JTE HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Data Privacy Using Anonymization Method on Open Data

Thi Minh Chau Le*, Tran Thi Van Nguyen
Ho Chi Minh City University of Technology and Education, Vietnam

*Corresponding author. *Email: chaultm@hcmute.edu.vn*

ARTICLE INFO

Received: 27/09/2023
Revised: 08/10/2023
Accepted: 10/10/2023
Published: 28/04/2024

KEYWORDS

Open Data;
Anonymity;
Privacy;
k – anonymity;
ℓ - diversity.

ABSTRACT

Open Data is a type of data shared between organizations, agencies, businesses, governments, etc. It is mainly used to serve community projects in many fields: health, environment, education, etc. Nowadays, countries around the world are following the trend of building smart cities and smart governments. They are applying Open Data in these projects and achieving many significant benefits. However, sharing data can lead to many problems. In recent studies, many authors have pointed out that besides the benefits that Open Data offers, there are also risks in terms of security, including revealing information of individuals, organizations, and businesses. Data security using anonymization methods such as k-anonymity or l-diversity has been researched and applied for many years. However, these methods are just mainly implemented and tested on traditional data sets of businesses and organizations, not the data on Open Data. Therefore, this topic will focus on understanding Open Data, data security methods based on anonymization mechanism, implementing some security methods based on anonymization mechanism on Open Data and analyzing and evaluating research results.

Bảo Mật Dữ Liệu theo Phương Pháp Nặc Danh Hóa trên Open Data

Lê Thị Minh Châu*, Nguyễn Trần Thi Văn
Trường Đại học Sư Phạm Kỹ Thuật TP. HCM, Việt Nam

*Tác giả liên hệ. *Email: chaultm@hcmute.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 27/09/2023
Ngày hoàn thiện: 08/10/2023
Ngày chấp nhận đăng: 10/10/2023
Ngày đăng: 28/04/2024

TỪ KHÓA

Dữ liệu mở;
Nặc danh hóa;
Tính riêng tư;
k – nặc danh;
ℓ - đa dạng.

TÓM TẮT

Open Data (Dữ Liệu Mở) là loại dữ liệu được chia sẻ giữa các tổ chức, cơ quan, doanh nghiệp, chính phủ… Mục đích chung là để phục vụ cho các dự án cộng đồng trong nhiều lĩnh vực: sức khỏe, môi trường, giáo dục…. Hiện nay, các nước trên thế giới đang theo xu hướng xây dựng thành phố thông minh, chính phủ thông minh, ứng dụng Open Data trong các dự án này và đạt được nhiều lợi ích đáng kể. Tuy nhiên, việc chia sẻ dùng chung dữ liệu có thể dẫn đến nhiều vấn đề. Trong những nghiên cứu gần đây các tác giả đã chỉ ra bên cạnh các lợi ích mà Open Data đem lại cũng tồn tại những rủi ro về tính bảo mật, làm lộ thông tin của các cá nhân, tổ chức, doanh nghiệp. Bảo mật dữ liệu bằng các phương pháp nặc danh hóa như: k-anonymity, l-diversity... đã được nghiên cứu và áp dụng trong nhiều năm nhưng đa phần vẫn trên tập dữ liệu truyền thống của các doanh nghiệp, tổ chức, chưa áp dụng lên các dữ liệu của hệ thống Dữ liệu mở. Do đó, đề tài này sẽ tập trung tìm hiểu Open Data, các phương pháp bảo mật dữ liệu theo cơ chế nặc danh, hiện thực một số phương pháp bảo mật theo cơ chế nặc danh hóa trên Open Data và phân tích, đánh giá kết quả nghiên cứu được.

Doi: https://doi.org/10.54644/jte.2024.1472

1. Giới thiệu

Open Data (Dữ Liệu Mở) là loại dữ liệu được chia sẻ giữa các tổ chức, cơ quan, doanh nghiệp, chính phủ… Bất kỳ ai cũng có thể sử dụng Open Data một cách rộng rãi mà không cần xin bản quyền hay giấy phép sử dụng. Mục đích chung là để phục vụ cho các dự án cộng đồng trong nhiều lĩnh vực: sức

khỏe, môi trường, giáo dục…. Hiện nay, các nước trên thế giới đang theo xu hướng xây dựng thành phố thông minh, chính phủ thông minh, ứng dụng Open Data trong các dự án này và đạt được nhiều lợi ích đáng kể.

Tuy nhiên, việc chia sẻ dùng chung dữ liệu có thể dẫn đến nhiều vấn đề về tính bảo mật, làm lộ thông tin của các tổ chức, doanh nghiệp chia sẻ dữ liệu. Do đó, đã có nhiều nghiên cứu về vấn đề bảo mật cho Open Data.

