ПРОГРАММИРОВАНИЕ, 2023, № 4, с. 58–74

АНАЛИЗ ДАННЫХ

УДК 004.622

ПРИМЕНЕНИЕ ИМИТАЦИОННОГО КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ К ЗАДАЧЕ ОБЕЗЛИЧИВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ. ОЦЕНКА СОСТОЯНИЯ И ОСНОВНЫЕ ПОЛОЖЕНИЯ

© 2023 г. А. В. Борисов[a],* (ORCID: 0000-0002-3124-2147),
А. В. Босов[a], (ORCID: 0000-0001-7163-341X),**
А. В. Иванов[a],* (ORCID: 0000-0001-7811-7645)**

[a]Федеральный исследовательский центр "Информатика и управление" РАН
119333, Москва, ул. Вавилова, д. 44, кор. 2, Россия

**E-mail: aborisov@ipiran.ru*

***E-mail: avbosov@ipiran.ru*

****E-mail: aivanov@ipiran.ru*

Поступила в редакцию 20.01.2023 г.
После доработки 25.02.2023 г.
Принята к публикации 02.03.2023 г.

В статье представлена первая часть исследования по проблеме автоматизированной обработки персональных данных с целью их обезличивания и анализа. Эта часть носит обзорный характер и ставит целью анализ состояния исследований в данной области и систематизацию имеющихся результатов. Представлены результаты анализа широкого круга вопросов обезличивания, сформировавшие системное понимание состояния исследований и обосновавшие выбор направления для дальнейшего изучения. Вначале сформулированы определения основных терминов и понятий, используемых в связи с обезличиванием персональных данных, в т.ч. в увязке с законодательством РФ. Направления исследований сгруппированы по четырем разделам: методы обезличивания, проблемы реализации, приложения обработки обезличенных данных, вопросы деобезличивания. По каждой из групп методов обезличивания – рандомизации, группировке, распределению данных и контролю приложений – даны описания основных алгоритмов, проанализированы их достоинства и недостатки. Проблемы реализации затрагивают такие понятия как полезность обезличенных данных, ограничения применимости универсальных алгоритмов и надежность в отношении сохранения анонимности субъектов персональных данных. В числе прикладных решений, сформировавших востребованность обработки обезличенных данных, обсуждаются медицинские, биологические, генетические исследования и охрана правопорядка. В заключительной части упоминаются наиболее резонансные факты деобезличивания и дается небольшой обзор прессы.

DOI: 10.31857/S0132347423040040, **EDN:** RDBXOS

1. МЕТОДЫ И АЛГОРИТМЫ ОБЕЗЛИЧИВАНИЯ

1.1. Основные термины и определения

К настоящему времени для исследований, проводимых научным сообществом в области обезличивания персональных данных (ПД), не сформирована единая общепринятая терминология. Во многом это объясняется различными правовыми режимами ПД в разных государствах. Основные понятия, использованные в данной работе, опираются на законодательную базу РФ и ориентированы на решение информационных и алгоритмических вопросов, а не на правовое регулирование. Принципиальное понимание "персональных данных" дается Федеральным законом "О персональных данных" от 27.07.2006 № 152-ФЗ. Для технического описания методов и алгоритмов далее используются следующие понятия:

• необезличенные ПД – множество ПД в исходном виде,

• персональная идентификационная информация (персональные идентификаторы) – любые цифровые идентификаторы человека (субъекта ПД), указывающие на него напрямую,

• обезличенные ПД – множество ПД, полученное в результате обработки необезличенных ПД в соответствии с некоторым алгоритмом обезличивания с целью препятствования деобезличиванию ПД,

• деобезличивание ПД – процесс обработки обезличенных ПД с целью (возможно, частично-

го) восстановления соответствующего набора необезличенных ПД и/или установления любого соответствия ПД и субъекта ПД (нарушения анонимности),

• чувствительная информация – ПД, которые не являются персональными идентификаторами, но представляют угрозу нарушения анонимности,

• нечувствительная информация – ПД, которые не представляют сколь-либо существенной угрозы нарушения анонимности и не могут быть использованы для установления любого соответствия ПД и субъекта ПД,

• несущественная информация – ПД, которые не содержат семантически значимых данных в отношении заявленной цели обезличивания,

• квазиидентификаторы – это любые элементы ПД, которые позволяют исключить из имеющегося набора обезличенных ПД данные, не относящиеся к конкретному субъекту ПД,

• псевдоданные (синтетические ПД) – это данные, полученные в результате компьютерного моделирования, реализующего модель, воспроизводящую статистические характеристики соответствующего набора необезличенных данных,

• уровень анонимности обезличенных ПД – оценка сложности проведения деобезличивания субъектов ПД,

• уровень полезности обезличенных ПД – свойство множества обезличенных ПД, заключающееся в возможности решения прикладных задач с использованием обезличенных ПД.

1.2. Направления исследований в области обезличивания ПД

Вопросы и задачи, затрагивающие тематику ПД, уже довольно длительное время привлекают исследовательский интерес научного сообщества. Пристальное внимание к области последние двадцать лет наблюдается в РФ: периодические сообщения об очередных утечках ПД привлекают внимание общества в целом и инспирируют весьма ответственные и актуальные правовые шаги госорганов. Но наше внимание направлено не на социальные и правовые аспекты, а на вопросы, составляющие исследовательский интерес. И здесь надо констатировать, что внимание российского научного сообщества к области минимально. Вместе с тем проблематика алгоритмов обезличивания и деобезличивания очень богата и предоставляет хорошие возможности для приложения исследовательских усилий.

К настоящему времени опубликовано огромное число работ, с той или иной степенью математической строгости представляющих соответствующие алгоритмы. Представленный далее материал в целом написан в контексте систематического обзора [1] с современными дополнениями [2–6]. Особо надо упомянуть отдельный класс задач, связанных с обезличиванием фото- и видеоматериалов [7–10], специфика которых очевидно связана с медийностью элементов ПД.

Большинство методов обезличивания представляют собой преобразования данных с целью обеспечения требуемого уровня анонимности (далее будет обсуждаться формальное определение этого понятия, пока ограничимся его интуитивно понятным смыслом). Такие методы основаны на уменьшении детализации данных, приводящей к потере эффективности последующих операций поиска и обработки информации, если они направлены на деобезличивание, и потенциально не препятствуют обработке данных, не носящей идентифицирующий характер, например агрегации, статистике и прочим обобщениям. Уменьшение детализации – это плата за обеспечение анонимности. Основная задача разработчиков алгоритмов обезличивания заключается в нахождении компромисса между требуемым уровнем приватности и остаточной полезностью обезличенных данных. Большинство результатов, имеющих адекватное математическое описание, посвящаются следующим вопросам:

• методы обезличивания ПД,

• проблемы реализации,

• целевые приложения обработки обезличенных ПД,

• возможности и варианты деобезличивания.

Наиболее распространенными подходами к обезличиванию являются следующие.

1. Рандомизация: преобразование данных путем добавления в них шума для маскировки значений записей (чаще всего числовых) [11, 12]. Считается, что уровень шума должен быть настолько велик, чтобы сделать невозможным восстановление исходных значений записей по имеющимся зашумленным данным.

2. K-анонимность и L-разнообразие: метод K-анонимности разработан как возможное средство борьбы с косвенным деобезличиванием по информации из открытых баз данных, которое оказывается возможным, если некоторые комбинации элементов ПД позволяют точно идентифицировать всю запись. Метод K-анонимности предполагает уменьшение детализации данных за счет группировки, достигаемой различными приемами: удалением информации, обобщением, маскингом и пр. Уменьшение детализации должно обеспечивать такое преобразование, чтобы каждой обезличенной записи соответствовало не менее K исходных записей [13–16]. Модель L-разнообразия разработана, чтобы нейтрализовать имеющийся у метода K-анонимности недостаток – возможность распознавания субъекта с точностью до группы размером K не обеспечивает защиту чувствительной информации до такого же

уровня, в особенности, когда чувствительная информация неоднородна внутри данной группы. Для обеспечения этого свойства предлагается концепция внутригруппового разнообразия чувствительной информации за счет специальной схемы обезличивания [17, 18].

3. Распределение данных: выполняется разбиение данных на несколько наборов с последующей распределенной работой с этими наборами. Разбиение может быть как горизонтальным (наборы данных разбиваются на некоторые подмножества), так и вертикальным (наборы данных разбиваются на некоторые подмножества атрибутов) [19–21]. Подобное разбиение обеспечивает приложениям доступ только к разделам данных, необходимым этим приложениям для выполнения целевых функций, без возможности обработки всего массива ПД.

4. Контроль работы приложений: нередкой является ситуация, когда анонимность нарушается из-за результатов работы приложений. Это выражается в нахождении новых правил ассоциации в процессе работы с данными, возможности классификации целей поиска при обработке поисковых запросов и пр. Данные обстоятельства вынуждают выполнять модификацию либо данных, либо приложений, работающих с данными. Примеры подобных исследований связаны с сокрытием ассоциативных правил [22], снижением чувствительности классификаторов [23] и контролем выполняемых запросов [24].

1.3. Методы и их свойства

1.3.1. Рандомизация. Традиционное применение находит при проведении анонимных опросов [25, 26], а также распространена на задачу безопасного поиска данных [11]. Метод заключается в следующем. Пусть набор значений одного элемента ПД представляет собой реализацию числовой выборки $X = \{x_1,\ldots,x_N\}$. Обезличивание заключается в добавлении к X случайного шума $Y = \{y_1,\ldots,y_N\}$ – набора независимых одинаково распределенных случайных величин с публично известным распределением $f_Y(y)$. В результате обезличивания для последующего статистического анализа доступна реализация выборки $Z = \{x_1 + y_1,\ldots,x_N + y_N\}$. Соответственно, задача обработки обезличенных данных заключается в восстановлении неизвестного распределения $f_X(y)$ по зашумленным наблюдениям Z. Для решения этой задачи используются различные методы [11, 12], которые можно разделить на эмпирические, методы непараметрического оценивания (построение ядерных оценок плотности распределения), методы параметрического оценивания (метод максимального правдоподобия, его реализация в виде EM-алгоритма). Задача значительно осложняется в случае необходимости восстановления совместного распределения нескольких полей [27], т.к. для достижения постоянной относительной точности восстановления многомерной плотности с ростом размерности требуется экспоненциальный рост длины выборки.

