Федеральное государственное автономное образовательное учреждение
высшего образования
«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(национальный исследовательский университет)»

На правах рукописи

Мищенко Евгений Юрьевич

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ ОБЕЗЛИЧИВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ И ОЦЕНКА ЭФФЕКТИВНОСТИ ИСПОЛЬЗУЕМЫХ МЕТОДОВ НА ОСНОВЕ МОДЕЛИ НАРУШИТЕЛЯ

Специальность 2.3.6 – Методы и системы защиты информации,
информационная безопасность

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
кандидат технических наук, доцент
Соколов Александр Николаевич

Челябинск – 2022

Оглавление

Введение

Актуальность темы исследования. Обеспечение безопасности персональных данных (ПД) является обязательным условием их обработки в соответствии с Федеральным законом «О персональных данных» [1]. Методы защиты ПД по критерию воздействия на злоумышленника можно разделить на активные (использование средств защиты информации (СЗИ), определяющих права доступа пользователя, но не влияющих на характеристики ПД) и пассивные (использование мер, не влияющих на права доступа пользователя, но изменяющих характеристики ПД). К активным относятся СЗИ, сертифицированные ФСТЭК России либо ФСБ России, интегрированные в операционную среду обработки ПД. К пассивным методам относятся шифрование, стеганография и обезличивание. Шифрование обеспечивается применением средств криптографической защиты информации (СКЗИ), сертифицированных ФСБ России либо на каждом рабочем месте, либо на границе информационной системы. Причем применение пассивных криптографических средств в общем случае не исключает необходимости применения активных СЗИ. Применение стеганографии не регламентировано нормативными актами, в то время как обезличивание регламентировано законом «О персональных данных» [1], согласно которому оно является методом обработки ПД, в результате которой невозможно без дополнительной информации определить принадлежность этих ПД физическому лицу (ФЛ).

Применение обезличивания ПД обусловлено необходимостью обработки, хранения и передачи ПД в научных, статистических и прочих целях в форме представления, не допускающей возможности нанесения ущерба ФЛ, которому эти ПД принадлежат, так как обезличивание скрывает эту принадлежность.

К достоинствам применения обезличивания можно отнести возможность реализации методов обезличивания путем модернизации прикладного программного обеспечения силами оператора ПД, что упрощает эксплуатацию информационных систем персональных данных (ИСПДн).

К недостаткам применения обезличивания можно отнести:

– необходимость защиты дополнительной информации, предназначенной для восстановления (деобезличивания) ПД, при ее хранении, передаче и использовании для доступа к ПД на рабочем месте;

– наличие у злоумышленника возможностей получения необходимой для деобезличивания дополнительной информации косвенными методами (путем подбора, вычисления или из открытых источников).

В России обезличивание ПД регламентируется Приказом Роскомнадзора от 05.09.2013 г. № 996 «Об утверждении требований и методов по обезличиванию персональных данных» [2], устанавливающего такие методы обезличивания, как введение идентификаторов, изменение состава или семантики и перемешивание. Но, предложенные характеристики методов обезличивания имеют исключительно качественный характер.

Актуальность моделирования процессов обезличивания обусловлена необходимостью решения проблем, возникающих при реализации методов обезличивания, к которым можно отнести:

– отсутствие методики обоснования выбора методов обезличивания и настройки их характеристик в зависимости от свойств базы ПД;

– отсутствие методики количественной оценки эффективности методов обезличивания ПД;

– отсутствие схемы безопасной передачи данных между разделенными частями обезличенной базы.

Степень разработанности темы исследования. Для исследований по рассматриваемой тематике характерны следующие этапы:

1) теоретическое обоснование необходимости обезличивания ПД по отдельным атрибутам с учетом их семантики и количества записей базы данных (БД);
2) разработка функциональных схем и алгоритмов обезличивания ПД;
3) программная реализация алгоритмов обезличивания ПД;

4) внедрение функциональных схем и алгоритмов обезличивания ПД;
5) разработка методик оценки эффективности алгоритмов обезличивания ПД;
6) внедрение методик оценки эффективности алгоритмов обезличивания ПД.

В работах зарубежных исследователей преимущественно рассматривается необратимое обезличивание ПД, при этом определенное внимание уделяется теоретическому обоснованию выбора атрибутов для обезличивания [3, 4] и оценке эффективности обезличивания [5].

Работы по теме исследования в России начались еще до ввода в действие приказа Роскомнадзора [2]. К наиболее ранним исследованиям можно отнести работы И.Ю. Кучина (обоснование, разработка, реализация и внедрение обезличивания по методу изменения состава/семантики, 2012 г.) [6], Е.С. Волокитиной (разработка, реализация, патент и внедрение обезличивания по методу введения идентификаторов, 2012 г.) [7, 8], М.И. Денисова и К.А. Чехонина (разработка и реализация алгоритма обезличивания по методу перемешивания, 2013 г.) [9].

После ввода в действие приказа Роскомнадзора [2] метод введения идентификаторов рассмотрен в работах А.А. Халафяна и А.А. Кошкарова (разработка, реализация и внедрение, 2015 г.) [10], А.А. Ноздриной и Д.В. Применко (разработка и реализация, 2016 г.) [11]. Но наиболее разработанным является метод перемешивания, различные алгоритмы которого приведены в работах В.В. Воронина и Н.Л. Нехай (разработка, реализация и внедрение, 2017 г.) [12], К.О. Бондаренко и В.А. Козлова (разработка и реализация, 2015 г.) [13], Е.А. Макаровой и Д.Г. Лагерева (разработка и реализация, 2016 г.) [14].

В особое направление следует выделить алгоритмы обезличивания, которые, строго говоря, выходят за рамки методов, установленных приказом Роскомнадзора [2], так как реализуются с применением криптографических средств, но, по сути, выполняют ту же задачу обезличивания (например, работы

Ю.В. Трифоновой и Р.Ф. Жаринова, 2014 г. [15], И.М. Ажмухамедова, Р.Ю. Демина и И.В. Сафарова, 2015 г. [16]).

С момента ввода в действие терминологии по обезличиванию ПД Приказом Роскомнадзора [2], а также Методическими рекомендациями [17] стали возможными строгое теоретическое обоснование и оценка эффективности алгоритмов обезличивания ПД, но эти этапы разработки практически не нашли отражения в работах исследователей. Исключение составляют работа И.П. Карповой (оценка эффективности обезличивания по методу перемешивания, 2013 г.) [18] и работа автора представленного исследования (обоснование, разработка, реализация, внедрение и оценка эффективности обезличивания по методу введения идентификаторов, 2018 г.) [19].

Таким образом, степень разработанности темы исследования для алгоритмов обезличивания по методу изменения состава/семантики и перемешивания является недостаточной с точки зрения этапов 1, 5 и 6, что позволяет сформулировать следующие цели и задачи.

Цели и задачи диссертационной работы. Целью диссертационной работы является разработка моделей процесса обезличивания ПД ФЛ и оценка эффективности реализации методов обезличивания.

Для достижения поставленной цели сформулированы следующие задачи:

1. Разработка модели идентификации ФЛ по отдельным атрибутам и их сочетаниям на основании количественных оценок вероятности идентификации для определения критерия необходимости обезличивания ПД.
2. Разработка модели нарушителя, реализующей алгоритм деобезличивания атрибутов ФЛ, обезличенных с помощью методов введения идентификаторов, изменения состава/семантики, перемешивания, для оценки эффективности реализации методов обезличивания ПД.
3. Разработка функциональной схемы процедуры обезличивания ПД для метода введения идентификаторов, обеспечивающей связь обезличенных ПД с

таблицей идентификаторов с использованием внешнего носителя идентификационной информации.

Объектом исследования являются системы защиты ПД в составе ИСПДн, реализованные методами обезличивания в соответствии с приказом Роскомнадзора [2].

Предметом исследования являются:

– зависимости вероятности идентификации ФЛ от семантики атрибутов ФЛ и их сочетаний при любом количестве записей (объеме) БД;

– алгоритмы действий нарушителя при деобезличивании ПД, обезличенных методами, основанными на искажении ПД;

– способы применения внешнего носителя идентификационной информации при реализации обезличивания ПД методом введения идентификаторов.

Научная новизна работы. В рамках проведенного исследования получены следующие новые научно обоснованные результаты:

1. Разработана математическая модель идентификации ФЛ, отличающаяся применением количественных оценок вероятности идентификации по атрибуту в целом и сформулирован критерий необходимости обезличивания ПД по любым идентификаторам или их совокупности при любом объеме БД (соответствует п.10 паспорта специальности «Модели и методы оценки защищенности информации и информационной безопасности объекта»).

2. Разработана функциональная модель нарушителя, реализующая итерационный алгоритм деобезличивания ПД для методов обезличивания, основанных на искажении ПД, и отличающаяся применением количественных оценок эффективности методов обезличивания (соответствует п.11 паспорта специальности «Модели и методы оценки эффективности систем (комплексов), средств и мер обеспечения информационной безопасности объектов защиты»).

3. Разработана функциональная схема передачи информации между базами ПД, разделенными методом введения идентификаторов, отличающаяся

применением внешнего носителя идентификационной информации и реализующая безопасную передачу информации между таблицей идентификаторов и обезличенными ПД (соответствует п.15 паспорта специальности «Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности»).

Теоретическая значимость работы заключается в развитии научно-методического аппарата для анализа процессов обезличивания ПД ФЛ и оценки эффективности реализации методов обезличивания. Теоретические положения, составляющие основу критерия необходимости обезличивания, могут использоваться для исследования баз ПД, которые имеют состав атрибутов и количество записей, отличающиеся от экспериментальной БД.

Практическая значимость работы заключается в:

1) возможности применения критерия необходимости обезличивания ПД для обоснования выбора обезличиваемых атрибутов ФЛ;

2) возможности использования регулирующими органами показателя вероятности идентификации по атрибуту ФЛ для создания нормативной базы параметров обезличивания ПД

3) возможности применения функциональной схемы передачи информации между частями базы ПД, разделенными методом введения идентификаторов, для построения системы защиты ИСПДн.

Методология и методы исследования. В диссертации представлены результаты исследования, полученные на основе функционального и математического моделирования с применением методов теории вероятностей и математической статистки. Модель идентификации разработана на основе частотного метода статистического анализа. Для подтверждения гипотез о характере зависимостей свойств атрибутов ФЛ от их семантики и от количества записей БД использован метод параметрической идентификации. Для

подтверждения оценки эффективности алгоритмов деобезличивания использован метод комбинаторного анализа.

Положения, выносимые на защиту:

1. Разработанная математическая модель идентификации ФЛ по произвольному атрибуту, использующая в качестве случайной величины количество записей в БД, содержащих любое искомое значение атрибута, устанавливает вид распределения этой случайной величины для исследованных атрибутов и их сочетаний, степенную зависимость вероятности идентификации от объема БД и определяет количественные, в т.ч. нормативные, значения вероятности идентификации ФЛ, а также условия критерия необходимости обезличивания ПД для БД произвольного объема.

2. Разработанная функциональная модель нарушителя, учитывающая количественные характеристики возможностей нарушителя и реализующая алгоритм деобезличивания ПД, основанный на поэлементном сравнении обезличенных атрибутов с имеющейся у нарушителя достоверной информацией об ограниченном количестве ФЛ, применима в качестве средства для определения оптимальных значений параметров методов изменения состава или семантики и перемешивания, а также количественной оценки эффективности этих методов.

3. Разработанная функциональная схема передачи информации обеспечивает безопасную передачу идентификаторов между разделенными частями базы ПД, обезличенной методом введения идентификаторов, путем применения внешнего носителя идентификационной информации; эффективность функциональной схемы подтверждена внедрением в сфере здравоохранения.

Достоверность полученных результатов обеспечивается использованием математических методов, адекватных задачам исследования, а также применимостью разработанных критериев для баз ПД различного объема и семантики.

Апробация работы. Различные аспекты выбора метода обезличивания и обезличиваемых идентификаторов, оценки эффективности обезличивания были апробированы на нескольких тематических конференциях:

– Научно-практическая конференция, посвященная 100-летию со дня рождения профессора Г.С. Черноруцкого и 75-летию ЮУрГУ «Актуальные проблемы автоматизации и управления» (Челябинск, 2013 г.);

– XVI Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2017» (Екатеринбург, 2017 г);

– 10-я научная конференция аспирантов и докторантов ЮУрГУ (Челябинск, 2018 г);

– XVII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2018» (Челябинск, 2018 г);

– XVIII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2019» (Магнитогорск, 2019 г);

– 12-я научная конференция аспирантов и докторантов ЮУрГУ (Челябинск, 2020 г);

– 2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) (Екатеринбург, 2021 г).

Публикации. Основные результаты исследования опубликованы в 13 печатных работах, 6 из которых – в рецензируемых журналах, определенных ВАК РФ и Аттестационным советом УрФУ, включая 1 статью в издании, индексируемом в международной наукометрической базе Scopus.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав основного материала, заключения, приложений, списка литературы.

Работа изложена на 165 страницах машинописного текста и включает 66 рисунков, 22 таблицы и 4 приложения. Список литературы содержит 93 наименования.

Во введении обоснована актуальность темы исследования, сформулированы цель, задачи, объект и предмет исследования, научная новизна, практическая ценность, выносимые на защиту результаты, степень достоверности и апробация работы.

В первой главе проведен анализ современного состояния разработанности и реализации методов обезличивания ПД. Определены принципы работы методов обезличивания в соответствии с нормативной базой РФ, их параметры и возникающие при их реализации проблемы. Отмечены сходство и различие проблем, а также значимость параметров для различных методов обезличивания ПД. Проведен анализ работ, описывающих существующие реализации методов обезличивания, их технологические особенности и возможности оценки эффективности. Определена цель исследования: разработка моделей процесса обезличивания ПД ФЛ и оценка эффективности реализации методов обезличивания. Приведено описание функциональной схемы взаимодействия частей базы ПД, разделенных методом введения идентификаторов, и результатов ее применения.

Во второй главе сформулирован критерий необходимости обезличивания ПД ФЛ по атрибутам, основанный на применении модели идентификации ФЛ и модели нарушителя безопасности ПД. Модель идентификации основана на вычислении вероятности идентификации ФЛ по атрибуту в целом для различных значимых атрибутов (идентификаторов) в зависимости от их синтаксиса и семантики, а также от количества записей БД. Сформулирована задача создания модели идентификации ФЛ. Достоверность полученных результатов обеспечена использованием известных методов теории вероятностей и математической статистики, адекватных задачам исследования, значительным объемом

исследуемых БД, согласованностью параметров модели идентификации ФЛ для БД различной структуры.

В третьей главе предложена модель нарушителя, реализующая алгоритм деобезличивания, и методика определения эффективности обезличивания для различных искажающих методов. К искажающим, в соответствии с [2], относятся метод изменения состава или семантики и метод перемешивания. Алгоритм деобезличивания реализован в виде автоматизированного поиска известных нарушителю неискаженных атрибутов в обезличенной базе и неавтоматизированного (вручную) составления таблицы смещений исходных элементов идентификаторов в конечное положение. Сформулирована и решена задача реализации модели нарушителя для оценки эффективности методов обезличивания. Разработана модель нарушителя и методика оценки эффективности применительно к искажающим методам обезличивания с одинаковым алгоритмом искажения для всех обезличиваемых элементов. Разработанная методика оценки эффективности применима для формулирования рекомендаций нормативных значений вероятности идентификации для обезличиваемых идентификаторов. Для обеспечения достоверности полученных результатов применены методы теории вероятностей и математической статистики, адекватные задачам исследования. При проведении экспериментов использованы ПД из базы с большим количеством записей.

В четвертой главе предложена функциональная схема передачи информации между частями базы ПД, разделенными методом введения идентификаторов, основанная на применении внешнего носителя идентификационной информации для связи обезличенных ПД с таблицей перекрестных ссылок. Предлагаемая схема представлена в виде функциональной модели, реализованной с использованием внешнего бумажного носителя, содержащего идентификатор ФЛ в виде штрих-кода. Описана реализация функциональной схемы взаимодействия частей базы ПД, разделенных методом введения идентификаторов, в сфере здравоохранения.

В заключении подведены итоги диссертационной работы, приведены основные результаты и сформулированы направления дальнейших исследований.

В приложениях приведены дополнительные материалы, подтверждающие результаты исследований.

Глава 1. Анализ современного состояния применения методов обезличивания персональных данных

1.1. Состояние разработанности методов обезличивания в нормативных документах по защите персональных данных

Активные методы защиты информации основаны на использовании сертифицированных СЗИ, интегрированных в операционную среду обработки ПД. Стоимость применения таких СЗИ пропорциональна количеству рабочих мест. Обезличивание ПД имеет главной целью обесценивание попыток злоумышленника использовать открытую информацию во вред физическому лицу. Поскольку этот метод защиты не требует применения СЗИ, достигается еще одна цель – значительная экономия средств.