1.1. Tình hình nghiên cứu trong nước

Gần đây trong nước cũng đã có các nghiên cứu về Open Data nhưng đa phần đều chỉ ở mức tìm hiểu, xây dựng và cung cấp các giải pháp cho các hệ thống Open Data. Ví dụ: Cổng Dữ Liệu Mở (Open Data Portal) của các công ty BKAV, VNPT…; cổng dữ liệu quốc gia của Chính Phủ và các tỉnh, thành phố. Có một vài nhóm nghiên cứu về các vấn đề bảo mật trong Open Data nhưng ở góc độ nào đó chỉ đáp ứng được việc bảo mật dữ liệu ở mức cơ bản.

Bảo mật dữ liệu bằng các phương pháp nặc danh hóa như: k-anonymity, l-diversity, t-closeness đã được nghiên cứu và áp dụng trong nhiều năm nhưng đa phần vẫn trên tập dữ liệu truyền thống của các doanh nghiệp, tổ chức, chưa áp dụng lên các dữ liệu của hệ thống Open Data.

Năm 2020 chỉ có đề tài NCKH cấp Thành Phố về Kỹ thuật ẩn danh bảo vệ tính riêng tư cho dữ liệu mở. Đề tài này được chủ trì bởi các tác giả Trường Đại học Bách Khoa TP. HCM thực hiện và đã được nghiệm thu bởi Sở Khoa học và Công nghệ TP.HCM. Nhóm tác giả đã nghiên cứu về kiến trúc Open Data, các framework, thuật toán và các kỹ thuật nặc danh hóa dữ liệu để bảo vệ tính riêng tư như: ARX, SECRETA, PSI, RAPPOR, Amnesia, k-anonymity, l-diversity, δ-presence… ập trung cho dữ liệu mở trong Smart City. Nhóm đã đề xuất framework, tích hợp thử nghiệm các giải pháp trên các tập dữ liệu khác nhau và đánh giá kết quả đạt được [https://dost.hochiminhcity.gov.vn/tiem-luc/ket-qua-nckh/ky-thuat-an-danh-bao-ve-tinh-rieng-tu-cho-du-lieu-mo/].

1.2. Tình hình nghiên cứu nước ngoài

Việc xây dựng các hệ thống Open Data đã được chính phủ nhiều nước và các tổ chức, doanh nghiệp áp dụng từ nhiều năm nay. Ngoài ra, trong những nghiên cứu gần đây các tác giả đã chỉ ra bên cạnh các lợi ích mà Open Data đem lại cũng tồn tại những rủi ro, mâu thuẫn về tính minh bạch, quyền riêng tư, bảo mật và độ tin cậy của Open Data [1], [2]. Để khắc phục những hạn chế này, các tác giả đã đưa ra một số giải pháp về kỹ thuật cũng như một số chính sách cho Open Data.

Trong [3], các tác giả đã chỉ ra việc xuất bản dữ liệu (data publishing) để phục vụ cho việc phân tích sẽ đánh đổi quyền riêng tư cá nhân với chất lượng kết quả đầu ra. Các tác giả cũng giới thiệu một thuật toán công bố dữ liệu đáp ứng mô hình bảo vệ tính riêng tư khác nhau. Các phép biến đổi được thực hiện một cách trung thực, không làm xáo trộn dữ liệu đầu vào hoặc tạo ra dữ liệu đầu ra tổng hợp. Thay vào đó, các bảng dữ liệu được rút ra một cách ngẫu nhiên từ tập dữ liệu đầu vào và tính chất duy nhất của các tính năng của dữ liệu sẽ bị thu giảm. Điều này cũng cung cấp một khái niệm trực quan về bảo vệ quyền riêng tư. Hơn nữa, cách tiếp cận này mang tính tổng quát vì có thể được tham số hóa bằng các hàm mục tiêu khác nhau để tối ưu hóa kết quả đầu ra cho các ứng dụng khác nhau. Các tác giả đã tích hợp 06 mô hình chất lượng dữ liệu nổi tiếng để thực hiện thử nghiệm, đánh giá, phân tích và cho ra kết quả dự đoán với độ chính xác cao.

Trong [6], tác giả đề xuất mô hình k – anonymity (k – nặc danh) để bảo vệ tính riêng tư của dữ liệu. Theo mô hình này, một tập dữ liệu thỏa tính chất k – anonymity nếu thông tin của mỗi cá nhân trong đó không thể phân biệt được với thông tin của ít nhất k – 1 cá nhân khác trong tập dữ liệu đó.

Tác giả đã đưa ra các định nghĩa, xây dựng mô hình k – anonymity và xác định các rủi ro về bảo mật và các loại tấn công trên mô hình này.

Trong bài báo này, nhóm tác giả sẽ tập trung tìm hiểu Open Data, các phương pháp bảo mật dữ liệu theo cơ chế nặc danh, hiện thực một số phương pháp bảo mật theo cơ chế nặc danh hóa trên Open Data và viết báo cáo phân tích, đánh giá kết quả nghiên cứu được.

