# АНОТАЦІЯ

Захист персональної інформації в задачах аналізу та обробки великих даних // Дипломна робота ОР «Магістр» // Сачик Тетяна Владиславівна // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра кібербезпеки, група СБм-61 // Тернопіль, 2019 // С. 109 , рис. – 14 , табл. – 47 , кресл. – 0 , додат. – 1.

Ключові слова: К-АНОНІМІЗАЦІЯ, КОНФІДЕЦІЙНІСТЬ, КВАЗІ-ІДЕНТИФІКАТОР, ЕФЕКТИВНІСТЬ, СИНТЕТИЧНИЙ ГЕНЕРАТОР.

Метою роботи – систематичне порівняння трьох відомих алгоритмів *k*-анонімізації для вимірювання їх продуктивності (з точки зору використання ресурсів) та їх ефективності (з точки зору корисності даних).

Основні результати роботи: в роботі досліджено поняття анонімізації, описано її моделі, обрано для дослідження алгоритми *k*-анонімізації, як однієї з базових моделей, запропоновано критерії якості алгоритмів *k*-анонімізації для подальшого прийняття рішення щодо вибору алгоритму, проведено порівняння трьох основних алгоритмів Datafly, Incognito, Modrian для двох наборів даних (реального та синтетичного) та для різних початкових налаштувань параметрів алгоритмів, сформовано рекомендації щодо застосування алггоритмів.

У першому розділі описується, що таке анонімізація, моделі конфідеційності та на прикладах розглядаються деякі атаки конфідеційності.

У другому розділі розглядаються методи анонімізації, три найбільш поширені алгоритми анонімізації та описується методологія порівняння цих алгоритмів.

Третій розділ експерементальний. У ньому порівнюються три алгоритими анонімізації за такими критеріями як – час анонімізації, узагальнена втрата інформації, метрика чутливості та середній розмір класу еквівалентності.

У четвертому розділі описується генератор реальних синтетичних даних та принцип його роботи.

У розділі “Обґрунтування економічної ефективності” підраховується вартість роботи та термін її окупності.

У розділі “Охорона праці та безпека в надзвичайних ситуаціях” зазначено, що дослідження відбувалося зі збереженням правил пожежної безпеки та всіх норм охорони праці.

У розділі “Екологія” описуються методи узагальнення екологічної інформації та зазначаються вимоги до мікроклімату приміщень.

У результаті підготовки дипломної роботи проведено серію експериментів та всебічний аналіз для виявлення факторів, що впливають на ефективність загальнодоступних реалізації алгоритмів анонімізації. Представлено за допомогою експериментальної оцінки умови, в яких один алгоритм перевершує інші за певним показником, залежно від вхідних даних та вимог конфіденційності.

# ANNOTATION

Personal information protection in big data analysis and processing problems // Sachyk Tetiana // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information System and Software Engineering, Department of Cybersecurity// Ternopil, 2019 // P. 109 , Tables – 47 , Fig. – 14 , Annexes. – 1 , References – 56.

Keywords: K-ANONYMITY, PRIVACY, QUASI-IDENTIFIERS, EFFICIENCY, SYNTHETIC DATA GENERATOR.

Project purpose: systematic comparison of three well-known k-anonymization algorithms to measure their efficiency (in terms of resources usage) and their effectiveness (in terms of data utility).

Main results: The concept of anonymization is investigated, its models are described, k-anonymization algorithms are selected as one of the basic models, k-anonymization quality criteria are proposed for further decision making, algorithm selection is performed for three basic algorithms, Datafly Incognito, Modrian two sets of data (real and synthetic) and different initial adjustments of algorithm parameters, recommendations for the application of algorithms were formed.

The first section describes anonymization, privacy models, and some examples of privacy attacks.

The second section discusses anonymization methods, the three most common anonymization algorithms, and describes a methodology for comparing these algorithms.

The third section is experimental. It compares three anonymization algorithms against such criteria as anonymization time, generalized information loss, sensitivity metric, and average equivalence class size.

The fourth section describes the real synthetic data generator and how it works.

In the economic section the cost of the work and its payback period are calculated.

The section "Occupational Health and Safety" states that the study was conducted in compliance with fire safety rules and all occupational safety standards.

The section "Ecology" describes the methods of generalizing environmental information and specifies the requirements for the microclimate of the premises.

As a result of the preparation of the thesis, a series of experiments and a comprehensive analysis were conducted to identify the factors that influence the effectiveness of publicly available anonymization algorithms. Provided by an experimental evaluation of the conditions in which one algorithm outperforms the others by a certain measure, depending on the input and privacy requirements.

ЗМІСТ

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

$C_{AVG}$ (Average Equivalence Class Size Metric) – показник розміру середнього класу еквівалентності;

COCOA – синтетичний генератор даних;

DM (Discernibility Metric) – метрика чутливості;

EQ (Equivalense class) – клас еквівалентності;

GDPR (General Data Protection Regulation) – загальний регламент щодо захисту даних;

GenILoss (Generalized Information Loss) – узагальнена втрата інформація;

ID – ідентифікатор;

MDAV (maximum distance to average vector method) – метод максимальної відстані до середнього векторного методу;

PPDM (Privacy-Preserving Data Mining) – збереження конфіденційності видобутих даних;

PPDP (Privacy-Preserving Data Publishing) – публікація даних щодо збереження конфіденційності;

QID (Quasi-identifiers) – квазі-ідентифікатор;

SA (Sensitive attribute) – чутливий атрибут;

SDC (Statistical disclosure control) – контроль статистичного розкриття інформації;

VGH (Value generalization hierarchy)– ієрархія узагальнення значень.

# ВСТУП

В даний час обсяги створених даних щороку зростають експоненціально [1]. Серед цих даних збільшується кількість особистої інформації, що міститься в ній. Цей факт привернув увагу тих, хто зацікавлений у створенні більше індивідуальних та персоналізованих сервісів на основі наявної демографічної інформації. З цієї причини підприємства та організації в різних секторах збирають особисті дані, якими можуть ділитися за різних обставин (з грошових, суспільних чи юридичних причин). Тим не менш, це явище викликало нові виклики щодо захисту конфіденційності людей, представлених у опублікованих наборах даних.

Публікація даних щодо збереження конфіденційності (PPDP) стала цікавою для дослідників та практиків. Одним з ключових припущень моделі PPDP є те, що серед одержувачів даних можуть бути зловмисники, які мають намір розкрити конфіденційну інформацію про осіб. Таким чином, мета методів PPDP – змінити дані, зробивши їх менш специфічними, таким чином, щоб захищалася конфіденційність приватних осіб; при цьому прагнучи зберегти корисність анонімізованих даних. Суть PPDP полягає у створенні наборів даних, які мають хорошу корисність для виконання різних завдань, як правило, всі потенційні сценарії використання даних невідомі на момент публікації. Наприклад, за ініціативи відкритих даних [2] неможливо ідентифікувати всіх одержувачів даних. Таким чином у обміні персональними даними, повинні застосовуватись механізми збереження конфіденційності [3].

На перший погляд, вирішення цієї проблеми видається доволі тривіальним: адже достатнього просто видалити стовпці, що містять прямі ідентифікатори, такі як імена та номери соціального страхування тощо. Тим не менше, було доведено, що такого підходу недостатньо для збереження конфіденційності. Ця проблема виникає тому, що все ще можливо поєднувати різні набори даних або мати базові знання про людей, щоб зробити висновки про особу. Повторна ідентифікація особи досягається за допомогою зв’язування атрибутів, відомих як квазі-ідентифікатори (QID), таких як стать, дата

народження або поштовий індекс. . Науковці з США довели, що поєднуючи відкриту інформацію з різних джерел можна однозначно ідентифікувати 70-90% людей.

Існує кілька моделей, які пропонують формальні гарантії щодо захисту конфіденційності особи при публікації даних. Зосередимось на *k*-анонімізації, оскільки на відміну від інших моделей (ℓ-різноманіття, *t*-близькість та диференційна конфіденційність), які мають обмеження в використанні, ця модель є простою для розуміння і базовою у багатьох сферах використання

Метою дипломної роботи є порівняння найбільш відомих методів *k*-анонімізації (Datafly, Incognito, Mondrian) з огляду на використання ресурсів та корисність залишкових даних.

Для досягнення поставленої мети потрібно розв’язати наступні задачі:

- ознайомитись з літературними джерелами в області дослідження;
- дослідити реальні набори даних та процес генерування синтетичних наборів даних;
- запропонувати критерії оцінки якості алгоритмів;
- обчислити кількісні показники критеріїв якості для реального та синтетичного наборів даних;
- провести порівняльний аналіз алгоритмів анонімізування для різних налаштувань моделі анонімізації.

Об'єкт дослідження – процес анонімізації реального та синтетичного набору даних.

Предмет дослідження – моделі та алгоритми *k*-анонімізації.

Наукова новизна. В роботі запроновано критерії для оцінки якості алгоритмів k-анонімізації, проведено порівняльний аналіз трьох найбільш популярних алгоритмів k-анонімізації: Datafly, Incognito, Mondrian для різних налаштувань моделі k-анонімізації.

Апробація результатів роботи. Окремі результати роботи доповідались на VII науково-технічній конференції «Інформаційні моделі, системи та технології», Тернопіль, ТНТУ, 11 – 12 грудня 2019 р.

# РОЗДІЛ 1 АНОНІМІЗАЦІЯ. МОДЕЛІ, ЗАГРОЗИ, АТАКИ

## 1.1 Регламент захисту даних GDPR

Загальний регламент щодо захисту даних (General Data Protection Regulation, GDPR) – це документ, за допомогою якого Європейський парламент, Рада Європейського союзу і Європейська комісія підсилюють і уніфікують захист персональних даних фізичних осіб в Європейському союзі. Якщо простіше – це правила регулювання процедури збору, обробки, зберігання і розповсюдження персональних даних. Їх основна мета – захистити особисті дані, згідно з правами людини. Усі дані, за якими особа можна відразу ідентифікувати – вважаються персональними. Наприклад, корпоративна пошта містить ім'я і прізвище – буде відноситися до персональних даних, а з будь-яким іншим змістом (info @ company) – ні; ім'я та номер телефону – персональні дані, телефон / адреса самі по собі – просто дані. Якщо з даних стає щось відомо про людину (його місце роботи, контакти та інше), то вони належать до персональних [4].

Відповідно GDPR обов'язкові до виконання вимоги:

- збирати персональні дані тільки за згодою суб'єкта;
- використовувати та обробляти їх відповідно до поставлених цілей;
- по досягнені цих цілей - дані знищувати;
- вилучати та знищувати дані за вимогою їх власника;
- забезпечувати безпеку зберігання даних;
- не поширювати дані без згоди суб'єкта.

Для того, щоб компанія відповідала регламенту GDPR в своїй діяльності, їй потрібно дотримуватись наступних правил:

1. Отримувати згоду суб'єкта персональних даних на їх обробку.
2. Анонімізувати дані, з метою їх захисту при поширенні.
3. Документувати і зберігати перелік усіх дій по виконанню GDPR.
4. Призначити відповідального за захист даних співробітника.

## 1.2 Анонімізація

Згідно GDPR, немає конкретних вимог щодо ступеня, порядку і способу захисту даних – кожен має право вибирати сам. Як визначено в п.1.1, одним із завдань GDPR є анонімізація даних у випадку їх поширення третій стороні або публічній публікації.

Анонімізація, або редагування даних – процес видалення або приховування персональних даних з метою їх подальшого використання.

Незважаючи на те, що, на перший погляд, видається, що проблему можна розв'язати шляхом простого видалення стовпців датасету з відверто персональною інформацію, насправді проблема значно глибша. І прикладом того є інцидент, який стався в штаті Массачусетс [5]. Державне агентство, Комісія з групового страхування (GIC), придбало медичне страхування для державних службовців. Задля надання реальних даних для дослідників, GIC вирішив опублікувати набір даних, що складається з відвідувань лікарні кожного державного службовця. Для захисту конфіденційності GIC видалила всі опубліковані набори даних усіх імен, адрес та інших явних ідентифікаторів. Професор інформатики Суїні [5] придбав неанонімну копію списку виборців держави за 20 доларів. З опублікованого набору даних та списків виборців дивовижним виявилося, що 87% населення США можна однозначно ідентифікувати, використовуючи поєднання дати народження, статі та поштового індексу. Ці атрибути, на основі яких відбувається повторна ідентифікація, називають квазі-ідентифікаторами (QID). В результаті цього дослідження Суїні зміг ідентифікувати за допомогою опублікованих наборів даних медичну документацію губернатора Вільяма Уельда. Висновок Суїні базувався на даних перепису населення 1990 року. Через десятиліття Голле [6] перерахував статистику на основі даних перепису населення 2000 року. Використовуючи ту саму комбінацію атрибутів QID, тобто дати народження, статі та поштового індексу, Голлу вдалося однозначно ідентифікувати 63% населення США.

У 2006 році Netflix, відомий сервіс онлайн-прокату фільмів, запропонував приз у розмірі 1 000 000 000 доларів тому, хто міг би покращити їхню систему рекомендацій щодо кіно на 10% [7]. Щоб полегшити роботу дослідникам, Netflix випустив набір даних, що містить рейтинги фільмів, які належать майже 500 000 підписників, а також назви фільмів та рейтинг кожного фільму [8]. Намагаючись захистити конфіденційність користувачів, усі явні ідентифікатори були видалені з опублікованого набору даних та замінені ідентифікованими випадковими ідентифікаторами. Однак робота [8] передбачає, що знаючи дати (±2 тижні) шести рейтингів фільмів, 99% людей у опублікованому наборі даних можуть бути ідентифіковані. Крім того, знаючи лише два фільми з датами рейтингу (±3 дні), 68% підписників ідентифікуються з набору даних. Автори в роботі [8] здійснили свою атаку, використовуючи загальнодоступний, неанонімний, зовнішній набір даних про огляди фільмів з веб-сайту Internet Movie Database (IMDb).

Наведені вище приклади демонструють атаку на конфіденційність, полегшену можливістю збору базових знань із зовнішніх джерел даних, що містять спільну інформацію про одну групу осіб у опублікованому наборі даних. Наступний приклад демонструє можливість однозначної ідентифікації особи за зовнішніми анонімними даними.

America Online (AOL) випустив набір даних, що містить 20 мільйонів пошукових запитів 650 000 користувачів. Метою цього випуску даних було "охопити бачення відкритої дослідницької спільноти". AOL анонімізував дані перед публікацією, видаляючи імена користувачів AOL, IP-адреси та інші ідентифікатори. Кожному користувачеві та його пошукам було задано випадкове число. У опублікованому наборі даних одному користувачеві не пощастило, щоб подати набір запитів, серед яких такі: «ландшафти в Лілберні, штат Джорджія», «кілька людей з прізвищем Арнольд» та «будинки, продані в тіньовому озері підрозділу округу Гвіннетт Грузія». Ці запити виявилися настільки унікальними серед усіх інших поданих запитів, що двоє репортерів New York Times швидко визначили, хто шукав «будинки, продані в тіньовому озері підрозділу округу Гвіннетт Грузія»; Пані Тельма Арнольд, 62-річна вдова, яка проживає в Лілберні,

штат Джорджія [9]. В інтерв'ю пані Тельма Арнольд тоді підтвердила пошукові запити. Цей інцидент призвів до вилучення опублікованих наборів даних і змусив власників даних неохоче оприлюднювати свої дані, принаймні в найближчому майбутньому.

## 1.3 Атаки конфідеційності

При публікації таблиць даних, що містять інформацію про певну групу осіб, не менш важливо захищати конфіденційність тих осіб, чиїми даними обмінюються. Термін конфіденційність у цьому контексті є двояким: (1) особа будь-якої особи в опублікованих даних не повинна розкриватися, і (2) будь-яка інформація, яка вважається секретною (уразливою), не повинна пов'язуватися з відповідною особою.

Існує два типи загроз, які ставлять під загрозу конфіденційність осіб, дані яких публікуються. Якщо модель анонімізації не враховує, скільки інформації може бути зібрано із зовнішніх джерел даних про осіб у опублікованих даних, то опубліковані дані сприятливі до атак на узгодження записів та атрибутів. Ці атаки, в першу чергу, націлені на таблиці реляційних даних, де рядок представляє унікальну особу, а стовпець є атрибутом, який описує певний тип інформації про осіб у таблиці. Однак атаки можуть поширюватися на інші типи даних, такі як дані транзакцій [10].

### 1.3.1 Узгодження записів

Узгодження записів належить до типу атак на конфіденційність, які мають на меті однозначну ідентифікацію групи осіб у опублікованій таблиці даних за допомогою раніше зібраної інформації про цих осіб. Суїні [5] представив сценарій атаки, коли зловмисник отримує інформацію про групу осіб у опублікованій таблиці даних, використовуючи деяке зовнішнє джерело даних. Такі отриманні ззовні знання, які називаються фоновими знаннями зловмисника, потім використовуються для виділення (або звуження) запису цільової жертви в опублікованій таблиці даних. Поєднання певних ознак, відомих як квазі-

ідентифікатори (QID), у опублікованій таблиці потенційно може призвести до повторної ідентифікації певної групи осіб. Атрибути QID є загальними атрибутами між опублікованими даними та зовнішніми даними.

Приклад 1. Припустимо, лікарня публікує записи своїх пацієнтів у дослідницький центр для аналізу даних. У таблиці 1.1 показана опублікована необроблена таблиця після видалення явних ідентифікаторів. Нехай атрибути ‘Рік народження’, ‘Стать’ та ‘Робота’ є атрибутами QID у цьому сценарії. Припустимо, що зловмисник дізнався за допомогою анонімної зовнішньої таблиці даних про те, що його сусід Боб має такі значення щодо атрибутів QID {1955, Ч, Архітектор}. Така комбінація однозначно ідентифікує пацієнта в записі 1. Отже, нападник тепер знає, що запис 1 у таблиці 1.1 належить Бобу і що йому встановлено діагноз ВІЛ.

Таблиця 1.1 - Необроблена таблиця пацієнтів

| № | Рік народження | Стать | Робота | Діагноз |
|---|---|---|---|---|
| 1 | 1955 | Ч | Архітектор | ВІЛ |
| 2 | 1953 | Ч | Юрист | Грип |
| 3 | 1955 | Ч | Письменник | ВІЛ |
| 4 | 1951 | Ч | Художник | Грип |
| 5 | 1961 | Ж | Художник | Грип |
| 6 | 1965 | Ж | Письменник | ВІЛ |
| 7 | 1965 | Ж | Письменник | ВІЛ |

### 1.3.2 Узгодження атрибутів

Узгодження атрибутів – атака спрямована на те, щоб пов'язати конфіденційні фрагменти інформації з відповідними особами.. Видавець даних, атрибут вважає чутливим залежно від серйозності інформації, описаної цим атрибутом. Приклади чутливих ознак включають зарплату, хворобу та добробут. Зауважимо, що не обов'язково всі доменні значення чутливого атрибута є чутливими; можна вказати лише деякі значення домену як чутливі. Цінності, які вважаються чутливими, не повинні асоціюватися з відповідними особами. Наприклад, у таблиці 1.1 атрибут ‘Діагноз’ = {ВІЛ, Грип} вважається чутливим.

Однак у деяких прикладах ми вважатимемо ‘Грип’ нечутливим. Навіть якщо запис пацієнта не є унікальним у опублікованій таблиці, зловмисник може вивести хворобу цільової жертви. Ми називаємо ступінь визначеності зловмисника, виводячи чутливе значення жертви довірою до висновку.

Приклад 2. Розглянемо записи пацієнтів у таблиці 1.1. Припустимо, що зловмисник знає, що запис про цільового пацієнта, Аліса, містить {1965, Ж, Письменник}. Незважаючи на існування двох записів з однаковим набором значень над атрибутами QID, обидва ці записи мають ‘ВІЛ’ на чутливому атрибуті ‘Діагноз’. Тому зловмисник може зі 100% впевненістю зробити висновок про те, що Аліса має ВІЛ, тобто довіра до висновку зловмисника становить 100%.

### 1.3.3 Атака мінімальності

Як показано у вищезгаданих двох пунктах, якщо недостатньо уваги приділяється захисту конфіденційності людей, то їх опублікована таблиця даних залишається вразливою для атак на узгодження записів та атрибутів. В результаті, необроблена (оригінальна) таблиця даних повинна бути перетворена в анонімну версію, яка відповідає деяким вимогам конфіденційності і, таким чином, запобігає цим атакам. Цей процес називається анонімізацією.

Вонг та ін. [11] виявили, що навіть після анонімізації таблиці даних зловмисник все ще може пов'язувати чутливі значення з відповідним індивідом. Ми демонструємо запропоновану атаку атворами у наступному прикладі.

Приклад 3. Розглянемо просту версію даних про пацієнтів у таблиці 1.2, яка має ‘Стать’ та ‘Робота’ як атрибути QID. Нехай ‘ВІЛ’ є єдиною чутливою цінністю для хвороби. Під атакою на узгодження з атрибутами таблиця 1.2 у своїй необробленій версії із 100% впевненістю виявляє, що якщо цільова жертва є жінкою, письменницею, то вона має ВІЛ. Припустимо, ми хочемо знизити достовірність висновку до 70%; Отже, таблиця 1.2 анонімізована до таблиці 1.3 відповідно до набору дерев таксономії на рисунку 1.1.

Таблиця 1.2 - Необроблена таблиця пацієнтів

| № | Стать | Робота | Діагноз |
|---|---|---|---|
| 1 | Ч | Юрист | Грип |
| 2 | Ч | Юрист | Грип |
| 3 | Ч | Юрист | Грип |
| 4 | Ж | Письменник | ВІЛ |
| 5 | Ж | Письменник | ВІЛ |

Таблиця 1.3 містить найвищі загальні значення для всіх атрибутів QID, і це призводить до наявності лише однієї групи QID {М-Ж, Працюючий}, в якій зловмисник може зробити з максимум 40% впевненості висновок про те, що пацієнт має ВІЛ.

Таблиця 1.3 - Анонімізована версія таблиці 1.2

| № | Стать | Робота | Діагноз |
|---|---|---|---|
| 1 | Ч-Ж | Працюючий | Грип |
| 2 | Ч-Ж | Працюючий | Грип |
| 3 | Ч-Ж | Працюючий | Грип |
| 4 | Ч-Ж | Працюючий | ВІЛ |
| 5 | Ч-Ж | Працюючий | ВІЛ |

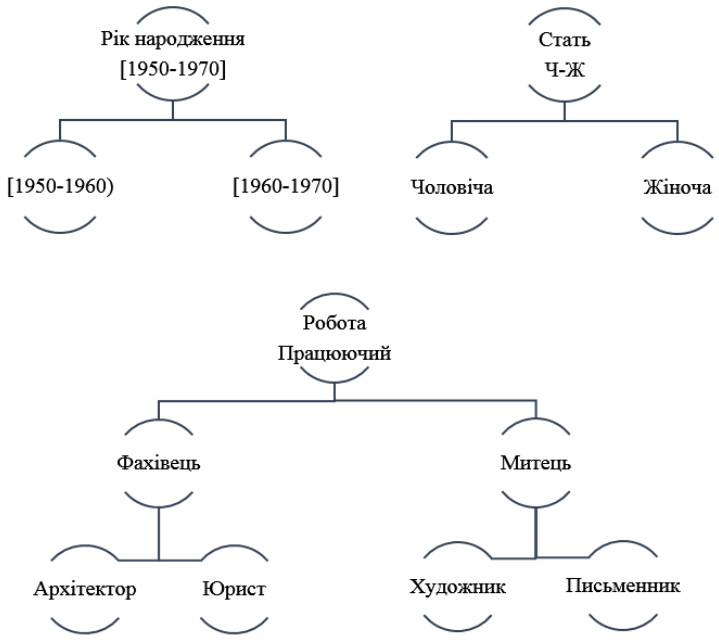


Рисунок 1.1 Дерева таксономії для даних пацієнтів

Оскільки таблиця 1.3 містить дуже загальну інформацію та відносно невелику достовірність логічного висновку, таблиця добре захищена від атак на узгодження атрибутів. Вонг та ін. [11] стверджував, що це не обов'язково може бути правдою. Їх міркування обгрунтовано розширює знання зловмисника не тільки на зовнішні таблиці даних, що містять ті ж атрибути QID, що і для анонімної таблиця, але також для алгоритму анонімізації, накладених вимог конфіденційності та набору дерев таксономії. Таким чином, у цьому прикладі зловмисник має такі відомості:

1. Зовнішня таблиця даних, аналогічна таблиці 1.2 мінус атрибут ‘Діагноз’ (для простоти, припустимо те саме впорядкування записів).

2. Вимога обмеження довіри до висновку про ВІЛ до 70%, оскільки це загальновідома інформація.

3. Опублікована анонімна таблиця, таблиця 1.3 .

4. Набір дерев таксономії, які зображено на рисунку 1.2.

З огляду на ці чотири відомості, зловмисник помічає, що опублікована таблиця, таблиця 1.3, була анонімізована. Таким чином, зловмисник легко виводить, що або група, що містить записи 1-3, порушує накладені вимоги конфіденційності, або група, що містить записи 4-5. Якщо в першій групі було два пацієнта з ВІЛ, то довіра до висновку цієї групи становила б 2/3 = 66,67%, що відповідає вимогам конфіденційності; отже, анонімізація не потрібна. Однак якщо друга група (записи 4-5) містить два значення ВІЛ, то довіра до висновку становить 100%. Зловмисник виводить, що анонімізація відбулася, оскільки в записах 4-5 осіб повинні бути пацієнти з ВІЛ; інакше анонімізація не потрібна.

Наведений вище приклад описує типовий сценарій атаки мінімальності [11]. Цей тип атак використовує той факт, що існуючі алгоритми анонімізації дотримуються принципу мінімальності при перетворенні таблиці даних [12]. За принципом мінімальності, необроблена таблиця даних перетворюється на її найменш анонімну версію. Це тому, що анонімізація таблиці даних накладає втрату інформації на вихідні значення даних; таким чином, для запобігання подальших спотворень не проводиться додаткова анонімізація на вже анонімній таблиці, яка задовольняє деяким заданим вимогам конфіденційності.

## 1.4 Моделі конфідеційності

Запропоновано кілька моделей, які пропонують формальні гарантії щодо захисту конфіденційності особи. Ці моделі були розроблені з урахуванням різних сценаріїв атак на дані. Наприклад, якщо припустити, що зловмисники мають різні рівні базових знань, що може призвести до розкриття інформації. Прикладами відомих моделей є *k*-анонімність [13], `ℓ-різноманітніття [14], та диференційна конфіденційність [15].

### 1.4.1 *k*-Анонімізація

Це була перша модель, запропонована для анонімізації мікроданих, і це основа, з якої були розроблені подальші розширення Ця модель конфіденційності зазначає, що кожен запис у опублікованій таблиці даних повинен відрізнятися щонайменше від *k*-1 інших записів за атрибутами QID. Отже, максимальна ймовірність успішної атаки узгодження записів пов'язана з 1/*k*. Група QID – це набір записів, які мають однакові значення для атрибутів QID. Таким чином, *k*-анонімізація приховує унікальний запис у групі QID розміром ≥ *k*. Міцність *k*-анонімізації покладається на *k* та атрибути QID: більш високі значення *k* мінімізують ймовірність успішної атаки, а більше QID-атрибутів забезпечує кращий захист від ширшого кола знань.

Наприклад таблиця 1.4 – анонімна версія таблиці 1.1, де *k*=2. У цьому випадку ми говоримо, що таблиця 1.4 є 2-анонімною, оскільки кожен запис у повному обсязі (над атрибутами QID) відображається у мінімум 2 рази в таблиці. Зауважимо, що таблицю 1.4 було досягнуто шляхом узагальнення деяких значень атрибутів у таблиці 1.1 відповідно до набору дерев таксономії на рисунку 1.2.