Учитывая необходимость обработки ПД с научными и иными целями в общем доступе, российское законодательство, как и законодательство США и ЕС, предъявляет жесткие требования к защите ПД, поэтому задача обезличивания ПД является актуальной. Главная особенность российского законодательства заключается в том, что, в соответствии с [1], обезличивание не позволяет определить принадлежность ПД физическому лицу без применения дополнительной информации, поэтому наиболее важным свойством обезличенных данных является возможность их деобезличивания, а отсутствие такой возможности признается лишь как частный случай. Таким образом, в России для процесса обезличивания одинаково актуальны прямая и обратная задачи.

Начиная с 2009 года, в России было предложено множество алгоритмов реализации процесса обезличивания, но только в 2013 году приказом Роскомнадзора [2] были определены, а в Методических рекомендациях [17] подробно описаны четыре основных метода обезличивания: введение идентификаторов, изменение состава или семантики, декомпозиция и перемешивание (Табл.1).

Таблица 1

Методы обезличивания ПД

Метод	Принцип работы	Секрет	Проблемы
Введение идентификаторов	– Группа идентифицирующих атрибутов заменяется абстрактным идентификатором – Группа хранится в отдельной таблице	Таблица перекрестных ссылок	– Выбор состава идентифицирующей группы – Генерация идентификатора – Обеспечение связи между таблицей и обезличенными данными
Изменение состава или семантики	Изменяется структура (количество, положение и размер полей) или изменяется значение идентифицирующих атрибутов внутри строки	Алгоритм модификации	– Выбор состава идентифицирующей группы – Генерация алгоритма модификации – Обеспечение секретности алгоритма модификации
Декомпозиция	База данных разделяется на много частей, информация о связях хранится в отдельной таблице	Таблица связей	– Выбор состава частей – Генерация алгоритма связывания – Обеспечение связи между частями
Перемешивание	Идентифицирующие атрибуты перемещаются в другие записи внутри группы записей	Алгоритм перемещения	– Выбор состава идентифицирующей группы – Генерация алгоритма перемещения – Обеспечение секретности алгоритма перемещения

Из таблицы следует, что:

– общей проблемой для всех методов является выбор группы идентифицирующих атрибутов в зависимости от количества записей БД, поскольку отсутствует методика количественной оценки влияния атрибута на процесс обезличивания ПД;

– методы изменения состава или семантики и перемешивания (искажающие методы) основаны на скрытии расположения идентифицирующей информации в массиве обезличенных данных, поэтому для них общей проблемой является

отсутствие методики количественной оценки защищенности алгоритма, поскольку искаженная идентифицирующая часть является общедоступной в любых режимах работы;

– методы введения идентификаторов и декомпозиции (разделяющие методы) основаны на отделении идентифицирующей информации от обезличенных данных, поэтому для них общей проблемой является обеспечение связи разделенных частей во время сеанса работы. При этом идентифицирующая часть недоступна для злоумышленника во время хранения, но может быть доступна во время прочих сеансов обработки (ввод, вывод). Следует отметить, что с точки зрения реализации алгоритмы этих двух методов принципиально очень близки, поэтому далее они рассматриваются совместно;

– общей проблемой, важной с точки зрения затрат, является необходимость модификации структуры БД и прикладного программного обеспечения для реализации метода обезличивания. В случае разработки программного обеспечения сторонней организацией требуемая модификация может оказаться экономически невыгодной и/или технически невозможной.

Рассмотрим существующие методики выбора атрибутов для обезличивания и варианты реализации методов обезличивания ПД с учетом имеющегося опыта их внедрения.

1.2. Процедура выбора атрибутов физического лица для обезличивания персональных данных

Выбор атрибутов ФЛ для процесса обезличивания не зависит от метода обезличивания. В отличие от законодательства РФ, которое никак не регламентирует использование конкретных атрибутов, зарубежные нормативные акты [20, 21], наоборот, не предъявляют требований к методам обезличивания, но уделяют особое внимание именно атрибутам. В частности, документ США «Руководство по защите конфиденциальности персональной информации» [20] не только классифицирует ПД по различным критериям (чувствительные/не чувствительные и т.п.), но и приводит прямое перечисление всех видов ПД,

которые подлежат обезличиванию. Однако для обоснования значимости атрибутов предлагаются экспертные оценки возможного ущерба, имеющие не количественный, а качественный характер.

Несмотря на то, что большинство зарубежных исследователей [3 – 5], [22 – 32] и ряд российских ученых [33 – 44] занимаются проблемой невосстановимого обезличивания ПД с целью его безопасного использования в общем доступе (анонимизация), ими были получены интересные результаты в плане выбора атрибутов для обезличивания, а также критериев и методик оценки результатов обезличивания и возможности косвенного восстановления ПД. Например, в связанных между собой работах [3] и [4], посвященных анонимизации больших массивов данных (SetValued Data), вводится понятие квази-идентификаторов для обработки сочетаний атрибутов, а также понятие степени потери информации (Generalized Loss Metric or Normalized Certainty Penalty – NCP), которое является критерием эффективности обезличивания и концептуальным аналогом понятия «вероятность идентификации», рассмотренного далее в Главе 2. Но наиболее близким количественным аналогом вероятности идентификации является термин «вероятность нарушения» (Breach Probability), введенный и вычисленный авторами работы [5].

Для выяснения закона частотного распределения характеристик атрибутов проведен анализ отечественных и зарубежных статистических исследований. В зарубежных источниках наиболее полные исследования на эту тему производились в США на базе большого количества различных типов данных. Для целей защиты ПД представляет интерес исследование частотного распределения количества населенных пунктов и фамилий в зависимости от количества ФЛ. В работах [45, 46] демонстрируется степенной характер $y(x) = Cx^{-a}$ закона распределения этих и ряда других величин. Для степенного закона характерно представление в виде прямой линии в логарифмическом масштабе по обеим координатам, что дает возможность предположить его наличие, не производя дополнительных расчетов. В работе [46] показано, что

наглядность представления в логарифмическом масштабе не гарантирует наличие степенного закона, и точность представления должна быть дополнительно обоснована. Автор работы [45] допускает, что для неанглоязычных стран с другими демографическими и языковыми особенностями указанные зависимости могут не соблюдаться. Поэтому для конкретных стран и языков требуются дополнительные исследования. В России подобные исследования [47 – 50] проводились на предмет частотного распределения количества городов в зависимости от численности населения. В работе [47] показано, что для российских населенных пунктов степенной закон соблюдается с высокой точностью.

Большое количество исследований в области медицины и социологии [51 – 53] посвящено изучению зависимостей различных характеристик ФЛ от их возраста. Но в большинстве случаев авторы ограничиваются годом рождения, либо месяцем рождения независимо от года (сезонные зависимости). Т.е. количество различных значений дат является целым числом либо от 0 до 100 (лет), либо от 1 до 12 (месяцев).

1.3. Способы реализации методов обезличивания персональных данных в России и за рубежом

Основанием для анализа состояния проблемы является исследование имеющихся научных публикаций по теме обезличивания ПД в России и за рубежом, а также собственные разработки автора. Все направления исследований можно условно разделить на четыре группы по применяемым методам обезличивания. Условность разделения возникает из-за различий терминологии законодательных актов в России и за рубежом.

1. Способы реализации метода введения идентификаторов. В зарубежных источниках достаточно полным аналогом этого метода является псевдонимизация (Pseudonymization), которая предлагает различные (в том числе криптографические) способы создания уникальных идентификаторов для связи с таблицей перекрестных ссылок и подробно описана в работах [22] и [23]. В

российских источниках эта группа способов представлена работами [54 – 63], часть которых рассматривает различные криптографические методы генерации идентификатора связи между таблицей перекрестных ссылок и обезличенной базой. Например, уникальный и релевантный идентификатор ФЛ получен путем использования односторонней криптографической функции от атрибутов фамилии, имени, отчества и даты рождения ФЛ О.А. Вишняковой и Д.Н. Лавровым [54]. Решение вопросов связи и безопасности сеансов обработки данных остаются за рамками этой и подобных работ.

В 2012 году Е.С. Волокитиной на основе патента [7] на способ идентификации субъекта ПД внедрен метод введения идентификаторов в сфере образования, более подробно представленный автором патента в работе [8]. В качестве идентификатора связи использовались сим-карты, установленные в мобильные телефоны. Схема реализации с использованием внешнего идентификатора представлена на рис. 1.

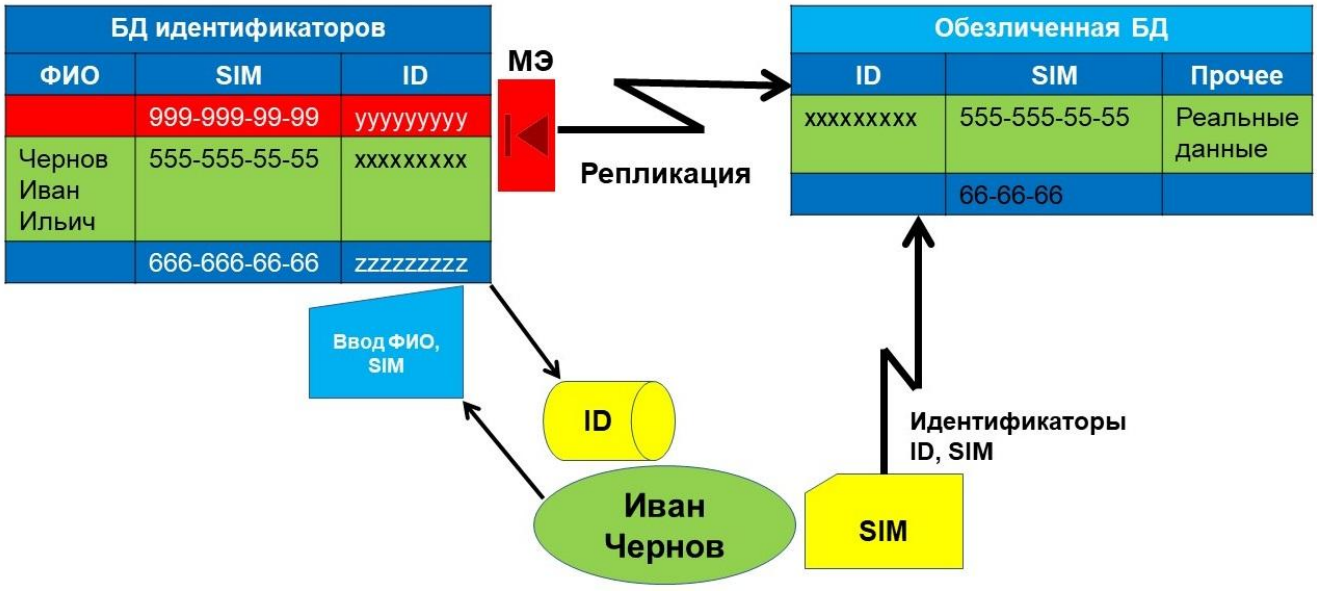

Рис. 1. Схема реализации с использованием SIM-карты

Представленный алгоритм успешно решает проблему безопасности во время сеансов обработки обезличенных данных, но использование дополнительного идентификатора усложняет процесс обработки и повышает затраты на эксплуатацию.

Для обеспечения экономической эффективности в 2010 году Д.Н. Ивановым и автором представленной работы предложена полезная модель, получившая патент [64] и внедренная на предприятиях сферы здравоохранения.

Запатентованный метод в качестве идентификатора связи предполагает использование бумажного носителя – бланка рецепта на получение лекарств. Предложена функциональная схема, которая обеспечивает безопасную обработку на всех этапах процесса обработки ПД и подробно описана в Главе 4.

А.А. Халафян и А.А. Кошкаров [10] в качестве идентификатора связи также предлагают использовать бланк рецепта на получение лекарств, но обезличенная БД хранится с применением облачных технологий. Доступ к ней обеспечивается с использованием каналов Интернет, что позволяет централизовать и ускорить обработку данных, но при этом на рабочих местах во время сеансов данные не обезличены, используются и идентификаторы, и чувствительные атрибуты. Таким образом, все рабочие места требуют установки СКЗИ и межсетевого экранирования, что не позволяет добиться снижения эксплуатационных расходов в аптечных пунктах. Тем не менее, этот метод был внедрен в распределенную информационную систему сферы здравоохранения в 2015 году.

В работах Н.В. Журилова и др. [55] и С.Е. Раузиной и др. [56] приведено описание применения электронных рецептов, внедренного в системах льготного лекарственного обеспечения Московской, Белгородской и др. областей России, где в качестве внешнего носителя используются средства мобильной связи со специальным программным обеспечением. Предметом исследования в этих работах является технология, но не информационная безопасность процессов обработки ПД. Тем не менее, внешние носители этого типа могут быть использованы в схеме полезной модели [64] при соблюдении требований защиты информации.

2. Способы реализации метода изменения состава или семантики. В зарубежных источниках достаточно близким аналогом этого метода является перестановка символов (Character Scrambling), в которой в качестве области

перестановки рассматривается не отдельный атрибут, а полная строка атрибутов-идентификаторов. Кроме того, в работе [22] в качестве объекта перестановки рассматривается не только символ, но и бит.

В Российской Федерации в соответствии с приказом [2], метод изменения состава и/или семантики предполагает внесение обратимых изменений (искажений) в идентифицирующие атрибуты каждой записи ПД без использования информации из других записей базы. При этом алгоритм искажений (замена символов на другие символы или перемещение символов внутри строки) не регламентируется. Изменения могут производиться по таблице подстановки или формуле смещения, при этом таблица/формула необязательно должна быть одинаковой для разных записей базы. Это направление представлено работами [6, 65], в которых предложен способ кодирования идентифицирующих атрибутов на базе разработанного алгоритма.

Отличительной особенностью работы И.Ю. Кучина [6] является аналитическое обоснование выбора состава идентифицирующей группы и обеспечение заданной степени анонимности в обезличенной базе. Предложенный алгоритм кодирует каждую запись базы отдельным кодом с использованием значений идентификаторов и параметров операционной системы. Этот алгоритм внедрен в сфере страхования, но он охватывает процесс хранения небольшого количества записей БД и только в СУБД в среде конкретной операционной системы.

В работе Серышева А. С. [65], внедренной в сфере общественного питания, предложено кодирование чувствительных атрибутов словарем.

3. Способы реализации метода перемешивания. В зарубежных источниках достаточно полный аналог этого метода имеет подобное название (Shuffling) и функции и упоминается во многих работах, в том числе в [22] и [23].

В Российской Федерации в соответствии с приказом [2] метод перемешивания состоит в обратимом изменении (искажении) положения идентифицирующих атрибутов каждой записи путем перемещения их в другие

записи, но с сохранением семантики перемещаемых атрибутов (их положения в строке). При этом алгоритм искажений не регламентируется, а группа записей, внутри которой производится перемещение, может быть произвольной по количеству. Различные группы записей могут отличаться друг от друга по размеру. Изменения могут производиться по таблице подстановки или формуле смещения, при этом таблица или формула смещения необязательно должна быть одинаковой для разных групп записей базы. Это направление представлено работами [9], [12 – 14] и [66 – 70], которые предлагают использование алгоритмов перемешивания, ориентированных на хранение ПД, либо на их передачу по открытым каналам связи.

Например, среди внедренных способов (в сфере здравоохранения) можно отметить алгоритм перемешивания, описанный в работе М.И. Денисова и К.А. Чехонина [9], где обезличенная база используется только внутри контролируемой зоны, а для передачи данных используются СКЗИ.

В.В. Воронин и Н.Л. Нехай [12] разработали алгоритм с многочисленными циклическими сдвигами в подмножествах каждого чувствительного атрибута и внедрили его в систему учета клиентов и работников предприятия по обслуживанию автотранспорта. Для хранения параметров перемешивания используются СКЗИ, что повышает затраты на систему защиты. В предложенном алгоритме перемешивание осуществляется не в одинаковых сегментах базы, а в группах записей разной величины для различных идентификаторов. Этот подход можно назвать несегментированным.

В других разработанных алгоритмах, рассчитанных на БД с большим количеством записей, используется сегментированный подход, то есть предварительное разделение базы на равные сегменты, в границах которых и производится перемешивание. Например, в работе К.О. Бондаренко и В.А. Козлова [13], размер сегмента составляет 256 записей, в сегменте сначала перемешиваются полные строки, а затем идентификаторы между строками. Используются таблицы подстановки, сгенерированные криптографическим

методом гаммирования [71]. Применение криптографии, с одной стороны, гарантирует защиту, даже во время сеанса обработки, но, с другой стороны, усложняет процесс добавления/удаления/поиска данных и повышает затраты на их защиту. Эти недостатки являются препятствием для внедрения предложенного способа обезличивания.

