Міністерство освіти і науки України
«Київський національний університет імені Тараса Шевченка»

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ДОПУСТИТИ ДО ЗАХИСТУ:
завідувач кафедри кібербезпеки
та захисту інформації
_____ Н.В. Лукова-Чуйко
«18» червня 2021р.

ПОЯСНЮВАЛЬНА ЗАПИСКА

дипломної роботи

бакалавра
(назва освітнього рівня)

галузь знань 12 Інформаційні технології
(шифр і назва галузі знань)

спеціальність 125 Кібербезпека
(код і назва спеціальності)

освітня програма Кібербезпека
(назва освітньої програми)

на тему: Розробка елементів системи захисту персональних даних

Виконавець: студент IV курсу, групи КБ-42

Гальміз Максим Вікторович
(підпис) (прізвище ім'я по-батькові)

	Прізвище, ініціали	**Підпис**
Керівник	Мирутенко Л.В.	

Нормоконтроль	Зюбіна Р. В.	

Київ 2021

Міністерство освіти і науки України
«Київський національний університет імені Тараса Шевченка»

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ЗАТВЕРДЖЕНО:
завідувач кафедри кібербезпеки
та захисту інформації
_____ Н.В. Лукова-Чуйко
«11» листопада 2020 р.

ЗАВДАННЯ

на виконання дипломної роботи

спеціальності 125 Кібербезпека
(код і назва спеціальності)

освітньої програми Кібербезпека
(назва освітньої програми)

Студенту **КБ-42** **Гальмізу Максиму Вікторовичу**
(група) (прізвище ім'я по-батькові)

Тема дипломної роботи Розробка елементів системи захисту персональних даних

1. ПІДСТАВИ ДЛЯ ПРОВЕДЕННЯ РОБОТИ

Тема дипломної роботи затверджена на засіданні кафедри кібербезпеки та захисту інформації протокол №2 від 08.10.2020 р.

2. ВИХІДНІ ДАНІ ДЛЯ ПРОВЕДЕННЯ РОБІТ

Нормативно правова база захисту персональних даних, методи захисту персональних даних, механізми деідентифікації даних

3. ЗМІСТ РОЗРАХУНКОВО-ПОЯСНЮВАЛЬНОЇ ЗАПИСКИ

Необхідно ознайомитися з методами захисту персональних даних, обрати механізм деідентифікування даних, проаналізувати підходи K-анонімізації та

розробити рекомендації щодо роботи з ними.

4. ВИМОГИ ДО РЕЗУЛЬТАТІВ ВИКОНАННЯ РОБОТИ

Практична цінність Реалізація підходів анонімізації захисту персональних даних та розробка рекомендацій щодо їх використання.

5. ДАТА ВИДАЧІ ЗАВДАННЯ

Дата видачі завдання:11 листопада 2020 року

Завдання видала _____ (підпис) Л.В. Мирутенко (ініціали, прізвище)

Завдання прийняв до виконання _____ (підпис) М.В. Гальміз (ініціали, прізвище)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів робіт	Строки виконання робіт (початок-кінець)	Відмітка про виконання
1	Уточнення постановки задачі	25.01.2021 – 30.01.2021	*виконано*
2	Аналіз літератури	31.01.2021 – 12.02.2021	*виконано*
3	Обґрунтування вибору рішення	13.02.2021 – 18.02.2021	*виконано*
4	Аналіз проблем захисту персональних даних	19.02.2021 – 07.03.2021	*виконано*
5	Дослідження вразливостей та загроз персональним даним	08.03.2021 – 25.03.2021	*виконано*
6	Дослідження механізмів захисту та деідентифікації даних	26.03.2021 – 10.04.2021	*виконано*
7	Проведення дослідження анонімізації та розробка рекомендацій	11.04.2021 – 17.05.2021	*виконано*
8	Оформлення пояснювальної записки	18.05.2021 – 08.06.2021	*виконано*
9	Підготовка до захисту дипломної роботи	09.06.2020 – 21.06.2021	*виконано*

Завдання видала _____ (підпис) Л.В. Мирутенко (ініціали, прізвище)

Завдання прийняв до виконання _____ (підпис) М.В. Гальміз (ініціали, прізвище)

Термін подання дипломної роботи до ЕК 08 червня 2021 року

РЕФЕРАТ

Дипломна робота складається зі вступу, трьох розділів, загальних висновків, списку використаних джерел, 1 додатка, має 61 сторінку основного тексту, 4 рисунка, 2 таблиці. Список використаних джерел містить 59 найменування і займає 5 сторінок.

Об'єктом дослідження є процес захисту персональних даних.

Предметом дослідження є методи та засоби захисту персональних даних.

Метою роботи є реалізація елементів захисту персональних даних з використанням підходів анонімізації.

Методи дослідження: спостереження, порівняння, розрахунок, аналіз і синтез.

Практична значимість полягає у реалізації підходів анонімізації для захисту персональних даних та розробка рекомендацій щодо їх використання.

Для захисту персональних даних подальші дослідження можуть задіяти інші підходи анонімізації, наприклад, локальне перекодування.

Ключові слова: персональні дані, захист персональних даних, GDPR, інформаційна безпека, вразливості, анонімізація, k-анонімізація.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

ACA	–	Adaptive Conjoint Analysis
AES	–	Advanced Encryption Standard
API	–	Application Programming Interface
CCA	–	Canonical-Correlation Analysis
CCPA	–	California Consumer Privacy Act
EDBP	–	European Data Protection Board
HIPAA	–	Health Insurance Portability and Accountability Act
HMAC	–	Hash-based Message Authentication Code
LGPD	–	Lei Geral de Proteção de Dados Pessoais
MITM	–	Man In The Middle
NSA	–	National Security Agency
PCLOB	–	Privacy and Civil Liberties Oversight Board
PIMS	–	Privacy Information Management Systemc
PPD	–	Presidential Policy Directive
RDP	–	Remote Desktop Protocol
SEC	–	Securities and Exchange Commission
TOMS	–	Test Operation Management Systems
ЕОМ	–	Електронно-Обчислювальна Машина
СЄС	–	Суд Європейського Союзу
СМІБ	–	Система Менеджменту Інформаційної Безпеки
СУІБ	–	Система Управління Інформаційною Безпекою

ЗМІСТ

РЕФЕРАТ 5
ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ 6
ЗМІСТ 7
ВСТУП 8
РОЗДІЛ 1 ОСОБЛИВОСТІ ВИКОРИСТАННЯ ПЕРСОНАЛЬНИХ ДАНИХ 9
1.1 Аналіз нормативно-правової бази захисту персональних даних 9
1.2 Специфіка використання персональних даних 14
Висновки за розділом 1 22
РОЗДІЛ 2 ЗАГРОЗИ ПЕРСОНАЛЬНИМ ДАНИМ ТА МЕТОДИ ЗАХИСТУ ВІД НИХ 24
2.1 Основні типи загроз персональним даним 24
2.2 Аналіз методів захисту персональних даних 27
2.3 Основні механізми деідентифікації даних 29
2.3.1 Маскування даних 31
2.3.2 Токенізація 33
2.3.3 Анонімізація 36
Висновки за розділом 2 43
РОЗДІЛ 3 ЗАСТОСУВАННЯ ЕЛЕМЕНТІВ АНОНІМІЗАЦІЇ ДЛЯ ЗАХИСТУ ПЕРСОНАЛЬНИХ ДАНИХ 44
3.1 Сценарії повторної ідентифікації для k-анонімізованого набору даних 45
3.2 Практична реалізація 51
Висновки за розділом 3 55
ВИСНОВКИ 57
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 58
ДОДАТКИ 64
ДОДАТОК А 64

ВСТУП

Актуальність дослідження полягає в існуванні етичних, правових та технічних питань в умовах пандемії щодо обсягу та типів персональних даних, які збираються, обробляються, обмінюються та використовуються закладами громадського здоров'я. В такій ситуації обробка персональних даних є необхідною для того, щоб здійснити заходи для стримування поширення вірусу або зменшення його наслідків. Ці виклики демонструють необхідність нових моделей прозорого управління даними та технологіями у зусиллях щодо боротьби з вірусом, а також у майбутніх надзвичайних ситуаціях у галузі охорони здоров'я.

Для досягнення заданої мети в роботі були поставлені наступні завдання:

- розглянути нормативно-правові акти у сфери захисту персональних даних;
- проаналізувати можливі загрози персональним даним;
- дослідити існуючі механізми захисту персональних даних;
- дослідити підходи анонімізації для захисту персональних даних;
- оцінити ризики повторної ідентифікації даних.

РОЗДІЛ 1
ОСОБЛИВОСТІ ВИКОРИСТАННЯ ПЕРСОНАЛЬНИХ ДАНИХ

1.1 Аналіз нормативно-правової бази захисту персональних даних

В 2020 році панувала пандемія covid-19. Роботодавці та уряди враховували конфіденційність, пристосовуючи практику на робочому місці, враховуючи, хто має лихоманку та інші симптоми, хто куди подорожував, хто з ким контактував та які члени громади виявили позитивні результати тесту на зараження.

В результаті потреби у відстеженні та трансуванні осіб, уряди та громадяни визнали неминучі компроміси між приватністю й винятковою увагою до громадського здоров'я та безпеки.

Навіть Європейська рада з питань захисту даних (EDBP) визнала, що заходи щодо захисту даних, такі як GDPR, не перешкоджають заходам, вжитим у боротьбіз пандемією коронавірусу. Боротьба з інфекційними хворобами є цінною метою, яку розділяють усі нації, що є в інтересах людства стримувати розповсюдження хвороб та використовувати сучасні методи. Відповідно, EDPB погодився, що «надзвичайна ситуація є правовою умовою, яка може узаконити обмеження свобод за умови, що ці обмеження пропорційні та обмежуються надзвичайним періодом» [1].

І хоча конфіденційність не вважається абсолютним правом у будь-якій юрисдикції, важливо визнати, що жодна демократична країна не зайняла такої позиції або не діяла таким чином, припускаючи, що приватне життя особи не має значення або не вимагається – навіть під час надзвичайної ситуації з пандемією. Навпаки, права на конфіденційність враховуються при боротьбі з вірусом майже скрізь.

Але Covid-19 був не єдиним потрясінням для системи конфіденційності в 2020 році. 16 липня Суд Європейського Союзу (CJEU) визнав недійсними рамки «Щита конфіденційності», узгоджені між Комісією ЄС та Міністерством торгівлі США для сприяння потоків даних в США [2]. Хоча СЄС підтвердив обґрунтованість

використання стандартних договірних положень для передачі персональних даних у країни, які ЄС ще не визнав «адекватними» (наприклад, США), СЄС наклав значні нові зобов'язання на організації, які хочуть передавати дані, а також органи захисту даних, які слід враховувати, схвалюючи такі передачі.

Зокрема, Суд погодився з австрійським захисником конфіденційності Максом Шремсом, що існує теоретична можливість того, що користувачі мереж соціальних мереж можуть таємно передавати свої повідомлення до Агентства національної безпеки США (NSA) без вигоди захисту приватності, доступної в Європі [3].

Не зважайте на те, що немає жодних доказів чи підстав вважати, що NSA зацікавлена у збиранні повідомлень середнім європейським користувачам соціальних мереж – іншими словами, від когось іншого, крім терористичного, шпигунського чи ворожого актора. І не дивлячись на те, що захист конфіденційності та можливості юридичного відшкодування відповідно до законів США про нагляд значно сильніші, ніж ті, які держави-члени ЄС надають своїм громадянам. Насправді Директива про президентську політику 28 вимагає від NSA та інших американських агентств, які беруть участь у розвідці сигналів, захищати конфіденційність осіб за межами США способом, достатньо порівнянним із захистом, який отримують громадяни США [4]. І, зокрема, незалежна Комісія з нагляду за конфіденційністю та громадянськими свободами (PCLOB) надала ретельну оцінку впровадженню PPD-28 16 жовтня 2018 року [5].

Примітно, що СЄС визнав, що американські наглядові гарантії та засоби захисту є менш ніж «по суті еквівалентними» тим, що застосовуються в ЄС, не порівнюючи закони та практики нагляду за державами-членами ЄС із законами США. Хоча оцінка Судом гарантій нагляду в США була поверхневою і вкрай неповною, його поводження з наглядом ЄС повністю відсутнє. СЄС не просто запитав, чи має яка-небудь держава-член ЄС наглядовий орган для вивчення та оцінки наслідків конфіденційності та громадянських прав електронного спостереження, як це робить PCLOB та Суд з нагляду за зовнішньою розвідкою – з повним дозволом національної безпеки для доступу до найглибших таємниць сигнали інтелекту.

Хоча рішення СЄС може здатися легковажним (не кажучи вже про небезпеку) для багатьох спостерігачів, його потенційний вплив на трильйони доларів трансатлантичної торгівлі – це щось інше. Комісія ЄС та Міністерство торгівлі США публічно віддані вирішенню цієї судової загадки. Тим часом компанії проходять метафізичний процес, намагаючись продемонструвати (собі, органам захисту даних і, зрештою, можливо, Максу Шремсу), що за рішенням Шремса вони можуть передавати персональні дані Сполученим Штатам – та іншим не «адекватним» країнам – без таких даних, які не пропорційно доступні для нагляду за національною безпекою з боку країни-імпортера [3]. Іншими словами, справжніх стандартів взагалі не існує.

На відміну від цього, Сполучені Штати надали модель міжнародного спілкування та поваги верховенства закону в Законі про CLOUD 2018 року [6]. Закон уповноважує провайдерів послуг зв'язку США виробляти вміст електронних комунікацій, які вони зберігають за межами Сполучених Штатів, у відповідь на юридичні запити США, а також робити те саме для іноземних урядів, які уклали виконавчі угоди із США (які лише Великобританія зробив до цього часу, починаючи з липня 2020 року).