Очевидным преимуществом метода рандомизации является его вычислительная экономичность, а существенным недостатком то, что зашумление малоэффективно для маскировки выбросов [28]. Свойства этого метода позволяют довольно эффективно использовать его в разных аналитических задачах. Так, в [11, 29, 30] они применялись для построения классификаторов, в [31, 32] – для построения безопасных ассоциативных правил. Подобная техника рандомизации также применялась при создании безопасных алгоритмов OLAP [33] и SVD-коллаборативной фильтрации в рекомендательных системах [34].

Для определения уровня анонимности при рандомизации используются различные числовые характеристики. В качестве простейшего показателя [11] в случае аддитивного шума с непрерывным распределением используются пара “длина доверительного интервала – уровень доверительной вероятности”. Например, если после статистических выводов оказалось, что оцениваемый параметр с вероятностью 0.8 принадлежит интервалу (1, 10), то в качестве этого показателя выступает пара (9, 0.8). Такой подход к анонимности имеет серьезный недостаток, состоящий в том, что он не учитывает собственное распределение значений поля. В некоторых случаях неудачный выбор распределения шума может привести даже не к повышению, а к снижению уровня анонимности.

В качестве альтернативных показателей уровня анонимности [12] могут выступать следующие характеристики:

• безусловная дифференциальная энтропия $h(B)$ обезличенных данных B с плотностью распределения вероятностей $f_B(b)$

$$h(B) = -\int_{\Omega_B} f_B(b)\log_2 f_B(b)\,db,$$

• величина $\Pi(B) = 2^{h(B)}$, имеющая смысл длины носителя равномерного распределения с тем же значением $h(B)$ (энтропия $h(B)$ случайной величины B, имеющей произвольную плотность распределения, совпадает с энтропией случайной величины с равномерным распределением на отрезке длиной $\Pi(B)$),

• условная дифференциальная энтропия $h(A|B)$ исходных данных A относительно обезли-

ченных данных B, и соответствующий показатель $\Pi(A|B) = 2^{h(A|B)}$,

$$h(A|B) = -\int_{\Omega_{A,B}} f_{A,B}(a,b)\log_2 f_{A|B}(a|b)\,dadb,$$

• условная потеря анонимности A при условии B: $\mathcal{P}(A|B) = 1 - \dfrac{\Pi(A \mid B)}{\Pi(A)}$,

• совместная информация A и B: $I(A, B) = h(A) - h(A|B)$.

Помимо исследований, сфокусированных на характеризации уровня анонимности, также рассматривалась задача поиска паллиатива между уровнем анонимности и потерей информации при обезличивании [35].

Кроме упомянутой неспособности скрыть аномальные значения в необезличенных ПД, рандомизация имеет и другие слабые стороны, которые можно использовать для деобезличивания. Основным подходом здесь является корреляционный анализ, который позволяет найти "линейную" зависимость между элементами обезличенных данных даже при добавлении аддитивного шума и уменьшить тем самым "размерность" обезличенных ПД, что и составляет угрозу нарушения анонимности. Наиболее распространенными являются алгоритмы главных компонент и спектральная фильтрация [36–38]. Как только корреляционная структура обезличенных данных прикидочно оценена (по выборке большого объема), можно пытаться очистить имеющиеся данные от шума для того, чтобы добиться деобезличивания. Другой вид атак на анонимность заключается в использовании метода максимального правдоподобия для оценки потенциальных шумов в данной выборке обезличенных данных по известному распределению шума.

В связи с рандомизацией нельзя не упомянуть использование вместо аддитивного шума мультипликативного. Техника здесь в основном базируется на аппарате многомерных проекций для сокращения размерности данных [39]. Он приближенно сохраняет "расстояние" между отдельными записями ПД, и поэтому обезличенные данные могут быть использованы для последующего статистического анализа. Этот подход использован в [40, 41] для кластерного анализа обезличенных ПД. Эта же техника применялась в [42] для решения задачи классификации, в [43] – для обеспечения безопасного распределенного анализа данных, в [44] – для обезличивания результатов переписи населения. Еще один вариант применения данного подхода связан с использованием преобразования Фурье, сохраняющего расстояние [45]. Как и использование аддитивных шумов, использование мультипликативных шумов не может рассматриваться как панацея при обезличивании. Опасности имеются в случае, если атакующий обладает некоторым априорным знанием о данных [46].

Наконец, аддитивное или мультипликативное зашумление – не единственные варианты случайного возмущения данных. Еще один возможный способ – перемешивание с целью достижения обезличивания [47]. Явным преимуществом перемешивания является то, что сохраняется разброс выборки (т.е. пара "минимум-максимум"), да и сама выборка тоже. В этом же и самый серьезный недостаток – актуальные, не измененные цифровые ПД могут оказаться квазиидентификаторами и стать источником для деобезличивания всего набора ПД или его части. Кроме того, разные перемешивания разных элементов ПД нарушают имеющиеся зависимости между ними, что уменьшает информативность обезличенных ПД (например, из-за нарушения корреляции между элементами ПД при их независимом перемешивании).

1.3.2. Группировки (K-анонимность, L-разнообразие, t-близость). Добавление шума – эффективный метод обезличивания, который может применяться в процессе сбора данных, т.к. отдельные элементы ПД получаются и зашумляются последовательно, независимо друг от друга. Но в этом состоит и слабость данного метода, поскольку выбросы в исходных данных можно определить только после окончания их сбора. Кроме того, аномальные измерения сложно замаскировать шумами. Таким образом, при обезличивании нужно обеспечивать одинаковую степень приватности всем элементам ПД. Другой слабой чертой рандомизации является наличие обширных открытых данных, с использованием которых оказывается возможным идентифицировать субъекта, к которому данная запись относится. В [28] показано, что совместная обработка открытых данных и многомерных обезличенных данных способна серьезным образом скомпрометировать последние. И особо уязвимыми при этом являются значения элементов ПД (выбросы). Именно эти обстоятельства подтолкнули исследователей к созданию методов обезличивания, основанных на группировке исходных ПД.

Группировки (K-анонимность). Во многих приложениях обезличивание проводилось путем простого исключения из записей ключевых идентификаторов типа имен, номеров документов и пр. Однако некоторые другие атрибуты (выше они определены как квазиидентификаторы) могут быть использованы для точной идентификации обезличенных записей. К таким атрибутам относятся, например, набор "возраст/дата рождения-индекс-пол". Если эти данные одновременно присутствуют в неискаженном виде в обез-

личенных и открытых данных, то по ним легко идентифицировать субъекта ПД. Идея метода *K*-анонимности заключается в снижении детализации квазиидентификаторов с использованием операций обобщения/группировки и исключения. Обобщение означает замену конкретного значения некоторым интервалом, например, замена даты рождения годом или интервалом лет. Исключение подразумевает отсутствие значения в обезличенных данных, например, исключение номеров банковских счетов. Очевидно, что снижение детализации одновременно с обеспечением высокого уровня анонимности, снижает уровень полезности обезличенных данных.

Концепция *K*-анонимности предполагает, что каждый кортеж квазиидентификаторов в наборе обезличенных данных неотличим от кортежей как минимум еще *K* субъектов ПД. Первый алгоритм *K*-анонимности был предложен в [48]. Он основан на некоторой иерархии обобщения предметной области квазиидентификаторов. Однако с практической реализацией метода, несмотря на кажущуюся простоту идеи, возникают проблемы, связанные с его ресурсозатратностью. В [16] представлено утверждение о том, что задача оптимальной *K*-анонимности (т.е. обеспечения набору данных *K*-анонимности с минимальной потерей информации) является NP-сложной, т.е. характеризуется полиномиальным ростом числа операций. Поэтому были разработаны достаточно эффективные с вычислительной точки зрения эвристические алгоритмы. Так, в [49] предложен алгоритм *K*-Optimize, использующий идею упорядочения полей квазиидентификатора. Алгоритм Incognito [15] разработан для *K*-минимальной группировки иерархии обобщения предметной области квазиидентификаторов путем ее агрегации “снизу вверх”. Еще два интересных метода нисходящей детализации и восходящего обобщения реализации *K*-анонимности были предложены в [50] и [51]. В [50] предлагается эвристический алгоритм обработки иерархии “сверху вниз”, в процессе которого некоторые поля детализируются, что увеличивает информацию, но снижает уровень анонимности, а в [51] предложен смежный метод “обобщения снизу вверх”.

Следует заметить, что обобщение и исключение не являются единственными возможными преобразованиями данных, используемых концепцией *K*-анонимности. Например, в [52] предлагается подход, реализующий агрегацию записей в кластеры с использованием средней величины по каждому полю кластера. Аналогичная кластеризация, но с применением факторного анализа представлена в [53]. Очень важна для дальнейшего исследования работа [54], в которой кластеризация дополнена техникой генерации псевдоданных из кластеризованных групп *K*-записей. Такой подход оказался эффективным для решения задач классификации. Кроме того, использование псевдоданных представляет собой дополнительный уровень защиты, т.к. сложно организовать атаки на синтетические данные.

Поскольку процесс обеспечения набору данных *K*-анонимности по своей сути является многошаговым итеративным поиском в пространстве многомерных решений, стандартные эвристические методы поиска также могут быть использованы достаточно эффективно. Так, в [55] применен алгоритм имитации отжига, а в [56] – генетический алгоритм.

