Khai phá tập phổ biến đảm bảo tính riêng tư cho dữ liệu phân mảnh dọc sử dụng hệ mật ElGamal

Nguyễn Văn Chung, Trần Đức Sự

***Tóm tắt*— Khai phá dữ liệu đảm bảo tính riêng tư ngày càng nhận sự quan tâm của cộng đồng nghiên cứu, đặc biệt là khai phá tập phổ biến có đảm bảo tính riêng tư là một chủ đề được đề cập nhiều trong thời gian gần đây. Bài báo này, tác giả đề xuất một phương pháp khai phá tập phổ biến đảm bảo tính riêng tư cho mô hình dữ liệu phân mảnh dọc trên 3 thành viên. Phương pháp đề xuất có hiệu quả tương đương và an toàn hơn các phương pháp hiện có, mỗi thành viên tham gia giao thức có thể chống lại sự thông đồng của 2 thành viên còn lại và không làm lộ thông tin trong quá trình thực thi giao thức.**

***Abstract*— Privacy-preserving data mining is increasingly receiving the attention of the research community, especially privacy-preserving frequent set mining, which is a topic that has been discussed a lot in recent times. In this paper, we propose a novel technique for privacy-preserving freuquent itemset mining in three party vertically distributed data. We show that the proposed technique is as efficient as the best existing protocols for performing the same task, and more secure than the most secured protocols with against collusion, 3 members joined the protocol the 2 members colluded not to reveal data of the other member.**

Từ khóa— đảm bảo tính riêng tư; khai phá dữ liệu bảo mật; khai phá tập phổ biến.

Keywords— privacy-preserving; frequent itemset mining; secure data mining.

I. Giới thiệu

Khai phá dữ liệu phân tán [1] trong đó dữ liệu được sở hữu bởi các thành viên khác nhau. Các thành viên này muốn kết hợp để khai phá trên bộ dữ liệu chung bởi nó mang lại lợi ích và kết quả chính xác hơn khai phá dữ liệu cục bộ, tuy nhiên mỗi thành viên tham gia muốn giữ bí mật về dữ dữ liệu của mình. Bài báo nghiên cứu phương pháp khai phá tập phổ biến trên dữ liệu phân tán mà không tiết lộ dữ liệu riêng tư của mỗi thành viên.

Khai phá luật kết hợp hay khai phá tập phổ biến có đảm bảo tính riêng tư là một chủ đề đang được nhiều nhà nghiên cứu quan tâm. Một số bài báo về chủ đề này như trong [2-7], tuy nhiên các bài báo này sử dụng kỹ thuật ẩn các luật nhạy cảm. Trong khai phá dữ liệu phân tán, một phương pháp đơn giản nhằm đảm bảo tính riêng tư của dữ liệu là làm nhiễu dữ liệu gốc. Quá trình chuyển đổi dữ liệu gốc thành dữ liệu mới ẩn đi một số thông tin gọi là quá trình gây nhiễu [8-12]. Quá trình khai phá trên dữ liệu đã gây nhiễu có thể làm giảm nguy cơ lộ thông tin nhạy cảm. Mặc dù các giải pháp này hiệu quả, nhưng kết quả khai phá trên dữ liệu bị nhiễu giảm độ chính xác trong quá trình khai phá dữ liệu.

Một số giải pháp khác dựa trên tính toán bảo mật cho việc khai phá các tập phổ biến có đảm bảo tính riêng tư như của Kantarcioglu và Clifton [13] đưa ra một giải pháp cho dữ liệu được phân mảnh ngang, sử dụng giao thức tính toán bảo mật chung Yao. Goldreich đã chỉ ra trong [14], các giao thức tính toán bảo mật chung rất tốn kém trong thực tế. Do đó, một số giải pháp khác cho dữ liệu được phân mảnh theo chiều ngang có hiệu suất tốt hơn được đề xuất trong [15-18].

Bài báo này, tác giả giới thiệu giao thức xử lý trên dữ liệu phân mảnh dọc trên 3 thành viên: Một phần thông tin về mỗi giao dịch sẽ nằm ở các thành viên, nhưng không thành viên nào chứa đầy đủ thông tin về một giao dịch bất kỳ. Trong [18] Vaidya và Clifton giới thiệu phương pháp tính tích vô hướng bảo mật nhiều thành viên khai

Bài báo được nhận ngày 01/7/2022 Bài báo được nhận xét bởi phản biện thứ nhất vào ngày 09/7/2022 và được chấp nhận đăng vào ngày 19/7/2022. Bài báo được nhận xét bởi phản biện thứ hai vào ngày 15/7/2022 và được chấp nhận đăng vào ngày 22/7/2022.

