DOI: 10.20998/2079-0023.2024.02.12
УДК 004.67

Д. В. КАЛІНІН, аспірант кафедри системного аналізу та інформаційно-аналітичних технологій, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна; e-mail: Denys.Kalinin@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0004-4431-7728
В. П. СЕВЕРИН, д-р техн. наук, професор, професор кафедри системного аналізу та інформаційно-аналітичних технологій, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна; e-mail: Valerii.Severyn@khpi.edu.ua; ORCID: https://orcid.org/0000-0002-2969-6780
М. І. БЕЗМЕНОВ, кандидат технічних наук (PhD), доцент, професор кафедри системного аналізу та інформаційно-аналітичних технологій, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна; e-mail: Mykola.Bezmenov@khpi.edu.ua; ORCID: https://orcid.org/0000-0002-2995-2350

МОДЕЛІ ПРИВАТНОСТІ ТА ТЕХНІКИ АНОНІМІЗАЦІЇ ТАБЛИЧНИХ МЕДИЧНИХ ДАНИХ

У сучасному світі питання приватності та захисту персональних даних набувають надзвичайної актуальності, особливо в медичній галузі, де використання великих обсягів даних для досліджень стає все більш поширеним. Використання персональних даних регулюється відповідними законами, які вимагають анонімізації даних для мінімізації ризиків ідентифікації осіб. Анонімізація є процесом, що дозволяє використовувати чутливі дані без ризику розкриття особистої інформації, зберігаючи при цьому їх корисність. У статті розглядаються основні моделі приватності та техніки анонімізації, що застосовуються для захисту табличних медичних даних. Моделі приватності включають *k*-анонімність (*k*-anonymity), *l*-диверсність (*l*-diversity) та *t*-близькість (*t*-closeness). Модель *k*-анонімності забезпечує щоб будь-яка комбінація квазіідентифікаторів була спільною для щонайменше *k* записів. Модель *l*-диверсності доповнює *k*-анонімність, вимагаючи наявності щонайменше *l* унікальних комбінацій значень чутливих атрибутів (SA) у кожному класі еквівалентності. Модель *t*-близькості враховує розподіл значень цих чутливих атрибутів, забезпечуючи, щоб відстань між розподілом SA у класі еквівалентності та загальним розподілом не перевищувала заданий поріг. Техніки анонімізації включають узагальнення (generalization), подавлення (suppression), перенесення (relocation), перестановку (permutation), пертурбацію (perturbation), розділення (slicing), диференційну приватність (differential privacy) та синтетичні дані (synthetic data). Узагальнення зменшує точність квазіідентифікаторів. Подавлення видаляє певні значення з набору даних для покращення його статистичних характеристик. Перенесення змінює обмежену кількість значень в даних з метою підвищення захисту. Перестановка змішує значення квазіідентифікаторів між записами, зберігаючи при цьому загальні статистичні особливості набору даних. Пертурбація додає шум до даних, що підвищує приватність. Ідея диференційної приватності також полягає у додаванні шуму, але це виконується на етапі обробки запитів за даними. Генерація синтетичних даних дозволяє створювати нові набори даних, як схожі за характеристиками на оригінальні дані.

Ключові слова: обробка даних, моделі штучного інтелекту, класифікація, ідентифікація, моделі приватності, медичні дані, табличні дані, захист чутливих даних, анонімізація даних, техніки анонімізації даних, диференційна приватність.

Вступ. У сучасному світі зберігання та обробка персональних даних невід'ємно пов'язані з питаннями приватності та захисту таких даних. Для медичної галузі дані питання є особливо актуальними, оскільки все більш поширеною практикою стає використання великих об'ємів медичних даних для проведення різних досліджень та аналізу, у тому числі з використанням штучного інтелекту. Використання персональних даних для другорядних цілей регулюється конкретними законами GDPR [1] та HIPAA [2], порушення яких може призвести до серйозних наслідків, часом навіть незворотних. Відповідно до цих актів чутливі дані можуть бути використані тільки після спеціальної їх обробки, в результаті якої ідентифікація осіб з набору стає неможливою або ризик такої ідентифікації не перевищує допустимий рівень. Даний процес має назву *анонімізація* (anonymization), іноді – деідентифікація. *Основна задача анонімізації* – мінімізувати ризики розкриття особистої інформації, зберігаючи при цьому максимальну корисність вихідних даних. Для вирішення цієї задачі використовуються спеціальні техніки (методи), що забезпечують умови для певних моделей приватності (на практиці це часто комбінації технік).