Анализ уровня анонимности, обеспечиваемой концепцией *K*-анонимности, был проведен в [57]. Эффективность защиты была проверена с помощью сценария атаки, когда атакующему была доступна некоторая дополнительная информация, например часть оригинальных данных. Для этого сценария предложена методика, позволяющая оценить среднее число записей, которые возможно идентифицировать. Эта методика может быть полезна в тех ситуациях, когда необходимо ответить на вопрос, нужно или нет использовать *K*-анонимность в той или иной конкретной задаче.

Не все субъекты одинаково относятся к уровню анонимности. Это означает, что параметр *K*-анонимности может варьироваться от записи к записи. Алгоритм с переменным параметром *K* был предложен в [58]. Он основан на создании групп различного объема, для которых гарантировано, что размер группы не меньше, чем максимальный параметр *K* в данной группе. А затем, для каждой группы генерируются синтетические данные с характерным распределением.

Другой любопытный подход к обезличиванию предложен в [59], когда субъекту предлагается указать уровень защиты чувствительной информации, например, запретить выдавать определенные атрибуты ПД. Соответствующая техника предполагает, что субъект может указать узел в иерархии обобщения для определения своего уровня анонимности.

Отдельным классом приложений для методов на основе операций обобщения является обезличивание динамических данных, таких как потоковые. Очевидно, что как только блок данных обработан и передан в открытое использование, его нельзя преобразовать заново, например изменить уровень обобщения. В то же время следующий опубликованный блок данных может скомпрометировать предыдущие. Такой пример приведен в [60], где предлагается сделать невозможным создание связей между квазиидентификаторами, присутствующими в обоих блоках. Для этой задачи в [61] предлагается метод управляемого уменьшения степени детализации с целью сохранения уровня *K*-анонимности на всем объединении имеющихся обезличенных данных.

Еще одна группа специфических задач для методов обобщения образуется для атрибутов, имеющих текстовые, бинарные или строковые типы. В отличие от потоковых данных здесь нет особенностей, связанных с последовательным поступлением информации, но есть выраженный характер чувствительности и значительные размерности. В качестве примера можно упомянуть чеки супермаркетов и электронные резюме. Типовой объект для K-анонимизации – это числовые и/или категорийные данные. Чеки представляют собой совокупность числовых и строковых данных, обладают весьма высокой размерностью, при этом данные разрежены (лишь малое число значений в строковой составляющей отличны от 0). Некоторые методы обезличивания, базирующиеся именно на свойстве разреженности данных, предложены в [62]. Непосредственное применение алгоритмов K-анонимизации к строковым и тем более бинарным данным невозможно хотя бы по причине их переменной длины. В [63] предложен метод, кластеризующий исходные строковые данные с последующим моделированием синтетических данных, имеющих те же статистические характеристики, что и строки в отдельных кластерах.

Группировки (L-разнообразие, t-близость). Концепция K-анонимности выглядит интересной из-за простоты понимания и большого числа алгоритмов реализации. Тем не менее, ПД, удовлетворяющие условию K-анонимности, могут стать объектом атак, связанных с однородностью (когда значения всех атрибутов чувствительной информации в группе совпадают) и атак, связанных со знанием фоновых значений (когда можно использовать ассоциацию между одним/несколькими квазиидентификаторами для сужения множества возможных значений атрибутов чувствительной информации). Так, в [17] приведен пример, в котором фоновое знание о малой подверженности японцев сердечным приступам может быть использовано для сужения информационной неопределенности чувствительного атрибута “Заболевание”. Более детальное обсуждение влияния фоновых значений приведено в [64].

Можно сказать, что K-анонимность эффективно предотвращает деобезличивание конкретного субъекта ПД, но не всегда защищает отдельные значения атрибутов чувствительной информации. Концепция L-разнообразия гарантирует не только минимальный размер группы, равный K, но и обеспечивает внутригрупповое разнообразие чувствительных значений. Связанные с L-разнообразием понятия и определения предложены в [17]. Принципиальная идея этой концепции состоит в том, что при любом обобщении атрибутов нечувствительной информации остающиеся атрибуты чувствительной информации содержат не менее, чем L различных значений. Дополнительное требование L-разнообразия в случае многомерных атрибутов чувствительной информации делает задачу K-анонимизации особо сложной из-за проблемы “проклятия размерности” [14].

t-близость является дальнейшим развитием концепции L-разнообразия. Дело в том, что в модели, используемой концепцией L-разнообразия, все значения одного атрибута трактуются одинаково, независимо от их распределения. Тем самым косвенно предполагается, что реализации значений атрибута равновероятны. В реальности такая ситуация исключительна. Неравномерность распределения значений может затруднить построение обобщений, обладающих свойством L-разнообразия. Более того, знание глобальных распределений различных атрибутов может позволять делать выводы, снижающие уровень анонимности, особенно для атрибутов чувствительной информации. Кроме того, не все возможные значения атрибута обладают одинаковой “степенью чувствительности”. Например, логическое значение поля, соответствующее какому-либо заболеванию, более чувствительно, если оно имеет положительное значение, чем отрицательное. В [18] предложена концепция t-близости, согласно которой распределение значений чувствительного поля внутри любой группы анонимности не должно отличаться от глобального распределения более, чем на порог t. В качестве характеристики близости распределений часто используется метрика Вассерштейна.

1.3.3. Распределение данных. Целью большинства распределенных методов обезличивания ПД является обеспечение возможности выполнения статистических выводов по всему массиву данных без нарушения уровня анонимности отдельных записей. Идея заключается в разделении имеющегося массива данных таким образом, чтобы обеспечить корректность выполнения агрегированных статистических выводов, не имея одновременного доступа ко всей совокупности данных. Используемые для этого методы бывают как горизонтальными, так и вертикальными. В случае горизонтального разделения все множество записей разбивается на несколько подмассивов с одинаковым набором полей. При вертикальном разбиении отдельные массивы имеют различные атрибуты одного общего набора записей. Задача обеспечения анонимности распределенных данных по своей природе близка к задачам криптографии и протоколов конфиденциальных вычислений (secure multi-party computations, приемлемый обзор представлен в [65]). Подход к обезличиванию на основе распределенной обработки данных достаточно традиционен. Он базируется на вычислении функций по их аргументам, принадлежащим разным массивам, без допуска владельцев массивов к данным друг друга. Многие алгоритмы анализа и обработки данных могут быть пред-

ставлены в этом контексте как наборы повторений вычислительных примитивов: скалярных произведений, “безопасных” сумм, экстремумов и пр. Сами алгоритмы и обеспечиваемая ими анонимность зависят от уровня доверия между владельцами наборов данных. В литературе можно обнаружить такие характеристики как “получестные” и нечестные владельцы информации.

“Получестные” владельцы заинтересованы в получении информации из других массивов и пытаются извлечь пользу из полученной ими информации во время безопасных вычислений, но при этом не отклоняются от протокола. То есть целью их действий является не выполнение вычислений, а получение информации и чужих данных в процессе вычислений, например, путем большого числа запросов на выполнение разных вычислительных операций и последующего анализа полученных результатов. Во многих ситуациях это можно считать реалистичной моделью состязательного поведения. Нечестные владельцы информации могут отклоняться от протокола безопасных вычислений, поставлять на вход протокола специально синтезированные данные с целью получения информации о других информационных массивах.

Ключевым элементом выполнения безопасных вычислений является протокол передачи данных ([66, 67]), его называют “забывчивым”. Согласно ему отправитель предоставляет пару (x_0, x_1), а получатель – одно из битовых значений $\sigma \in \{0,1\}$. В результате получатель знает только x_σ, а отправитель не знает ничего. Для реализации этого протокола есть простые решения. В одном из них [66] отправитель генерирует два случайных открытых ключа, K_0, K_1, а получатель генерирует ключ K_σ. Получатель отправляет открытую часть своего ключа, шифруя ее одним из ключей, полученных от отправителя. Отправитель расшифровывает этот ключ, используя закрытые части своих ключей. При этом он получает два варианта ключа, один из которых действителен, а второй содержит мусорные данные. Отправитель шифрует x_0 с K_0, x_1 с K_1 и возвращает зашифрованные данные получателю. На этом этапе получатель может расшифровать только x_σ, т.к. он имеет только этот ключ дешифрования. Это “получестный” алгоритм, т.к. все промежуточные шаги требуют доверия между сторонами. Например, предполагается, что, когда получатель отправляет два ключа отправителю, он действительно знает ключ дешифрования только для одного из них. В случае нечестных владельцев информации необходимо убедиться, что отправитель выбирает открытые ключи согласно протоколу. Эффективный метод проверки этого факта приведен в [68]. В этой же работе протокол обобщен на случай нескольких владельцев информации. Забывчивый протокол является одним из простейших блоков в алгоритмах безопасных вычислений и может повторяться для вычисления заданной функции, поэтому важна его вычислительная эффективность. Численно эффективные методы для обоих видов владельцев информации также представлены в [68]. Более сложные проблемы в этой области включают вычисление различных вероятностных функций по аргументам с различными владельцами [69]. Алгоритмический аппарат этих действий развит как для случая двух владельцев информации, так и для произвольного числа владельцев [70]. Протокол забывчивой передачи данных может использоваться при выполнении некоторых вычислительных “примитивов”, связанных с вычислением расстояний между многомерными векторами, а также скалярных произведений [71]. Достаточно полный набор этих методов представлен в [72]. Многие из них основаны на отправке измененных или зашифрованных аргументов друг другу для вычисления функции с различными альтернативными значениями, последующей забывчивой передачей данных и получения в конце концов корректного конечного результата. В [72] в этом ключе рассмотрены задачи кластеризации, классификации, поиска ассоциативных правил, суммаризации и обобщения данных. Другой набор безопасных “примитивов” распределенного анализа данных представлен в [73]: он включает операции суммирования, объединения, пересечения, вычисления мощности множества и скалярного произведения. Все эти операции могут быть успешно применены при безопасной статистической обработке как вертикально, так и горизонтально распределенных наборов данных.