2. Dữ liệu mở và một số phương pháp nặc danh hóa dữ liệu

2.1. Open Data

Theo định nghĩa của Quỹ Kiến Thức Mở OKF (Open Knowledge Foundation) [4], Open Data có thể được sử dụng, chỉnh sửa và chia sẻ bởi bất kỳ ai với bất kỳ mục đích nào.

Một số quy định cho Open Data:

- Dữ liệu phải luôn có sẵn và được truy cập miễn phí, ngoại trừ phí vận hành hệ thống.
- Dữ liệu phải tồn tại ở dạng có thể chỉnh sửa và sử dụng một cách thuận tiện.
- Dữ liệu phải được cung cấp theo giấy phép cho phép tái sử dụng, phân phối lại và kết hợp với các bộ dữ liệu khác.
- Mọi người đều có quyền sử dụng, tái sử dụng và phân phối dữ liệu dưới mọi hình thức cho bất kỳ mục đích nào.

Trong [5], mục tiêu liên quan đến Open Data được chia thành 03 loại:

- Đổi mới và tăng trưởng kinh tế: Các dịch vụ thông tin được xây dựng trên dữ liệu hành chính công rất đa dạng: các nhà cung cấp dịch vụ tài chính sử dụng số liệu công bố chính thức của chính phủ để làm dữ liệu đầu vào, các công ty khí tượng sử dụng dữ liệu thời tiết để dự báo cho các ngành công nghiệp dầu mỏ, dữ liệu quy hoạch nhà ở được kết hợp với các nguồn khác để đưa ra lời khuyên cho các khách hàng như các nhà phát triển bất động sản…
- Trách nhiệm chính trị và sự tham gia dân chủ: việc công bố dữ liệu mở của các tổ chức chính phủ sẽ giúp làm tăng trách nhiệm giải trình và thúc đẩy sự tham gia có hiểu biết của công chúng vào các hoạt động của chính phủ, góp phần xây dựng chính phủ minh bạch.
- Hiệu quả của chính sách công: dữ liệu mở sẽ giúp tiết kiệm tài nguyên và cải thiện các dịch vụ công. Ví dụ, Ủy ban Châu Âu cho biết dữ liệu mở sẽ cải thiện các dịch vụ y tế và quản lý giao thông, đồng thời giúp giải quyết các thách thức môi trường, chẳng hạn như thông qua giám sát mức tiêu thụ năng lượng. Ở cấp quốc gia, một chiến lược ngày càng phổ biến là công bố dữ liệu hiệu quả hoạt động của các tổ chức được nhà nước tài trợ. Việc tiết lộ thông tin kiểm tra và các dữ liệu khác được cho là sẽ cải thiện hiệu suất của những người nhận tiền thuế, như trường học (điểm kiểm tra) và bệnh viện (tử vong, thời gian chờ đợi). Công dân với tư cách là khách hàng được cho là sẽ đưa ra những lựa chọn sáng suốt hơn khi được cung cấp dữ liệu hiệu suất đó.

Các mối lo ngại về việc phát hành thông tin cá nhân dưới dạng dữ liệu mở:

- Tác động đáng lo ngại với những người sử dụng dịch vụ chính sách công: mọi người cung cấp dữ liệu cho các cơ quan công quyền sợ rằng thông tin của họ sẽ được lưu trữ hoặc sẽ bị công khai.Vì vậy, mọi người có thể hạn chế liên hệ với khu vực công nếu họ lo ngại thông tin cá nhân của mình sẽ không được giữ bí mật.
- Thiếu kiểm soát đối với các thông tin cá nhân: Việc phát hành công khai thông tin cá nhân dưới dạng dữ liệu mở có thể đặc biệt rắc rối vì chính sách dữ liệu mở ở dạng tự do nhất ngụ ý rằng số lượng người dùng lại không giới hạn có thể sử dụng dữ liệu cho bất kỳ mục đích nào.
- Sử dụng dữ liệu mở cho sự phân loại xã hội hoặc các hành vi phân biệt đối xử: nếu chính phủ công bố dữ liệu cá nhân, các nhà môi giới dữ liệu có thể sẽ nằm trong số những người tái sử dụng lại dữ liệu. Nhà môi giới dữ liệu là các công ty thu thập thông tin cá nhân của người tiêu dùng và bán lại hoặc chia sẻ thông tin đó với người khác.” Ví dụ: thông tin có thể được sử dụng để tiếp thị trực tiếp, chấm điểm tín dụng hoặc sàng lọc người xin việc. Thông tin cá nhân cũng có thể được sử dụng để phân biệt đối xử không công bằng. Ví dụ, một công ty có thể sử dụng thông tin về “Người hút thuốc trong hộ gia đình” để kết luận rằng những người trong hộ gia đình đó không nên được cung cấp bảo hiểm.

Một số hệ thống quản trị dữ liệu mở phổ biến hiện nay gồn có: CKAN (được sử dụng bởi các hệ thống chính phủ Anh, Hà Lan, Mỹ…), DSPACE (được sử dụng bởi Ngân hàng thế giới, Đại học Cambirdge, Đại học MIT…).