Мета статті. Описати основні моделі приватності та техніки анонімізації і провести аналіз їх переваг та недоліків у контексті анонімізації табличних медичних даних.

Моделі приватності. У процесі анонімізації реляційних даних виділяють наступні категорії атрибутів: прямі ідентифікатори (direct identifiers, DI), квазіідентифікатори (quasi-identifiers, QI), чутливі атрибути (sensitive attributes, SA) та нечутливі атрибути. *Прямі ідентифікатори* – це атрибути, які дозволяють однозначно визначити конкретну особу. *Квазіідентифікатори* – атрибути, що лише у комбінації з іншими квазіідентифікаторами можуть бути використані для ідентифікації даних. Під *чутливими* (*захищеними*) *атрибутами* розуміють такі атрибути, які мають залишатися приватними та захищеними від потенційного зловмисника. *Нечутливі атрибути* у свою чергу можуть вважатися безпечними для публічного використання. Орієнтуючись на дану класифікацію, можна описати основні моделі приватності, які використовуються для анонімізації даних.

Модель приватності під назвою *k-анонімність* (*k*-anonymity) було запропоновано в роботі [3]. Дана модель вимагає, аби при будь-якій комбінації QIs щонайменше *k* записів мали однаковий набір значень квазіідентифікаторів, тим самим забезпечуючи анонімність у рамках відповідної групи (класу еквівалентності). Очевидно, що чим більше значення *k*, тим вищий рівень захисту ми отримуємо. Основними недоліками даної моделі є неефективність проти атак на основі однорідності (homogeneity attack) значень SAs в

межах класу еквівалентності, а також суттєве погіршення корисності анонімізованих даних при великих значеннях *k*.

Проблему беззахисності даних до атак на основі однорідності дозволяє вирішити модель *l-диверсності* (*l*-diversity), що була представлена як доповнення до моделі *k*-анонімності в [4]. Термін «*l*-диверсність» визначає, що в рамках кожного класу еквівалентності присутні хоча б *l* унікальних комбінацій значень для набору чутливих атрибутів. Маючи *l*-диверсність значень SAs, у загальному випадку можна стверджувати, що зловмисник буде мати не менше ніж *l* варіантів значень SAs при спробі отримати чутливу інформацію в рамках відповідного класу розміром *k*. Основною метою даної моделі приватності є захист чутливих атрибутів, але на практиці доволі часто виникають ситуації, коли *l*-диверсність не гарантує, що ці *l* варіантів є дійсно унікальними. Так, наприклад, деякі діагнози можуть мати декілька різних назв, але семантично бути еквівалентними, що призведе до розкриття чутливої інформації при формальному дотриманні вимог моделі. Також важливим недоліком даної моделі є її несприйнятливість до розподілу даних в SAs. Так, наприклад, якщо деякий набір значень SAs домінує в своєму класі еквівалентності, він тим самим збільшує шанси зловмисника отримати чутливу інформацію відповідно до наявного розподілу.

Модель *t-близькості* (*t*-closeness) була введена в рамках роботи [5], де було визначено, що клас еквівалентності називається таким, що має *t*-близькість, у випадку, коли відстань між розподілом чутливого атрибуту у своєму класі та його розподілом у повному наборі даних не перевищує заданий поріг *t*. Таким чином, можна казати про таку залежність, коли менше значення *t* відповідає більш високому рівню захисту чутливих даних. Для обчислення відстані між розподілами найчастіше використовують відстань Кульбака – Лейблера або відстань Васерштейна. Дана модель гармонійно доповнює моделі *k*-анонімності та *l*-диверсності, надаючи рішення проблеми викриття чутливих даних, що спричинена специфічним розподілом значень SAs у групі. Але звісно, що на практиці можуть виникати унікальні ситуації, коли та чи інша модель не зможе гарантувати задовільні результати.