phá luật kết hợp trên dữ liệu phân mảnh dọc, hạn chế của phương pháp này là độ phức tạp tính toán và chi phí truyền thông tương ứng là O(mn) và $O(mn^2)$, trong đó n là số thành viên trong giao thức và m là số giao dịch. Vaidya và cộng sự [19] đưa ra một giải pháp đại số cho dữ liệu được phân mảnh dọc, tuy nhiên giải pháp này có thể làm lộ thông tin riêng tư của các thành viên. Hơn nữa, để xử lý một tập phổ biến của các ứng viên, chi phí tính toán của nó là bình phương về số lượng giao dịch. Zhong [20] đã đề xuất một phương pháp khai phá luật kết hợp dựa trên kỹ thuật mã hóa đồng cấu trên dữ liệu phân mảnh dọc, với độ phức tạp tính toán là O(n) và chi phí truyền thông là O(mn). Hạn chế của giao thức này là có thể tiết lộ thông tin riêng tư: Giao thức yêu cầu một thành viên P tạo cặp khóa (khóa công khai và khóa bí mật) dựa trên thuật toán mã hóa khóa công khai. Dữ liệu riêng tư của tất cả các thành viên khác được mã hóa dựa trên khóa công khai từ P. Nếu P không trung thực và thông đồng với các thành viên không trung thực khác, dữ liệu riêng tư của thành viên trung thực có thể bị tiết lộ. Li và cộng sự [21] đề xuất phương pháp khai phá tập phổ biến trên dữ liệu phân mảnh dọc sử dụng kỹ thuật mã hóa đồng cấu. Bằng cách mã hóa các mục sử dụng mã thay thế và thêm các giao dịch giả để giảm thiểu các cuộc tấn công phân tích tần số vào mật mã thay thế, với mức độ đảm bảo tính riêng tư tương đương với giải pháp của Vaidya và Clifton [18].

Dựa trên sự tìm hiểu của tác giả, giao thức của Zhong [20] có độ phức tạp tính toán và chi phí truyền thông thấp nhất để thực hiện khai phá luật kết hợp trên dữ liệu phân mảnh dọc. Giao thức của Vaidya và Clifton [18] và Li cùng cộng sự [21] bảo vệ tính riêng tư tốt nhất, bảo vệ tính riêng tư của mỗi thành viên và chống được sự thông đồng của n - 2 thành viên không trung thực, tuy nhiên trong trường hợp chỉ có 3 thanh viên nếu 2 thành viên thông đồng với nhau sẽ làm lộ thông tin riêng tư của thành viên còn lại.

Các giao thức đã đề xuất đều có khả năng chống thông đồng n - 2 thành viên không trung thực, với trường hợp có 3 thành viên các giao thức này không phù hợp. Bài báo này đề xuất một phương pháp khai phá tập phổ biến đảm bảo tính riêng tư có cùng chi phí truyền thông với giao thức của Zhong, đảm bảo tính riêng tư tốt hơn giao thức của Vaidya và Cliftons và Li cùng cộng sự. Giao thức đề xuất áp dụng cho 3 thành viên, có thể bảo vệ tính riêng tư của mỗi thành viên và chống được sự thông đồng của 2 thành viên không trung thực.

II. ĐẶT VẤN ĐỀ

A. Luật kết hợp và tập phổ biến

Vấn đề khai phá luật kết hợp được trình bày trong [1]. Đặt $I = \{i_1, i_2, ..., i_d\}$ là tập các hạng mục; DB là cơ sở dữ liệu giao dịch, trong đó mỗi giao dịch T_j (j = 1, .., m) là một tập các mục sao cho $T_j \subseteq I$. Cho X là một tập mục trong I, một giao dịch T_j chứa X khi và chỉ khi $X \subseteq T_j$. Luật kết hợp là một liên kết có dạng $X \rightarrow Y$, trong đó $X \subseteq I$, $Y \subseteq I$ và $X \cap Y = \phi$.

Luật kết hợp có độ hỗ trợ s% nếu tỷ lệ phần trăm của các giao dịch chứa cả X và Y trong DB là s%. Luật có độ tin cậy c% nếu tỷ lệ phần trăm của các giao dịch chứa cả X và Y trong tổng số giao dịch chứa X là c%.

Đặt X.count là độ hỗ trợ của tập mục X, ký hiệu số lượng giao dịch có chứa bộ X trong DB và $|DB|$ là tổng số giao dịch trong DB. Các luật kết hợp mạnh là các luật thỏa mãn độ hỗ trợ tối thiểu (s_{min}) và độ tin cậy tối thiểu (c_{min}) được xác định bởi người dùng.

Một tập các mục gọi là một tập mục, một tập mục chứa k mục gọi là k-tập mục (k-Itemset). Độ hỗ trợ của một tập mục là số lượng giao dịch có chứa các tập mục đó. Độ hỗ trợ tối thiểu được xác định là $t = s_{min}|DB|$. Một tập mục được gọi là tập mục phổ biến nếu độ hỗ trợ của nó lớn hơn độ hỗ trợ tối thiểu. Khai phá luật kết hợp là một quá trình 2 bước: (1) Tìm tất cả các tập phổ biến; (2) Tạo luật kết hợp từ các tập phổ biến. Như vậy trong khai phá luật kết hợp vấn đề quan trọng là việc tìm ra các tập phổ biến.

