PHÁT HIỆN TẤN CÔNG CÓ ĐẢM BẢO TÍNH RIÊNG TƯ TỪ CÁC NGUỒN DỮ LIỆU MẠNG PHÂN TÁN

Nguyễn Văn Chung[1*], Nguyễn Văn Tảo[2], Trần Đức Sự[3]
[1]Trường Cao đẳng kinh tế - kỹ thuật Vĩnh Phúc,
[2]Trường Đại học Công nghệ thông tin & Truyền thông - ĐH Thái Nguyên,
[3]Ban cơ yếu Chính Phủ

TÓM TẮT

Vấn đề phát hiện tấn công có đảm bảo tính riêng tư ngày càng trở nên quan trọng. Nhiều trường hợp để phát hiện tấn công cần phải kết hợp các mạng lại với nhau, trong khi giữ được tính riêng tư của từng tập dữ liệu. Bài báo đề xuất một giải pháp phát hiện tấn công có đảm bảo tính riêng tư dựa trên khai phá luật kết hợp. Để xây dựng giải pháp, bài báo đề xuất giao thức tính tổng bảo mật cải tiến nhằm nâng cao hiệu quả trong việc khai phá luật kết hợp có đảm bảo tính riêng tư trên tập dữ liệu phân tán ngang..

Từ khóa: *Tập phổ biến, luật kết hợp, tính riêng tư, phát hiện tấn công, tổng bảo mật.*

Ngày nhận bài: 15/01/2019; Ngày hoàn thiện: 18/02/2019; Ngày duyệt đăng: 28/02/2019

ATTACK DETECTION PRIVACY PRESERVING FROM DATA DISTRIBUTED NETWORK

Nguyen Van Chung[1*], Nguyen Van Tao[2], Tran Duc Su[3]
[1]Vinh Phuc Technical and Economic College,
[2]University of Information and Communication Technology - TNU,
[3]Essential Government Committee

ABSTRACT

The problem of detection privacy attack privacy preserving is becoming increasingly important. Many cases to detect attacks need to combine networks, while maintaining the privacy of each data set. The paper studies and proposes a method detecting attacks with ensure the privacy-based mining association rules. To build a solution, the paper proposes an improved security total protocol to improve the efficiency of association rule mining to ensure privacy on horizontal distributed data sets..

Keywords: *Frequent itemsets, association rule, privacy, attack detection, Secure Sum*

Received: 15/01/2019; Revised: 18/02/2019; Approved: 28/02/2019

* Corresponding author: *Tel: 0978 955677; Email: nguyenvanchung.vtec@gmail.com*

GIỚI THIỆU

Ngày nay cùng với sự phát triển mạnh mẽ của mạng Internet, thì tội phạm máy tính cũng gia tăng. Các hình thức tấn công mạng ngày càng tinh vi và nguy hiểm hơn khiến việc bảo đảm an toàn, an ninh thông tin gặp nhiều thách thức. Nhiều giải pháp, công nghệ an ninh mạng đã được phát triển và đã có những đóng góp nhất định trong việc hạn chế các tấn công xảy ra. Một trong những công nghệ an ninh mạng mới, được sử dụng hiệu quả trong thời gian gần đây là công nghệ giám sát an toàn mạng. Quá trình hoạt động đòi hỏi các hệ thống giám sát an toàn mạng phải thu thập các thông tin từ nhiều nguồn dữ liệu khác nhau để thực hiện các thuật toán phân tích nhằm phát hiện tấn công mạng. Tuy nhiên, các tổ chức mong muốn việc giám sát phát hiện tấn công cho các hệ thống mạng của họ nhưng không muốn làm lộ các thông tin riêng tư trên hệ thống mạng của họ, do đó vấn đề đặt ra là làm thế nào để cho phép quá trình phân tích phát hiện tấn công trong khi vẫn đảm bảo thông tin riêng tư cho hệ thống của các tổ chức.