4. Прочие способы обезличивания. В зарубежных источниках прочие направления исследований имеют своей целью невосстановимое искажение базы ПД, например, из 11 методов обезличивания, приведенных в работах [20] и [22], все, кроме 3-х методов, рассмотренных выше в пп. 1 – 3, вносят невосстановимые искажения и поэтому не имеют реализаций в РФ. Российские источники [15 – 19], [33, 72] предполагают, в основном, использование криптографических методов (off-line либо on-line обезличивание с помощью инструментов БД), которые могут быть отнесены к обезличиванию с достаточной степенью условности, поскольку обеспечивают и невозможность идентификации ФЛ по обработанным данным, и их обратное восстановление, но формально не входят в набор методов, установленных Роскомнадзором в [2], либо используют эти методы частично. Например, работа Ю.В. Трифоновой и Р.Ф. Жаринова [15] предлагает использование встроенных криптографических средств СУБД CryptDB. В качестве примера частичного использования метода идентификаторов можно привести работу И.М. Ажмухамедова, Р.Ю. Деминой и И.В. Сафарова [16], где применено шифрование таблицы перекрестных ссылок с последующим блокированием.

1.4. Оценки эффективности реализации методов обезличивания персональных данных

Оценку эффективности реализации методов обезличивания можно проводить в различных аспектах, важных с точки зрения внедрения: технологическом (производительность), функциональном (защищенность), экономическом (окупаемость). Универсальная методика количественной оценки

эффективности обезличивания отсутствует, что затрудняет сравнение процедур обезличивания данных, осуществляемого различными методами.

С целью решения проблемы в [73] предложены методики расчета эффективности обезличивания данных с использованием показателей вероятности идентификации и степени обезличивания для методов введения идентификаторов [74], изменения состава или семантики [75] и перемешивания [76]. Реализация метода введения идентификаторов в сфере здравоохранения на базе методики [74] и полезной модели [64] подробно описана в Главе 4.

Опыт внедрения методов обезличивания в России за последние 5 лет нельзя назвать достаточным, тем не менее, он позволяет оценить некоторые тенденции. Наиболее важной представляется зависимость эффективности обезличивания от метода обезличивания и количества записей БД. Например, в работе И.П. Карповой [18] приведены расчеты производительности алгоритма, реализующего метод перемешивания с несегментированным подходом, из которых можно сделать вывод о нецелесообразности его применения для БД с большим количеством записей. Причина состоит в необходимости деобезличивания всей базы для возможности внесения даже небольших изменений.

Результаты, полученные автором представленной работы с использованием методической базы оценки функциональной эффективности, позволили сделать вывод, что наиболее эффективным с точки зрения обеспечения безопасности ПД является метод введения идентификаторов.

Надежные оценки экономической эффективности внедрений сделать сложно, поскольку цель получения экономического эффекта преследовалась не во всех случаях. Для более широкого внедрения различных методов обезличивания ПД в работах автора представленной работы [77, 78] предложено оценивать эффективность обезличивания ПД как трудоемкость их деобезличивания нарушителем, действующим в рамках специально разработанной модели нарушителя. Разработка моделей нарушителя безопасности ПД описана в работах ряда российских исследователей, например, модель нарушителя для медицинских

учреждений предложена в работе Баранковой И.И. [79], но в ней используются только качественные показатели. Модель нарушителя, предложенная в [77], имеет итерационный характер, трудоемкость деобезличивания можно оценить количественно числом необходимых итераций. Предполагается, что предложенная количественная оценка позволит по мере накопления опытных данных сформировать соответствующую нормативную базу.

1.5. Постановка задач исследования

Методы обезличивания ПД имеют явное преимущество перед активными методами защиты с применением СЗИ, а также перед криптографическими методами, поскольку сложность их реализации не зависит ни от количества рабочих мест, ни от степени защищенности каналов передачи данных. Явным недостатком методов обезличивания является сложность обеспечения безопасности дополнительной информации, необходимой для обработки обезличенных данных, на некоторых этапах процесса обработки информации. Методы обратимого обезличивания, предусмотренные российскими нормативными актами [2, 16], можно сгруппировать следующим образом:

1) методы, использующие искажение идентификаторов с сохранением их в составе обезличенных данных;

2) методы, использующие разделение идентификаторов и обезличенных данных на разные хранилища.

Для обеих групп методов обезличивания актуальна проблема выбора атрибутов для обезличивания (идентификаторов), например, «фамилия», «имя», «дата рождения» и т.д. Для решения этой проблемы сформулирована задача создания модели идентификации ФЛ: требуется количественно оценить возможность идентификации ФЛ по каждому атрибуту и по некоторым совокупностям атрибутов. При этом атрибуты с наибольшими значениями вероятности идентификации войдут в группу идентификаторов, подлежащих дальнейшей обработке в рамках обезличивания. Поскольку состав группы атрибутов также зависит от количества записей БД, результаты расчета

вероятности идентификации должны позволить сделать вывод об этой зависимости для каждого атрибута.

Для первой группы методов основная проблема реализации – оценка эффективности алгоритма искажения и его защищенности. Для решения этой проблемы сформулирована задача разработки модели нарушителя: требуется сформировать таблицу смещений элементов идентификаторов, используя достаточное для неавтоматизированной обработки количество записей, известных нарушителю ФЛ, заведомо входящих в обезличенную базу. С точки зрения нарушителя, условием полного деобезличивания является совпадение значений всех идентификаторов всех ФЛ, известных нарушителю, со значениями, полученными в процессе восстановления искаженных значений этих идентификаторов. Условием частичного деобезличивания является совпадение значений части идентификаторов ФЛ, известных нарушителю, в соответствии с параметром целесообразности дальнейшей обработки, заложенным в модель.

Для второй группы методов основная проблема реализации – обеспечение защищенной передачи информации между разделенными частями БД. Для решения этой проблемы сформулирована задача реализации функциональной схемы: требуется обезличить базу ПД методом введения идентификаторов и обеспечить безопасную передачу идентификаторов между удаленной таблицей перекрестных ссылок и базой обезличенных данных без добавления в процесс обработки данных дополнительных внешних носителей и СКЗИ.

Глава 2. Критерий необходимости обезличивания персональных данных

Задачей этой главы является формулирование критерия необходимости обезличивания ФЛ по атрибуту, либо по сочетанию атрибутов. Критерий необходимости обезличивания должен учитывать характеристики атрибута, количество записей БД и возможности нарушителя. В соответствии с этим критерием, по каждому исследованному атрибуту, либо по сочетанию атрибутов, предлагается решение о необходимости обезличивания для БД. Кроме того, для некоторых атрибутов рассмотрена зависимость принятия решения от количества записей БД.

2.1. Процесс идентификации физического лица. Общая модель нарушителя

В соответствии с федеральным законом [1], ПД – это любая информация, относящаяся к ФЛ. Иными словами, если информация относится к ФЛ, то это его ПД, и они этому лицу принадлежат. Любая обработка ПД включает процесс проверки отношения ПД к ФЛ, называемый идентификацией. Для описания процесса идентификации используем следующие понятия, введенные в работах [80 – 82]:

1) База данных B – набор информации о ФЛ, состоящий из записей L_i:

$$B = \{L_1, L_2, \ldots L_i, \ldots L_V\},$$
$$L_i = \{A_1, A_2, \ldots A_j, \ldots A_K\} \quad (1)$$

где: $i = 1,\ldots, V$ – номер записи;

V – количество записей БД (объем базы);

A_j – атрибут ФЛ (поле БД);

$J = 1,\ldots, K$ – номер атрибута;

K – количество атрибутов ФЛ в БД.

В общем случае в базе содержится несколько записей об одном ФЛ, но для решения задачи идентификации будем полагать, что в базе существует только одна запись об одном ФЛ, то есть V – количество ФЛ.

2) Обезличенная БД B' состоит из записей L_i' и по количеству записей V совпадает с базой B:

$$B' = \{L_1', L_2', \ldots L_i', \ldots L_V'\},$$

$$L_i' = \{A_1', A_2', \ldots A_j', \ldots A_K'\},$$

где A_j' – обезличенный атрибут ФЛ;

$j = 1,\ldots, K'$ – номер атрибута в обезличенной БД;

K' – количество атрибутов ФЛ в обезличенной БД.

Но количество атрибутов K' в записи L_i' зависит от метода обезличивания: для разделяющих методов $K' < K$, для искажающих методов $K' = K$.

3) Маркер поиска R – набор информации об определяемом ФЛ (известные атрибуты неизвестного человека, которого надо идентифицировать), состоящий из одной записи:

$$R = \{A_1, A_2, \ldots A_j, \ldots A_M\},$$

где A_j – искомое значение атрибута ФЛ;

$j = 1, \ldots, M'$– номер искомого атрибута;

M – количество атрибутов поиска ($M < K$).

Набор R задает цель поиска. В общем случае целью может быть идентификация группы ФЛ, но далее в качестве цели поиска рассматривается только одно ФЛ.

Под идентификацией подразумевается процесс сравнения двух различных наборов информации (один – заданный маркером R, другой – запись L_i из БД поиска B или B') с целью выявления их однозначного соответствия друг другу. В общем случае атрибут маркера A_1 семантически не совпадает с атрибутом записи базы A_1. Положим, что маркер R нормирован так, что каждому его атрибуту найден семантически соответствующий атрибут базы (иначе идентификация невозможна). Если ПД определяемого ФЛ содержатся в БД, то для каждой записи БД возможны два варианта результатов сравнения с маркером R:

1) значения всех сравниваемых атрибутов набора R совпадают с атрибутами одной записи L_i набора B: в этом случае принимается гипотеза принадлежности всех атрибутов записи L_i определяемому ФЛ. Если сравнение с другими записями базы B покажет, что таких совпадений будет более одного (есть совпадения в нескольких записях L_i), то сравниваемые атрибуты набора R

принадлежат многим лицам, и гипотеза принадлежности всех атрибутов записи L_i определяемому ФЛ принимается с определенной вероятностью. Это означает, что набор, заданный маркером, недостаточен для идентификации;

2) не все или никакие сравниваемые атрибуты набора R не совпадают с атрибутами ни одной записи L_i набора B: в этом случае отвергается гипотеза принадлежности всех атрибутов записи L_i определяемому ФЛ. Если сравнение со всеми записями БД покажет, что совпадения отсутствуют, то это означает, что заданный маркером набор избыточен для идентификации.

Понятия идентификации и деобезличивания отличаются тем, кто является субъектом процесса поиска. Если поиск ПД осуществляется в необезличенной базе B, то субъектом поиска является санкционированный пользователь, а процесс сравнения атрибутов наборов R и B является идентификацией. При этом ожидается безусловный вариант результата поиска (1), а результат (2) считается ошибочным.

Если поиск ПД осуществляется в обезличенной базе B', то субъектом поиска является несанкционированный пользователь – нарушитель, а процесс сравнения атрибутов наборов R и B' является деобезличиванием. При этом ожидается результат поиска (2) или вероятный вариант результата поиска (1). В этом состоит цель процедуры обезличивания в соответствии с федеральным законом [1].

Результат поиска в обезличенной базе B' зависит как от вероятности идентификации по атрибутам, так и от возможностей нарушителя. Для построения общей модели нарушителя определены следующие условия, которые не зависят от метода обезличивания:

1) нарушитель имеет неограниченный доступ к полному объему обезличенной базы B', так как в соответствии с законом [1] обезличенные данные не являются ПД и не подлежат защите;

2) нарушитель знает структуру (размер полей и записи в целом) исходной базы B, так как в соответствии с требованием обратимости приказа [2] размер полей при обезличивании сохраняется;

3) нарушитель знает, что обработке (разделению или искажению) подверглись только идентификаторы, т.е. прочие данные не искажены;

4) нарушитель не знает секрета обезличивания базы (алгоритма формирования идентификатора связи – для разделяющих методов, алгоритма модификации – для искажающих методов), за исключением того, что этот алгоритм является единым для всех записей базы, в соответствии с определением в законе [1];

5) нарушитель знает некоторые ПД определяемого ФЛ и хочет узнать другие идентификаторы и прочие ПД этого лица;

6) нарушитель знает, что в базе есть ПД (как идентификаторы, так и прочие данные) определенного количества известных ему лиц, причем прочие (неискаженные) данные он может найти в базе;

7) нарушитель планирует, используя известные данные (п.6) и соответствующий алгоритм деобезличивания, вычислить алгоритм обезличивания (п.4) и получить в результате поиска ограниченную группу записей ФЛ для дальнейшего уточнения организационными методами;

8) нарушитель не может однозначно выбрать определяемое ФЛ из группы записей, которую он получил в результате поиска (п.7), если размер группы превышает некоторое количество разных ФЛ;

9) нарушитель не может разрабатывать специальное программное обеспечение для вскрытия алгоритма искажений, но может пользоваться готовыми программными средствами (например, для поиска или сравнения символов).

Таким образом, параметрами модели идентификации являются:

– G – количество лиц, ПД которых заведомо известны нарушителю (п.6);

– U – параметр целесообразности действий нарушителя, определяющий максимальное количество найденных записей, при котором целесообразна их дальнейшая обработка (п.8.);

– алгоритм деобезличивания, применяемый нарушителем (п.7).

Указанные параметры могут быть использованы для классификации уровня компетенции нарушителя.

Если нарушитель обладает дополнительной информацией, например, знает метод обезличивания, параметры модели следует уточнить.

2.2. Вероятность идентификации физического лица и критерий необходимости обезличивания по атрибуту

Для количественной оценки значимости каждого атрибута в процессе идентификации использовано понятие вероятности идентификации ФЛ по конкретному значению этого атрибута, введенное в работе [82]. Атрибут базы A_j имеет набор значений, которые составляют множество:

$$A_j = \{A_{j1}, \ldots, A_{jk}, \ldots A_{jQ}\},$$

где j – номер атрибута, определенный в п.2.1;

$k = 1, \ldots, Q$ – номер значения атрибута;

Q – количество различных значений атрибута.

Если в качестве случайной величины принять значение A_{jk} атрибута A_j ФЛ (например, значение атрибута A_1, равное «Мищенко»), то факт принадлежности этого значения атрибута автору представленного исследования согласно (1) означает, что значение входит в состав записи $L_{\text{ФЛ}}$, соответствующей этому лицу в БД, а вероятность этого события определяется как

$$p(A_{jk} \in L_{\text{ФЛ}}) = 1/V, \quad (2)$$

где V – количество записей (объем) БД.

Формула верна для любого атрибута.

Если представить это событие как последовательность двух событий – выявление ФЛ в группе записей с одинаковым значением атрибута «фамилия» (событие X) после того, как эта группа была выбрана из базы (событие Y), то его вероятность определится как

$$p(XY) = p(X|Y) \cdot p(Y). \quad (3)$$

Согласно модели нарушителя, интерес с точки зрения идентификации представляет событие X и его вероятность, определяемая как

$$p(X|Y) = 1 / q_{jk}, \quad (4)$$

где q_{jk} – количество записей в базе, для которых значение атрибута A_j равно A_{jk}, что соответствует (2).

Так как для разных атрибутов количество различных значений Q отличается, для j-го атрибута использовано обозначение Q_j. При этом для любого атрибута выполняется условие $Q_j < V$, а (4) можно записать как

$$p(A_{jk} \in L_{ФЛ}) = 1 / q_{jk}, \quad (5)$$

то есть, вероятность идентификации ФЛ по значению атрибута A_{jk} – величина, обратная количеству записей q_{jk} с данным значением атрибута, найденных в БД.

Минимальное значение $q_{jmin} = 1$, а максимальное значение для одного A_{jk} будет в случае, если для всех остальных значений $(Q_j – 1)$ этого атрибута $q_{jk} = 1$. Максимальное значение можно вычислить как

$$q_{jmax} = V – Q_j + 1.$$

Максимальное значение $p(A_{jk} \in L_{ФЛ}) = 1$ может быть гарантировано только в случае, когда в базе ФЛ найдена ровно одна запись. Если же таких записей найдено несколько, то это означает, что такое же значение атрибута имеют несколько ФЛ, поэтому $p(A_{jk} \in L_{ФЛ})$ будет меньше 1.

Значение $p(A_{jk} \in L_{ФЛ}) = 0$ соответствует ситуации, когда невозможно сопоставить ПД из базы определяемому ФЛ, что может произойти, если с данным значением атрибута не сопоставлено ни одной записи.

Эксперимент показал, что наиболее вероятна ситуация, когда конкретному ФЛ можно сопоставить ПД нескольких «прочих» ФЛ. Чем больше «прочих» ФЛ, тем выше эффективность обезличивания, при этом значение $p(A_{jk} \in L_{ФЛ})$ обратно пропорционально количеству «прочих» ФЛ и при его увеличении будет стремиться к значению «0», но никогда его не достигнет, поскольку максимально возможное количество найденных записей – это полное количество ФЛ в базе.

Если для идентификации используется M атрибутов, то результатом поиска является M множеств записей размером q_{jk}, а общей характеристикой поиска будет совокупность записей множества, являющегося пересечением всех M множеств и имеющего размер q_{int}. Нижней границей является $q_{int} = 0$ (ни одной записи не найдено для всего набора атрибутов маркера). Оптимальным является $q_{int} = 1$ (найдена ровно одна запись) – идентификация считается успешной. Если найдено несколько записей ($q_{int} > 1$) – идентификация считается вероятной.