B. Biểu diễn ma trận

Trong khai phá luật kết hợp, tập giao dịch có thể biểu diễn bằng ma trận nhị phân M. Mỗi hàng của ma trận tương ứng với một giao dịch, mỗi cột tương ứng với một mục. Biểu diễn một phần tử của ma trận bằng M(i, j), trong đó M(i, j) = 1 nếu giao dịch thứ i chứa mục thứ j, bằng 0 nếu ngược lại.

Tính toán độ hỗ trợ cho tập X: gọi C_X là tập các cột tương ứng với một tập mục, ta viết X = {I_j | j∈C_X}. Như đã trình bày trong [22], độ hỗ trợ X.count có thể được biểu diễn bằng giá trị của tất cả các cột trong tập C_X như sau:

$$s = \sum_{i=1}^{m} \prod_{j \in C_X} M(i,j)$$

Do đó, để kiểm tra X có phổ biến không, ta cần tính giá trị s và so sánh với ngưỡng t.

C. Xác định tập phổ biến trong dữ liệu phân mảnh dọc

Giả sử DB được phân mảnh dọc trên 3 thành viên P_1, P_2, P_3 trong đó mỗi thành viên sở hữu một tập các cột của ma trận M tương ứng là C_1, C_2, C_3 với $C_1 \cup C_2 \cup C_3$ = C và $C_i \cap C_j = \phi$ $(\forall i \neq j)$.

Mục đích là thiết kế giao thức phân tán để các thành viên có thể tìm các tập phổ biến từ DB trong khi vẫn đảm bảo tính riêng tư của dữ liệu được sở hữu bởi mỗi thành viên. Tính riêng tư ở đây nghĩa là bảo vệ các thông tin cá nhân, độ hỗ trợ cục bộ của các tập phổ biến trên các cột dữ liệu của mỗi thành viên.

Xét tập mục X $= \{I_j \mid j \in C_X\} \subset$ I. Lưu ý tập cột C_X của tập mục này được phân chia thành các tập con: $C_X \cap C_1$, $C_X \cap C_2$, $C_X \cap C_3$, được sở hữu lần lượt bởi các thành viên P_1, P_2, P_3.

Ta có:

$$s = \sum_{i=1}^{m} \prod_{j \in C_X \cap C_1} M(i,j) \prod_{j \in C_X \cap C_3} M(i,j) \prod_{j \in C_X \cap C_3} M(i,j)$$

trong đó $\prod_{j \in C_X \cap C_1} M(i,j)$, $\prod_{j \in C_X \cap C_3} M(i,j)$, $\prod_{j \in C_X \cap C_3} M(i,j)$ lần lượt được tính bởi P_1, P_2, P_3.

Nếu đặt $u_{ik} = \prod_{j \in C_X \cap C_k} M(i,j)$ thì mỗi thành viên P_i sở hữu một vectơ cột nhị phân riêng tư U_i = (u_{i1}, ..., u_{im}), trong đó mỗi $u_{ij} \in \{0,1\}$ (i = 1, ..., 3 và j = 1, ..., m) và độ hỗ trợ của tập X là $s = \sum_{j=1}^{m} \prod_{i=1}^{3} u_{ij}$.

Giao thức khai phá tập phổ biến X đảm bảo tính riêng tư được định nghĩa như sau:

Định nghĩa 1. Giao thức khai phá tập phổ biến X đảm bảo tính riêng tư là giao thức tính toán độ hỗ trợ s của tập X và kiểm tra điều kiện $s \geq t$ (với ngưỡng công khai t = s_{min} |DB| cho trước) trong khi mỗi thành viên P_i không tiết lộ vectơ nhị phân riêng U_i cho các thành viên khác.

D. Mô hình hệ thống

Trong nghiên cứu này, giao thức đề xuất cho việc khai phá tập phổ biến đảm bảo tính riêng tư được giả thiết thực hiện trong mô hình bán trung thực. Nói cách khác, các thành viên tuân theo quy tắc tính toán của giao thức nhưng tồn tại một số thành viên không trung thực sẵn sàng chia sẻ dữ liệu với nhau để khai thác dữ liệu riêng tư của các thành viên trung thực. Do đó yêu cầu được đặt ra đối với giao thức đề xuất là không có thông tin riêng tư nào của các thành viên trung thực bị tiết lộ trừ kết quả đầu ra của giao thức.

Giả sử tất cả các thành viên đều trực tuyến, trong đó giữa hai thành viên bất kỳ tồn tại kênh kết nối được xác thực (không nhất thiết phải an toàn). Trước khi thực hiện giao thức, mỗi thành viên có một bộ khóa bí mật x_i và một bộ khóa công khai y_i tương ứng theo sơ đồ mã hóa ElGamal, trong đó khóa công khai của mỗi thành viên được công khai cho các thành viên khác trong hệ thống.

