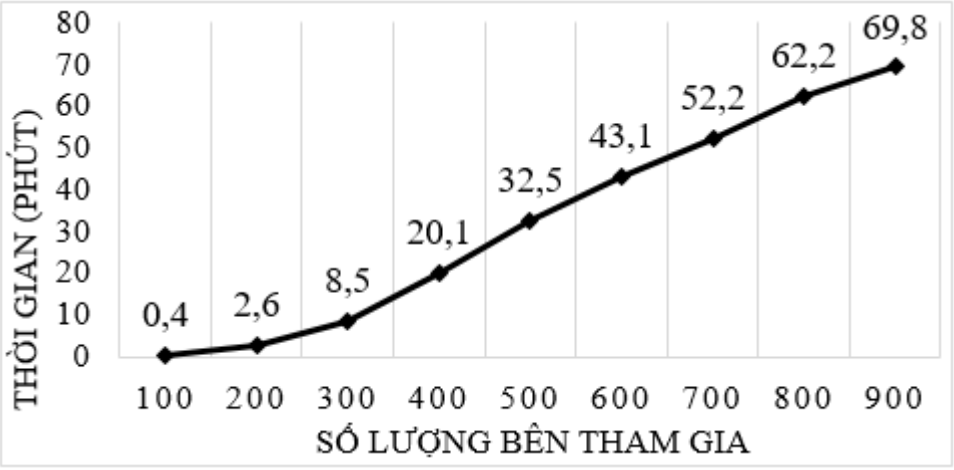
Do đó, việc phân tích hiệu suất của giải pháp gồm 943 người dùng tham gia đánh giá mỗi bên và 500 mục tin. Các giải pháp được thực thi độc lập trong cùng một môi trường với số lượng mục tin thay đổi từ 100 đến 500 và lấy kết quả thời gian tính toán trung bình. Số lượng các mục tin của mỗi bên được giả định là bằng nhau ($m_1 = m_2 = \frac{m}{2}$). Chi tiết về thời gian tính toán của mỗi người dùng và của máy chủ trong quá trình xây dựng hệ tư vấn người dùng được thể hiện trong Hình 3.7.

Hình 3.7 biểu diễn thời gian tính toán của mỗi người dùng và máy chủ trong quá trình xây dựng hệ tư vấn người dùng. Kết quả cho thấy với 943 người dùng tham gia xếp hạng, khi số lượng mục tin là 100, mỗi người dùng mất 0,2 phút và máy chủ mất 2,9 phút để tính toán. Khi số lượng mục tin là 500, mỗi người dùng mất 3,8 phút và máy chủ mất 70 phút.



Hình 3.7: Thời gian tính toán của người dùng và máy chủ trong quá trình xây dựng hệ tư vấn người dùng cho mô hình dữ liệu 2PFD

Để đánh giá về sự thay đổi thời gian tính toán của máy chủ trong quá trình xây dựng hệ tư vấn người dùng theo sự thay đổi của số lượng các bên tham gia, Luận án tiến hành thực nghiệm giải pháp độc lập với số lượng các bên tham gia thay đổi từ 100 đến 900 và lấy kết quả thời gian tính toán trung bình. Số lượng các mục tin của mỗi bên được giả định là bằng nhau ($m_1 = m_2 = \frac{m}{2} = 250$). Chi tiết về thời gian tính toán của máy chủ khi số lượng các bên tham gia thay đổi được thể hiện trong Hình 3.8.



Hình 3.8: Thời gian tính toán của máy chủ trong quá trình xây dựng hệ tư vấn cho mô hình dữ liệu 2PFD khi số lượng các bên tham gia thay đổi

Hình 3.8 biểu diễn thời gian tính toán của máy chủ trong quá trình xây dựng hệ tư vấn người dùng. Kết quả cho thấy với cùng số lượng mục tin là 500, khi có 100 người

dùng tham gia, máy chủ tư vấn cần 0,4 phút để tính toán. Khi có 900 người dùng tham gia, máy chủ cần 69,8 phút để tính toán, cho thấy tính hiệu quả của giải pháp.

Kết quả thực nghiệm cho thấy tính hiệu quả của giải pháp đề xuất. Thời gian thực thi của mỗi người dùng và máy chủ sẽ tăng theo hàm số luỹ thừa của số lượng mục tin. Với cùng một số lượng mục tin, khi số lượng người dùng tăng lên thì thời gian thực thi của các máy chủ sẽ tăng lên còn thời gian thực thi của mỗi người dùng là không đổi. Tuy nhiên trong thực tế máy chủ thường có cầu hình cao, hơn nữa các tính toán mà máy chủ cần thực hiện chủ yếu là phép cộng điểm trên đường cong Elliptic nên hiệu suất của giải pháp vẫn đảm bảo khi hệ tư vấn có hàng vạn người dùng. Trong trường hợp số lượng người dùng lớn hơn nữa, ta có thể đảm bảo hiệu suất của giải pháp bằng cách tăng cấu hình của máy chủ.

3.4.6. Ví dụ minh họa

Để mô tả rõ ràng giải pháp đề xuất, ma trận xếp hạng kích thước 3×4 như Bảng 3.7 được sử dụng. Trong đó, người dùng và mục tin được biểu diễn dưới dạng $U_i = \{U_1, U_2, U_3\}$, $V_i = \{V_1, V_2, V_3\}$ và $i_j = \{i_1, i_2, i_3, i_4\}$ tương ứng. Trước khi bắt đầu, các bên tham gia thống nhất các tham số đường cong Elliptic $(\mathbb{E}, G, d, p_e)$.

Vì giai đoạn sinh tư vấn cho người dùng giống với giải pháp được đề xuất trong mục 3.3.2 và đã được minh họa quá trình tính toán bằng số cụ thể trong mục 3.3.3 nên trong phần này luận án chỉ minh họa cụ thể giai đoạn 1: Tính trung bình xếp hạng của các mục tin và độ tương tự giữa cặp mục tin.