Горизонтальное разделение. В этом случае разные владельцы имеют разные наборы записей ПД с одинаковым (или очень близким) набором атрибутов. Большое число методов безопасного статистического анализа горизонтально распределенных данных основываются на общих методах, предложенных в [72, 73]. В других работах исследуются более тонкие методы: алгоритм ID3 построения дерева решений с использованием приближений наилучших атрибутов разбиения [74], наивный байесовский классификатор [75], классификатор Вапника–Червоненкиса с нелинейными ядрами [76]. “Предельный” случай рассмотрен в [77]: каждая запись, участвующая в совместной обработке, принадлежит отдельному владельцу. Кроме того, на случай горизонтально распределенных данных были обобщены такие алгоритмы анализа данных как поиск ассоциативных правил [78], кластеризация [79–81], коллаборативная фильтрация [82].

С обработкой горизонтально распределенных данных связана задача совместной индексации данных различных провайдеров. Сложность за-

ключается в том, что, с одной стороны, для ее выполнения нужна информационная кооперация, а с другой – провайдеры являются конкурентами друг друга. В [83] исследован алгоритм, выполняя который злоумышленник по результатам контекстных запросов к поисковой системе может восстановить интересующие его ПД. Решение возникшей проблемы предложено искать с помощью создания централизованного безопасного индекса в комплексе с распределенным механизмом контроля доступа. Предложенные меры позволяют обеспечить строгую приватность даже в случае сговора нескольких злоумышленников и в ситуации полного раскрытия индекса.

Вертикальное разделение. В этом случае остаются полезными многие вычислительные примитивы, например скалярное произведение, пересечение множеств и пр. Например, в [71] предложено использовать скалярное произведение для часто выполняемой операции пересчета элементов множества. Она же может быть выполнена с помощью безопасной процедуры нахождения пересечения множеств [73]. Один из методов поиска ассоциативных правил [84] использует скалярное произведение по вертикальному битовому столбцу, представляющему использование набора элементов в транзакциях, чтобы вычислить частоту соответствующих наборов элементов. Эта же вычислительно эффективная процедура может использоваться для многократного пересчета элементов множества.

Многие методы анализа данных распространены на случай их вертикальной распределенности: построение дерева решений [20], наивный байесовский классификатор [85], алгоритм Вапника–Червоненкиса [21], метод кластеризации k-средних [86]. Теоретическим вопросам реализации вычисления различных функций от вертикально распределенных аргументов с использованием методов криптографии посвящена работа [19].

Распределенные алгоритмы K-анонимности. Во многих практических задачах имеется необходимость обеспечить K-анонимность информационных массивов с различными собственниками. В [87] предложен K-анонимный протокол для вертикально распределенных данных с двумя владельцами. Общая идея заключается в определении для данных обоих владельцев некоторого общего квазиидентификатора, поскольку нужно атрибуты разных массивов каким-то образом синхронизировать для соответствующих субъектов ПД. Похожая идея предложена в [88], где собственникам предлагается предварительно договориться о правилах обобщения/объединения. В [89] задача K-анонимности рассматривалась для горизонтально распределенных данных. При этом исследован максимально распределенный вариант, т.е. в предположении, что каждая запись общего массива данных имеет своего собственника. Запись содержала квазиидентификаторы и атрибуты чувствительной информации, которые в итоге предполагается шифровать.

Проблема обеспечения K-анонимности важна и в других задачах безопасной обработки распределенных обезличенных данных. Так, известна задача сокрытия идентификационной информации при работе сервисов, связанных с определением местоположения [90–92]. Для такого контента K-анонимность идентификатора пользователя нужна даже в случае, когда информация о его местоположении становится доступной. В частности, это может быть в том случае, когда пользователь отправляет сообщение из некоторой точки данной локации/окрестности. Аналогичные проблемы могут возникать и в телекоммуникационных протоколах, в которых анонимность отправителя/получателя может быть не защищена. Здесь можно говорить о K-анонимности по отправителю K-анонимности по получателю.

1.3.4. Контроль работы приложений. Часто деобезличивание ПД оказывается возможным без прямого доступа к данным, а “благодаря” возможностям, предоставляемым приложениями, которые должны обеспечивать защищенный доступ к ПД. Все обсуждение здесь тесно связано с известной проблемой контроля разглашения [24] в статистических базах данных, хотя надо учитывать, что современные достижения в методах интеллектуального анализа данных предоставляют все более изощренные способы, чтобы злоумышленники могли делать выводы о ПД. В случае, когда ПД передаются в какой-либо форме оператором-собственником другому оператору, правила ассоциации могут представлять собой чувствительную информацию, например для целевого маркетинга, и эта информация должна быть защищена. К проблемам контроля раскрытия информации (в нашей терминологии – нарушению анонимности ПД) в приложениях относятся сокрытие ассоциативных правил, снижение эффективности классификации и обработки запросов.

Сокрытие ассоциативных правил. Ассоциативные правила являются весьма важными в сфере бизнеса при построении таргетированной рекламы. Для сокрытия ассоциативных правил используются два метода. Первый – искажение – описан в [93], где предложено параметр для транзакции заменять другим значением, а в битовом случае – инвертировать. Второй метод – блокировка [94] – оставляет параметр транзакции пустым. Оба метода имеют ряд побочных эффектов, влияющих на нечувствительные правила в данных. Многие из них могут быть искажены или утеряны вместе с чувствительными правилами

(атрибутами чувствительной информации) или могут появиться новые фантомные правила.

В [95] представлено формальное доказательство того факта, что применение метода искажения для сокрытия ассоциативных правил является NP-сложной задачей. В работе предлагается метод битового инвертирования $1 \to 0$ для снижения уровня поддержки чувствительных правил. Полезность предложенного подхода характеризуется числом нечувствительных правил, которые были нейтрализованы одновременно с чувствительным. Этот подход в [96] был усовершенствован. Детальное описание различных методов искажения данных для сокрытия ассоциативных правил можно найти в [93].

Основная идея метода блокировки представлена в [97]. Его главное качество заключается в том, что вместо искажения исходных данных он не дает доступа к ним. Некоторые интересные алгоритмы были представлены в [98], а затем расширены в [94] путем рассмотрения задачи реконструкции скрытых правил. Другой подход к построению алгоритмов сокрытия ассоциативных правил, ставящий целью снижение потерь нечувствительных правил/появления фантомных правил, представлен в [99]. В [100] обсуждалось влияние техники блокировки на определение злоумышленником атрибутов чувствительной информации – скрывались только те правила, которые могли раскрыть именно чувствительные атрибуты.

Снижение эффективности классификаторов. Одними из важнейших приложений, результаты которого могут понижать уровень анонимности, являются различные классификаторы. Основной сложностью противодействия здесь является такое преобразование данных, при котором точность классификации будет снижена без снижения уровня полезности для других видов приложений. В [23] и [101] были предложены методики снижения качества классификации в контексте правил классификации и приложений, связанных с построением дерева решений. Определение “скупого снижения” (parsimonious downgrading) введено в [101] в контексте блокировки выдачи результатов классификации с последующим изучением ее влияния на общую полезность. Система Rational Downgrader [23] была разработана с использованием этих принципов.

К снижению эффективности классификаторов могут адаптироваться методы сокрытия ассоциативных правил. В [102], предложен метод снижения эффективности классификаторов, базирующихся на правилах, и показана его эффективность.

Обработка запросов (и контроль логических выводов). Многие массивы ПД не являются полностью открытыми, но имеют открытый интерфейс. В этом есть очевидная угроза того, что злоумышленник может получить чувствительную информацию путем построения “правильной” серии запросов. Уровень такой угрозы может быть характеризован как “полное раскрытие”, когда злоумышленник получает точные значения чувствительных атрибутов, или “частичное раскрытие”, когда удается только сузить множество возможных значений атрибута. Есть два подхода противодействия такой угрозе. Первый – аудит запросов – состоит в том, что один или несколько запросов из последовательности могут быть отклонены с тем, чтобы сохранить анонимность исходных данных (см. примеры в [103, 104]). Второй подход состоит в управлении логическими заключениями. В этом случае либо исходные данные, либо результат запроса искажаются таким образом, чтобы сохранить уровень анонимности исходных данных (см. примеры в [33, 105, 106], методы искажения результатов запросов – в [19, 107–110]).

2. ВОПРОСЫ РЕАЛИЗАЦИИ АЛГОРИТМОВ ОБЕЗЛИЧИВАНИЯ

2.1. Сохранение полезности обезличенных данных

Потеря части информации является естественным следствием обезличивания ПД. Вопрос состоит в том, является ли утерянная часть полезной с точки зрения конечного потребителя обезличенной информации. В [14] представлен негативный ответ: обеспечение требуемого уровня анонимности приводит к исключению слишком большого числа атрибутов ПД. Вопрос об оценке уровня полезности часто сопровождает конкретную процедуру обезличивания. Так, для некоторых методов обезличивания [17, 49, 50, 111] предложены вполне конкретные показатели для потери информации от процесса обезличивания. Примерами таких показателей являются “высота обобщения” (общее число шагов обобщения, которые приводят исходные данные к нынешнему обезличенному виду), размеры анонимных групп в терминах K-анонимизации, меры различимости значений полей, относительная потеря информации при обезличивании и пр. Ряд предлагаемых метрик типа классификационной метрики [56] были специально разработаны для оценки полезности обезличенных данных для решения конкретных прикладных задач статистического анализа.

Задача обезличивания, сохраняющего полезность, и последующего безопасного интеллектуального анализа данных впервые была поставлена в [112]. Основная идея заключалась в отдельном опубликовании частных таблиц, имеющих полезность, но при этом обеспечивающих нужный уровень анонимности.