Bài báo này xem xét bài toán phân tích dữ liệu dựa trên luật kết hợp nhằm phát hiện các tấn công mạng máy tính trong khi đảm bảo tính riêng tư cho các dữ liệu thu thập được từ các hệ thống mạng. Về lĩnh vực này đã có các nghiên cứu như: khai phá luật kết hợp có đảm bảo tình riêng tư với dữ liệu mờ sử dụng giao thức tính tổng bảo mật [1], khai phá luật kết hợp có đảm bảo tình riêng tư trong việc phát hiện và phòng ngừa tấn công [2]. Để giải quyết vấn đề đặt ra. trong bài báo này chúng tôi đề xuất một giao thức tính tổng bảo mật mới hiệu quả hơn các phương pháp cũ và ứng dụng trong bài toán khai phá dữ liệu tấn công có đảm bảo tính riêng tư.

TỔNG QUAN

Luật kết hợp

Cho F = {F_1, F_2, ..., F_n} là tập các thuộc tính, D là một tập các giao dịch cơ sở dữ liệu, trong đó mỗi giao tác T là tập các thuộc tính sao cho T $\subseteq$ F. Mỗi giao dịch được kết hợp với một định danh, được gọi là TID, cho A là một bộ các thuộc tính, một giao dịch T được cho là chứa A khi và chỉ khi A $\subseteq$ T. Một luật kết hợp là một liên kết của mẫu A$\rightarrow$B, trong đó A $\subset$ F, B $\subset$ F, và A $\cap$ B = φ. Luật A$\rightarrow$B lưu giữ trong tập giao dịch D với độ hỗ trợ s, trong đó s là phần trăm của các giao dịch trong D có chứa A $\cup$ B, đây là xác suất P(A/B). Luật A$\rightarrow$B có độ tin cậy c trong tập giao dịch D, trong đó c là tỷ lệ phần trăm của các giao dịch trong D chứa A cũng có B. Điều này được coi là xác suất có điều kiện P(B/A), trong đó:

$$\text{Support } (A \rightarrow B) = P(A \cup B)$$

$$\text{Confidence } (A \rightarrow B) = P(B/A) = \frac{P(A,B)}{P(A)}$$

Các luật đáp ứng cả ngưỡng hỗ trợ tối thiểu (min_sup) và ngưỡng tin cậy tối thiểu (min_conf) được gọi là mạnh. Tần suất xảy ra của tập thuộc tính là số lượng các giao dịch chứa tập thuộc tính. Nếu sự hỗ trợ tương đối của một tập thuộc tính F đáp ứng ngưỡng tối thiểu xác định, thì F là tập phổ biến. Tập k-thuộc tính phổ biến ký hiệu bởi L_k. Từ đẳng thức trên, chúng ta có:

$$\text{Confidence } (A \rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)} \quad (1)$$

Đẳng thức cho thấy độ tin cậy của luật A $\rightarrow$ B có thể dễ tính được từ các giá trị hỗ trợ của A và A$\cup$ B. Tức là, khi xác định được các giá trị hỗ trợ của A, B và A $\cup$ B thì sẽ dễ dàng nhận ra các luật kết hợp A $\rightarrow$ B và B $\rightarrow$ A và kiểm tra xem chúng có mạnh hay không. Như vậy, vấn đề của khai phá luật kết hợp có thể được coi là khai phá các tập phổ biến.

Nói chung, khai phá luật kết hợp có thể được xem là một quá trình hai bước [2]:

Bước 1. Tìm tất cả các tập phổ biến từ cơ sở dữ liệu, tức là tìm tất cả các tập D thỏa mãn s(D) $\geq$ min_sup

Bước 2. Sinh ra các luật kết hợp từ các tập phổ biến. Các luật này phải đáp ứng được min_sup và min_conf.