2.2. Mô hình k – anonymity

Bảng 1. *Dữ liệu y tế của bệnh nhân [7]*

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Bảng 2. *Dữ liệu bệnh nhân đã được nặc danh hóa với k = 4 [7]*

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130*	< 30	*	Heart Disease
3	130*	< 30	*	Viral Infection
4	130*	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Theo nghiên cứu trong [6], [7], nhiều tổ chức đang tăng cường xuất bản dữ liệu vi mô (microdata) về thông tin chưa tổng hợp về các cá nhân. Các bảng này có thể chứa dữ liệu y tế, đăng ký cử tri, điều tra dân số và khách hàng. Microdata là nguồn thông tin có giá trị cho việc phân bổ công quỹ, nghiên cứu y tế và phân tích xu hướng kinh doanh. Tuy nhiên, nếu các cá nhân có thể được xác định duy nhất trong dữ liệu vi mô thì thông tin của họ (chẳng hạn như tình trạng sức khỏe của họ) sẽ bị tiết lộ, điều này không được phép xảy ra.

Để tránh việc xác định các bảng ghi trong dữ liệu vi mô, thông tin nhận dạng duy nhất các cá thể như tên và số an sinh xã hội sẽ được xóa khỏi bảng dữ liệu. Tuy nhiên, việc tinh chỉnh này vẫn không đảm

bảo được tính riên tư của các cá nhân trong dữ liệu. Một nghiên cứu gần đây ước tính rằng 87% dân số Mỹ có thể được xác định bằng cách sử dụng các thuộc tính dường như vô hại như giới tính, ngày sinh và mã ZIP gồm 5 chữ số. Trên thực tế, 03 thuộc tính này được dùng để liên kết hồ sơ đăng ký cử tri của Massachusetts (bao gồm tên, giới tính, mã zip và ngày sinh) với dữ liệu y tế được cho là ẩn danh từ GIC1 (bao gồm giới tính, mã ZIP, ngày sinh và chuẩn đoán bệnh). "Cuộc tấn công liên kết" này đã xác định được hồ sơ y tế của thống đốc bang Massachussets trong dữ liệu y tế.

Tập các thuộc tính (như giới tính, ngày sinh và zip code) có thể được liên kết với dữ liệu bên ngoài để nhận dạng duy nhất các cá nhân trong một tập hợp dữ liệu. Tập các thuốc tính này được gọi là quasi – identifier.

Cho RT(A1,...,An) là một bảng dữ liệu và QIRT là thuộc tính quasi-identifier tương ứng. RT được gọi là thỏa mãn tính chất k – anonymity nếu và chỉ nếu mỗi chuỗi giá trị trong RT[QI_{RT}] xuất hiện với ít nhất k lần trong RT[QI_{RT}].

2.3. Mô hình ℓ - diversity

Trong [7], tác giả đã chứng minh mô hình k – anonymity không đảm bảo chống lại 02 loại tấn công sau:

- **Tấn công đồng nhất**

Giả sử Alice và Bob là những hàng xóm đối nghịch nhau. Bob bị bệnh nhập viện. Alice phát hiện rằng hồ sơ bệnh nhân nội trú hiện tại do bệnh viện công bố (Hình 6) có chứa dữ liệu của Bob. Vì Alice là hàng xóm của Bob nên biết được Bob là một nam giới người Mỹ 31 tuổi sống trong vùng có mã ZIP 13053. Do đó, Alice biết số thứ tự của Bob là 9 hoặc 10, 11, 12. Tất cả bệnh nhân có thứ tự này đều bị bệnh ung thư nên Alice có thể kết luận Bob cũng bị ung thư.

⇨ k – anonymity có thể tạo ra các nhóm bị rò rỉ thông tin do thiếu sự đa dạng trong thuộc tính nhạy cảm.

- **Tấn công dựa trên tri thức nền tảng**

Alice có một người bạn tên là Umeko nhập viện cùng bệnh viện với Bob và hồ sơ bệnh của Umeko cũng xuất hiện trong bảng dữ liệu được bệnh viên công bố (Hình 6). Alice biết rằng Umeko là một phụ nữ Nhật 21 tuổi sống tại khu vực có mã ZIP 13086. Dựa vào thông tin này, Alice biết được thông tin về Umeko nằm trong dòng 1 hoặc 2, 3, 4 của bảng dữ liệu. Nếu không có thêm thông tin gì thì Alice không chắc là liệu Umeko có bị nhiễm virus hay mắc bệnh tim hay không? Tuy nhiên, ai cũng biết rằng người Nhật có tỷ lệ mắc bệnh tim cực kỳ thấp. Vì vậy, Alice kết luận gần như chắc chắn rằng Umeko bị nhiễm virus.

⇨ k – anonymity không chống lại được các cuộc tấn công dựa trên tri thức nền tảng

- **Bayes – Optimal Privacy**

Quyền riêng tư tối ưu Bayes liên quan đến việc lập mô hình kiến thức nền tảng dưới dạng phân bố xác suất trên các thuộc tính và sử dụng kỹ thuật suy luận Bayes để suy luận về quyền riêng tư.