Еще один интересный подход к обезличиванию с сохранением полезности представлен в [113]. Основная идея состоит в том, чтобы зада-

вать для разных атрибутов разную полезность. В отличие от обсуждаемых выше методов обезличивания, которые называют глобальными из-за того, что они каждый атрибут исходных данных отображают на то же множество значений, этот подход предлагает разбиение исходного пространства данных на несколько областей, учитывающих полезность, с последующим отображением значения атрибута в зависимости от принадлежности той или иной области. Концептуально схожий подход, но на основе адаптации уровня анонимности данных в зависимости от частоты обращения к ним пользователей, предложен в [114].

Надо отметить, что всегда уровень полезности обезличенных данных рассматривается только в контексте конкретных прикладных задач и выбор функции для определения уровня полезности адаптируется именно к этим задачам. Например, в [50] для анализа свойств обезличенных данных, удовлетворяющих условию K-анонимности, рассматривалась величина потери информации (т.е. разности информационных энтропий до и после анонимизации). Подобный показатель позволяет оценить полезность данных для решения задач классификации. В [115] был предложен метод, сохраняющий точность базовых запросов (например, сохранение коэффициентов корреляции).

2.2. Ограничение применимости методов обезличивания

Возможности применения многих методов анализа обезличенных данных изначально ограничены “проклятием размерности” из-за наличия общедоступной информации. Например, в [14] представлен анализ одного метода K-анонимизации с ростом размерности данных. Рост размерности начинает оказывать значимое влияние, когда доступен большой объем справочной информации, в результате применения которой граница между квазиидентификаторами и атрибутами чувствительной информации становится размытой. Как уже упоминалось, подобный факт является стимулом для развития таких методов обезличивания, как L-разнообразие и t-близость. Но и без этого с ростом размерности данных обеспечение K-анонимности становится фактически невозможным из-за соответствующего роста групп анонимности. При постоянном числе записей субъектов ПД в исходном массиве данных появление новых атрибутов неизбежно ведет к разреженному характеру заполнения групп анонимности, поскольку чем больше атрибутов, тем меньшее число субъектов ПД (записей) попадает в ту или иную группу. В [14] делается важный пессимистический вывод о том, что для сохранения уровня анонимности может потребоваться исключение большого числа атрибутов. Такой же вывод в [17] был сделан в отношении модели L-разнообразия. Ясно, что такие операции снижают уровень полезности обезличенных данных.

Свойства методов рандомизации были исследованы в [28]. Предполагалось, что распределение исходных данных известно и они обезличены путем добавления шума. Задача восстановления данных решалась в постановке оценки максимального правдоподобия. Уровень шума предполагалось выбрать таким образом, чтобы оценка максимального правдоподобия указывала на другие записи, нежели на истинные. В работе показано, что с ростом размерности данных вероятность “обмануть” таким образом метод максимального правдоподобия также уменьшается.

Таким образом, проблема роста размерности данных является фундаментальным вызовом для методов обезличивания ПД, обеспечивающих высокий уровень анонимности. Маловероятно, что можно найти более эффективные методы для сохранения уровня анонимности, в случае доступности не только детальной информации о фоновых значениях, но и о некоторых выбранных группах субъектов ПД. Косвенные примеры таких нарушений приведены в [116, 117]. Описаны ситуации, когда информация из нескольких источников может быть скомпилирована для создания высокоразмерного представления, нарушающего ограничения анонимности для субъекта ПД. Очень ярким примером такой компиляции может служить совместная обработка потоковых данных о текущем геопозиционировании группы субъектов и одновременно получаемых данных из социальных сетей, раскрывающая текущее местоположение конкретного субъекта.

2.3. Чувствительность методов обезличивания к угрозам деобезличивания

В литературе обсуждается весьма широкий спектр возможных атак на обезличенные данные. Ограничиваясь кратким качественным анализом этих атак, сгруппируем их по результатам, которые могут быть достигнуты. Обобщенно их можно назвать угрозами, и исследовать каждый метод обезличивания по отношению к этим угрозам. Более детальный анализ приведен в работе [118], предлагающей именно такой подход. Для наших целей достаточно выделить три типа возможных угроз.

Угроза выделения (Singling Out) заключается в возможности изолировать некоторые или все записи, идентифицирующие субъект ПД в имеющемся наборе данных.

Угроза связывания (Linkability) заключается в возможности сопоставить как минимум две записи относительно одного и того же субъекта ПД или группы субъектов (в одном и том же или разных массивах данных). Если злоумышленник мо-

жет определить (например, с помощью корреляционного анализа), что две записи относятся к одной и той же группе лиц, но не может выделить отдельных участников в этой группе, то метод устойчив к выделению, но неустойчив к связыванию.

Угроза умозаключения, угроза логических выводов (Inference) заключается в возможности выделить, со значительной вероятностью, значение атрибута из общего множества значений атрибутов, для одной записи или их группы.

Все методы обезличивания, примененные отдельно, в той или иной степени подвержены всем перечисленным типам угроз. Некоторые из угроз могут быть устранены полностью или частично в конкретной ситуации, например, путем использования комбинации методов. Гарантированно противодействовать всем угрозам для некоторого универсального массива ПД не представляется возможным, поэтому такое большое внимание уделяется проектированию каждого конкретного приложения обезличивания ПД.

3. ПРИЛОЖЕНИЯ И ПРИМЕРЫ ОБЕЗЛИЧИВАНИЯ

3.1. Медицинские базы данных

Медицинская информация – возможно, самый очевидный источник ПД, обладающих высоким уровнем полезности. И при этом так же очевидна чувствительность этой информации к возможным утечкам. В качестве примеров систем обезличивания медицинской информации приведем Scrub [119] и Datafly [120], чей заслуженный возраст подчеркивает внимание к теме.

Scrub была создана для обезличивания клинических записей и писем, представленных в текстовой форме. Подобные записи обычно содержат ссылки на пациентов, членов их семей, адреса, телефоны и пр., но вместе с этим часто встречаются “загадочные” ссылки в виде сокращений, понятных только специалистам. Scrub параллельно использует несколько алгоритмов обнаружения текстов имен, адресов или номеров телефонов, применяет конкурентно локальные источники данных, специфические для данной области. В [119] показано, что система способна удалить из данных более 99% персональной идентификационной информации.

Datafly – одно из первых практических приложений преобразований, сохраняющих уровень анонимности. Эта система была разработана для предотвращения идентификации собственников медицинских карт, которые могут содержать множество атрибутов, в том числе персональные идентификаторы и квазиидентификаторы. Система была разработана в качестве ответа на опасения, что процесс удаления только персональных идентификаторов недостаточен для гарантии анонимности. Система имеет те же истоки, которые породили концепцию *K*-анонимности, но при создании Datafly самой концепции не было. Довольно простой подход в Datafly использует установку минимального размера ячейки для каждого атрибута. Уровень анонимности (отметим, что именно здесь уже формулируется это понятие) определяется в Datafly относительно именно установленного размера. Таким образом, значения в записях обобщаются до уровня неоднозначности размера ячейки, а не до точных значений. Идентифицирующие поля удаляются из данных. Кроме того, исключаются выбросы значений. Пользователь имеет возможность задать уровень анонимности в зависимости от профиля соответствующего получателя данных. Общий уровень анонимности определяется числовым параметром, лежащим в пределах от 0 до 1. Значение 0 соответствует исходным (неизмененным) данным, значение 1 – максимальному уровню обобщения. С момента создания Datafly прошло много времени и было выполнено большое число исследований в области обезличивания, в настоящий момент можно составить солидный список улучшений/модернизаций. И, конечно, к настоящему времени накопилось огромное число приложений медицинских ПД. Мы привели эти два примера с той лишь целью, чтобы подчеркнуть солидный исторический возраст проблематики.

3.2. Выявление возможного применения биологического оружия

Одна из актуальных задач применения обезличенных медицинских данных – своевременное обнаружение применения биологического оружия. Наверное, уместно учесть события с COVID-19 последних лет и добавить сюда “а также обнаружения и мониторинга опасных вирусных заболеваний”. Применение биологического оружия внешне, вызывает симптомы, сходные с симптомами других распространенных респираторных заболеваний: кашель, головная боль и жар. При отсутствии предварительных сведений о факте биологической атаки поставщики медицинских услуг могут диагностировать у пациента симптомы одного из наиболее распространенных и относительно безопасных респираторных заболеваний. Главная задача заключается в быстрой идентификации настоящей биологической атаки или пандемии в отличие от вспышки обычного респираторного заболевания. Во многих случаях необычное число или сочетание типовых симптомов в данной местности может указывать на такую атаку. Следовательно, для выявления необходимо отслеживать частоту появления “безопасных распространенных” заболеваний, а значит соответствующие данные необходимо обезличе-

но собирать/регистрировать/обрабатывать в органах общественного здравоохранения. Решение, предложенное в [8] – это "выборочное раскрытие" (снижение уровня анонимности по интересующим атрибутам), которое изначально допускает только ограниченный доступ к данным. Однако в случае подозрительной активности оно позволяет "углубиться" в базовые данные для уверенной идентификации вспышек опасных эпидемий.

3.3. Приложения по охране общественного порядка

Функционирование приложений, используемых для обеспечения общественного порядка, часто связано с обработкой результатов различных систем наблюдения. В [121] дается широкий обзор того, как эффективно использовать результаты наблюдения с сохранением уровня анонимности. Вот некоторые примеры таких приложений.

Система проверки учетных данных пытается сопоставить субъект данных с лицом, представляющим учетные данные. Кража паспортных данных или номера СНИЛС представляет серьезную угрозу для общественного порядка. В подходе к проверке учетных данных в [121] делается попытка использовать семантику, связанную с этими номерами, чтобы определить, действительно ли лицо, представляющее учетные данные, владеет им.

Система профилактики краж личных данных "ангел идентичности" (identity angel) [9] предлагает более активный подход. В ней анализируются данные данные из онлайн источников и обнаруживаются люди, которым грозит опасность кражи личных данных. Эта информация может быть использована для уведомления.