Thuật toán Apriori

Như được trình bày trong [3, 4], thuật toán Apriori được sử dụng để tìm ra tất cả các tập phổ biến

1. Duyệt toàn bộ cơ sở dữ liệu giao dịch để có được độ hỗ trợ S của l-itemset, so sánh S với min_sup, để có được 1-itemset (L_1)

2. Sử dụng L_{k-1} nối (join) L_{k-1} để sinh ra ứng viên k-itemset. Loại bỏ các itemsets không phải là tập phổ biến thu được k-itemset

3. Duyệt toàn bộ cơ sở dữ liệu giao dịch để có được độ hỗ trợ của mỗi ứng viên k-itemset, so sánh S với min_sup để thu được tập phổ biến k-itemset (L_k)

4. Lặp lại từ bước 2 cho đến khi tập ứng viên (C) trống (không tìm thấy tập phổ biến)

5. Với mỗi tập phổ biến I, sinh tất cả các tập con s không rỗng của I

6. Với mỗi tập con s không rỗng của I, sinh ra các luật **s => (I-s)** nếu độ tin cậy (Confidence) của nó > =min_conf

Kỹ thuật bảo vệ tính riêng tư sử dụng Secure Sum

Cho một hệ thống gồm M site, và một đối tượng ký hiệu bởi V. V_i là một ví dụ của Site S_i ($0 \le i$? M). Tính toán $\sum_{i=0}^{M-1} Vi$ theo cách mà các V_i không thể biết được các thông tin của bên khác hoặc các bên cũng không thể biết được thông tin của S_i, trừ khi một số site thông đồng với nhau.

Phương pháp nặc danh được đưa ra trong quy trình Secure Sum [3, 4, 5] và được mô tả trong thuật toán phía dưới. Phương pháp này gọi là "chia sẻ và che dấu" được sử dụng để bảo vệ sự nặc danh của V_i, và cố gắng để giảm chi phí truyền thông.

Procedure Secure Sum()

Given an object *V*. V_i is *V*'s instance at site S_i $(0 \le i < M)$

Caculate securely the sum $\sum_{i=0}^{M-1} V_i$

Input: (1) $\{S_i\}_{0 \le i < M}$: A set of sites, $M \ge 3$

(2) V_i: An instance of *V* at S_i $(0 \le i < M)$

Output: Sucure sum $\sum_{i=0}^{M-1} V_i$

Secure Sum begin

Phare1: share V_i among $M - i$ site

Foreach site S_i $(1 \le i < M)$ **do**

Divide V_i randomly into such $(M - i)$ parts as $\{ V_{i,i}, V_{i,i+1}, \ldots, V_{i,m-1}\}$;

For $j = i+1, i+2, \ldots, M-1$ **do**

Send $V_{i,j}$ to S_j;

Phare2: send the masked share of oneself to S_0

Foreach site S_i $(1 \le i < M)$ **do**

$V_i' \leftarrow V_{i,i} + \sum_{j=1}^{i-1} V_{i,j}$;

Send V_i' to S_0;

For site S_0 **do return** $V_0 + \sum_{i=1}^{M-1} V_i'$

end

PHƯƠNG PHÁP PHÁT HIỆN TẤN CÔNG DỰA TRÊN LUẬT KẾT HỢP CÓ ĐẢM BẢO TÍNH RIÊNG TƯ

Định nghĩa bài toán

Cho N thành viên (P_1 …. P_n), mỗi thành viên có tập dữ liệu tấn công gồm các thuộc tính được trích rút từ gói tin TCP/IP [8]: Flag (rời rạc), serror_rate (liên tục), srv_serror_rate (liên tục), same_srv_rate (liên tục), diff_srv_rate (liên tục), dst_host_srv_count (liên tục), dst_host_same_srv_rate (liên tục), dst_host_diff_srv_rate (liên tục), dst_host_serror_rate (liên tục), Dst_host_srv_serror_rate (liên tục). Thành viên P_i có n_j bản ghi, các thành viên này cần phải kết hợp lại với nhau để tìm ra tấn công trong khi đảm bảo tính riêng tư cho từng tập dữ liệu.