Таблиця 1.4 - Анонімізована версія таблиці 1.1 ($k$=2)

| № | Рік народження | Стать | Робота | Діагноз |
|---|---|---|---|---|
| 1 | [1950-1960) | Ч | Фахівець | ВІЛ |
| 2 | [1950-1960) | Ч | Фахівець | Грип |
| 3 | [1950-1960) | Ч | Митець | ВІЛ |
| 4 | [1950-1960) | Ч | Митець | Грип |
| 5 | [1960-1970) | Ж | Митець | Грип |
| 6 | [1960-1970) | Ж | Митець | ВІЛ |
| 7 | [1960-1970) | Ж | Митець | ВІЛ |

### 1.4.2 ℓ -Різноманітність

Для запобігання атаки на узгодження атрибутів Machanavajjhala та ін. [14] запропонував поняття ℓ -різноманіття. У різноманітній таблиці ℓ кожна група QID повинна мати принаймні ℓ «добре представлених» записи стосовно чутливих значень. ℓ-різноманіття – це концепція конфіденційності, яку можна реалізувати більш ніж одним способом. Найпростіша інтерпретація "добре представлених" полягає в тому, щоб кожна група QID у таблиці мала мінімум ℓ чутливих значень. Розглянемо таблицю 1.4, яка є 2-анонімною версією таблиці 1.1. Таблиця 1.4 містить 3 групи QID. Припустимо, що всі доменні значення чутливого атрибута ‘Діагноз’ є чутливими і не повинні асоціюватися зі своїми відповідними пацієнтами. Оскільки кожна група QID у таблиці 1.4 містить два різних чутливих значення, як вважається, таблиця 1.4 є 2-різноманітною.

Інша інтерпретація моделі ℓ -різноманіття є ентропія ℓ -різноманіття, яка вимірює, наскільки рівномірно розподілені чутливі значення в таблиці. Більш рівномірно розподілений чутливий атрибут між табличними записами передбачає більшу невизначеність у виведенні чутливих значень про осіб. Навпаки, менш рівномірно розподілений чутливий атрибут означає, що в таблиці більше значень, що трапляються, ніж інші, таким чином, даючи зловмиснику більший шанс вивести ці часті значення. Якщо таблиця розділена на групи QID, то ентропія ℓ -різноманіття застосовується до чутливих значень у межах кожної групи QID.

### 1.4.3 Диференційна конфідеційність

Усі вищезазначені моделі конфіденційності покладаються на оцінку фонових знань зловмисника, щоб запобігти атакам на узгодження записів та атрибутів. Такий тип моделей конфіденційності називають синтаксичними моделями, оскільки анонімні дані повинні відповідати деяким синтаксичним умовам. Щоб усунути дію сили будь-якого зловмисника і уникнути розробки моделей конфіденційності із захисною силою, прив'язаною до такої влади, Дворк запропонував диференційну конфіденційність [15]. Диференційна конфіденційність – це імовірнісна модель конфіденційності, яка працює незалежно від фонових знань та обчислювальної сили будь-якого зловмисника. У цьому дусі диференційна конфіденційність обмежує ймовірність отримання однакової відповіді з двох різних наборів вхідних даних, D і D’, які відрізняються лише одним записом.

# РОЗДІЛ 2 МЕТОДИКА ПОРІВНЯННЯ АЛГОРИТМІВ K-АНОНІМІЗАЦІЇ

## 2.1 Методи анонімізації

Алгоритм анонімізації може використовувати різні методи для досягнення бажаного рівня конфіденційності. Якщо метою – збереження правдивості даних, то найбільш вдалим механізмом є детерміновані механізми [16]. У літературі запропоновано кілька методів анонімізації; у цьому розділі ми зупинимось на двох основних техніках: приховування та узагальнення [17].

### 2.1.1 Приховування

Приховування полягає у заміні деяких вихідних даних спеціальним значенням (наприклад, "*"), щоб вказати, що ці дані не розкриваються. Приховування виконується різними способами, залежно від бажаного балансу між корисністю анонімних даних та складністю застосованого алгоритму. Приховування значення [18] видаляє всі екземпляри значення, яке слід приховати з таблиці даних. Локальне приховування [19], яке також називається приховуванням комірок, з іншого боку, може зберігати деякі екземпляри значення, яке слід приховати з таблиці даних. Інтуїтивно, місцеве приховування несе менші втрати інформації, ніж приховування значення; однак колишній тип поставляється ціною високої обчислювальної складності. Приховування записів видаляє цілі записи.

### 2.1.2 Узагальнення

Узагальнення (також зване перекодування) полягає у заміні значень атрибута менш конкретними, але послідовними значеннями; часто використовують ієрархію узагальнення значень (VGH), таку як показано на рисунку 2.1. Значення на нижньому рівні (праворуч) знаходяться в основній області атрибута, що відповідають найбільш конкретним значенням (вихідні значення). Найвищий рівень (зліва), що показує значення «*», відповідає

максимальному узагальненню або повному придушенню значення. Перекодування може бути здійснено за глобальною (повноцінне узагальнення) або локальною схемою. Локальне перекодування може застосовувати різні правила до одних і тих же екземплярів атрибутів таким чином, що одні екземпляри залишаються з певними значеннями, а інші узагальнюються. Навпаки, глобальне перекодування полягає у застосуванні одного і того ж узагальнення до всіх екземплярів атрибута, таким чином, що всі значення узагальнюються до одного рівня VGH. Глобальне перекодування далі класифікується на два типи: одновимірне [12, 20], яке розглядає кожен атрибут групи QID незалежно; і багатовимірність [21], яка перекодує домен n-векторів, які є поперечним добутком доменів окремих атрибутів QID [16].

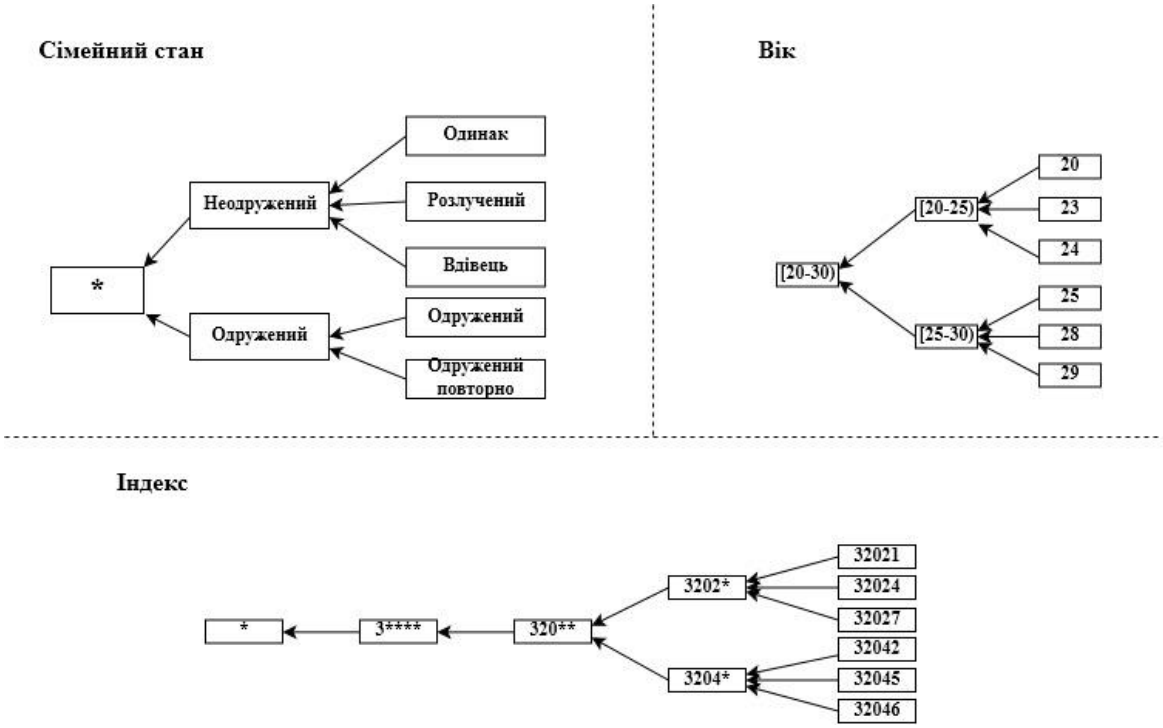


Рисунок 2.1 - Узагальнені значення VGH для сімейного стану, віку та індексу

Таблиця 2.1 - Мікродані судимості

| | ID | QIDs | | | SA |
|---|---|---|---|---|---|
| **№** | **Ім’я** | **Сімейний стан** | **Вік** | **Індекс** | **Злочин** |
| 1 | Джой | Розлучений | 29 | 32042 | Вбивство |
| 2 | Джим | Одиинак | 20 | 32021 | Крадіжка |
| 3 | Сью | Вдівець | 24 | 32024 | Продаж наркотиків |
| 4 | Еб | Розлучений | 28 | 32046 | Напад |
| 5 | Боб | Вдівець | 25 | 32045 | Піратство |
| 6 | Емі | Одинак | 23 | 32027 | Непорядність |

В якості наочного прикладу розглянемо таблицю 2.1, що показує таблицю судимості. Серед атрибутів ім'ям є ідентифікатор (ID), сімейний стан, вік та поштовий індекс – QID; а злочин – чутливий атрибут (SA). У таблиці 2.2 показано 3-анонімну версію таблиці 2.1, що означає, що кожен кортеж має щонайменше два інших кортежі, що мають однакові значення QID. Для досягнення анонімності ідентифікатор видалено, а QID узагальнено за допомогою одновимірної схеми: ‘Сімейний стан’ замінено на менш конкретний, але семантично послідовний опис, вікові значення замінено діапазонами значень та останньою цифрою поштового індексу було замінено на "*".

Таблиця 2.2 - Версія 3-анонімної Таблиці 1 (*k*=3)

| QIDs | | | | | SA |
|---|---|---|---|---|---|
| **№** | **EQ** | **Сімейний стан** | **Вік** | **Індекс** | **Злочин** |
| 1 | 1 | Неодружений | [25:30) | 3204* | Вбивство |
| 4 | | Неодружений | [25:30) | 3204* | Напад |
| 5 | | Неодружений | [25:30) | 3204* | Піратство |
| 2 | 2 | Неодружений | [20:25) | 3202* | Крадіжка |
| 3 | | Неодружений | [20:25) | 3202* | Продаж наркотиків |
| 6 | | Неодружений | [20:25) | 3202* | Непорядність |

## 2.2 Алгоритми k-анонімізації

Існує велика кількість алгоритмів *k*-анонімізації, запропонованих у літературі в різних областях конфіденційності даних: PPDP, збереження конфіденційності, обмін даними (PPDM) та контроль статистичного розкриття інформації (SDC). У роботі [17] запропоновано підхід із використанням генетичного алгоритму, який спрямований на збереження класифікаційної інформації в анонімізованих даних. Ітераційний алгоритм узагальнення знизу представлений в [22], пропонуючи мінімальну *k*-анонімізацію для класифікації. kACTUS [23] - ще один алгоритм *k*-анонімізації, орієнтований на збереження конфіденційності в класифікаційних завданнях з використанням багатовимірного приховування. Спеціалізація зверху вниз [24] – це алгоритм, який виходить із найбільш узагальненого стану таблиці та спеціалізує його відповідно до метрики пошуку, пропонуючи мінімальну *k*-анонімізацію. *k*-optimize [25] пропонує оптимальну анонімність із використанням узагальнення підрівнів та приховування запису методами обрізки. В області SDC широко використовуються методи, що використовують методи мікроагрегації для анонімізації. Деякі з найбільш релевантних методів – це метод максимальної відстані до середнього векторного методу (MDAV) [26], багатоваріантна мікроагрегація фіксованого розміру [27], мінімальна розбивка дерев [28], загальна MDAV [29] та MDAV змінної величини [30].

У порівняльному дослідженні вибрано три алгоритми *k*-анонімізації з використанням узагальнення та приховування. Їх вибрали, виходячи з таких причин:

- ці алгоритми широко цитуються в літературі;
- ці алгоритми використовують різні стратегії анонімізації, що дозволяють отримати більш всебічну оцінку;
- доступна публічна реалізація цих алгоритмів;
- ці алгоритми можуть бути оцінені в одних і тих же рамках, що забезпечує більш справедливе порівняння.

2.2.1 Datafly алгоритм

Datafly [20] - жадібний евристичний алгоритм, який виконує одновимірне повноцінне узагальнення. На рисунку 2.2 зображені основні етапи алгоритму Datafly. Він підраховує частоту для набору QID, і якщо *k*-анонімність ще не задоволена, він узагальнює атрибут, що має найрізноманітніші значення, поки *k*-анонімність не буде задоволена. Якщо цей алгоритм гарантує *k*-анонімне перетворення, він не забезпечує мінімального узагальнення [31].

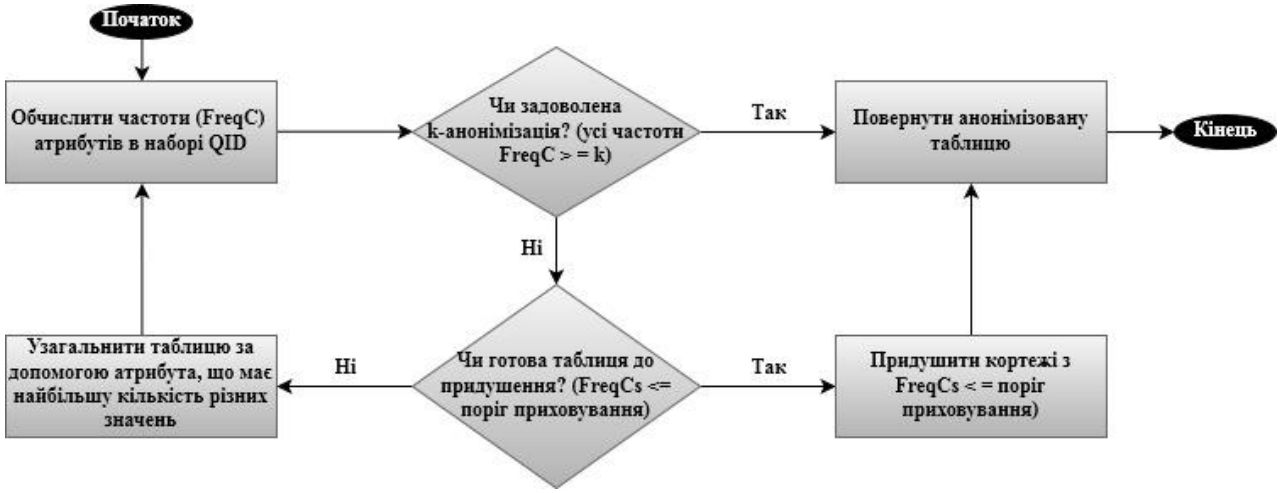


Рисунок 2.2 - Основний процес алгоритму Datafly

Розглянемо застосування алгоритму на прикладі. Ми показуємо посередницькі анонімізації, що виконуються з даними (тобто ітераціями) та отриманим кінцевим рішенням. Вхідними параметрами для цієї анонімізації є: *k*=2, поріг приховування=0, QIDs={Сімейний стан, Вік, Індекс}. Узагальнення зображено на рисунку 2.1.

У першій ітерації обчислимо частоту наборів QID таблиці. Бачимо, що таблиця 2.3 не задовільняє *k*-значення, тобто так і залишилося 6 кортежів. У атрибуті ‘Сімейний стан’ 3 чітких значення, у ‘Вік’ – 6, ‘Індекс’ – 6. Узагальнюємо таблицю за допомогою атрибута ‘Вік’.

Таблиця 2.3 - Перша ітерація алгоритму Datafly

| QIDs | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Розлучений | 29 | 32042 | 1 |
| Одинак | 20 | 32021 | 1 |
| Вдівець | 24 | 32024 | 1 |
| Розлучений | 28 | 32046 | 1 |
| Вдівець | 25 | 32045 | 1 |
| Одинак | 23 | 32027 | 1 |

У другій ітерації знову обчислимо частоту наборів QID таблиці. Бачимо, що таблиця 2.4 не задовільняє *k*-значення, тобто так і залишилося 6 кортежів. У атрибуті ‘Сімейний стан’ 3 чітких значення, у ‘Вік’ – 2, ‘Індекс’ – 6. Узагальнюємо таблицю за допомогою атрибута ‘Індекс’.

Таблиця 2.4 - Друга ітерація алгоритму Datafly

| QIDs | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Розлучений | [25:30) | 32042 | 1 |
| Одинак | [20:25) | 32021 | 1 |
| Вдівець | [20:25) | 32024 | 1 |
| Розлучений | [25:30) | 32046 | 1 |
| Вдівець | [25:30) | 32045 | 1 |
| Одинак | [20:25) | 32027 | 1 |

У третій ітерації знову обчислимо частоту наборів QID таблиці. Бачимо, що таблиця 2.5 не задовільняє *k*-значення залишилося 2 кортежа. У атрибуті ‘Сімейний стан’ 3 чітких значення, у ‘Вік’ – 2, ‘Індекс’ – 2. Узагальнюємо таблицю за допомогою атрибута ‘Сімейний стан’.

Таблиця 2.5 - Третя ітерація алгоритму Datafly

| QIDs | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Розлучений | [25:30) | 3204* | 2 |
| Одинак | [20:25) | 3202* | 2 |
| Вдівець | [20:25) | 3202* | 1 |
| Вдівець | [25:30) | 3204* | 1 |

У четвертій ітерації обчислимо частоту наборів QID таблиці. Таблиця 2.6 задовільняє *k*-значення і утворюємо анонімізовану таблицю (таблиця 2.7).

Таблиця 2.6 - Четверта ітерація алгоритму Datafly

| **QIDs** | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Неодружений | [25:30) | 3204* | 3 |
| Неодружений | [20:25) | 3202* | 3 |

Таблиця 2.7 - Анонімізована таблиця алгоритмом Datafly

| **QIDs** | | | | | **SA** |
|---|---|---|---|---|---|
| **№** | **EQ** | **Сімейний стан** | **Вік** | **Індекс** | **Злочин** |
| 1 | 1 | Неодружений | [25:30) | 3204* | Вбивство |
| 4 | | Неодружений | [25:30) | 3204* | Напад |
| 5 | | Неодружений | [25:30) | 3204* | Піратство |
| 2 | 2 | Неодружений | [20:25) | 3202* | Крадіжка |
| 3 | | Неодружений | [20:25) | 3202* | Продаж наркотиків |
| 6 | | Неодружений | [20:25) | 3202* | Непорядність |

Очевидно, що k-анонімізація проведена з параметром k=3.

### 2.2.2 Incognito алгоритм

Incognito [12] - це одновимірний алгоритм узагальнення повного домену, який будує решітку узагальнення та обводить її за допомогою пошуку знизу вгору. На рисунку 2.3 зображені основні етапи алгоритму Incognito.

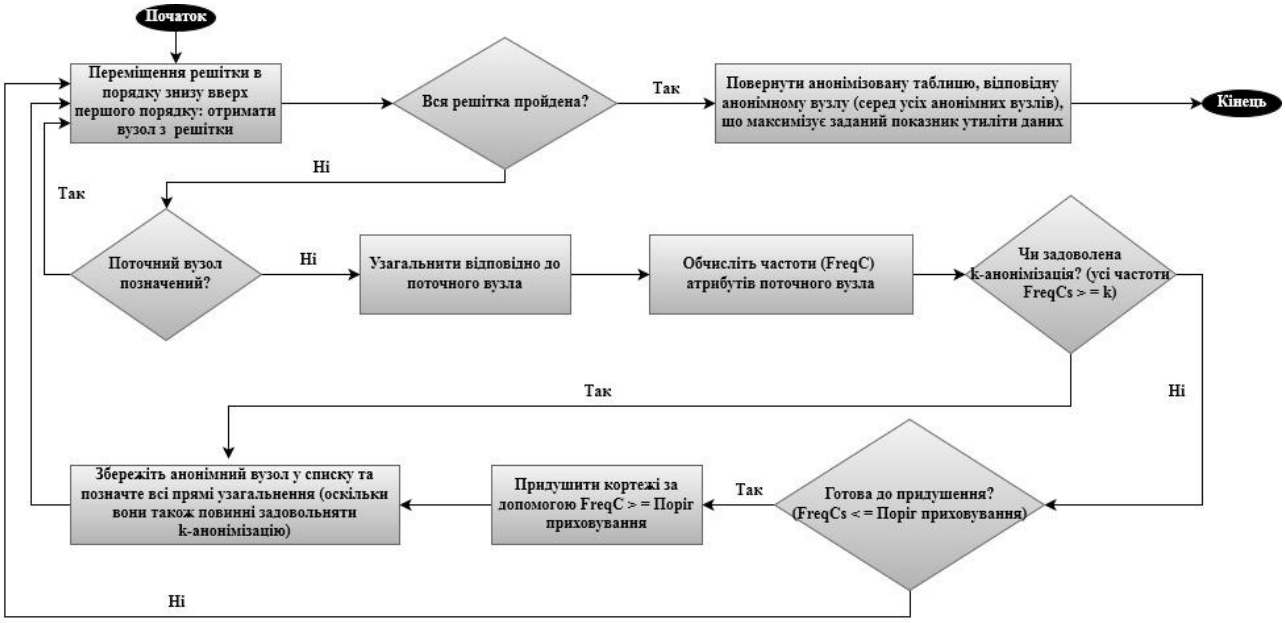


Рисунок 2.3 - Основні етапи алгоритму Incognito

Кількість дійсних узагальнень для атрибута визначається глибиною його VGH. Враховуючи VGH на рисунку 2.1, це: два для сімейного стану (M) та віку (A); чотири для поштового індексу (Z). Приклад узагальнюючої решітки, створеної для цього набору QID, можна побачити на рисунку 2.4.

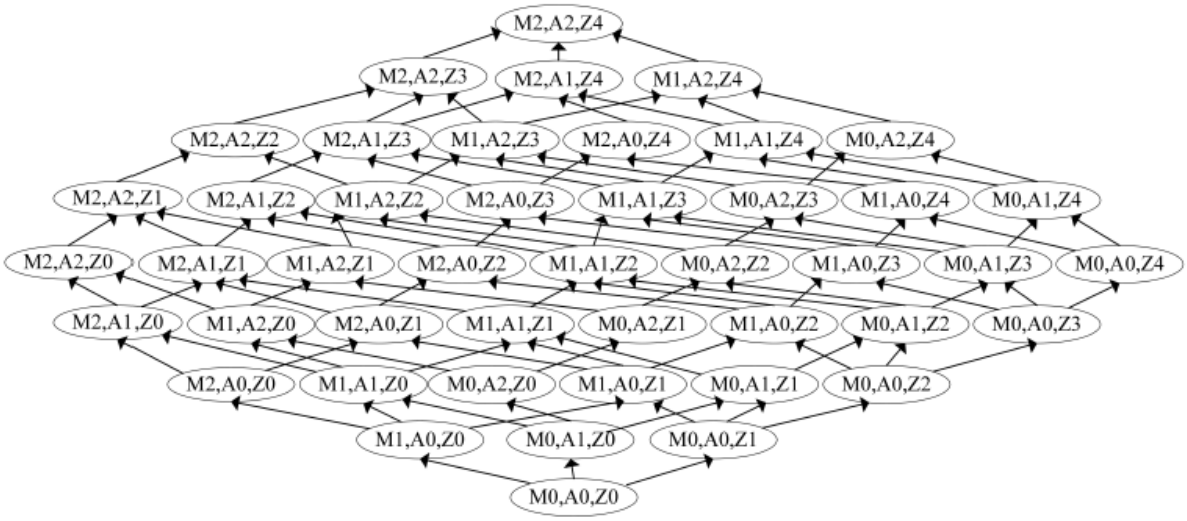


Рисунок 2.4 - Узагальнююча решітка для набору QID

Кожен вузол решітки представляє рішення для анонімізації. Наприклад, вузол <M1, A1, Z1> означає, що три QID були узагальнені один раз (на один рівень вище в VGH), що є рішенням, показаним у таблиці 2.2. Для анонімізації

даних, Incognito використовує прогнозування для зменшення пошуковий простір. Це означає, що, обходячи решітку, якщо буде виявлено, що вузол задовольняє *k*-анонімність, всі його прямі узагальнення можна підрізати, оскільки гарантується, що вони також задовольняють *k*-анонімність. На відміну від Datafly, Incognito виробляє оптимальне рішення. Це означає, що анонімізоване рішення містить максимальну кількість інформації відповідно до обраної інформаційної метрики. Остаточне рішення вибирається з набору рішень, які всі відповідають заданій вимозі конфіденційності (наприклад, *k*). Під час оцінки, яку ми оцінюємо, Incognito вибирає рішення, яке дає максимальну кількість еквалайзерів (класів нквівалентності).

Розглянемо приклад того, як виконується анонімізація за допомогою Incognito. . Вхідними параметрами для цієї анонімізації є: *k*=2, поріг приховування=0, QIDs={Сімейний стан, Вік, Індекс}.

Вибираємо вузол M0, A1, Z0. Він є неанонімним, оскільки тільки один атрибут анонімізований. Обчислюємо частоту атрибутів у вузлі. Вузол не задовільняє *k*-значення (таблиця 2.8).

Таблиця 2.8 - Перший етап алгоритму Incognito

| **QIDs** | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Розлучений | [25:30) | 32042 | 1 |
| Одинак | [20:25) | 32021 | 1 |
| Вдівець | [20:25) | 32024 | 1 |
| Розлучений | [25:30) | 32046 | 1 |
| Вдівець | [25:30) | 32045 | 1 |
| Одинак | [20:25) | 32027 | 1 |

Вибираємо наступний вузол M1, A1, Z1. Він анонімний. Далі обчислюємо частоту атрибутів у вузлі. Вузол задовільняє *k*-значення (таблиця 2.9).

Таблиця 2.9 - Другий етап алгоритму Incognito

| QIDs | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Неодружений | [25:30) | 3204* | 3 |
| Неодружений | [20:25) | 3202* | 3 |

Вибираємо вузол M0, A2, Z2. Він анонімний. Далі обчислюємо частоту атрибутів у вузлі. Вузол задовільняє *k*-значення (таблиця 2.10).

Таблиця 2.10 - Третій етап алгоритму Incognito

| QIDs | | | |
|---|---|---|---|
| **Сімейний стан** | **Вік** | **Індекс** | **FreqC** |
| Розлучений | [20:30) | 320** | 2 |
| Одинак | [20:30) | 320** | 2 |
| Вдівець | [20:30) | 320** | 2 |

Після проходження всієї решітки вибираємо вузол (стан узагальнення), який максимізує заданий показник утиліти даних (наприклад, той, який дає максимальну кількість еквалайзерів)(таблиця 2.11).

Таблиця 2.11 - Анонімізована таблиця алгоритмом Incognito

| QIDs | | | | | SA |
|---|---|---|---|---|---|
| **№** | **EQ** | **Сімейний стан** | **Вік** | **Індекс** | **Злочин** |
| 1 | 1 | Розлучений | [20:30) | 320** | Вбивство |
| 4 | | Розлучений | [20:30) | 320** | Напад |
| 2 | 2 | Одинак | [20:30) | 320** | Крадіжка |
| 6 | | Одинак | [20:30) | 320** | Непорядність |
| 3 | 3 | Вдівець | [20:30) | 320** | Продаж наркотиків |
| 5 | | Вдівець | [20:30) | 320** | Піратство |

### 2.2.3 Mondrian алгоритм

Mondrian [21] - жадібний багатовимірний алгоритм, який розділяє доменний простір рекурсивно на декілька регіонів, кожна з яких містить щонайменше *k* записів.