U_i/i_j	i_1	i_2	V_i/i_j	i_3	i_4
U_1	3	5	V_1	0	4
U_2	0	1	V_2	5	1
U_3	2	3	V_3	2	2

Bảng 3.7: Ví dụ về ma trận xếp hạng trong mô hình 2PFD

Trước khi bắt đầu giải pháp, mỗi U_i trong đó $i = \{1,2,3\}$, chọn các khóa bí mật $x_i, ksu_{i,j} \in_R \mathbb{Z}_d^*, 1 \leq j \leq n_{k1} = \lceil \frac{1}{2} + \sqrt{2(2+5) + \frac{1}{4}} \rceil = 5$, tính khóa công khai $X_i = x_i G$, $KPU_{i,j} = ksu_{i,j} G$ và gửi cho máy chủ.

Người dùng V_i trong đó $i = \{1,2,3\}$, chọn ngẫu nhiên các khóa bí mật $y_i, ksv_{i,j} \in_R \mathbb{Z}d^*, 1 \leq j \leq n_{k2} = \lceil \frac{1}{2} + \sqrt{2(2+5) + \frac{1}{4}} \rceil = 5$, tính khóa công khai $X_i = x_i G$, $KPU_{i,j} = ksu_{i,j} G$ và gửi cho máy chủ.

Đầu tiên, những người dùng $U_i = \{U_1, U_2, U_3\}$ và máy chủ thực thi giao thức SMASC với tập mục tin là $i_j = \{i_1, i_2\}$ theo các bước được minh họa như trong mục 2.2.8 và thu được các giá trị sau:

$$R_1 = 2,5, \quad R_2 = 3, \quad \sum_{i=1}^{3} r_{i,1}^2 \quad = 13, \quad \sum_{i=1}^{3} r_{i,2}^2 = 37, \quad S(i_1, i_2) = S(i_2, i_1) = 0,99.$$

Tiếp theo, những người dùng $V_i = \{V_1, V_2, V_3\}$ và máy chủ thực thi giao thức SMASC với tập mục tin là $i_j = \{i_3, i_4\}$ theo các bước được minh họa như trong mục 2.2.8 và thu được kết quả sau:

$$R_3 = 2,5, \quad R_4 = 2,33, \quad \sum_{i=1}^{3} r_{i,3}^2 \quad = 29, \quad \sum_{i=1}^{3} r_{i,4}^2 = 21, \quad S(i_3, i_4) = S(i_4, i_3) = 0,36.$$

Tiếp theo, thực hiện tính các giá trị tổng các tích xếp hạng giữa các mục tin thuộc hai miền dữ liệu khác nhau như sau:

Máy chủ tính $n_{k3} = \lceil \frac{1}{2} + \sqrt{2.2.2 + \frac{1}{4}} \rceil = 4$ khóa công khai dùng chung:

$$\begin{aligned} K_1 &= KU_1 + KV_1, \\ K_2 &= KU_1 + KV_2, \\ K_3 &= KU_1 + KV_3, \\ K_4 &= KU_1 + KV_4. \end{aligned}$$

Và gửi $\{K_j\}_{j \in [1,4]}$ cho tất cả các bên tham gia.

Những người dùng U_i mã hoá các giá trị xếp hạng của mình và gửi cho máy chủ:

$$\begin{aligned} (C_1^{(1,j)}, C_2^{(1,j)}) &= \{(3G + c_{1,1}X_1, c_{1,1}G), (5G + c_{1,2}X_1, c_{1,2}G)\}; \\ (C_1^{(2,j)}, C_2^{(2,j)}) &= \{(0G + c_{2,1}X_2, c_{2,1}G), (1G + c_{2,2}X_2, c_{2,2}G)\}; \\ (C_1^{(3,j)}, C_2^{(3,j)}) &= \{(2G + c_{3,1}X_3, c_{3,1}G), (3G + c_{3,2}X_3, c_{3,2}G)\}. \end{aligned}$$

Người dùng V_i lấy $C_1^{(i,j)}, C_2^{(i,j)}$ từ máy chủ và tính $n_{s3} = 2 \times 2 = 4$ cặp giá trị sau:

$(Q_1^{(1,j)}, Q_2^{(1,j)}) = \{(0C_1^{(1,1)} + ksv_{1,2}K_1 - rv_{1,1}C_2^{(1,1)} - ksv_{1,1}K_2, rv_{1,1}Y_1 - 0X_1),$
$(4C_1^{(1,1)} + ksv_{1,3}K_1 - rv_{1,2}C_2^{(1,1)} - ksv_{1,1}K_3, rv_{1,2}Y_1 - 4X_1),$
$(0C_1^{(1,2)} + ksv_{1,4}K_1 - rv_{1,3}C_2^{(1,2)} - ksv_{1,1}K_4, rv_{1,3}Y_1 - 0X_1),$
$(4C_1^{(1,2)} + ksv_{1,3}K_2 - rv_{1,4}C_2^{(1,2)} - ksv_{1,2}K_3, rv_{1,4}Y_1 - 4X_1)\};$

$(Q_1^{(2,j)}, Q_2^{(2,j)}) = \{(0G + 5c_{2,1}X_2 + ksv_{2,2}K_1 - rv_{2,1}c_{2,1}G - ksv_{2,1}K_2, rv_{2,1}Y_2 - 5X_2),$

$(0G+1c_{2,1}X_2+ksv_{2,3}K_1-rv_{2,2}c_{2,1}G-ksv_{2,1}K_3, rv_{2,2}Y_2-1X_2)$,
$(5G+5c_{2,2}X_2+ksv_{2,4}K_1-rv_{2,3}c_{2,2}G-ksv_{2,1}K_4, rv_{2,3}Y_2-5X_2)$,
$(1G+1c_{2,2}X_2+ksv_{2,3}K_2-rv_{2,4}c_{2,2}G-ksv_{2,2}K_3, rv_{2,4}Y_2-1X_2)\}$;

$(Q_1^{(3,j)}, Q_2^{(3,j)}) = \{(4G+2c_{3,1}X_3+ksv_{3,2}K_1-rv_{3,1}c_{3,1}G-ksv_{3,1}K_2, rv_{3,1}Y_3-2X_3)$,
$(4G+2c_{3,1}X_3+ksv_{3,3}K_1-rv_{3,2}c_{3,1}G-ksv_{3,1}K_3, rv_{3,2}Y_3-2X_3)$,
$(6G+2c_{3,2}X_3+ksv_{3,4}K_1-rv_{3,3}c_{3,2}G-ksv_{3,1}K_4, rv_{3,3}Y_3-2X_3)$,
$(6G+2c_{3,2}X_3+ksv_{3,3}K_2-rv_{3,4}c_{3,2}G-ksv_{3,2}K_3, rv_{3,4}Y_3-1X_3)\}$.