Задача выбора идентификаторов для процедуры обезличивания требует расчета некоторого критерия, подобного $p(A_{jk} \in L_{ФЛ})$, но не по конкретному значению идентификатора, а по некоторой общей величине, характеризующей всё множество значений атрибута. Для этого в соответствии с (5) предложен аналог q_{jk}, характеризующий атрибут в целом.

Идея предлагаемой модели обезличивания состоит в том, что в качестве случайной величины принято значение q_j – количество записей, найденных нарушителем в БД, содержащих любое искомое значение j-го атрибута, при этом значение q_{jk} – это количество записей в БД, для которых j-й атрибут принимает одинаковые значения A_{jk}, где $k = 1, \ldots, Q_j$, Q_j – количество различных значений j-го атрибута в БД. В случае с атрибутом «фамилия» – это количество однофамильцев.

Далее определяется $p(q_j = q_{jk})$ – вероятность того, что произвольное ФЛ попадает в эту группу. Если значение q_{jk} уникально, то вероятность этого события

$$p(q_j = q_{jk}) = q_{jk} / V.$$

В общем случае q_{jk} может повторяться некоторое количество раз n_{jk} (например, 2 группы по 100 фамилий «Иванов» и «Петров»), при этом

$$p(q_j = q_{jk}) = q_{jk} \cdot n_{jk} / V.$$

В этом случае

$$V = \sum_{k=1}^{H} q_{jk} \cdot n_{jk}, \quad (5)$$

$$Q_j = \sum_{k=1}^{H} n_{jk}, \quad (6)$$

где H – количество различных значений q_{jk} и n_{jk}.

В разработанной модели, аналогично (3), вероятность идентификации ФЛ $p_{jk} = p(XY)$ зависит не только от размера группы q_{jk}, но и от количества групп одинакового размера n_{jk}, и определяется выражением

$$p_{jk} = p(A_{jk} \in L_{ФЛ}) \cdot p(q_j = q_{jk}) / n_{jk},$$

где $p(Y) = p(q_j = q_{jk}) / n_{jk}$, $p(X\backslash Y) = p(A_{jk} \in L_{ФЛ})$.

Необходимый для оценки значимости атрибута A_j общий показатель должен учитывать все значения этого атрибута из набора

$$A_j = \{A_{j1}, \ldots, A_{jk}, \ldots A_{jQ_j}\},$$

где $k = 1,\ldots,Q_j$ – номер значения атрибута в наборе различных значений;

Q_j – количество различных значений j-го атрибута.

В разработанной модели этот показатель определен как вероятность идентификации произвольного ФЛ по атрибуту A_j в целом по базе, и вычисляется по формуле:

$$W_j = \sum_k p_{jk} = \sum_k p(A_{jk} \in L_{ФЛ}) \, p(q_j = q_{jk}) / n_{jk} = Q_j / V \qquad (7)$$

где W_j – вероятность идентификации по атрибуту A_j.

Минимальное значение $Q_j = 1$ означает, что все значения атрибута A_j в базе одинаковые, идентифицировать ФЛ по этому атрибуту невозможно, и атрибут обезличиванию не подлежит.

Максимальное значение $Q_j = V$ означает, что все значения атрибута A_j в базе уникальные, ФЛ идентифицируется однозначно. Таким образом, из (7) следует, что чем больше значение W_j, тем ближе база к однозначной идентификации по атрибуту A_j. Фактически W_j является границей целесообразности обезличивания по атрибуту A_j , то есть частью объема БД, которой соответствует количество записей q_{jW} (полученных в результате поиска), превышение которого означает нецелесообразность процесса идентификации (деобезличивания).

Таким образом, q_{jW} – это верхняя граница параметра модели нарушителя U, а W_j – это вероятность того, что количество записей, полученное при поиске по атрибуту A_j, превышает q_{jW}.

При рассмотрении количества полученных записей U как характеристики возможностей нарушителя по обработке записей БД, можно вычислить допустимое значение W_U вероятности того, что количество записей БД, содержащих произвольное значение атрибута ФЛ, превысит возможности нарушителя U. Для обеспечения безопасного уровня обезличивания регулирующий орган должен установить нормативное значение вероятности идентификации $W_{норм} \geq W_U$.

Количество записей U – это максимальное количество, которое нарушитель может обработать с применением организационных мер. В результате поиска атрибута A_j по значению A_{jk} нарушитель получает количество записей q_{jk}, при этом показателем эффективности действий для нарушителя является условие $q_{jk} < U$. Но, поскольку существуют такие значения атрибута, при которых выполняется условие $q_{jk} > U$, то для произвольного значения атрибута показателем эффективности для нарушителя является условие, при котором значение вероятности этого события $p(q_{jk} > U)$ не слишком велико. Оператор ИСПДн, напротив, должен обеспечить, чтобы эта вероятность была больше некоторого нормативного значения $W_{норм}$.

Предложенный подход позволил сформулировать критерий необходимости обезличивания по атрибуту ФЛ.

Обезличивание ПД ФЛ по атрибуту A_j, где $j = 1, \ldots, K$, необходимо при выполнении хотя бы одного из двух условий:

1) вероятность идентификации ФЛ по атрибуту больше, чем допустимое (нормативное) значение вероятности того, что количество записей БД, содержащих произвольное значение атрибута ФЛ, превысит возможности нарушителя:

$$W_j > W_{норм} \geq W_U = p(q_{jk} > U), \quad (8)$$

где W_j – вероятность идентификации ФЛ по j-му атрибуту;

$W_{норм}$ – нормативное (в перспективе заданное регулирующим органом) значение вероятности;

W_U – вероятность того, что количество записей, содержащих искомое значение j-го атрибута, больше количества записей, соответствующих возможностям нарушителя;

2) количество записей, соответствующее возможностям нарушителя, больше, чем количество записей, содержащих любое искомое значение атрибута:

$$U > q_j. \quad (9)$$

Условие $W_{норм} = W_U$ из (8) теоретически выполнимо, но в реальных условиях оно трудно достижимо из-за разницы подходов к этому вопросу нарушителя и контролирующего органа. Для нарушителя параметр U не зависит от номера атрибута и его характеристик, поэтому вероятность W_U не имеет практического значения, поскольку зависит от параметров атрибута. Для контролирующего органа, наоборот, задачей является регламентирование нормативной вероятности $W_{норм}$ как универсального показателя для всех атрибутов, а условие (9) касается количества записей с одинаковым значением для конкретного атрибута.

На параметр U, не зависимо от атрибутов, влияют внешние условия, одно из которых – время поиска, рассмотренное в [74], предполагает, что $U = 20$. Это означает, что нарушитель не может за приемлемое время однозначно выбрать определяемое ФЛ из группы записей более 20 разных лиц, которую он получил в результате поиска по имеющимся значениям идентификаторов. Под приемлемым понимается максимальное время, которое нарушитель готов затратить на дальнейшую обработку полученной группы записей.

Набор атрибутов, подлежащих обезличиванию, в первую очередь должен быть достаточным для однозначной идентификации конкретного ФЛ (например, «фамилия» и «дата рождения»), то есть, общий показатель W для этого набора должен быть равен 1 [74 – 76]. Но достаточность для идентификации выявленного набора атрибутов не гарантирует надежного обезличивания прочих данных по оставшимся атрибутам. Например, если среди прочих данных окажутся такие атрибуты как «имя», «адрес проживания», «номер телефона», «место работы», то для некоторых ФЛ сочетание этих атрибутов может быть достаточным для идентификации ФЛ без атрибута «фамилия». Поэтому показатель W для любого

набора из оставшихся атрибутов должен быть гарантированно меньше 1, т.е. меньше $W_{норм}$.

В представленной работе в качестве предварительного значения принято значение $W_{норм} = 0{,}05$. Сформулирована гипотеза о том, что это значение для всех атрибутов соответствует параметру модели нарушителя $U = 20$.

В модифицируемую группу должны быть включены все реквизиты (например, номер паспорта, ИНН, СНИЛС и т.д.), для которых $W = 1$ [74]. Поэтому далее в работе эти атрибуты не рассматриваются.

Необходимо учитывать, что обработка ПД автоматизированным способом в реальных условиях производится в рамках нескольких таблиц БД реляционного типа, одна часть которых является справочными (условно постоянные), а другая часть – изменяемыми данными функционального характера (переменные). Эти части связаны посредством специальных служебных идентификаторов. Такие связующие идентификаторы не могут быть модифицированы независимо в различных таблицах и поэтому не должны включаться в модифицируемую группу.

В реальных условиях значение атрибута всегда является дискретным, так как может принимать конечное множество значений, размер которого соответствует количеству записей V, когда все значения являются уникальными, и, при этом может быть аппроксимировано некоторой непрерывной функцией.

Построение модели идентификации решает две задачи:

1) определение функции распределения случайной величины q_{jk} для каждого атрибута по полученной зависимости определение значения q_{jk}, для которого соблюдается условие (8), принятие решения о целесообразности обезличивания по j-му атрибуту;

2) определение зависимости вероятности идентификации W по атрибутам от количества записей БД. Необходимо отметить, что количество записей БД V может изменяться во времени, а целая база может состоять из нескольких частей меньшего объема с теми же атрибутами. Поэтому расчеты и выводы, сделанные

для количества записей базы V, будут отличаться от расчетов и выводов для баз других объемов.

Для любого атрибута A_j используется следующий алгоритм применения критерия необходимости обезличивания:

1) определить множество различных значений j-го атрибута в БД

$$A_j = \{A_{j1}, \dots, A_{jk}, \dots, A_{jQ}\};$$

2) определить количество различных значений атрибута Q_j;

3) вычислить W_j – вероятность идентификации ФЛ по j-му атрибуту (7);

4) определить для каждого значения атрибута количество записей q_{jk}, имеющих атрибут с данным значением;

5) построить дискретную последовательность частот n_{jk} – количества повторов значений q_{jk};

6) для полученной последовательности подобрать непрерывную функцию $y_j = f_j(x)$, аппроксимирующую дискретную случайную величину q_j, и соответствующую ей функцию плотности вероятности $p_j(x)$, отвечающую критериям согласия (первого и второго рода);

7) определить количество записей q_{jW}, соответствующее значению W_j, вычисленному на шаге 3, как решение уравнения

$$W_j = \int_{q_{jW}}^{\infty} p_j(x)dx, \qquad (10)$$

где $p_j(x)$ – функция плотности вероятности, полученная на предыдущем шаге,
и количество записей $q_{j\text{норм}}$, соответствующее значению $W_{\text{норм}}$, путем подстановки $W_j = W_{\text{норм}} = 0{,}05$ (предполагается, что это значение соответствует высокой компетенции нарушителя $U = 20$;

8) сравнить значения q_{jW}, $q_{j\text{норм}}$ между собой и с параметром U;

9) сделать вывод о целесообразности обезличивания по j-му атрибуту для БД с количеством записей V;

10) произвести действия (1) – (9) с сочетаниями тех атрибутов, для которых в отдельности обезличивание нецелесообразно, и сделать расчет для выбранного сочетания атрибутов;

11) произвести действия (1) – (10) для рассмотренных выше атрибутов для БД других объемов V_b, где b – номер базы/части базы, имеющей количество записей, отличное от V;

12) повторить шаги (5) и (6) для аппроксимации зависимости W_{jb} от V_b.

Для обеспечения достоверности результатов эксперимента исследованы две базы ПД:

1) БД B_1 объемом 310 тыс. записей ФЛ – жителей одного города, где одному ФЛ соответствовала одна запись базы. Для обработки были использованы следующие атрибуты (поля БД):

– фамилия (A_1);

– имя (A_2);

– отчество (A_3);

– наименование улицы (часть адреса проживания – A_4);

– номер дома (часть адреса проживания – A_5);

– номер квартиры (часть адреса проживания – A_6).

2) БД B_2 объемом 329 тыс. записей ФЛ – жителей одного региона, где одному ФЛ соответствовала одна запись базы. Для обработки были использованы следующие атрибуты (поля БД):

– объединенное поле «фамилия + имя + отчество», из которого был выделен атрибут «фамилия» (A_7);

– дата рождения (A_8).

Каждый из указанных атрибутов рассматривается как текстовая случайная величина, для которой определено множество различных значений и частота ряда характеристик.

Для атрибутов «фамилия», «имя», «отчество», адресов проживания рассмотрена зависимость вероятности идентификации от количества записей БД. С этой целью произведен ряд случайных выборок различного объема из экспериментальной БД и построена функция зависимости W_j от V. Была проверена и подтверждена выдвинутая ранее в [83 – 84] гипотеза о степенном

законе распределения значений указанных атрибутов по количеству записей независимо от размера базы V, а также гипотеза о степенной зависимости вероятности идентификации W по этим атрибутам от размера базы V.

Для атрибута «дата рождения» определена частота количества различных значений дат. При этом рассмотрена полная дата рождения в формате «гггг-мм-дд» (фактически это возраст человека в днях в более сложном представлении). Кроме того, исследована зависимость вероятности идентификации по этому атрибуту от количества записей БД.

2.3. Распределение характеристик атрибутов по количеству физических лиц

Для всех рассмотренных атрибутов, функция распределения строилась в диапазоне количества записей от $q_{jmin} = 1$ до q_{jmax}. В соответствии с [85] диапазон был разделен, как правило, на $d = 10$ равных частей, согласно неравенству

$$d > 1 + log_2 N,$$

где N – количество различных значений q_j.

Для предварительной оценки характера зависимости количества повторов n_{jk} от количества записей q_{jk} в рамках дискретной последовательности методом наименьших квадратов автоматизированными средствами Excel строились аппроксимирующие графики линейного, экспоненциального, логарифмического и степенного типа. Дополнительно строились диаграммы двух видов: в логарифмическом масштабе по оси ординат (проверка гипотезы экспоненциальной зависимости) и по обеим осям (проверка гипотезы степенной зависимости). Для всех рассмотренных атрибутов график экспоненциальной зависимости (диаграмма в логарифмическом масштабе по ординате) имел вид, близкий к линейному, только в средней части диапазона. Для всех атрибутов, за исключением даты рождения, диаграммы в логарифмическом масштабе по обеим координатам имели близкий к линейному вид в большей части диапазона, исключая участок больших значений (эффект так называемого «размытия хвоста», в соответствии с [45]).

Для проверки соответствия экспериментальной и теоретической зависимостей согласно [85, 86] для нескольких коэффициентов экспоненциальной и степенной функции $y(x)$ использовались три критерия согласия:

1) критерий хи-квадрат χ^2, вычисляемый как

$$\chi^2 = Q_j \sum_k ((n_{jk}/Q_j - y(q_{jk}))^2 / y(q_{jk})), \quad (11)$$

где $k = 1, \ldots, d_j$ – количество диапазонов диаграммы распределения q_j, остальные обозначения соответствуют (5) и (6);

2) критерий отношения правдоподобия S_{on}, вычисляемый как

$$S_{on} = -2 \sum_k (n_{jk} \ln(y(q_{jk}) / n_{jk}/Q_j)), \quad (12)$$

где $k = 1, \ldots, d_j$ – количество диапазонов диаграммы частот n_{jk}, остальные обозначения соответствуют (5) и (9);

3) критерий Колмогорова $T(q_{jk})$, вычисляемый как

$$T(q_{jk}) = \sqrt{Q_j} \mathrm{MAX}(F_{y(qjk)} - F_{njk}), \quad (13)$$

где $F_{y(qjk)}$ и F_{njk} – функции распределения для проверяемой и экспериментальной зависимостей, соответственно.

Все три критерия проверялись при заданном уровне значимости $\alpha = 0{,}05$. Эта проверка позволила избежать ошибки первого рода (отвергнуть верную гипотезу).

Далее для устранения ошибки второго рода (принять неверную гипотезу) все функции, прошедшие проверку первого рода, подвергались проверке по методу распознавания зависимостей на основе обратного отображения в соответствии с [87]. В результате сравнения выбиралась функция $y(x)$ с минимальной дисперсией ошибок s^2, вычисляемой как

$$s^2 = \sum_{k=1}^{dj} (Y(q_{jk}) - q_{jk})^2, \quad (14)$$

где $Y(x)$ – функция, обратная к $y(x)$.

Для атрибутов текстовой семантики с неограниченным набором значений («фамилия», «имя» и т.п.) в качестве примеров функции $y(q_{jk})$ для аппроксимации рассматривались монотонные функции: степенная $y(x) = ax^b$,

логарифмическая $y(x) = a\ln(x) + c$ и экспоненциальная $y(x) = ae^{-cx}$ с различными параметрами.

Для атрибутов числовой семантики с условно ограниченным набором значений («номер дома», «дата рождения» и т.п.) с возможным наличием максимума дополнительно рассматривалась гамма-функция вида $y(x) = ax^b e^{-cx}$.

Особенность атрибута «дата рождения», в отличие от атрибутов «фамилия» или «номер дома» состоит в том, что с увеличением объема БД количество различных значений дат тоже растет, но есть верхний предел

365 дней · 100 лет = 36500 дат.