Система веб-камер наблюдения (система видеонаблюдения). Одним из возможных методов наблюдения является использование общедоступных веб-камер, которые можно использовать для обнаружения необычной активности [10, 121]. Надо заметить, что это гораздо больше нарушает анонимность, чем другие виды наблюдения, потому что зафиксированная в веб-камерах информация о субъекте включает изображение его лица, которое очевидно является персональным идентификатором. При этом исключить данную информацию гораздо сложнее в сравнении с исключением некоторой совокупности атрибутов из записи о субъекте ПД. Обезличивание фото- и видеоинформации предлагается проводить, извлекая из изображений только информацию о числе лиц и используя только ее для обнаружения необычной активности. В [10] была выдвинута гипотеза, что необычную активность можно обнаружить только по числу лиц, а не с использованием детальной информации о конкретных людях. Фактически, такой подход использует понижение уровня информации, доступной в веб-камерах, в зависимости от предметной области, чтобы сделать подход безопасным с точки зрения сохранения уровня анонимности.

Если в системе, связывающей камеры, есть функциональность по распознаванию лиц, угроза нарушения анонимности становится существенно выше. Для обезличивания самым простым решением было бы полное затемнение всех лиц, что выглядело бы странно для конечного пользователя. Более сбалансированный подход [7] заключается в использовании выборочного понижения качества изображения таким образом, чтобы ограничить способность надежного применения программного обеспечения распознавания лиц, сохраняя при этом детали лица на изображениях. Алгоритм называется k-Same и его ключевая идея состоит в том, чтобы идентифицировать лица, которые в некоторой степени похожи, а затем синтезировать "новые лица", как комбинации из похожих. Таким образом, личность владельца лица до некоторой степени анонимна, но видео остается полезным. Этот метод чем-то похож на K-анонимность, за исключением того, что он создает новые синтезированные данные для приложения.

3.4. Генетические данные

В последние годы были достигнуты успехи в области секвенирования ДНК и судебно-медицинской экспертизы с использованием ДНК. В результате базы данных ДНК очень быстро растут как в медицинском, так и в правоохранительном сообществе. Данные ДНК считаются крайне чувствительными, так как представляют собой информацию, почти однозначно идентифицирующую отдельного человека, а также важную медико-биологическую информацию о нем. Как и в случае роста размерности ПД, простого удаления непосредственно идентификационных данных, сопровождающих ДНК, недостаточно для предотвращения деобезличивания. В [122] показано, что программное обеспечение, обрабатывающее данные ДНК, может идентифицировать субъекта ПД независимо от другой информации. Такое программное обеспечение обрабатывает общедоступные медицинские данные и знания о различных конкретных заболеваниях, чтобы идентифицировать записи ДНК. В [122] показано, что идентифицировать можно 98–100% субъектов. Идентификация осуществляется путем взятия образцов ДНК человека и последующего построения генетического профиля, соответствующего полу, генетическим заболеваниям, месту сбора ДНК и т.д. Этот генетический профиль весьма эффективен при деобезличивании. Один из способов обеспе-

чить анонимность генетических данных – это использование решеток обобщения [123], являющихся, по сути, реализацией концепции *K*-анонимизации. Другой подход, обсуждаемый в [62], создает синтетические данные, которые имеют совокупные характеристики исходных данных, но сохраняют уровень анонимности исходных записей.

Одним из способов нарушения конфиденциальности геномных данных является последовательная реидентификация (trial re-identification), при которой уникальность шаблонов пациентов [116, 117] используется для деобезличивания. Предпосылка этих работ заключается в том, что пациенты часто посещают и оставляют генетические данные в "информационно разнесенных" местах: клиниках, лабораториях и пр. Больницы обычно отделяют клинические данные от генетических и предоставляют генетические данные для исследовательских целей. При этом данные кажутся обезличенными, а схема посещения пациентов кодируется на сайте, с которого они получены. В [116, 117] показано, что эта информация может быть объединена с общедоступными данными, чтобы выполнить деобезличивание.

4. ЗАКЛЮЧЕНИЕ

Выполненный обзор проблематики автоматизации обезличивания ПД был бы неполным без упоминания хотя бы некоторых резонансных случаев нарушения анонимности обезличенных ПД. Именно отдельные, уникальные примеры скандального характера инициировали и поддержали усилия исследователей в этой области, дали основу для придания этим исследованиям фундаментального характера. Заметим, кстати, что далеко не все резонансные ситуации носили и носят противоправный характер. Не всегда источником атаки на уровень конфиденциальности оказываются злоумышленники, действующие с целью незаконного извлечения какой-то прибыли. Чаще виновник оказывается таковым по причине некомпетентности в области ПД вообще и обезличивания в частности. К сожалению, также не всегда даже резонансные случаи нарушения анонимности подвергаются методическому исследованию. Информация о таких историях поступает чаще из газетных публикаций и по большей части направлена на привлечение внимания, чем на обстоятельное доказательное и аналитическое исследование. Но любопытно хотя бы обозначить фигурантов этих историй. Так, статья [124] посвящена истории скандала между подразделением Google, компанией DeepMind и медицинским фондом Royal Free London NHS Foundation Trust. Нашумевший случай сбора персональной информации профилей Facebook компанией Cambridge Analytica описан в [125]. Статья [126] касается случая с небольшим индейским племенем хавасупай и исследователями Университета Аризоны. Случай, инспирированный специалистом в области ПД Латаньей Суинни, продемонстрировавшей возможность деобезличивания в отношении губернатора Массачусетса Уильяма Вэлда, приведен в [127, 128]. Довольно аккуратные статистические исследования данных в отношении граждан США выполнены в [130] и [131]. В статье [132] декларирован факт деобезличивания медицинской информации, опубликованной департаментом здравоохранения Австралии, а в [133] – возможность деобезличивания субъектов ПД по общедоступному набору данных об аренде велосипедов в Лондоне. Проблемы возможного деобезличивания субъектов по данным электронных проездных документов в Риге представлены в статье [134]. Наконец, одному из наиболее громких и "математически проработанных" случаев деобезличивания, связанному с базой данных Netflix, посвящена работа [135].

Эти и многие другие факты угроз нарушения анонимности обезличенных ПД, прежде всего, вызывают большой общественный резонанс. Но и научное сообщество, как следует из представленного материала, уделяет весьма большое внимание проблеме обезличивания ПД. Причем для исследований используется вполне современный математический аппарат для решения задач, которые можно сгруппировать по следующим направлениям:

- обеспечение безопасного сбора персональных данных,
- построение надежных алгоритмов обезличивания персональных данных (существующих и новых типов),
- оценка полезности обезличенных данных,
- поиск уязвимостей существующих и перспективных алгоритмов обезличивания,
- поиск квазиидентификаторов в обезличенных данных,
- поиск ассоциативных правил в обезличенных данных,
- разработка методов безопасной обработки персональных данных,
- разработка методов безопасной распределенной обработки персональных данных и безопасных вычислений.

В этих исследованиях имеются возможности для применения интеллектуальных усилий специалистов самых разных компетенций. Так, авторы обратили внимание на два аспекта, характерных для всех методов обезличивания. Во-первых, в рамках каждого метода обсуждается некоторый показатель его надежности – границы, в рамках которых можно говорить о поддержании уровня анонимности. Но некоторая вероятность деобезличивания имеется всегда. Говорить об исключе-

нии возможности нарушения анонимности можно лишь для методов, основанных на искусственном синтезе обезличенных данных. Во-вторых, компьютерному синтезу обезличенных данных внимания уделяется довольно мало, практически отсутствуют сколь-либо универсальные методики. Вторая часть нашего исследования посвящена применению различных методов обезличивания, в том числе компьютерного синтеза данных. В ней будет представлена математическая модель процесса, даны формальные определения, предложен метод обезличивания, основанный на комбинации простых статистических характеристик и ядерных оценок Розенблатта–Парзена, выполнены численные эксперименты.

6. БЛАГОДАРНОСТИ

Работа выполнялась с использованием инфраструктуры Центра коллективного пользования "Высокопроизводительные вычисления и большие данные" (ЦКП "Информатика" ФИЦ ИУ РАН, Москва).