Що важливо, Закон CLOUD вимагає від урядів іноземних держав, які бажають укласти такі угоди, продемонструвати свою повагу до верховенства закону та міжнародних прав людини, приватності, свободи слова, мінімізації даних (еквівалентно тому, що США застосовують під наглядом зовнішньої розвідки. Закону), принципи недискримінації, підзвітності, прозорості, незалежного нагляду та багатьох інших детальних гарантій. Більше того, коли запитувані дані стосуються особи, яка не входить до США, яка знаходиться за межами США, від повідний постачальник послуг зв'язку має право подати клопотання до федерального суду про скасування урядового запиту. Постачальник може зробити це, якщо він вважає, що закони США та закони іноземного уряду конфліктують – включаючи закони про конфіденційність – стосовно попиту уряду на комунікації своїх клієнтів. Після подання клопотання суд повинен провести детальний аналіз «комітету» для вирішення колізії міжнародних законів щодо прав на конфіденційність даних [6].

Правові стандарти для такого аналізу комітету викладені із суттєвою конкретністю в Законі CLOUD. Суди повинні враховувати характер розглянутого юридичного конфлікту, суттєвість передбачуваного порушення іноземного законодавства, відповідні інтереси двох країн у даному питанні, контакти постачальника послуг та особи, про яку йдеться, зі Сполученими Штатами Держав, а також важливість інформації про особу для кримінального питання чи інтересу національної безпеки, про яку йде мова.

Вдумливий характер цього комітетного аналізу, який вимагає Закон CLOUD, не відповідає ні рішенню ЄС від Шремса, ні Загальному регламенту ЄС про захист даних.

Європі та решті світу було б добре забезпечити вивчення американської моделі гарантій, контролю та рівноваги, незалежного нагляду та міжнародної спільноти щодо доступу уряду до електронних комунікацій.

У будь-якому випадку, розвиток конфіденційності у Сполучених Штатах стосувався не лише доступу уряду до інформації. За останні півтора року американські регулятори та судові органи отримали найбільші штрафи та судові винагороди, коли-небудь зібрані за нібито порушення вимог щодо конфіденційності та безпеки даних [7].

Федеральна торгова комісія отримала угоду на суму 5 мільярдів доларів США та запровадила безпрецедентні вимоги корпоративного управління після розслідування справи в Кембриджській аналітиці [8]. А приватні позивачі отримали врегулювання на суму понад півмільярда доларів у зв’язку з нібито порушенням законодавства про конфіденційність державної біометричної інформації. Федеральна торгова комісія та державні прокурори (а в деяких випадках і приватні позивачі) зібрали фінансові збитки на сотні мільйонів доларів за порушення даних, а також про передбачувані порушення Закону про захист конфіденційності дітей в Інтернеті. Важливо зазначити, що багато з цих проваджень ґрунтуються на теорії, згідно з якою «контролер даних» несе юридичну відповідальність за нібито інвазивні або зловживані дії третіх осіб, які працюють на платформі «контролера» або через неї [9]. Цю тенденцію до розширеної відповідальності варто спостерігати.

Навіть Комісія з цінних паперів та бірж США (SEC) все більше зосереджується на цифровій практиці та ризиках [10]. Зараз SEC активно дотримується точності та надійності розкриття інформації про конфіденційність та кібербезпеку державними компаніями. Компанії можуть зіткнутися з регуляторними діями, якщо вони істотно занижують свої цифрові ризики, або уникають обговорення значущих інцидентів, які вже переживали, або якщо вони публічно завищують свої практики безпеки даних або конфіденційності. Як результат, багато компаній – особливо технологічні компанії – значно розширюють дискусію про те, як американські та міжнародні закони про конфіденційність, такі як GDPR або нещодавно введений в дію закон про конфіденційність Бразилії (LGPD), впливають або можуть вплинути на їхній загальний профіль регуляторного ризику або економічну життєздатність своїх поточних та майбутніх бізнес-моделей [11, 12].

Але, мабуть, найважливішим розвитком конфіденційності США є новий Закон про конфіденційність споживачів у Каліфорнії (CCPA) [13]. Він набув чинності в 2020 році, а з 1 липня може застосовуватися Генеральним прокурором штату.

CCPA вимагає не лише розголошення конфіденційності, надання прав на конфіденційність та введення обмежень щодо конфіденційності, порівнянних із GDPR. Але, типово по-американськи, CCPA боксує перспективою встановлених законом збитків (незалежно від фактичної шкоди), щоб стимулювати адвокатів подавати судові спори за порушення даних, що зачіпають особисту інформацію жителів Каліфорнії, – але лише в тому випадку, якщо порушення є наслідком невиконання компанією вимог.

Однак навіть CCPA може бути недостатньо для Каліфорнії. Його родоначальник, директор з нерухомості Аластер Мактаггарт, висунув нову ініціативу щодо конфіденційності, яка замінить і вийде за рамки CCPA – Каліфорнійський закон про права на конфіденційність (CPRA) [14]. CPRA була представлена перед виборцями штату в рамках виборів у листопаді 2020 року (повністю обходячи законодавчий орган штату).

Зрештою, настільки ж новаторським для Сполучених Штатів був CCPA, його наслідком може бути спонукання інших держав діяти, і, можливо, тоді Конгрес США нарешті прийме всеосяжний національний закон про конфіденційність.

Найбільш розумним результатом для Америки було б встановлення стабільних федеральних стандартів конфіденційності та безпеки, які визначають та націлюють фактичні пошкодження, спричинені зловживанням даними. Якщо Covid-19 навчив нас чомусь щодо конфіденційності, це означає, що з одного боку можуть бути реальні компроміси (за охорону здоров'я, безпеку, інновації, економіку та безпеку), а також реальну шкоду (для кишенькових книжок і особиста гідність та автономія), з іншого боку, від регулювання або занадто багато, або замало.

Правові відносини у сфері захисту персональних даних в Україні регулюються Конституцією України, чинними міжнародними договорами, згода на обов'язковість яких надана Верховною Радою України, Законом України «Про захист персональних даних», іншими законодавчими та прийнятими на їх виконання підзаконними нормативно-правовими актами [15].

Відповідно до закріпленого європейського та євроатлантичного курсу у Конституції України на постійній основі здійснюється моніторинг законодавства Європейського Союзу у сфері захисту персональних даних [16].

Європейський парламент прийняв Загальний регламент захисту даних (GDPR) у квітні 2016 року, замінивши застарілу директиву про захист даних від 1995 року [17, 18]. Він передбачає положення, яке вимагається від підприємств для захисту персональних даних та конфіденційності громадян ЄС за операціями, що відбуваються в країнах-членах ЄС. GDPR також регулює експорт персональних даних за межі ЄС.

1.2 Специфіка використання персональних даних

Початком GDPR – є термін «персональні дані». Якщо обробка даних стосується персональних даних, застосовується Загальне положення про захист

даних. Термін визначений у ст. 4 [19]. Персональні дані – це будь-яка інформація про фізичну особу, яку ідентифіковано чи можна ідентифікувати.

Суб'єкти даних можна ідентифікувати, якщо їх можливо прямо або опосередковано ідентифікувати, особливо за допомогою посилання на ідентифікатор, таке як ім'я, ідентифікаційний номер, дані про місцезнаходження, онлайн-ідентифікатор або одну з кількох спеціальних характеристик, що виражають фізичну, фізіологічну, генетичну, ментальну, комерційну, культурну чи соціальну ідентичність цих осіб. На практиці вони також включають усі дані, які є або можуть бути призначені людині будь-яким способом. Наприклад, телефон, кредитна картка або номер особи, дані рахунку, номерний знак, зовнішній вигляд, номер клієнта або адреса – це все персональні дані.

Оскільки визначення включає «будь-яку інформацію», слід припустити, що термін «персональні дані» слід тлумачити якомога ширше. Це також пропонується у судовій практиці Суду ЄС, який також розглядає менш явну інформацію, таку як записи робочого часу, що включає інформацію про час, коли працівник починає та закінчує свій робочий день, а також перерви чи час, який не потрапляє у робочий, як персональні дані. Крім того, письмові відповіді кандидата під час тестування та будь-які зауваження екзаменатора щодо цих відповідей є «персональними даними», якщо кандидата можна теоретично ідентифікувати. Те саме стосується і IP-адрес. Якщо особа має юридичну можливість зобов'язати постачальника передавати додаткову інформацію, яка дозволяє йому ідентифікувати користувача за IP-адресою, то така інформація – це також персональні дані. Крім того, слід зазначити, що особисті дані не повинні бути об'єктивними. Суб'єктивна інформація, такі як думки, судження чи оцінки, може бути персональними даними. Таким чином, сюди входить оцінка кредитоспроможності особи або оцінка результатів роботи роботодавця.

Регламент передбачає, що інформація для довідки про персонал повинна стосуватися фізичної особи. Іншими словами, захист даних не поширюється на інформацію про такі юридичні особи, як корпорації, фонди та установи. Для фізичних осіб, навпаки, захист починається і згасає з дієздатності. В основному,

людина отримує цю здатність своїм народженням і втрачає її після смерті. Отже, дані повинні бути присвоєні ідентифікованим живим особам, щоб вважатися особистими.

На додаток до загальних персональних даних, слід враховувати перш за все спеціальні категорії персональних даних (також відомих як конфіденційні персональні дані), які є надзвичайно актуальними, оскільки вони підлягають вищому рівню захисту. Ці дані включають генетичні, біометричні та дані про здоров'я, а також персональні дані, що виявляють расове та етнічне походження, політичні думки, релігійні чи ідеологічні переконання або членство в профспілках.

Турбота громадськості щодо конфіденційності зростає з кожним новим гучним порушенням даних. Згідно звіту про конфіденційність даних та безпеку даних RSA, для якого RSA опитував 7500 споживачів у Франції, Німеччині, Італії, Великобританії та США, 80% споживачів заявили, що найбільше занепокоєння викликає втрата банківських та фінансових даних [20]. Втрата інформації про безпеку (наприклад, паролі) та інформації про особу (наприклад, паспорти чи водійські права) була названою проблемою для 76% респондентів.

62% респондентів RSA заявляють, що будуть звинувачувати компанію за втрату даних у разі порушення, а не порушника чи хакера. Автори звіту дійшли висновку, що «По мірі того, як споживачі стають краще поінформованими, вони очікують більшої прозорості та чуйності від розпорядників своїх даних».

Відсутність довіри до того, як компанії ставляться до особистої інформації, змусила деяких споживачів вжити власних контрзаходів. Згідно з повідомленням, 41% респондентів заявили, що навмисно підробляють дані під час реєстрації на послугах в Інтернеті. Проблеми безпеки, бажання уникнути небажаного маркетингу або ризик перепродажу їх даних були серед головних проблем.

Звіт також показує, що споживачі не зможуть легко пробачити компанію, коли трапиться порушення, що викриває їхні персональні дані. Сімдесят два відсотки респондентів у США заявили, що будуть бойкотувати компанію, яка, імовірно, знехтувала захистом даних. П'ятдесят відсотків усіх респондентів сказали, що вони

частіше купуватимуть у компанії, яка може довести, що серйозно ставиться до захисту даних.

GDPR визначає кілька ролей, які відповідають за забезпечення відповідності: контролер даних, процесор даних та службовець із захисту даних (DPO) [21]. Контролер даних визначає, як обробляються персональні дані та цілі, для яких вони обробляються. Контролер також відповідає за те, щоб сторонні підрядники дотримуються вимог.

Обробники даних можуть бути внутрішніми групами, які ведуть та обробляють записи персональних даних, або будь-якою аутсорсинговою фірмою, яка виконує всю або частину цих видів діяльності. GDPR передбачає відповідальність процесорів за порушення або невиконання вимог. Тоді можливо, що ваша компанія та партнер з обробки, наприклад хмарний провайдер, нестимуть штрафні санкції, навіть якщо винна повністю на партнері по обробці.

GDPR вимагає від контролера та процесора призначення DPO для нагляду за стратегією безпеки даних та дотриманням GDPR. Компанії зобов’язані мати службовеця із захисту даних, якщо вони обробляють або зберігають великі обсяги даних громадян ЄС, обробляють або зберігають спеціальні персональні дані, регулярно контролюють суб’єктів даних або є державним органом. Деякі державні установи, такі як правоохоронні органи, можуть бути звільнені від вимоги організації з обмеженою відповідальністю.

GDPR покладає однакову відповідальність на контролерів даних (організація, яка володіє даними), і процесори даних (зовнішні організації, які допомагають керувати цими даними). Сторонній процесор, який не відповідає вимогам, означає, що організація не відповідає вимогам. У новому регламенті також встановлені суворі правила щодо повідомлення про порушення, котрі всі члени ланцюга повинні мати можливість виконувати. Організації також повинні інформувати клієнтів про їх права згідно з GDPR.

Це означає, що всі існуючі контракти з процесорами (наприклад, хмарними провайдерами, постачальниками SaaS або постачальниками послуг з оплати праці) та клієнтами повинні визначати обов’язки. Контракти також повинні визначати

послідовні процеси щодо управління та захисту даних та способу повідомлення про порушення.

Якщо компанія не відповідає вимогам GDPR, допускаються суворі штрафи до 20 мільйонів євро або 4% світового річного обороту, залежно від того, що вище, за невиконання. Однак більшість накладених на цей час штрафів були відносно невеликими. Згідно з GDPR Enforcement Tracker, ЄС виніс 282 штрафа станом на травень 2020 року [22]. Переважна більшість цих штрафів складають низькі тисячі та десятки тисяч євро. Найбільший штраф було накладено на Google, накладений у січні на 50 мільйонів євро, згідно з опитуванням про порушення даних GDPR DLA Piper від січня 2020 року [23]. Цей штраф був винесений за недостатню прозорість та дійсну згоду. Регулятори визнали, що у них немає ресурсів для обробки повідомлених порушень, які вони отримали, тому для створення ідентифікованих прецедентів знадобиться час.