С увеличением количества записей БД доля одиночных дат (дат, встречающихся в базе один раз) падает, что подтверждается расчетами: для базы объемом 5,5 тыс. записей доля равна 87%, для базы 65 тыс. записей – 30%, для базы 329 тыс. записей – менее 3%. За одиночными начинает падать доля двойных, затем тройных и т.д. дат. Таким образом, распределение по атрибуту «дата» не может быть ни экспоненциальным (хотя на малых объемах очень близко к нему), ни степенным. Исходя из поведения функции выдвинута гипотеза о гамма-распределении вида $y(x) = ax^b e^{-cx}$, параметры которого описаны в [88].

В части диапазона, близкой к $q_{j\max}$, для некоторых атрибутов наблюдаются значительные отклонения от монотонности (эффект «размытия хвоста»), аналогичные выявленным в [45], что увеличивает погрешность вычислений. Для устранения возможной погрешности принято решение использовать объединение нескольких частей диапазона.

Полученные результаты по каждому атрибуту приведены в разделах 2.3.1 – 2.3.8.

2.3.1. Распределение характеристик атрибута «фамилия»

Для атрибута «фамилия» (A_1) базы данных B_1 (количество записей V = 310132 ФЛ) определено множество различных значений. Количество различных значений Q_1 = 45099, для каждого из которых определено количество записей q_{1k}

в диапазоне от $q_{1\text{мин}} = 1$ до $q_{1\text{макс}} = 1892$ (количество записей со значением «Иванов»).

По формуле (7) вычислено значение

$$W_1 = Q_1 / V = 0{,}1454.$$

На рис. 2 в логарифмическом масштабе представлена диаграмма частот значений n_{1d} в зависимости от q_{1k} для атрибута «фамилия» (A_1) базы данных B_1 при $d = 10$ интервалов, где n_{1d} – сумма n_{1k} по каждому интервалу. На рис. 2a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 2b – по обеим осям (при линейном тренде для степенной функции).

На рис. 2a виден линейный характер в средней части диапазона, на рис. 2b – в большей части диапазона. На обеих диаграммах наблюдается эффект «размытия хвоста» на правом краю диапазона, поэтому принято решение об изменении размера интервалов.

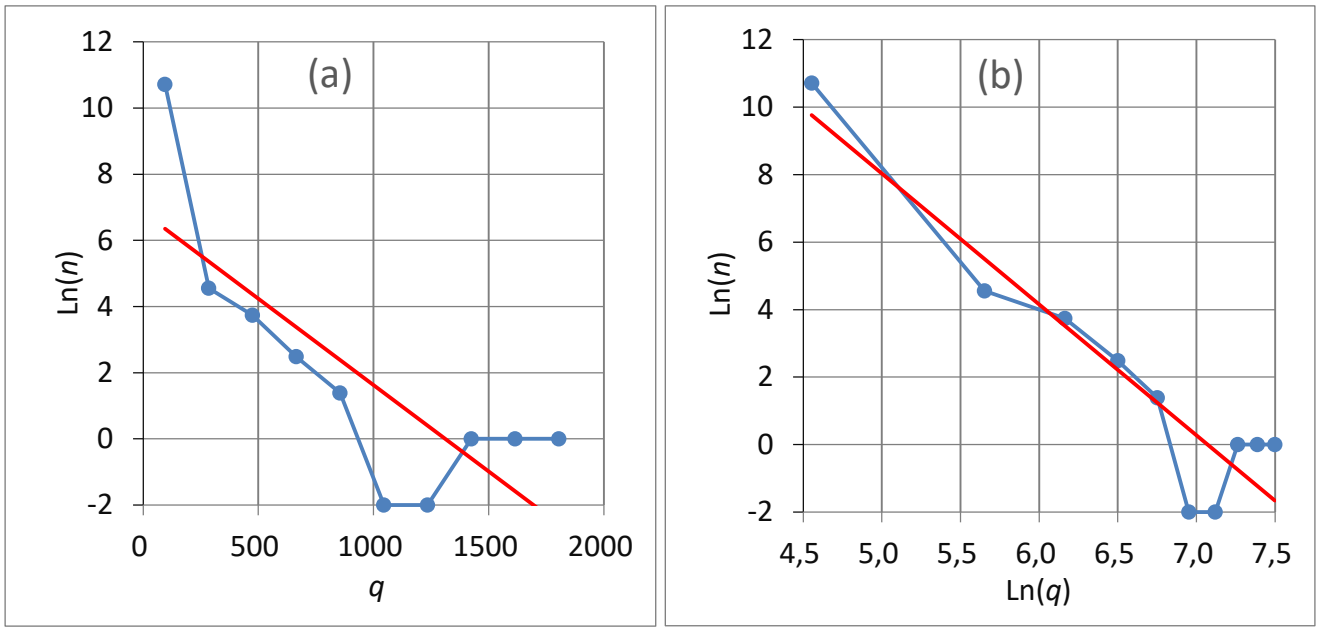

Рис. 2. Диаграмма частот значений n атрибута «фамилия» по всему диапазону в логарифмическом масштабе: a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В табл. 2 приведены результаты аппроксимации дискретного распределения случайной величины q_1 для атрибута «фамилия» (A_1) базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 194800e^{-0,047x}$ и степенной функцией $f_2(x) = 2,46 \cdot 10^9 x^{-3,163}$, где x – среднее значение интервала на оси количества записей q_1, а значение функций f_1 и f_2 – количество фамилий n_1 в интервале.

Таблица 2

Распределение значений фамилий в диапазоне 1 … 1982

x	B_1	f_1	f_2
32	43296	43292,2394	42739,44
96	1340	2138,21932	1323,41
160	283	105,607424	263,0187
224	74	5,21598876	90,73685
288	37	0,25761957	40,97885
352	18	0,01272392	21,7222
416	17	0,00062844	12,80637
480	13	$3,1039 \cdot 10^{-5}$	8,14427
576	7	$3,407 \cdot 10^{-7}$	4,575113
1266	1	$2,8067 \cdot 10^{-21}$	0,378984
χ^2	16,919	$3,59 \cdot 10^{20}$	11,921
S_{on}	16,919	2254,824	11,225
$T(q)$	1,3581	2,057195195	0,06915

В трех нижних строках табл. 2 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по (11) – (13) при заданном уровне значимости $\alpha = 0,05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 3 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_1 для атрибута A_1 «фамилия» базы B_1 (см. рис.2b).

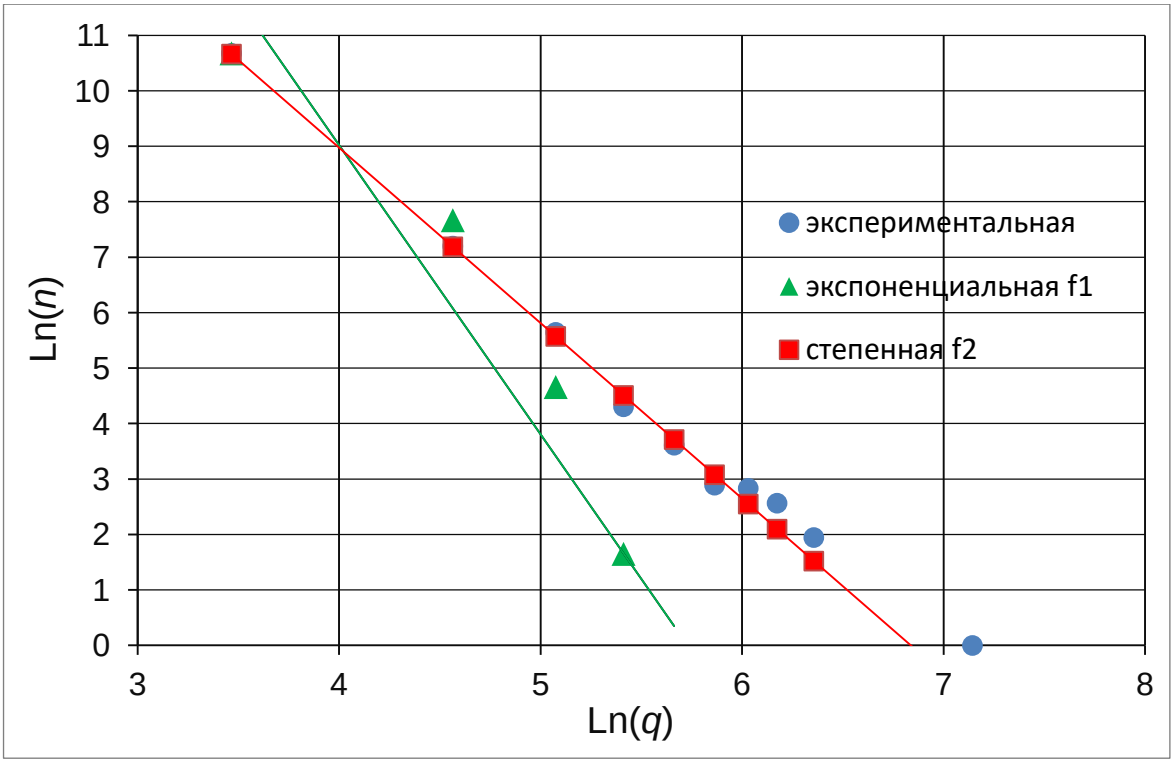

Рис. 3. Функции f_1 и f_2 в сравнении с дискретной последовательностью для атрибута A_1, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения q_{1W} и $q_{1норм}$. На рис. 4 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_1, $W_{норм}$ и U.

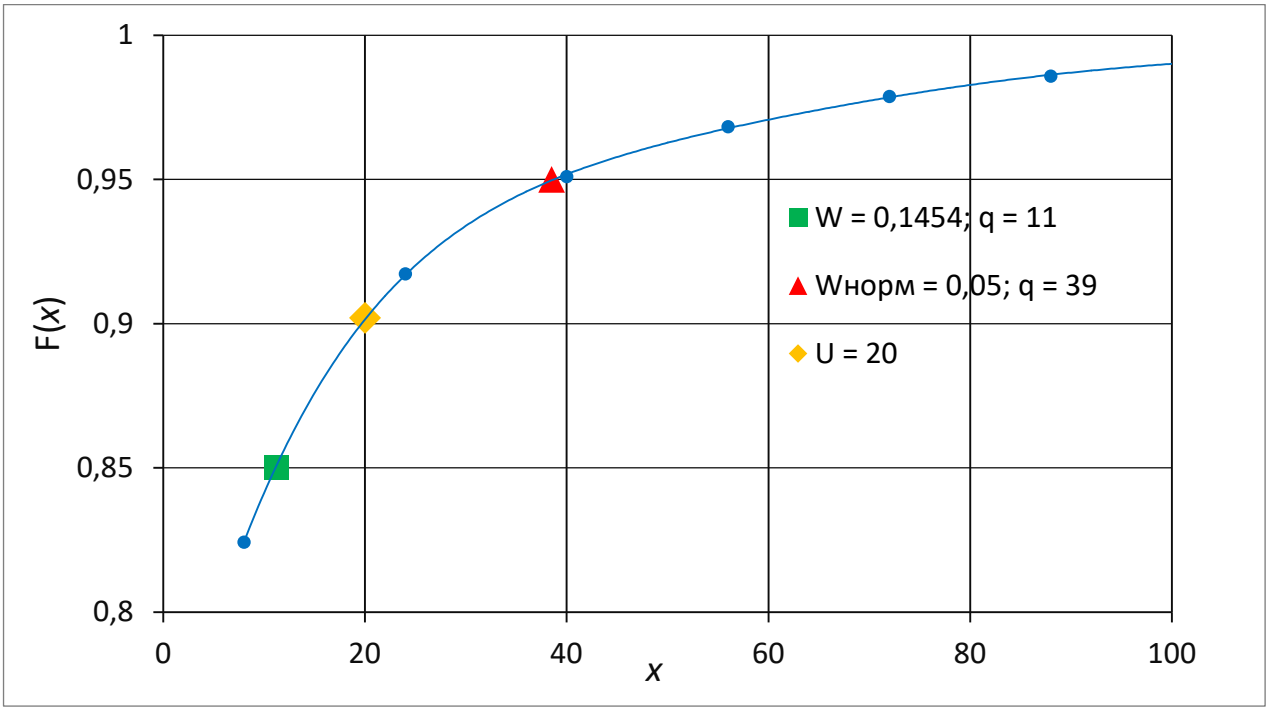

Рис. 4. Распределение вероятности для атрибута «фамилия» в сравнении с W_1, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 4 можно сделать следующие выводы:

– возможности нарушителя позволяют ему успешно работать с необезличенным атрибутом A_1 ($U > q_{1W} = 11{,}2$), поэтому атрибут A_1 «фамилия» необходимо обезличивать;

– атрибут A_1 необходимо обезличивать в соответствии с нормативным значением, поскольку $W_1 > W_{норм}$;

– нормативное значение для атрибута A_1 является избыточным относительно возможностей нарушителя ($U < q_{1норм} = 38{,}5$).

С целью подтверждения достоверности полученных результатов (и независимости их от конкретной базы данных) аналогичные вычисления были произведены для атрибута «фамилия» (A_7) базы данных B_2 в одной из выборок объемом $V_6 = 87551$ записей. Количество различных значений $Q_7 = 19763$, для каждого из которых определено количество записей q_{7k} в диапазоне от $q_{7мин} = 1$ до $q_{7макс} = 534$.

По формуле (7) вычислено значение

$$W_7 = Q_7 / V_6 = 0{,}2257.$$

Для атрибута «фамилия» (A_7) базы данных B_2 для d = 10 интервалов в логарифмическом масштабе координат построена диаграмма распределения значений n_{7d} от q_{7k}, где n_{7d} – сумма n_{7k} по каждому интервалу. При этом наблюдался эффект «размытия хвоста» на правом краю диапазона, поэтому было принято решение об изменении размера и количества интервалов (Табл.3).

В табл. 3 приведены результаты аппроксимации дискретного распределения случайной величины q_7 для атрибута «фамилия» (A_7) базы данных B_2 для d = 8 интервалов экспоненциальной функцией $f_1(x) = 26690 \cdot e^{-0{,}048x}$ и степенной функцией $f_2(x) = 4{,}508 \cdot 10^7 x^{-3{,}033}$, где x – среднее значение интервала на оси количества записей q_7, а значение функций f_1 и f_2 – количество фамилий n_7 в интервале.

Таблица 3

Распределение значений фамилий в диапазоне 1 … 534

x	B_2	f_1	f_2
13	18870	14300,42	18854,53
39	694	4105,342	673,4525
65	110	1178,555	143,0341
91	40	338,3376	51,55056
117	22	97,12941	24,05461
143	16	27,88375	13,08795
169	10	8,004823	7,885434
358	1	0,000919	0,809243
χ^2	14,067	6770,77759	12,29430034
S_{on}	14,067	7821,261	13,08324
$T(Q)$	1,3581	33,98943	0,294181

В трех нижних строках табл. 3 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_2) [85] и

рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 5 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_7 для атрибута A_7 «фамилия» базы B_2.

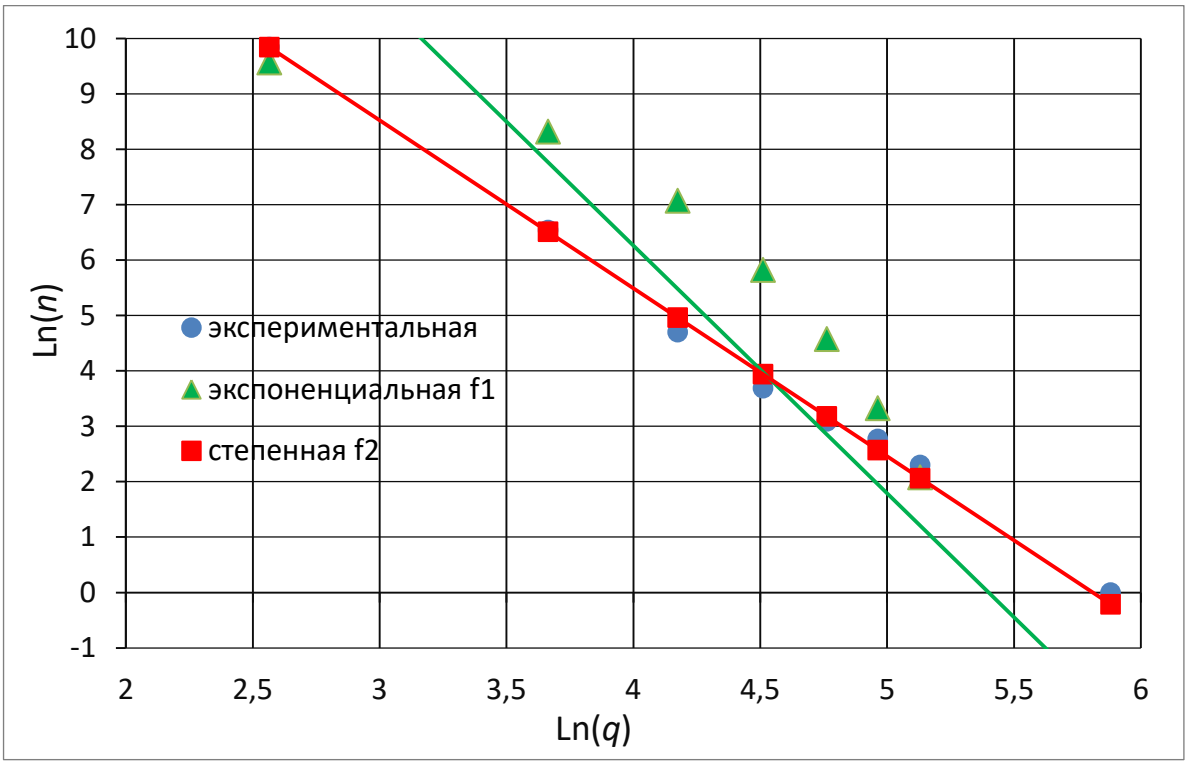

Рис. 5. Функции f_1 и f_2 в сравнении с дискретной последовательностью для атрибута A_7, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения q_{7W} и $q_{7\text{норм}}$. На рис. 6 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_7, $W_{\text{норм}}$ и U.