СПИСОК ЛИТЕРАТУРЫ

1. *Aggarwal C.C., Yu P.S.* A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In: *Aggarwal C.C., Yu P.S.* (eds) Privacy-Preserving Data Mining. Advances in Database Systems. 2008. V. 34. Springer, Boston, MA.
2. *Domingo-Ferrer J., Farràs O., Ribes-González J., Sánchez D.* Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges // Computer Communications. 2019. V. 140–141. P. 38–60.
3. *Sahi M.A. et al.* Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions // IEEE Access. 2018. V. 6. P. 464–478. https://doi.org/10.1109/ACCESS.2017.2767561
4. *Spiekermann S., Cranor L.F.* Engineering Privacy // IEEE Transactions on Software Engineering. 2009. V. 35. № 1. P. 67–82. https://doi.org/10.1109/TSE.2008.88
5. *Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin Y., Theodoridis Y.* State-of-the-art in privacy preserving data mining // ACM SIGMOD Record. 2004. V. 33. № 1.
6. *Guide to Basic Data Anonymization Technique*. Personal Data Protection Commission, Singapore. 2018.
7. *Newton E., Sweeney L., Malin B.* Preserving Privacy by De-identifying Facial Images // IEEE Transactions on Knowledge and Data Engineering. 2005.
8. *Sweeney L.* Privacy-Preserving Bio-terrorism Surveillance // AAAI Spring Symposium, AI Technologies for Homeland Security. 2005.
9. *Sweeney L.* AI Technologies to Defeat Identity Theft Vulnerabilities // AAAI Spring Symposium, AI Technologies for Homeland Security. 2005.
10. *Sweeney L., Gross R.* Mining Images in Publicly-Available Cameras for Homeland Security // AAAI Spring Symposium, AI Technologies for Homeland Security. 2005.
11. *Agrawal R., Srikant R.* Privacy-Preserving Data Mining // Proceedings of the ACM SIGMOD Conference. 2000.
12. *Agrawal D., Aggarwal C.C.* On the Design and Quantification of Privacy-Preserving Data Mining Algorithms // ACM PODS Conference. 2002.
13. *Aggarwal G., Feder T., Kenthapadi K., Motwani R., Panigrahy R., Thomas D., Zhu A.* Approximation Algorithms for k-anonymity. Journal of Privacy Technology. 2005. № 20051120001.
14. *Aggarwal C.C.* On k-anonymity and the curse of dimensionality // VLDB Conference. 2005.
15. *LeFevre K., DeWitt D., Ramakrishnan R.* Incognito: Full Domain K-Anonymity // ACM SIGMOD Conference. 2005.
16. *Meyerson A., Williams R.* On the complexity of optimal k-anonymity // ACM PODS Conference. 2004.
17. *Machanavajjhala A., Gehrke J., Kifer D., Venkitasubramaniam M.* L-Diversity: Privacy Beyond k-Anonymity // ICDE Conference. 2006.
18. *Li N., Li T., Venkatasubramanian S.* t-Closeness: Privacy beyond k-anonymity and l-diversity // ICDE Conference. 2007.
19. *Dwork C., Nissim K.* Privacy-Preserving Data Mining on Vertically Partitioned Databases // CRYPTO. 2004.
20. *Vaidya J., Clifton C.* Privacy-Preserving Decision Trees over vertically partitioned data // Lecture Notes in Computer Science. 2005. V. 3654.
21. *Yu H., Vaidya J., Jiang X.* Privacy-Preserving SVM Classification on Vertically Partitioned Data // PA-KDD Conference. 2006.
22. *Verykios V.S., Elmagarmid A., Bertino E., Saygin Y., Dasseni E.* Association Rule Hiding // IEEE Transactions on Knowledge and Data Engineering. 2004. V. 16. № 4.
23. *Moskowitz I., Chang L.* A decision theoretic system for information downgrading // Joint Conference on Information Sciences. 2000.
24. *Adam N., Wortmann J.C.* Security-Control Methods for Statistical Databases: A Comparison Study // ACM Computing Surveys. 1989. V. 21. № 4.
25. *Liew C.K., Choi U.J., Liew C.J.* A data distortion by probability distribution // ACM TODS. 1985. V. 10. № 3. P. 395–411.
26. *Warner S.L.* Randomized Response: A survey technique for eliminating evasive answer bias // Journal of American Statistical Association. 1965. V. 60. № 309. P. 63–69.
27. *Silverman B.W.* Density Estimation for Statistics and Data Analysis. Chapman and Hall. 1986.
28. *Aggarwal C.C.* On Randomization, Public Information and the Curse of Dimensionality // ICDE Conference. 2007.
29. *Gambs S., Kegl B., Aimeur E.* Privacy-Preserving Boosting // Knowledge Discovery and Data Mining Journal. 2007. V. 14. № 1. P. 131–170.
30. *Zhang P., Tong Y., Tang S., Yang D.* Privacy-Preserving Naive Bayes Classifier // Lecture Notes in Computer Science. 2005. V. 3584.

31. *Evfimievski A., Srikant R., Agrawal R., Gehrke J.* Privacy-Preserving Mining of Association Rules // ACM KDD Conference. 2002.
32. *Rizvi S., Haritsa J.* Maintaining Data Privacy in Association Rule Mining // VLDB Conference. 2002.
33. *Agrawal R., Srikant R., Thomas D.* Privacy-Preserving OLAP // Proceedings of the ACM SIGMOD Conference. 2005.
34. *Polat H., Du W.* SVD-based collaborative filtering with privacy // ACM SAC Symposium. 2005.
35. *Bertino E., Fovino I., Provenza L.* A Framework for Evaluating Privacy-Preserving Data Mining Algorithms // Data Mining and Knowledge Discovery Journal. 2005. V. 11. P. 121–154.
36. *Evfimievski A., Gehrke J., Srikant R.* Limiting Privacy Breaches in Privacy Preserving Data Mining // ACM PODS Conference. 2003.
37. *Huang Z., Du W., Chen B.* Deriving Private Information from Randomized Data // ACM SIGMOD Conference. 2005. P. 37–48.
38. *Kargupta H., Datta S.,Wang Q., Sivakumar K.* On the Privacy Preserving Properties of Radom Data Perturbation Techniques // ICDM Conference. 2003. P. 99–106.
39. *Johnson W., Lindenstrauss J.* Extensions of Lipshitz Mapping into Hilbert Space // Contemporary Math. 1984. V. 26. P. 189–206.
40. *Oliveira S.R.M., Zaiane O.* Privacy Preserving Clustering by Data Transformation // Proc. 18th Brazilian Symp. Databases. 2003. P. 304–318.
41. *Oliveira S.R.M., Zaiane O.* Data Perturbation by Rotation for Privacy-Preserving Clustering // Technical Report TR04-17, Department of Computing Science, University of Alberta, Edmonton, AB, Canada. 2004.
42. *Chen K., Liu L.* Privacy-preserving data classification with rotation perturbation // ICDM Conference. 2005.
43. *Liu K., Kargupta H., Ryan J.* Random Projection Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining // IEEE Transactions on Knowledge and Data Engineering. 2006. V. 18. № 1.
44. *Kim J., Winkler W.* Multiplicative Noise for Masking Continuous Data // Technical Report Statistics 2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C. 2003.
45. *Mukherjee S., Chen Z., Gangopadhyay S.* A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier based transforms // VLDB Journal. 2006.
46. *Liu K., Giannella C. Kargupta H.* An Attacker's View of Distance Preserving Maps for Privacy-Preserving Data Mining // PKDD Conference. 2006.
47. *Fienberg S., McIntyre J.* Data Swapping: Variations on a Theme by Dalenius and Reiss // Technical Report, National Institute of Statistical Sciences. 2003.
48. *Samarati P.* Protecting Respondents' Identities in Microdata Release // IEEE Trans. Knowl. Data Eng. 2001. V. 13. № 6. P. 1010–1027.
49. *Bayardo R.J., Agrawal R.* Data Privacy through Optimal k-Anonymization // Proceedings of the ICDE Conference. 2005. P. 217–228.
50. *Fung B., Wang K., Yu P.* Top-Down Specialization for Information and Privacy Preservation // ICDE Conference. 2005.
51. *Wang K., Yu P., Chakraborty S.* Bottom-Up Generalization: A Data Mining Solution to Privacy Protection // ICDM Conference. 2004.
52. *Domingo-Ferrer J., Mateo-Sanz J.* Practical data-oriented micro-aggregation for statistical disclosure control // IEEE TKDE. 2002. V. 14. № 1.
53. *Aggarwal G., Feder T., Kenthapadi K., Khuller S., Motwani R., Panigrahy R., Thomas D., Zhu A.* Achieving Anonymity via Clustering // ACM PODS Conference. 2006.
54. *Aggarwal C.C., Yu P.S.* A Condensation approach to privacy preserving data mining // EDBT Conference. 2004.
55. *Winkler W.* Using simulated annealing for k-anonymity // Technical Report 7, US Census Bureau, Washington D.C. 20233. 2002.
56. *Iyengar V.S.* Transforming Data to Satisfy Privacy Constraints // KDD Conference. 2002.
57. *Lakshmanan L., Ng R., Ramesh G.* To Do or Not To Do: The Dilemma of Disclosing Anonymized Data // ACM SIGMOD Conference. 2005.
58. *Aggarwal C.C., Yu P.S.* On Variable Constraints in Privacy-Preserving Data Mining // SIAM Conference. 2005.
59. *Xiao X., Tao Y.* Personalized Privacy Preservation // ACM SIGMOD Conference. 2006.
60. *Wang K., Fung B.C.M.* Anonymization for Sequential Releases // ACM KDD Conference. 2006.
61. *Pei J., Xu J., Wang Z., Wang W., Wang K.* Maintaining k-Anonymity against Incremental Updates // Symposium on Scientific and Statistical Database Management. 2007.
62. *Aggarwal C.C., Yu P.S.* On Privacy-Preservation of Text and Sparse Binary Data with Sketches // SIAM Conference on Data Mining. 2007.
63. *Aggarwal C.C., Yu P.S.* On Anonymization of String Data // SIAM Conference on Data Mining. 2007.
64. *Martin D., Kifer D., Machanavajjhala A., Gehrke J., Halpern J.* Worst-Case Background Knowledge // ICDE Conference. 2007.
65. *Pinkas B.* Cryptographic Techniques for Privacy-Preserving Data Mining // ACM SIGKDD Explorations. 2002. V. 4. № 2.
66. *Even S., Goldreich O., Lempel A.* A Randomized Protocol for Signing Contracts // Communications of the ACM. 1985. V. 28.
67. *Rabin M.O.* How to exchange secrets by oblivious transfer // Washington D.C. 20233TR-81, Aiken Corporation Laboratory. 1981.
68. *Naor M., Pinkas B.* Efficient Oblivious Transfer Protocols // SODA Conference. 2001.
69. *Yao A.C.* How to Generate and Exchange Secrets // FOCS Conference. 1986.
70. *Chaum D., Crepeau C., Damgard I.* Multiparty unconditionally secure protocols // ACM STOC Conference. 1988.
71. *Ioannidis I., Grama A., Atallah M.* A secure protocol for computing dot-products in clustered and distributed

environments // International Conference on Parallel Processing. 2002.