«Запитайте двох різних регуляторів, як слід розраховувати штрафи GDPR, і ви отримаєте дві різні відповіді. Нам ще далеко від того, щоб мати юридичну впевненість у цьому вирішальному питанні», – сказав Патрік Ван Іке, голова міжнародної практики захисту даних DLA Piper [24]. Наразі здатність проявляти добросовісні зусилля дотримуватися їх має захистити компанії від жорстких покарань. У своєму виступі у 2018 році Ліз Денхем, уповноважений з питань інформації у Великобританії, сказав про це організаціям, стурбованим штрафами GDPR: «... Я сподіваюся, що ви вже знаєте, що примусове виконання є крайнім засобом .... Великі штрафи будуть зарезервовані для тих організацій, які наполегливо, навмисно чи з необережності порушують закон. Ті організації, які самостійно звітують, співпрацюють з нами для вирішення питань та демонструють ефективний механізм підзвітності, можуть розраховувати, що це буде фактором, коли ми розглядаємо будь-які регулятивні дії» [25].

Вимоги GDPR змушують компанії змінювати спосіб обробки, зберігання та захисту персональних даних клієнтів. Наприклад, компаніям дозволено зберігати та обробляти персональні дані лише тоді, коли особа дасть згоду та “не до вше, ніж це необхідно для цілей, для яких обробляються персональні дані”. Персональні дані

також повинні переноситися з однієї компанії в іншу, а компанії повинні видаляти персональні дані або анонімізувати їх за запитом. Це пояснюється як право на забуття. Є деякі винятки. Наприклад, GDPR не замінює жодної законодавчої вимоги про те, щоб організація зберігала певні дані. Це включатиме вимоги до медичної картки HIPAA.

Анонімізація повинна означати, що дані очищено від будь-яких потенційно ідентифікуючих характеристик перед публікацією та / або передачею третім особам. Однак справжню анонімізацію важко здійснити, особливо коли треба зберегти корисність набору даних для повторного використання. Можливість задовільно видалити ідентифікаційні дані приведе до можливості їх псевдонімізуванню, що залишає достатньо для третьої сторони для ідентифікації окремих предметів дослідження. Суб'єкти дослідження повинні знати, чи будуть їхні дані включені до опублікованого набору даних та чи буде відбуватися анонімізація в тому числі з метою мінімізації їх ідентифікації.

Вимоги певних організацій безпосередньо вплинуть на безпеку даних. Одне з них полягає в тому, що компанії повинні мати можливість забезпечити "розумний" рівень захисту даних та конфіденційності для громадян ЄС. Те, що GDPR означає під «розумним», не є чітко визначеним. Це дає керівному органу GDPR велику свободу дій щодо оцінки штрафів за порушення даних та їх недотримання. Викличною вимогою може бути те, що компанії повинні повідомляти про порушення даних контролюючим органам та особам, які постраждали від порушення, протягом 72 годин з моменту виявлення порушення. Інша вимога, що проводить оцінку впливу, призначена допомогти зменшити ризик порушень шляхом виявлення вразливих місць та способів їх усунення.

За словами Метта Фішера, лідера інформаційних технологій та старшого віце-президента Snow Software, понад 39 000 додатків містять особисті дані. «Ефект айсберга представляє серйозний ризик для дотримання організаціями вимог GDPR, оскільки багато хто зосереджений на 10% додатків, що містять особисті дані, які видно на поверхні води», – говорить він [26].

Загальний регламент ЄС про захист даних (GDPR) вимагає від організацій вживати відповідних технічних та організаційних заходів, включаючи політику, процедури та процеси, для захисту персональних даних, які вони обробляють.

ISO 27001, міжнародний стандарт СУІБ (системи управління інформаційною безпекою), забезпечує чудову вихідну точку для досягнення технічних та експлуатаційних вимог, необхідних для зменшення ризику порушення [27].

Тим часом ISO 27701 визначає вимоги та надає вказівки щодо створення, впровадження, підтримання та постійного вдосконалення PIMS (система управління інформацією про конфіденційність) на основі вимог, цілей контролю та контролю в ISO 27001 та розширена набором вимоги щодо конфіденційності, цілі контролю та засоби контролю [28].

Організації, які впровадили ISO 27001, зможуть використовувати ISO 27701 для розширення своїх СМІБ на охорону конфіденційності, включаючи обробку даних.

Впровадження обох стандартів допоможе задовольнити і продемонструвати відповідність вимогам щодо конфіденційності та захисту інформації GDPR [29].

Стаття 32 GDPR конкретно вимагає від організацій:

- вжити заходів щодо псевдонімізації та шифрування персональних даних;
- забезпечувати постійну конфіденційність, цілісність, доступність та стійкість систем обробки та послуг;
- своєчасно відновлювати доступність та доступ до персональних даних у разі фізичного чи технічного інциденту;
- впровадити процес регулярного тестування, оцінки та оцінки ефективності технічних та організаційних заходів для забезпечення безпеки обробки.

Стаття 32 додатково вимагає виявлення та пом’якшення ризиків «від випадкового або незаконного знищення, втрати, зміни, несанкціонованого розкриття або доступу до персональних даних».

СУІБ, що відповідає ISO 27001, відповідатиме всім вищезазначеним вимогам.

Стаття 32 GDPR є основним положенням, яка вимагає технічні заходи для захисту даних. Хоча вона наводить приклади заходів безпеки та контролю, ця стаття не містить детальних вказівок щодо того, що ви повинні зробити для цього.

Натомість GDPR змушує компанії розглянути існуючі найкращі практики та рекомендації, такі як ISO 27001, щоб мінімізувати ризик порушення даних.

ISO 27001 описує найкращі практики щодо СУІБ, систематичного підходу, що складається з людей, процесів та технологій, що допомагає захищати та керувати всією інформацією організації за допомогою управління ризиками.

СУІБ, пристосований до ISO 27001, приносить багато організаційних переваг, таких як:

- здатність надати переконливі докази того, що вжито необхідних заходів для дотримання вимог щодо захисту даних, що містяться в GDPR;
- захист усієї корпоративної інформації та інтелектуальної власності – не лише особистих даних;
- можливість зменшувати, контролювати та переглядати ризики, а також не відставати від постійно зростаючої загрози безпеці даних;
- культура обізнаності навколо інформаційної безпеки.

Компанії часто помилково вважають, що додавання шарів ультрасучасних технологій допоможе запобігти порушенням даних.

Без комплексної програми захисту інформації, яка також враховує людей і процеси, технологія не зможе забезпечити належний захист.

Погані процеси компанії та проблеми, пов’язані з персоналом, є одними з найпоширеніших моментів збою в безпеці даних.

Без зобов’язань керівництва (важливого критерію відповідності ISO 27001), доведено, що найкраще розроблені плани інформаційної безпеки не справляються.

Відповідність ISO 27001 означає, що компанія постійно переглядає та оновлює свої СУІБ відповідно до змін у середовищі загроз та розвитку бізнесу.

Без ефективної системи управління засоби контролю часто залишаються ізольовано, стають зайвими та непрацездатними.

Отримання сертифікації за стандартом ISO 27001 допомагає бізнесу отримати зовнішню, експертну оцінку ефективності своїх планів інформаційної безпеки, тим самим переконавшись, що заходи, які він ввів, працюють.

Ігнорування або недотримання повної вимоги GDPR може призвести до великих витрат для організації. СУІБ, що відповідає вимогам ISO 27001, може допомогти досягти відповідності GDPR економічно ефективним способом.

Окрім досягнення відповідності вимогам ISO 27001, організація повинна відповідати певним додатковим вимогам GDPR, на які поширюється система конфіденційності, наприклад ISO 27701. Впровадження обох стандартів дозволить відповідати вимогам щодо конфіденційності та захисту інформації GDPR та іншим законам про захист даних.

Висновки за розділом 1

Загальний регламент ЄС про захист даних (GDPR) намагається знайти баланс між тим, щоб бути достатньо сильним, щоб надати людям чіткий та відчутний захист, та бути достатньо гнучким, щоб враховувати законні інтереси бізнесу та громадськості. Він захищає великий набір даних, включаючи не лише особисту інформацію, таку як імена, посвідчення особи та номери соціального страхування, а також медичні дані, біометричні дані, політичні думки тощо.

GDPR зосереджується на конфіденційності даних та захисті персональної інформації. Це вимагає від організацій докладати більше зусиль для отримання явної згоди на збір даних та забезпечення законної обробки всіх даних. Однак йому бракує технічних деталей щодо того, як підтримувати належний рівень захисту даних або пом'якшувати внутрішні та зовнішні загрози. Такою деталлю є анонімізація, в якої немає фактичного визначення.

Таким чином в даній роботі, згідно досліджуваної мети, необхідно розглянути наступні задачі:

- проаналізувати основні типи загроз персональним даним;
- проаналізувати методи захисту персональних даних;

- визначити сценарії повторної ідентифікації даних після анонімізації;
- на основі цих сценаріїв виконати порівняльний аналіз ризиків повторної ідентифікації;
- розробити рекомендації щодо вибору підходу анонімізації.

РОЗДІЛ 2
ЗАГРОЗИ ПЕРСОНАЛЬНИМ ДАНИМ ТА МЕТОДИ ЗАХИСТУ ВІД НИХ

2.1 Основні типи загроз персональним даним

В зв’зку з розвитком інформаційних технологій з’являється все більше можливостей отримання доступу до інформації. Несанкціонована, необережна або необізнана обробка персональних даних може завдати великої шкоди людям і компаніям.

По-перше, метою захисту персональних даних є не просто захист даних особи, а захист основних прав і свобод осіб, пов’язаних із цими даними. Захищаючи особисті дані, можна гарантувати, що права та свободи людей не порушуються. Наприклад, неправильна обробка персональних даних може призвести до ситуації, коли особа залишається поза увагою щодо можливості працевлаштування або, що ще гірше, втрачає поточну роботу.

По-друге, недотримання правил захисту персональних даних може призвести до ще більш суворих ситуацій, коли можливо витягнути всі гроші з банківського рахунку людини або навіть спричинити ситуацію, що загрожує життю, маніпулюючи інформацією про стан здоров’я.

По-третє, нормативні акти щодо захисту даних необхідні для забезпечення справедливої та зручної для споживачів торгівлі та надання послуг. Норми захисту персональних даних спричиняють ситуацію, коли, наприклад, особисті дані не можуть вільно продаватися, а це означає, що люди мають більший контроль над тим, хто робить їх пропозиції та які пропозиції вони роблять.

Якщо особистідані просочуються, це може завдати компаніям значної шкоди їх репутації, а також спричинити штрафні санкції, саме тому важливо дотримуватись норм захисту даних особи.

Зі збільшенням використання Інтернету та мобільних додатків, досягнень аналітики та Інтернету речей, потреба в безпеці даних є більш важливою, ніж будь-

коли, враховуючи ризики нової відкритої системної вразливості та кібератак, а також величезні можливості для поєднання даних та відстеження кінцевих користувачів.

Проте безпека не полягає лише у застосуванні одного або декількох заходів, тому як жоден із них сам по собі не може забезпечити належний рівень захисту персональних даних. Навпаки, безпека персональних даних повинна слідувати ретельному та постійно контролюваному механізму контролю, як технічного, так і організаційного, що відповідає характеру обробки даних та пов'язаним з цим ризикам.

Щоб захистити найцінніший актив від загроз, необхідно їх зрозуміти. Розглянемо основні види загроз персональниї даних.

1. Фішингові атаки є однією з найбільших причин порушення даних у всьому світі. Останні дані звіту про розслідування порушень даних Verizon за 2019 рік свідчать про те, що найбільш успішні порушення стосуються фішингу та використання викрадених облікових даних [30]. Фішинг, який націлюють на приватних осіб, створюють злочинні компанії, які мають на меті обдурити людей завантаженням шкідливих програм та виведенням конфіденційних даних.

2. Шкідливе програмне забезпечення може з'їсти пропускну здатність, пошкодити файли та спричинити простої, втрату продуктивності та втрату цінних даних. Подібним чином програми-вимагателі можуть вразити будь-яку незахищену мережу.

Вимагальні програми є однією з найбільш руйнівних і найпродуктивніших загроз безпеці на сьогоднішній день, що може послабити будь-який бізнес. Програми-вимагачі блокують організації та кінцевих користувачів від своїх комп'ютерів, даних та мереж. Це зупиняє критичні комп'ютерні системи, поки жертва не заплатить викуп. Незважаючи на те, що програми-вимагателі зазвичай націлені на бізнес через фішинг-атаки, хакери часто використовують такі методи, як хробаки, щоб заразити всі комп'ютери, які підключаються до мережі.

Червоподібні можливості програми-вимагателя дозволяють швидко та експоненційно атакувати комп’ютери в мережах клієнтів, включаючи всіх, з ким ведеться бізнес.

3. RDP, який зазвичай використовується для зручності, дозволяє адміністраторам віддалено підключатися до комп’ютерів своїх користувачів. На жаль, кіберзлочинці сумно відомі тим, що експлуатують RDP для викрадення конфіденційних даних та встановлення задніх дверей та інших скалічуючих компромісів, таких як програми-вимагателі. Хакери розробили програми злому паролю, які можуть спробувати мільйони можливих варіантів паролів для доступу до віддалених комп’ютерів.

4. Публічний Wi-Fi може зашкодити організаціям. Хакери мають безліч способів викрасти дані та перехопити комунікації за допомогою атак «людина посередині» (MITM), пакетних снайперів, підробленого доступу та інших.

5. Людська помилка: коли хтось випадково видаляє файл або переміщує його в неправильне місце, дані можуть втратитись. Без резервних копій інформація пропадає.

6. Фізичний викрадення або втрата пристроїв також є цілком реальною загрозою, якщо вона потрапляє в чужі руки.

7. Співробітники компаній можуть використовують персональні пристрої для роботи, що може піддавати мережу та дані компанії зовнішнім загрозам.