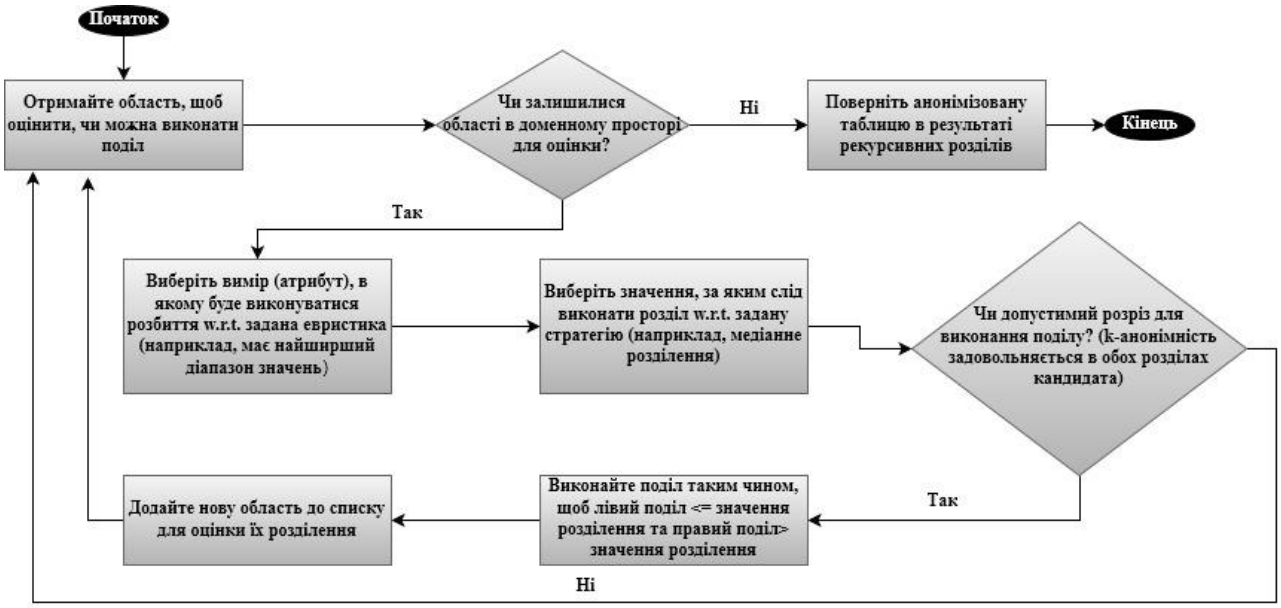


Рисунок 2.5 - Основні етапи алгоритму Mondrian

На рисунку 2.5 зображені основні етапи алгоритму Mondrian. Він починається з найменш конкретного (найбільш узагальненого) значення атрибутів у наборі QID та спеціалізується на тому, що поділ виконуються на даних. Щоб вибрати розмірність (тобто атрибут), за якою виконати поділ, Mondrian використовує атрибут з найширшим (нормалізованим) діапазоном значень. Якщо декілька розмірів мають однакову ширину, вибирається перший, який дозволяє допустимий зріз (тобто зріз не викликає порушення *k*-анонімності). Після вибору розміру, Mondrian використовує підхід середнього розподілу для вибору значення розділення, значення, за яким буде виконуватися поділ. Для пошуку медіани для атрибута використовується підхід набору частот [21]. Це означає, що дані скануються, додаючи частоти для кожного унікального значення в атрибуті, поки не буде знайдено медіанне положення. Значення, за яким знайдено медіану, стає розділеним значенням.

Розглянемо приклад анонімізації таблиці за допомогою Mondrian. У цьому прикладі ми графічно показуємо двовимірне подання значень QID; виникнення значення представляється у вигляді точки. Ми також представляємо деякі рекурсивні розділи, виконані з отриманими даними та кінцевим рішенням. Для більш чіткого подання ми вибрали лише 2 (з 3) атрибутів як QID. Вхідними

параметрами для цієї анонімізації є: *k*=2, поріг приховування=0, QID={Вік, Сімейний статус}. Евристика, використана в цьому прикладі, для того щоб знати, де з'являються перегородки, це допомагає вибрати розмірність із найширшим (нормалізованим) діапазоном значень. Аналогічно, стратегія вибору значення розділення полягає у виконанні медіанного розподілу.

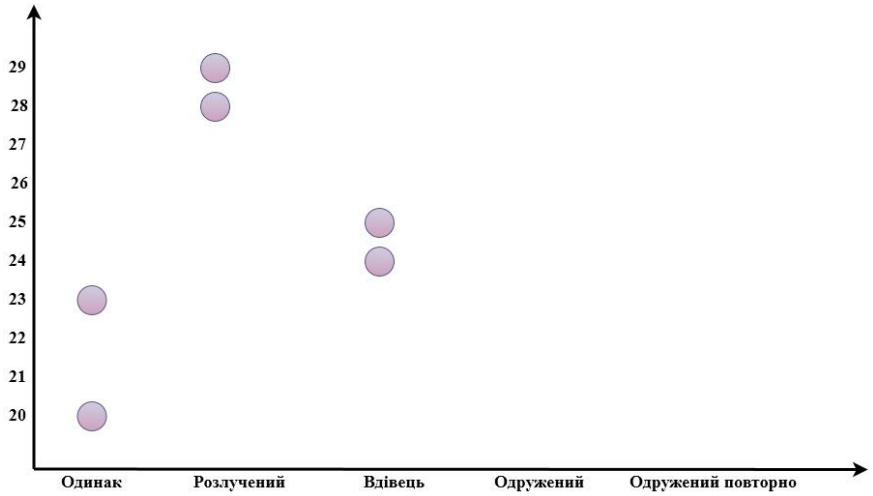


Рисунок 2.6 - Просторове представлення вихідних даних

На рисунку 2.6 зображено просторове представлення вхідних даних. У першій ітерації область що підлягає оцінці [20-29] [Одинак-Одружений повторно]. Вибираємо розмір для виконання розділу. Нормований діапазон значень 'Вік' -1, 'Сімейний стан' -1. Обидва мають однаковий діапазон, вибираємо 'Сімейний стан'. Далі вибираємо значення поділу. Середнє для 'Сімейного стан' – 'Розлучений', і це допустимий розріз. Отримані поділи – лівий [Одинак-Розлучений], правий (Розлучений-Повторно одружений]. Результат першого поділу зображено на рисунку 2.7.

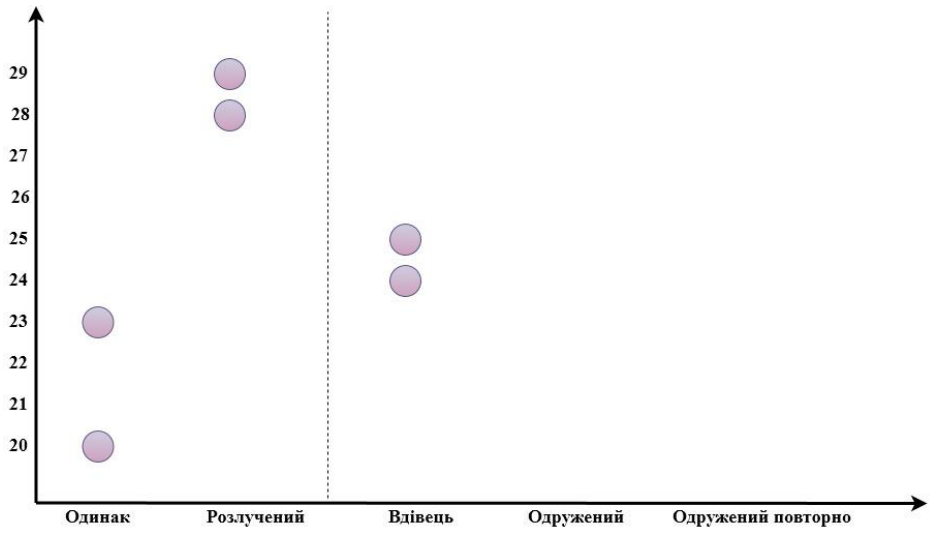


Рисунок 2.7 - Просторове представлення після першого поділу

У другій ітерації область що підлягає оцінці [20-29] [Одинак-Розлучений]. Вибираємо розмір для виконання розділу. Нормований діапазон значень – ‘Вік’ 1, ‘Сімейний стан’ 0,2 і обираємо ‘Вік’. Далі вибираємо значення розділення. Середнє для ‘Вік’, знайдене в ’23’, і це допустимий розріз. Отримані поділи – лівий [20-23], правий (23-29]. Результат другого поділу зображено на рисунку 2.8.

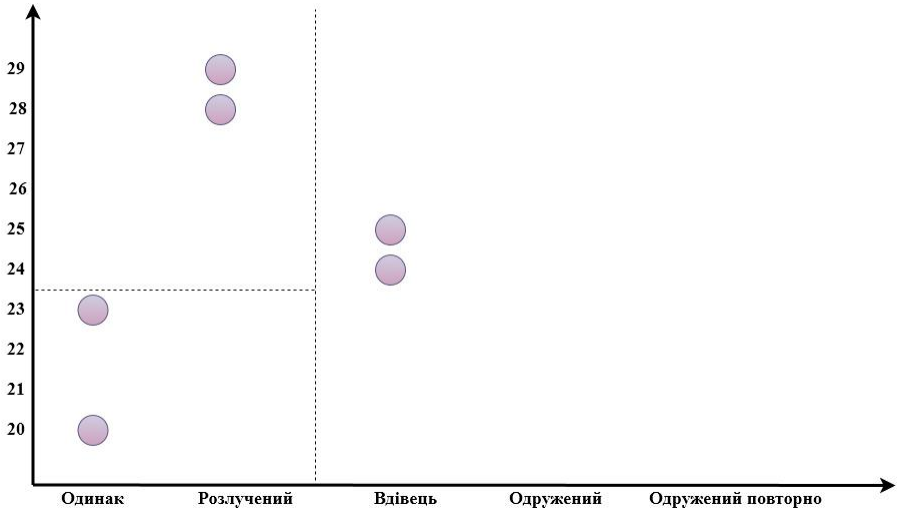


Рисунок 2.8 - Просторове представлення після другого поділу

У третій ітерації область що підлягає оцінці [20-29] (Розлучений-Повторно одружений]. Середнє значення категорії ‘Вік’ рівне ‘24’, але це не допустимий

розріз. Середнє значення для ‘Сімейний стан’ знайдено в ‘Вдівець’, але це також не допустимий розріз. Отже, немає допустимих скорочень для цієї області даних.

У четвертій ітерації область що підлягає оцінці [20-23] [Одинак-Розлучений]. Середнє значення для ‘Вік’ знайдено ‘20’, але це не допустимий розріз. Середнє значення для ‘Сімейний стан’ знайдено в ‘Одинак’, але це також не допустимий розріз. Отже, немає допустимих скорочень для цієї області даних.

У п’ятій ітерації область що підлягає оцінці (23-29] [Одинак-Розлучений]. Середнє значення для ‘Вік’ знайдено в ‘28’, але це не допустимий розріз. Середнє значення для ‘Сімейний стан’ знайдено в ‘Розлучений’, але це також не допустимий розріз. Отже, немає допустимих скорочень для цієї області даних.

Оскільки скорочень більше немає, формуємо анонімізовану таблицю 2.12.

Таблиця 2.12 - Анонімізована таблиця алгоритмом Mondrian

| **QIDs** | | | | | **SA** |
|---|---|---|---|---|---|
| **№** | **EQ** | **Сімейний стан** | **Вік** | **Індекс** | **Злочин** |
| 1 | 1 | Одинак, Розлучений | (23:30) | 32042 | Вбивство |
| 4 | | Одинак, Розлучений | (23:30) | 32046 | Напад |
| 2 | 2 | Одинак, Розлучений | [20:23] | 32021 | Крадіжка |
| 6 | | Одинак, Розлучений | [20:23] | 32027 | Непорядність |
| 3 | 3 | Вдівець - Одружений пов. | [20:30) | 32024 | Продаж наркотиків |
| 5 | | Вдівець - Одружений пов | [20:30) | 32045 | Піратство |

Отримані анонімізовані дані (k=3) можна використати для публікації.

## 2.3 Набір даних

### 2.3.1 Реальний набір даних

Перший набір даних – це набір даних перепису для дорослих осіб із сховища машинного навчання UCI [32], який став еталонним показником для

оцінки алгоритмів *k*-анонімізації. Ми підготували цей набір даних, видаливши записи з пропущеними значеннями, таким чином залишивши 30 162 дійсних записів, як у [17, 21, 33, 34]. Опис набору даних для ‘Дорослі особи’ наведено в таблиці 2.13. У цій таблиці представлені атрибути, потужність множини їх значень (кількість різних значень) та узагальнення VGH, визначені для кожного атрибута. У цьому останньому стовпчику ми вказуємо у дужках кількість рівнів VGH (тобто наявних узагальнень) кожного конкретного атрибута.

Таблиця 2.13 - Набір даних ‘Дорослі особи’

| № | Атрибут | Потужність множини | Узагальнення |
|---|---|---|---|
| 1 | Вік | 74 | Дерево таксономії (4) у 5-,10-, 20-річному діапазоні |
| 2 | Стать | 2 | Дерево таксономії (1) |
| 3 | Раса | 5 | Дерево таксономії (1) |
| 4 | Матеріальний стан | 7 | Дерево таксономії (2) |
| 5 | Рідна країна | 41 | Дерево таксономії (2) |
| 6 | Робочий клас | 8 | Дерево таксономії (2) |
| 7 | Вид діяльності | 14 | Дерево таксономії (2) |
| 8 | Освіта | 16 | Дерево таксономії (3) |
| 9 | Клас заробітньої плати | 2 | Дерево таксономії (1) |

### 2.3.2 Синтетичний набір даних

Другий набір даних – це ‘Ірландський’ набір даних, який ми синтетично генерували за допомогою Benerator [35], інструменту з відкритим кодом Java. Цей набір даних був створений з використанням розподілів частоти під час перепису в Ірландії 2011 року [36] як вагових коефіцієнтів у процесі генерації даних. Більш детальну інформацію про цей процес генерації можна знайти в розділі 4. Оригінальний набір даних перепису Ірландії складався з 3,550,246 записів, і він був зменшений до різних розмірів: 5k, 10k, 20k, 30k, 50k та 100k записів. На рисунку 6 показано порівняння між розподілом від вихідних даних перепису Ірландії та синтетичним набором даних із 500k записів за віковим атрибутом. Видно, що синтетичні дані зберігають високий рівень точності

порівняно з початковим розподілом. Опис цих наборів даних показано в таблиці 2.14, яка має таку ж структуру, як і раніше описана таблиця наборів даних для дорослих осіб.

Таблиця 2.14 - Набір даних ‘Ірландський’

| № | Атрибут | Потужність множини | Узагальнення |
|---|---|---|---|
| 1 | Вік | 70 | Дерево таксономії (4) у 5-,10-, 20-річному діапазоні |
| 2 | Стать | 2 | Дерево таксономії (1) |
| 3 | Країна | 34 | Дерево таксономії (3) |
| 4 | Матеріальний стан | 7 | Дерево таксономії (2) |
| 5 | Рідна країна | 56 | Дерево таксономії (3) |
| 6 | Економічний статус | 9 | Дерево таксономії (3) |
| 7 | Промислова група | 22 | Дерево таксономії (2) |
| 8 | Освіта | 10 | Дерево таксономії (3) |
| 9 | Ступінь освіти | 48 | Дерево таксономії (2) |

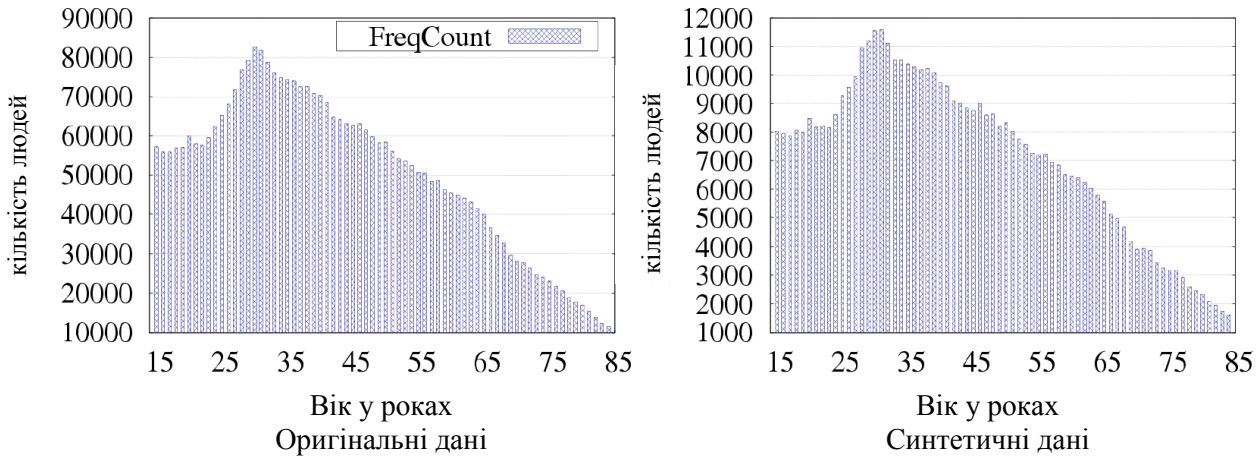


Рисунок 2.9 - Розподіл даних для оригінальних та синтетичних наборів даних за віковим атрибутом

## 2.4 Методологія порівняння

Проведення справедливого порівняння алгоритмів анонімізації по суті є складним завданням, оскільки кожен запропонований алгоритм використовує різні метрики та налаштування. Продуктивність алгоритмів може відрізнятися між різними комбінаціями наборів даних та вхідних параметрів (наприклад, алгоритм може добре працювати в деяких експериментальних конфігураціях та погано працювати в інших). Як результат, важливо оцінити алгоритми, визначивши загальну конфігурацію, яка відображає параметри, що використовуються в існуючих оцінках. Крім того, порівняння вимагає використання критеріїв, які можуть бути широко застосовними для вимірювання різних аспектів алгоритмів (наприклад, ефективності та корисності даних). Як раніше обговорювалося в розділі 1, в GPDP всі потенційні сценарії використання часто невідомі. Таким чином, з точки зору корисності даних ми сконцентрувались на метриках, які можна застосувати до декількох сценаріїв (тобто загальних показників) та до декількох типів алгоритмів. Нижче ми описуємо показники, використані в нашій методології порівняння.

### 2.4.1 Ефективність алгоритму

Алгоритм повинен оцінюватися ресурсами, необхідними для проведення анонімізації. Це важливий аспект, оскільки процес анонімізації може бути інтенсивним у споживанні ресурсів. Якщо ресурси обмежені, вони представляють обмеження у виборі алгоритму. Навіть коли алгоритм досягає хорошого рівня корисності анонімізованих даних, якщо він не є ефективним щодо споживання пам’яті, використання процесора або часу виконання, він може не бути практичним для використання.

Для вимірювання часу виконання можна було відслідковувати минулий час для різних етапів процесу анонімізації: Завантаження даних (або в пам'ять, або в базу даних), сама анонімізація та вихід даних. Оскільки кроки завантаження та виведення даних не змінюються серед алгоритмів, ми зупинимося лише на вимірюванні часу анонімізації.

Для аналізу продуктивності алгоритмів щодо часу анонімізації ми вивчаємо залежність між часом анонімізації та вартістю наступних функціональних характеристик алгоритмів: для Datafly – кількість виконуваних операцій узагальнення; для Incognito – кількість вузлів або станів узагальнення, оцінені у створеній решітці; а для Mondrian – кількість розділів, виконаних за даними.

Що стосується обчислювальних ресурсів, то пам'ять та використання процесора також є хорошими показниками ефективності. Інколи вони можуть залежати від реалізації алгоритму. Наприклад, деякі реалізації завантажують усі дані в основну пам'ять, щоб пришвидшити виконання анонімізації. Однак такі типи підходів не будуть масштабованими для великих обсягів даних. У цьому порівняльному дослідженні ми вимірюємо споживання пам’яті під час процесу анонімізації. Ми не повідомляємо про результати використання процесора, враховуючи, що жоден алгоритм не був інтенсивним процесором.

Враховуючи, що кількість анонімованих даних може бути значною, важливо також оцінити масштабованість алгоритмів. Залежно від запланованих навантажень, користувач повинен знати ефективність алгоритму для обробки великих наборів даних. У нашій роботі масштабованість оцінюється також щодо використання пам'яті та часу анонімізації. Ми зосереджуємось на аналізі тенденцій, які слідують за алгоритмами в міру збільшення розміру вхідного набору даних.

### 2.4.2 Корисність даних

Відсутність стандартизованих показників ускладнює порівняння алгоритмів між собою. Наприклад, деякі показники залежать від наявності VGH для обчислення спотворень даних [20, 31]. Оскільки вони використовують кількість “стрибків” або глибину VGH, ці показники не застосовуються до інших типів алгоритмів (наприклад, таких, що базуються на розділенні чи кластеризації). У статистичному співтоваристві корисність даних часто вимірюється шляхом оцінки змін у розподілі базових даних (наприклад, KL-дивергенція [37], норма L1 [38]) або вимірювання однорідності в кластеризації

(наприклад, сума квадратів [ 27, 39]). Крім того, інші показники вимірюють якість анонімізації на основі корисності даних у конкретному сценарії використання. Наприклад, для відповіді на запити [21, 40], для асоціативних правил [41] або для класифікаторів навчання [17]. Ці показники не дуже підходять для PPDP, оскільки видавець даних не знає точних сценаріїв використання щойно опублікованих даних [42]. В іншому випадку видавець даних може просто виконати цільове завдання над вихідними даними та обмінятися результатами із зацікавленими сторонами замість того, щоб випускати анонімізовані дані. Цей брак знань щодо можливих застосувань мотивував потребу в загальних цілях метрики корисності, які використовують синтаксичні властивості як проксі для корисності [43].

Для наших критеріїв оцінки ми вибрали набір метрик загального призначення, оскільки вважаємо, що використання метрик, які можна широко застосувати до більшості алгоритмів анонімізації, є хорошим кроком до стандартизації їх порівняння. Нижче ми описуємо ці показники в межах даної роботи.

2.4.2.1 Узагальнена втрата інформації GenILoss

Ця метрика фіксує штраф, понесений під час узагальнення конкретного атрибуту, шляхом кількісного визначення частки доменних значень, узагальнених [17]. В нашій оцінці ми використовуємо нормований варіант цієї метрики, який був представлений в [44]. Нехай $L_i$ і $U_i$ - нижня і верхня межі атрибута $i$. Запис комірки для атрибута $i$ узагальнюється на інтервал $ij$, визначений нижньою лінійкою $L_{ij}$ та верхньою межею $U_{ij}$. Загальна втрата інформації анонімізованої таблиці $T^*$ може бути обчислена як:

$$GenILoss(T^*) = \frac{1}{|T| \cdot n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i} \qquad (2.1)$$

де $T$ - початкова таблиця, $n$ - кількість атрибутів, а $|T|$ - кількість записів.

Цей показник заснований на концепції, що значення комірок даних, які представляють більший діапазон значень, є менш точними, ніж ті, які представляють менший діапазон значень (наприклад, не одружений є менш конкретним, ніж одинокий або розлучений). Бажані нижчі значення: 0 означає відсутність перетворення (вихідні дані) і 1 означає повне приховування або максимальний рівень узагальнення даних. Хоча ця метрика застосовується лише в алгоритмах, які використовують ієрархію узагальнення, ми застосовуємо її і для неієрархічних алгоритмів (наприклад, Mondrian), оскільки будь-який інтервал узагальнених значень можна кількісно визначити таким чином. Крім того, для обчислення штрафу за категоричними атрибутами (наприклад, сімейним становищем) за вищенаведеною формулою ми використовуємо підхід відображення кожного значення до числового значення (як пояснено в [17]). Наприклад, для атрибута ‘Сімейний стан’ (VGH, зображеного на рисунку 2.1), ‘Одинак’ відображається на 1, розділене відображається на 2 і так далі, поки ‘Одружений повторно’ не відображається на 6. Отже, статус ‘Неодружений’ представлений в інтервалі [1-4], який охоплює статуси від ‘Одинак’ до ‘Вдівець’.

Для ілюстрації цього показника розглянемо таблицю 2.2 та VGH, зображені на рисунку 2.1. Для ‘Сімейний стан’, який є категоричним атрибутом, GenILoss для комірок зі значенням, що не перебуває у шлюбі, становить $\frac{4-1}{6-1} = \frac{3}{5}$ Для ‘Вік’, який є числовим атрибутом, GenILoss для комірок зі значеннями в [25-30) становить $\frac{29-25}{29-20} = \frac{4}{9}$. Нарешті, для ‘Індекс’ GenILoss для комірок зі значенням 3204* становить $\frac{6-4}{6-1} = \frac{2}{5}$. Дотримуючись формули GenILoss для інших комірок, оцінка GenILoss для всієї таблиці становить $\frac{1}{6*3} \times \frac{78}{9} = \frac{13}{27} = 0{,}48$.

2.4.2.2 Метрика чутливості DM

Ця метрика вимірює те, наскільки запис відрізняється від інших, шляхом призначення штрафу кожному запису, рівного розміру класу еквівалентності, до якого він належить [25]. Якщо запис приховано, то йому призначається штраф, рівний розміру вхідної таблиці. Загальний бал DM для *k*-анонімізованої таблиці *T** визначається:

$$DM(T^*) = \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + \sum_{\forall EQ s.t. |EQ| < k} |T| \cdot |EQ| \qquad (2.2)$$

де $T$ - початкова таблиця, $|T|$ - кількість записів, а $|EQ|$ - розмір класів еквівалентності (анонімізовані групи), створених після виконання анонімізації. Ідея цієї метрики полягає в тому, що чим більший клас еквівалентності тим більшими є втрати інформації, тому бажані нижчі значення для цього показника. Для ілюстрації цієї метрики розглянемо таблицю 2.2, оскільки обидва еквівалентні таблиці в анонімізованій таблиці мають розмір 3, показник DM для всієї таблиці обчислюється як: $3^2 + 3^2 = 18$.

2.4.2.3 Показник середнього розміру класу еквівалентності $C_{AVG}$

Ця метрика вимірює, наскільки добре створення класу еквівалентності *EQ* підходить до найкращого випадку, коли кожен запис узагальнений в *EQ* з *k* записів [21]. Мета полягає в тому, щоб мінімізувати штраф: значення 1 вказувало б на ідеальну анонімізацію, в якій розмір *EQ* є заданим *k* значенням. Загальний бал $C_{AVG}$ для анонімованої таблиці *T** задається:

$$C_{AVG}(T^*) = \frac{|T|}{|EQs| \cdot k} \qquad (2.3)$$

де $T$ - початкова таблиця, $|T|$ - кількість записів, $|EQs|$ - загальна кількість створених класів еквівалентності, а $k$ - вимога конфіденційності. Для ілюстрації цього показника розглянемо таблицю 2.2, що показує 2 класи еквівалентності, бал $C_{AVG}$ для всієї таблиці обчислюється як: $\frac{6}{2*3} = 1$.

# РОЗДІЛ 3 ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

## 3.1 Навколишнє середовище

Експерименти проводилися на завантаженій машині з 64-розрядним процесором Ubuntu 12.04 LTS, з процесором Intel Xeon E5-2430 при тактовій частоті 2,20 ГГц та 24 ГБ оперативної пам’яті, використовуючи Oracle Hotspot JVM версії 7. Розмір пам'яті машини дозволив нам виконувати експерименти, не запускаючи основні збирання сміття на Java (MaGC). Отже, отримання акуратних та стабільних вимірювань ефективності, оскільки будь-який вплив на ефективність MaGC [9] було усунено. Для трьох алгоритмів ми використовували реалізовані в загальнодоступному виконанні інструменти UT Dallas Anonymousmization Toolbox [46]. Той факт, що всі алгоритми реалізовані за допомогою загальної основи (наприклад, використання однакових структур даних для анонімізації), дозволяє справедливо порівняти ефективність. У цій реалізації проміжні набори даних про анонімізацію зберігаються в базі даних. Тому пам'ять (ОЗУ) споживається лише структурами даних, що використовуються під час анонімізації, такими як решітка узагальнення, перелік атрибутів, що входять до набору QID, та відповідні їм VGH. Ці реалізації, написані на Java, не мають апріорних оптимізацій (тобто попереднього обчислення частотних наборів), які могли б дати перевагу одному алгоритму над іншими. Ми ввели додаткову логіку в панель інструментів для вимірювання ефективності алгоритмів. Так само ми розробили окремий компонент для обчислення показників утиліти даних для вимірювання ефективності алгоритмів. Ми використовували MySQL 5.5.34 як базу даних для зберігання посередницьких станів анонімізації замість вбудованої бази даних SQLite, що є за замовчуванням у цій панелі інструментів. Причиною такого вибору є те, що MySQL є більш масштабованим, ніж SQLite, коли маємо справу з дуже великими наборами даних.

## 3.2 Налаштування експерименту

Для нашого експерименту ми використовували набори даних ‘Дорослі особи’ та ‘Ірландський’, які описано в розділі 2.3. Конфігурації, які використовуються в цих експериментах, наведені в таблиці 3.1.

Таблиця 3.1 - Експериментальні конфігурації.

| № | Експеримент | Параметри | Набір даних (розмір) |
|---|---|---|---|
| 1 | Різна кількість QID | *k*-значення=2<br>$\lvert QIDs \rvert \in [1..5]$ | ‘Дорослі особи’ (30 162)<br>‘Ірландський’ (30 000) |
| 2 | Різне значення *k* | *k*-значення $\in \{2, 5, 10, 25, 50, 100, 250, 500, 1000\}$<br>$\lvert QIDs \rvert = 3$ | ‘Дорослі особи’ (30 162)<br>‘Ірландський’ (30 000) |
| 3 | Різноманітний розмір набору даних | *k*-значення=50<br>$\lvert QIDs \rvert = 3$ | ‘Ірландський’ (5k, 10k, 20k, 30k, 50k, 100k) |

Параметри, які змінюються в цих експериментах,:

- $\lvert QIDs \rvert$: визначає кількість атрибутів, які є частиною набору QID.
- *k*-значення: визначає рівень конфіденційності, який повинен задовольняти алгоритм анонімізації. Він представляє мінімальний розмір для EQ в анонімізованому рішенні.
- Розмір набору даних відповідає кількості записів у наборі даних.