Và gửi các giá trị này cho máy chủ.

Người dùng U_i lấy $Q_1^{(i,j)}$, $Q_2^{(i,j)}$ từ máy chủ và tính như sau:

$A_{1,j} = \Big\{(ksv_{1,2}+ksu_{1,1})K_1-(ksu_{1,1}+ksv_{1,1})K_2$,
$12G+(ksu_{1,1}+ksv_{1,3})K_1-(ksu_{1,1}+ksv_{1,1})K_3$,
$(ksu_{1,1}+ksv_{1,4})K_1-(ksu_{1,1}+ksv_{1,1})K_4$,
$20G+(ksu_{1,1}+ksv_{1,3})K_2-(ksu_{1,1}+ksv_{1,2})K_3\Big\}$;

$A_{2,j} = \Big\{(ksv_{2,2}+ksu_{2,1})K_1-(ksu_{2,1}+ksv_{2,1})K_2$,
$(ksu_{2,1}+ksv_{2,3})K_1-(ksu_{2,1}+ksv_{2,1})K_3$,
$5G+(ksu_{2,1}+ksv_{2,4})K_1-(ksu_{2,1}+ksv_{2,1})K_4$,
$G+(ksu_{2,1}+ksv_{2,3})K_2-(ksu_{2,1}+ksv_{2,2})K_3\Big\}$;

$A_{3,j} = \Big\{4G+(ksv_{3,2}+ksu_{3,1})K_1-(ksu_{3,1}+ksv_{3,1})K_2$,
$4G+(ksu_{3,1}+ksv_{3,3})K_1-(ksu_{3,1}+ksv_{3,1})K_3$,
$6G+(ksu_{3,1}+ksv_{3,4})K_1-(ksu_{3,1}+ksv_{3,1})K_4$,
$6G+(ksu_{3,1}+ksv_{3,3})K_2-(ksu_{3,1}+ksv_{3,2})K_3\Big\}$;

Và gửi các giá trị này cho máy chủ.

Máy chủ tính giải mã bản mã tổng hợp tương ứng với các giá trị cần tính như sau:
$A_1 = 4G+(ksu_{1,1}+ksv_{1,3}+ksu_{2,1}+ksv_{2,3}+ksu_{3,1}+ksv_{3,3})K_1$
$-(ksu_{1,1}+ksv_{1,1}+ksu_{2,1}+ksv_{2,1}+ksu_{3,1}+ksv_{3,1})K_2 = 4G$.

$A_2 = 12G+4G+(ksv_{1,2}+ksu_{1,1}+ksv_{2,2}+ksu_{2,1}+ksv_{3,2}+ksu_{3,1})K_1$
$-(ksu_{1,1}+ksv_{1,1}+ksu_{2,1}+ksv_{2,1}+ksu_{3,1}+ksv_{3,1})K_2 = 16G$.

$A_3 = 5G+6G+(ksu_{1,1}+ksv_{1,4}+ksu_{2,1}+ksv_{2,4}+ksu_{3,1}+ksv_{3,4})K_1$
$-(ksu_{1,1}+ksv_{1,1}+ksu_{2,1}+ksv_{2,1}+ksu_{3,1}+ksv_{3,1})K_2 = 16G$.

$A_4 = 20G+G+6G+(ksu_{1,1}+ksv_{1,3}+ksu_{2,1}+ksv_{2,3}+ksu_{3,1}+ksv_{3,3})K_2$
$-(ksu_{1,1}+ksv_{1,2}+ksu_{2,1}+ksv_{2,2}+ksu_{3,1}+ksv_{3,2})K_3 = 27G$.

Máy chủ tính logarit rời rạc để thu được các giá trị:

$suv = \{\sum_{i=1}^{3} r_{i,1}r_{i,3} = 4, \sum_{i=1}^{3} r_{i,1}r_{i,4} = 16, \sum_{i=1}^{3} r_{i,2}r_{i,3} = 16, \sum_{i=1}^{3} r_{i,2}r_{i,4} = 27\}$.

Cuối cùng, máy chủ tính độ tương tự:

$S(i_1,i_3) = S(i_3,i_1) = \frac{suv_1}{\sqrt{\sum_{i=1}^{3} r_{i,1}^2} \cdot \sqrt{\sum_{i=1}^{3} r_{i,3}^2}} = \frac{4}{\sqrt{13} \cdot \sqrt{29}} = 0,21.$

Tương tự: $S(i_1,i_4) = S(i_4,i_1) = 0,97, \quad S(i_2,i_3) = S(i_3,i_2) = 0,49, \quad S(i_2,i_4) = S(i_4,i_2) = 0,97.$

Vì hệ mật mã đường cong Elliptic không thể xử lý số thập phân, máy chủ nhân các giá trị này với 100: $R_1 = 250, R_2 = 300, R_3 = 250, R_4 = 233, S(i_2,i_1) = 99, S(i_3,i_1) = 21, S(i_4,i_1) = 97, S(i_3,i_2) = 49, S(i_4,i_2) = 97, S(i_4,i_3) = 36.$

3.5. Kết luận chương

Chương này đã trình bày về hai giải pháp hệ tư vấn người dùng có đãm bảo tính riêng tư được đề xuất: Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ và Giải pháp hệ tư vấn người dùng có đảm bảo tính rêng tư cho mô hình dữ liệu 2PFD. Trong các giải pháp này có sử dụng các giao thức đề xuất trong Chương 2 cho giai đoạn thứ nhất của hệ tư vấn người dùng, ở giai đoạn thứ hai và đề xuất hai phương pháp sinh tư vấn có đảm bảo tính riêng tư: sinh tư vấn dựa trên CBF và sinh tư vấn dựa trên CF với việc cải tiến giai đoạn này trong giải pháp gốc sang sử dụng hệ mật mã trên đường cong Elliptic để đồng bộ với giai đoạn thứ nhất của giải pháp đề xuất đồng thời cũng góp phần nâng cao được tính hiệu quả của giải pháp.