Đầu tiên, ta giả sử rằng T là một mẫu dữ liệu lấy ngẫu nhiên từ tập dân số Ω. Thứ hai giả định rằng chỉ có một thuộc tính nhạy cảm duy nhất. Trong kịch bản tấn công, Alice có một phần kiến thức về sự phân bố các thuộc tính nhạy cảm và không nhạy cảm. Chúng ta giả định trường hợp xấu nhất Alice biết phân bố chung đầy đủ f của Q và S (tần suất xuất hiện của chúng trong tập dân số Ω). Cô ấy biết rằng Bob tương ứng với dòng t ∈ T đã được tổng quát hóa thành dòng t* trong bảng T* và cô ấy cũng biết giá trị các thuộc tính không nhạy cảm của Bob t[Q] = q. Mục tiêu của Alice là dùng kiến thức nền tảng của mình để khám phá thông tin nhạy cảm của Bob t[S].

Niềm tin trước đó của Alice, α_((q,s)), rằng thuộc tính nhạy cảm của Bob là s và thuộc tính không nhạy cảm của Bob là q, chỉ là kiến thức nền tảng của Alice:

$$\alpha_{(q,s)} = P_f(t[S] = s | t[Q] = q) \quad (1)$$

Sau khi Alice xem được bảng T*, niềm tin của cô ấy về thuộc tính nhạy cảm của Bob thay đổi. Niềm tin sau này của Alice là β(q, s, T*):

$$\beta_{(q,s,T*)} = P_f\left(t[S] = s \mid t[Q] = q \wedge \exists t^* \in T*, t \overset{*}{\rightarrow} t^*\right) \quad (2)$$

Cho q là giá trị của thuộc tính không nhạy cảm Q trong bảng cơ sở T; q* là giá trị tổng quát hóa của q trong bảng được xuất bảng T*; s là giá trị có thể của thuộc tính nhạy cảm ; n_((q^*,s^')) là số bộ t* € T* trong đó t*[Q] = q* và t*[S] = s^'; và cho f(s ′ | q*) là xác suất có điều kiện của thuộc tính nhạy cảm dựa trên thực tế là thuộc tính không nhạy cảm Q có thể được khái quát hóa thành q*. Khi đó, niềm tin của Alice được giữ nguyên:

$$\beta_{(q,s,T*)} = \frac{n_{(q*,s)}\frac{f(s|q)}{f(s|q*)}}{\sum_{s' \in S} n_{(q*,s')}\frac{f(s'|q)}{f(s'|q*)}} \quad (3)$$

Tiết lộ tích cực (Possitive Disclosure)

Việc xuất bản bảng T* được lấy từ bảng T mang lại kết quả tiết lộ tích cực nếu kẻ tấn công có thể xác định chính xác giá trị của thuộc tính nhạy cảm với xác suất cao; tức là với δ > 0, sẽ có sự tiết lộ tích cực nếu β(q, s, T*) > 1 – δ và tồn tại t € T sao cho t[Q] = q và t[S] = s.

Tiết lộ tiêu cực (Negative Disclosure)

Việc xuất bản bảng T* được lấy từ T dẫn đến tiết lộ tiêu cực nếu kẻ tấn công có thể loại bỏ chính xác một số giá trị có thể có của thuộc tính nhạy cảm (với xác suất cao); tức là cho trước ϵ > 0 sẽ có tiết lộ tiêu cực nếu β(q, s, T*) < ϵ và tồn tại t € T sao cho t[Q] = q nhưng t[S] ≠ s.

Cuộc tấn công vào tính đồng nhất trong đó Alice xác định rằng Bob mắc bệnh ung thư là một ví dụ về tiết lộ tích cực. Tương tự, ngay cả khi không có kiến thức nền tảng, Alice vẫn có thể suy luận rằng Umeko không bị ung thư. Đây là một ví dụ về tiết lộ tiêu cực.

Tiết lộ tích cực không phải lúc nào cũng gây tổn hại lớn. Nếu niềm tin trước đó là α_((q,s))>1-δ thì kẻ tấn công sẽ không học được điều gì mới. Tương tự, những tiết lộ tiêu cực không phải lúc nào cũng xấu: việc phát hiện Bob không mắc bệnh Ebola có thể không nghiêm trọng lắm vì niềm tin trước đó về sự thật này là rất nhỏ.

Hạn chế của Bayes Optimal Privacy

– Không đủ kiến thức: nhà xuất bản dữ liệu khó có thể biết được sự phân bố đầy đủ f của các thuộc tính nhạy cảm và không nhạy cảm trên tập dân số Ω.