Đề xuất cải tiến giao thức bảo vệ tính riêng tư sử dụng Secure Sum

Tư tưởng của thuật toán được thực hiện trong 2 giai đoạn

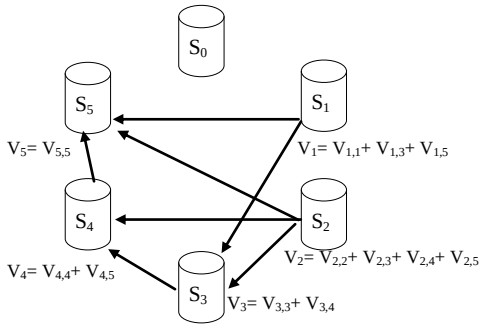
Giai đoạn 1: Các thành viên (site) chia nhỏ ngẫu nhiên tổng của mình thành các $V_{i,j}$, giữ

lại một phần và gửi các phần còn lại cho các site khác trừ những site trước nó

Giai đoạn 2: Các site tính toán tổng những $V_{i,k}$ của các site khác gửi đến và $V_{i,i}$ của mình sau đó gửi cho S_0 để tổng hợp lại

Đề xuất Secure Sum cải tiến (ASecureSum)

Ý tưởng của việc cải tiến tập trung vào giai đoạn 1 của thuật toán. Trong giai đoạn này các site ngẫu nhiên tổng của mình thành các $V_{i,j}$, giữ lại một phần và gửi các phần còn lại cho các site khác trừ những site trước nó. Trong giai đoạn này khác với thuật toán trước, trước khi gửi mỗi site sẽ chọn ngẫu nhiên một số thành viên trong các site còn lại để gửi thay vì gửi cho tất cả các site.. Ví dụ có 6 site, S_0, S_1, S_2, S_3, S_4, S_5



Ví dụ: Giai đoạn 1, giao thức ASecureSum cải tiến có 6 thành viên

Giai đoạn 1:

Trong giai này các Si không gửi các Vi,k của mình cho tất cả các Site sau nó, mà Si sẽ gửi ngẫu nhiên co một số site bất kì, ngẫu nhiên ở đây được chia làm 2 loại:

- Ngẫu nhiên về số lượng: Có nghĩa là số lượng Site mà Si sẽ gửi tới là ngẫu nhiên từ 1 đến M

- Ngẫu nhiên về đối tượng: Có nghĩa là sẽ không biết chắc chắn đối tượng nào sẽ được gửi tới

Giai đoạn 2:

Giai đoạn này thực hiện giống giai đoạn 2 của Secure Sum:

- Các Si tính V`i = Vi,i+$\sum V_{j,k}$ sau đó gửi cho S0

- S0 tính tổng V = V0+ V`1+ V`2 + V`3 + V`4 + V`5

Đánh giá khả năng đảm bảo tính riêng tư của giao thức cải tiến và chi phí truyền thông

- Mức độ đảm bảo tính riêng tư

Trường hợp 1 (đối với Site Si): S0 không gửi bất cứ dữ liệu nào liên quan đến V0 đến bất kỳ site nào khác. Trước khi gửi đến site khác, Vì vậy, V0 không thể được biết đến trừ khi tất cả các site khác thông đồng với nhau.

Trường hợp 2: Giả sử một site Sj nào đó muốn viết Vi của Si (j#i) thì chắc chắn Sj không thể biết Vi,I của Si, vì thế Sj phải thông đồng để biết các giá trị còn lại của Vi, nhưng vì Si gửi ngẫu nhiên nên Sj không thể biết chính xác những đối tượng nào và bao nhiêu đối tượng để xác định cần phải thông đồng, vì thế muốn biết chắc chắn thì Sj phải thông đồng với tất cả các site trừ Si.