• Các giải pháp này không những đảm bảo tính đúng đắn mà còn đảm bảo tính riêng tư cho hồ sơ và lịch sử xếp hạng của người dùng khỏi bất kỳ bên thứ ba nào hoặc những người dùng khác.

• Kết quả thực nghiệm chứng minh rằng giải pháp đề xuất cho mô hình phân tán đầy đủ không những đáp ứng yêu cầu ứng dụng thực tế mà còn vượt trội so với các giải pháp trước đây [7, 75], giải pháp mới đề xuất cho mô hình 2PFD có hiệu suất tốt, hoàn toàn có thể áp dụng vào trong các hệ thống thực tế.

• Các giải pháp này có thể được triển khai thực tế dưới dạng mô hình máy chủ - máy trạm, đây là một mô hình phổ biến hiện nay.

Nội dung của chương này liên quan đến các công trình [CT3, CT4, CT6] trong Danh mục các công trình khoa học sử dụng trong Luận án.

KẾT LUẬN

Trong luận án này, nghiên cứu sinh đã nghiên cứu giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng. Nghiên cứu sinh nhận thấy rằng công việc thiết kế các giải pháp như vậy không đơn giản ngay cả khi sử dụng các kỹ thuật mã hóa tiên tiến như mã hóa đồng cấu và tính toán bảo mật với mô hình dữ liệu phân tán, đặc biệt là mô hình dữ liệu phân tán đầy đủ và mô hình dữ liệu phân tán đầy đủ hai bên.

Các giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư sử dụng mã hoá đồng cấu hay tính toán bảo mật có thể đảm bảo được tính đúng đắn của kết quả tính toán, vì thế đảm bảo chất lượng của hệ tư vấn tương đương với hệ tư vấn nguyên thuỷ. Bên cạnh đó các giải pháp này cũng đảm bảo tính riêng tư cho dữ liệu của người dùng ở mức độ cao. Tuy nhiên, hiệu suất của các giải pháp này thường thấp do các hoạt động mã hoá.

Do đó, trong luận án này, nghiên cứu sinh đã đề xuất giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư mới giúp đảm bảo được tính riêng tư dữ liệu của người dùng ở mức độ cao, đồng thời duy trì được tính đúng đắn của các tư vấn tương đương với các giải pháp điển hình hiện có và hiệu suất tốt hơn.

Sau một thời gian nghiên cứu và dưới sự hướng dẫn tận tình của tập thể hướng dẫn khoa học, mặc dù còn nhiều hạn chế về chuyên môn, thời gian, song nghiên cứu sinh đã nỗ lực hết mình để hoàn thành các mục tiêu của đề tài *"Phát triển giao thức tính toán bảo mật và ứng dụng cho hệ tư vấn người dùng có đảm bảo tính riêng tư"*.

I. Kết quả đạt được

Một số kết quả chính mà Luận án đã đạt được bao gồm:

1. Phát triển 02 giao thức tính toán bảo mật

• Giao thức tính đồng thời nhiều giá trị trung bình và độ tương tự bảo mật trong mô hình dữ liệu phân tán đầy đủ bảo toàn tính đúng đắn của kết quả đầu ra tương đương với phép toán nguyên thuỷ khi chưa áp dụng biện pháp bảo mật, có hiệu quả tính toán và truyền thông tốt hơn các giao thức đã được công bố trước đây trong cùng mô hình tính toán. Đồng thời, việc chứng minh về mặt lý thuyết cũng đã chỉ ra mức độ an toàn tương đương với giao thức tính trung bình và độ tương tự điển hình có mức độ an toàn cao.

Giao thức đề xuất đảm bảo tính riêng tư của các bên tham gia trung thực trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng của số lượng lớn người dùng bán trung thực $(n-2)$, thì tính riêng tư của các bên tham gia trung thực vẫn được đảm bảo. Giao thức này còn bảo vệ tính riêng tư của

các bên tham gia trung thực trước máy chủ độc hại muốn tìm hiểu thông tin về giá trị bí mật của các bên tham gia theo kịch bản được mô tả chi tiết trong công trình [55].

Hơn nữa, theo các giả định thiết lập của giao thức, giao thức đề xuất có thể chống lại một số tấn công điển hình đối với các hệ thống sử dụng lược đồ mã hoá khoá công khai (CPA, CCA1, CCA2).

• Giao thức tính đồng thời nhiều giá trị tần suất bảo mật trong mô hình 2PFD bảo toàn tính đúng đắn của kết quả đầu ra tương đương với phép toán nguyên thuỷ khi chưa áp dụng biện pháp bảo mật, có hiệu quả tính toán và truyền thông tốt hơn các giao thức đã được công bố trước đây trong cùng mô hình tính toán. Đồng thời, việc chứng minh về mặt lý thuyết cũng đã chỉ ra mức độ an toàn tương đương với các giao thức điển hình có mức an toàn cao.

Giao thức đề xuất đảm bảo tính riêng tư của các bên tham gia trung thực trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng của số lượng lớn người dùng bán trung thực $(n-2)$, thì tính riêng tư của các bên tham gia trung thực vẫn được đảm bảo.

Hơn nữa, theo các giả định thiết lập của giao thức, giao thức đề xuất có thể chống lại một số tấn công điển hình đối với các hệ thống sử dụng lược đồ mã hoá khoá công khai (CPA, CCA1, CCA2).

2. Thiết kế giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư

• Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình dữ liệu phân tán đầy đủ.

• Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư trong mô hình dữ liệu 2PFD.

Các giải pháp này có thể dễ dàng được triển khai dưới dạng mô hình máy chủ - máy trạm, một mô hình phổ biến hiện nay.

Đối với mỗi đề xuất:

+ Chứng minh tính đúng đắn của kết quả tư vấn được bảo toàn tương đương với giải pháp nguyên thuỷ khi chưa áp dụng biện pháp bảo mật.

+ Chứng minh được các giải pháp đề xuất đảm bảo tính riêng tư của các bên tham gia trung thực trong mô hình bán trung thực khi không có sự thông đồng giữa các bên. Trong trường hợp có sự thông đồng của số lượng lớn người dùng bán trung thực, thì tính riêng tư của các bên tham gia trung thực vẫn được đảm bảo, tương đương với độ an toàn của giải pháp hiện có có độ an toàn cao.