Định nghĩa về đảm bảo tính riêng tư đối với giao thức khai phá tập phổ biến đảm bảo tính riêng tư trong mô hình bán trung thực được phát biểu như sau:

Định nghĩa 2. Giả sử để xác định tập X là tập mục phổ biến, mỗi thành viên P_i có vectơ đầu vào riêng tư U_i và sử dụng bộ khóa bí mật x_i với một bộ khóa công khai y_i tương ứng. Một giao thức khai phá tập phổ biến X (được nêu trong định nghĩa 1) đảm bảo tính riêng tư trong mô hình bán trung thực nếu với mọi tập I ⊂ {1,..k} thỏa mãn | I | = t, tồn tại thuật toán xác suất theo thời gian đa thức M sao cho:

$$\{M(s, [U_i, x_i]_{i \in I}, [y_j]_{j \notin I}\} \stackrel{c}{\equiv} \{view_{\{P_i\}_{i \in I}}([U_i, x_i]_{i=1}^{3})\}$$

trong đó $\stackrel{c}{\equiv}$ là ký hiệu không thể phân biệt về mặt tính toán.

Định nghĩa 2 phát biểu rằng: Giao thức khai phá tập phổ biến an toàn trong mô hình bán trung thực nếu như chúng ta có thể mô phỏng quá trình tính toán các thông điệp mà những thành viên nguy hại quan sát được trong quá trình tham gia

giao thức mà chỉ cần sử dụng kết quả đầu ra s, các giá trị bí mật của những thành viên nguy hại và các khóa công khai. Như vậy, nếu chỉ ra tồn tại một bộ mô phỏng M như trên thì giao thức đề xuất được gọi là an toàn.

E. Công cụ bảo mật

Công cụ bảo mật được bài báo sử dụng là mã hóa ElGamal, kỹ thuật tái ngẫu nhiên và kỹ thuật giải mã chung. Những kỹ thuật này cũng được sử dụng cho một số bài toán được trình bày trong [23-25].

1. Kỹ thuật giải mã chung và mã hóa ElGamal

Giả sử mỗi thành viên có khóa bí mật $x_i \in [1, q - 1]$, khóa công khai $y_i = g^{x_i}$ (giả sử các thành viên được đánh số từ 1 đến n).

Ta có:

$$x = \sum_{i=1}^{n} x_i \; y = \prod_{i=1}^{n} y_i = g^x$$

Trong giao thức các thành viên sử dụng giá trị y làm khóa công khai để mã hóa dữ liệu của mình, bài báo sử dụng các biến thể của mã hóa ElGamal trong đó các thông điệp m được biến đổi thành g^m [33] trước khi mã hóa.

$$C = (a, h) = (g^m y^r, g^r)$$

Giải mã bản hóa này cần khóa bí mật x, trong khi không có thành viên nào biết x. Để giải mã bản mã C tất cả các thành viên tham gia sẽ cùng thực hiện giải mã phân tán. Mỗi thành viên tham gia tính toán và công bố h^{x_i}. Bản rõ sau đó có thể được giải mã bằng công thức:

$$\frac{a}{\prod h^{x_i}} = \frac{g^m g^{r\sum x_i}}{g^{r\sum x_i}} = g^m$$

2. Kỹ thuật tái ngẫu nhiên

Kỹ thuật tái ngẫu nhiên được giới thiệu trong [26] được sử dụng để bảo vệ tính riêng tư. Kỹ thuật tái ngẫu nhiên là giao thức nhiều thành viên liên quan đến một số máy chủ hỗn hợp. Đầu vào của giao thức là danh sách bản mã, đầu ra là một danh sách mới được hoán vị các bản mã đầu vào. Các thao tác được sử dụng là hoán vị thứ tự các mục và mã hóa lại, do đó nó ngẫu nhiên sắp xếp lại thứ tự các mục. Nếu tái ngẫu nhiên và hoán vị một chuỗi bản mã, thì sẽ nhận được một chuỗi bản mã khác với cùng một tập hợp bản rõ nhưng theo một thứ tự khác. Tính bảo mật của kỹ thuật này được đặc trưng bởi: Khi quan sát vào hai chuỗi mã hóa này, đối thủ không thể xác định được bất kỳ thông tin nào về sự tương đương giữa bản mã cũ và bản mã mới tương ứng.

Giao thức được đề xuất sử dụng kỹ thuật tái ngẫu nhiên dựa trên hệ mật ElGamal, trong đó mỗi thành viên đóng vai trò là một máy chủ hỗn hợp thực hiện nhiệm vụ hoán vị.

Kỹ thuật tái ngẫu nhiên được mô tả như sau: Đầu vào là danh sách bản mã ElGamal $\{(a_1, h_1), \ldots (a_m, h_m)\}$ được mã hóa bởi khóa công khai y, khóa bí mật tương ứng là x. Sau đó, mỗi thành viên tham gia lần lượt thực hiện một hoán vị ngẫu nhiên trên danh sách đầu vào bằng cách mã hóa lại các bản mã và đưa ra chúng một cách ngẫu nhiên. Đầu ra là chuỗi $\{(a_{\pi(1)}, h_{\pi(1)}), \ldots a_{\pi(m)}, h_{\pi(m)})\}$, điều này thể hiện mã hóa lại ngẫu nhiên của $\{(a_1, h_1), \ldots (a_m, h_m)\}$ và π là hàm hoán vị ngẫu nhiên trên tập k phần tử.