– Tri thức của kẻ tấn công là không xác định: kẻ tấn công cũng có thể khó kiến thức về sự phân bố chung đầy đủ giữa các thuộc tính nhạy cảm và không nhạy cảm. Tuy nhiên, nhà xuất bản dữ liệu không biết được kẻ tấn công biết được bao nhiêu. Ví dụ, trong cuộc tấn công dựa vào kiến thức nền tảng, nhà xuất bản cũng không biết được thông tin Alice biết rằng người Nhật có tỷ lệ bệnh tim thấp.

– Ngoài ra, kẻ tấn công có thể có được tri thức bằng cách khác hoặc có thể kết hợp nhiều kiến thức nền tảng để suy luận. Ví dụ: con trai của Bob nói cho Alice biết Bob không bị tiểu đường hoặc Bob mắc căn bệnh (a) rất dễ xảy ra với những người ở độ tuổi [30 – 50] nhưng (b) rất hiếm gặp với những người trong độ tuổi đó nếu là bác sĩ. Alice biết Bob là bác sĩ. Do đó, Bayes – Optimal Privacy không đủ đảm bảo chống lại các tấn công này.

ℓ - diversity

Một khối q* - block gọi là ℓ - diversity nếu chứa ít nhất ℓ giá trị "được thể hiện tốt" cho các thuộc tính nhạy cảm S. Một bảng dữ liệu gọi là ℓ - diversity nếu mỗi q* - block trong bảng đều là ℓ - diversity.

Entropy ℓ - diversity

Một bảng dữ liệu là Entroy ℓ - diversity nếu với mỗi q* - block

$$\sum_{s \in S} p_{(q*,s)} log\left(p_{(q*,s')}\right) \geq log(\ell) \quad (4)$$

với $p_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s' \in S} n_{(q*,s')}}$ là phần trăm các dòng dữ liệu trong q*-block có thuộc tính nhạy cảm là s.

- **Recursive (c, ℓ) – diversity**

Trong một khối q* - block nhất định, gọi ri là số lần thuộc tính nhạy cảm thường xuyên thứ i xuất hiện trong khối q* - block đó. Cho một hằng số c, khối q* - block thỏa mãn recursive (c, ℓ) nếu r1 < c (r ℓ + r ℓ+1 + … + rm). Một bảng T* thỏa mãn recursive (c, ℓ) nếu mỗi q* - block thỏa mãn recursive ℓ - diversity.

3. Giải pháp đề tài

Phần này sẽ mô tả về việc xây dựng chương trình thực nghiệm và giải pháp nặc danh hóa Dữ liệu mở, cách tiếp cận để tìm lời giải cho vấn đề nghiên cứu.

Có 02 mô hình chuyển đổi dữ liệu nặc danh: multidimentional generation và local generation. Chương trình thực nghiệm áp dụng giải pháp k – anonymity cho các tập dữ liệu mở được lấy từ nhiều nguồn khác nhau theo phương pháp local generation với mục đích tối ưu hóa về mặt thời gian.

Việc đánh giá chất lượng dữ liệu sau khi được nặc danh hóa thường được thực hiện thông qua các độ đo:

- **Cell – oriented general – purpose model**
  - **Granularity / loss [12]**: độ đo này tóm tắt mức độ mà các giá trị thuộc tính được chuyển đổi bao trùm miền trị ban đầu của thuộc tính.
  - **Precision [13]**: mô hình này ước tính chất lượng dữ liệu dựa trên mức độ tổng quát hóa được chuẩn hóa của các giá trị thuộc tính được chuyển đổi.
- **Attribute – oriented general – purpose model**
  - **Non – uniform entroy [14]**: mô hình này định lượng việc mất thông tin dựa trên thông tin tương hỗ (mutual information), đo lường lượng thông tin có thể được thu được về giá trị gốc của các biến trong tập dữ liệu đầu vào bằng các quan sát các giá trị của các biến trong tập dữ liệu đầu ra.
  - **Height**: mô hình này định lượng việc mất thông tin bằng tổng các mức độ khái quát hóa được áp dụng cho tất cả các giá trị thuộc tính.
- **Record – oriented general – purpose model**
  - **Average equivalence class size [15]**: mô hình này ước tính chất lượng dữ liệu bằng cách tính toán kích thước trung bình của các lớp bản ghi mà không thể phân biệt được, không tính đến các giá trị thuộc tính thực tế trong tập dữ liệu đầu ra.
  - **Discernibility [16], [17]**: mô hình này cũng ước tính chất lượng dữ liệu dựa trên kích thước của các lớp tương đương trong tập dữ liệu đầu ra. Những bản ghi bị thu giảm (suppressed) sẽ bị hạn chế, Mô hình không tính đến các giá trị thuộc tính thực tế trong tập dữ liệu đầu ra.
  - **Ambiguity [18]**: mô hình này định lượng mức độ mơ hồ của các bản ghi trong tập dữ liệu đầu ra.
  - **Entropy – based model [19]**

Mô hình k – anonymity đã được chứng minh bảo mật tốt dữ liệu. Do đó, chương trình thực nghiệm sẽ đánh giá chất lượng dữ liệu sau khi áp dụng phương pháp nặc danh hóa bằng cách áp dụng độ đo Granularity / loss. Mục tiêu là tập dữ liệu sau khi chuyển đổi đạt tính chất k – nặc danh càng nhiều càng tốt, kẻ tấn công sẽ khó xác định chính xác các cá nhân nếu chỉ dựa vào các thuộc tính QI.