Техніки анонімізації даних. Техніки, що розглядаються в даному розділі, базуються на вимогах вищеописаних моделей приватності: *k*-анонімності, *l*-диверсності та *t*-близькості, а також їх варіацій.

Техніка узагальнення (generalization) [3] використовується для зменшення точності певного QI. Так, наприклад, говорячи про конкретні дати (дата народження, дата відвідування лікаря тощо), дана техніка дозволяє змінити ступінь деталізації даних до необхідного рівня (місяця, року), тим самим забезпечуючи захист від ре-ідентифікації. Для атрибутів, значення яких беруться з деякої наперед відомої скінченної множини, зазвичай будується відповідна ієрархія, яка використовується для заміни даних менш конкретними узагальненнями. Важливою особливістю даної техніки є збереження правдивості даних. Серед недоліків можна виділити втрату корисності даних у випадках надмірного узагальнення. Для вирішення даної проблеми було запропоновано техніку адаптивного узагальнення (adaptive generalization) [6], що дозволяє мінімізувати негативний ефект на корисність даних.

Техніка подавлення даних (suppression) [7] часто застосовується у алгоритмах на базі *k*-анонімності. Дана техніка полягає у вилученні певних значень з набору даних з метою покращення статистичних характеристик цього набору. Так, наприклад, у випадку, коли один запис виділяється через аномальні значення певних атрибутів, для досягнення задовільного рівня захисту анонімізованих даних, виникає потреба у надмірному узагальненні таких атрибутів, що в результаті призводить до суттєвих втрат корисності всього набору даних. Вилучення необхідних значень чи відповідних записів з набору дозволяє вирішити дану проблему. З іншої сторони, надмірне застосування техніки вилучення також негативно впливає на корисність даних.

У роботі [8] було запропоновано *техніку перенесення* (relocation), яка полягає у перенесенні або зміні обмеженої кількості значень набору даних для забезпечення більш високого рівня захисту від ре-ідентифікації. Очевидно, що недоліком даної техніки є надання анонімізованим даним певного рівню неправдивості, що може призвести до погіршення загальної їх якості.

Окремо можна виділити *техніку перестановки* (permutation), яка представляю собою змішування значень певного квазіідентифікатора між різними записами, зберігаючи при цьому загальні статистичні особливості оригінальних даних та забезпечуючи високий рівень захисту. Треба зазначити, що зловживання даною технікою може призвести до порушень зв'язків між атрибутами, що у свою чергу може погіршити цінність таких анонімізованих даних для аналізу, чутливого до таких зв'язків.

Техніка пертурбації (perturbation) використовується для додавання адитивного або мультиплікативного шуму до оригінальних даних (частіше до SAs). У випадку адитивного шуму додається значення з фіксованого діапазону, в той час як для мультиплікативного випадково обирається коефіцієнт множення. Дана техніка може бути неефективною для анонімізації даних з аномаліями, а тому її часто використовують у комбінації з іншими техніками для забезпечення достатнього рівня приватності. Окрім цього, вибір невдалого діапазону чи коефіцієнту може призвести до суттєвого зниження корисності вихідних даних.

Техніка бакетизації (bucketization) [9] спрямована на забезпечення вимог моделі приватності *l*-диверсності. Спочатку дані розбиваються на групи (класи еквівалентності) за значеннями QIs, після чого до кожної групи застосовується вищезгадана техніка перестановки, яка перемішує значення чутливого атрибуту у межах своєї групи. Такий підхід дозволяє отримати вищий рівень корисності даних у порівнянні з технікою узагальнення, але при цьому він також має певні недоліки. Так, наприклад, бакетизація не захищає дані від атак на розкриття членства конкретної особи (membership disclosure) через те, що квазіідентифікатори залишаються незмінними.