III. GIAO THỨC TÍNH ĐỘ HỖ TRỢ BẢO MẬT

A. Tổng quan

Giả sử X là một tập phổ biến, ta có $t \le s \le m$, tồn tại giá trị 0 trong danh sách $\lambda = \{\lambda_1 = s - 1 - t, \lambda_2 = s - 2 - t, \ldots, \lambda_k = s - k - t\}$, trong đó $k = m - t$. Với mục đích giữ bí mật giá trị s, ý tưởng cơ bản của giao thức như sau: Đặt p và q là hai số nguyên tố sao cho $p = 2q + 1$, gọi G là tập con của $\mathbb{Z}_p^*$ và g là phần tử sinh của G. Tất cả các tính toán trong phần này thực hiện trong $\mathbb{Z}_p$.

Việc xem xét giá trị $s = \sum_{j=1}^{m} \prod_{i=1}^{3} u_{ij} \ge t$, trong đó mỗi $U_i = (u_{i1}, \ldots, u_{im})(u_{ij} \in \{0, 1\})$ thuộc sở hữu của thành viên P_i. Đặt $\lambda_j = \sum_{i=1}^{n} u_{ij} - n$ (j=1, …, m), chú ý $\prod_{i=1}^{n} u_{ij} = 1$, miễn là $\lambda_j = 0$. Vì vậy, s là số $\lambda_j = 0$. Bài báo thiết kế một giao thức để tính giá trị này, với ý tưởng là các thành viên có được hoán vị ngẫu nhiên của tập $(g^{\lambda_{\pi(1)}}, \ldots, g^{\lambda_{\pi(m)}})$, trong đó $(\lambda_{\pi(1)}, \ldots, \lambda_{\pi(m)})$ là một hoán vị ngẫu nhiên của $(\lambda_1, \ldots, \lambda_m)$. Cần xây dựng một giao thức để thực hiện hàm sau:

$$(U_1, U_2, \ldots, U_n) \mapsto (g^{\lambda_{\pi(1)}}, \ldots, g^{\lambda_{\pi(m)}})$$

Nếu thu được kết quả này sẽ đạt được hai mục tiêu. Mục tiêu thứ nhất là các thành viên có thể tính được giá trị $\lambda_i = 0$ tương đương với $g^{\lambda_i} = g^0 = 1$. Sau khi có được độ hỗ trợ, có thể so sánh nó với ngưỡng t. Mục tiêu thứ hai là khi $\lambda_i \neq 0$ thì

g^{λ_i} là một số ngẫu nhiên, vì vậy cũng đạt được mục tiêu riêng tư, vì các thành viên không thể biết g^{λ_j} được tạo từ λ_i nào.

B. Thiết kế giao thức

Ý tưởng của giao thức: Với mỗi giá trị u_{ij}, thành viên giữ u_{ij} mã hóa giá trị này bằng khóa công khai của mình. Lưu ý nếu không có sự trợ giúp của tất cả các thành viên, không một thành viên nào có thể giải mã bất kỳ bản mã nào. Bằng tính chất cộng đồng cấu của sơ đồ Elgamal, các thành viên lặp n vòng để kết nối tất cả các bản mã của các giá trị nhị phân u_{ij} thu được bản mã của $\sum_{i=1}^{n} u_{ij}$. Kết thúc bước này, các thành viên thu được m bản mã của $\lambda_1, \dots, \lambda_m$.

Các thành viên thực hiện lặp n vòng để hoán vị và ngẫu nhiên tập bản mã $\lambda'_1, \dots, \lambda'_m$. Các thành viên cùng thực hiện giải mã các bản mã mới nhận được, theo thứ tự độc lập của các bản mã ban đầu. Sau đó đếm số lượng bản giải mã bằng 1 (g^0) (giá trị này bằng với độ hỗ trợ).

Giao thức tính độ hỗ trợ bảo mật:

Đầu vào: Có 3 thành viên, mỗi thành viên P_i có vectơ riêng $U_i = (u_{i1}, \dots, u_{im})$

$(u_{ij} \in \{0,1\};\ i = 1, \dots, 3;\ j = 1, \dots, m)$

Đầu ra: $s = \sum_{j=1}^{m} \prod_{i=1}^{3} u_{ij}$

Giai đoạn 1. Mã hóa

For j = 1,...,m

- For i = 1, … , 3

P_i tính $C_i(j) = (a_{ij}, h_{ij}) = (g^{u_{ij}} y^{\alpha_{ij}}, g^{\alpha_{ij}})$, trong đó α_{ij} được chọn ngẫu nhiên từ [1, q - 1].

- P_1 tính $C_j = (a_j, h_j) = (g^{-n} a_{1j}, h_{1j})$, gửi C_j đến P_2,

- P_2 tính $C_j = (a_j, h_j) = (a_j a_{2j}, h_j h_{2j})$, gửi C_j đến P_3,

- P_3 tính $C_j = (a_j, h_j) = (a_j a_{3j}, h_j h_{3j})$, gửi C_j đến P_1.