Bằng việc này chúng ta đã chỉ ra rằng Sj không thể biết Vi, hoặc đoán biết Vi trừ khi nó thông đồng với tất cả site khác, vậy mức độ đảm bảo tính riêng tư ở đây vẫn được giữ nguyên là M-2

- Chi phí truyền thông:

Hệ thống gồm M site, T là thời gian trung bình để gửi một thông điệp từ site này đến site khác ta có.

Trường hợp xấu nhất, các site vẫn gửi đầy đủ thông điệp cho các site còn lại sau nó (theo Secure Sum) thì số thông điệp là M(M-1)/2

Trường hợp tốt nhất : Mỗi site chỉ gửi đến cho một site khác thì số thông điệp sẽ là M-2 (giai đoạn 1) +M-1(Giai đoạn 2) =2M-3 thông điệp

Số thông điệp sẽ nằm trong khoảng (2M-3 đến M(M-1)/2)

Từ chứng minh trên ta thấy rằng đối với tiến trình thực hiện Secure Sum cải tiến có chi phí truyền thông thấp hơn so với Secure Sum ban đầu

Vậy với những phân tích và chứng minh trên có thể thấy việc cải tiến Secure Sum của bài báo có mức đảm bảo tính riêng tư tốt trong

khi vẫn giữ được và có chi phí truyền thông thấp hơp thuật toán Secure Sum ban đầu.

Khai phá tập phổ biến có đảm bảo tính riêng tư dựa trên giao thức Secure Sum cải tiến

Bước quan trọng để tìm ra các luật kết hợp là tìm ra các tập phổ biến vì vậy tác giả trình bày giao thức tính tập phổ biến có đảm bảo tính riêng tư.

Input: Mỗi thành viên P1, P2, ….., Pn có các tập dữ liệu D1, D2, ….., Dn

Ouput: Các tập phổ biến của tập dữ liệu D = D1 ∪ D2 ∪ ….. ∪ Dn

1. Xác định tập phổ biến 1- Itemset (L1)

- Mỗi thành viên duyệt CSDL Di để tính support supi của tập 1- Itemset (L1)

- Các thành viên tham gia để thực hiện giao thức S=ASecureSum($\sum_{i=1}^{n} sup_i$)

- Mỗi thành viên so sánh: If S>= min_sup đưa vào tập 1-Itemset (L1) else loại bỏ

2. Mỗi thành viên sử dụng Lk-1 nối (join) Lk-1 để sinh ra tập candidate k-itemset (C), loại bỏ các itemsets không phải là tập phổ biến thu được k-itemset

3. Xác định tập phổ biến k –itemset (Lk)

- Mỗi thành viên duyệt CSDL Di để tính support supi của tập k- Itemset (Lk)

- Các thành viên tham gia để thực hiện giao thức S=ASecureSum($\sum_{i=1}^{n} sup_i$)

- Mỗi thành viên so sánh: If S>= min_sup đưa vào tập k- Itemset (Lk) else loại bỏ

4. Lặp lại từ bước 2 cho đến khi C trống (không tìm thấy tập phổ biến nào khác).

ĐÁNH GIÁ HIỆU QUẢ, THỬ NGHIỆM TẬP DỮ LIỆU KDD99

Để so sánh hiệu quả (thời gian thực thi) giữa giao thức khai phá luật kết hợp có đảm bảo tính riêng tư dựa trên Asecuresum với giao thức dựa trên Secure Sum, bài báo sử dụng tập dữ liệu KDD Cup 99 [6, 7] được tạo ra bằng cách xử lý phần dữ liệu TCPDUMP lấy được trong 7 tuần từ hệ thống phát hiện xâm nhập DARPA 1998 bởi MIT Lincoln Labs. Trong tập dữ liệu KDD Cup 1999 ta trích chọn 10% trong số dữ liệu này để làm thực nghiệm, bao gồm 91060 bản ghi. Chia tập KDD Cup 99 rút gọn thành 20 phần, mỗi phần 4553 bản ghi. Thực hiện quá trình tính toán mô phỏng trên phần mềm NS2, mỗi nút mạng là một thành viên, trên môi trường hệ điều hành Windows 10 64bit, máy tính 20 Core (mỗi core tương ứng với một nút mạng) tốc độ 2.3GHz, độ hỗ trợ = 40%, độ tin cậy = 70%.