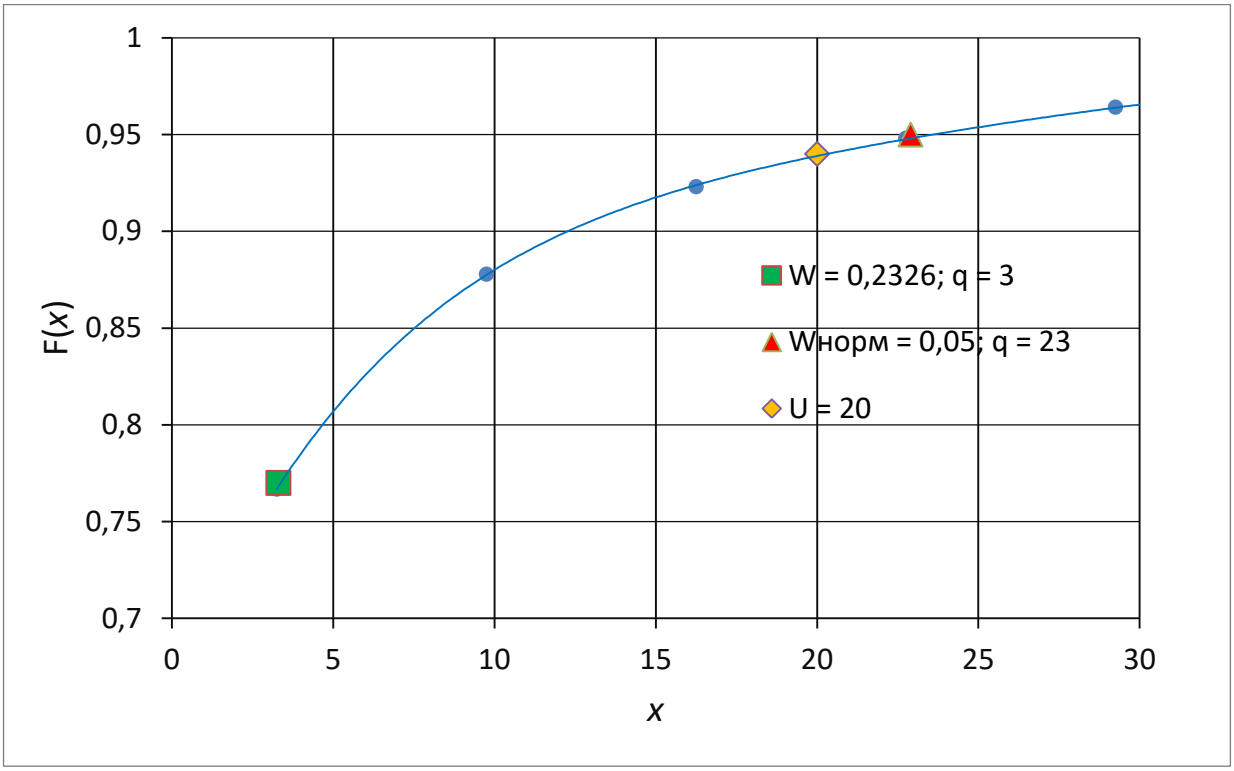

Рис. 6. Распределение вероятности для атрибута «фамилия» в сравнении с W_7, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 6 можно сделать следующие выводы:

– возможности нарушителя позволяют ему успешно работать с необезличенным атрибутом A_7 ($U > q_{7W} = 3{,}3$), поэтому атрибут A_7 («фамилия») необходимо обезличивать;

– атрибут A_7 необходимо обезличивать в соответствии с нормативным значением, поскольку $W_7 > W_{норм}$;

– нормативное значение для атрибута A_7 является избыточным относительно возможностей нарушителя ($U < q_{7норм} = 22{,}8$).

Произведенные вычисления для атрибута A_7 («фамилия») позволяют сделать следующие выводы:

– о подтверждении степенного вида зависимости на примере ПД из другой базы (достоверность эксперимента);

– о подтверждении степенного вида зависимости атрибута «фамилия» для другого количества записей базы;

– об увеличении значения вероятности обезличивания по этому атрибуту с уменьшением количества записей базы.

Подробнее зависимость вероятности обезличивания от количества записей базы рассмотрена в п.2.4.

2.3.2. Распределение характеристик атрибута «имя»

Атрибут «имя» (A_2) базы данных B_1 рассматривался аналогично атрибуту «фамилия». Для женских и мужских имен построена общая диаграмма частот. Количество различных значений $Q_2 = 755$, для каждого из которых определено количество записей q_{2k} в диапазоне от $q_{2\text{мин}} = 1$ до $q_{2\text{макс}} = 8278$ (количество записей со значением «Татьяна»).

По формуле (7) вычислено значение

$$W_2 = Q_2 / V = 0{,}002434.$$

На рис. 7 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{2d} в зависимости от q_{2k} для атрибута «имя» (A_2) базы данных B_1 при $d = 10$ интервалов, где n_{2d} – сумма n_{2k} по каждому интервалу. На рис. 7a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 7b – по обеим осям (при линейном тренде для степенной функции).

На рис. 7a виден линейный характер в средней части диапазона, на рис. 7b – в большей части диапазона. Эффект «размытия хвоста» не наблюдается.

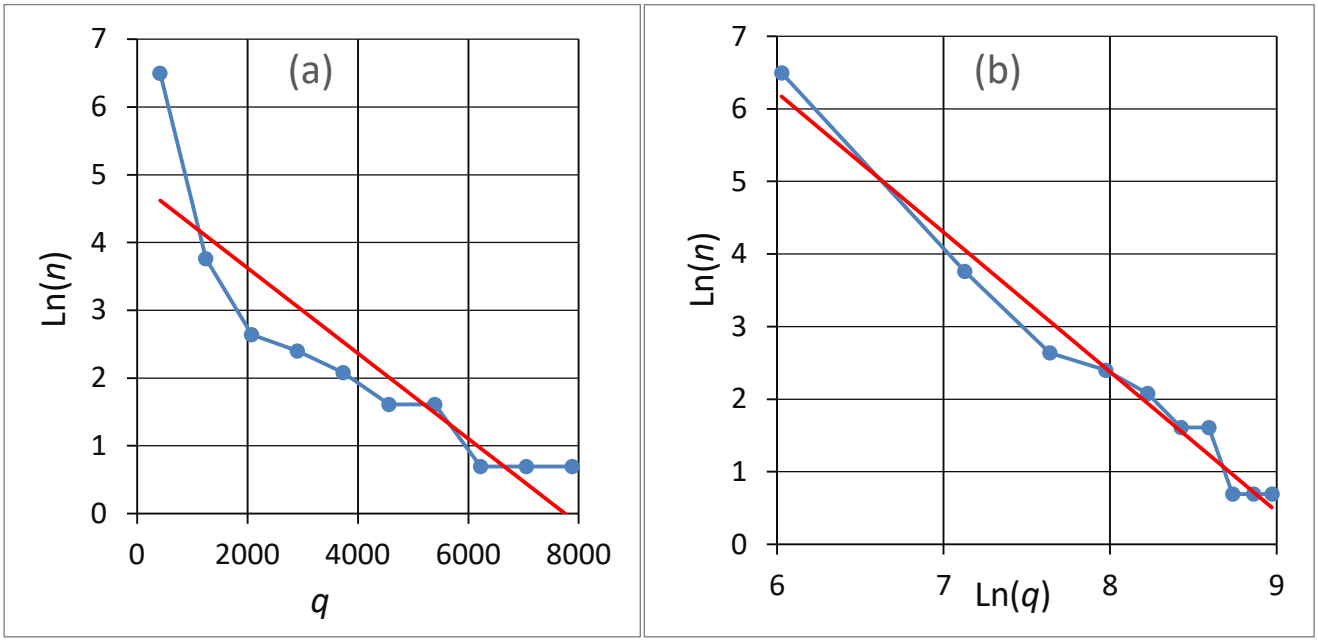

Рис. 7. Диаграмма частот значений *n* атрибута «имя» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, *q* – количество записей, имеющих одинаковое значение атрибута, *n* – количество различных значений атрибута

В табл. 4 приведены результаты аппроксимации дискретного распределения случайной величины q_2 для атрибута «имя» (A_2) базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 7100 \cdot e^{-0{,}0011x}$ и степенной функцией $f_2(x) = 2{,}94 \cdot 10^9 x^{-2{,}158}$, где x – среднее значение интервала на оси количества записей q_2, а значение функций f_1 и f_2 – количество имён n_2 в интервале.

Таблица 4

Распределение значений имен в диапазоне 1 … 8300

x	***B*₁**	***f*₁**	***f*₂**
415	663	443,4479066	659,1657058
1245	43	177,9638314	61,56968572
2075	14	71,4201709	20,4464227
2905	11	28,66223306	9,891747777
3735	8	11,50268326	5,750945726
4565	5	4,616239142	3,729659545
5395	5	1,852581988	2,600787386
6225	2	0,743475352	1,90980782
7055	2	0,298370384	1,457759338

x	B_1	f_1	f_2
7885	2	0,119741543	1,146685534
χ^2	16,919	309,6142181	12,13729295
S_{on}	16,919	343,1867645	12,10223238
$T(q)$	1,3581	7,677127392	0,535454236

В трех нижних строках табл. 4 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 8 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_2 для атрибута A_2 «имя» базы B_1 (см. рис. 7b).

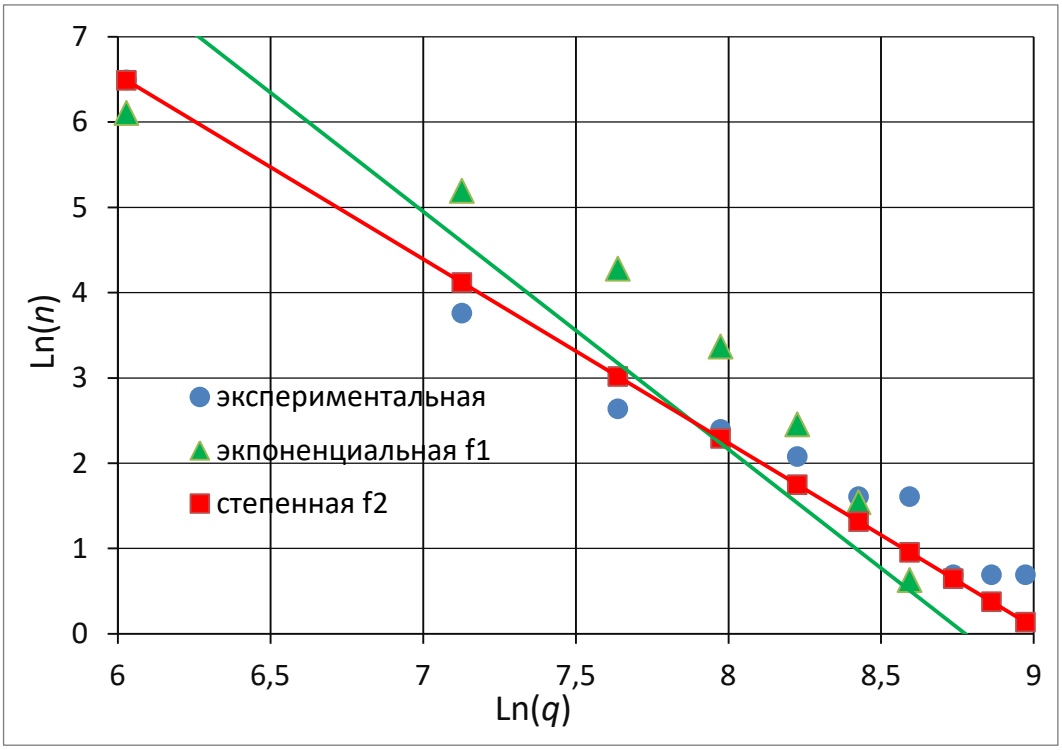

Рис. 8. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для атрибута A_2, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения q_{2W} и $q_{2норм}$. На рис. 9 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_2, $W_{норм}$ и U.

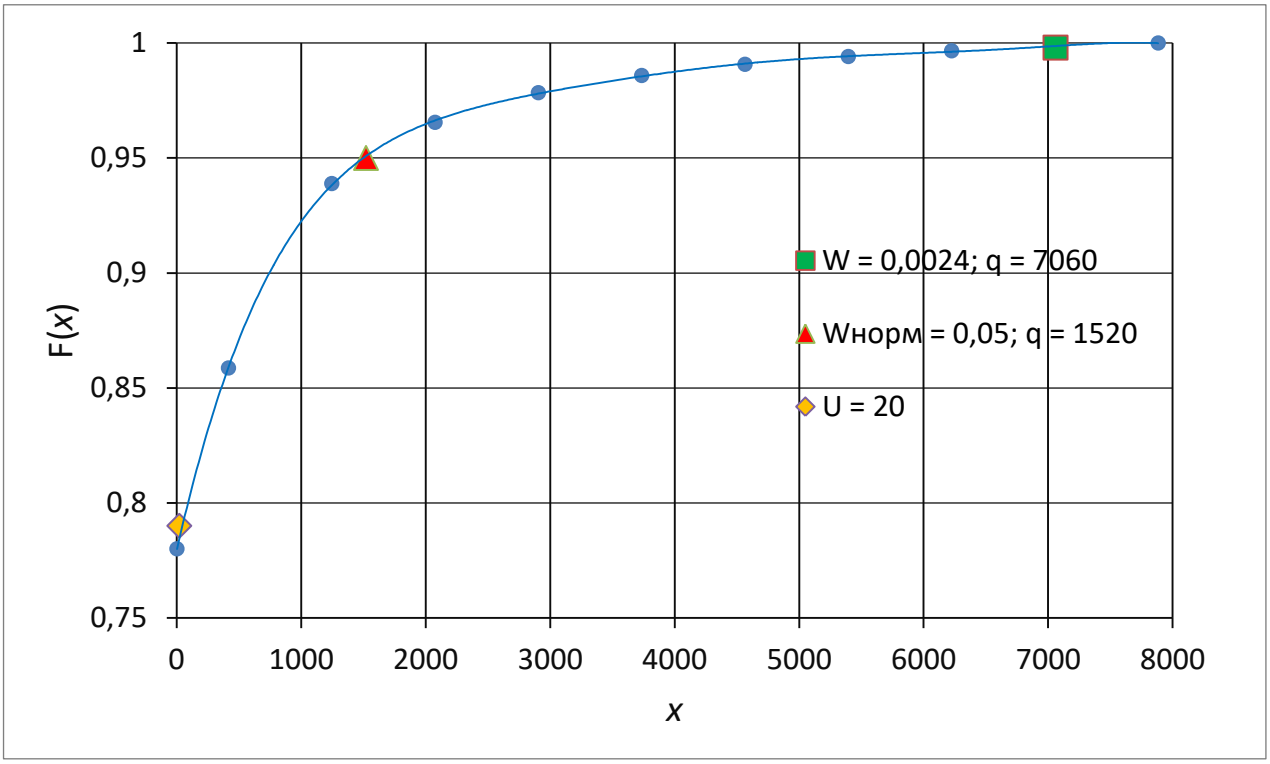

Рис. 9. График распределения вероятности для атрибута «имя» в сравнении с W_2, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 9 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_2 ($U < q_{2W} = 7060$), поэтому атрибут A_2 («имя») обезличивать не требуется;

– атрибут A_2 не требуется обезличивать в соответствии с нормативным значением, поскольку $W_2 < W_{норм}$;

– нормативное значение для атрибута A_2 является избыточным относительно возможностей нарушителя, поскольку $U < q_{2норм} = 1520$.

2.3.3. Распределение характеристик атрибута «отчество»

Атрибут «отчество» (A_3) базы данных B_1 рассматривался аналогично атрибуту «имя». Для женских и мужских отчеств построена общая диаграмма частот в диапазоне от $q_{3\text{мин}} = 1$ до $q_{3\text{макс}} = 18390$ (количество записей со значением «Александровна(вич)»). Количество различных значений $Q_3 = 349$.