+ Chứng minh được về mặt lý thuyết rằng chi phí truyền thông và chi phí tính toán của các giải pháp đề xuất tốt hơn một số giải pháp điển hình hiện có.

Cuối cùng, chương trình mô phỏng hoạt động của các giao thức, giải pháp được triển khai cài đặt và xây dựng để thực nghiệm nhằm chỉ ra tính khả thi của các giao thức, giải pháp đã phát triển. Kết quả thực nghiệm cho thấy rằng các giao thức, giải pháp của luận án có hiệu suất tốt hơn cả về chi phí truyền thông và chi phí tính toán so với các giải pháp trong các công trình đã công bố trước đây. Vì vậy, những giải pháp này hoàn toàn có thể được triển khai trong thực tế.

II. Các hướng nghiên cứu tiếp theo

Mặc dù các giao thức được phát triển có thể sử dụng cho một số ứng dụng hệ tư vấn người dùng có đảm bảo tính riêng tư, tuy nhiên, vẫn còn một số vấn đề có thể được cải thiện trong tương lai:

1. Tiếp tục cải tiến để nâng cao mức độ an toàn của các giải pháp. Mặc dù có thể sử dụng một số giải pháp sẵn có để tích hợp vào các giải pháp của Luận án nhằm chống lại các kẻ tấn công độc hại, nhưng các giải pháp hiện tại có thể khá tốn kém cho các ứng dụng thực tế.

2. Nghiên cứu thêm về các hệ mật mã đồng cấu khác để có thể phát triển các giao thức SMC có khả năng kháng lượng tử, chỉ thực thi với dữ liệu có xác thực.

3. Trong giải pháp tính tần suất được đề xuất trong 2PFD, một nửa số người dùng cần hai lần tương tác với trung tâm tính toán, do đó, cần tiếp tục nghiên cứu, cải tiến để có thể giảm bớt số lần tương tác này.

4. Các giao thức SMC có thể được ứng dụng trong các lĩnh vực khác nhau, vì thế có thể nghiên cứu để mở rộng lĩnh vực ứng dụng thực tế của các giao thức đã được đề xuất trong Luận án. Điều này không chỉ giúp các tác vụ tính toán phân tán được thực hiện dễ dàng hơn mà còn bảo vệ sự an toàn của các bên tham gia.

5. Giải pháp hệ tư vấn người dùng có đảm bảo tính riêng tư được đề xuất chủ yếu dựa trên độ tương tự giữa các mục tin, độ chính xác của tư vấn còn thấp so với một số kỹ thuật tư vấn khác như hệ tư vấn dựa trên mô hình. Vì thế, hoàn toàn có thể nghiên cứu để đề xuất giải pháp đảm bảo tính riêng tư cho hệ tư vấn người dùng sử dụng các kỹ thuật tư vấn phức tạp hơn khác.

6. Các bài toán phân tán khác nhau sẽ có những yêu cầu khác nhau về hàm cần tính toán, mô hình dữ liệu phân tán, mức độ an toàn. Do đó, dựa trên các yêu cầu của các vấn đề thực tế hoặc các mô hình dữ liệu phân tán mới hoặc các mức độ an toàn khác nhau, các giao thức tính toán bảo mật mới cần được phát triển.

DANH MỤC CÔNG TRÌNH CÔNG BỐ

CT1. Van Vu-Thi, Dung Luong-The and Quan Hoang-Van, 2020, An Elliptic Curve-based Protocol for Privacy Preserving Frequency Computation in 2-Part Fully Distributed Setting, 12th International Conference on Knowledge and Systems Engineering (KSE), doi: 10.1109/KSE50997.2020.9287423 (**Scopus indexed**), Can Tho, Vietnam.

CT2. Van Vu-Thi, Dung Luong The, Quan Hoang Van, Luong Tran Thi and Tho Hoang Duc, Jun. 2022, Privacy-Preserving Decision Tree Solution in the 2-Part Fully Distributed Setting, Journal of Science and Technology on Information security, vol. 1, no. 15, pp. 92-101.

CT3. Van Vu-Thi, Dung Luong-The and Quan Hoang-Van, 2022, An efficient Privacy-Preserving Recommender System, 14th International Conference on Knowledge and Systems Engineering (KSE), doi: 10.1109/KSE56063.2022.9953800 (**Scopus indexed**), Nha Trang, Vietnam.

CT4. Van Vu-Thi and Dung Luong-The, 2023, An Elliptic Curve-based Privacy-Preserving Recommender System, The International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM2023), Lecture Notes in Computer Science (**Scopus Q2**), vol 14376, pp. 334-345, Japan, Springer, Cham, doi: 10.1007/978-3-031-46781-3_28.

CT5. Van Vu-Thi, Dung Luong-The and Duong Luong Ngoc, 2024, A secure Multi-Frequency Computation Protocol in 2-Part Fully Distributed Setting, Journal of Science and Technology on Information security, vol. 2, no. 22, pp. 40-48, doi: 10.54654/isj.v2i22.1051

CT6. Dung Luong-The and Van Vu-Thi, 16 February 2025, An Efficient Privacy-Preserving Recommender System Based on Elliptic Curve, Annals of Operations Research (**SCI/ISI indexed, Scopus Q1**), Springer Netherlands, doi: 10.1007/s10479-025-06515-w.

MỘT SỐ CÔNG TRÌNH ĐÃ CÔNG BỐ KHÁC

CT7. Van Vu-Thi, Dung Luong-The, Quan Hoang-Van and Luong Tran-Thi, Jan. 2022, An efficient secure sum of multi-scalar products protocol base on elliptic curve, Journal of Science and Technology on Information security, vol. 2, no. 14, pp. 18-25.

CT8. Vũ Thị Vân, Lương Thế Dũng, Hoàng Văn Quân và Trần Thị Lượng, 11/2022, Giải pháp hệ gợi ý có đảm bảo tính riêng tư hiệu quả và thực tiễn, Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Nhà xuất bản Khoa học tự nhiên và Công nghệ.

TÀI LIỆU THAM KHẢO

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, June 2005.

[2] Esma Aïmeur, Gilles Brassard, José M Fernandez, and Flavien Serge Mani Onana. Alambic: a privacy-preserving recommender system for electronic commerce. *International Journal of Information Security*, 7(5):307–334, 2008.