Giai đoạn 2. Ngẫu nhiên hóa và hoán vị

For i = 1, ..., 3

For j = 1, ..., m

- P_i tính $R_j = (R_j^{(1)}, R_j^{(2)}) = (a_{\pi_i(j)} y^{\delta_{\pi_i(j)}}, h_{\pi_i(j)} g^{\delta_{\pi_i(j)}})$. Trong đó π_i là một hoán vị trên {1, ..., m} và $\delta_{\pi_i(j)}$ được chọn ngẫu nhiên từ [1, q - 1].

- P_i đặt $C_j = R_j$, sau đó gửi C_j cho $P_{i+1(mod\ 3)}$

Giai đoạn 3. Tính toán cách thành phần

For j = 1, ..., m

- For i = 1, …, 3

P_i tính $h_{ij} = (h_j)^{x_i}$

- P_1 đặt $h_j = h_{1j}$ sau đó gửi h_j cho P_2
- P_2 tính $h_j = h_j h_{2j}$ sau đó gửi h_j cho P_3
- P_3 tính $h_j = h_j h_{3j}$ sau đó gửi h_j cho P_1

Giai đoạn 4. Giải mã

P_1 tính:

- s = 0
- For j = 1, ..., m

+ $d_j = a_j / h_j$

+ Nếu $d_j = 1$ thì $s = s + 1$

- Trả về giá trị s.

C. Phân tích tính chính xác

Định lý 1. Trong Giao thức tính độ hỗ trợ bảo mật, số lượng bản rõ "1" trong danh sách giải mã $\{d_1, d_2, \dots, d_m\}$ chính là độ hỗ trợ.

Chứng minh:

Tại giai đoạn 1: Các thành viên thu được

$$C_j = (a_j, h_j) = (g^{\sum_{i=1}^{3} u_{ij} - n} y^{\sum_{i=1}^{3} \alpha_{ij}}, g^{\sum_{i=1}^{3} \alpha_{ij}}) = (g^{\lambda_j} y^{\theta_j}, g^{\theta_j})$$

trong đó $\lambda_j = \sum_{i=1}^{3} u_{ij} - n$ và $\theta_j = \sum_{i=1}^{3} \alpha_{ij}$

Tại giai đoạn 2: Giao thức nhận tập (C_1, ..., C_m) và hoán vị tập này 3 lần. Do đó, các thành viên thu được: $C_j = (a_{\pi(j)}, h_{\pi(j)})$, trong đó $\pi(j)$ là kết quả của hoán vị 3 lần.

$$a_{\pi(j)} = g^{\lambda_{\pi(j)}} y^{\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)}}$$
$$h_{\pi(j)} = g^{\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)}}$$

Tại giai đoạn 3: Kết quả của giai đoạn này là $\prod_{i=1}^{3} (h_j)^{x_i}$ trong đó h_j nhận được từ giai đoạn 2.

Tại giai đoạn 4: Các thành viên thu được

$$d_j = a_j / \prod_{i=1}^{3} (h_j)^{x_i}$$

$$= \frac{g^{\lambda_{\pi(j)}} y^{(\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)})}}{g^{(\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)}) \sum_{i=1}^{3} x_i}}$$
$$= \frac{g^{\lambda_{\pi(j)}} g^{\sum_{i=1}^{3} x_i (\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)})}}{g^{(\theta_{\pi(j)} \sum_{i=1}^{3} \delta_{\pi_i(j)}) \sum_{i=1}^{3} x_i}}$$
$$= g^{\lambda_{\pi(j)}}$$

Nếu $d_j = 1$, C_j là mã hóa của g^0 nghĩa là $\lambda_j = 0$. Điều đó có nghĩa là là bộ dữ liệu xảy ra tại một bản ghi của tập dữ liệu và do đó tần số f được tăng lên 1. Kết quả của giai đoạn 4 đúng.

D. Phân tích tính riêng tư

Định lý 2. Giao thức tính độ hỗ trợ bảo mật bảo vệ tính riêng tư của 3 thành viên tham gia giao thức, chống lại sự thông đồng của 2 thành viên không trung thực.

Chứng minh: Trong giao thức đề xuất, mỗi thành viên gửi một vectơ m giá trị u_{ij} được mã hóa ElGamal. Cách duy nhất để biết giá trị thực của một thành viên là giải mã các bản mã hóa này bằng khóa riêng *x*. Trong giao thức được đề xuất, mỗi thành viên tự tạo cặp khóa của mình sau đó tất cả các thành viên cùng tính toán khóa công khai *y*. Do đó không thành viên nào biết khóa riêng *x* tương ứng với y, cách duy nhất để giải mã vectơ được mã hóa là tất cả các thành viên cùng tham gia tính toán bằng khóa bí mật của mình. Do đó, bất kỳ nhóm thành viên thông đồng nào cũng không thể giải mã được giá trị mã hóa cụ thể mà không có sự hợp tác của tất cả các thành viên khác. Kỹ thuật tái ngẫu nhiên sử dụng ý tưởng tương tự, đó là hoán vị ngẫu nhiên được thực hiện bởi tất cả các thành viên. Theo cách này, có 3 thành viên tham gia giao thức thì ngay cả 2 thành viên thông đồng cũng không thể biết được sự tương đương giữa các bản mã đầu vào và các bản mã được mã hóa lại ở đầu ra.