8. Відсутність належного захисту брандмауера можуть призвести до спроб вторгнення, які можуть негативно вплинути на сервери (і навіть можуть бути успішними).

9. Погодні події або перебої з електропостачанням можуть збити системи або пошкодити їх, залишаючи без критичних систем або даних.

10. Доступ до мережі або даних компанії за допомогою незахищеного віддаленого доступу може піддати будь-яким ризикам, які можуть ховатися на домашньому ПК працівника.

11. Ботнети беруть на себе будь-який пристрій (навіть «розумний холодильник»), щоб надсилати спам або виконувати хакерські дії, спрямовані на інші мережі. Вони в основному перетворюють пристрій на «хакерську машину».

2.2 Аналіз методів захисту персональних даних

Для забезпечення безпеки персональних даних важливо знати, які дані обробляються, чому вони обробляються та на яких підставах. Крім того, важливо визначити, які заходи безпеки застосовуються. Це можливо завдяки аудиту захисту даних, який визначає потік даних та чи дотримуються правила захисту даних. Аудит може бути проведений шляхом відповіді на набір конкретних питань, які були підготовлені з цією метою.

Як правило, виділяють чотири класи методів захисту персональних даних в інформаційних системах. По-перше, фізичні методи, по-друге, апаратні, по-третє, програмні, нарешті, по-четверте, організаційні [31].

До організаційних методів захисту персональних даних відносять наступні засоби:

- розробка внутрішніх нормативних документів, в яких повинні бути встановлені правила роботи з конфіденційною інформацією і комп’ютерною технікою;
- періодичні перевірки персоналу і інструктаж стосовно збереження конфіденційних даних;
- підписання додаткових угод до трудового договору, в яких чітко прописана відповідальність працівника за неправомірне використання або розголошення відомостей, які стали відомі в процесі здійснення професійної діяльності;
- розмежування зон відповідальності для виключення тих ситуацій, коли найбільш важлива інформація знаходиться в доступі тільки одного співробітника;

- організація робіт в загальних програмах документообігу і відстеження за особливо важливими файлами;
- впровадження програмних комплексів, які захищають інформацію від знищення або копіювання будь-яким користувачем системи, в тому числі топ-менеджером компанії;
- складання планів, які можуть відновити систему в тому випадку, якщо вона вийде з ладу.

Фізичний захист здійснюється за допомогою таких речей як служба охорони, система захисту вікон і дверей, лазерні та оптичні системи, які реагують на перетин зловмисником світлових променів. Тобто фізичні методи захисту мають на увазі під собою фізичну заборону доступу до персональних даних.

Апаратні методи захисту можливо реалізувати за допомогою спеціальних пристроїв. До таких засобів можна віднести різні схеми блокування від несанкціонованого використання персональних даних. Апаратні засоби застосовуються в складі ЕОМ.

Нарешті, програмний захист здійснюється за допомогою програм, до яких можна віднести операційну систему, антивіруси, спеціальні програми захисту та інші.

Мабуть, саме апаратно-програмні засоби захисту персональних даних найбільшою мірою дозволяють захищати персональні дані від несанкціонованого доступу до них.

Апаратно-програмний захист досягається застосуванням таких способів захисту як:

1. Захист від несанкціонованого використання персональних даних з боку користувачів і програм, в тому числі і при наявності доступів.
2. Захист від некоректного використання наявних ресурсів.
3. Висока ступінь якості використовуваних апаратно-програмних засобів.

В цілому, перелік технічних заходів щодо захисту персональних даних в інформаційній системі виглядає наступним чином:

- недопущення несанкціонованого доступу до персональних даних за допомогою антивірусного програмного забезпечення і системи паролів;
- діяльність з виявлення фактів несанкціонованого доступу і використання персональних даних (наприклад, оновлення антивірусного програмного забезпечення);
- охорона, а також регламентування використання технічних засобів, за допомогою яких відбувається обробка персональних даних з метою недопущення порушення їх функціонування;
- забезпечення віддаленого зберігання і резервного копіювання найбільш важливих інформаційних даних на регулярній основі;
- резервування і дублювання всіх підсистем, які містять важливу інформацію;
- перерозподіл ресурсів мережі в тому випадку, якщо порушена працездатність її окремих елементів;
- забезпечення можливості застосовувати резервні системи електричного живлення;
- забезпечення безпеки інформаційних даних та належного захисту в разі виникнення пожежі або пошкодження комп'ютерного обладнання водою;
- установка такого програмного забезпечення, яке зможе забезпечити належний захист інформаційних баз даних у разі несанкціонованого доступу.

2.3 Основні механізми деідентифікації даних

Елементом системи захисту персональних даних є організаційні заходи, які спрямовані на протидію загрозам для системи і метою яких є мінімізація можливих збитків користувачів і власників системи.

Кожна організація має власні способи та методи захисту персональних даних. У сучасну епоху щорічних придбань, реорганізацій та «синергетичних переїздів» цілком ймовірно, що компанії будь-якого розміру мають кілька забутих серверів та

баз даних, які утримуються без поважних причин. Якщо деідентифікувати дані кількість інформаційних порушень знизиться. Розглянемо механізми деідентифікації даних.

Перший метод, маскування даних – це той, який у багатьох організаціях використовується як загальний висновок для всіх методів [32]. Зазвичай він може підтримувати деідентифікацію багатьох типів даних, чи то прямі чи непрямі ідентифікатори та різноманітні підходи збереження формату, збереження семантики, послідовні та складені варіанти маскування. І часто зазвичай підтримують API хуки, щоб дозволити користувацьке маскування або включення користувацьких стилів словників для заміни.

Токенізація або псевдонімізація – це підхід, який замінює вихідне значення символьним значенням, яке, як правило, математично не пов’язане з початковим значенням. Підходи можуть бути повторюваними чи не повторюваними та оборотними. Ступінь безпеки, доступний підходом токенізації, часто пов’язаний з підходом до управління ключовим матеріалом, який використовується для генерації або створення маркерів. Вони можуть варіюватися від вбудованого в програмне забезпечення до підходів, що базуються на апаратній безпеці.

Третій підхід – анонімізація – це один з тих підходів, які часто вважаються більш сучасними, коли вони прагнуть досягти певної доказової гарантії корисності, одночасно забезпечуючи належну ступінь анонімізації або анонімності для суб’єкта даних. Деякі типові підходи включають k-анонімність або диференційовану конфіденційність. Насправді будь-який із них може придушити або усунути відхилення від даних, але важко отримати справжню ідентичність суб’єкта даних від загальної сукупності.

Виробництво даних – це техніка, яка зростає в популярності та намагається уникнути ризиків викриття особистих чи конфіденційних даних, використовуючи повністю сфабриковані або замінні дані. Цей прийом може бути єдиним варіантом у випадках, скажімо, дуже високочутливих даних, коли ризики просто занадто високі, щоб спробувати деідентифікувати вміст. Але це має ряд плюсів і мінусів, залежно від складності даних та цільових випадків використання.

Застосування деяких оцінок ризику повторної ідентифікації персональних даних є виправданим залежно від зон даних, користувачів та випадків використання. Це стає все більш актуальним, оскільки створюється все більше правил щодо конфіденційності даних, які закликають організації зрозуміти ризики повторної ідентифікації та застосувати деякі можливості для оцінки цих ризиків.

2.3.1 Маскування даних

Концепції маскування даних для багатьох організацій може бути загальним терміном, яку вони використовують для опису всіх своїх можливостей де-ідентифікації. Як вже зазначалось, кожна організація, ймовірно, матиме власний спосіб, коли мова заходить про методи конфіденційності даних, які вони використовують.

Коли з’являлося більше норм щодо конфіденційності даних, було докладено більше зусиль для створення конкретного розуміння різних методів, більш детальної диференціації між маскуванням та редагуванням та токенізацією та анонімізацією. Оскільки це може бути багато речей, теоретичне маскування даних зазвичай можна застосовувати як до непрямих, так і до прямих ідентифікаторів. Але потрібно чітко визначити, який з результатів заміщення значень даних матиме найбільшу цінність у випадках використання даних. Отже, це можуть бути речі, такі як узагальнення даних, генерація випадкових чи послідовних чисел, перетасування значень, що відсортовані за датою, і навіть пошук значень заміни за попередньо визначеними словниками, редагування, замінюючи дані нічим або якимись пробілами чи спеціальними символами. Вони можуть включати контекстуально точні підходи як семантичні, так і підходи до збереження формату. І може мати певне накладення на підходи до токенізації, створюючи спеціальний маркер, який не базується на вихідному значенні.

Маскувальні рішення, що просувають такий тип можливостей, мають менше ключових можливостей управління, ніж складніші засоби токенізації. Вони можуть застосовуватися тоді, як правило, до прямих та / або непрямих ідентифікаторів.

Однію з ключових можливостей під час оцінки того, хто має інструмент маскування, є можливість використання семантичного збереження.

Складене маскування – це здатність або витягувати, або імпортувати, або використовувати існуючі взаємозв’язки між даними у джерелах даних, щоб вигадані значення зберігали ці зв’язки або цілісність даних. Прикладом може бути те, що якщо існує дата вступу та дата виписки в лікарні, розуміється взаємозв’язок між цими значеннями і не створюю дату виписки до дати вступу або дати реєстрації. І тоді послідовне маскування даних – типова характеристика чи можливість, яка гарантує, що всі екземпляри вихідного значення можуть бути замінені або вигадані однаково. Незалежно від того, скільки разів здійснюється маскування або де б це значення не існувало в ландшафті даних.

Це значна техніка та можливість, особливо у випадках використання тестових даних. Прикладом деяких простих підходів до маскування даних може включати надання випадкових значень заміни, навіть якщо вони є із заздалегідь виз наченого словника чи списку. Або навіть пряма редакція таких елементів, як стать чи вік людини. Це підтримувало б випадки використання, коли конкретні значення даних заміни насправді не мають значення для даного випадку використання, як у простому блоці або, можливо, дуже простому середовищі функціонального тестування.

Більш складним маскуванням даних може бути формулювання нового номера кредитної картки з існуючого номера, який не тільки підтримує збереження перших шести значень ідентифікатора банку, але й виробляє заміну, яка проходить перевірку Luhn [33]. Це те, що відрізняє його від деяких більш досконалих методів токенізації або методів анонімізації порівняно з базовою технікою маскування.

Далі припускається, що існує набір даних про клієнтів та їх замовлення, і необхідно здійснити маскування даних для використання тестового управління даними. Визначається, що для цього цільового тестового середовища, де також існує низка інших засобів контролю безпеки, існує кілька прямих ідентифікаторі. Вони повинні бути замасковані, щоб стати чимось іншим, але це потрібно робити послідовно. Можливе використання заміни значення з таблиці пошуку адрес, які не

обов’язково повинні стосуватися існуючих клієнтів або вихідних даних. Для значення ідентифікатора замовника можна застосувати повторюваний алгоритм, такий, що маскує або створює іншу літеру для букви або цифри. Потрібно було б перевірити, чи інструмент маскувального механізму підтримує можливість це для кожгого унікального значення. Для імен можливо використовуватися підхід пошуку даних із таблиці заміни.

Можливе створення хеш-значення з чогось на кшталт ідентифікатора касти у вихідних записах, яке є унікальним і пов’язане з іменем замовника. Це може бути використано як частина підходу до пошуку хешу у заздалегідь визначеній таблиці пошуку імен заміни. Якщо бажано менше, ніж більше повторюваних імен у вихідних даних, тоді домен значень заміни повинен бути побудований більш громіським, ніж загальний домен імен клієнтів у джерелі. І тоді для адресних даних можна вибрати подібний пошук значень заміни, але обраних випадково. Особливо, якщо визначено, що в якості результатів у процесі тестування насправді потрібні лише формати даних, і що немає вимоги для перевірки справжності адрес.

Замовлення, той самий підхід, що використовується для ідентифікатора касти в таблиці даних клієнта, може бути використаний для заміни зовнішнього ключа ідентифікатора касти в таблиці замовлення.

2.3.2 Токенізація

Токенізація – це механізм, який замінює вихідне значення індивідуальним значенням заміщення, яке не має математичного зв’язку з вихідним значенням даних. Іноді токени називають також псевдонімами або процесом псевдонімізації. Багато нових норм конфіденційності даних згадують псевдонімізацію як один із методів, який, за очікуванням організацій, можна використовувати як частину своїх заходів щодо захисту персональних даних або як частину загальної системи управління тестовими операціями (TOMS) чи технічних та організаційних заходів, місце для захисту персональних даних.

Токенізація – це не те саме, що шифрування. Шифрування зазвичай використовує якийсь математичний або криптографічний алгоритм для захисту даних, роблячи їх нечитабельними. Вона використовуватиме набори ключів, які є частиною як процесу шифрування, так і дешифрування уповноваженими сторонами. Це означає, що авторизовані користувачі або будь-хто, хто отримує криптографічні ключі, можуть розшифрувати захищені дані, щоб зробити їх читабельними. Мережа шифрування підтримує математичне відношення до вихідної точки даних. А це означає, що методи шифрування настільки ж добрі, наскільки міцні їх алгоритм і / або засоби управління навколо клавіш.

Токенізація заснована на заміні конфіденційних даних незалежно від їх характеру: інформацію, що ідентифікує особу, інформацію про стан здоров'я, інформацію про особисту кредитну картку, будь-якого типу чи формату із нечутливою заміною. Оригінальна інформація не міститься в маркері. Таким чином, сам маркер не може бути використаний в будь-якому звороті маркера назад до справжніх даних. Токени можуть бути різними стилями виводу, включаючи ті, які можуть зберегти такі речі, як довжина, тип і формат.

Застосування токенізації може значно пом'якшити вплив на персональні дані, якщо порушення даних відбувається хоча б для прямих ідентифікаторів, де це найчастіше використовується. Отже, дивлячись на деякі характеристики та розміри, токенізаційні підходи опрацьовують наявні дані, хоча вони не походять від них, вони не є математичними розрахунками з вихідних значень. Вони можуть переглядати вихідні дані, щоб отримати деяке розуміння вимог до збереження формату, але вони, як правило, не включають в це алгоритмічно.