На практиці забезпечити виконання умов моделей *k*-анонімності, *l*-диверсності та *t*-близькості разом буває дуже складно без суттєвих втрат корисності даних, а іноді й взагалі неможливо. Покращити процес анонімізації в даному випадку допомагає використання *техніки мікроагрегації* (microaggregation) [10]. Дана техніка в загальному випадку застосовується до числових QIs, хоча також може бути використана для атрибутів на базі категорій при застосуванні конвертації в числовий формат [11]. Спочатку дані розбиваються на певні кластери таким чином, щоб дані записів в одному кластері були максимально між собою, а дані записів із різних кластерів мали максимальну відмінність. При цьому, кожен кластер буде мати щонайменше *k* записів. Далі, значення QIs в рамках конкретного кластера обробляються найбільш доцільним способом (наприклад, замінюються на середнє значення), в результаті чого забезпечується захист даних від ре-ідентифікації. Проблемою даного підходу є пошук оптимального рішення у випадку багатовимірних даних, оскільки такий алгоритм має клас складності NP та потребує значних ресурсів для обчислення.

Техніку розділення (slicing) було запропоновано в роботі [12], як альтернативу технікам узагальнення та бакетизації. Ідея техніки розділення представляє собою сегментацію набору даних одночасно по атрибутам (вертикальна) і по записам (горизонтальна). Вертикальна сегментація утворює нові колонки через групування сильно зв'язаних між собою атрибутів, в той час як горизонтальна – створює групи даних (buckets) на основі записів. Після цього застосовується перемішування або сортування даних у кожному стовпці для того, щоб приховати зв'язок з атрибутами з інших стовпців. У результаті анонімізовані дані мають кращу корисність у порівнянні з застосуванням технік узагальнення та бакетизації саме за рахунок збереження зв'язків між зв'язаними атрибутами.

Техніка диференційної приватності (differential privacy, DP) [13] – це математична техніка, яка як і техніка пертурбації, базується на ідеї додавання шуму до даних. На відміну від техніки пертурбації, де шум інтегрується безпосередньо в набір даних, механізм диференційної приватності застосовується на рівні запитів за даними, надаючи певний рівень випадковості до кожного результату. Виділяють локальну та глобальну диференційні приватності, що відрізняються між собою архітектурними особливостями системи, яка обробляє запити. У роботі [14] також було запропоновано варіант реалізації DP із застосуванням глибоких нейронних мереж.

До технік анонімізації даних також можна віднести генерацію синтетичних даних (synthetic data). На відміну від вищезазначених технік, генерація синтетичних даних надає абсолютно нові дані, які тільки схожі за структурою та розподілом на оригінальні дані, але при цьому не містять ніякої персональної інформації, оскільки вони не можуть бути асоційовані з конкретною реальною особою. І хоча проблема створення штучних даних не є новою, за останнє десятиріччя досягнення у сфері генеративного штучного інтелекту сприяли поширенню використання моделей штучного інтелекту для вирішення даної задачі.

Коли мова йде про структуровані дані, для навчання та безпосередньо генерації нових даних використовуються такі групи моделей ШІ, як варіаційні автокодувальники (Variational Autoencoders, VAEs) [15] та генеративні змагальні мережі (Generative Adversarial Networks, GANs) [16]. Дані моделі вивчають структуру та внутрішні залежності даних, на яких вони навчаються, а також, що важливо для задачі анонімізації, аналізують правила розподілу цих даних. Варіаційні автокодувальники представляють собою групу генеративних моделей, які працюють на принципі стискання даних до формату представлення, що називається латентним простором. Даний простір надалі використовується для відтворення вхідних даних. У свою чергу модель генеративної змагальної мережі складається з двох нейронних мереж – мережі генератора (G) та мережі дискримінатора (D), між якими виникає антагоністична гра (гра з нульовою сумою). Мережа G створює зразки штучних даних, а мережа D намагається вгадати, чи є даний зразок штучним. При цьому, генератор постійно вдосконалює створювані штучні зразки таким чином, щоб дискримінатор не зміг відрізнити штучні дані від реальних, а дискримінатор вчиться точніше робити висновки.