Діапазон значень, що використовуються в цих експериментах, був обраний на основі значень, використаних в оригінальних роботах Mondrian і Incognito. Крім того, ми розширили деякі з цих параметрів з більш високими значеннями (наприклад, у *k*-значеннях) на основі стандартів узагальнення бюро перепису (наприклад, будь-який розкритий географічний регіон повинен містити щонайменше 10 000 або 100 000 осіб) [46].

Поріг приховування визначає відсоток записів, які можуть бути придушені в процесі анонімізації, щоб все-таки розглянути набір даних як *k*-анонімний. Для всіх експериментів встановлено нуль, тому всі записи враховуються в процесі анонімізації.

## 3.3 Експеримент 1: різна кількість QID

У цьому експерименті ми аналізуємо ефективність алгоритмів в обох наборах даних у міру збільшення кількості QID. Ми використовуємо позначення $|\text{QIDs}|$ для позначення кількості атрибутів, що входять до набору QID. Квазі-ідентифікатор розміру $n$ складається з перших $n$ атрибутів, перелічених для набору даних, як показано в таблицях 2.13 та 2.14. Експерименти проводилися шляхом зміни $|\text{QIDs}| \in [1..5]$.

### 3.3.1 Час анонімізації

На рисунках 3.1 та 3.2 відображаються результати за час анонімізації при збільшенні кількості QID. Перше спостереження на цих рисунках полягає в тому, що для набору даних ‘Ірландський’ Datafly та Incognito показують час анонімізації, близький до 0, коли $|\text{QIDs}| \in \{1,2\}$.

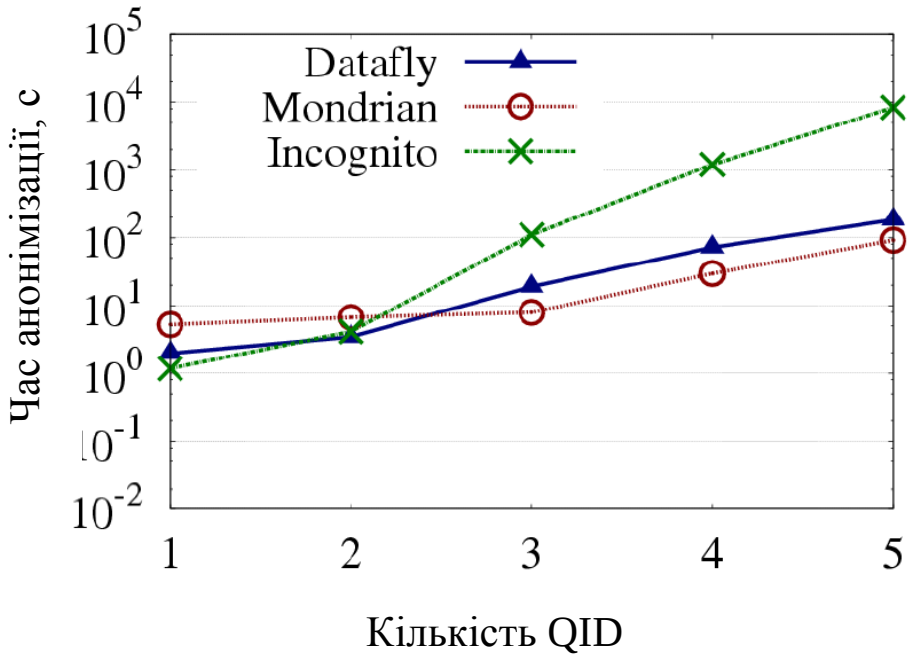


Рисунок 3.1 - Результати затрати часу для анонімізації набору ‘Дорослі особи’ при збільшенні кількості QID

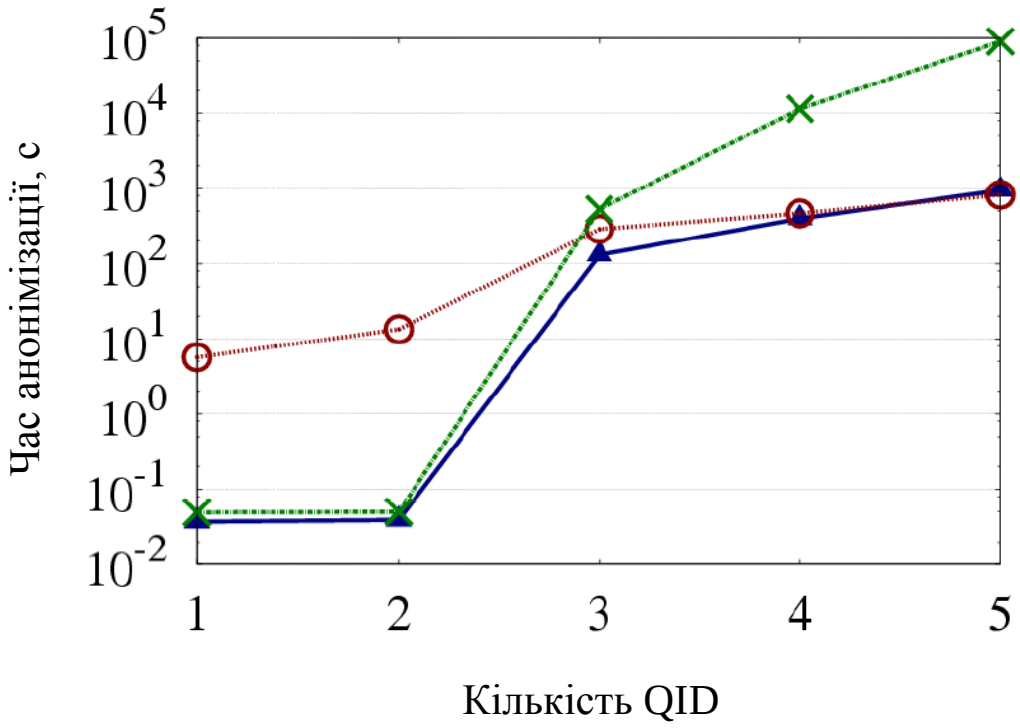


Рисунок 3.2 - Результати затрати часу для анонімізації набору ‘Ірландський’ при збільшенні кількості QID

Така поведінка викликана тим, що вихідний набір даних вже задовольняє 2-анонімності для цієї конфігурації, тому ці два алгоритми не виконують жодного узагальнення. Mondrian навпаки анонімізував набір даних, оскільки цей алгоритм не перевіряє *k*-анонімність як його критерій зупинки. Натомість він намагається зробити всі можливі розділи для даних, поки не буде дозволено більше скорочень (навіть якщо вихідний набір даних вже задовольняє *k*-анонімність). В алгоритмі Incognito – найгірший показник для обох наборів даних майже в усіх експериментальних конфігураціях (крім тих, де не було узагальнено). Як і очікувалося, Incognito показав складність у часі O $(2^{|QID|})$[12, 47]. Це можна спостерігати на рисунках, де Incognito показує найбільший приріст часу анонімізації щодо кількості QID незалежно від набору даних. Наприклад, для набору ‘Дорослі особи’ Incognito виконує анонімізацію в 13 і 90 разів повільніше, ніж Mondrian, і в 5 і 45 разів повільніше, ніж Datafly, коли $|QIDs| \geq 3$. Це збільшення пояснюється тим, що простір пошуку анонімізаційних рішень (вузлів) стає ширшим, оскільки більше атрибутів є

частиною набору QID, що збільшує час, необхідний для проходження решітки узагальнення та перевірки наявності *k*-анонімності в кожному окремому стані. Цей приріст, який відбувається в різних ступенях для кожного набору даних, залежить від глибини VGH, визначеної для кожного атрибута, яка визначає загальну кількість вузлів, що підлягають оцінці в гіршому випадку (оскільки це може бути менше, якщо решітка буде обрізана) . Наприклад, розглянемо випадок, коли $|QIDs|$ збільшується від 3 до 4, що показує перше значне збільшення (зверніть увагу на логарифмічну шкалу на осі y). Для ‘Дорослі особи’ кількість оцінюваних вузлів зростає з 17 (з 20) до 50 (з 60), тоді як для ‘Ірландський’ вона зростає з 10 (з 40) до 91 (з 120); саме тому Incognito демонструє більш істотний підйом для набору ‘Ірландський’.

Ми помітили, що Incognito працює добре лише тоді, коли кількість квазі-ідентифікаторів невелика. Крім того, важливо враховувати не тільки кількість атрибутів у наборі QID, але й кількість можливих станів узагальнення, які може мати кожен атрибут (тобто глибина VGH), оскільки цей фактор впливає на швидкість збільшення за час анонімізації. Mondrian і Datafly демонструють прийнятну продуктивність (менше 16 хвилин, коли $|QIDs|=5$). Загалом, можна зазначити, що при порівнянні продуктивності трьох алгоритмів для обох наборів даних алгоритми працюють краще для ‘Дорослі особи’, ніж для ‘Ірландський’; оскільки вартість функціональних характеристик (згаданих у розділі 2.4.1) для ‘Ірландський’ вище. Крім того, найкращий виконавець щодо часу анонімності відрізняється у кожному наборі даних: Mondrian найшвидший для ‘Дорослі особи’ коли $|QIDs| \geq 3$, тоді як Datafly – найшвидший для ‘Ірландський’.

### 3.3.2 Споживання пам'яті.

На рисунках 3.3 та 3.4 показані результати споживання пам'яті в процесі анонімізації при збільшенні кількості QID.

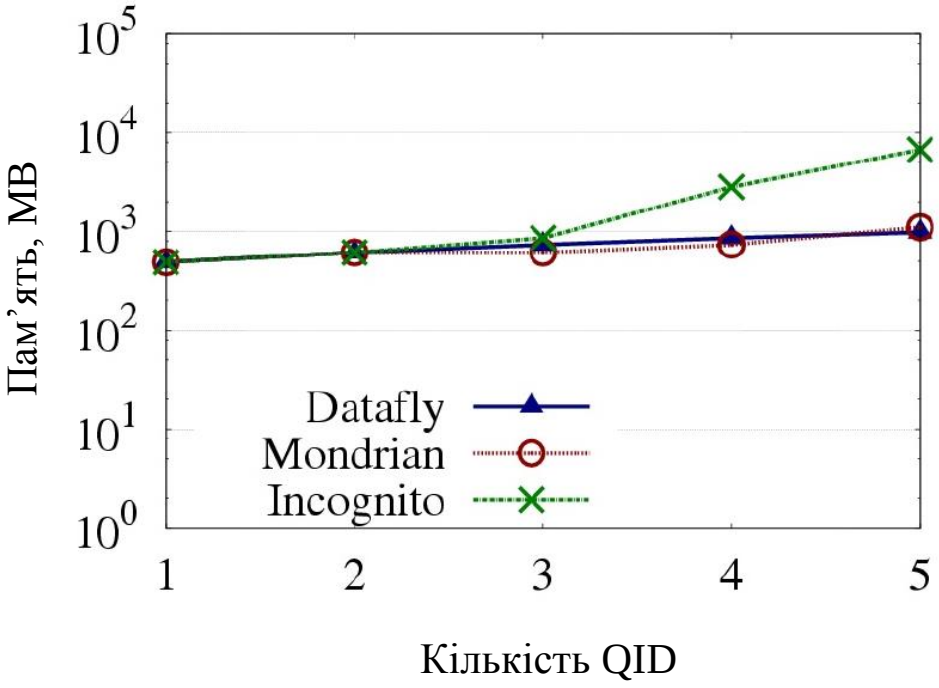


Рисунок 3.3 - Результати споживання пам'яті під час анонімізації для набору ‘Дорослі особи’ при збільшенні кількості QID

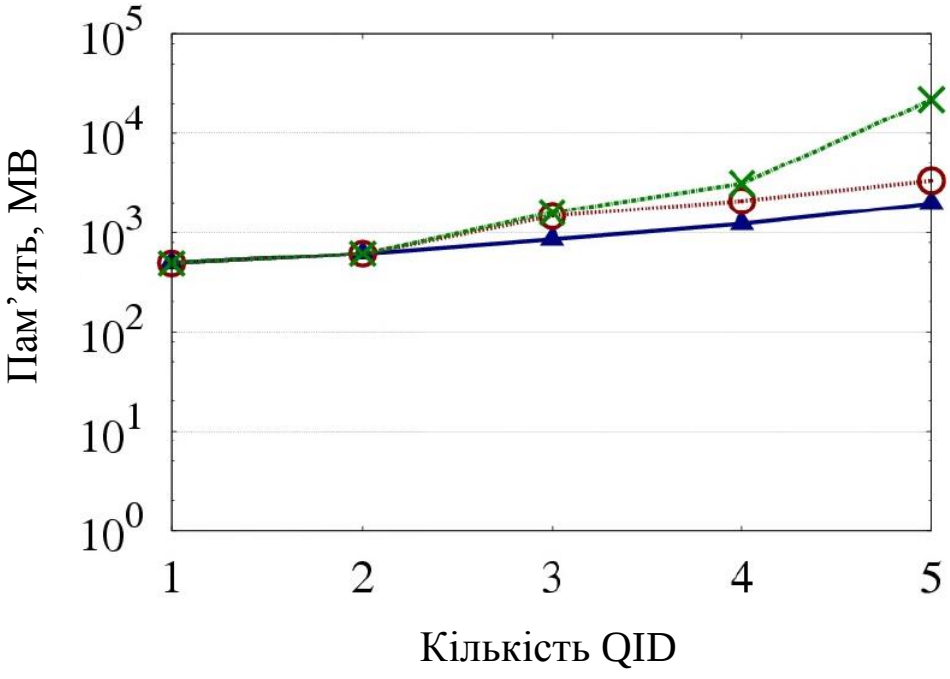


Рисунок 3.4 - Результати споживання пам'яті під час анонімізації для набору ‘Ірландський’ при збільшенні кількості QID

Datafly та Mondrian демонструють незначну розбіжність у споживанні пам'яті, але залишаються відносно стабільними, оскільки кількість QID-кодів збільшується. Однак у випадку Incognito один недолік, який можна спостерігати, - це те, що він є інтенсивним у пам’яті щодо збільшення кількості QID; досягнення 21 ГБ пам'яті, коли $|QIDs|=5$. Зростання споживання пам’яті в Incognito, можливо, пояснюється збільшенням кількості вузлів у генералізаційній решітці.

3.3.3 Узагальнена втрата інформації (GenILoss).

На рисунках 3.5 та 3.6 представлені результати утиліти даних щодо метрики GenILoss зі збільшенням кількості QID. Можна зазначити, що тенденції продуктивності для кожного алгоритму різні для обох наборів даних. Datafly та Mondrian демонструють нестабільну поведінку у ‘Дорослі особи’, тоді як для ‘Ірландський’ вони в основному демонструють зростаючу тенденцію. Більше того, штрафи за втрату інформації для набору даних для ‘Дорослі особи’ є вищими, ніж в ‘Ірландський’ для всіх алгоритмів. Наприклад, Mondrian добре працює для ‘Ірландський’, але він має найгірші результати для ‘Дорослі особи’ (коли $|QIDs| \in \{2,3\}$). Причиною таких високих значень у Mondrian для ‘Дорослі особи’є спосіб виконання розділення (медіанний підхід), а також перекошений розподіл даних у QID. Більш конкретно, той факт, що Mondrian розбиває дані за допомогою одного атрибута, замість двох або трьох наявних у наборі QID, змушує решту атрибутів у наборі QID зберігати їх найменші конкретні значення. Це спричиняє високу пеню за ці ознаки. Наприклад, коли $|QIDs| = 3$({Вік, Стать, Раса}), ‘Вік’ є єдиним атрибутом, де дозволено скорочення, оскільки здійснити розріз у ‘Стать’ та ‘Раса’ неможливо, оскільки середня умова розподілу не виконується.

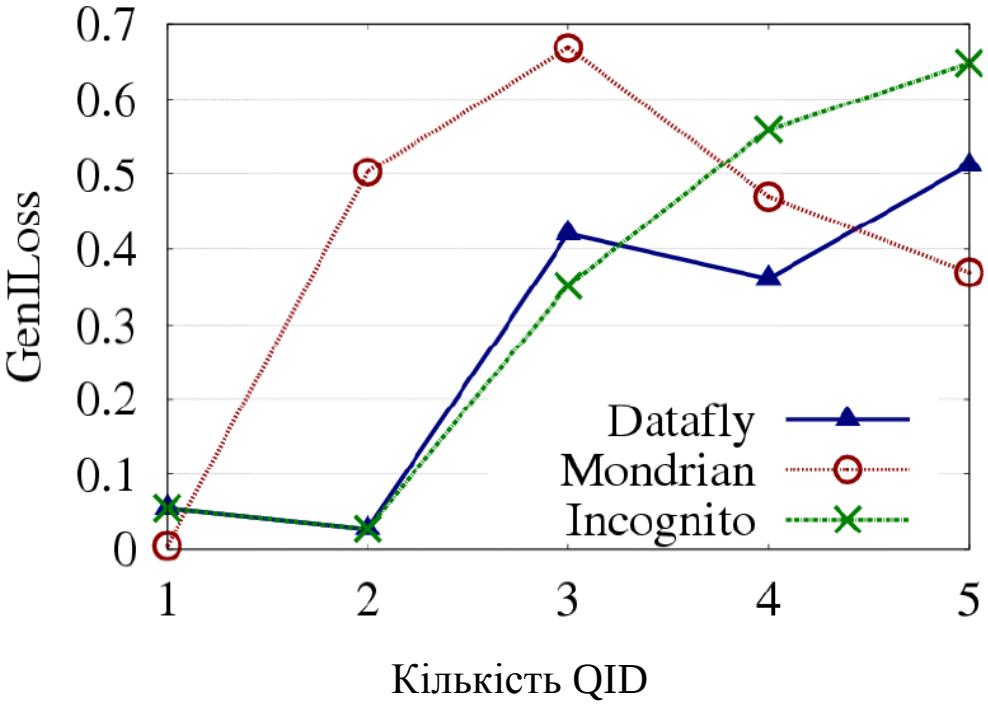


Рисунок 3.5 - GenILoss для набору ‘Дорослі особи’ при збільшенні кількості QID

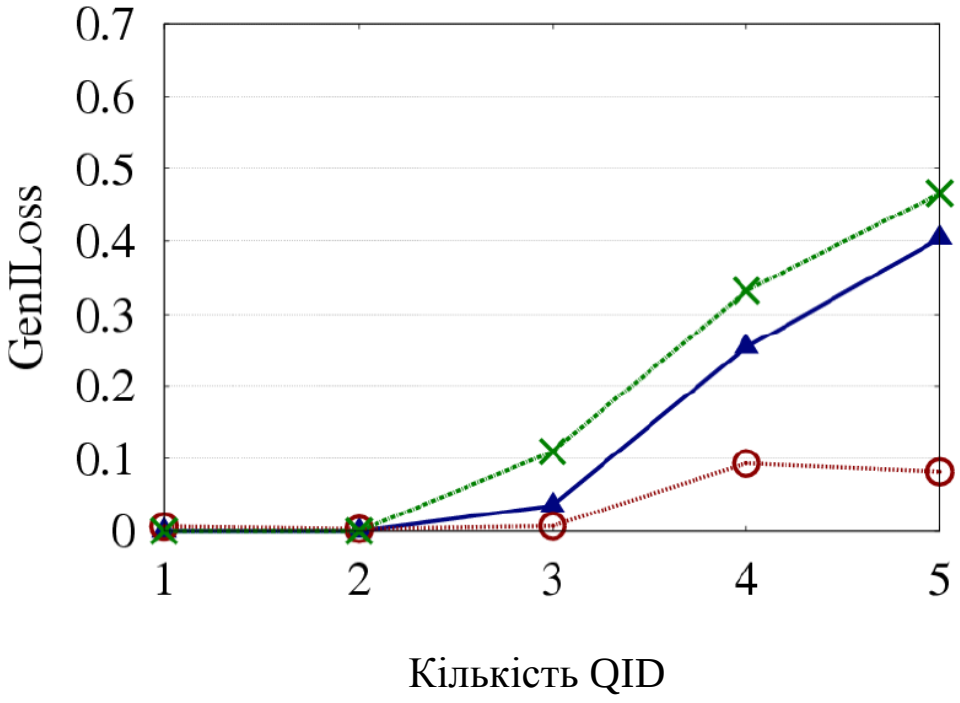


Рисунок 3.6 - GenILoss для набору ‘Ірландський’ при збільшенні кількості QID

Підхід, який використовується для пошуку медіани, використовує частотні набори, про які йдеться у розділі 2.2.3. Розглянемо першу ітерацію Mondrian, коли дані скануються, щоб знайти середнє значення для атрибута ‘Стать’. У своїй області ‘Стать’ має два значення: ‘жіноча’ з числом частот 9 782 записів і ‘чоловіча’ з 20 370 записами. Коли всі записи враховані, медіана виявляється в положенні 15,081; яка досягається, коли до числа додається частота, встановлена для чоловічої статі, отже, вибране розділене значення є ‘чоловіча’. Тим не менш, виконувати розріз за цим значенням не дозволяється, оскільки це не відповідає умові мінімального розміру *k* записів для кожного нового поділу. Отже, дані не можна розділити за допомогою цього атрибута, і його значення зберігають найменший специфічний стан, характеристика, зазначена в розділі 2.2.3. Така поведінка зберігається в подальших ітераціях, оскільки розподіл даних залишається однаковим навіть на нижчих рівнях.

Аналогічна ситуація трапляється і з атрибутом ‘Раса’, який має такий розподіл для дорослих: ‘Американець-індіанець-ескімос’ – 286, ‘Азіат-Ісландець’ –895, ‘Темношкірий’ – 2,817, ‘Інші’ – 231 та ‘Світлошкірий’ – 25,933. Вибране значення розділення – ‘Світлошкірий’, де середнє значення знайдено, але знову ж таки, розріз не допустимий.

Оскільки методика середнього розподілу має на меті отримати рівномірне заповнення, ця методика спричиняє великі втрати інформації при перекосі даних. Така поведінка демонструє, що Mondrian краще працює при рівномірному розподілі, оскільки більшість QID можуть використовуватися для розділення даних. В іншому випадку бал GenILoss для Mondrian високий, оскільки ті атрибути, які не вдалося використати для розділення, зберігають свої найменші конкретні значення.

Для двох інших алгоритмів, Datafly та Incognito, помітною тенденцією в обох наборах даних є те, що оцінка GenILoss має тенденцію до збільшення. Така поведінка обумовлена тим, що для досягнення *k*-анонімності більшість атрибутів потрібно додатково узагальнити, що погіршує корисність даних. Це підтверджує прокляття розмірності, вивчене в роботі [4], яке полягає в тому, що ефективність анонімізованих даних знижується при збільшенні розмірності.

3.3.4 Метрика чутливості (DM).

На рисунках 3.7 та 3.8 представлені результати для корисності даних щодо метрики DM при збільшенні кількості QID. Для набору даних для ‘Дорослі особи’ зрозуміло, що Mondrian є найкращим виконавцем, оскільки цей алгоритм має на меті створення більш тонких еквалайзерів (EQ невеликих розмірів). Однак для набору даних ‘Ірландський’ Mondrian виконує найгірше, коли $|\text{QIDs}| \in \{1,2\}$. Це пов’язано з тим, що вихідний набір даних вже задовольняє *k*-анонімності для цієї конфігурації; тим не менш, Mondrian створює перегородки. Така ситуація призводить до зменшення кількості еквалайзерів (вже сформованих у вихідних даних), але в результаті збільшується розмір кожного окремого еквалайзера (кількість записів у кожному еквіваленті), що не бажано для метрики DM. Наприклад, коли $|\text{QIDs}|=1$, кількість еквалайзерів у вихідних даних становить 70, тоді як в анонімізованих даних, що використовують Mondrian, це 57.

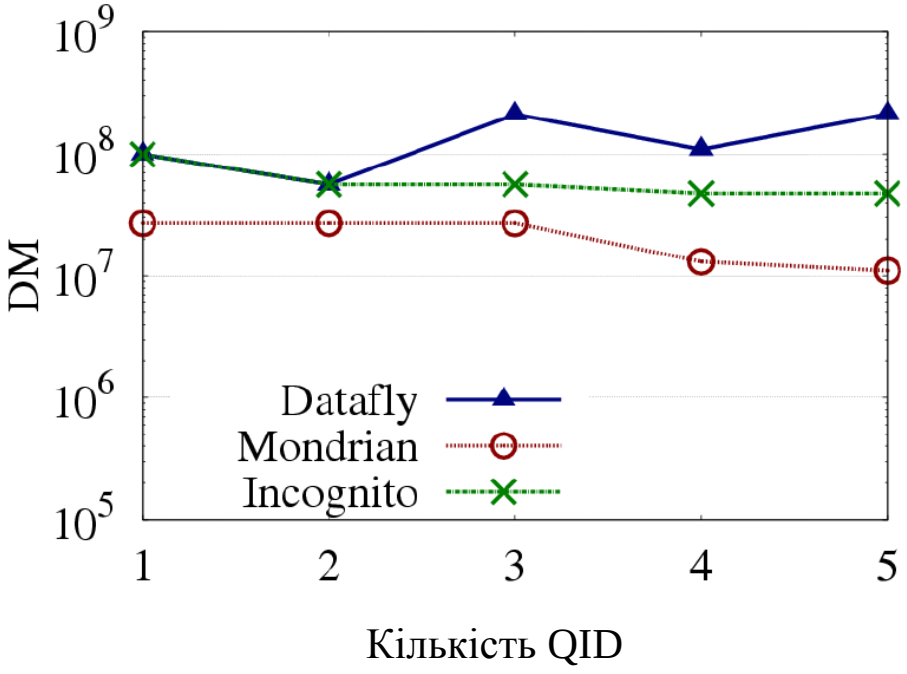


Рисунок 3.7 - DM для набору ‘Дорослі особи’ при збільшенні кількості QID

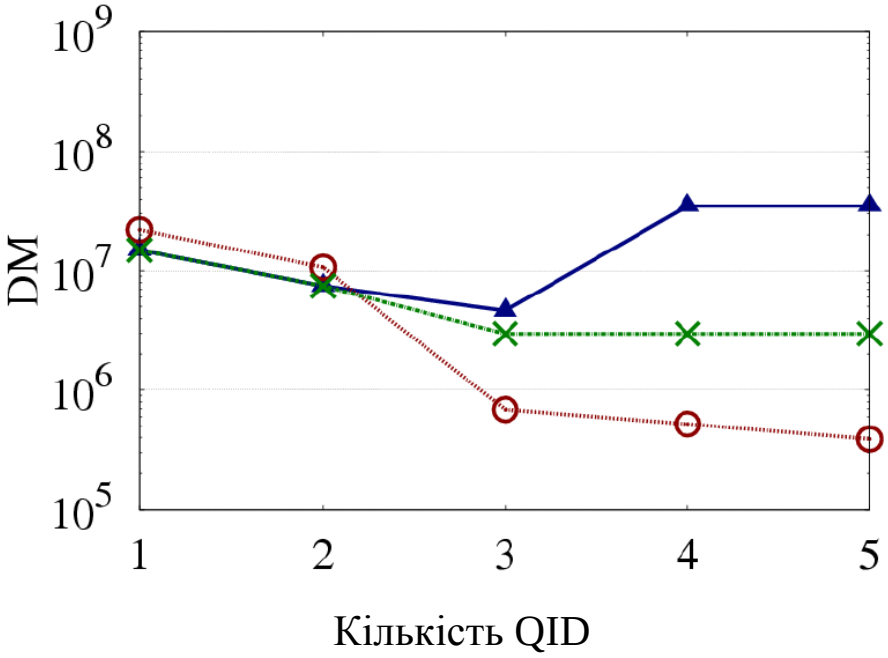


Рисунок 3.8 - DM для набору ‘Ірландський’ при збільшенні кількості QID

Ще одне спостереження полягає в тому, що DM не фіксує перетворення, виконані на QID. Наприклад, коли $|QIDs| \in [1..3]$ для ‘Дорослі особи’, Mondrian показує те саме значення DM (той самий рівень корисності даних). Така поведінка пов'язана з тим, що для цих трьох експериментальних конфігурацій Mondrian виконує розділи, використовуючи лише один атрибут (вік); ситуація, яка була пояснена в попередній метриці GenILoss. Отже, для цих трьох конфігурацій завжди створюються одні й ті ж еквалайзери (анонімізовані групи). Після того, як до набору QID додаються інші атрибути (тобто, кількість QID збільшується), Mondrian може робити розділи на всіх, покращуючи таким чином значення DM.