Маркер є сурогатною цінністю, псевдонімом, заміною базової вартості. Найбезпечніші підходи до токенізації використовуватимуть високо контрольоване та безпечне управління ключами, або те, що в галузі згадується як процес генерації ключових матеріалів, і це може бути за допомогою модулів апаратного захисту або інструментів управління життєвим циклом програмного забезпечення або певної їх комбінації. І як вже зазначалося, вони дуже часто застосовуються до прямих ідентифікаторів суб'єкта даних. Деякі приклади криптографічних схем включають

односторонній хеш: HMAC, що використовує SHA-256, який створить довге 64-бітове значення шифрування AES, яке може бути двостороннім [34]. Тож дуже схожий механізм шифрування. Деякі виміри криптографічних токенів включають фізичну структуру маркера. Маркери за визначенням можуть виглядати як вихідне значення в типі та довжині даних, одночасно деідентифікуючи конфіденційну інформацію, що дозволяє їй пройти через свій життєвий цикл без будь-яких або обмежених змін у системах, або вони можуть бути довгими стилями HMAC.

Насправді все залежить від вимог до випадків використання, і оскільки зазначена зворотність і незворотність є ключовим варіантом при токенізації, це слід враховувати. Найбезпечніші середовища, як правило, шукають використання незворотних підходів. Найпоширеніші оборотні підходи забезпечуються використанням захищеного сховища токенів, де можна надійно отримати відповідність маркерів оригінальним значенням. Однак цей підхід, як правило, використовується лише у виробничих середовищах, де фактичне отримання справжніх значень даних можливе лише уповноваженим особам персоналу, що є обов'язковою вимогою використання. І можливо, всі інші користувачі цих даних повинні мати доступ лише до значень маркерів. Невиробничі середовища ніколи не повинні застосовувати оборотні підходи до токенізації. І як вже відзначалося, найбезпечніші підходи до токенізації використовують ключі або ключовий матеріал при їх обробці для створення сурогатних значень, а також застосовуватимуть високозахищені підходи до управління ключами, такі як модулі апаратної безпеки або програмні системи управління ключами.

Токенізація – одна з ключових складових процесу деідентифікації, яка підтримуватиме цілі конфіденційності. Це також досягає цього за допомогою декількох ключових характеристик, що можна зробити безповоротно, використовуючи незворотні підходи до токенізації для прямих ідентифікаторів. І це може бути зроблено повторно, як для даних, так і для подальших вправ з ідентифікації, що забезпечує певний рівень корисності, який, ймовірно, буде важливим для будь-якої роботи з моделювання сегментації та класифікації. Крім того, у цьому випадку використання існує ще кілька додаткових вимог.

Припускається, що дані матимуть велику пропускну здатність, тому середовище обробки повинно бути чимось подібним до Hadoop, щоб мати можливість розподіленого підходу до обробки [35]. Знадобиться високобезпечний підхід до токенізації, тобто бажано мати модуль апаратної безпеки та / або підходи до управління життєвим циклом програмного забезпечення, а також зашифровані зони даних для обробки. І стилі маркерів, які необхідно створити, повинні бути гнучкими, тобто використовуватимуться випадки, коли хеш-значення підходу HMAC є цілком дійсним, але інші, де буде потрібно зберегти певний формат та всю семантику вихідних даних, тоді як все ще токенізуючи баланс змісту.

Прикладом є деякі вхідні дані, для яких визначаються елементи, які потрібно токенізувати, і зосередження відбувається лише на двох. Перший – це номер кредитної картки, інший – деякий порядковий номер цього рахунку. Тоді створюється середовище для обробки даних з відповідними компонентами, що підтримують усі вимоги до обробки та безпеки, зі збереженою високою пропускною здатністю та надійним управлінням ключами, виробляючи відповідні вихідні дані маркера для кожного з цільових прямих ідентифікаторів. У цьому випадку створюється оборотня токенізація для цього номера кредитної картки та номера рахунку, а середовище даних виглядатиме приблизно так: хеш-значення для фактичного номера картки буде схожий на вигляд, але не той же маркер від початкового значення, зберігаючи, можливо, деякі провідні символи. Як зазначалося, на додаток до токенізації існуватиме ряд приватизаційних методів. Отже, у випадку якого-небудь порушення даних у цій зоні, теоретично можливим був би лише підхід грубої сили, що відміняє токенізацію, і це вимагало б, щоб порушники мали доступ до ключового матеріалу, який є надійно захищений.

2.3.3 Анонімізація

Анонімізація – це термін, який найчастіше використовується для звернення до типово непрямих ідентифікаторів, які створились в результаті поєднання методів токенізації даних та / або маскування. Інші непрямі ідентифікатори не виявляють

непрямих способів повторної ідентифікації даного суб'єкта даних. Наприклад, накопичення місця розташування магазину та статі, а також придбаних товарів та інформації про час доби, пов'язаної навіть із символічним ідентифікатором клієнта, має достатньо деталей у поєднанні з простою інформацією, що спостерігається зовні, щоб призвести до повторної ідентифікації людини. Це особливо вірно, якщо ці непрямі атрибути є винятком серед популяції або надзвичайно унікальними або спостережуваними зовні атрибутами про людину чи її поведінку. Анонімізація часто вимагає тонкого балансу або розуміння того, наскільки справжні статистичні розподіли даних мають важливе значення для успіху випадку використання фокусу. Анонімізація набуває все більшої важливості у багатьох організаціях, оскільки здійснюється та створюється все більше ініціатив та середовищ для розвитку машинного навчання та інтелектуального розвитку. Можливе отримання деяких характеристик анонімізації (Рис. 2.1), які можуть бути опрацьовані та застосовані для різного збурення даних, і це є частиною науки про деідентифікацію, яка використовується для захисту чутливості.

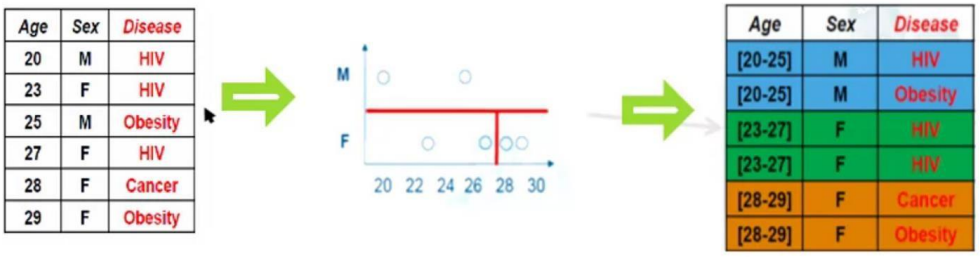

Рисунок 2.1 – Приклад анонімізації даних

K-анонімність – це перетворення даних таким чином, що кожен запис неможливо відрізнити принаймні від k-1 ідентичності.

Диференціальна конфіденційність – це техніка, яка досягається шляхом рандомізації різних значень, додавання шуму, таким чином, що відбувається узагальнення даних або збільшення подібних появ цих атрибутів у сукупності. Але до такої міри, яка суттєво не спотворює статистику населення.

Це складні методи, які існують протягом багатьох років на інструментальних стрічках статистиків та науковців. Є атрибути, які самі по собі, як правило, не розкривають особистість людини, але в поєднанні з іншими непрямими атрибутами

та знанням того, що всі вони прив’язані до єдиного, навіть якщо це символічний псевдонім, можуть виявити особистість суб’єктів даних або особисті або чутливі характеристики щодо їх. До того ж, залежно від рівня їх унікальності та повної сукупності даних може призвести до повторної ідентифікації суб’єкта, навіть якщо можливо бачити лише пов’язаний з ними псевдонім. Це підводить нас до важливої концепції, щоб чітко пояснити, що концепція суб’єкта даних повторно ідентифікується.

Ризики зростають, коли будь-який з бітів інформації може бути спостережуваний зовні або може бути пов’язаний з іншими загальнодоступними даними або якщо дані надзвичайно чутливі. Не має значення, чи ця здатність повторно ідентифікувати щось, що може статися лише в безпечних межах топології або віддалених можливостей, все ще порушує порушення прав суб’єктів даних. Припускається, що є дата народження та поштовий індекс як частина адреси та показник етнічної приналежності у вихідних даних. Ця інформація є критично важливою. Можна вирішити, що ці дані потребують подальшого збурення.

Використання елементів конфіденційних даних без будь-якої деідентифікації було б можливим, але це вимагає застосування більш суворого захисту зони даних та/або контролю доступу користувачів навколо. Зони потребують обробки як виробничу зону, а також усі суворі та накладні витрати, які це може означати. Припускається, що у цьому випадку застосовується деяка анонімізацію. Тож, повернувшись до прикладу використання даних, існують ті самі вимоги щодо обсягу пропускної здатностіобробки даних, які вимагаються, вбудувавши анонімізацію або набори бібліотек анонімізації в робочий процес обробки. Для тих самих архітектурних конструкцій для великого обсягу та високої швидкості обробки.

Необхідно анонімізувати деякі з наявних непрямих ідентифікаторів, перетворивши дати народження у віковий діапазон, і це іноді називають збіркою цієї інформації та скоротивши інформацію про поштовий індекс таким чином, щоб він виглядав на ширшій області або, можливо, на столичній статистичній області, що, як правило, є першими трьома цифрами поштового індексу з п’ятизначного коду. З огляду на високу чутливість необхідно повністю усунути чи видалити інформацію

про етнічну приналежність, оскільки вона є надто чутливою і насправді може сприяти аналітичним результатам, які включають деякі форми упередженості.

Існує ряд інших, більш складних методів анонімізації, які, наприклад, можуть внести більше шуму в дані, не порушуючи статистичного розподілу. Але всі підходи повинні бути переглянуті та оцінені з користувачами стосовно випадків використання, а головне з розробниками політики конфіденційності даних, щоб знайти необхідне середовище захисту даних та корисності. У цьому випадку, структурні дані, які змінять дати народження на діапазон, який скорочує та розширює географічну ідентифікацію поштового індексу або адреси, повністю редагує або видаляє інформацію про етнічну приналежність.

Шифрування даних – це загальноприйнята техніка, що використовується у виробничих середовищах для конфіденційних даних і навіть у невиробничих середовищах як основний елемент механізмів захисту даних, які будуть використовуватися на рівнях зберігання даних та зв’язку. Освоєння можливостей шифрування даних є основною вимогою захисту, і, як і всі методи, які наведено в цьому розділі, повинні бути включені в комплексні можливості, які організація розглядає для створення повного рішення із захищеними даними для відповідної топології даних та випадків фокусування.

Інший підхід до збереження конфіденційності даних, який набирає популярності – це мінімізація відбитків персональних даних в організації. Хоча і не техніка збурення даних сама по собі. Зменшуючи, де в організації зберігається фактичний особистий вміст. Ця конструкція може мати далекосяжні наслідки захисту даних. Візьмемо для прикладу підхід, який вирішує зробити загальними всі дані про клієнтів у єдиному оцінюваному сховищі, яке має підключення до каналів, та створює зовнішні ключі до всіх систем, яким може знадобитися доступ до цієї інформації про клієнта.

Таким чином, супутникові системи ніколи не мають жодного фактичного вмісту суб’єкта даних. Тож про будь-яку їх копію невиробничого середовища буде менше турботи при оцінці чи застосуванні деідентифікації. Здебільшого, необхідно розглянути непрямі типи ідентифікаторів. Відсутність копій персональних даних,

що надлишково розповсюджуються по організації в багатьох системах, не тільки зменшує необхідну роботу із захисту, але також може значно зменшити іншу роботу, пов’язану з багатьма зобов’язаннями суб’єкта даних, які переважають у багатьох правилах конфіденційності, таких як збереження персональних даних, надання детальної інформації про всю обробку персональних даних і навіть видалення персональних даних, коли вимагаються права на забуття.

Існує ряд методів анонімізації даних. Хоча багато з цих методів призначені для достатньої маскування даних, деякі, можливо, доведеться використовувати разом з іншими, щоб забезпечити анонімність як прямих, так і непрямих ідентифікаторів. Кілька найпоширеніших методів анонімізації даних включають:

Маскування символів: під час маскування символів або “маскування” формат даних зберігається, але вибрані символи замінюються символом маски, наприклад “x” або “#”. Прикладом маскування символів може бути зміна дати народження 24.03.1955 р. На ## / ## / 19 ##.

Перетасовка даних: цей метод, також відомий як обмін даними, передбачає перестановку даних таким чином, щоб атрибути даних залишалися присутніми, але не відповідали їх початковим записам. Перетасовка даних часто прирівнюється до перемішування колоди карт. Цей метод ефективний, коли немає необхідності оцінювати дані на основі взаємозв’язку між інформацією, що міститься в кожному записі.

Заміна даних: при заміні дані зі стовпця повністю замінюються випадковими значеннями зі списку фальшивих, але схожих даних. Наприклад, прізвища можуть бути замінені на інші, нерелевантні прізвища, або номери кредитних карток можуть бути замінені випадковим рядком із 16 чисел. Для належного використання цього методу користувачі повинні мати списки, що дорівнюють або перевищують обсяг даних, які вони намагаються анонімізувати.

Узагальнення: узагальнення працює шляхом усунення специфіки даних та заміни їх на більш загальну, але все ще актуальну інформацію. Це часто досягається використанням діапазонів. Замість того, щоб говорити про 33 роки, узагальнені дані

можуть говорити про те, що людині від 30 до 40 років. Для адрес можна вказати лише назви доріг.

Відхилення числа та дати: алгоритми можна використовувати для зміни значення числових даних на випадкові відсотки. Цей невеликий крок може мати велике значення, якщо його застосувати належним чином.