[3] Muhammad Ammad-Ud-Din, Elena Ivannikova, Suleiman A Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint arXiv:1901.09888*, 2019.

[4] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Antonio Ferrara, and Fedelucio Narducci. User-controlled federated matrix factorization for recommender systems. *Journal of Intelligent Information Systems*, 58(2):287–309, 2022.

[5] Muhammad Asad, Saima Shaukat, Ehsan Javanmardi, Jin Nakazato, and Manabu Tsukada. A comprehensive survey on privacy-preserving techniques in federated recommendation systems. *Applied Sciences*, 13(10):6201, 2023.

[6] Shafika Assaad, Rawan Chaaban, Fida Tannous, and Christy Costanian. Dietary habits and helicobacter pylori infection: a cross sectional study at a lebanese hospital. *BMC gastroenterology*, 18:1–13, 2018.

[7] Shahriar Badsha. *Privacy preserving recommender systems*. PhD thesis, RMIT University, Melbourne, Victoria, Australia, 48-77, 2018.

[8] Shahriar Badsha, Xun Yi, and Ibrahim Khalil. A practical privacy-preserving recommender system. *Data Science and Engineering*, 1:161–177, 2016.

[9] Rabeya Bosri, Mohammad Shahriar Rahman, Md Zakirul Alam Bhuiyan, and Abdullah Al Omar. Integrating blockchain with artificial intelligence for privacy-preserving recommender systems. *IEEE Transactions on Network Science and Engineering*, 8(2):1009–1018, 2020.

[10] Hương Bùi Song. *Nghiên cứu mối liên quan giữa kháng thể kháng Nucleosome và C1q với mức độ hoạt động của bệnh và tổn thương thận trong Lupus ban đỏ hệ thống trẻ em*. PhD thesis, 2019.

[11] Ran Canetti. Decisional diffie-hellman assumption., 2005.

[12] Di Chai, Leye Wang, Kai Chen, and Qiang Yang. Secure federated matrix factorization. *IEEE Intelligent Systems*, 36(5):11–20, 2020.

[13] Yu Fang Chung, Kuo Hsuan Huang, Feipei Lai, and Tzer Shyong Chen. Id-based digital signature scheme on the elliptic curve cryptosystem. *Computer Standards & Interfaces*, 29(6):601–604, 2007.

[14] Ran Cohen. Secure multiparty computation: Introduction. *Seminar on Secure MultiParty Computation, Tel Aviv University*, 2017.

[15] Changyu Dong and Liqun Chen. A fast secure dot product protocol with application to privacy preserving association rule mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 606–617. Springer, 2014.

[16] Yongjie Du, Deyun Zhou, Yu Xie, Jiao Shi, and Maoguo Gong. Federated matrix factorization for privacy-preserving recommender systems. *Applied soft computing*, 111:107700, 2021.

[17] Lương Thế Dũng. *Nghiên cứu xây dựng một số giải pháp đảm bảo an toàn thông tin trong quá trình khai phá dữ liệu*. PhD thesis, Viện Khoa học và Công nghệ quân sự, 2011.

[18] Zekeriya Erkin, Michael Beye, Thijs Veugen, and Reginald L Lagendijk. Efficiently computing private recommendations. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5864–5867. IEEE, 2011.

[19] Zekeriya Erkin, Thijs Veugen, Tomas Toft, and Reginald L Lagendijk. Generating private recommendations efficiently using homomorphic encryption and data packing. *IEEE transactions on information forensics and security*, 7(3):1053–1066, 2012.

[20] Chen Gao, Xiangning Chen, Fuli Feng, Kai Zhao, Xiangnan He, Yong Li, and Depeng Jin. Cross-domain recommendation without sharing user-relevant data. In *The world wide web conference*, pages 491–502, 2019.

[21] Chen Gao, Chao Huang, Yue Yu, Huandong Wang, Yong Li, and Depeng Jin. Privacy-preserving cross-domain location recommendation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–21, 2019.

[22] Chen Gao, Yong Li, Fuli Feng, Xiangning Chen, Kai Zhao, Xiangnan He, and

Depeng Jin. Cross-domain recommendation with bridge-item embeddings. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1):1–23, 2021.

[23] Bart Goethals, Sven Laur, Helger Lipmaa, and Taneli Mielikäinen. On private scalar product computation for privacy-preserving data mining. In *Information Security and Cryptology–ICISC 2004: 7th International Conference, Seoul, Korea, December 2-3, 2004, Revised Selected Papers 7*, pages 104–120. Springer, 2005.

[24] Oded Goldreich. *Foundations of Cryptography, Volume 2*. Cambridge university press Cambridge, 2004.

[25] Fatma Esra Gunes, Nural Bekiroglu, Nese Imeryuz, and Mehmet Agirbasli. Relation between eating habits and a high body mass index among freshman students: a cross-sectional study. *Journal of the American College of Nutrition*, 31(3):167–174, 2012.

[26] LUlu Han, Weiqi Luo, Anjia Yang, Yudan Cheng, Rongxing Lu, Junzuo Lai, and Yandong Zheng. Ad hocprivacy-preserving cross-domain point-of-interests recommendation based on friendship in lbss. *Available at SSRN 4356599*, 2023.

[27] Tu Bao Ho et al. Privacy preserving frequency mining in 2-part fully distributed setting. *IEICE TRANSACTIONS on Information and Systems*, 93(10):2702–2708, 2010.

[28] Hongsheng Hu, Gillian Dobbie, Zoran Salcic, Meng Liu, Jianbing Zhang, Lingjuan Lyu, and Xuyun Zhang. Differentially private locality sensitive hashing based federated recommender system. *Concurrency and Computation: Practice and Experience*, 35(14):e6233, 2023.

[29] Weiming Huang, Baisong Liu, and Hao Tang. Privacy protection for recommendation system: a survey. In *Journal of Physics: Conference Series*, volume 1325, page 012087. IOP Publishing, 2019.

[30] Mubashir Imran, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Alexander Zhou, and Kai Zheng. Refrs: Resource-efficient federated recommender system for dynamic and diversified user preferences. *ACM Transactions on Information Systems*, 41(3):1–30, 2023.