Повний процес генерації синтетичних даних складається з таких етапів, як тренування моделі на реальних даних, безпосередньо генерація нових наборів даних, а також оцінка якості та корисності цих вихідних даних. Важливо зазначити, що ключовим фактором можливості використання синтетичних даних є ступінь коректності процесу генерації таких даних у контексті вирішення певної задачі. Так, наприклад, якщо деяка система потребує великої кількості даних певної структури з метою аналізу продуктивності цієї системи, то в такому випадку характеристики самих даних та їх розподіл не будуть мати великої різниці, а тому не має потреби вимагати від процесу генерації високого рівня правдоподібності вихідних наборів даних.

Якісні синтетичні дані дозволяють вирішити багато проблем, пов'язаних з обмеженнями використання та розповсюдження персональних даних, а в деяких випадках просто не мають альтернатив. На практиці синтетичні дані часто використовуються для тренування та перевірки моделей штучного інтелекту. Використання синтетичних даних також має ряд недоліків. Так, налаштування процесу генерації синтетичних даних є вкрай складною задачею, де особливу увагу треба приділяти оцінці згенерованих даних, їх якості та корисності відносно вирішення конкретної задачі.

Висновки. У роботі було розглянуто проблему зберігання та використання структурованих медичних даних, а також їх анонімізацію, як механізм забезпечення приватності персональної інформації. У контексті анонімізації було розглянуто три основні моделі приватності *k*-анонімність, *l*-диверсність та *t*-близькість. Були описані ролі даних моделей в забезпеченні анонімності, а також їх недоліки, що зумовили необхідність в створенні інших моделей приватності та

таких технік анонімізації, як диференційна приватність та генерація синтетичних даних. Було також розглянуто техніки узагальнення, подавлення, перенесення, перестановки, пертурбації та розділення. Дані техніки використовуються для обробки даних таким чином, щоб виконати вимоги відповідних моделей приватності. Не дивлячись на те, що поєднання розглянутих моделей приватності та технік анонімізації дозволяють гарантувати достатньо високий рівень захисту від реідентифікації, особливу увагу треба приділити оцінці корисності вихідних даних. Так, застосування неякісних анонімізованих даних не тільки буде неефективним, а може навіть призвести до шкідливих наслідків. Такі техніки, як диференційна приватність та генерація синтетичних даних, відрізняються своїм підходом до забезпечення приватності даних та мають широке практичне застосування. Більш того, активний розвиток генеративних нейронних мереж відкрив нові можливості для генерації синтетичних даних, що робить даний напрям перспективним для подальшого дослідження.