E. Phân tích hiệu suất

Chi phí truyền thông. Các giao tiếp chính được thực hiện ở các giai đoạn 1, 2 và 3. Có 2m thông điệp được truyền trong mỗi lần 3 vòng lặp của giai đoạn 1. Trong mỗi vòng lặp i, thành viên i sẽ gửi một bộ mã hóa cho thành viên tiếp theo. Tương tự, việc truyền dữ liệu trong giai đoạn 2 và 3 là 2m thông điệp. Toàn bộ giao thức đối xứng, nên tất cả các thành viên đều truyền lượng dữ liệu bằng nhau. Vì vậy để tính tổng chi phí truyền thông, chỉ cần nhân chi phí của một thành viên với 3. Chi phí truyền thông của giao thức trong bảng sau:

BẢNG 1. CHI PHÍ TRUYỀN THÔNG

Số vòng lặp	*3*
Số thông báo	18*m*
Số bit	18*mK*

Độ phức tạp tính toán. Chi phí tính toán của tất cả các thành viên trong giai đoạn 1 và giai đoạn 2 là lũy thừa mô đun 12m và nhân mô đun 12m. Chi phí tính toán của các thành viên hầu hết là các phép nhân mô-đun 3m và lũy thừa mô-đun 3m trong giai đoạn 3. Thành viên 1 sử dụng phép đảo ngược nhân mô-đun m trong giai đoạn 4. Tổng độ phức tạp tính toán là lũy thừa mô-đun O(m), Phép nhân mô đun O(m) và đảo nhân mô đun O(m). Các chi phí tính toán này không bao gồm chi phí sinh khóa và tính toán các tham số y. Độ phức tạp tính toán của giao thức được minh họa dưới bảng sau:

BẢNG 2. ĐỘ PHỨC TẠP TÍNH TOÁN CỦA GIAO THỨC TÍNH ĐỘ HỖ TRỢ BẢO MẬT

	Phép lũy thừa	**Phép nhân**	**Phép nghịch đảo**
Số phép toán	15*m*	15*m*	*m*
Số vòng lặp	*3*	*3*	1
Bình quân	5*m*	5*m*	*m*
Ước lượng	*O*(*m*)	*O*(*m*)	*O*(*m*)

Giao thức này thực hiện một vòng lặp với 3 thành viên, các tính toán của mỗi thành viên bằng nhau. Do đó, nó được coi là giao thức lặp 3 vòng, mỗi vòng bao gồm độ phức tạp tính toán của phép nhân mô đun O(m) và lũy thừa mô đun O(m).

IV. KẾT LUẬN

Bài báo này đề xuất giao thức khai phá tập phổ biến có đảm bảo tính riêng tư cho dữ liệu phân mảnh dọc trên 3 thành viên. Giao thức đề xuất dựa trên sơ đồ mã hóa ElGamal đảm bảo tính riêng tư và độ chính xác tốt. Bài báo đã chứng minh giao thức đề xuất có hiệu quả tương đương với giao thức hiện có và an toàn hơn các phương pháp hiện tại để chống lại sự thông đồng của 2 thành viên, không làm lộ ra thông tin riêng tư của thành viên còn lại. Ngoài ra nó cũng bảo vệ các thông tin cá nhân, độ hỗ trợ cục bộ của các tập phổ biến trên các cột dữ liệu của mỗi thành viên.

TÀI LIỆU THAM KHẢO

[1] D. W. Cheung, J. Han, V. T. Ng, A. W. Fu, and Y. Fu (1996). "A fast distributed algorithm for mining association rules," in *DIS '96: Proceedings of the fourth international conference on on Parallel and distributed information systems*, pp. 31-43, IEEE Computer Society.

[2] Y.-H. Wu, C.M. Chiang, and A. L. P. Chen (2007). "Hiding sensitive association rules with limited side effects," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 1, pp. 29-42.

[3] Dehkordi MNM, Badie K, Zadeh AKA (2009). A novel method for privacy preserving in association rule mining based on genetic algorithms. J Softw 4(6):555-562.

[4] Kasthuri S, Meyyappan T (2013). Detection of sensitive items in market basket database using association rule mining for privacy preserving. In: IEEE international conference on pattern recognition, informatics and mobile engineering (PRIME).

[5] Lin, C.W., Hong, T.P., Hsu, H.C (2014). Reducing side effects of hiding sensitive itemsets in privacy preserving data mining. The Scientific World Journal 2014, 267-289.

[6] Afzali, G. A., & Mohammadi, S. (2018). Privacy preserving big data mining: association rule hiding using fuzzy logic approach. *IET Information Security*, *12*(1), 15-24.

[7] Telikani, A., Gandomi, A. H., Shahbahrami, A., & Dehkordi, M. N. (2020). Privacy-preserving in association rule mining using an improved discrete binary artificial bee colony. *Expert Systems with Applications*, *144*, 113097.