По формуле (7) вычислено значение

$$W_3 = Q_3 / V = 0{,}001116.$$

На рис. 10 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{3d} в зависимости от q_{3k} для атрибута «отчество» (A_3) базы данных B_1 при $d = 10$ интервалов, где n_{3d} – сумма n_{3k} по каждому интервалу. На рис. 10a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 10b – по обеим осям (при линейном тренде для степенной функции).

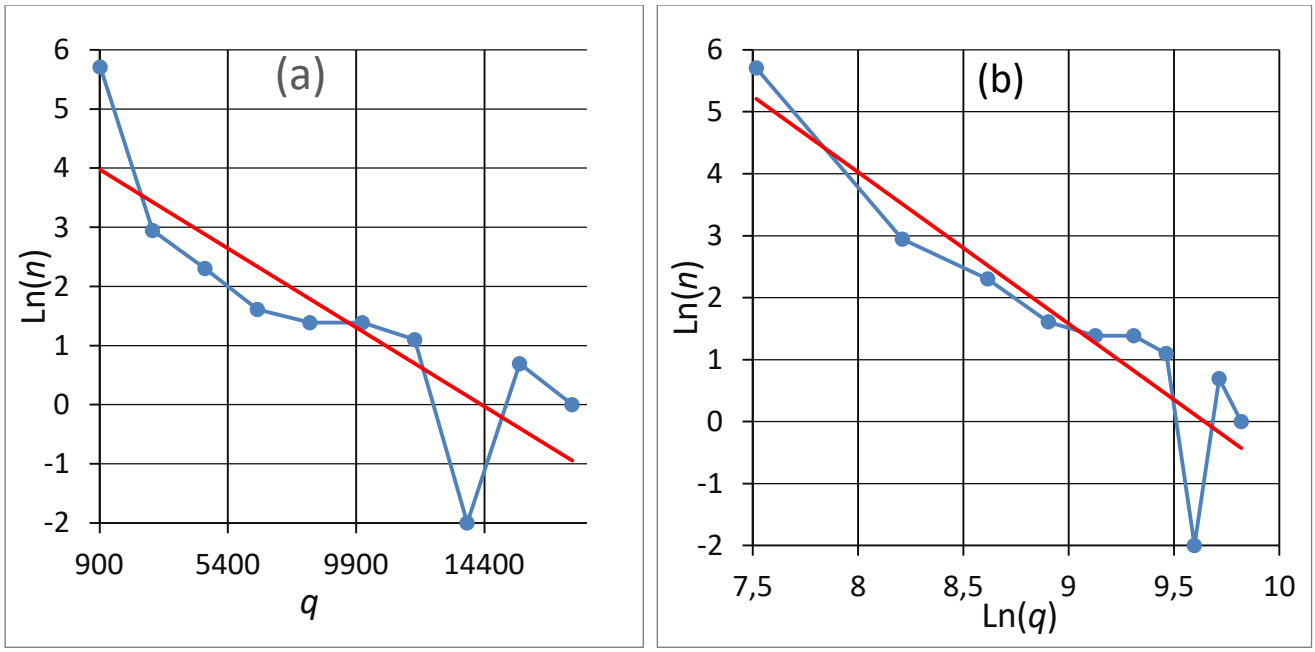

Рис. 10. Диаграмма частот значений *n* атрибута «отчество» по полному диапазону в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, *q* – количество записей, имеющих одинаковое значение атрибута, *n* – количество различных значений атрибута

На рис. 10a виден линейный характер в средней части диапазона, на рис. 10b – в большей части диапазона. На обеих диаграммах наблюдается эффект

«размытия хвоста» на правом краю диапазона, поэтому принято решение об изменении размера интервалов (табл.5).

В табл. 5 приведены результаты аппроксимации дискретного распределения случайной величины q_3 для атрибута «отчество» (A_3) базы данных B_1 для d = 10 интервалов экспоненциальной функцией $f_1(x) = 430 \cdot e^{-0,0009x}$ и степенной функцией $f_2(x) = 1,50 \cdot 10^7 x^{-1,697}$, где x – среднее значение интервала на оси количества записей q_3, а значение функций f_1 и f_2 – количество отчеств n_3 в интервале.

Таблица 5

Распределение значений отчеств в диапазоне 1 … 18390

x	B_1	f_1	f_2
643	258	241,0694843	257,4826664
1930	40	75,70053355	39,87413414
3217	22	23,77144829	16,75473521
4504	7	7,464699747	9,465107808
5791	6	2,344061735	6,178587754
7078	4	0,736081236	4,395252216
8365	3	0,231143906	3,310223441
9652	3	0,072583708	2,596488709
10939	2	0,022792704	2,099602078
15000	1	0,000589512	1,228726295
χ^2	16,919	2087,505826	2,426949441
S_{on}	16,919	85,72656629	2,363232768
$T(Q)$	1,3581	1,110073	0,187993

В трех нижних строках табл. 5 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости α = 0,05. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы

видно, что экспоненциальная функция f_1 отвечает только одному критерию согласия, а степенная функция f_2 отвечает всем трем.

На рис. 11 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_3 для атрибута A_3 «отчество» базы B_1 (см. рис.10b).

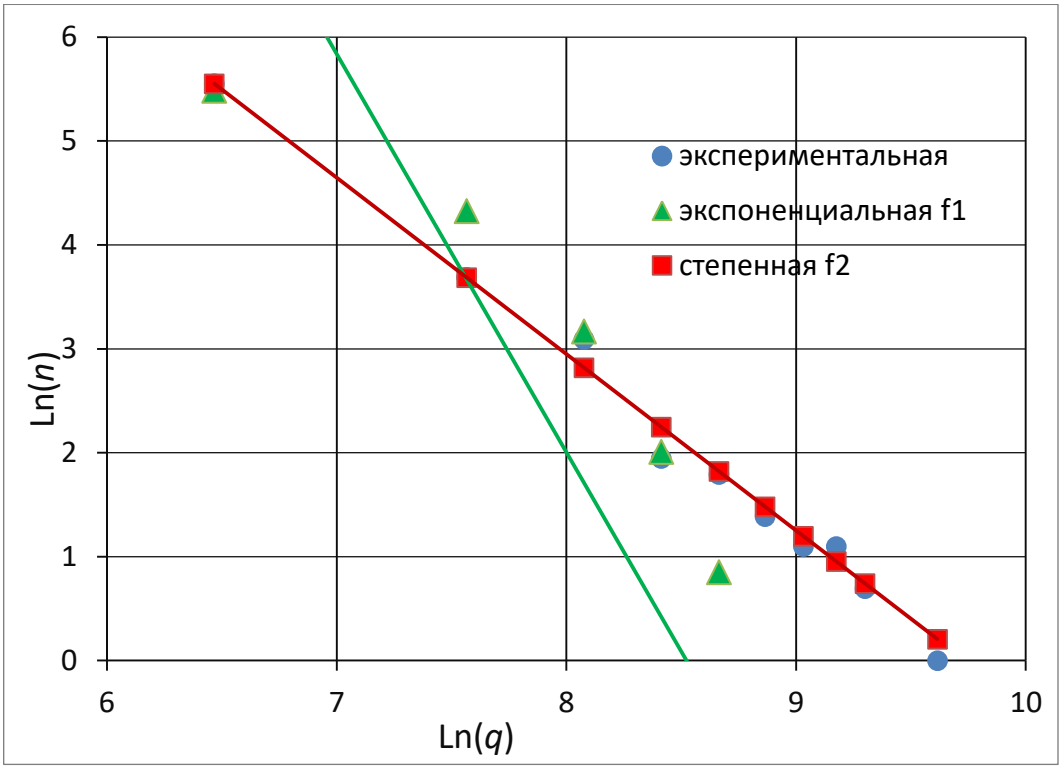

Рис. 11. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для атрибута A_3, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения q_{3W} и $q_{3норм}$. На рис. 12 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_3, $W_{норм}$ и U.

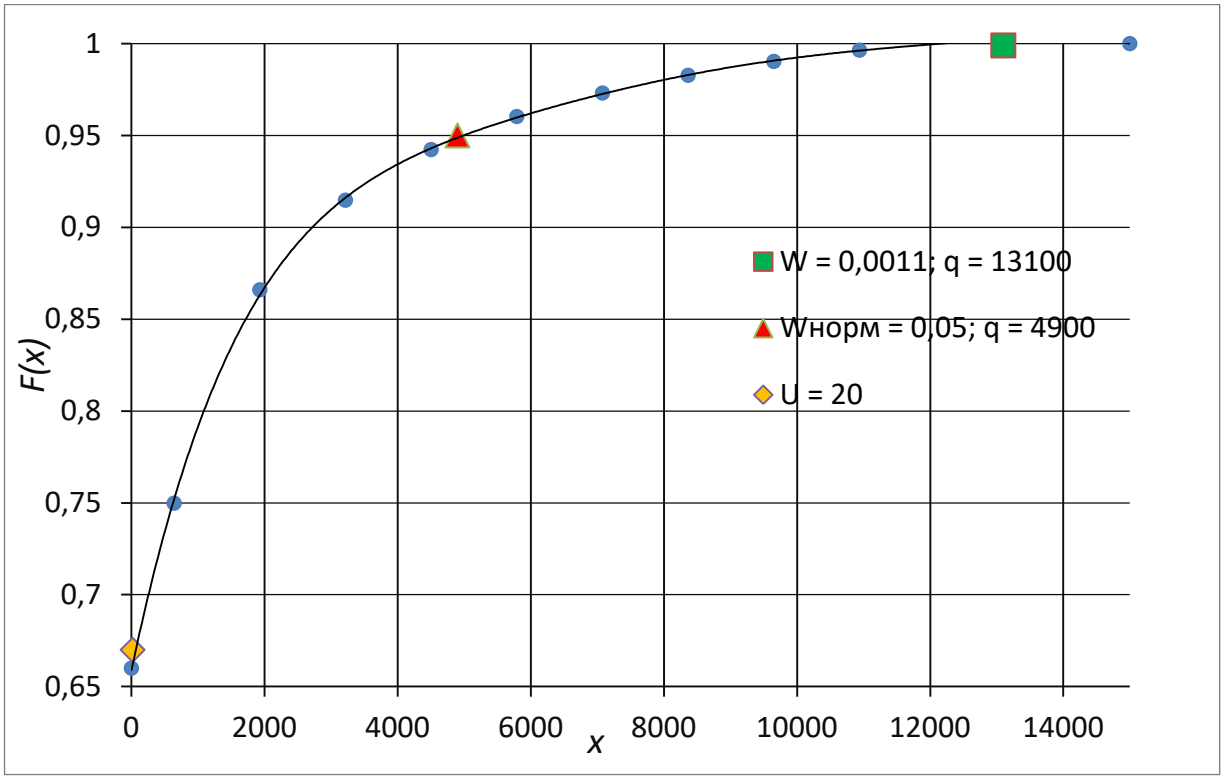

Рис. 12. График распределения вероятности для атрибута «отчество» в сравнении с W_3, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 12 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_3 ($U < q_{3W} = 13100$), поэтому атрибут A_3 («отчество») обезличивать не требуется;

– атрибут A_3 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_3 < W_{норм}$;

– нормативное значение для атрибута A_3 является избыточным относительно возможностей нарушителя ($U < q_{3норм} = 4900$).

2.3.4. Распределение характеристик атрибута «наименование улицы»

Атрибут «наименование улицы» (A_4) базы данных B_1 рассматривался аналогично атрибуту «имя». Определено количество различных значений атрибута $Q_4 = 888$. Количество записей для каждого из этих значений находится в диапазоне от $q_{4мин} = 1$ до $q_{4макс} = 8810$ (количество записей со значением «Комсомольский пр»).

По формуле (7) вычислено значение

$$W_4 = Q_4/V = 0{,}002863297.$$

На рис. 13 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{4d} в зависимости от q_{4k} для атрибута «отчество» (A_3) базы данных B_1 при $d = 10$ интервалов, где n_{4d} – сумма n_{4k} по каждому интервалу. На рис. 13a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 13b – по обеим осям (при линейном тренде для степенной функции). Эффект «размытия хвоста» не наблюдается.

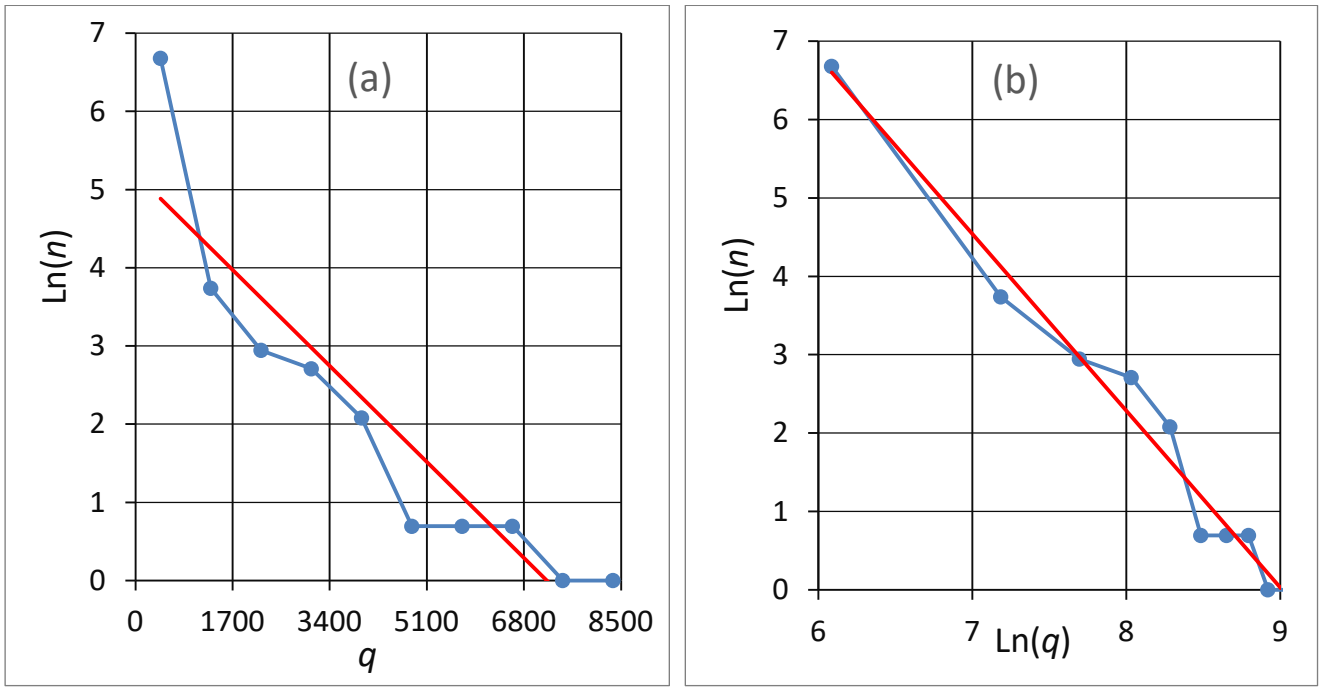

Рис. 13. Диаграмма частот значений n атрибута «наименование улицы» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

На рис. 13a виден линейный характер в средней части диапазона, на рис. 13b – в большей части диапазона.

В табл. 6 приведены результаты аппроксимации дискретного распределения случайной величины q_4 для атрибута «наименование улицы» (A_4) базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 7100 \cdot e^{-0{,}0011x}$ и степенной

функцией $f_2(x) = 2{,}94 \cdot 10^9 x^{-2{,}158}$, где x – среднее значение интервала на оси количества записей q_4, а значение функций f_1 и f_2 – количество улиц n_4 в интервале.

Таблица 6

Распределение значений наименований улиц в диапазоне 1 … 8300

x	B_1	f_1	f_2
440	796	796,20753	775,28927
1320	42	276,95634	65,02468
2200	19	96,337714	20,539385
3080	15	33,510535	9,6144091
3960	8	11,656452	5,453718
4840	2	4,0546318	3,4680223
5720	2	1,410381	2,3790719
6600	2	0,4905931	1,7226692
7480	1	0,1706501	1,2988859
8360	1	0,0593597	1,0106381
χ^2	16,919	284,4548	13,78276
S_{on}	16,919	328,055762	14,5201579
$T(Q)$	1,3581	7,27770809	0,630412034

В трех нижних строках табл. 6 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 14 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_4 для атрибута A_4 «наименование улицы» базы B_1 (см. рис. 13b).