Скремблювання: при правильному скремблюванні літери змішуються та переставляються так інтенсивно, що вихідні дані неможливо визначити. Спрощеним прикладом цього є перетворення імені «Даніель» у «Леніада», «Жаклін» у «Какайнейджю» тощо.

Числове розмиття: замість того, щоб повністю замасковано, розмиті значення даних змінюються, тому їх просто вистачає від їх фактичного значення, щоб захистити особистість особи. Числове розмиття може виконуватися різними способами, включаючи звітування округлих значень або середніх показників групи.

Придушення: у певних випадках у наборі даних можуть існувати стовпці даних та / або записи, які жодним чином не допомагають оцінювачу даних, але містять ідентифікаційну інформацію. У цих випадках найкраще придушити або видалити стовпці та / або записи. Важливо, щоб дані були повністю видалені з електронної таблиці, а не просто приховані.

Синтетичні дані: на відміну від інших методів анонімізації даних, набори синтетичних даних є імітаційними версіями фактичних даних, а не модифікованими даними. Ці синтетичні набори даних мають багато спільного з реальними даними, наприклад, формат та взаємозв’язки між атрибутами даних, і використовуються, коли для тестування системи потрібен великий обсяг даних, а фактичні дані не можуть бути використані.

Хоча анонімізація даних може призвести до значного прогресу для компаній у різних секторах, вона не позбавлена обмежень та ризиків. При неправильному виконанні або зі слабкими алгоритмами погана анонімізація може призвести до:

Розкриття особистості: також відоме як виділення, розкриття особистості – це термін, що використовується для опису ситуацій, в яких можна ідентифікувати всіх або деяких осіб у наборі даних.

Розкриття атрибутів: це здатність визначити, чи зберігається атрибут у наборі даних конкретною особою. Наприклад, анонімний набір даних може показати, що всі працівники відділу продажів певного офісу прибувають після 10 години ранку. Якщо відомо, що конкретний працівник знаходиться у відділі продажів цього офісу, ви знаєте, що вони прибувають після 10 години ранку. , навіть якщо їх конкретна ідентичність маскується в наборах даних.

Зв’язок: Зв’язок відноситься до того, коли можна підключити кілька точок даних, будь то в одному наборі даних або окремих наборах даних, щоб створити більш згуртовану картину конкретної особи.

Розкриття висновків. Розкриття висновків відбувається, коли ви можете впевнено зробити висновок про значення атрибута на основі інших атрибутів.

Правильна анонімізація даних може зайняти багато часу і складно, але дуже важливо, щоб ці методи виконувались досвідченими професіоналами точно, якщо компанії хочуть підтримувати відповідність нормативним актам та відвернути зловмисників. Існує ряд інструментів та програмного забезпечення для анонімності даних, які можуть допомогти компаніям подолати бар’єри щодо анонімізації даних та надійно скористатися багатьма перевагами, які може запропонувати збір даних.

Висновки за розділом 2

В даному розділі було проаналізовано основні загрози персональним даним та методи захисту від них. Найбільшу увагу приділено механізмам деідентифікації даних, які дозволяють використовувати дані багатьма способами, які не можуть забезпечити жодні інші засоби безпеки, такі як шифрування або контроль доступу, водночас зменшуючи ризики конфіденційності.

Захист даних – це процес захисту файлів, баз даних та облікових записів у мережі шляхом прийняття набору елементів керування, програм та методів, що визначають відносну важливість різних наборів даних, їх чутливість, вимоги до відповідності нормативним документам, а потім застосовують відповідні засоби захисту для їх захисту.

Будь-яка система, яка обробляє або зберігає персональні дані, повинна захищати персональні дані, наприклад, шляхом деідентифікації, зведення до мінімальної форми, необхідної для цілей контролера даних або токенізація, яка замінює персональні дані випадковими токенами.

Процес, який застосовує до даних деідентифікацію, також відомий як анонімізація даних. Загальні стратегії включають видалення або маскування особистих ідентифікаторів, таких як особисте ім’я , та придушення або узагальнення квазіідентифікаторів, таких як дата народження. Зворотний процес використання неідентифікованих даних для ідентифікації осіб відомий як повторна ідентифікація даних. Успішні повторні ідентифікації ставлять під сумнів ефективність деідентифікації.

РОЗДІЛ 3
ЗАСТОСУВАННЯ ЕЛЕМЕНТІВ АНОНІМІЗАЦІЇ ДЛЯ ЗАХИСТУ ПЕРСОНАЛЬНИХ ДАНИХ

Ефективним підходом до анонімізації даних є k-анонімність [36-39]. За допомогою k-анонімності оригінальний набір даних, що містить особисту інформацію про здоров'я, може бути трансформований так, що зловмиснику важко визначити особу осіб у цьому наборі даних. K-анонімований набір даних має властивість, що кожен запис подібний принаймні до іншого k -1 інших записів щодо потенційно ідентифікуючих змінних. Наприклад, якщо k = 5, а потенційно ідентифікуючими змінними є вік і стать, тоді k-анонімований набір даних має принаймні 5 записів для кожної комбінації значень віку та статі. Найбільш поширені реалізації k-анонімності використовують такі методи трансформації, як узагальнення, глобальне перекодування та придушення [36, 37, 39-42].

Будь-який запис у k-анонімованому наборі даних має максимальну ймовірність $1/k$ бути повторно ідентифікованим [41]. На практиці власник даних вибирає значення k, пропорційне імовірності повторної ідентифікації, яку вони готові допустити – пороговий ризик. Більш високі значення k означають меншу ймовірність повторної ідентифікації, але також більше спотворень даних, а отже, більшої втрати інформації через k-анонімізацію. Взагалі, надмірна анонімність може зробити розкриті дані менш корисними для одержувачів, оскільки певний аналіз стає неможливим або аналіз дає необ'єктивні та неправильні результати [43-48].

В ідеалі, фактична ймовірність повторної ідентифікації k-анонімізованого набору даних була б близькою до $1/k$, оскільки це врівноважує толерантність до зберігача даних до ступеня спотворення, яке вводиться внаслідок k-анонімізації. Однак, якщо фактична ймовірність набагато нижча, ніж $1/k$ тоді k-анонімність може бути надмірно захисною, а отже, призводить до непотрібних надмірних спотворень даних.

3.1 Сценарії повторної ідентифікації для k-анонімізованого набору даних

Проблема k-анонімності полягає у повторній ідентифікації однієї особи в анонізованому наборі даних. Існує два сценарії повторної ідентифікації для однієї особи [49-51]:

- повторне встановлення конкретної особи (відоме як сценарій повторної ідентифікації прокурора). Зловмисник (наприклад, прокурор) знав, що конкретна особа (наприклад, обвинувачений) існує в анонімній базі даних, і бажає з'ясувати, який запис належить цій особі.
- Повторне ідентифікування довільної особи (відоме як сценарій повторної ідентифікації журналіста). Зловмисну байдуже, яку особу повторно ідентифікувати, він зацікавлений у можливості заявити, що це можна зробити. У цьому випадку зловмисник бажає повторно встановити особу, щоб дискредитувати організацію, що розкриває дані.

Набір пацієнтів у файлі, який слід розкрити, позначається символом s. Перш ніж файл s може бути розкритий, його слід анонімувати. Деякі записи у файлі будуть вимкнені під час анонімізації, тому інша підмножина пацієнтів s' буде представлена в анонімізованій версії цього файлу. Нехай анонімізований файл позначається символом ζ. Існує індивідуальне відображення між записами в ζ та особами в s'

За сценарієм прокурора, конкретна особа переосмислюється, скажімо, VIP. Зловмисник зрівняє VIP із записами ζ на квазіідентифікаторах. Такі змінні, як стать, дата народження, поштовий індекс та раса, зазвичай використовуються як квазіідентифікатори. Записи в ζ які мають однакові значення на квазіідентифікаторах, називаються класом еквівалентності.

Нехай буде кількість записів, ζ які мають точно такі самі значення квазіідентифікатора, що і VIP f Тоді ризик повторної ідентифікації для VIP є $1/f$.

Наприклад, якщо особа, яку повторно ідентифікують, є чоловіком 50 років, то f це кількість записів про чоловіків 50 років у ζ. Зловмисник має ймовірність $1/f$ отримати правильний збіг.

Оскільки власник даних апріорі не знає, до якого класу еквівалентності збігається VIP, можна припустити гірший сценарій. За гіршого сценарію, у зловмисника буде VIP, який відповідає найменшому класу еквівалентності в ζ, який у k-анонімованому наборі даних матиме розмір щонайменше k. Отже, ймовірність повторної ідентифікації буде не більше $1/k$.

Отже, за сценарієм повторної ідентифікації прокурора k-анонімність може гарантувати, що ризик повторної ідентифікації приблизно дорівнює пороговому ризику, як передбачається зберігачем даних. Однак це не так за сценарієм повторної ідентифікації журналістів.

Нехай існує велика кінцева популяція пацієнтів, що позначається множиною U. Тоді $s' \subseteq s \subseteq U$. Зловмисник мав би доступ до ідентифікаційної бази даних про населення U і використовував цю ідентифікаційну базу даних для порівняння з пацієнтами в ζ. База даних ідентифікації позначається $\mathcal{Z}$, а записи $\mathcal{Z}$ мають індивідуальне відображення для осіб у U.

На прикладі рисунку 3.1, існує набір даних про 14 осіб, який потрібно розкрити. Після 2-анонімізації залишилось лише 11 записів, оскільки три потрібно було придушити.

Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
|s| 14	Fred Thompson	Male	1967	-ve

2-Anonymization

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
|s'| 14	Male	1960-1969	-ve

Disclosed (k-Anonymized) Database (ζ)

Matching

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
|U| 31	Natasha Markhov	Female	1941

Рисунок 3.1 – 2-анонімізація

Зловмисник виконує зіставлення, в той час як власник даних виконує 2-анонімізацію.

Зловмисник отримує ідентифікаційну базу даних із 31 записом. Це Z база даних. Потім зловмисник намагається повторно ідентифікувати, зіставляючи довільний запис із записами ζ року народження та статі. У цьому прикладі, як тільки довільну особу ідентифікують, зловмисник отримає результат тесту цієї особи.

Дискретна змінна, сформована шляхом перехресної класифікації всіх значень на квазіідентифікаторах ζ, може приймати J різні значення. Нехай $X_{\zeta,i}$ позначає значення запису i в ζ наборі даних. Наприклад, якщо є два квазіідентифікатори, такі як стать та вік, то можливо отримати $X_{\zeta,1} = "MALE, 50", X_{\zeta,1} = "MALE, 53"$ тощо. Аналогічно нехай $X_{Z,i}$ позначають значення запису i в Z наборі даних.

Розміри різних класів еквівалентності задаються як:

$$f_j = \sum_{i \in \acute{s}} I(X_{\zeta,i} = j),\ j = 1,\dots,j \qquad (3.1.1)$$

де f_j – розмір ζ класу еквівалентності та $I(\cdot)$ є показниковою функцією. Розмір класу еквівалентності задається як:

$$F_j = \sum_{i \in U} I(X_{Z,i} = j),\ j = 1, \dots, j \qquad (3.1.2)$$

де F_j – це розмір класу еквівалентності в Z.

За сценарієм повторної ідентифікації журналіста ймовірність повторної ідентифікації запису у класі еквівалентності j становить $1/F_j$ [52-53]. Однак розумний зловмисник зосередився на записах у класах еквівалентності з найбільшою ймовірністю повторної ідентифікації. Класи еквівалентності з найменшим значенням для F_j мають найбільшу ймовірність бути повторно ідентифікованими, і тому припускається, що розумний зловмисник зосередиться на них. Тоді ймовірність повторної ідентифікації довільної особи розумним зловмисником визначається:

$$\theta_{max} = 1 \Big/ \min_j(F_j) \qquad (3.1.3)$$

Якщо розглянути рисунок 3.1, знову ж таки, у 2-анонімованому файлі вік був перетворений на 10-річні інтервали. У цьому прикладі можливо спостерігати $\theta_{\max} = 0.25$, оскільки найменший клас еквівалентності у Z має 4 записи (ідентифікаційні номери 1, 4, 12 та 27). З 2-анонімізацією зберігач даних використовував пороговий ризик 0.5, але фактичний ризик повторної ідентифікації $\theta_{\max}$ становить половину від цього. Цей консерватизм може здатися гарною ідеєю, але насправді він має великий негативний вплив на якість даних. У цьому прикладі 2-анонімізація призвела до перетворення віку на десятирічні інтервали та придушення більше п’ятої частини записів, які потрібно було розкрити (3 із 14 записів потрібно було придушити). За більшістю стандартів втрата п’ятої частини набору даних через анонімізацію вважається великою втратою інформації.

Тепер розглянемо інший підхід: k-map. За допомогою k-map передбачається, що власник даних може k-анонімізувати саму базу даних ідентифікації (і, отже,

безпосередньо керувати F_j значеннями). Скажімо, база даних ідентифікації є k-анонімізованою для створення Z'. Властивість k-map стверджує, що кожен запис у ζ схожий на принаймні k записів у Z'[54]. Це проілюстровано в рисунку 3.2. Тут власник даних 2-анонімізує ідентифікаційну базу даних безпосередньо, а потім реалізує перетворення набору даних, що підлягає розкриттю. У цьому прикладі, $\theta_{\max} = 0.5$, оскільки найменші класи еквівалентності Z' для записів від 1 до 14 мають два записи. Крім того, ступінь втрати інформації значно зменшується: у розкритому наборіданих відсутні записи, а вік перетворюється на інтервали 5 років, а не 10-річні інтервали. Використовуючи властивість k-map, видно, що фактичний ризик повторної ідентифікації є тим, що передбачав власник даних, і одночасно зменшилась втрата інформації.