4. Kết quả thực nghiệm

Phần này trình bày các kết quả thu được từ nghiên cứu.

4.1. Ngôn ngữ lập trình và công nghệ sử dụng

Chương trình thực nghiệm được tham khảo từ [4] và được viết bằng ngôn ngữ Java.

Việc xây dựng và dịch chương trình được thực hiện dựa trên công cụ Apache Ant.

Mô hình nặc danh hóa dữ liệu k – anonymity với k = 3 và k = 5 (phổ biến) được triển khai sử dụng thư viện ARX.

Chương trình được thực nghiệm trên môi trường máy tính cá nhân với cấu hình CPU Intel R Core TM i7 2.2 Ghz và 32 GB RAM.

4.2. Các tập dữ liệu

Chương trình thực nghiệm áp dụng trên 04 tập dữ liệu sau:

1. Thông tin thu nhập của các cá nhân được lấy từ UCI Machine Learning Repository [8] (adult.csv): 09 thuộc tính và 30,162 dòng dữ liệu.
2. Bảng khảo sát Cộng đồng Hoa Kỳ [9] (ss13acs_int.csv): 30 thuộc tính và 68,725 dòng dữ liệu.
3. Kết quả kết quả khảo sát sức khoẻ của người dân Hoa Kỳ [10] (ihis_int.csv): 09 thuộc tính và 1,193,504 dòng dữ liệu.
4. Thông tin sử dụng thời gian của các cá nhân [11] (atus_int.csv): 09 thuộc tính và 539,253 dòng dữ liệu.

Bảng 3. *Số liệu đo được trong quá trình chạy thực nghiệm*

Tập dữ liệu	**adult_int.csv**		**ss13acs int.csv**		**ihis_int.csv**		**atus_int.csv**	
k	3	5	3	5	3	5	3	5
Số thuộc tính QI	9	9	30	30	9	9	9	9
Thời gian thực thi (ms)	6,622	9,027	1,728,577	1,824,912	134,632	216,957	20,606	31,282
Granularity/Loss	0,9165	0,9340	0,9507	0,9303	0,9738	0,9579	0,9836	0,9661

Kết quả thực nghiệm cho thấy thời gian chạy bị ảnh hưởng nhiều bởi số lượng dòng của tập dữ liệu. Số lượng thuộc tính QI không ảnh hưởng nhiều đến thời gian thực thi.

Giải pháp nặc danh hóa theo phương pháp local generation cho kết quả chất lượng dữ liệu khá tốt, độ đo Granularity/Loss đều đạt tỷ lệ cao (gần bằng 1) thể hiện các giá trị thuộc tính được nặc danh hóa bao trùm gần như toàn bộ miền trị ban đầu của thuộc tính.

5. Kết luận

Về cơ bản, nghiên cứu đã thực hiện:

Tìm hiểu các khái niệm về Open Data, đặc trưng của Open Data và các vấn đề liên quan.

Tìm hiểu các phương pháp và kỹ thuật nặc danh hóa dữ liệu: k – anonymity, ℓ - diversity…

Tìm hiểu các tập dữ liệu thực nghiệm.

Chạy chương trình thực nghiệm trên các tập dữ liệu mở.

Đo lường được thời gian nặc danh hóa và chất lượng dữ liệu sau khi thực hiện.

Ưu điểm

Chương trình thực nghiệm trên các tập dữ liệu có các đặc điểm đa dạng khác nhau: số lượng thuộc tính và số dòng dữ liệu.

Thiết lập và đánh giá các kết quả dựa trên các thông số phổ biến.

Nhược điểm

Chương trình thực nghiệm chỉ mới hiện thực kỹ thuật k – anonymity trên các tập dữ liệu, chưa thực nghiệm với các kỹ thuật nâng cao như ℓ - diversity, t – closseness.

Kết quả thực nghiệm chỉ xét trên thời gian thực thi của mô hình và chất lượng dữ liệu theo độ đo Granularity/Loss, chưa đánh giá các vấn đề khác như giá trị thông tin, luật kết hợp của tập dữ liệu sau khi được nặc danh hóa và đánh giá chất lượng dữ liệu theo các độ đo khác.

Tập dữ liệu được lấy từ tập Open Data mẫu của nước ngoài. Vì Open Data ở Việt Nam còn ít, chưa có đầy đủ và đa dạng nên chưa thử nghiệm trên các tập Open Data trong nước.