Ще один аспект, який ми визначили, що впливає на корисність даних при використанні Mondrian, - це критерії, які використовуються для вибору розміру (атрибута), у якому виконується розділ (тобто атрибут з найширшим діапазоном значень), зокрема, коли кілька атрибутів мають той самий діапазон значень. У разі реалізації UT Dallas Toolbox він вибирає перший атрибут; як обговорювалося розділі 2.2.3. У наших експериментах ми помітили, що це

рішення впливає на кількість створених поділів, зменшуючи кількість поділів у деяких випадках.

Що стосується Datafly та Incognito, вони представляють однакові значення у вигляді DM, коли $|\text{QIDs}| \in \{1,2\}$, оскільки обидва алгоритми досягають однакового рішення для анонімізації. Що стосується більшої кількості QID, Incognito перевершує Datafly як перший виявляє узагальнення, що дає максимальну кількість EQ (як пояснено у розділі 2.2.2). Таким чином, з точки зору показників, заснованих на розмірах груп (таких як DM та $C_{AVG}$), очікується, що Incognito буде працювати краще, ніж Datafly. Однак Incognito не перевершує Mondrian, що виграє від його багатовимірного підходу.

Порівнюючи корисність даних алгоритмів між наборами даних щодо DM, ми спостерігаємо, що вони краще для набору ‘Ірландський’. Причиною такої поведінки є розподіл набору даних, який дозволяє алгоритмам створювати більш точні EQ, отже, зменшуючи бал DM.

### 3.3.5 Середній розмір класу еквівалентності $C_{AVG}$

На рисунках 3.9 та 3.10 представлені результати для метрики $C_{AVG}$ зі збільшенням кількості QID.

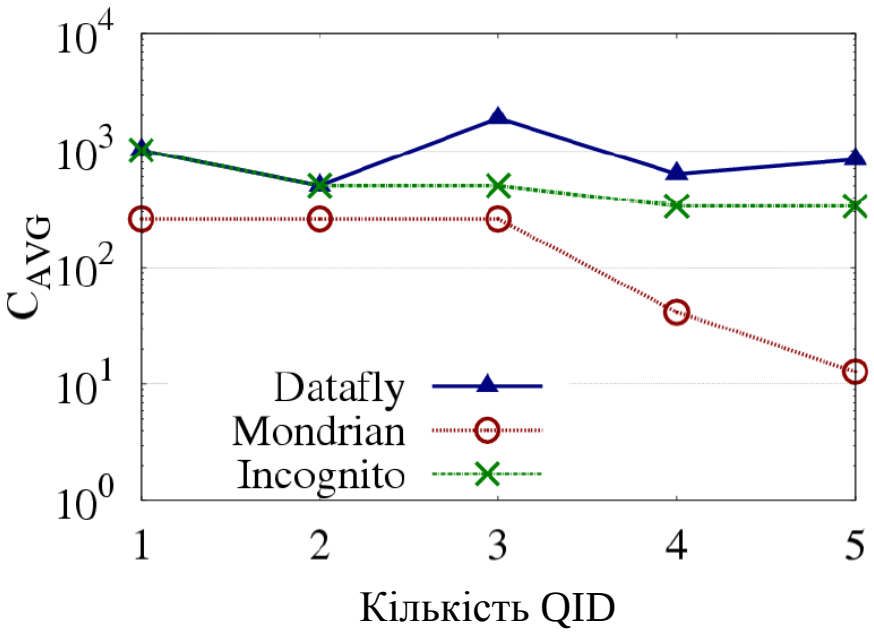


Рисунок 3.9 - $C_{AVG}$ для набору ‘Дорослі особи’ при збільшенні кількості QID

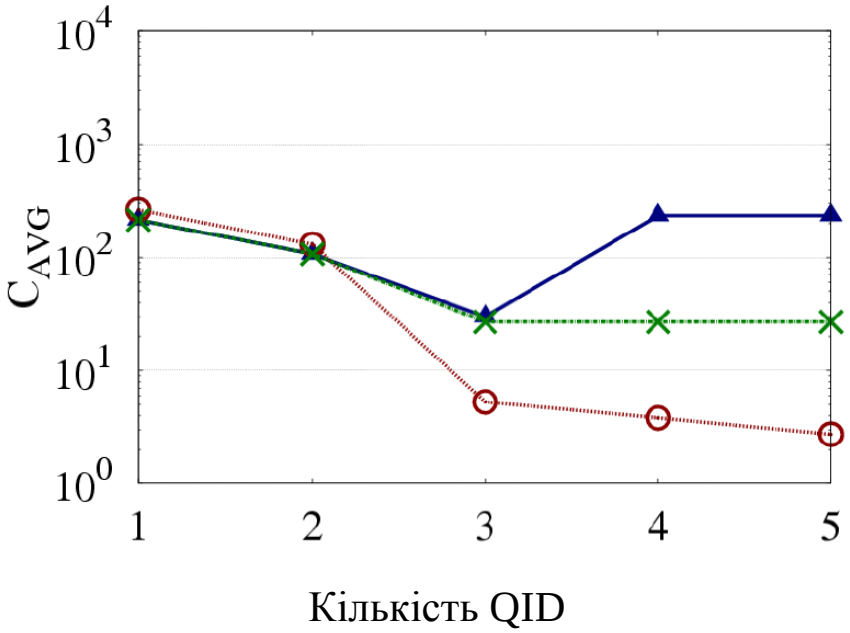


Рисунок 3.10 - $C_{AV\,G}$ для набору ‘Ірландський’ при збільшенні кількості QID

Як видно на рисунках, тенденції цієї метрики схожі на результати DM для обох наборів даних, але з різною величиною. Mondrian, здається, переважає у більшості сценаріїв, завдяки кількості EQ, створених під час анонімізації; аспект, який Mondrian прагне максимально використати. Кожного разу, коли оцінка для $C_{AV\,G}$ однакова для двох або більше алгоритмів, це означає, що ці алгоритми виробляють однакову кількість еквалайзерів. Однак, це не обов'язково означає, що розмір їхніх еквалайзерів однаковий. Це показує, що $C_{AV\,G}$ не фіксує розподіл записів серед еквалайзерів.

### 3.3.6 Результати порівняння алгоритмів для експерименту 1

Для цього експерименту алгоритми мають кращу ефективність для ‘Дорослі особи’, ніж для ‘Ірландський’ з точки зору ефективності. Для часу анонімізації та споживання пам’яті Datafly та Mondrian працюють краще. Обидва алгоритми анонімізували дані в розумний час для оцінюваних розмірів QID. Хоча Datafly та Incognito демонстрували експоненціальний ріст, показник Incognito набагато більший. Крім того, не тільки кількість QID впливає на часову складність, але й на кількість можливих станів анонімізації (тобто висоти VGH).

Щодо корисності даних, алгоритми показали кращу ефективність в наборі даних ‘Ірландський’ завдяки розподілу даних QID. Щодо показників утиліти даних на основі розміру еквалайзерів, таких як, DM та $C_{AV\ G}$, Mondrian перевершує інші алгоритми. Однак для GenILoss було показано, що Mondrian істотно залежить від вихідних даних, оскільки медіанне розділення не може бути виконане для деяких атрибутів, що призводить до високих значень GenILoss..

## 3.4 Експеримент 2: різні значення *k* в *k* -анонімізації

У цьому експерименті ми аналізуємо ефективність алгоритмів для обох наборів даних, оскільки значення *k* збільшується, використовуючи конфігурацію $|QIDs|=3$.

### 3.4.1 Час анонімізації

На рисунках 3.11 та 3.12 показані результати часу анонімізації при збільшенні значення *k*. Зі збільшенням значення *k* інтуїтивно очікується, що кількість узагальнень, необхідних для задоволення *k*-анонімізаціїі, збільшуватиметься (і, отже, час анонімізації). Це тому, що складніше задовольнити більш високий рівень конфіденційності. Як видно з рисунків, це збільшення мінімальне для Datafly, де тенденція виглядає майже стійкою для обох наборів даних, враховуючи відмінності в кількості виконаних узагальнень; від 4 до 5 для ‘Дорослі особи’ і від 2 до 7 для.

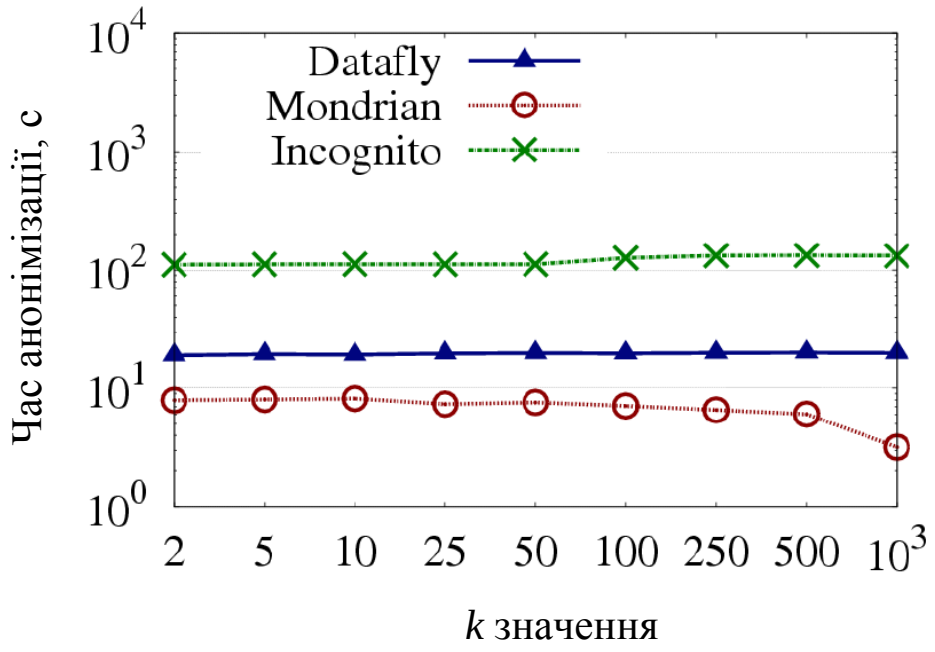


Рисунок 3.11 - Результати затрати часу для анонімізації набору ‘Дорослі особи’ при збільшенні значення *k*

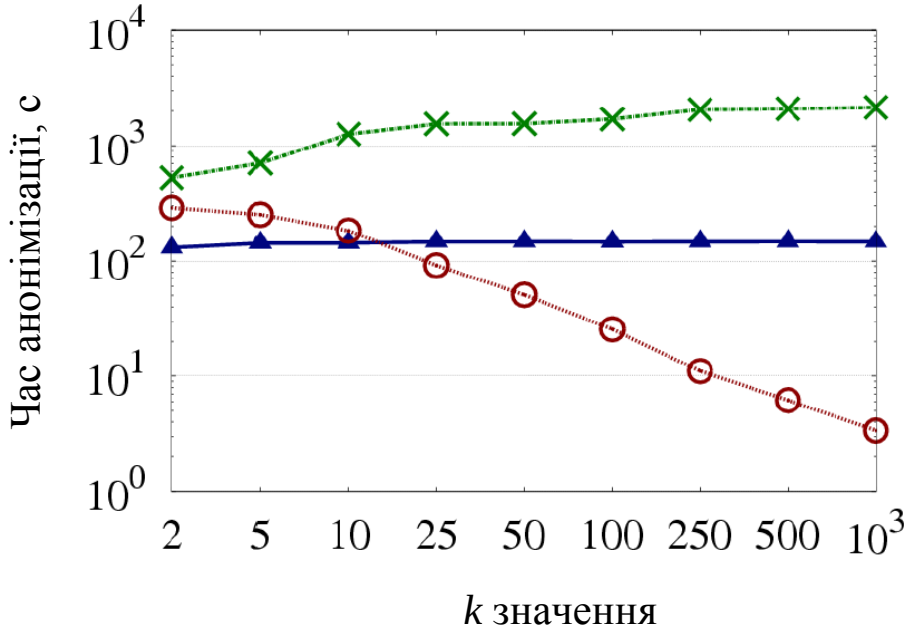


Рисунок 3.12 - Результати затрати часу для анонімізації набору ‘Ірландський’ при збільшенні значення *k*

Incognito демонструє низьку чутливість до зростання *k* для набору ‘Дорослі особи’, де найнижчий час анонімізації становить 110,692 секунди (трохи нижче 2 хвилин), а найвищий - 132,59 секунди (трохи вище 2 хвилин). Incognito оцінює між 17 і 20 (з 20) вузлів із решітки узагальнення для ‘Дорослі особи’. Однак для ‘Ірландський’ різниця в кількості оцінюваних вузлів вища, між 10 і 39 (з 40). Тому Incognito демонструє зростаючу тенденцію для набору ‘Ірландський’, де найнижчий час анонімізації становить 528,50 секунди (трохи нижче 9 хвилин), а найвищий - 2146,87 секунди (трохи нижче 36 хвилин). Невеликі відмінності у кількості узагальнень чи оцінених вузлів вказують на те, що зміна параметра *k* при фіксованому $|QIDs|$ не робить істотного впливу на час анонімізації для Datafly та Incognito для ‘Дорослі особи’.

З іншого боку, Mondrian поводиться по-різному, демонструючи тенденцію зменшення обох наборів даних. Це тому, що кількість можливих поділів, які можна виконати (які задовольняють *k*), зменшується зі збільшенням значення *k*. Наприклад, для верхньої та нижньої меж *k*, Mondrian створює між 15 та 57 розділами для ‘Дорослі особи’ та між 32 та 5664 розділами для ‘Ірландський’. Велика різниця між кількістю поділів, виконаних для цих наборів, показує вплив, який має розподіл даних QID на ефективність Mondrian, оскільки тенденція зменшення цього алгоритму для ‘Ірландський’ є більш різкою.

### 3.4.2 Споживання пам'яті.

На рисунках 3.13 та 3.14 представлені результати споживання пам'яті зі збільшенням значення *k*.

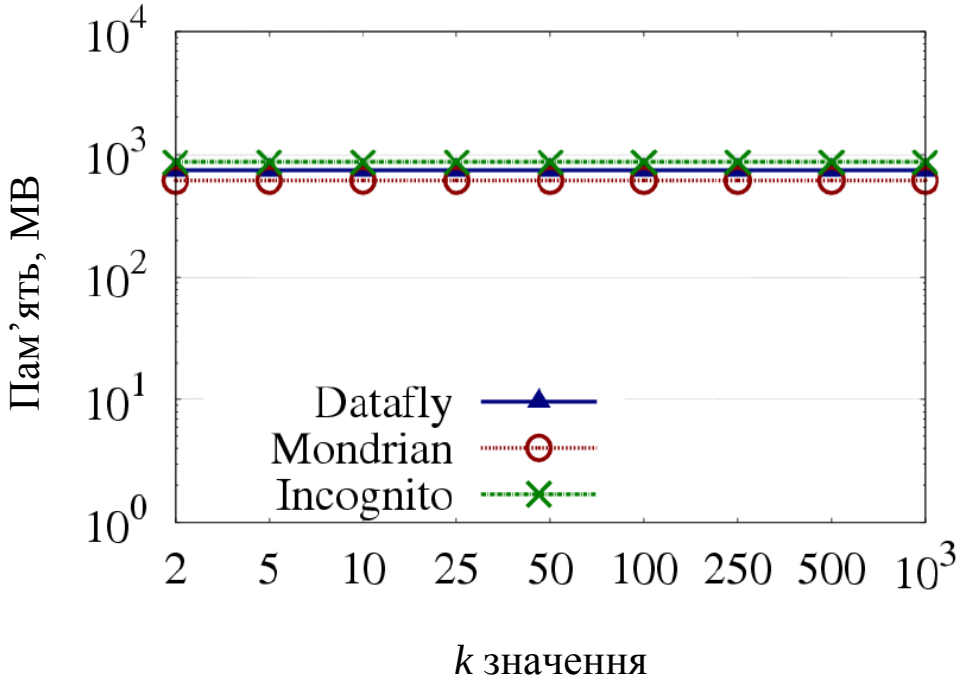


Рисунок 3.13 - Результати споживання пам'яті під час анонімізації для набору ‘Дорослі особи’ при збільшенні значення *k*

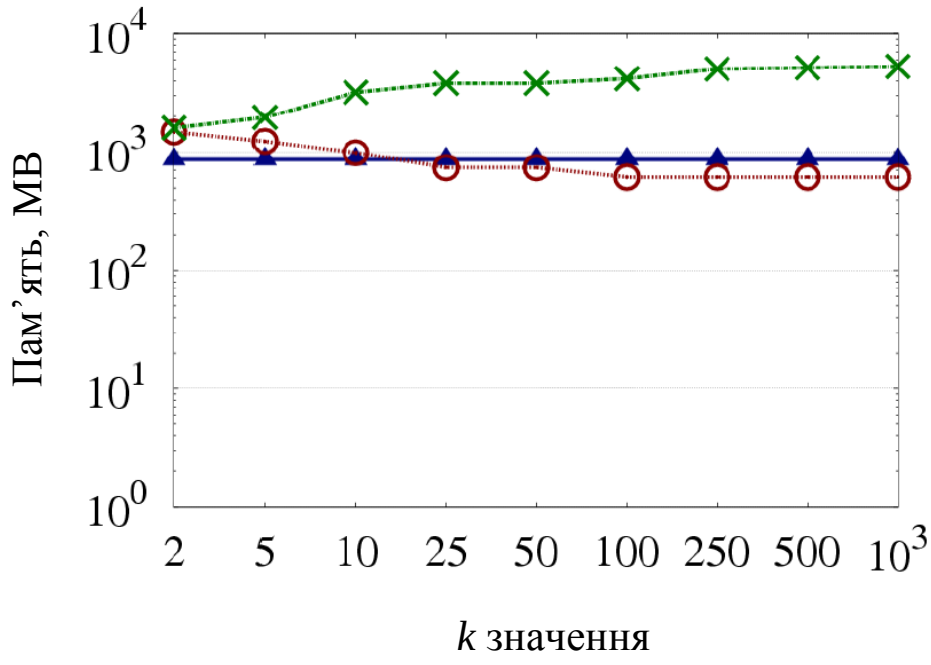


Рисунок 3.14 - Результати споживання пам'яті під час анонімізації для набору ‘Ірландський’ при збільшенні значення *k*

Як видно з рисунків, зміна *k* не має істотного впливу на споживання пам'яті для алгоритмів. Тенденції трьох алгоритмів відносно стабільні. Mondrian навіть показав невелике зменшення споживання пам’яті, оскільки *k* збільшується, тоді як Incognito показав невелике збільшення для одного з оцінюваних наборів даних. Datafly підтримував стабільну поведінку для обох наборів даних.

3.4.3 Узагальнена втрата інформації (GenILoss).

Результати, що показують корисність анонімізованих даних щодо GenILoss при збільшенні значення *k*, зображені на рисунках 3.15 та 3.16. У наборі даних для дорослих Mondrian показує однакові втрати інформації для всіх *k*-значень. Така поведінка очікується, оскільки ця метрика фіксує втрату точності в узагальнених атрибутах QID, а для цієї експериментальної конфігурації (з різним значенням *k* та $|QIDs|=3$) Mondrian використовує лише один атрибут для розділення, який викликає той самий результат GenILoss незалежно від використовуваного *k* значення. Ця ситуація також призводить до того, що показник GenILoss у Mondrian буде більшим для набору ‘Дорослі особи’, ніж для ‘Ірландський’. Це відбувається тому, що Mondrian починає розподіл даних з атрибутами в їх найменш специфічному стані, щоб поступово спеціалізувати їх на кожному поділі. Отже, якщо розділяється лише один атрибут, інші атрибути зберігають найбільш загальне значення, несучи максимальне покарання за ці атрибути. Для того ж набору даних (для дорослих) Datafly демонструє раптове збільшення балу GenILoss, коли значення *k* змінюється з 25 на 50. Це відбувається тому, що Datafly виробляв одне рішення для анонімізації, коли $2 \leq k \leq 25$ та інше, коли $k \geq 50$. Incognito є найкращим виконавцем стосовно GenILoss для цього набору даних, хоча, коли $k = 100$, три алгоритми досягають однакової оцінки GenILoss.

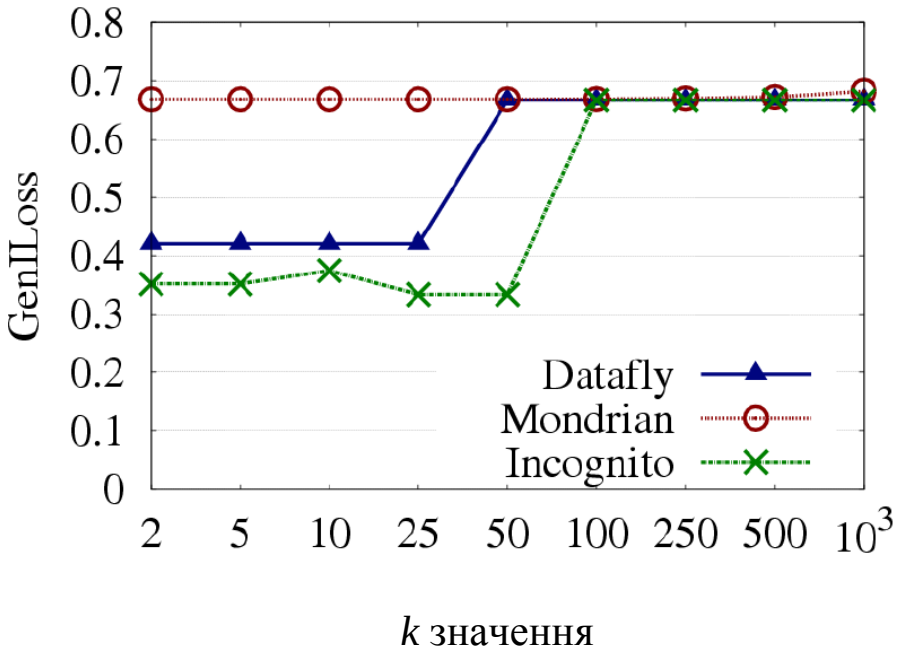


Рисунок 3.15 - GenILoss для набору ‘Дорослі особи’ при збільшенні значення $k$

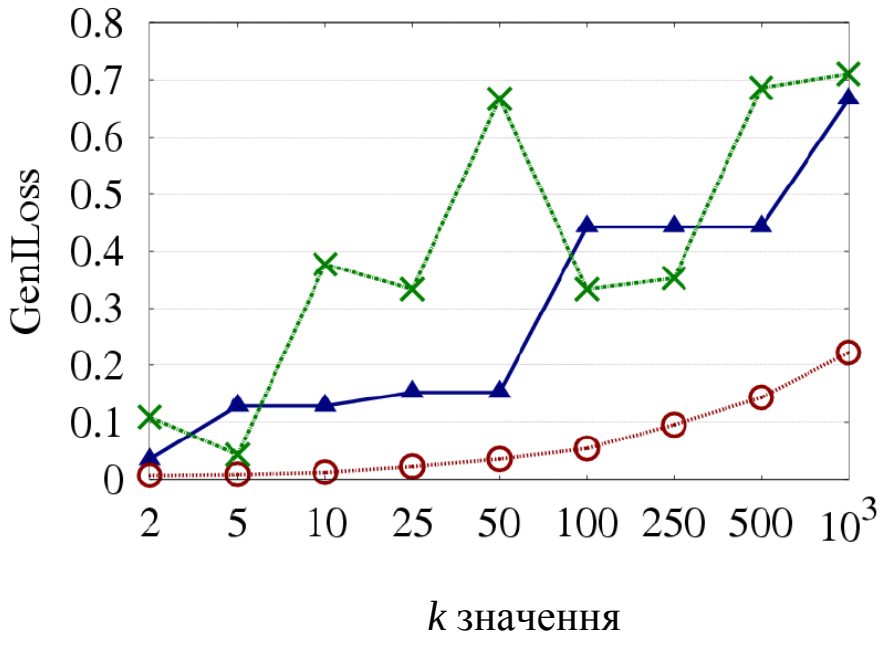


Рисунок 3.16 - GenILoss для набору ‘Ірландський’ при збільшенні значення $k$

Для набору даних ‘Ірландський’ Incognito показує помилкову поведінку. Високі значення (наприклад, піки, показані при $k \in \{50,500,1000\}$), викликані тим, що два з трьох атрибутів у наборі QID (тобто ‘Стать’ та ‘Країна’) були

узагальнені до їх максимального рівня. Включення атрибуту ‘Стать’ в узагальнення має більш високий бал GenILoss порівняно з тим, який отримують інші атрибути, що мають більш глибокий VGH. Це демонструє важливість не тільки врахування кількості атрибутів у наборі QID (що фіксується в цьому експерименті), але і VGH, визначеного для QID; оскільки включення певних атрибутів до узагальнення може погіршити корисність даних. На основі цієї інформації видавці даних можуть розглянути можливість виключення атрибутів (коли це можливо), які найбільше погіршують корисність даних. Іншою альтернативою є реструктуризація VGH для цих атрибутів, щоб забезпечити більш детальну деталізацію узагальнення та уникнути деградації даних. Для того ж набору даних (ірландський) показник Datafly істотно зростає, коли *k* змінюється від 50 до 100. Це відбувається тому, що кількість узагальнень для атрибута ‘Вік’ збільшилася вдвічі (тобто вікові значення були згруповані в більш широкі інтервали) , що робить отримані значення менш конкретними. Mondrian показаний найкращим виконавцем для набору даних ‘Ірландський’, причому значення GenILoss поступово збільшується порівняно з іншими алгоритмами. Це викликано тим, що в міру збільшення *k* можливе менше поділів у наборі даних. Таким чином, значення QID залишаються згрупованими в більш широкі інтервали. Незважаючи на це, для цього набору даних Mondrian показує найменші втрати інформації, оскільки для цієї експериментальної конфігурації (з різним значенням *k* та $\left|\text{QIDs}\right|=3$) він зміг виконувати поділи за всіма атрибутами в наборі QID.

### 3.4.4 Метрика чутливості (DM).

На рисунках 3.17 та 3.18 ми можемо спостерігати, що загальна тенденція значення DM – це збільшення для всіх трьох алгоритмів (набагато менше для ‘Дорослі особи’, ніж для ‘Ірландський’).