Kết quả khi thực nghiệm trên hai giao thức ASecuresum và Secure Sum là giống nhau, đã tìm ra 67187 luật. Thời gian thực hiện khi số lượng các thành viên thay đổi từ 1 đến 20 như trong bảng 1.

Bảng 1. So sánh hiệu quả về thời gian của giao thức Secure Sum và giao thức ASecuresum

Số lượng thành viên		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Thời gian (s)	**Secure Sum**	5,05	5,37	5,103	4,848	4,606	4,376	4,157	3,949	3,752	3,564
	ASecuresum	5,05	5,35	5,059	4,781	4,518	4,27	4,035	3,813	3,603	3,405
Số lượng thành viên		**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Thời gian (s)	**Secure Sum**	3,386	3,217	3,056	2,903	2,758	2,62	2,489	2,365	2,247	2,135
	ASecuresum	3,218	3,041	2,874	2,716	2,567	2,426	2,293	2,167	2,048	1,935

KẾT LUẬN

Bài báo đã nghiên cứu và cải tiến giao thức tính tổng bảo mật nhiều thành viên hiệu quả nhất hiện nay (Seucresum) và kết quả cải tiến được áp dụng trong bài toán khai phá luật phát hiện tấn công có đảm bảo tính riêng tư. Thực nghiệm chỉ ra cho thấy thuật toán cải tiến (ASeucresum) có độ chính xác không thay đổi, nhưng cải tiến được thời gian đáng kể so với thuật toán cũ (Seucresum).

TÀI LIỆU THAM KHẢO

1. M. D. Chachkamy, B.Sadeghiyan (2013), "Privacy Preserving Association Rule Mining in Collaborative Intrusion Detection Systems with Fuzzy Data", *International Journal of Information and Communication Technology Research*, Volume 3 No. 9, pp. 272 – 276.

2. V.Ragunath, C.R.Dhivya (2014), "Privacy Preserved Association Rule Mining For Attack Detection and Prevention", *International Journal of Innovative Research In Computer and Communication Engineering*, Vol.2, pp. 3650 -3654.

3. C. Clifton, M. Kantarcioglu, J. Vaidya, X.Lin, and M.Y.Zhu (2002), "Tools for privacy preserving distributed data mining" *SIGKDD Explor. Newsl*, Volume 4(2) pp. 28–34.

4. R. Sheikh, B. Kumar (2009), "Privacy-Preserving k-Secure Sum Protocol" *International Journal of Computer Science and Information Security*, Vol. 6, No. 2, pp. 184-188.

5. Charu C. Aggarwal, Philip S. Yu (2008), *Privacy Preserving Data Mining Models and Algorithms*, Springer Science + Business Media, LLC.

6. S. Hettich, S.D. Bay (1999), *The UCI KDD Archive*, University of California, USA.

7. Preeti Aggarwal, Sudhir Kumar Sharma (2015), "Analysis of KDD Dataset Attributes-Class wise For Intrusion Detection", *3rd International Conference on Recent Trends in Computing*, pp. 842 – 851.

8. H. Güneş Kayacık, A. Nur Zincir-Heywood, M. I. Heywood (2005), *Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets*, Dalhousie University, Faculty of Computer Science, Nova Scotia.

9. M.R.A. Huth (2002), *Secure Communicating Systems*, CAMBRIDGE UniversityPress.