[31] SK Hafizul Islam and GP Biswas. A pairing-free identity-based authenticated group key agreement protocol for imbalanced mobile networks. *Annals of télécommunications-annales des telecommunications*, 67:547–558, 2012.

[32] Somesh Jha, Luis Kruger, and Patrick McDaniel. Privacy preserving clustering. In *Computer Security–ESORICS 2005: 10th European Symposium on Research in*

Computer Security, Milan, Italy, September 12-14, 2005. Proceedings 10, pages 397–417. Springer, 2005.

[33] Jonathan Katz and Yehuda Lindell. *Introduction to modern cryptography: principles and protocols*. Chapman and hall/CRC, 2007.

[34] Harmanjeet Kaur, Neeraj Kumar, and Shalini Batra. An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system. *Future Generation Computer Systems*, 86:297–307, 2018.

[35] Hiroaki Kikuchi, Hiroyasu Kizawa, and Minako Tada. Privacy-preserving collaborative filtering schemes. In *2009 International Conference on Availability, Reliability and Security*, pages 911–916. IEEE, 2009.

[36] Hiroaki Kikuchi, Kei Nagai, Wakaha Ogata, and Masakatsu Nishigaki. Privacy-preserving similarity evaluation and application to remote biometrics authentication. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 3–14. Springer, 2008.

[37] Donggyu Kim, Garam Lee, and Sungwoo Oh. Toward privacy-preserving text embedding similarity with homomorphic encryption. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 25–36, 2022.

[38] Jinsu Kim, Dongyoung Koo, Yuna Kim, Hyunsoo Yoon, Junbum Shin, and Sungwook Kim. Efficient privacy-preserving matrix factorization for recommendation via fully homomorphic encryption. *ACM Transactions on Privacy and Security (TOPS)*, 21(4):1–30, 2018.

[39] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[40] Neal Koblitz, Alfred Menezes, and Scott Vanstone. The state of elliptic curve cryptography. *Designs, codes and cryptography*, 19:173–193, 2000.

[41] Qi Le, Enmao Diao, Xinran Wang, Ali Anwar, Vahid Tarokh, and Jie Ding. Personalized federated recommender systems with private and partially federated autoencoders. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 1157–1163. IEEE, 2022.

[42] Matt Lepinski and S Kent. Rfc 5114-additional diffie-hellman groups for use with ietf standards. *See Section 2.1. 1024-bit MODP Group with 160-bit Prime Order Subgroup*, 2008.

[43] Liangjie Lin, Yuchen Tian, and Yang Liu. A blockchain-based privacy-preserving recommendation mechanism. In *2021 IEEE 5th international conference on cryptography, security and privacy (CSP)*, pages 74–78. IEEE, 2021.

[44] Yehuda Lindell. Secure multiparty computation. *Communications of the ACM*, 64(1):86–96, 2020.

[45] Manfred Lochter and Johannes Merkle. Elliptic curve cryptography (ecc) brainpool standard curves and curve generation. Technical report, 2010.

[46] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K Liu, and Jun Shao. Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4):46–50, 2014.

[47] Junwei Luo, Xuechao Yang, Xun Yi, and Fengling Han. Privacy-preserving recommendation system based on user classification. *Journal of Information Security and Applications*, 79:103630, 2023.

[48] Junwei Luo, Xun Yi, Fengling Han, and Xuechao Yang. An efficient privacy-preserving recommender system in wireless networks. *Wireless Networks*, pages 1–12, 2022.

[49] The Dung Luong and Dang Hung Tran. A new method of privacy preserving computation over 2-part fully distributed data. In *The 9th International Conference on Computing and InformationTechnology (IC2IT2013) 9th-10th May 2013 King Mongkut's University of Technology North Bangkok*, pages 115–123. Springer, 2013.

[50] The Dung Luong and Dang Hung Tran. Erratum: Privacy preserving frequency-based learning algorithms in two-part partitioned record model. In *Knowledge and Systems Engineering: Proceedings of the Fifth International Conference KSE 2013, Volume 2*, pages 445–445. Springer, 2014.

[51] Nguyễn Thị Thu Luyến. *Ứng dụng thuật toán cây quyết định cho hệ thống gợi ý sản phẩm trong thương mại điện tử.* PhD thesis, 2024.

[52] Manal Hamed Mahmoud and Amal Said Taha. The association between eating habits and body mass index among nursing students. *J Nurs Health Sci*, 6(3):14–26, 2017.

[53] Sitikantha Mallik and Abhaya Kumar Sahoo. A comparison study of different privacy preserving techniques in collaborative filtering based recommender system. In *Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018*, pages 193–203. Springer, 2020.

[54] Shagufta Mehnaz, Gowtham Bellala, and Elisa Bertino. A secure sum protocol and its application to privacy-preserving multi-party analytics. In *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies*, pages 219–230, 2017.

[55] E Mu, C Shao, and V Miglani. Privacy preserving collaborative filtering. 2017.

[56] Nguyễn Thái Nghe, Mai Nhựt Tự, and Nguyễn Hữu Hoà. Một tiếp cận đa quan hệ cho hệ thống gợi ý. *PROCEEDING of Publishing House for Science and Technology*, 2017.

[57] Taiwo Blessing Ogunseyi, Cossi Blaise Avoussoukpo, and Yiqiang Jiang. Privacy-preserving matrix factorization for cross-domain recommendation. *IEEE Access*, 9:91027–91037, 2021.

[58] Taiwo Blessing Ogunseyi, Tang Bo, and Cheng Yang. A privacy-preserving framework for cross-domain recommender systems. *Computers & Electrical Engineering*, 93:107213, 2021.

[59] Abdullah Al Omar, Rabeya Bosri, Mohammad Shahriar Rahman, Nasima Begum, and Md Zakirul Alam Bhuiyan. Towards privacy-preserving recommender system with blockchains. In *Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference, DependSys 2019, Guangzhou, China, November 12–15, 2019, Proceedings 5*, pages 106–118. Springer, 2019.

[60] Sonam Penjor, Einar B Thorsteinsson, Ian Price, and Natasha M Loi. Parenting style, distress, and problematic alcohol use in bhutan. *Cogent Psychology*, 6(1):1579503, 2019.

[61] Vasileios Perifanis, George Drosatos, Giorgos Stamatelatos, and Pavlos S Efraimidis. Fedpoirec: Privacy-preserving federated poi recommendation with social influence. *Information Sciences*, 623:767–790, 2023.