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Anonymization

Anonymized Identification Database (Z')

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
~~27~~	~~Almond Zipf~~	~~Male~~	~~1954~~
1	John Smith	Male	1955-1959
4	Hercules Green	Male	1955-1959
12	Douglas Henry	Male	1955-1959
2	Alan Smith	Male	1960-1964
15	Joe Doe	Male	1960-1964
14	Fred Thompson	Male	1965-1969
22	Deitmar Plank	Male	1965-1969
26	Anderson Heft	Male	1965-1969
16	Mark Fractus	Male	1970-1974
20	William Cooper	Male	1970-1974
23	Anderson Hoyt	Male	1970-1974
9	Bill Nash	Male	1975-1979
10	Albert Blackwell	Male	1975-1979
5	Alicia Freds	Female	1940-1944
31	Natasha Markhov	Female	1940-1944
~~28~~	~~Alex Long~~	~~Female~~	~~1952~~
3	Alice Brown	Female	1955-1959
29	Britney Goldman	Female	1955-1959
11	Beverly McCulsky	Female	1960-1964
18	Jane Doe	Female	1960-1964
7	Marie Kirkpatrick	Female	1965-1969
19	Nina Brown	Female	1965-1969
21	Kathy Last	Female	1965-1969
~~24~~	~~Alexandra Knight~~	~~Female~~	~~1974~~
6	Gill Stringer	Female	1975-1979
13	Freda Shields	Female	1975-1979
17	Lillian Barley	Female	1975-1979
25	Helene Arnold	Female	1975-1979
8	Leslie Hall	Female	1985-1989
30	Lisa Marie	Female	1985-1989

2-Map

Disclosed Database (ς)

	QUASI-IDENTIFIERS		
ID	Gender	Year of Birth	Test Result
1	Male	1955-1959	+ve
2	Male	1960-1964	-ve
3	Female	1955-1959	-ve
4	Male	1955-1959	-ve
5	Female	1940-1944	-ve
6	Female	1975-1979	-ve
7	Female	1965-1969	+ve
8	Female	1985-1989	-ve
9	Male	1975-1979	-ve
10	Male	1975-1979	-ve
11	Female	1960-1964	-ve
12	Male	1955-1959	+ve
13	Female	1975-1979	-ve
14	Male	1965-1969	-ve

Рисунок 3.2 – Анонімізація k-map

На практиці модель k-map не використовується, оскільки передбачається, що власник даних не має доступу до ідентифікаційної бази даних, але зловмисник має. Отже, замість цього використовується модель k-анонімності.

Є вагомі причини, чому власник даних не мав би ідентифікаційної бази даних. Часто отримати базу даних населення досить дорого. Також, цілком ймовірно, що власнику даних доведеться захищати кілька груп населення, отже, примножуючи витрати. Наприклад, побудова єдиної специфічної бази даних з використанням напівпублічних реєстрів, які можна використовувати для повторної

ідентифікації атак у Канаді, коштує від 150 000 до 188 000 доларів [55]. Комерційні бази даних можуть бути порівняно дорогими. Крім того, зловмисник може вчиняти незаконні дії, щоб отримати доступ до реєстрів населення. Наприклад, законодавство про приватне життя та Закон про вибори в Канаді обмежують використання списків виборців для участі та підтримки виборчої діяльності. Відомий принаймні один випадок, коли благодійна організація, яка нібито підтримує терористичну групу, змогла отримати канадські списки виборців для збору коштів [56-58]. Законний власник даних не буде брати участь у таких діях.

3.2 Практична реалізація

Було проведено імітаційне дослідження, щоб оцінити фактичну ймовірність повторної ідентифікації для k-анонімізованих наборів даних за сценарієм повторної ідентифікації журналіста та повторної ідентифікації прокурора. Використано значення k = 5.

Для моделювання використовували 2 випадкових набора даних. Перший (Рис.3.3) – це список із 10000 записів. Код для створення цього списку наведено в додатку А. Квазіідентифікаторами були: порядковий номер, вік, стать, країна та раса. Для другого списка із 5962 записів – стать, вік, раса, сімейний стан, освіта, країна, робочий клас, ремесло та зарплатній рівень.

Анонімізація даних та розрахунок ризиків здійснився за допомогою застосунка в загальному доступі ARX Data Anonymization Tool [59].

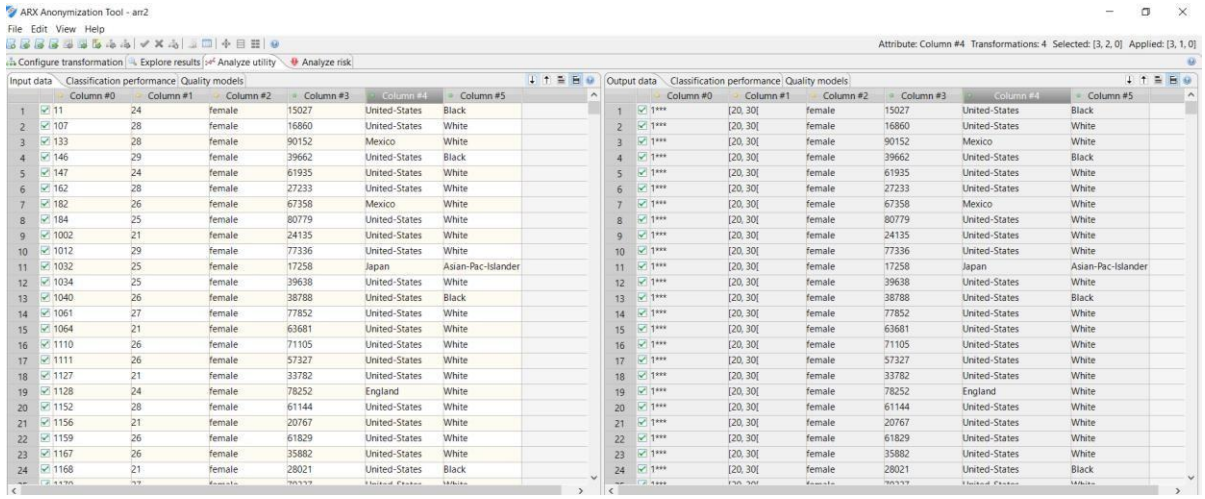

Рисунок 2.3 – анонімізація 1 набору даних

Базовий підхід оцінки ризику (1/k), який є поточною практикою, є досить неефективним і демонструє широкий розрив між фактичним ризиком при k = 5. Цей розрив досить помітний для малих фракцій вибірки та зникає для великих фракцій вибірки. При вищих частках вибірки немає різниці між базовим підходом та іншими з точки зору фактичного ризику (всі вони зводяться до 0,2).

Результати першого набору даних наведено в таблиці 3.1.

Таблиця 1.1

Вимірювання	Значення [%]
Найнижчий ризик прокурора	0,9 %
Записи, на які впливає найнижчий ризик	1,1 %
Середній ризик прокурора	1,1 %
Найвищий ризик прокурора	1,4 %
Записи, на які впливає найвищий ризик	0,7 %
Розрахунковий прокурорський ризик;	1,4 %
Оцінений журналістський ризик	1,4 %

Результати другого набору даних наведено в таблиці 3.2.

Таблица 3.2

Вимірювання	Значення [%]
Найнижчий ризик прокурора	5 %
Записи, на які впливає найнижчий ризик	0,4 %
Середній ризик прокурора	18,2 %
Найвищий ризик прокурора	20 %
Записи, на які впливає найвищий ризик	60,8 %
Розрахунковий прокурорський ризик;	20 %
Оцінений журналістський ризик	20 %

Придушення призводить до викидання даних, які було дорого зібрано, і потенційно призведе до значної втрати статистичної потужності при будь-якому наступному аналізі. Крім того, якщо придушення записів не є абсолютно випадковим, результати аналізу будуть упереджені. Якщо взяти простий приклад одного квазіідентифікатора, записи будуть придушені для рідкісних та екстремальних значень цієї змінної. Тому, за визначенням, схема придушення не буде повністю випадковою.

Деякі алгоритми k-анонімізації пригнічують окремі комірки, а не повні записи. На практиці це може мати не такий позитивний вплив на здатність аналізу даних, як можна було б сподіватися. Одним із загальноприйнятих підходів до боротьби з пригніченими клітинами є повний аналіз випадків (CCA), завдяки якому в аналіз включаються лише записи без пригнічених значень. Видалення повних записів з будь-якими пригніченими значеннями є підходом за замовчуванням у більшості статистичних пакетів і є звичайною практикою в епідеміологічному аналізі. Відомо, що CCA може призвести до відкидання великих пропорцій набору даних. Наприклад, якщо лише 2% значень випадково відсутні в кожній з 10 змінних, можна втратити в середньому 18,3% спостережень за допомогою CCA, а з 5

змінними, у яких випадково відсутні 10% значень, 41% спостереження будуть втрачені в середньому за допомогою CCA. Іншим популярним підходом є доступний аналіз випадків (ACA), за допомогою якого використовуються записи з повними значеннями змінних, що використовуються в певному аналізі. Наприклад, при побудові матриці кореляції використовуються різні записи для кожної пари змінних залежно від наявності обох значень. Однак це може призвести до безглуздих результатів. Як CCA, так і ACA є доречними лише за твердого припущення, що придушення є абсолютно випадковим. Отже, повний запис або придушення окремих клітин негативно впливають на якість набору даних.

Спосіб застосування k-анонімності залежить від сценарію повторної ідентифікації, від якого захищається. Для захисту від сценарію повторної ідентифікації прокурора слід використовувати k-анонімність. Якщо сценарій прокурора не застосовується, тоді k-анонімність не рекомендується. Якщо обидва сценарії правдоподібні, тоді слід використовувати k-анонімність, оскільки це є найбільш захисним. Тому важливо прийняти рішення про те, чи застосовний сценарій прокурора.

Зловмисник застосовуватиме сценарій повторної ідентифікації прокурора лише в тому випадку, якщо він / вона впевнений у тому, що VIP-особа має протокол ζ. Існує три способи, якими зловмисник може мати таку впевненість:

- розкритий набір даних представляє всю сукупність (наприклад, реєстр сукупності) або має велику частку вибірки. Якщо розкривається ціла популяція, то зловмисник мав би бути впевнений, що VIP-особа є у розкритому наборі даних. Крім того, велика частка вибірки означає, що VIP, швидше за все, буде в розкритому наборі даних.
- Якщо це легко можна визначити, хто є у розкритому зразку. Наприклад, вибіркою може бути набір даних опитування співбесід, проведеного в компанії, і загальновідомо, хто брав участь у цих співбесідах, оскільки учасники пропустили півдня роботи. У такому випадку це відомо як компанії, так і внутрішньому зловмиснику, який знаходиться у розкритому наборі даних.

- Особи у розкритому наборі даних самостійно виявляють, що вони є частиною вибірки. Наприклад, випробовувані в клінічних випробуваннях зазвичай повідомляють родині, друзям і навіть знайомим про те, що вони беруть участь у випробуванні. Хтось із знайомих може спробувати повторно ідентифікувати одну з цих предметів, що саморозкриваються. Однак не завжди люди знають, що їхні дані містяться в наборі даних. Наприклад, для досліджень, коли було відмовлено від згоди або якщо пацієнти надають широкий дозвіл на використання своїх даних або зразків тканин для дослідження, пацієнти можуть не знати, що їх дані містяться в певному наборі даних, не надаючи можливості для саморозкриття їх включення.

Якщо застосовується будь-яка з перерахованих вище умов, тоді необхідний захист від сценарію прокурора.

Існують інші підходи, запропоновані для досягнення k-анонімності, які не було розглянуто, наприклад, локальне перекодування. При локальному перекодуванні спостереження можуть мати різні інтервали відповіді, що перекриваються. Наприклад, одне спостереження може мати вік 27 років, закодований до інтервалу 20–29, а інше спостереження може мати вік 27 років, закодований до інтервалу 25–35. Це робить будь-який аналіз даних k-анонімізованого набору даних більш складним, ніж наявність однакових інтервалів перекодування для всіх спостережень, і виключає використання загальноприйнятих методів статистичного моделювання. Зараз реалізація k-анонімності використовувала замість цього глобальне перекодування, що гарантувало однакові інтервали відповідей у всіх спостереженнях.

Висновки за розділом 3

Продемонстровано два сценарії повторної ідентифікації, проти яких була розроблена k-анонімність, відомі як сценарії прокуратора та журналістів.

Зроблено висновок, що базова модель k-анонімності, яка представляє сучасну практику, буде добре працювати для захисту від сценарію повторної ідентифікації прокурора. Однак емпіричні результати показують, що початкова модель k-

анонімності дуже консервативна з точки зору ризику повторної ідентифікації за сценарієм повторної ідентифікації журналіста. Цей консерватизм призводить до значних втрат інформації. Втрата інформації посилюється для невеликих фракцій вибірки.

Тому важливо точно розуміти типи атак повторної ідентифікації, які можуть бути запущені на наборі даних, і різні способи правильної анонімізації даних перед їх розкриттям.

Методи анонімізації призводять до спотворень даних. Надмірна анонімізація може знизити якість даних, роблячи їх непридатними для певного аналізу, і, можливо, призвести до неправильних або упереджених результатів. Тому важливо збалансувати обсяг анонімізації, що проводиться, та обсяг втрати інформації.

У цьому розділі була розглянута k-анонімность, яка є популярним підходом до захисту конфіденційності. Розглянуто два сценарії повторної ідентифікації, від яких покликана захищати k-анонімність. Для одного зі сценаріїв показано, що реальний ризик повторної ідентифікації за базової k-анонімності набагато нижчий за пороговий ризик, який бере на себе власник даних, і що це призводить до надмірного обсягу втрати інформації, особливо при малих частках вибірки. Потім оцінено три альтернативні підходи і виявили, що один із них послідовно гарантує, що ризик повторної ідентифікації досить близький до фактичного ризику і завжди має менші втрати інформації, ніж базовий підхід.