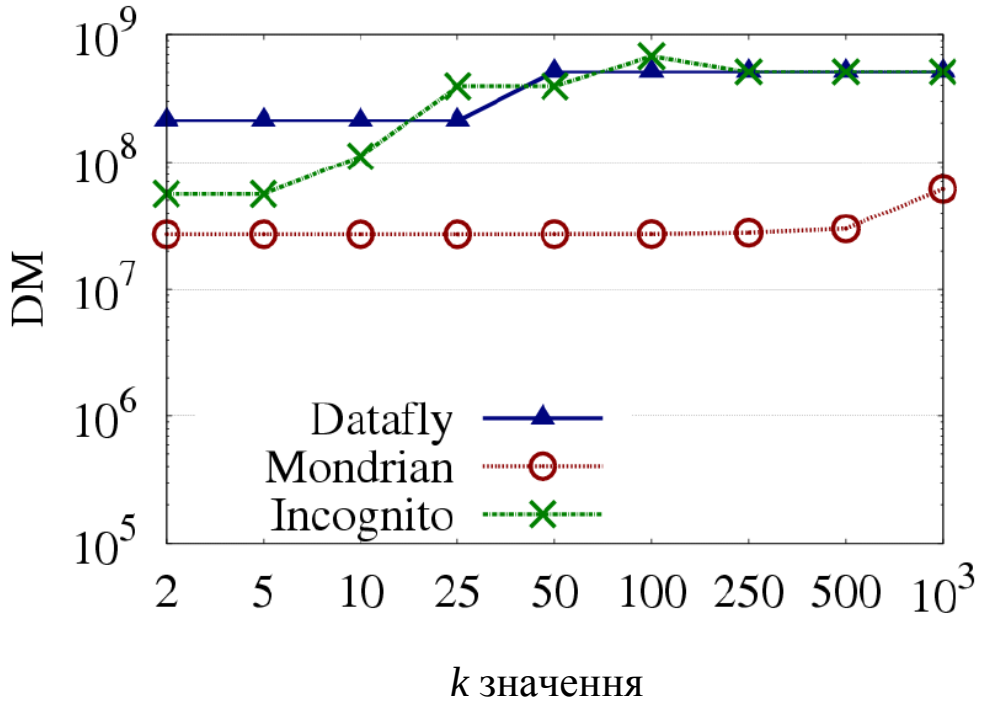


Рисунок 3.17 - DM для набору ‘Дорослі особи’ при збільшенні значення *k*

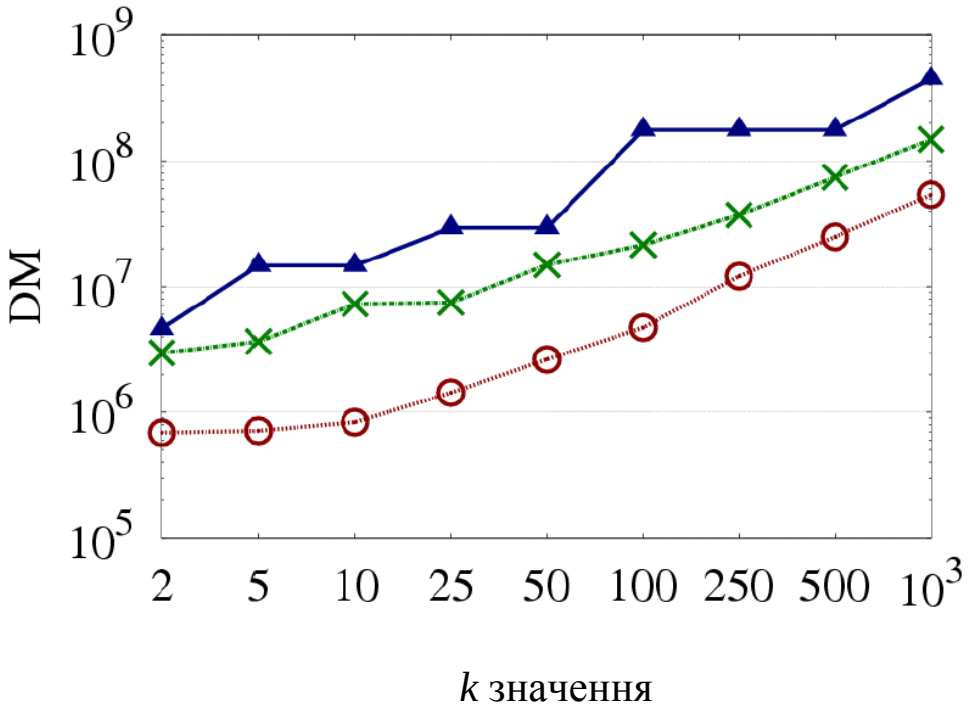


Рисунок 3.18 - DM для набору ‘Ірландський’ при збільшенні значення *k*

Зі збільшенням значення *k* все більше записів є частиною еквалайзера, і, таким чином, записи менш відрізняються один від одного. У наборі даних для

‘Дорослі особи’ тенденції залишаються більш стійкими, демонструючи низьку чутливість до зростання значення *k*. Mondrian є найкращим виконавцем, оскільки мета цього алгоритму та його механізму розподілу – максимально збільшити кількість еквалайзерів (мінімізуючи їх розмір). У цьому експерименті, при збільшенні *k* значення, Mondrian створює менше еквалайзерів, але більшого розміру; ситуація, чітко показана для набору даних ‘Ірландський’ (тобто тенденція, що зростає). Інтуїтивно зрозуміло, що Incognito має бути другим найкращим виконавцем, оскільки реалізація цього алгоритму вибирає оптимальне рішення, таке, яке дає максимальну кількість еквалайзерів. Це стосується більшості випадків, за винятком набору даних ‘Дорослі особи’, коли *k* = 25 і 100. У цьому випадку Datafly працює краще, ніж Incognito, оскільки, хоча Incognito дав більше еквалайзерів, ніж Datafly, один з еквалайзерів великий, що суттєво вплинула на показник DM Incognito. Наприклад, коли k = 25, Datafly створює 8 еквалайзерів, де найбільший EQ має 9 362 записів. Incognito створює 10 еквалайзерів, але найбільший еквівалент має 18 038 записів; що призводить до більш високого загального показника DM.

### 3.4.5 Середній розмір класу еквівалентності ($C_{AV\ G}$).

На рисунках 3.19 та 3.20 представлені результати корисності даних щодо $C_{AV\ G}$ у міру збільшення значення *k*. Часто трапляється так, що $C_{AV\ G}$ і DM демонструють подібні тенденції щодо продуктивності алгоритмів, оскільки обидві метрики вимірюють корисність даних на основі створених EQ. Однак у цій експериментальній конфігурації (з різним значенням *k* та $\left|\text{QIDs}\right|=3$) вони показують різні результати. Знижуючи тенденції, які можна спостерігати для метрики $C_{AV\ G}$, вказують на те, що зі збільшенням *k* середній розмір створених еквалайзерів наближається до ідеального сценарію, де розмір еквалайзерів дорівнює заданому *k*. Цей ідеальний сценарій трапляється, коли кожен запис узагальнюється та групується в EQ, що складається з *k* записів.

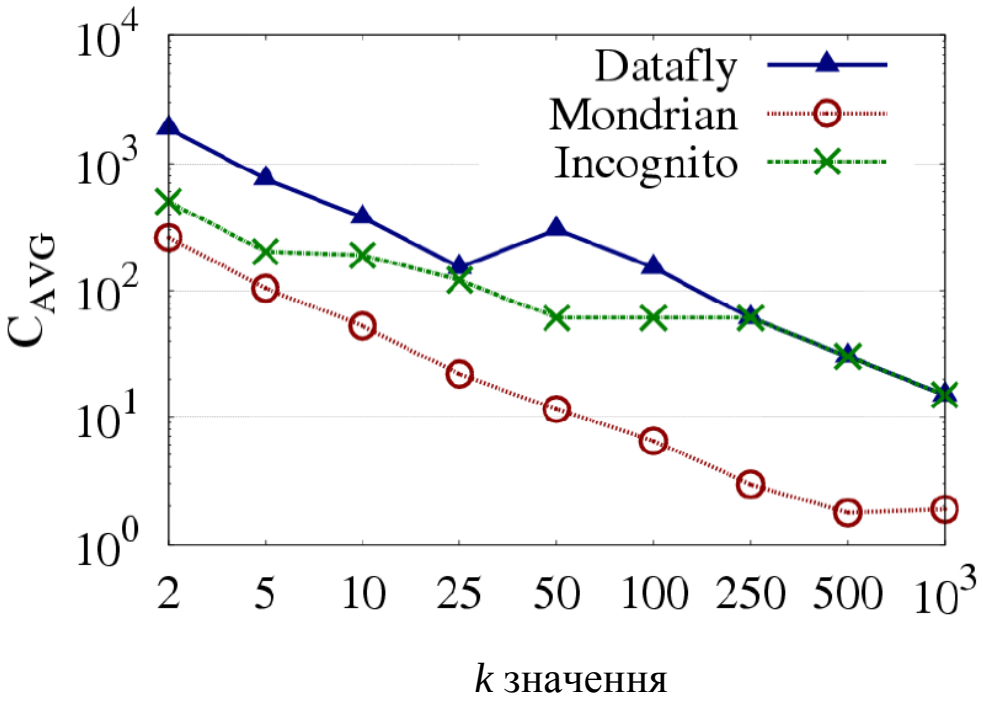


Рисунок 3.19 - $C_{AVG}$ для набору ‘Дорослі особи’ при збільшенні значення *k*

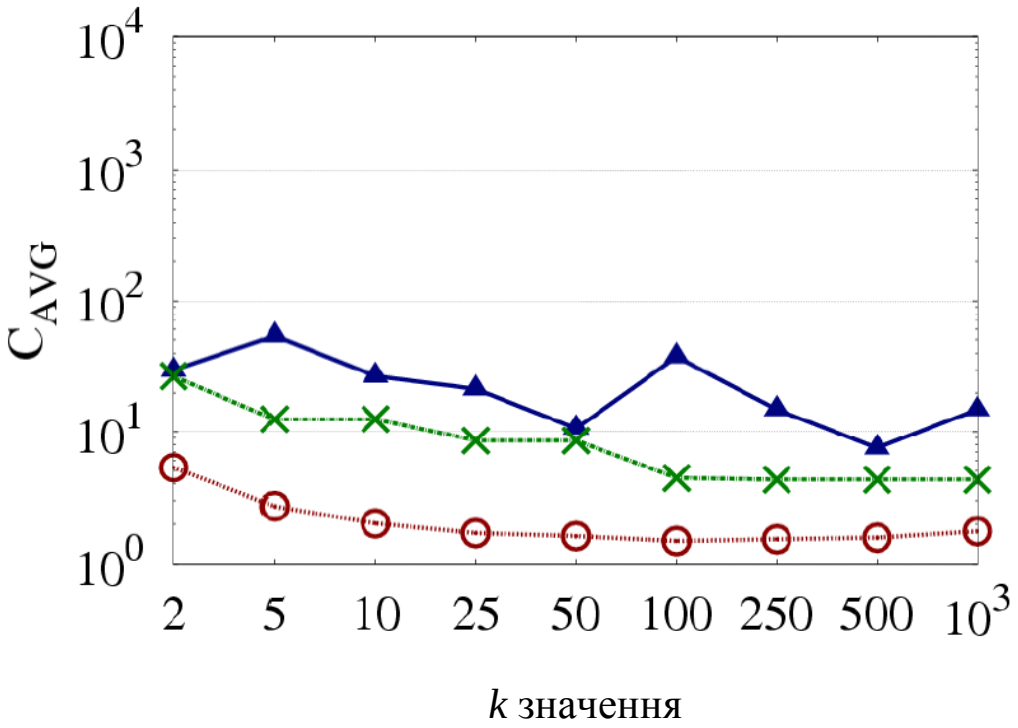


Рисунок 3.20 - $C_{AVG}$ для набору ‘Ірландський’ при збільшенні значення *k*

Результати для набору даних ‘Ірландський’ мають нижчі значення, ніж для ‘Дорослі особи’, це вказує на те, що еквалайзери, створені для набору даних

‘Ірландський’, є більш точними (менший розмір). Наприклад, розглянемо Mondrian, який, схоже, перевершує інші алгоритми. Коли *k*=2, Mondrian створює 58 еквалайзерів для ‘Дорослі особи’ і 2833 еквалайзерів для ‘Ірландський’. Це означає, що середній розмір EQ для дорослих становить 520,03, тоді як для ірландських - 10,58.

### 3.4.6 Результати порівняння алгоритмів для експерименту 2

Цей другий набір експериментів запропонував кілька цікавих висновків. Збільшення *k* мало або зовсім не впливає на ефективність алгоритмів (крім Mondrian, як було пояснено раніше), враховуючи, що виконується аналогічна кількість узагальнень/кількість шуканих вузлів. Ми можемо спостерігати, що час анонімізації зменшується із кількістю поділів. Можна також спостерігати, що Mondrian перевершує інші алгоритми щодо показників, які вимірюють розмір та кількість створених еквалайзерів (наприклад, DM та $C_{AV\ G}$). Однак для показників, які фіксують перетворення атрибутів QID (тобто GenILoss), Mondrian може працювати гірше. Це тому, що, коли дані спотворені, Mondrian не може виконувати поділи за всіма атрибутами.

## 3.5 Експеримент 3: Різноманітний розмір набору даних

У цьому експерименті ми аналізуємо масштабованість алгоритмів з точки зору часу анонімізації та споживання пам’яті у міру збільшення розміру набору даних. Ми використовуємо набір даних ‘Ірландський’ різних розмірів. Зважаючи на те, що цей набір даних генерується синтетично, ми можемо використовувати той самий розподіл даних у наших експериментах.

### 3.5.1 Час анонімізації

На рисунку 3.21 представлені результати за час анонімізації алгоритмів у міру збільшення розміру набору даних. На рисунку видно, що три алгоритми слідують плавній тенденції зростання. Зі збільшенням розміру набору даних алгоритми виконують менше узагальнень / поділів, що, як очікується, скоротить

час анонімізації. Однак тенденція до зростання, показана на рисунку, пояснюється самим обсягом даних; час, необхідний для анонімізації записів. Mondrian є найкращим виконавцем. Найбільший набір даних займає близько 370,17 секунд (трохи вище 6 хвилин) для анонімізації. Часова складність Mondrian становить O(*nlogn*) [44, 17], де *n* - кількість кортежів у вихідній таблиці. Для одного і того ж набору даних Datafly та Incognito займають приблизно 448 секунд (трохи вище 7 хвилин) і 4 444 сек (трохи вище 72 хвилин) відповідно.

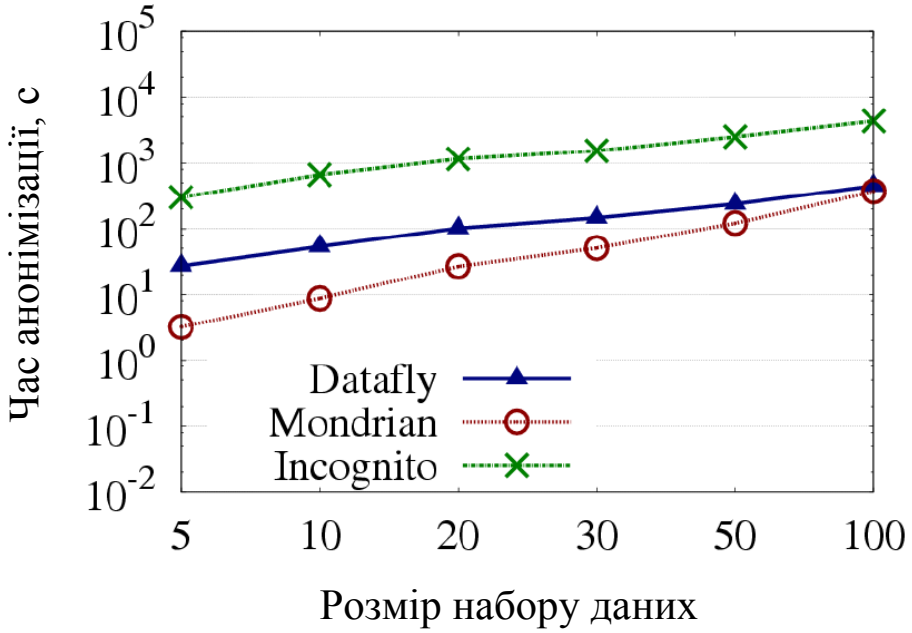


Рисунок 3.21 - Результати затрати часу для анонімізації при збільшені розміру набору ‘Ірландський’

### 3.5.2. Споживання пам'яті

На рисунку 3.22 представлено порівняння алгоритмів анонімізації з точки зору споживання пам'яті у міру збільшення розміру набору даних.

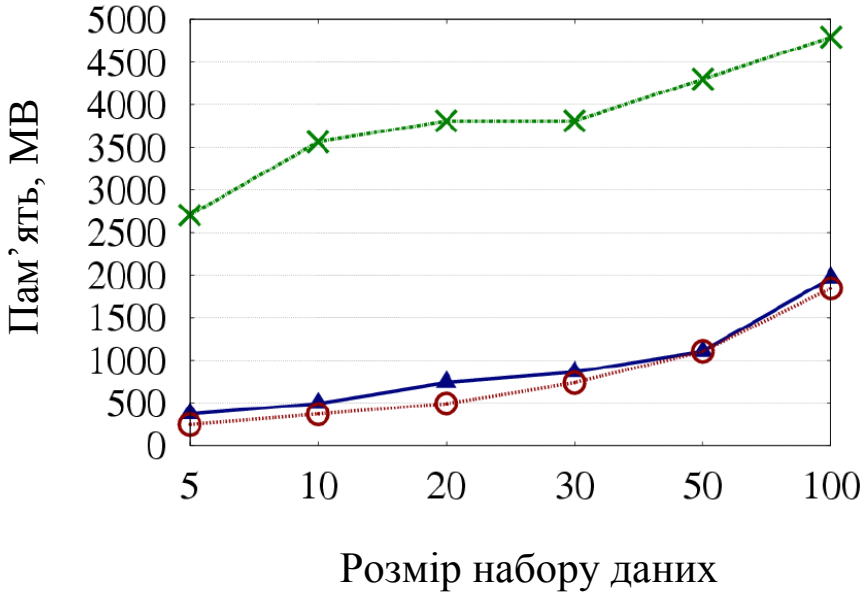


Рисунок 3.22 - Результати споживання пам'яті під час анонімізації при збільшені розміру набору ‘Ірландський’

Три алгоритми мають тенденцію до зростання, що пов'язано головним чином із збільшенням обсягу даних. Слід зазначити, що жоден Java MaGC не відбувався під час процесу анонімізації, який міг би вплинути на ефективність алгоритмів. Отже, можна спостерігати максимальне споживання пам'яті кожного алгоритму. Mondrian та Datafly демонструють подібні тенденції зростання. Incognito споживає більше пам'яті порівняно з іншими алгоритмами завдяки узагальнюючій структурі решітки, яка підтримується в основній пам'яті. Інші два алгоритми потребують пам'яті для виконання підрахунку частоти, щоб перевірити, чи задовільність *k* (для Datafly), та обчислення середнього значення для розділення (для Mondrian).

### 3.5.3 Результати порівняння алгоритмів для експерименту 3

У третьому наборі експериментів для показників ефективності для різних розмірів набору даних найкращими виконавцями залишалися Mondrian, Datafly. Цей висновок підводить нас до того, що на продуктивність алгоритмів не впливають зміни їх навантаження, доки розподіл даних є постійним між навантаженнями.

## 3.6 Порівняльний аналіз алгоритмів *k*-анонімізації

Результати у порівняльному дослідженні демонструють, як використання численних показників дозволяє ширше розуміти ефективність алгоритмів у процесі анонімізації даних.

Існують значні відмінності в роботі серед оцінюваних алгоритмів. Оцінка показала, що жоден алгоритм не перевершує інші за всіма показниками. Крім того, найкращий алгоритм для одного набору даних іноді був далеко не найкращим для іншого набору даних. Було помічено, що і вимоги конфіденційності, і розподіл даних мають великий вплив на продуктивність алгоритмів. Ці відмінності зазначені у таблиці 3.2, де показані результати за всіма показниками, що використовуються як критерії оцінювання (ефективність та корисність даних) для однієї експериментальної конфігурації (з $\left|\text{QIDs}\right| = 3$ і $k = 2$).

Таблиця 3.2 - Порівняння ефективності для всіх показників

| Набір даних | Алгоритм | Час анонімізації, с | Споживання пам'яті, МВ | GenILoss | DM | $C_{AVG}$ |
|---|---|---|---|---|---|---|
| Дорослі особи | Datafly | 16,3 | 172 | 0,42 | $2{,}1\ 10^{-8}$ | 1885 |
| | Incognito | 110,7 | 473 | 0,35 | $5{,}7\ 10^{-9}$ | 263,9 |
| | Mondrian | 7,8 | 860 | 0,67 | $2{,}7\ 10^{-9}$ | 490,1 |
| Ірландський | Datafly | 116,3 | 1357,5 | $3{,}5\ 10^{-2}$ | $4{,}6\ 10^{-6}$ | 29,7 |
| | Incognito | 528,5 | 1597 | $1{,}1\ 10^{-1}$ | $2{,}9\ 10^{-6}$ | 26,73 |
| | Mondrian | 301,3 | 1484,2 | $7{,}7\ 10^{-3}$ | $7{,}8\ 10^{-7}$ | 5,346 |

Показано, що Incognito забирає багато часу та займає пам'ять щодо розміру QID (тобто, показує експоненціальне зростання). Цей час був би більшим для більш глибоких VGH, оскільки простір пошуку ширший. Це випливає з часу,

необхідного для проходження решітки та перевірки на k-анонімність у кожному окремому стані.

На корисність даних Mondrian значно впливає розподіл даних та механізм, який використовується для розділення. Підхід середнього розподілу придатний для рівномірного розподілу, оскільки більшість атрибутів можна використовувати для розділення даних. Навпаки, коли дані спотворені, дані не можуть бути розподілені між усіма атрибутами, які впливають на показник втрати інформації. Таким чином, Mondrian в цій справі погано працює. Крім того, при анонімізації категоричних значень (які не мають чітко визначеного впорядкування) порядок, в якому обробляються значення для обчислення медіани, відіграє важливу роль у виконанні алгоритму, оскільки це може бути причиною того, що розділення є неможливо. Отже, це впливає на показник корисності даних. Mondrian припиняє його виконання, коли більше поділів не дозволено. Очікується, що ця стратегія виграє метрикою утиліти даних, яка базується на розмірі класу еквівалентності (наприклад, DM), оскільки в процесі анонімізації будуть створені більш тонкі EQ. Однак для наборів даних, які спочатку задовольняли *k*-анонімізацію, DM не покращується. Це пояснюється тим, що кількість створених класів еквівалентності може бути нижчою за кількість класів еквівалентності, які вже є у вихідному наборі даних.

Показано, що Datafly є другим найкращим виконавцем за часом анонімності, споживанням пам’яті та GenILoss. Для показників на основі розміру класу еквівалентності (тобто, DM і $C_{AV\,G}$) алгоритм був найгірший майже у всіх конфігураціях.

Щодо стратегій, що використовуються алгоритмами: Datafly та Incognito використовують узагальнення на основі ієрархії; і Mondrian використовує узагальнення на основі поділів. З точки зору практики, стратегії мають значення в двох аспектах. По-перше, у типі рішення для анонімізації, яке будуть генерувати алгоритми. По-друге, необхідні умови, які повинні надати практикуючі спеціалісти, щоб застосувати алгоритми. Що стосується виробленого рішення, Datafly та Incognito, які базуються на визначених користувачем ієрархіях узагальнення (тобто VGH), будуть виробляти

анонімізацію, яка дотримується обмежень, визначених у VGH. Mondrian навпаки, це неконтрольована анонімізація, оскільки розділення усуває всі ієрархічні обмеження, тим самим створюючи діапазони значень динамічно. Виходячи з цієї характеристики, Mondrian був би більш підходящим (і практичним) для анонімізації числових наборів даних [5]. Це тому, що у випадку категоричних атрибутів будь-яка семантика, пов’язана зі значеннями, може бути порушена (наприклад, згрупувати країни, що належать до одного континенту), оскільки дані поділяються без контролю над створеними групами.

Щодо ефективності алгоритмів: (1) З точки зору $\left|\text{QIDs}\right|$, Datafly та Mondrian - найкращі варіанти, коли кількість QID є великою. Incognito навпаки, він погано виконує заходи щодо ефективності, особливо коли решітка узагальнення велика і в атрибутах є мінливість (чіткі значення). (2) Що стосується рівня *k*-анонімності, Mondrian виступив найкраще у міру збільшення необхідного *k*. Загалом, три алгоритми добре впоралися із збільшенням *k*, оскільки не було жодного значного збільшення часу для алгоритмів. Mondrian навіть показав покращення часу, коли рівень *k* зростав. Щодо пам'яті, то три алгоритми показали відносно стабільне споживання. (3) Що стосується розміру набору даних, Mondrian та Datafly виступили краще, оскільки розмір набору даних збільшувався, демонструючи менші масштаби зростання часу анонімізації та споживання пам’яті.

Щодо ефективності алгоритмів: Mondrian був показаний найкращим виконавцем щодо показників на основі розміру групи (тобто, DM та $C_{AV\,G}$). Це означає, що Mondrian пропонує більш точну деталізацію в класі еквівалентності, що, як очікується, підвищить точність результатів даних. Винятком є випадок, коли вихідний набір даних вже задовольняє k-анонімініть, оскільки корисність даних щодо цих показників може зменшуватися. Для показника, що фіксує перетворення атрибутів (тобто, GenILoss), хоча не існує чіткого найкращого виконавця, Incognito та Mondrian показали хороші результати. Зокрема, Mondrian краще підходить, якщо вихідні дані дотримуються рівномірного розподілу.

# РОЗДІЛ 4 СПЕЦІАЛЬНА ЧАСТИНА

## 4.1 Вибір набору даних

Зростання доступності мікроданих викликає зацікавленість організацій збирати та обмінюватися цими даними з метою їх аналізу. Однак чинне законодавство вимагає захисту персональних даних від неналежного розкриття інформації. Методи анонімізації допомагають безпечно поширювати дані, зберігаючи достатню корисність для їх повторного використання. З цієї причини в літературі існує безліч методів анонімізації мікроданих. Кожна з цих методик має особливі переваги над іншими (наприклад, покращення корисності даних або обчислювальних ресурсів). Однак їх ефективність може змінюватись при їх тестуванні з різними наборами даних. Цей факт ускладнює узагальнення висновків оцінки продуктивності для вирішення питання, який алгоритм найкраще відповідає їх вимогам.

Усі оцінки в тій чи іншій мірі обмежені (через час / зусилля / обмеження витрат). Поширене обмеження – кількість та різноманітність використовуваних наборів даних, оскільки процес отримання якісних наборів даних може бути обтяжливим і трудомістким. Наприклад, доступ до реальних мікроданих є дуже обмеженим для захисту конфіденційності приватних осіб. Агенції надають доступ лише на певний період, і як тільки цей термін закінчується, надані файли повинні бути знищені, що не завжди дозволяє відтворити експериментальні результати.

Дослідники часто використовують реальні набори даних, які є загальнодоступними, або синтетичні дані, що генеруються спеціальним чином. Обидва рішення в деякій мірі є упередженими, оскільки обмежуються специфічним розподілом даних: або реальним, що складається з єдиного сценарію (наприклад, демографії однієї країни), або ідеальні синтетичні дані, яких ніколи не можна знайти в реальному світі. Більше того, застосування більшості реальних наборів даних часто обмежується для тестування методів анонімізації, оскільки набори даних, доступні в дослідженнях, зазвичай

узагальнені та попередньо анонімізуються для захисту конфіденційності людей (що саме полягає у використанні методів збереження конфіденційності). Таким чином, цим даним бракує достатньої різноманітності для імітації сценаріїв атак або для оцінки надійності запропонованих технологій (наприклад, розрідженість даних). З цих причин синтетичні дані є цінним ресурсом для проведення тестування в декількох областях, оскільки вони маніпулюють даними для задоволення конкретних характеристик, які не знайдені в реальних даних, але все-таки потребують врахування для тестування (гіпотетичний сценарій майбутнього) стійкості щодо атак. Щоб бути рівноцінними замінниками реальних даних, необхідно зберегти функціональні залежності даних.

Дослідницька робота [48] була зосереджена на розробці техніки для створення реалістичних синтетичних наборів даних, застосовних до домену конфідеційності.

## 4.2 COCOA: Генератор синтетичний даних

Метою дослідницької роботи [48] було розробити основу для генерації синтетичного набору даних (COCOA), яка може бути використана для диверсифікації набору характеристик, наявних у мікроданих. Ця стратегія допоможе покращити тестування методів анонімізації. На рисунку 4.1 зображено концептуальний вигляд рішення. Видно, як COCOA слідує за ітераційним процесом для створення набору наборів даних, заснованих на базі інформації, що надається користувачем. Інформаційна база складається з усіх вхідних параметрів, необхідних вибраному набору даних (наприклад, розмір набору даних)

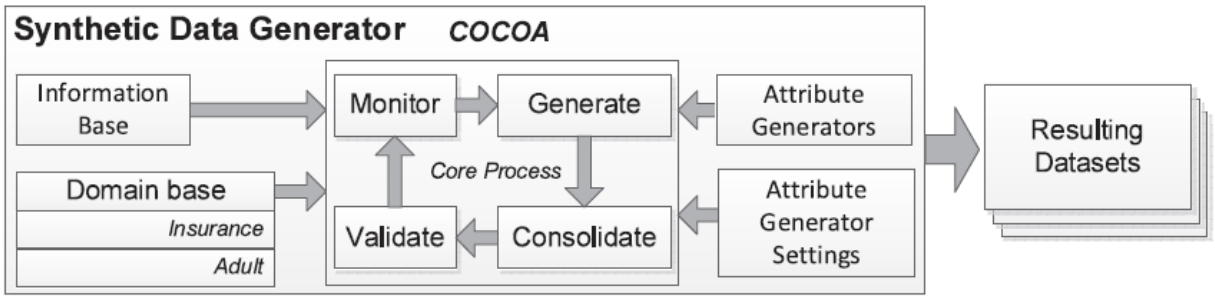


Рисунок 4.1 - Концептуальний вигляд COCOA

Ключовим елементом COCOA є його доменна база, яка охоплює експертні знання про підтримувані бізнес-сфери (наприклад, перепис населення, охорона здоров'я, фінанси). Цей елемент дозволяє COCOA бути легко розширюваним і здатним включати кілька бізнес-кейсів (навіть для одного і того ж домену, наприклад, перепис Ірландії, перепис США), що може бути придатним для різних тестових сценаріїв. У цьому контексті домен визначає правила та обмеження, необхідні для створення набору даних, що зберігає функціональні залежності бізнесу. Кожен домен характеризується іменем, набором атрибутів та відповідними їм генераторами атрибутів.