72. *Du W., Atallah M.* Secure Multi-party Computation: A Review and Open Problems // CERIAS Technical Report 2001-51, Purdue University. 2001.
73. *Clifton C., Kantarcioglou M., Lin X., Zhu M.* Tools for privacy preserving distributed data mining // ACM SIGKDD Explorations. 2002. V. 4. № 2.
74. *Lindell Y., Pinkas B.* Privacy-Preserving Data Mining // CRYPTO. 2000.
75. *Kantarcioglu M., Vaidya J.* Privacy-Preserving Naive Bayes Classifier for Horizontally Partitioned Data // IEEE Workshop on Privacy-Preserving Data Mining. 2003.
76. *Yu H., Jiang X., Vaidya J.* Privacy-Preserving SVM using nonlinear Kernels on Horizontally Partitioned Data // SAC Conference. 2006.
77. *Yang Z., Zhong S., Wright R.* Privacy-Preserving Classification of Customer Data without Loss of Accuracy // SDM Conference. 2006.
78. *Kantarcioglu M., Clifton C.* Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data // IEEE TKDE Journal. 2004. V. 16. № 9.
79. *Inan A., Saygin Y., Savas E., Hintoglu A., Levi A.* Privacy-Preserving Clustering on Horizontally Partitioned Data // Data Engineering Workshops. 2006.
80. *Jagannathan G., Wright R.* Privacy-Preserving Distributed k-means clustering over arbitrarily partitioned data // ACM KDD Conference. 2005.
81. *Jagannathan G., Pillaipakkamnatt K., Wright R.* A New Privacy-Preserving Distributed k-Clustering Algorithm // SIAM Conference on Data Mining. 2006.
82. *Polat H., Du W.* Privacy-Preserving Top-N Recommendations on Horizontally Partitioned Data // Web Intelligence. 2005.
83. *Bawa M., Bayardo R.J., Agrawal R.* Privacy-Preserving Indexing of Documents on the Network // VLDB Conference. 2003.
84. *Vaidya J., Clifton C.* Privacy-Preserving Association Rule Mining in Vertically Partitioned Databases // ACM KDD Conference. 2002.
85. *Vaidya J., Clifton C.* Privacy-Preserving Naive Bayes Classifier over vertically partitioned data // SIAM Conference. 2004.
86. *Vaidya J., Clifton C.* Privacy-Preserving k-means clustering over vertically partitioned Data // ACM KDD Conference. 2003.
87. *Jiang W., Clifton C.* Privacy-preserving distributed k-Anonymity // Proceedings of the IFIP 11.3 Working Conference on Data and Applications Security. 2005.
88. *Wang K., Fung B.C.M., Dong G.* Integrating Private Databases for Data Analysis // Lecture Notes in Computer Science. 2005. V. 3495.
89. *Zhong S., Yang Z., Wright R.* Privacy-enhancing k-anonymization of customer data // Proc. of the ACM SIGMOD-SIGACT-SIGART Principles of Database Systems, Baltimore, MD. 2005.
90. *Bettini C., Wang X.S., Jajodia S.* Protecting Privacy against Location Based Personal Identification // Proc. of Secure Data Management Workshop, Trondheim, Norway. 2005.
91. *Gedik B., Liu L.* A customizable k-anonymity model for protecting location privacy // ICDCS Conference. 2005.
92. *Mimoto T., Kiyomoto Sh., Miyaji A.* Secure Data Management Technology // In Security Infrastructure Technology for Integrated Utilization of Big Data (*T. Mimoto and A. Miyaji* eds.), Singapore, Springer Open. 2020.
93. *Oliveira S.R.M., Zaiane O., Saygin Y.* Secure Association-Rule Sharing // PAKDD Conference. 2004.
94. *Saygin Y., Verykios V., Clifton C.* Using Unknowns to prevent discovery of Association Rules // ACM SIGMOD Record. 2001. V. 30. № 4.
95. *Atallah M., Elmagarmid A., Ibrahim M., Bertino E., Verykios V.* Disclosure limitation of sensitive rules // Workshop on Knowledge and Data Engineering Exchange. 1999.
96. *Dasseni E., Verykios V., Elmagarmid A., Bertino E.* Hiding Association Rules using Confidence and Support // 4th Information Hiding Workshop. 2001.
97. *Chang L., Moskowitz I.* An integrated framework for database inference and privacy protection. Data and Applications Security. Kluwer. 2000.
98. *Saygin Y., Verykios V., Elmagarmid A.* Privacy-Preserving Association Rule Mining // 12th International Workshop on Research Issues in Data Engineering. 2002.
99. *Wu Y.-H., Chiang C.-M., Chen A.L.P.* Hiding Sensitive Association Rules with Limited Side Effects // IEEE Transactions on Knowledge and Data Engineering. 2007. V. 19. № 1.
100. *Aggarwal C., Pei J., Zhang B.* A Framework for Privacy Preservation against Adversarial Data Mining // ACM KDD Conference. 2006.
101. *Chang L., Moskowitz I.* Parsimonious downgrading and decision trees applied to the inference problem // New Security Paradigms Workshop. 1998.
102. *Natwichai J., Li X., Orlowska M.* A Reconstruction-based Algorithm for Classification Rules Hiding // Australasian Database Conference. 2006.
103. *Kenthapadi K., Mishra N., Nissim K.* Simulatable Auditing // ACM PODS Conference. 2005.
104. *Nabar S., Marthi B., Kenthapadi K., Mishra N., Motwani R.* Towards Robustness in Query Auditing // VLDB Conference. 2006.
105. *Chawla S., Dwork C., McSherry F., Smith A., Wee H.* Towards Privacy in Public Databases // TCC. 2005.
106. *Mishra N., Sandler M.* Privacy vs Pseudorandom Sketches // ACM PODS Conference. 2006.
107. *Blum A., Dwork C., McSherry F., Nissim K.* Practical Privacy: The SuLQ Framework // ACM PODS Conference. 2005.
108. *Dinur I., Nissim K.* Revealing Information while preserving privacy // ACM PODS Conference. 2003.
109. *Dwork C., Kenthapadi K., McSherry F., Mironov I., Naor M.* Our Data, Ourselves: Privacy via Distributed Noise Generation // EUROCRYPT. 2006.
110. *Dwork C., McSherry F., Nissim K., Smith A.* Calibrating Noise to Sensitivity in Private Data Analysis // TCC. 2006.

111. *Wang K., Fung B.C.M., Yu P.* Template based Privacy-Preservation in classification problems // ICDM Conference. 2005.

112. *Kifer D., Gehrke J.* Injecting utility into anonymized datasets // SIGMOD Conference. 2006. P. 217–228.

113. *Xu J., Wang W., Pei J., Wang X., Shi B., Fu A.W.C.* Utility Based Anonymization using Local Recoding // ACM KDD Conference. 2006.

114. *LeFevre K., DeWitt D., Ramakrishnan R.* Workload Aware Anonymization // KDD Conference. 2006.

115. *Koudas N., Srivastava D., Yu T., Zhang Q.* Aggregate Query Answering on Anonymized Tables // ICDE Conference. 2007.

116. *Malin B., Sweeney L.* Re-identification of DNA through an automated linkage process // Proc. AMIA Symp. 2001. P. 423–427.

117. *Malin B.* Why methods for genomic data privacy fail and what we can do to fix it // AAAS Annual Meeting, Seattle, WA. 2004.

118. ARTICLE 29 DATA PROTECTION WORKING PARTY. Opinion 05/2014 on Anonymisation Techniques. Adopted on 10 April 2014.

119. *Sweeney L.* Replacing Personally Identifiable Information in Medical Records, the Scrub System // Proc. AMIA Annu Fall Symp. 1996. P. 333–337.

120. *Sweeney L.* Guaranteeing Anonymity while Sharing Data, the Datafly System // Proc. AMIA Annu Fall Symp. 1997. P. 51–55.

121. *Sweeney L.* Privacy Technologies for Homeland Security // Testimony before the Privacy and Integrity Advisory Committee of the Department of Homeland Security, Boston, MA, June 15. 2005.

122. *Malin B., Sweeney L.* Detrmining the identifiability of DNA database entries // Proc. AMIA Symp. 2000. P. 537–541.

123. *Malin B.* Protecting DNA Sequence Anonymity with Generalization Lattices // Methods of Information in Medicine. 2005. V. 44. № 5. P. 687–692.

124. *Hodson H.* Revealed: Google AI has access to huge haul of NHS patient data // New Scientist, 29 Apr 2016.

125. *Cadwalladr C., Graham-Harrison E.* Revealed: 50 million facebook profiles harvested for Cambridge Analytica in major data breach // The Guardian, 17 Mar 2018.

126. Harmon A. Indian tribe wins fight to limit research of its DNA // New York Times. 2010, April, 22.

127. *Meyer M.* Law, Ethics & Science of Re-identification Demonstrations // Bill of Health: Examining the Intersection of Health Law, Biotechnology and Bioethics, Petrie Flom Center at Harvard University. 2021.

128. *Ohm P.* Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization // UCLA Law Review. 2010. V. 57. P. 1700–1777.

129. *de Montjoye Y.-A., Radaelli L., Singh V.K., Pentland A.* Unique in the shopping mall: on the reidentifiability of credit card metadata // Science. 2015. V. 347. P. 536–539.

130. *Golle P.* Revisiting the uniqueness of simple demographics in the U.S. population // Workshop on privacy in the electronic society, New York, Association for Computive Machinery. 2006.

131. *Rocher L., Hendrickx J.M., de Montjoye Y.-A.* Estimating the success of re-identifications in incomplete datasets using generative models // Nat. Commun.. 2019. V. 10. № 1 (3069).

132. *Culnane C., Rubinstein B.I.P., Teague V.* Health data in an open world // Preprint at: https://arxiv.org/abs/1712.05627. 2017.

133. *Siddle J.* I know where you were last summer: London's public bike data is telling everyone where you've been // vartree.blogspot.com. 2014.

134. *Lavrenovs A., Podins K.* Privacy violations in Riga open data public transport system // 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), Vilnius, Lithuania. 2016. P. 1–6.

135. *Narayanan A., Shmatikov V.* Robust De-anonymization of Large Sparse Datasets // IEEE Symposium on Security and Privacy. 2008. P. 111–125.