ВИСНОВКИ

Методи анонімізації призводять до спотворень даних. Надмірна анонімізація може знизити якість даних, роблячи їх непридатними для певного аналізу, і, можливо, призвести до неправильних або упереджених результатів. Тому важливо збалансувати обсяг анонімізації, що проводиться, та обсяг втрати інформації.

K-анонімізація є популярним підходом до захисту конфіденційності. Було продемонстровано два сценарії повторної ідентифікації, від яких покликана захищати k-анонімність. Для одного зі сценаріїв показано, що реальний ризик повторної ідентифікації за базової k-анонімності набагато нижчий за пороговий ризик, який бере на себе власник даних, і що це призводить до надмірного обсягу втрати інформації, особливо при малих частках вибірки.

Для захисту даних власникам рекомендується визначати, які сценарії повторної ідентифікації застосовуються в кожному конкретному випадку, та анонімізувати дані перед розкриттям, використовуючи базову модель k-анонімності або модифіковану модель k-анонімності відповідно. Для захисту від сценарію повторної ідентифікації прокурора слід використовувати k-анонімність. Якщо сценарій прокурора не застосовується, тоді k-анонімність не рекомендується, і замість цього слід використовувати k-map. Якщо обидва сценарії можливі, слід використовувати k-анонімність, оскільки цей підхід є найбільш захисним.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Statement by the EDPB Chair on the processing of personal data in the context of the COVID-19 outbreak. European Data Protection Board. 16 March 2020. Access: https://edpb.europa.eu/news/news/2020/statement-edpb-chair-processing-personal-data-context-covid-19-outbreak_en

2. FAQs – EU-U.S. Privacy Shield Program Update. Privacy Shield Framework. Access: https://www.privacyshield.gov/article?id=EU-U-S-Privacy-Shield-Program-Update

3. The Court of Justice invalidates Decision 2016/1250 on the adequacy of the Potection provided by the EU-US Data Protection Shield / Court of Justice of the European Union / PRESS RELEASE №91/20 Luxembourg, 16 July 2020. Access: https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091en.pdf

4. Presidential Policy Directive 28 (PPD-28) Signals Intelligence Activities. Homeland Security. Access: https://www.dhs.gov/publication/presidential-policy-directive-28-ppd-28-signals-intelligence-activities

5. Report to the President on the Implementation of Presidential Policy Directive 28: Signals Intelligence Activies. PRIVACY & CIVIL LIBERTIES OVERSIGHT BOARD. January 17, 2014. Access: https://documents.pclob.gov/prod/Documents/OversightReport/16f31ea4-3536-43d6-ba51-b19f99c86589/PPD-28%20Report%20(for%20FOIA%20Release).pdf

6. The CLOUD Act. Electronic Privacy Information Center. Access: https://epic.org/privacy/cloud-act/

7. ALAN CHARLES RAUL. An Early Recap of Privacy in 2020: A US Perspective. 29 September, 2020. Access: https://datamatters.sidley.com/an-early-recap-of-privacy-in-2020-a-us-perspective

8. FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook. FEDERAL TRADE COMMISSION. July 24, 2019. Access:

https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions

9. Alan Charles Raul. The Privacy, Data Protection and Cybersecurity Law Review: Global Overview. 14 October 2020. Access: https://thelawreviews.co.uk/title/the-privacy-data-protection-and-cybersecurity-law-review/global-overview

10. Mike Seery, Anthony Perez. SEC Increases Focus on Digital Assets. April 05, 2021. Access: https://www.acaglobal.com/insights/sec-increases-focus-digital-assets

11. Fábio Lacaz. BRAZIL'S DATA PROTECTION LAW: A BRIEF OVERVIEW. 02.03.2021. Access: https://inplp.com/latest-news/article/brazils-data-protection-law-a-brief-overview/

12. Renata Neeser. Is the Brazilian Data Protection Law (LGPD) Really Taking Off? June 8, 2021. Access: https://www.jdsupra.com/legalnews/is-the-brazilian-data-protection-law-1094165/

13. TITLE 1.81.5. California Consumer Privacy Act of 2018 [1798.100 - 1798.199.100] Title 1.81.5 added by Stats. 2018, Ch. 55, Sec. 3.

14. Brian H. Lam. The Next Act for the Architect of the California Consumer Privacy Act: The California Privacy Rights Act. January 30, 2020. Access: https://www.natlawreview.com/article/next-act-architect-california-consumer-privacy-act-california-privacy-rights-act

15. ЗАКОН УКРАЇНИ «Про захист персональних даних» від 23.04.2021. Режим доступу до документа: https://zakon.rada.gov.ua/laws/show/2297-17#Text

16. Окремі питання законодавства України та міжнародних стандартів у сфері захисту персональних даних при обробці персональних даних, які містяться в архівних документах. 10.12.2020. Веб-сайт. URL: https://www.ombudsman.gov.ua/ua/all-news/pr/okrem%D1%96-pitannya-zakonodavstva-ukra%D1%97ni-ta-m%D1%96zhnarodnix-standart%D1%96v-u-sfer%D1%96-zaxistu-personalnix-danix-pri-obrobcz%D1%96-personalnix-danix,-yak%D1%96-m%D1%96styatsya-v-arx%D1%96vnix-dokumentax/

17. РЕГЛАМЕНТ ЄВРОПЕЙСЬКОГО ПАРЛАМЕНТУ І РАДИ (ЄС) 2016/679 від 27 квітня 2016 року про захист фізичних осіб у зв'язку з опрацюванням

персональних даних і про вільний рух таких даних, та про скасування Директиви 95/46/ЄС (Загальний регламент про захист даних). 984_008-16 від 27.04.2016. Режим доступу до документа: https://zakon.rada.gov.ua/laws/show/984_008-16#Text

18. General Data Protection Regulation. [Electronic resource] — URL: https://gdpr-info.eu/

19. Definitions. Art. 4 GDPR. [Electronic resource] — Access: https://gdpr-info.eu/art-4-gdpr/

20. PRIVACY & SECURITY REPORT, RSA DATA. FIELDWORK OCCURRED 15 DECEMBER 2017 – 3 JANUARY, 2018. Access: https://www.rsa.com/content/dam/en/e-book/rsa-data-privacy-report.pdf

21. Monpi Neog Lobo, Responsibilities of a Controller, Processor & Data Protection Officer. DECEMBER 3, 2020. [Electronic resource] — Access: https://www.wsiworld.com/blog/responsibilities-of-a-controller-processor-and-data-protection-officer-according-to-gdpr

22. GDPR Enforcement Tracker: [Electronic resource] — URL: https://www.enforcementtracker.com/

23. DLA Piper: [Electronic resource] — URL: https://www.dlapiper.com/en/ukraine/

24. Petter Bjerke, Marlene Winther Plas, €114 million in fines have been imposed by European authorities under GDPR. DLA Piper Norway, 20 Jan 2020. [Electronic resource] — Access: https://norway.dlapiper.com/en/news/eu114-million-fines-have-been-imposed-european-authorities-under-gdpr

25. Blog: GDPR – sorting the fact from the fiction, ICO. 9 August 2017 [Electronic resource] — Access: https://ico.org.uk/about-the-ico/news-and-events/blog-gdpr-sorting-the-fact-from-the-fiction/

26. Matt Fisher, From the Desk of Matt Fisher – ICYMI, May 3, 2021. [Electronic resource] — Access: https://www.healthcarenowradio.com/from-the-desk-of-matt-fisher-icymi5321/

27. ДСТУ ISO/IEC 27001:2015 Інформаційні технології. Методи захисту системи управління інформаційною безпекою. Вимоги (ISO/IEC 27001:2013; Cor 1:2014, IDT)

28. Стандарт ISO/IEC 27701:2019. Методи і засоби забезпечення безпеки // Розширення до ISO / IEC 27001 та ISO / IEC 27002 щодо управління інформацією про конфіденційність.

29. Security of processing. Art. 32 GDPR. Access: https://gdpr-info.eu/art-32-gdpr/

30. Звіт компанії «Verizon» за 2019 рік : 2019 Data Breach Investigations Report. Режим доступу до документа: https://enterprise.verizon.com/resources/reports/2019/2019-data-breach-investigations-report.pdf

31. Методи і способи захисту інформації [Електронний ресурс]. — Режим доступу: https://pidru4niki.com/1801051351329/ekonomika/metodi_sposobi_zahistu_informatsiyi

32. Data masking and hiding [Electronic resource]. — Access: https://docs.apigee.com/api-platform/security/data-masking

33. Маскування даних [Електронний ресурс]. — Режим доступу: https://en.wikipedia.org/wiki/Data_masking

34. HMAC [Електронний ресурс]. — Access: https://uk.wikipedia.org/wiki/HMAC

35. Apache Hadoop [Electronic resource]. — Access: https://uk.wikipedia.org/wiki/Apache_Hadoop

36. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalisation and suppressionSRI International; 1998.

37. Samarati P. Protecting respondents' identities in microdata release IEEE Transactions on Knowledge and Data Engineering 2001; 13 (6) : 1010-1027.

38. Sweeney L. k-anonymity: a model for protecting privacy International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 2002; 10 (5) : 557-570.

39. Ciriani V, De Capitani di Vimercati SSF, Samarati P. k-AnonymitySpringer: Secure Data Management in Decentralized Systems; 2007.

40. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002; 10 (5) : 571-588.

41. Bayardo R, Agrawal R. Data Privacy through Optimal k-Anonymization Proceedings of the 21st International Conference on Data Engineering 2005 : 217-228.

42. Iyengar V. Transforming data to satisfy privacy constraints Proceedings of the ACM SIGKDD Int Conf Data Mining Knowledge Discov 2002 : 279-288.

43. Purdham K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records Env Planning 2007; 39 : 1101-1118.

44. Fefferman N, O'Neil E, Naumova E. Confidentiality and confidence: Is data aggregation a means to achieve both J Public Health Pol 2005;16:430-449.

45. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance J Am Med Inform Assoc 2002; 9(6) : s115-s119.

46. Kamlet MS, Klepper S, Frank R. Mixing micro and macro data: Statistical issues and implication for data collection and reporting Proceedings of the 1985 Public Health Conference on Records and Statistics 1985.

47. Clause S, Triller D, Bornhorst C, Hamilton R, Cosler L. Conforming to HIPAA regulations and compilation of research data Am J Health-Sys Pharm 2004; 61 : 1025-1031.

48. Abrahamowicz M, du Berger R, Krewski D, Burnett R, Bartlett G, Tamblyn R, Leffondre K. Bias due to aggregation of individual covariates in the Cox regression model Am J Epidemiol 2004; 160 (7) : 696-706.

49. Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D, Walford N. The case for samples of anonymized records from the 1991 census J Royal Statist Soc, Ser A (Statistics in Society) 1991; 154(2) : 305-340.

50. Elliot M, Dale A. Scenarios of attack: the data intruders perspective on statistical disclosure risk Netherlands Official Statistics 1999; 14 (Spring):6-10.

51. De Waal A, Willenborg L. A view on statistical disclosure control for microdata Surv Methodol 1996; 22(1) : 95-103.

52. Willenborg L, de Waal T. Elements of Statistical Disclosure ControlSpringer-Verlag; 2001.

53. Benedetti R, Franconi L. Statistical and technological solutions for controlled data dissemination, Proceedings of New Techniques and Technologies for Statistics (vol. 1), 1998; 225–32.

54. Sweeney L. Computational disclosure control: A primer on data privacy protection Massachusetts Institute of Technology; 2001.

55. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information J Med Internet Res 2006;8(4):e28.

56. Bell S. Alleged LTTE front had voter listsNational Post; 2006. July 22.

57. Bell S. Privacy chief probes how group got voter listsNational Post; 2006. July 25.

58. Freeze C, Clark C. Voters lists «most disturbing» items seized in Tamil raids, documents say. Globe and Mail. May 7, 2008.

59. ARX Data Anonymization Tool. URL: https://arx.deidentifier.org/

ДОДАТКИ

ДОДАТОК А

```
package javaapplication9;

import com.spire.xls.ExcelVersion;

import com.spire.xls.Workbook;

import com.spire.xls.Worksheet;

import java.util.Random;

import java.util.Arrays;

public class JavaApplication9 {

    public static void main(String[] args) {

        Random random = new Random();

        String[] sex = new String[]{"male", "female"};

        String[] country = new String[]{"United-States","Scotland", "South", "Mexico", "Japan", "Philippines", "Puerto-Rico", "Cambodia", "Canada", "China", "Columbia", "Cuba", "Dominican-Rapublic", "Ecuador", "EI-Salvador", "England", "France", "Germany", "Greece", "Guatemala", "Haiti", "Honduras", "Hong", "Hungary", "India", "Iran", "Ireland", "Italy", "Jamaica", "Laos", "Nigaragua", "Outlying-US(Guam-USVI-etc)", "Peru", "Poland", "Portugal", "Taiwan", "Thailand", "Trinadad&Tobago", "Vietnam", "Yugoslavia"};

        String[] race = new String[]{"Amer-Indian-Eskimo", "Asian-Pac-Island", "Black", "White", "Other"};

        String[][] myArray = new String[10000][6];

        for (int i=0; i < 10000; i++){
```

```
            myArray[i][0] = String.valueOf(i);
            myArray[i][1] = String.valueOf(20 + random.nextInt(60));
            myArray[i][2] = sex[random.nextInt(sex.length)];
            myArray[i][3] = String.valueOf(random.nextInt(90000) + 9999);
            myArray[i][4] = country[random.nextInt(country.length)];
            myArray[i][5] = race[random.nextInt(race.length)];
        }
        //Create a Workbook instance
        Workbook wb = new Workbook();
            //Get the first worksheet
        Worksheet sheet = wb.getWorksheets().get(0);
        //Write the array to the worksheet from the specified cell
        sheet.insertArray(myArray, 1, 1);
        //Save the file
        wb.saveToFile("InsertArrays2.xlsx", ExcelVersion.Version2016);
    }
}
```