Для створення даних домени використовують наявний набір генераторів атрибутів. Ці елементи підтримують логіку, яка пропонує неоднозначні стратегії для генерування значень атрибутів. Наприклад, генератор може зосередитись на диверсифікації даних в атрибуті шляхом встановлення його в розподіл даних (наприклад, звичайний розподіл). У цьому прикладі можуть бути об'єднані кілька генераторів, які пропонують різні розподіли даних (оскільки відповідний розподіл може змінюватися залежно від сценарію використання). У випадку, якщо генератору атрибутів потрібні будь-які конкретні параметри для роботи належним чином (наприклад, значення за замовчуванням для його застосовуваних параметрів), ця інформація (наприклад, середнє значення та дисперсія для нормального розподілу) також може бути захоплена рамкою (як налаштування генератора атрибутів).

## 4.3 Генератори атрибутів

Як було зазначено у розділі 4.2, генератори атрибутів підтримують логіку, яка пропонує різні стратегії для генерування значень атрибутів. Серед альтернативних варіантів розробки генераторів атрибутів для COCOA автори спочатку зосередилися на впровадженні наступних трьох:

- генератор на основі розподілу;
- генератор на основі атрибутів;
- генератор на основі розподілу та атрибутів.

Генератор на основі розподілу – тип генератора який виробляє незалежні дані, тому він застосовується для атрибутів, значення яких не залежать від інших. У цьому випадку стратегія, яка використовується для диверсифікації даних, полягає у імітації розподілу ймовірностей. Цей генератор також може бути хорошим пристосуванням для даних, що надходять із попередньо консолідованого реального джерела даних (подібність до наявних наборів даних). Це пояснюється тим, що об'єктивними випадками цих випадків є звичайна диверсифікація частот існуючих значень без зміни фактичних значень або потужності множини атрибута. Генератору на основі розподілу потрібні два параметри: розподіл (який буде використовуватися для повторного розподілу частоти значень) та стратегії асортименту (яка буде використовуватися для впорядкування значень-кандидатів перед застосуванням розподілу). Для початкової версії COCOA підтримуються 11 поширених розповсюджених даних: нормальний, бета, chi, chisquare, експоненціальний (exp), гамма, геометричний, логарифмічний (log), пуассон, t-студент (Tstu) та рівномірний (uni ) [49]. Додатковим підтримуваним розподілом є "оригінальний". Коли він використовується, генератор лише відображає заданий вхідний розподіл. Це корисно для генерування нових наборів даних (різних розмірів) для існуючих доменів без зміни оригінального розподілу, наявного в реальних даних. Переосмисливши стратегії сортування, COCOA в даний час підтримує три: алфавітно-числове, зворотнє алфавітно-числове та без сортування. Використання алфавітно-числового та зворотнього алфавітно-числового

дозволяє користувачеві не тільки логічно сортувати категоричні значення, але й допомагає надалі урізноманітнити перевірену поведінку шляхом змішування підтримуваних видів та розподілів. Наприклад, використання буквено-числового сортування та експоненціального розподілу може генерувати J-подібний розподіл, тоді як зворотний J-подібний розподіл може бути легко сформований шляхом переходу на зворотнє алфавітно-числове сортування. Стратегія без сортування залишає оригінальний порядок даних.

Генератор на основі атрибутів – це тип генератора який виробляє залежні дані. Він застосовується у випадках, коли існують функціональні залежності, які необхідно зберегти між атрибутами (щоб згенеровані дані могли бути реалістичними), а також у випадках, коли атрибут потрібно отримати від інших. Наприклад, розглянемо набір даних з інформацією про осіб, які належать страховій компанії. Мета – отримати атрибут класу, який визначає групи з більш високим ризиком нещасного випадку. Одним із способів побудови цього класу (наприклад, низький, середній, високий) є використання значень від розподілів, таких як вік, професія та захоплення. Параметри, необхідні для цього типу генератора, можуть змінюватися, але зазвичай вони будуть приймати необхідну вхідну інформацію з інших атрибутів. COCOA використовує генератори в такому порядку, щоб генератор на основі атрибутів мав необхідну інформацію (тобто атрибути, від яких залежить) перед його виконанням.

Генератор на основі розподілу та атрибутів – це тип генератора, який також виробляє залежні дані. Це тому, що це гібрид попередніх двох генераторів. Корисно зафіксувати більш складні зв'язки, коли на генерацію значення впливає не тільки розподіл частоти, але й значення одного або декількох атрибутів. Наприклад, місце, де здійснюється діяльність, залежить не тільки від сформованої діяльності, але і від певного розподілу. Наприклад, спираючись на історичну інформацію, футбол в основному практикується на зовнішніх полях, а в меншій мірі - в інших місцях, таких як критий майданчик або пляж. Інший приклад – зарплата людини, на яку впливають численні фактори, такі як її професія та багаторічний стаж роботи.. Цей тип відносин може бути легко захоплений у COCOA цим типом генератора.

# РОЗДІЛ 5 ОБҐРУНТУВАННЯ ЕКОНОМІЧНОЇ ЕФЕКТИВНОСТІ

Метою дипломної роботи є порівняння найбільш відомих методів *k*-анонімізації (Datafly, Incognito, Mondrian) з огляду на використання ресурсів та корисність залишкових даних.

## 5.1 Розрахунок норм часу на виконання науково-дослідної роботи

Ефективне використання часу має велике значення тому, що коефіцієнт корисної дії залежить від оптимального використання часу.

Аналіз алгоритмів анонімізації поділено на декілька етапів, що дозволяє полегшити і структурувати виконання поставленого завдання.

Основні етапи такі:

1. пошук літературних джерел з області дослідження;

2. дослідження наборів даних;

3. порівняльний аналіз алгоритмів анонімізування за різними параметрами.

Для оцінки тривалості виконання окремих робіт використовують нормативи часу.

Витрати часу по окремих операціях технологічного процесу відображені в таблиці 5.1.

Таблиця 5.1 – Операції технологічного процесу та їх час виконання

| № п/п | Назва операції (стадії) | Виконавець | Середній час виконання операції, год. |
|---|---|---|---|
| 1. | Пошук літературних джерел з області дослідження. | інженер | 26 |
| 2. | Дослідження наборів даних. | інженер | 48 |
| 3. | Порівняльний аналіз алгоритмів анонімізування за різними параметрами. | інженер | 52 |
| Разом | | | 126 |

Загальні затрати часу на реалізацію даної роботи становить 126 години, найбільш трудомістким є сам порівняльний аналіз – 52 годин.

## 5.2 Визначення витрат на оплату праці та відрахувань на соціальні заходи

Відповідно до Закону України “Про оплату праці” заробітна плата – це “винагорода, обчислена, як правило, у грошовому виразі, яку власник або уповноважений ним орган виплачує працівникові за виконану ним роботу”.

Розмір заробітної плати залежить від складності та умов виконуваної роботи, професійно-ділових якостей працівника, результатів його. Заробітна плата складається з основної та додаткової оплати праці.

Основна заробітна плата нараховується за виконану роботу за тарифними ставками, відрядними розцінками чи посадовими окладами.

Додаткова заробітна плата – це складова заробітної плати працівників, до якої включають витрати на оплату праці, не пов’язані з виплатами за фактично відпрацьований час. Нараховують додаткову заробітну плату залежно від досягнутих і запланованих показників, кваліфікації виконавців. Джерелом додаткової оплати праці є фонд матеріального стимулювання, який створюється за рахунок прибутку.

При розрахунку заробітної плати кількість робочих днів у місяці слід в середньому приймати – 24,5 дні/міс., або ж 196 год./міс. (тривалість робочого дня – 8 год.).

Місячний оклад кожного працівника слід враховувати згідно існуючих на даний час тарифних окладів. Згідно закону України «Про Державний бюджет України на 2019 рік», зокрема статтею восьмою мінімальна заробітна плата у погодинному розмірі становить 25,13 грн. Рекомендовані тарифні ставки: керівник дипломної роботи – 30,00...50,00 грн./год., інженер – 25,13...30,00 грн./год., консультант – 25,13...30,00 грн./год., технік – 25,13...30,00 грн./год., лаборант – 25,13...26,00 грн./год.

Основна заробітна плата розраховується за формулою:

$$З_{осн.} = Т_с \cdot К_г \,, \quad (5.1)$$

де $Т_с$ – тарифна ставка, грн.; $К_г$ – кількість відпрацьованих годин.

Оскільки всі види робіт в виконує розробник, то основна заробітна плата буде розраховуватись тільки за однією формулою

$$З_{осн.} = 25{,}13 \cdot 126 = 3166{,}38 \text{ грн.}$$

Додаткова заробітна плата становить 10–15 % від суми основної заробітної плати.

$$З_{дод.} = З_{осн.} \cdot К_{допл.}, \quad (5.2)$$

де $К_{допл}$ – коефіцієнт додаткових виплат працівникам, 0,1–0,15 (візьмемо його рівним 0,15).

$$З_{дод} = 3166{,}38 \cdot 0{,}15 = 474{,}65 \text{ грн.}$$

Звідси загальні витрати на оплату праці ($В_{о.п.}$) визначаються за формулою:

$$В_{о.п.} = З_{осн.} + З_{дод.} \quad (5.3)$$

$$В_{о.п.} = 3166{,}38 + 474{,}65 = 3641{,}03 \text{ грн.}$$

Крім того, слід визначити відрахування на соціальні заходи:

- єдиний соціальний внесок ЄСВ (прибутковий податок) – 22%;
- військовий збір – 1,5%.

У сумі зазначені відрахування становлять 23,5 %.

Отже, сума відрахувань на соціальні заходи буде становити:

$$В_{с.з.} = Ф_{оп} \cdot 0{,}235 \quad (5.4)$$

де $Ф_{оп}$ – фонд оплати праці, грн.

$$В_{с.з.} = 3641{,}03 \cdot 0{,}235 = 855{,}64 \text{ грн.}$$

Проведені розрахунки витрат на оплату праці наведено у таблицю 5.2.

Таблиця 5.2 – Розрахунки витрат на оплату праці

| з/п | Категорія працівників | Основна заробітна плата, грн. | | | Додаткова заробітна плата, грн. | Відрахування $Ф_{ОП}$, грн. | Всього витрати на плату праці, грн. (6=3+4+5) |
|---|---|---|---|---|---|---|---|
| | | Тарифна ставка, грн. | Кількість відпрацьованих год. | Фактично нарах. з/пл., грн. | | | |
| А | Б | 1 | 2 | 3 | 4 | 5 | 6 |
| 1. | Інженер (розробник) | 25,13 | 126 | 3166,38 | 474,65 | 855,64 | 4496,67 |

З таблиці розрахунки витрат на оплату праці видно що всього витрати на плату праці становить 4496,67грн.

## 5.3 Розрахунок матеріальних витрат

Матеріальні витрати визначаються як добуток кількості витрачених матеріалів та їх ціни:

$$M_{вi} = q_i \cdot p_i \,, \tag{5.5}$$

де: $q_i$ – кількість витраченого матеріалу і–го виду; $p_i$ – ціна матеріалу і–го виду.

Звідси, загальні матеріальні витрати можна визначити:

$$З_{м.в.} = \sum M_{вi} \,. \tag{5.6}$$

Розрахунки занесемо у таблицю 5.3.

Таблиця 5.3 – Розрахунки матеріальних витрат

| Найменування матеріальних ресурсів | Один. виміру | Норма витрат | Ціна за один., грн. | Затрати матер., грн. | Транс-портно-заготівель-ні витрати, грн. | Загальна сума витрат на матер., грн. |
|---|---|---|---|---|---|---|
| 1. Основні матеріали | | | | | | |
| Використання мережі Internet | години | 120 | – | 120 | – | 120 |
| 2. Допоміжні витрати | | | | | | |
| Папір формату А4 | шт. | 160 | 0,3 | 48 | – | 48 |
| Разом: | | | | | | 168 |

Загальні матеріальні витрати на Internet і Папір формату А4 становить 168 грн.

5.4 Розрахунок витрат на електроенергію

Затрати на електроенергію 1–ці обладнання визначаються за формулою:

$$З_в = W \cdot T \cdot S, \quad (5.7)$$

де $W$ – необхідна потужність, кВт; $T$ – кількість годин на реалізацію розробки; $S$ – вартість кіловат-години електроенергії.

Вартість кіловат-години електроенергії слід приймати згідно існуючих на даний час тарифів. Отже, 1 кВт з ПДВ коштує 1,68грн.

Потужність комп’ютера для створення дипломної роботи – 90 Вт, кількість годин роботи обладнання згідно таблиці 5.1 –126 години.

Тоді,

$$З_в = 0{,}09 \cdot 126 \cdot 1{,}68 = 19{,}05 \text{ грн.}$$

Згідно формули затрати на електроенергію де необхідна потужність множиться на кількість годин на реалізацію розробки і множиться на вартість кіловат-години електроенергії що в висновку дорівнює 19,05 грн.

## 5.5 Розрахунок суми амортизаційних відрахувань

Характерною особливістю застосування основних фондів у процесі виробництва є їх відновлення. Для відновлення засобів праці у натуральному виразі необхідне їх відшкодування у вартісній формі, яке здійснюється шляхом амортизації.

Амортизація – це процес перенесення вартості основних фондів на вартість новоствореної продукції з метою їхнього повного відновлення.

Для визначення амортизаційних використовується формула:

$$A = \frac{Б_В \cdot Н_А}{100\%}, \qquad (5.8)$$

де $A$ – амортизаційні відрахування за звітний період, грн.; $Б_В$ – балансова вартість групи основних фондів на початок звітного періоду, грн.; $Н_А$ – норма амортизації.

Комп’ютери та оргтехніка належать до четвертої групи основних фондів. Для цієї групи річна норма амортизації дорівнює 60 % (квартальна – 15 %).

Для даної дипломної роботи засобом розробки є комп’ютер. Його сума становить 12480 грн. Отже, амортизаційні відрахування будуть рівні:

$$A = 12480 \cdot 5\% / 100\% = 624{,}00 \text{ грн.}$$

Оскільки робота виконувалась 126 години, то амортизаційні відрахування будуть становити:

$$A = 624{,}00 \cdot 126 / 126 = 624{,}00 \text{ грн.}$$

Згідно формули для визначення амортизаційних де $Б_В$ множиться $Н_А$ і ділиться на 100% амортизація розробки становить 624,00 грн.

5.6 Обчислення накладних витрат

Накладні витрати пов’язані з обслуговуванням виробництва, утриманням апарату управління спілкою та створення необхідних умов праці.

В залежності від організаційно-правової форми діяльності господарюючого суб’єкта, накладні витрати можуть становити 20–60 % від суми основної та додаткової заробітної плати працівників.

$$H_в = B_{о.п.} \cdot 0{,}2\ldots0{,}6\ , \qquad (5.9)$$

де $H_в$ – накладні витрати.

Отже, накладні витрати:

$$H_в = 4496{,}67 \cdot 0{,}2 = 897{,}33 \text{ грн.}$$

Накладні витрати згідно розрахунку формули, становить 897,33 грн.

5.7 Складання кошторису витрат та визначення собівартості науково-дослідницької роботи

Результати проведених вище розрахунків зведемо у таблицю 5.4.

Таблиця 5.4 – Кошторис витрат на НДР

| Зміст витрат | Сума, грн. | В % до загальної суми |
|---|---|---|
| Витрати на оплату праці $B_{о.п}$ | 3641,03 | 58,68 |
| Відрахування на соціальні заходи $B_{с.з}$ | 855,64 | 13,79 |
| Матеріальні витрати $З_{м.в}$ | 168,00 | 2,71 |
| Витрати на електроенергію $З_в$ | 19,05 | 0,30 |
| Амортизаційні відрахування $А$ | 624,00 | 10,06 |
| Накладні витрати $H_в$ | 897,33 | 14,46 |
| Собівартість $C_в$ | 6205,05 | 100,00 |

Собівартість ($C_в$) роботи розрахуємо за формулою:

$$C_в = B_{о.п.} + B_{с.з.} + З_{м.в.} + З_в + А + Н_в \text{ .} \quad (5.10)$$

Отже, собівартість роботи дорівнює:

$C_в$ = 3641,03+855,64+168+19,05+624,00+897,33= 6205,05 грн.

Загальний кошторис витрат та визначення собівартості науково-дослідницької роботи становить 6205,05 грн.

## 5.8 Розрахунок ціни науково-дослідної роботи

Ціну науково-дослідної роботи можна визначити можна визначити за формулою:

$$Ц = C_В \cdot (1 + Р_{рен}) \cdot (1 + ПДВ) \quad (5.11)$$

де $Р_{рен.}$ – рівень рентабельності, 30 %, $ПДВ$ – ставка податку на додану вартість, (20 %).

Звідси ціна на роботу складе:

$Ц$ = 6205,05· (1 + 0,3) · (1 + 0,2) = 9679,88 грн.

Загальний розрахунок ціни програмного продукту становить 9679,88 грн.

## 5.9 Визначення економічної ефективності і терміну окупності капітальних вкладень

Ефективність виробництва – це узагальнене і повне відображення кінцевих результатів використання робочої сили, засобів та предметів праці на підприємстві за певний проміжок часу.

Економічна ефективність ($Е_р$) полягає у відношенні результату виробництва до затрачених ресурсів:

$$Е_р = \frac{П}{С_В}, \quad (5.12)$$

де $П$ – прибуток; $С_В$ – собівартість.

Плановий прибуток ($П_{пл}$) знаходимо за формулою:

$$П_{пл} = Ц - С_в \text{ .} \quad (5.13)$$

Розраховуємо плановий прибуток:

$$П_{пл} = 9679{,}88 – 6205{,}05 = 3474{,}83 \text{ грн.}$$

Отже, формула для визначення економічної ефективності набуде вигляду:

$$E_p = \frac{П_{пл}}{C_в}. \qquad (5.14)$$

Тоді,

$$E_p = 3474{,}83 / 6205{,}05 = 0{,}56.$$

Поряд із економічною ефективністю розраховують термін окупності капітальних вкладень ($T_p$):

$$T_p = \frac{1}{E_p}\ , \qquad (5.15)$$

Термін окупності дорівнює:

$$T_p = 1 / 0{,}56 = 1{,}8 \text{ р.}$$

Згідно формул плановий прибуток від розробки становить 3474,83 грн., економічна ефективність дорівнює 0,56 а термін окупності становить 1,8 роки що вважається доцільним та економічно вигідним.

# РОЗДІЛ 6 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

## 6.1 Охорона праці

Дослідження захисту персональної інформації в задачах аналізу та обробки великих даних відбувалося з використанням персональних комп'ютерів. Тому надзвичайно важливим фактором безпеки праці є дотримання правил користування комп'ютерною технікою, норм та правил охорони праці. Потрібно забезпечити користувачам максимально комфортні та безпечні умови для їх перебування в приміщенні та якісного, ефективного виконання поставлених завдань.

Працівники установ, де проводиться захист персональної інформації, повинні дотримуватися правил внутрішнього трудового розпорядку, виконувати правила особистої гігієни та гігієни приміщення, володіти знаннями з техніки безпеки користування технікою, електробезпеки, пожежної безпеки та не виконувати дії, які суперечать правилам охорони праці.

Будівлі та приміщення, де розміщені робочі місця для працівників, що проводять захист персональної інформації, повинні відповідати вимогам нормативно-технічної та експлуатаційної документації виробника персональних комп'ютерів ДСанПіН 3.3.2-007-98 [50].

Відповідно до встановлених санітарно-гігієнічних норм (ГОСТ 12.1.005-88) [51] регламентуються вимоги до приміщення, де буде проводитись захист персональної інформації:

- площа, відведена на одне робоче місце має становити не менше 6 кв.м, а об'єм – не менше 20 куб.м;
- для облаштування потрібно використовувати лише ті матеріали, які пройшли відповідну сертифікацію, не містять шкідливих речовин та дозволені для використання в приміщеннях;

- ергономічне розташування робочого місця, виробничих меблів з урахуванням антропометричних характеристик людини; раціональне компонування обладнання на робочих місцях;
- достатня освітленість робочих місць;
- гарантії електро- та пожежо- безпеки;
- приміщення не повинно межувати з джерелами шуму і вібрації , що перевищують допустимі норми;
- щодня має здійснюватися вологе прибирання та регулярне провітрювання приміщення.

Для внутрішнього оздоблення приміщень з персональними комп‘ютерами слід обирати світлі нейтральні кольори стін. Покриття підлоги та поверхня має бути рівною, неслизькою, з антистатичними властивостями.

Основним нормативним документом, що регулює забезпечення охорони праці користувачів комп’ютерної техніки є «Державні санітарні норми електронно-обчислювальних машин» ДСанПіН 3.3.2.007-98. У виробничих приміщеннях та на робочих місцях з ВДТ та ПК мають бути забезпечені оптимальні значення параметрів мікроклімату – температури повітря, відносної вологості, швидкості руху повітря. Для цього приміщення, в яких розташовані комп'ютеризовані робочі місця повинні бути обладнані системами опалення, кондиціонування, які автоматично підтримують задані параметри мікроклімату.

Згідно ГОСТ 12.1.005-88 “ Загальні санітарно-гігієнічні вимоги до повітря робочої зон”, температура навколишнього середовища повинна бути в межах +18 - + 22 ℃, відносна вологість повітря близько 55% швидкість руху повітря – 0,1-0,2 м/сек. Рівні позитивних і негативних іонів у повітрі мають відповідати санітарно-гігієнічним нормам №2152-80. Допустима інтенсивність шуму на робочих місцях з ЕОМ має відповідати вимогам ДСанПіН 3.3.2-007-98: оптимальна — до 45 дБ, гранична — до 60 дБ.

Потрібно створити сприятливі умови для зорової роботи, які б мінімізували втому очей, виникнення професійних захворювань та сприяли підвищенню продуктивності праці. Тому освітлення повинне відповідати

вимогам ДБН В.2.5-28:2018 «Природне і штучне освітлення» [52]. Основною вимогою є необхідність створення на робочій поверхні освітленості, що відповідає характеру зорової роботи і знаходиться в межах встановлених норм. Освітлення у приміщенні має бути суміщеним. Відтак, недостача денного природнього освітлення компенсується необхідною для приміщення кількістю штучного освітлення. Як джерело штучного освітлення в приміщеннях, де встановлено комп'ютерну техніку, бажано використовувати люмінесцентні лампи. Освітленість робочого місця у горизонтальній площині на висоті 0,8 м від рівня підлоги повинна бути не менше 400 лк. Для захисту від прямих сонячних променів повинні бути передбачені сонцезахисні пристрої, жалюзі, штори. Безпосередньо при виконанні роботи з обладнанням можливе використання місцевого освітлення в комбінації з загальним.

Одним із важливих параметрів охорони праці в ході експлуатації ЕОМ є ергономіка користування. Розробляючи програму, що містить текстову інформацію, важливо врахувати фізіологічні властивості користувачів. Важливо ретельно підійти до питання вибору розмірів іконок та рисунків, кольорової гами. Так, для кращого сприймання та розрізнення поданої інформації на дисплеї, рекомендовано виконати текст в темних тонах на світлому фоні. Таким чином збільшується акцент на поданій інформації.

При досліджені захисту персональної інформації в задачах аналізу та обробки великих даних користуються лініями електромереж. Персональні комп'ютери і периферійні пристрої повинні підключатися до електромережі тільки за допомогою справних штепсельних з'єднань і електророзеток заводського виготовлення. В них, окрім контактів фазового та нульового робочого провідників, мають бути спеціальні контакти для підключення нульового захисного провідника. Конструкція вилки має бути такою, щоб приєднання нульового захисного провідника відбувалося раніше, ніж приєднання фазового та нульового робочого провідників. Усі електроприлади, згідно з ДНАОП 0.00-1.21-98 [53], повинні бути заземленні за допомогою нульового захисного провідника.

Заземлені конструкції, що знаходяться в приміщеннях, де розміщені робочі місця (батареї опалення, водопровідні труби, кабелі із заземленим відкритим екраном), мають бути надійно захищені діелектричними щитками або сітками з метою недопущення потрапляння працівника під напругу.

Під час монтажу та експлуатації ліній електромережі необхідно повністю унеможливити виникнення електричного джерела загоряння внаслідок короткого замикання та перевантаження проводів, обмежувати застосування проводів з легкозаймистою ізоляцією.

Приміщення мають бути оснащені системою автоматичної пожежної сигналізації і вогнегасниками відповідно до наказу «Про затвердження правил експлуатації та типових норм належності вогнегасників» 15.01.2018 № 25. Проходи до засобів пожежогасіння мають бути вільними. Згідно техніки пожежної безпеки пристрої повинні розташовуватися не ближче одного метра від джерел тепла. Також на них не повинні падати прямі сонячні промені, щоб виключити можливість перегріву компонентів та вбудованих акумуляторів. Адже уже відомі приклади небезпек, які спричинили нагріті до критичних температур акумуляторні батареї.

Щоразу, як виникає необхідність для роботи з електронними пристроями, потрібно дотримуватися відповідних правил:

- перед початком роботи потрібно оглянути робоче місце на наявність пошкоджень, диму, неприємного запаху; перевірити правильність під’єднання обладнання до електричної мережі;
- виконувати роботу лише передбачену регламентом робіт;
- підтримувати порядок і чистоту робочого місця;
- устаткування, що використовується та працює від електромережі, повинне бути заземленим;
- при будь-яких випадках порушень роботи технічного обладнання або програмного забезпечення негайно викликати представника технічної служби з питань експлуатації обчислювальної техніки.

Таким чином дослідження захисту персональної інформації в задачах аналізу та обробки великих даних відбувалося з збереженням правил пожежної безпеки та всіх норм охорони праці.

## 6.2 Фактори, що впливають на функціональний стан користувачів комп'ютерів

Надійність системи "людина – комп'ютер" значною мірою визначається функціональним станом людини. Психофізіологічні та емоційні перенапруження, втома людини-оператора можуть призвести в комп'ютеризованих системах керування до помилок і як наслідок – до значних економічних втрат.

Згідно зі статистичними даними від 40 до 75% аварій літаків зумовлено людським фактором. Відмови комп'ютеризованої системи керування рухом залізничного транспорту, на гірничо-збагачувальних комбінатах з вини операторів становлять понад 50% їх загальної кількості, причому значна їх частина спричинена невідповідністю функціонального стану оператора складності виконуваної роботи.

Визначення та вивчення факторів, що впливають на функціональний стан користувачів комп'ютерів дозволить виділити основні причини виникнення станів напруженості, стомлення, стресу і здійснити відповідні профілактичні заходи.

Трудова діяльність користувачів комп'ютерів відбувається у певному виробничому середовищі, яке впливає на їх функціональний стан. Найбільш значимі – фізичні фактори виробничого середовища, до яких належать електромагнітні хвилі різних частотних діапазонів, електростатичні поля, шум, параметри мікроклімату та ціла низка світлотехнічних показників.

Трудовий процес суттєво впливає на психофізіологічні можливості користувачів комп'ютерів, оскільки їх діяльність характеризується значними статичними фізичними навантаженнями; недостатньою руховою активністю; напруженнями сенсорного апарату, вищих нервових центрів, які забезпечують

функції уваги, мислення, регуляції рухів. Окрім того, трудовий процес користувачів комп’ютерів відзначається значними інформаційними навантаженнями.

Професійні якості та виробничий досвід, які визначають внутрішні засоби діяльності, обумовлюють надійну та безпомилкову діяльність користувачів комп’ютерів, дозволяють знаходити безпечні методи розв’язання виробничих завдань навіть у нестандартних ситуаціях.