Список використаної літератури

1. Document 32016R0679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *EUR-Lex. Access to European Union law*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (дата звернення: 30.10.2024).
2. *Summary of the HIPAA Privacy Rule*. USA: United States Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (дата звернення: 30.10.2024).
3. Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*. 2002. Vol. 10, is. 5. P. 557–570. URL: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf (дата звернення: 30.10.2024).
4. Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M.. L-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data*. 2007. Vol. 1, is. 1. P. 3 – es.
5. Li N., Li T., Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering*. 2007. P. 106–115. URL: https://ieeexplore.ieee.org/document/4221659 (дата звернення: 01.11.2024).
6. Majeed A., Ullah F., Lee S. Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*. 2017. Vol. 17, is. 5. Article 1059. URL: https://www.mdpi.com/1424-8220/17/5/1059 (дата звернення: 01.11.2024).
7. Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. *Semantic Scholar*. 1998. URL: https://www.semanticscholar.org/paper/Protecting-privacy-when-disclosing-information%3A-and-Samarati-Sweeney/7df12c498fecedac4ab6034d3a8032a6d1366ca6 (дата звернення: 30.10.2024).
8. Nergiz M. E., Gök M. Z. Hybrid k-anonymity. *Computers & security*. 2014. Vol. 44. P. 51–63. URL: https://www.researchgate.net/publication/261139007_Hybrid_k-Anonymity (дата звернення: 30.10.2024).
9. Martin D. J., Kifer D., Machanavajjhala A., Gehrke J., Halpern J. Y. Worst-Case Background Knowledge for Privacy-Preserving Data Publishing. *IEEE 23rd International Conference on Data Engineering*. 2007. P. 126–135. URL: https://www.academia.edu/1520959/Worst_Case_Background_Knowledge_for_Privacy_Preserving_Data_Publishing (дата звернення: 30.10.2024).
10. Domingo-Ferrer J., Mateo-Sanz J. M. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*. 2002. Vol. 14, no. 1. P. 189–201.
11. Gal T. S., Tucker Th. C., Gangopadhyay A., Chen Z.. A data recipient centered de-identification method to retain statistical attributes. *Journal of biomedical informatics*. 2014. Vol. 50. P. 32–45.
12. Li T., Li N., Zhang J., Molloy I. Slicing: A New Approach to Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*. 2010. Vol. 24, is. 3. P. 561–574.
13. Dwork C. Differential Privacy. *Automata, Languages and Programming*. 2006. P. 1–12.
14. Abadi M., Chu A., Goodfellow I., McMahan H. B., Mironov I., Talwar K., Zhang L. Deep Learning with Differential Privacy. *CCS'16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016. P. 24–28.
15. Kingma D. P., Wellin M. Auto-Encoding Variational Bayes. *arXiv*. 2013. Article 1312.6114. URL: https://www.semanticscholar.org/reader/5f5dc5b9a2ba710937e2c413b37b053cd673df02 (дата звернення: 01.11.2024).
16. Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair Sh., Courville A., Bengio Y. Generative adversarial networks. *arXiv*. 2014. Article 1406.2661. URL: https://dl.acm.org/doi/pdf/10.1145/3422622 (дата звернення: 01.11.2024).

References

1. Document 32016R0679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *EUR-Lex. Access to European Union law*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (accessed 30.10.2024).
2. *Summary of the HIPAA Privacy Rule*. USA: United States Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (accessed 30.10.2024).
3. Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*. 2002б vol. 10, is. 5, pp. 557–570. URL: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf (accessed 30.10.2024).
4. Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M.. L-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data*. 2007, vol. 1, is. 1, pp. 3 – es.
5. Li N., Li T., Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering*. 2007, pp. 106–115. URL: https://ieeexplore.ieee.org/document/4221659 (accessed 01.11.2024).
6. Majeed A., Ullah F., Lee S. Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*. 2017, vol. 17, is. 5, article 1059. URL: https://www.mdpi.com/1424-8220/17/5/1059 (accessed 01.11.2024).
7. Samarati P., Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. *Semantic Scholar*. 1998. URL: https://www.semanticscholar.org/paper/Protecting-privacy-when-disclosing-information%3A-and-Samarati-Sweeney/7df12c498fecedac4ab6034d3a8032a6d1366ca6 (accessed 30.10.2024).
8. Nergiz M. E., Gök M. Z. Hybrid k-anonymity. *Computers & security*. 2014, vol. 44, pp. 51–63. URL: https://www.researchgate.net/publication/261139007_Hybrid_k-Anonymity (accessed 30.10.2024).
9. Martin D. J., Kifer D., Machanavajjhala A., Gehrke J., Halpern J. Y. Worst-Case Background Knowledge for Privacy-Preserving Data Publishing. *IEEE 23rd International Conference on Data Engineering*. 2007, pp. 126–135. URL: https://www.academia.edu/1520959/Worst_Case_Background_Kno