[8] S. R. M. Oliveira and O. R. Za¨ıane (2002). "Privacy preserving frequent itemset mining," in *CRPIT '14: Proceedings of the IEEE international conference on Privacy, security and data mining*, pp. 43-54, Australian Computer Society, Inc.

[9] S. Agrawal and J. R. Haritsa (2005). "A framework for high-accuracy privacypreserving mining," in *ICDE '05: Proceedings of the 21st International Conference on Data Engineering*, pp. 193-204, IEEE Computer Society.

[10] Baotou T (2010). Research on privacy preserving classification data mining based on random perturbation Xiaolin Zhang Hongjing Bi. 1-6.

[11] Polat H. and Wenliang Du (2003). Privacy-preserving collaborative filtering using randomized perturbation techniques. *Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, IEEE Comput. Soc, 625–628.

[12] Chaytor R. and Wang K. (2010). Small domain randomization: same privacy, more utility. *Proc VLDB Endow*, 3(1–2), 608–618.

[13] M. Kantarcioglu and C. Clifton (2004). "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 16, no. 9, pp. 1026-1037.

[14] O. Goldreich (2004). *The Foundations of Cryptography, volume 2, chapter 7: General Cryptographic Protocols*. Cambridge University Press, 2nd ed.

[15] Malik, M.B., Ghazi, M.A., Ali, R (2012). Privacy preserving data mining techniques: current scenario and future prospects, in: Proceedings of Third International Conference on Computer and Communication Technology (ICCCT), IEEE. pp. 26-32.

[16] Moses, T.J., Elavarasi, K., Jayachitra, J. (2014). Privacy preserving mining of association rules in horizontally distributed databases. International Journal of Management, IT and Engineering 4, pp. 209-222.

[17] Nanavati, N.R., Lalwani, P., Jinwala, D.C. (2014). Analysis and evaluation of schemes for secure sum in collaborative frequent itemset mining across horizontally partitioned data. Journal of Engineering 2014, pp. 110-120.

[18] Hien, V. D. (2022). An Efficient Solution for Privacy-preserving Naïve Bayes Classification in Fully Distributed Data Model. Tạp Chí Khoa học - Công nghệ Trong lĩnh vực An toàn thông Tin, 1(15), 56-61.

[19] J. Vaidya and C. Clifton (2005). "Secure set intersection cardinality with application to association rule mining," *J. Comput. Secur.*, vol. 13, no. 4, pp. 593-622.

[20] J. Vaidya and C. Clifton (2002). "Privacy preserving association rule mining in vertically partitioned data," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 639-644, ACM.

[21] S. Zhong (2007). "Privacy-preserving algorithms for distributed mining of frequent itemsets," *Information Sciences*, vol. 177, no. 2, pp. 490-503.

[22] Li, L., Lu, R., Choo, K. K. R., Datta, A., & Shao, J. (2016). Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Transactions on Information Forensics and Security*, No. *11, Vol.* 8, pp. 1847-1861.

[23] J. Vaidya and C. Clifton (2002). "Privacy preserving association rule mining in vertically partitioned data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM.

[24] Zhong, Z. Yang, and T. Chen (2009). "k-anonymous data collection," *Inf. Sci.*, vol. 179, no. 17, pp. 2948-2963.

[25] M. Hirt and K. Sako (2000). "Efficient receipt-free voting based on homomorphic encryption," in *Proceedings of EuroCrypt 2000, LNCS series*, pp. 539-556, Springer-Verlag.

[26] S. Zhong, Z. Yang, and R. N. Wright (2005). "Privacy-enhancing kanonymization of customer data," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 139-147, ACM.

[27] M. Michels and P. Horster (1996). "Some remarks on a receipt-free and universally verifiable mix-type voting scheme," in *Proceedings of the International Conference on the Theory and Applications of Cryptology and Information Security: Advances in Cryptology*, pp. 125-132, Springer-Verlag.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Văn Chung

Đơn vị công tác: Trường Cao đẳng Kinh tế - Kỹ thuật Vĩnh Phúc.

Email: nguyenvanchung.vtec@gmail.com

Quá trình đào tạo: Tốt nghiệp Đại học Sư phạm Kỹ thuật Nam Định vào năm 2009; Tốt nghiệp Thạc sĩ Khoa học máy tính tại Đại học Công nghệ thông tin và Truyền thông Thái Nguyên vào năm 2014.

Hướng nghiên cứu hiện nay: Mật mã; an toàn thông tin.

Trần Đức Sự

Đơn vị công tác: Ban Cơ yếu Chính phủ.

Email: tdsu@bcy.gov.vn

Quá trình đào tạo: Tốt nghiệp Đại học Bách Khoa Hà Nội năm 1988; Nhận bằng Tiến sĩ vào năm 2004; được công nhận hàm Phó Giáo sư vào năm 2014.

Hướng nghiên cứu hiện nay: Mật mã; an toàn thông tin; khai phá dữ liệu bảo mật.