Зовнішні засоби діяльності, які в основному визначаються ергономічними показниками щодо організації робочого місця, формою та параметрами його елементів, просторового розташування основного і допоміжного устаткування, можуть суттєво знизити фізичні та психофізіологічні навантаження, що діють на користувачів комп’ютерів.

Оскільки робота користувачів комп’ютерів найчастіше проходить за активної взаємодії з іншими людьми, то виникають питання раціоналізації міжособових відносин. Цей комплекс питань порушує як психологічні, так і соціально-психологічні аспекти трудових взаємовідносин, які також є факторами "ризику", що відчутно впливають на функціональний стан користувачів комп’ютерів.

Дослідження, проведені фахівцями Всесвітньої організації охорони здоров'я (ВООЗ) показали, що у професійних операторів та канцелярських службовців, які у своїй діяльності використовують ВДТ, частіше зустрічаються порушення органів зору, опорно-рухового апарату, центральної нервової, серцево-судинної, імунної та статевої систем, захворювання шкіри. Необхідно зазначити, що вже в перші роки впровадження ВДТ в Європі та США була зафіксована значна кількість скарг операторського персоналу на загальне недомагання, передчасне стомлювання, головний біль, порушення функцій органів зору, які здійснювали несприятливий психофізіологічний вплив на самопочуття та працездатність операторів. Однак, в той час основна увага приділялась розвитку техніки, а людина залишалась без необхідного захисту.

В умовах сучасного виробництва, яке характеризується масовим характером та широким застосуванням комп'ютерної техніки попередні

пріоритети зазнали суттєвої трасформації. У центрі уваги вітчизняних та зарубіжних фахівців є питання щодо визначення характеру та умов праці користувачів комп'ютерів, функціональних змін у динаміці виконання трудових завдань, захворюваності та стану здоров'я, розробки засобів захисту.

Дослідження медиків-гігієністів, психологів, світлотехніків та фахівців з охорони праці та ергономіки показали, що сучасна професія користувача ВДТ належить до розумової праці, яка характеризується: високою напруженістю зорових функцій; одноманітною позою; великою кількістю стереотипних висококоординованих рухів, що виконуються лише м'язами кистей рук на фоні малої загальної рухової активності; значним нервово-емоційним компонентом, особливо в умовах дефіциту часу; роботою з великими масивами інформації, що викликає активізацію уваги та інших вищих психічних функцій. Крім того, при роботі з дисплеями на електронно-променевих трубках виникає вплив на користувача цілої низки факторів фізичної природи — електростатичні поля, радіочастотне та рентгенівське випромінювання тощо.

Встановлено, що стан організму користувача значно залежить від типу роботи з ВДТ та умов її виконання. В загальному усі користувачі комп'ютерів поділяються на професіоналів та непрофесіоналів. До останніх можна віднести осіб, які використовують комп'ютер епізодично і він є для них не основним, а тільки допоміжним засобом (науково-технічні працівники, бібліотекарі, студенти, школярі, торгівельні працівники та ін.).

Діяльність професіоналів можна поділити на три групи:

1. Діяльність, яка пов'язана з виконанням нескладних багаторазово повторюваних операцій, що не вимагають великого розумового напруження. Наприклад, робота операторів комп'ютерного набору, працівників довідкових служб.

2. Діяльність, яка пов'язана із здійсненням логічних операцій, що постійно повторюються. Це робота інженера-економіста, інженера-проектувальника, оператора автоматизованого виробництва.

3. Діяльність, коли в процесі роботи необхідно приймати рішення за відсутності заздалегідь відомого алгоритму. Наприклад, робота інженера-програміста, диспетчерів руху залізничного транспорту, аеропортів тощо.

Необхідно зазначити, що такий поділ досить умовний, оскільки дане питання ще не достатньо розроблене і потребує детального вивчення. Проте, зрозуміло, що для кожної категорії користувачів комп'ютерів характерні свої особливості впливу комплексу несприятливих факторів трудового процесу та умов праці.

Робота з комп'ютером характеризується значною розумовою напругою й нервово-емоційним навантаженням операторів, високою напруженістю зорової роботи й досить великим навантаженням на м'язи рук при роботі із клавіатурою ЕОМ. Велике значення має раціональна конструкція й розташування елементів робочого місця, що важливо для підтримки оптимальної робочої пози людини-оператора.

Отже, у процесі роботи з комп'ютером необхідно дотримувати правильний режим праці й відпочинку. У противному випадку в персоналу відзначаються значна напруга зорового апарата з появою скарг на незадоволеність роботою, головні болі, дратівливість, порушення сну, утома й хворобливі відчуття в очах, у попереку, в області шиї та руках. Вимоги до виробничих приміщень. Фарбування приміщень і меблів повинна сприяти створенню сприятливих умов для зорового сприйняття, гарного настрою. Джерела світла, такі як світильники й вікна, які дають відбиття від поверхні екрана, значно погіршують точність знаків і спричиняють перешкоди фізіологічного характеру, які можуть виразитися в значній напрузі, особливо при тривалій роботі. Відбиття, включаючи відбиття від вторинних джерел світла, повинне бути зведене до мінімуму. Для захисту від надлишкової яскравості вікон можуть бути застосовані штори й екрани[54].

# РОЗДІЛ 7 ЕКОЛОГІЯ

## 7.1 Методи узагальнення екологічної інформації.

Узагальнення екологічної інформації – це стадія роботи з зібраною екологічною інформацією, коли оброблений екологічний матеріал потребує узагальнення, наочного подання і відображення складних екологічних ситуацій. Методами наочного подання та викладання фізичних величин, що використовують для більш раціонального та систематизованого викладення цифрової інформації – є статистичні таблиці і статистичні графіки, тобто табличний і графічний метод.

Статистичні таблиці – це форма раціонального та систематизованого викладення цифрової інформації. Основною перевагою цифрової інформації, зведеної в таблиці, є компактність, наочність, виразність. Інформація стає легкодоступною і рельєфною, компактною і раціональною.

Мета побудови таблиць багатогранна:

- систематизація цифрової інформації;
- полегшення і прискорення ефекту сприйняття;
- інтенсифікація пізнавального процесу;
- економія місця при викладенні інформації.

Таблиці складають не лише на заключному етапі дослідження. В процесі обробки статистичних даних користуються допоміжними, робочими таблицями. їх слід відрізняти від допоміжних розрахункових таблиць (логарифмічних, таблиць коефіцієнтів). Статистичними таблицями вважають тільки ті, що містять наслідки статистичного аналізу еколого-економічних явищ і процесів.

Таблиця за своїм логічним змістом розглядається як «статистичне речення», що має свій підмет і присудок. Підмет таблиці характеризує об'єкт дослідження, а присудок – це система показників, що відображує підмет як об'єкт.

Статистична таблиця має ряд горизонтальних рядків і вертикальних граф. Перетин рядків і граф утворює клітини таблиці. Ліві бічні і верхні клітини призначені для словесних заголовків, а решта для числових.

Статистичний графік являє собою рисунок, який описує статистичні сукупності умовною мовою геометричних знаків тієї чи іншої форми: крапок, ліній, площин, фігур та різних їх комбінацій.

Статистичні графіки – це спосіб умовного зображення цифрової інформації у вигляді крапок, ліній, стовпчиків, кругів або фігур.

Мета побудови потрійна:

- популяризація цифрової інформації;
- забезпечення доступності сприйняття інформації;
- узагальнення цифрової інформації.

Призначення графіків багатогранне:

- порівняння між собою різних величин;
- характеристика складу, структури і структурних зрушень сукупностей;
- з'ясування ступеня розповсюдження явищ в просторі;
- вивчення взаємозв'язку між явищами і їх ознаками;
- виявлення хронологічних явищ і їх ознак;
- дослідження темпів, тенденцій, закономірностей і перспектив розвитку явищ.

Графічні зображення в статистиці можуть бути представлені і негеометричними знаками – силуетами чи малюнками. Наприклад, динаміку книжкової продукції на графіку можна зобразити у вигляді книжкових полиць, інфляційні процеси – у вигляді банкнотів тощо.

У більшості випадків статистичних графіків використовують не об'ємне зображення, складне за побудовою, а площинне. Площинне зображення досить різноманітне за формою і водночас має ті ж самі складові елементи.

Поле графіка — це простір, у якому розміщуються геометричні або інші графічні знаки, що утворюють графік. Розмір поля графіка залежить від його

призначення і характеризується розміром та пропорціями сторін. З погляду естетичних вимог і зорового сприйняття зображених даних рекомендується співвідношення сторін: від 1:1,3 до 1:1,5. Найзручнішим для візуального сприйняття вважається формат, сторони якого знаходяться у співвідношенні 1:2. Таке співвідношення одержують, коли довша сторона прямокутника дорівнює діагоналі квадрата, побудованій на короткій стороні прямокутника. Ідеальні графіки прямокутної форми зі співвідношенням сторін 3:5, 5:8, 8:13 і т. д. Такі співвідношення сторін відомі під назвою «правило золотого перетину», згідно з яким висота прямокутника відноситься до його основи як основа до висоти плюс основа. Якщо статистичні графіки представлені у формі рівнобічного трикутника, то його основа повинна відноситися до висоти, як 1:3. Розмір графіка повинен відповідати його призначенню.

Просторові орієнтири в статистичних графіках використовують для визначення порядку розміщення геометричних знаків у полі графіка. Вони задаються системою координатних сіток контурних ліній, які ділять це поле на частини. Як правило, в статистиці використовується система прямокутників координат, але іноді може застосовуватися і полярна система (колові графіки).

Масштабні орієнтири визначаються системою масштабних шкал або спеціальними знаками для визначення розмірів графічних знаків.

Експлікація графіка являє собою словесне пояснення основних елементів графіка та його змісту. Вона включає назву графіка, надписи вздовж масштабних шкал, окремі пояснювальні надписи, що розкривають зміст елементів графічного образу. Статистичний графік – це знакова модель, без експлікації його не можна зрозуміти, тобто перенести знання із формалізованої системи характеристики дійсності на саму дійсність.

Види графіків поділяють: за призначенням (аналітичні дають можливість порівняння графічних образів, ілюстративні допомагають порівнянням геометричних фігур, показують зміну розмірів явищ, інформаційні вміщують інформацію лише про об'єкт вивчення), за формою графічного образу (крапкові, лінійні, площинні, просторові, зображувальні), за способом побудови (діаграми креслення із геометричних фігур і знаків, які замінюють цифри, картограми

показник відображений штриховкою на карті або плані, картодіаграми діаграми, які накладені на карті або плані території)[55].

7.2 Вимоги до мікроклімату, вмісту аероіонів і шкідливих хімічних речовин у повітрі приміщень експлуатації моніторів і ПЕОМ.

У виробничіх приміщеннях, у яких робота на ПЕОМ є допоміжною, температура, відносна вологість та швидкість руху повітря на робочих місцях повинна відповідати допустимим значенням, а для приміщень, в яких така робота є основною – оптимальним значенням по діючим санітарним нормам (таблиця 7.1)

Таблиця 7.1 – Оптимальні значення

| Період року | Температура повітря, ˚С | | Відносна вологість, % | | Швидкість руху повітря, м/с | |
|---|---|---|---|---|---|---|
| | Допуст. | Оптим. | Допуст. | Оптим. | Допуст. | Оптим. |
| Холодний | 21 – 25 | 22 – 24 | 75 | 40 – 60 | 0,1 | 0,1 |
| Теплий | 22 – 28 | 23 – 25 | 55 | 40 – 60 | 0,1 – 0,2 | 0,1 |

Рівні позитивних та негативних аероіонів у повітрі приміщень з ПЕОМ повинні відповідати нормам, що наведені у таблиці 7.2.

Таблиця 7.2 - Вимоги до вмісту аероіонів

| Рівні іонізації повітря | Кількість іонів в 1 см. куб. | |
|---|---|---|
| | п+ | п- |
| Мінімально необхідні | 400 | 600 |
| Оптимальні | 1500- | 3000- |
| Максимально | 50000 | 50000 |

Вміст шкідливих хімічних речовин у повітрі виробничих приміщень, в яких робота на ПЕОМ є допоміжною, не повинно перевищувати гранично допустимих концентрацій у робочій зоні ГДКр.з., а для приміщень, в яких робота з ПЕОМ є основною – не повинно перевищувати і гранично допустимих концентрацій забруднюючих речовин в атмосферному повітрі населених місць (середньодобові значення) ГДКс д. Витяги із нормативних документів наведені в таблиці 7.3 [56].

Таблиця 7.3 - Вимоги до вмісту шкідливих речовин

| Назва речовини | Клас небезпечності | $ГДК_{р.з.}$,мг/м$^3$ | $ГДК_{с.з.}$,мг/м$^3$ |
|---|---|---|---|
| Пил неорганічний | 3 | 0 | 0 |
| Оксид вуглецю | 4 | 2 | 3 |
| Сірнистий ангідрит | 3 | 1 | 0 |
| Диоксид азоту | 2 | 2 | 0 |
| Пари бензину | 4 | 1 | 1 |
| Ацетон | 4 | 2 | 0 |
| Аерозоль свинцю | 1 | 0 | 0 |
| Флюс каніфольний | 4 | 1 | 0 |
| Озон | 1 | 0 | 0 |

# ВИСНОВКИ

У цій роботі проведено систематичну оцінку продуктивності з точки зору ефективності та корисності даних трьох найбільш відомих алгоритмів *k*-анонімізації. Використовуючи загальнодоступні реалізації алгоритмів у загальній структурі (для справедливого порівняння), визначено сценарії, в яких алгоритми працювали добре чи погано, з точки зору цікавої метрики. Результати продемонстрували, що не існує найкращого алгоритму анонімізації для всіх сценаріїв, але на роботу найкращого алгоритму в даній ситуації впливає безліч факторів. На основі аналізу сформувалось уявлення про фактори, які слід враховувати під час вибору алгоритму анонімізації, та обговорено особливості (сильні та слабкі сторони) набору загальних метрик корисних даних. Крім того, ці результати підтверджують складність процесу відбору алгоритмів анонімізації, що відображає необхідність створення методологій, які допомагають користувачам у процесі визначення того, який алгоритм анонімізації найкраще підходить для конкретного сценарію публікації.

В перших чотирьох розділах дипломної роботи магістра було вирішено такі завдання:

1. Описано що таке анонімізація і для чого вона потрібна.

2. Розглянуто атаки конфідеційності ти їх приклади в реальному житті, моделі конфідеційності.

3. Розглянуто методи анонімізації та три найбільш відомі алгоритми *k*-анонімізації.

4. Проведено ряд експерементів для порівняння алгоритмів за різних умов: різна кількість QID та різні значення *k* в *k*–анонімізації та за такими критеріями: час анонімізації, споживання пам’яті, узагальнена втрата інформації, метрика чутливості та середній розмір класу еквівалентності.

5. Розглянуто набори даних (реальний та синтетичний) які використовувались в експерементах. Описано генератор синтетичних даних та принцип його роботи.

# БІБЛІОГРАФІЯ


1. Gantz J. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. / J. Gantz, D. Reinsel.., 2012. – (Technical report, IDC, sponsored by EMC)
2. OpenData websites URL: http://www.data.gov/ .
3. Information Commissioner's Office. Data Sharing Code of Practice. Technical report, ICO. – 2011.
4. Регламент захисту даних GDPR, і як просто його дотримуватися URL: https://evergreens.com.ua/ua/articles/general-data-protection-regulation.html.
5. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. / Sweeney. // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. – 2002. – №10(5). – С. 571–588
6. Golle P. Revisiting the uniqueness of simple demographics in the us population / Golle. // 5th ACM Workshop on Privacy in Electronic Society (WPES). – 2006. – С. 77–80.
7. Hafner K. And if You Liked the Movie, a Netflix Contest May Reward You Handsomely / Hafner. // New York Times, October 6. – 2006.
8. Narayanan A. Robust de-anonymization of large sparse datasets / A. Narayanan, V. Shmatikov. // IEEE Symposium on Security and Privacy (SP). – 2008. – С. 111–125
9. Barbaro M. A face is exposed for AOL searcher no. 4417749 / M. Barbaro, T. Zeller. // New York Times, August 9. – 2006.
10. Anonymizing transaction databases for publication. / Y.Xu, K. Wang, C. Fu, P. Yu. // 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). – 2008. – С. 767–775.
11. Minimality attack in privacy preserving data publishing / R.Wong, C. Fu, K. Wang, J. Pei. // 33rd International Conference on Very Large Data Bases (VLDB). – 2007. – С. 543–554.

12. LeFevre K. Incognito: Efficient full-domain k-anonymity / K. LeFevre, D. DeWitt, R. Ramakrishnan. // ACM SIGMOD International Conference on Management of Data (SIGMOD). – 2005. – С. 49–60.

13. Sweeney. L. k-Anonymity: A Model for Protecting Privacy / Sweeney.. // Int. J. Uncertain. Fuzziness Knowl.- Based Syst. – 2002. – №10(5). – С. 557–570.

14. l-Diversity: Privacy Beyond k-Anonymity / A.Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam. // ACM Trans. Knowl. Discov. Data,. – 2007. – №1

15. Dwork C. Differential Privacy / Dwork. // Automata, Languages and Programming, 4052:1–12. – 2006.

16. Privacy-Preserving Data Publishing / B.Chen, D. Kifer, K. LeFevre, A. Machanavajjhala. // Foundations and Trends in Databases, 2(1–2) – 2009. С. 1–167.

17. Dwork C. A firm foundation for private data analysis / Dwork. // . Communications of the ACM, 54(1). – 2011. – С. 86–95.

18. Wang K. Handicapping attacker’s confidence: An alternative to k-anonymization. / K. Wang, C. Fung, S. Yu. // Knowledge and Information Systems (KAIS), 11(3). – 2007. – С. 345–368.

19. Privacy-preserving trajectory data publishing by local suppression / R.Chen, B. Fung, N. Mohammed, B. Desai. // Information Sciences: Special Issue on Data Mining for Information Security, 231. – 2013. – С. 83–97.

20. Sweeney. L. k-Anonymity: A Model for Protecting Privacy / Sweeney.. // Int. J. Uncertain. Fuzziness Knowl.- Based Syst. – 2002. – №10(5). – С. 571–588.

21. LeFevre K. Mondrian Multidimensional K-Anonymity / K. LeFevre, D. DeWitt, R. Ramakrishnan. // 22nd International Conference on Data Engineering, ICDE ’06. – 2006. – С. 25.

22. Wang K. Bottom-Up Generalization: A Data Mining Solution to Privacy Protection. / K. Wang, P. Yu, S. Chakraborty. // 4th IEEE International Conference on Data Mining, ICDM ’04,. – 2004. – С. 249–256.

23. Efficient Multidimensional Suppression for K-Anonymity. / S.Kisilevich, L. Rokach, Y. Elovici, B. Shapira. // IEEE Trans. Knowl. Data Eng., 22(3). – 2010. – С. 334–347.

24. Fung B. Top-Down Specialization for Information and Privacy Preservation / B. Fung, K. Wang, P. Yu. // 21st International Conference on Data Engineering ICDE ’05. – 2005. – С. 205–216.

25. Bayardo R. Data Privacy Through Optimal k-Anonymization / R. Bayardo, R. Agrawal. // 21st International Conference on Data Engineering, ICDE ’05. – С. 217–228.

26. Argus User’s Manual version 3.2 / [A. Hundepool, A. de Wetering, R. Ramaswamy та ін.]. – 2003.

27. Domingo-Ferrer J. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. / J. Domingo-Ferrer, J. Mateo-Sanz. // . IEEE Trans. on Knowl. and Data Eng., 14(1). – 2002. – С. 189–201.

28. Laszlo M. Minimum Spanning Tree Partitioning Algorithm for Microaggregation. / M. Laszlo, S. Mukherjee. // . IEEE Trans. on Knowl. and Data Eng., 17(7). – 2005. – С. 902–911.

29. Domingo-Ferrer J. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation / J. Domingo-Ferrer, V. Torra. // Data Min. Knowl. Discov., 11(2). – 2005. – С. 195–212.

30. Solanas A. V-MDAV: A Multivariate MicroaggregationWith Variable Group Size. / A. Solanas, A. Mart´ınez-Ballest´e. // 17th COMPSTAT Symposium of the IASC. – 2006. – С. 917–925.

31. Samarati P. Protecting Respondents’ Identities in Microdata Release / Samarati. // IEEE Trans. on Knowl. and Data Eng., 13(6). – 2001. – С. 1010–1027.

32. Bache K. UCI Machine Learning Repository / K. Bache, M. Lichman., 2013.

33. Li N. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. / N. Li, T. Li. // 23rd International Conference on Data Engineering, ICDE ’07. – 2007. – С. 106–115.

34. Utility-Based Anonymization for Privacy Preservation with Less Information Loss. / [J. Xu, W. Wang, J. Pei та ін.]. // SIGKDD Explor. Newsl., 8(2). – 2006. – С. 21–30.

35. Bergmann. V. Data Benerator Tool URL: http://databene.org/databene-benerator/

36. Central Statistics Office URL: http://www.cso.ie/en/databases/.

37. Kifer D. Injecting Utility into Anonymized Datasets / D. Kifer, J. Gehrke. // 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD ’06. – 2006. – С. 217–228.

38. Agrawal D. On the Design and Quantification of Privacy Preserving Data Mining Algorithms / D. Agrawal, C. Aggarwa. // 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS ’01. – 2001. – С. 247–255.

39. Density-based Microaggregation for Statistical Disclosure Control / J.Lin, T. Wen, J. Hsieh, P. Chang. // Expert Syst. Appl., 37(4). – 2010. – С. 3256–3263.

40. Aggregate Query Answering on Anonymized Tables / Q.Zhang, N. Koudas, D. Srivastava, T. Yu. // 23rd International Conference on Data Engineering, ICDE ’07. – 2007. – С. 116-125.

41. Privacy Preserving Mining of Association Rules. / A.Evfimievski, R. Srikant, A. Agrawal, J. Gehrke. // 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’02. – 2002. – С. 217–228..

42. Domingo-Ferrer J. . Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure / J. Domingo-Ferrer, J. Mateo-Sanz, V. Torra. // of ETK-NTTS 2001, Luxemburg: Eurostat. – 2001. – С. 807–826.

43. Brickell J. The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing / J. Brickell, V. Shmatikov. // 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’08. – 2008. – С. 70–78

44. Nergiz M. Thoughts on k-Anonymization / M. Nergiz, C. Clifton. // Data and Knowledge Engineering, 63(3). – 2007. – С. 622–645.

45. UTD Anonymization ToolBox URL: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/

46. Clifton C. On Syntactic Anonymity and Differential Privacy / C. Clifton, T. Tassa. // Transactions on Data Privacy, 6(2). – 2013. – С. 161–183.

47. k-Anonymity. Secure Data Management In Decentralized Systems / V.Ciriani, S. De Capitani di Vimercati, S. Foresti, P. Samarati., 2007. – 323 с.

48. COCOA: A Synthetic Data Generator for Testing Anonymization Techniques / V.Ayala-Rivera, A. Portillo Dominguez, C. Thorpe, L. Murphy. – 2016

49. Walck C. Handbook on statistical distributions for experimentalists / Walck., 2007.

50. ДСанПіН 3.3.2-007-98 Державні санітарні правила і норми. Гігієнічні вимоги до організації роботи з візуальними дисплейними терміналами електронно-обчислювальних машин. Київ, 1998.

51. ГОСТ 12.1.005-88 Загальні санітарно-гігієнічні вимоги до повітря робочої зони. Москва, 1988.

52. ДБН В.2.5-28:2018 Природне і штучне освітлення. Київ, 2018.

53. НПАОП 40.1-1.21-98 (ДНАОП 0.00-1.21-98) Правила безпечної експлуатації електроустановок споживачів. Київ, 1998.

54. Безпека життєдіяльності. Захист населення і територій при надзвичайних ситуаціях: Навчальний посібник / В. В.Денисов, І. А. Денісова, В. В. Гутен, О. І. Монтвіла. – Москва: ІКЦ «МарТ», 2003. – 608 с.

55. Тарасова В. В. Екологічна статистика: Навчальний посібник / В. В. Тарасова. – Київ: Центр учбової літератури, 2008. – 392 с.

56. Козлов С. С. Конспект лекцій з дисципліни «Охорона праці» / С. С. Козлов. – Київ, 2013. – 94 с.

# ДОДАТКИ

Додаток А

УДК 004.056.53

**Т. Сачик, Н. Загородна**
(Тернопільський національний технічний університет імені Івана Пулюя)

# ЗАХИСТ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ В ЗАДАЧАХ АНАЛІЗУ ТА ОБРОБКИ ВЕЛИКИХ ДАНИХ

**T. Sachyk, N. Zagorodna**
(Ternopil Ivan Puluj National Technical University, Ukraine)

# PROTECTION OF PERSONAL INFORMATION IN THE OBJECTIVES OFANALYSIS AND PROCESSING OF BIG DATA

Інформаційні технології не лише полегшують наше повсякденне життя, але й збирають та зберігають величезні обсяги приватної інформації користувачів. І, якщо частина цих даних має цілком загальний характер, то інші персональні дані, включаючи прізвища людей, дати народження, номера страхових полісів і рахунків, дозволяють ідентифікувати особу. Часто компанії можуть публікувати результати своїх досліджень або ж передавати зібрані дані стороннім особам (науково-дослідним центрам) для аналізу. Необхідність публічного поширення або ж передача третій стороні приватних даних поставила нові виклики щодо захисту. Деякі організації, включаючи Інститутську наглядову раду (IRB, США) і Європейське агентство по оцінці лікарських засобів (EMA), вимагають, щоб дослідники і компанії-виробники ліків анонімізували свої дані, перш ніж публікувати результати своїх досліджень, з метою захисту персональних даних їх учасників і їх права на недоторканність приватного життя. Як результат, публікація даних, що зберігають конфіденційність, стала активною дослідницькою сферою. Отже, існує необхідність пошуку інструментів анонімізації персональних даних з метою мінімізації негативних наслідків можливого порушення їх конфіденційності.

Анонімізація, або редагування даних – процес видалення або приховування персональних даних з метою їх подальшого використання. На перший погляд, вирішення цієї проблеми видається доволі тривіальним: адже достатнього просто видалити стовпці, що містять прямі ідентифікатори, такі як імена та номери соціального страхування тощо. Тим не менше, було доведено, що такого підходу недостатньо для збереження конфіденційності. Ця проблема виникає тому, що все ще можливо поєднувати різні набори даних або мати базові знання про людей, щоб зробити висновки про особу. Повторна ідентифікація особи досягається за допомогою зв’язування атрибутів, відомих як квазі-ідентифікатори (QID), таких як стать, дата народження або поштовий індекс. Науковці з США довели, що поєднуючи відкриту інформацію з різних джерел можна однозначно ідентифікувати 70-90% людей.

Існує кілька моделей, які пропонують формальні гарантії щодо захисту конфіденційності особи при публікації даних. Зосередимось на *k*-анонімізації, оскільки на відміну від інших моделей (ℓ-різноманіття, *t*-близькість та диференційна конфіденційність), які мають обмеження в використанні, ця модель є простою для розуміння і базовою у багатьох сферах використання. Більше того автори [1] вказують на актуальність моделі k-анонімності як основи для побудови більш надійних моделей.

У моделі *k*-анонімізації кожна людина представлена у вигляді набору атрибутів, включаючи QID –атрибути, які можуть бути пов'язані із зовнішньою інформацією з метою однозначної ідентифікації особи. Захист конфіденційності в методі *k*-анонімізації полягає в тому, щоб гарантувати, що кожен набір QID відображається принаймні в *k* записах у наборі даних, або, що дані будь-якої конкретної особи не відрізняються від даних принаймні *k* - 1 інших осіб щодо QID. Мета методу полягає в тому, щоб зробити квазі-ідентифікатори неточними та менш інформативними. *k*-анонімізація зазвичай досягається шляхом

узагальнення та приховування даних (наприклад, опусканням імен осіб і заміною п'ятизначних поштових індексів лише їх першими двома цифрами) з метою створити класи еквівалентності, що мають однакові QID. Тому метою нашого дослідження є порівняння найбільш відомих методів k-анонімізації (Datafly, Incognito, Mondrian) з огляду на використання ресурсів та корисність залишкових даних.

1. B. Kenig and T. Tassa. A Practical Approximation Algorithm for Optimal k-Anonymity. Data Min. Knowl. Discov., 25(1):134–168, 2012.