Hướng phát triển

Dựa trên kết quả nghiên cứu được, đề tài có thể được phát triển theo hướng áp dụng thử nghiệm các phương pháp và kỹ thuật nặc danh hóa khác như ℓ - diversity, t – closseness…; thực nghiệm trên nhiều tập dữ liệu Open Data với đa dạng thuộc tính QI và nhiều bộ dữ liệu hơn; áp dụng các kỹ thuật lên các tập Open Data trong nước; đánh giá kết quả nặc danh hóa dữ liệu trên các độ đo khác; áp dụng các công nghệ lưu trữ và xử lý Big Data như Apache Hadoop, Apache Spark để tăng tốc quá trình xử lý.

Lời cám ơn

Công trình này thuộc đề án năm 2022 được tài trợ kinh phí bởi Trường Đại học Sư phạm Kỹ thuật Thành Phố Hồ Chí Minh.

Xung đột lợi ích

Các tác giả tuyên bố không có xung đột lợi ích trong bài báo này.

TÀI LIỆU THAM KHẢO

[1] A. Goben and R. J. Sandusky, "Open Data Repositories Current Risks and opportunities," C&RL News, 2020, doi: 10.5860/crln.81.1.62.
[2] N. Patoulias, "Political risks analysis using open data," Open University Cyprus, 2019. [Online]. Available: http://hdl.handle.net/11128/4185.
[3] R. Bild, K. A. Kuhn, and F. Prasser, "SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees," *Proceedings on Privacy Enhancing Technologies*, 2018, doi: 0.1515/popets-2018-0004.
[4] L. T. Hieu and D. T. Khanh, "An Elastic Anonymization Framework for Open Data," FDSE 2020, CCIS 1306, pp. 108–119, 2020, doi: 10.1007/978-981-33-4370-2_8.
[5] F. Z. Borgesius, J. Gray, and M. V. Eechoud, "Open Data, Privacy, And Fair Information Principles: Towards A Balancing Framework," *Berkeley Technology Law Journal*, vol. 30, no. 3, pp. 2073-2131, 2015, doi: 10.15779/Z389S18.
[6] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002, doi: 10.1142/S0218488502001648.
[7] A. Machanavajjhala, J. Gehrke, and D. Kifer, "ℓ-Diversity: Privacy Beyond k-Anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, doi: 10.1109/ICDE.2006.1.
[8] B. Becker and R. Kohavi. UCI Machine Learning Repository, doi: 10.24432/C5XW20.
[9] American Community Survey Main - U.S. Census Bureau. Accessed: Oct. 01, 2015. [Online]. Available: http://www.census.gov/acs/www/.
[10] Integrated Health Interview Series, doi: 10.18128/D070.V6.4.
[11] American Time Use Survey. Accessed: 2019. [Online]. Available: https://www.bls.gov/tus/data/datafiles-2019.htm.
[12] V. S. Iyengar *et al.*, "Transforming Data to Satisfy Privacy Constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 2022, doi: 10.1145/775047.775089.
[13] L. Sweeney *et al.*, "Achieving k-anonymity privacy protection using generalization and suppression," vol. 10, no. 05, pp. 571-588, 2002, doi: 10.1142/S021848850200165X.
[14] A. Gionis and T. Tassa, "k-Anonymization with Minimal Loss of Information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 2, 2009, doi: 10.1109/TKDE.2008.129.
[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, doi: 10.1109/ICDE.2006.101.
[16] R. J. Bayardo and R. Agrawa, "Data privacy through optimal k-anonymization," in *21st International Conference on Data Engineering (ICDE'05)*, 2005, doi: 10.1109/ICDE.2005.42.
[17] K. E. Emam *et al.*, "A Globally Optimal k-Anonymity Method for the De-Identification of Health Data," *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 670–682, 2009, doi: 10.1197/jamia.M3144.
[18] J. Goldberger and T. Tassa, "Efficient Anonymizations with Enhanced Utility," *Transactions on Data Privacy*, vol. 3, pp. 149–175, 2010, doi: 10.1109/ICDMW.2009.15.
[19] Z. Wan and Y. Vorobeychik, "A Game Theoretic Framework for Analyzing Re-Identification Risk," 2015, doi: 10.1371/journal.pone.0120592.

TÓM TẮT TIỂU SỬ CỦA CÁC TÁC GIẢ BẰNG TIẾNG ANH.

Lê Thị Minh Châu graduated from university in Infomatics in 2005 at the Open University and received a master's degree in computer science in 2012 at Ho Chi Minh City University of Technology. I am currently working at the Faculty of Information Technology, HCMC University of Technology and Education. My research interests include security, privacy, Big Data AI, and related technologies. My email: chaultm@hcmute.edu.vn. ORCID: https://orcid.org/0009-0004-8372-9098

Nguyễn Trần Thi Văn got my BS degree in Information Technology in 2002 at the University of Natural Sciences – National University Hochiminh City and finished my Masters program in Computer Sciences in 2015 at the University of Information Technology – National University Hochiminh City. I have been working as a Lecturer at the Falculty of Information Technology, University of Technology and Education Hochiminh City since 2003. My research interests include software engineering, mobile application development, data mining, social network analysis and machine learning. My email: nttvan@hcmute.edu.vn