wledge_for_Privacy_Preserving_Data_Publishing (accessed 30.10.2024).
10. Domingo-Ferrer J., Mateo-Sanz J. M. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*. 2002, vol. 14, no. 1, pp. 189–201.
11. Gal T. S., Tucker Th. C., Gangopadhyay A., Chen Z.. A data recipient centered de-identification method to retain statistical attributes. *Journal of biomedical informatics*. 2014, vol. 50, pp. 32–45.
12. Li T., Li N., Zhang J., Molloy I. Slicing: A New Approach to Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*. 2010, vol. 24, is. 3, pp. 561–574.
13. Dwork C. Differential Privacy. *Automata, Languages and Programming*. 2006, pp. 1–12.
14. Abadi M., Chu A., Goodfellow I., McMahan H. B., Mironov I., Talwar K., Zhang L. Deep Learning with Differential Privacy. *CCS'16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 24–28.
15. Kingma D. P., Wellin M. Auto-Encoding Variational Bayes. *arXiv*. 2013, article 1312.6114. URL: https://www.semanticscholar.org/reader/5f5dc5b9a2ba710937e2c413b37b053cd673df02 (accessed 01.11.2024).
16. Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair Sh., Courville A., Bengio Y. Generative adversarial networks. *arXiv*. 2014, article 1406.2661. URL: https://dl.acm.org/doi/pdf/10.1145/3422622 (accessed 01.11.2024).

Надійшла (received) 05.11.2024

UDC 004.67

D. V. KALININ, Postgraduate of Department of System Analysis and Information-Analytical Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine; e-mail: Denys.Kalinin@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0004-4431-7728

V. P. SEVERYN, Doctor of Technical Sciences, Professor, Professor of Department System Analysis and Information-Analytical Technologies, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine; e-mail: Valerii.Severyn@khpi.edu.ua; ORCID: https://orcid.org/0000-0002-2969-6780

M. I. BEZMENOV, Candidate of Technical Sciences (PhD), Docent, Professor of the Department of System Analysis and Information-Analytical Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine; e-mail: Mykola.Bezmenov@khpi.edu.ua; ORCID: https://orcid.org/0000-0002-2995-2350

PRIVACY MODELS AND ANONYMIZATION TECHNIQUES FOR TABULAR HEALTHCARE DATA

In today's world, issues of privacy and personal data protection are becoming extremely relevant, especially in the healthcare field, where the use of large volumes of data for research is becoming increasingly common. The use of personal data is regulated by relevant laws that require data anonymization to minimize the risks of identifying individuals. Anonymization is a process that allows the use of sensitive data without the risk of disclosing personal information while maintaining its utility. This article discusses the main privacy models and anonymization techniques used to protect tabular healthcare data. Privacy models include *k*-anonymity, *l*-diversity, and *t*-closeness. The *k*-anonymity model ensures that any combination of quasi-identifiers is shared by at least *k* records. The *l*-diversity model complements *k*-anonymity by requiring at least *l* unique combinations of sensitive attribute (SA) values in each equivalence class. The *t*-closeness model considers the distribution of these sensitive attribute values, ensuring that the distance between the SA distribution in the equivalence class and the overall distribution does not exceed a specified threshold. Anonymization techniques include generalization, suppression, relocation, permutation, perturbation, slicing, differential privacy, and synthetic data. Generalization reduces the precision of quasi-identifiers. Suppression removes certain values from the dataset to improve its statistical characteristics. Relocation changes a limited number of values in the data to enhance protection. Permutation mixes the values of quasi-identifiers between records while preserving the overall statistical features of the dataset. Perturbation adds noise to the data, increasing privacy. The idea of differential privacy also involves adding noise, but this is done at the query processing stage. Generating synthetic data allows the creation of new datasets that are similar in characteristics to the original data.

Keywords: healthcare data, tabular data, data anonymization, privacy models, *k*-anonymity, *l*-diversity, *t*-closeness, data anonymization techniques, differential privacy.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Калінін Денис Вікторович / Kalinin Denys Viktorovych
Автор 2 / Author 2: Северин Валерій Петрович / Severyn Valerii Petrovych
Автор 3 / Author 3: Безменов Микола Іванович / Bezmenov Mykola Ivanovych