[62] Vasileios Perifanis and Pavlos S Efraimidis. Federated neural collaborative filtering. *Knowledge-Based Systems*, 242:108441, 2022.

[63] Hendrik Pfaff. A survey on current recommender systems. *Frankfurt University of Applied Sciences, Faculty of Computer Science and Engineering*, 2021.

[64] H. Polat and Wenliang Du. Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE International Conference on Data Mining*, ICDM-03. IEEE Comput. Soc, 2003.

[65] Dhanya Pramod. Privacy-preserving techniques in recommender systems: state-of-

the-art review and future research agenda. *Data Technologies and Applications*, 57(1):32–55, 2022.

[66] Yogachandran Rahulamathavan. Privacy-preserving similarity calculation of speaker features using fully homomorphic encryption. *arXiv preprint arXiv:2202.07994*, 2022.

[67] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[68] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[69] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[70] Babak Siabi, Mehdi Berenjkoub, and Willy Susilo. Optimally efficient secure scalar product with applications in cloud computing. *IEEE Access*, 7:42798–42815, 2019.

[71] Minako Tada, Hiroaki Kikuchi, and Sutheera Puntheeranurak. Privacy-preserving collaborative filtering protocol based on similarity between items. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, 2010.

[72] Qiang Tang and Husen Wang. Privacy-preserving hybrid recommender system. In *Proceedings of the Fifth ACM International Workshop on Security in Cloud Computing*, pages 59–66, 2017.

[73] Trần Nguyễn Minh Thư and Nguyễn Thuỳ Quỳnh. Đặc trưng tâm trạng trong hệ thống gợi ý nhạc. *PROCEEDING of Publishing House for Science and Technology*, 2019.

[74] Jaideep Vaidya and Chris Clifton. Privacy preserving naive bayes classifier for vertically partitioned data. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 522–526. SIAM, 2004.

[75] Pranav Verma, Harshul Vaishnav, Anish Mathuria, and Sourish Dasgupta. An enhanced privacy-preserving recommender system. In *Security and Privacy: Second ISEA International Conference, ISEA-ISAP 2018, Jaipur, India, January, 9–11, 2019, Revised Selected Papers 2*, pages 247–260. Springer, 2019.

[76] Thijs Veugen, Frank Blom, Sebastiaan JA De Hoogh, and Zekeriya Erkin. Secure comparison protocols in the semi-honest model. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1217–1228, 2015.

[77] Duy-Hien Vu. Privacy-preserving naive bayes classification in semi-fully distributed data model. *Computers & Security*, 115:102630, 2022.

[78] Duy-Hien Vu, Trong-Sinh Vu, and The-Dung Luong. An efficient and practical approach for privacy-preserving naive bayes classification. *Journal of information Security and Applications*, 68:103215, 2022.

[79] Xiaoya Wang, Tianqing Zhu, Wei Ren, Dongmei Zhang, and Ping Xiong. Migrating federated learning to centralized learning with the leverage of unlabeled data. *Knowledge and Information Systems*, 65(9):3725–3752, 2023.

[80] Uwe Wolfradt, Susanne Hempel, and Jeremy NV Miles. Perceived parenting styles, depersonalisation, anxiety and coping behaviour in adolescents. *Personality and individual differences*, 34(3):521–532, 2003.

[81] Jiang Wu and Douglas R Stinson. On the security of the elgamal encryption scheme and damgard's variant. *Cryptology ePrint Archive*, 2008.

[82] Zenghua Xie, Xu Guo, Liyuan Han, Xin Wang, Qianqian Yan, Chang Shu, Zhenyi Fan, and Miaomiao Zhao. Differences in primary and secondary stroke prevention strategies for chinese men and women. *Preventive Medicine Reports*, 33:102219, 2023.

[83] Kun Xu, Weidong Zhang, and Zheng Yan. A privacy-preserving mobile application recommender system based on trust evaluation. *Journal of computational science*, 26:87–107, 2018.

[84] Dengcheng Yan, Yuchuan Zhao, Zhongxiu Yang, Ying Jin, and Yiwen Zhang. Fedcdr: Privacy-preserving federated cross-domain recommendation. *Digital Communications and Networks*, 8(4):552–560, 2022.

[85] Alper Yargic and Alper Bilge. Privacy-preserving multi-criteria collaborative filtering. *Information Processing & Management*, 56(3):994–1009, 2019.

[86] Tianjiao Yin, Yuezu Lv, and Wenwu Yu. Accurate privacy preserving average consensus. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(4):690–694, 2019.

PHỤ LỤC: THUẬT TOÁN VÉT CẠN TÍNH LOGARIT RỜI RẠC TRÊN ĐƯỜNG CONG ELLIPTIC

Thuật toán vét cạn được biết đến là thuật toán tồi nhất để tính logarit rời rạc. Tuy nhiên, thuật toán này tỏ ra khá hiệu quả trong trường hợp cần tính toán nhiều giá trị logarit rời rạc cùng một lúc và không gian nghiệm của bài toán logarit rời rạc bị giới hạn bởi các số nguyên vừa và nhỏ.

Thuật toán 3: Thuật toán vét cạn tính logarit rời rạc trên đường cong Elliptic

---

Đầu vào: Đường cong Elliptic $\mathbb{E}(\mathbb{Z}_d)$ với điểm vô cực O và d là một số nguyên tố lớn, G là một điểm cơ sở của đường cong E với bậc d, danh sách m điểm $B_i(x_i, y_i)$.
Đầu ra: Các giá trị $m_i \in \{0, \ldots, \max\}$ thoả mãn $B_i = b_i G$.

$K \leftarrow O(0,0)$;
$Check[] \leftarrow 0$;
$Result[] \leftarrow 0$;
$Num \leftarrow 0$;
for $\{j = 0$ to $max\}$
  if $\{Num = m\}$ **return** $Result$;
  else
    for $\{i = 0$ to $m\}$
      if $Num = m$ **return** $Result$;
      if $(Check[Num] == 0) \wedge (K == B_i)$
        $Check[Num] \leftarrow 1$;
        $Result[i] \leftarrow Num$;
        $Num \leftarrow Num + 1$;
      $K \leftarrow K + G$;