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

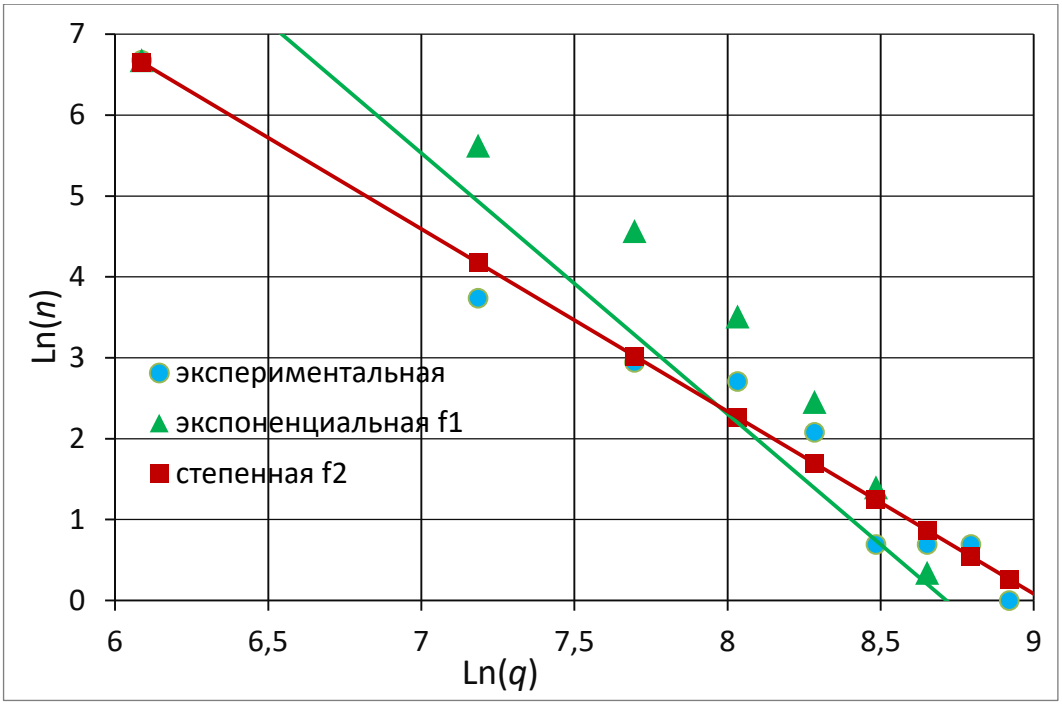

Рис. 14. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для атрибута A_4, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате решения уравнения (10) получены значения q_{4W} и $q_{4норм}$. На рис. 15 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_4, $W_{норм}$ и U.

Из рис. 15 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_4 ($U < q_{4W} = 6480$), поэтому атрибут A_4 («наименование улицы») обезличивать не требуется;

– атрибут A_4 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_4 < W_{норм}$;

– нормативное значение для атрибута A_4 является избыточным относительно возможностей нарушителя ($U < q_{4норм} = 1330$).

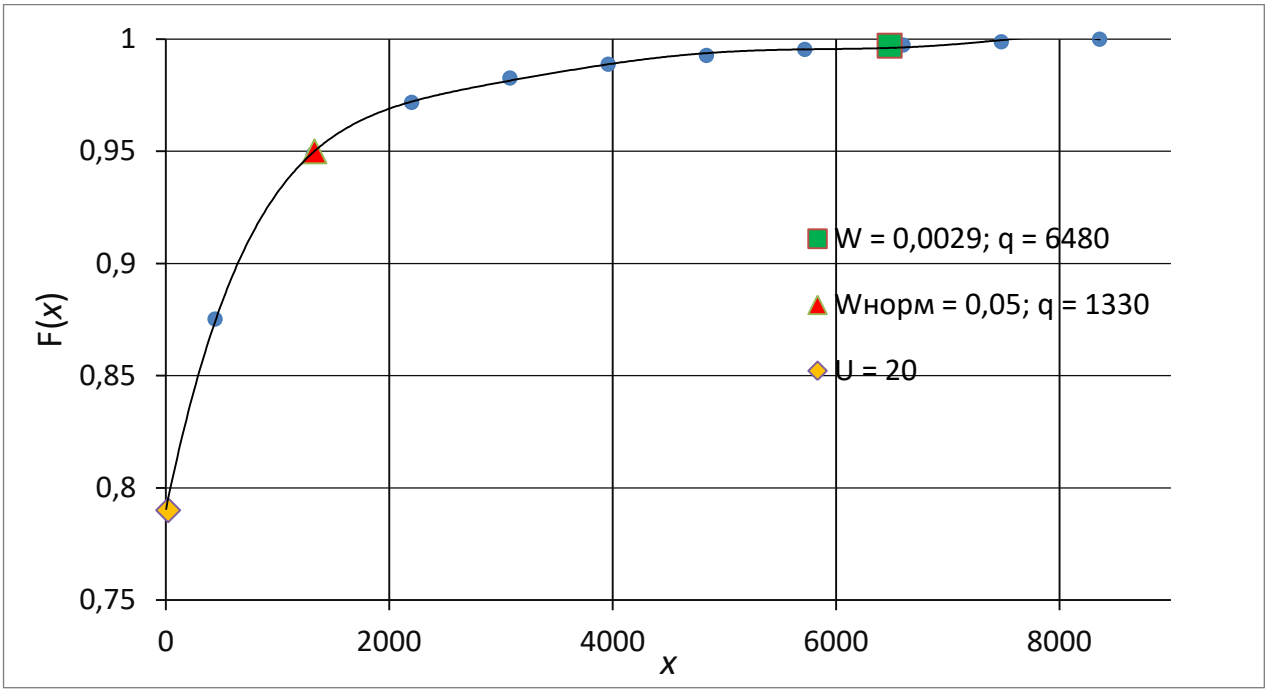

Рис. 15. График распределения вероятности для атрибута «наименование улицы» в сравнении с W_4, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

2.3.5. Распределение характеристик атрибута «номер дома»

Атрибут «номер дома» (A_5) базы данных B_1 в отличие от атрибутов $A_1 - A_4$ имеет в большинстве случаев цифровой формат (исключение составляют номера домов вида «1Б»), кроме того, диапазон его значений ограничен и неравномерен (домов с номером «1» существует в реальности больше, чем домов с номером «100»). Определено количество различных значений атрибута $Q_5 = 731$. Количество записей для каждого из этих значений находится в диапазоне от $q_{5мин} = 1$ до $q_{5макс} = 6207$ (количество записей со значением «4»).

По формуле (7) вычислено значение

$$W_5 = Q_5 / V = 0{,}002357.$$

На рис. 16 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{5d} в зависимости от q_{5k} для атрибута «номер дома» (A_5) базы данных B_1 при $d = 10$ интервалов, где n_{5d} – сумма n_{5k} по каждому интервалу. На рис. 16a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 16b – по

обеим осям (при линейном тренде для степенной функции). Эффект «размытия хвоста» не наблюдается.

На рис. 16a виден линейный характер в средней части диапазона, на рис. 16b – в большей части диапазона. По этой причине, а также с учетом ограниченного диапазона при дальнейшем рассмотрении для аппроксимации вместо экспоненциальной функции применяется гамма-функция вида $y(x) = cx^{ax}e^{-bx}$.

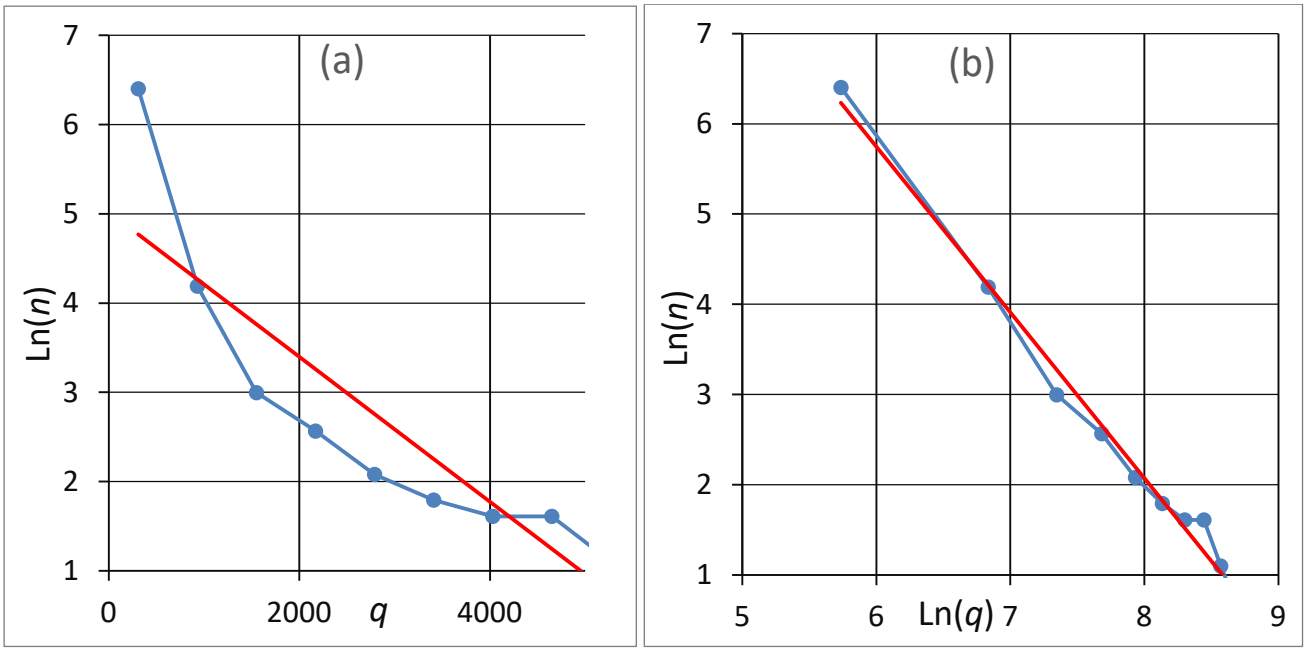

Рис. 16. Диаграмма частот значений n атрибута «номер дома» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В табл. 7 приведены результаты аппроксимации дискретного распределения случайной величины q_5 для атрибута «номер дома» (A_5) базы данных B_1 для 10 интервалов гамма-функцией $f_1(x) = 420 \cdot x^{0,14} \cdot e^{-0,0016x}$ и степенной функцией $f_2(x) = 4{,}00 \cdot 10^7 x^{-1,937}$, где x – среднее значение интервала на оси количества записей q_5, а значение функций f_1 и f_2 – количество номеров домов n_5 в интервале.

Таблица 7

Распределение значений номеров домов в диапазоне 1 … 6200

x	B_1	f_1	f_2
310	603	570,9918	597,8019

x	B_1	f_1	f_2
930	66	94,84602	71,18253
1550	20	35,17215	26,46381
2170	13	13,04304	13,79121
2790	8	4,836807	8,475973
3410	6	1,793654	5,746186
4030	5	0,665148	4,157661
4650	5	0,24666	3,151146
5270	3	0,09147	2,472736
5890	2	0,03392	1,993479
χ^2	16,919	350,7108	3,447975
S_{on}	16,919	86,62501047	3,402200251
$T(Q)$	1,3581	0,912335638	0,319665336

В трех нижних строках табл. 7 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что гамма-функция f_1 отвечает только одному критерию согласия, а степенная функция f_2 отвечает всем трем.

На рис. 17 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_5 для атрибута A_5 «номер дома» базы B_1 (см. рис. 16b).

В результате проверки по критериям первого рода вид гамма-функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

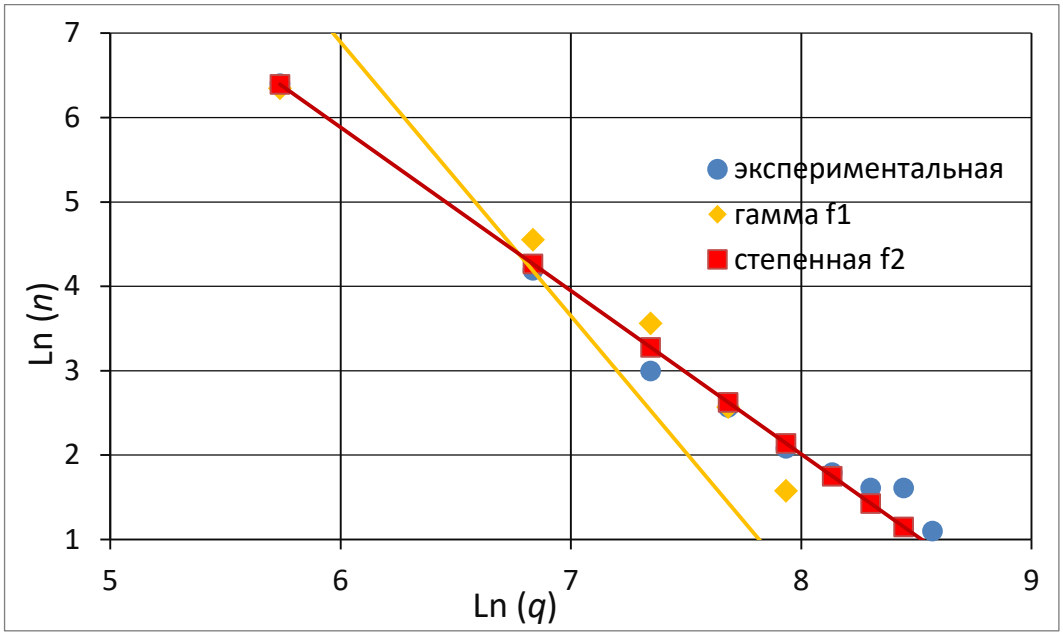

Рис. 17. Функции f_1 и f_2 в сравнении с дискретной последовательностью для атрибута A_5, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате решения уравнения (10) получены значения q_{5W} и $q_{5норм}$. На рис. 18 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_5, $W_{норм}$ и U.

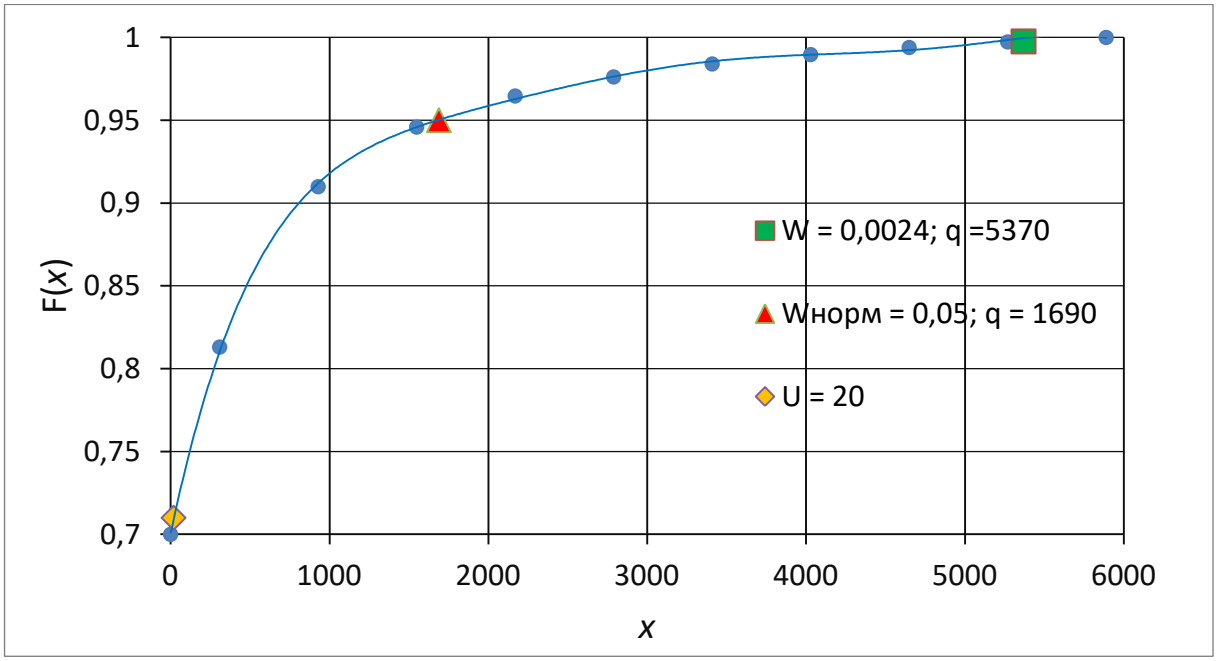

Рис. 18. График распределения вероятности для атрибута «номер дома» в сравнении с W_5, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 18 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_5 ($U < q_{5W} = 5370$), поэтому атрибут A_5 («номер дома») обезличивать не требуется;

– атрибут A_5 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_5 < W_{норм}$;

– нормативное значение для атрибута A_5 является избыточным относительно возможностей нарушителя ($U < q_{5норм} = 1690$).

2.3.6. Распределение характеристик атрибута «номер квартиры»

Атрибут «номер квартиры» (A_6) базы данных B_1 рассматривался аналогично «номеру дома», поскольку подобен ему семантически. Определено количество различных значений атрибута $Q_6 = 978$. Количество записей для каждого из этих значений находится в диапазоне от $q_{6мин} = 1$ до $q_{6макс} = 3889$ (количество записей со значением «2»).

По формуле (7) вычислено значение

$$W_6 = Q_6 / V = 0{,}0031535.$$

На рис. 19 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{6d} в зависимости от q_{6k} для атрибута «номер квартиры» (A_6) базы данных B_1 при $d = 10$ интервалов, где n_{6d} – сумма n_{6k} по каждому интервалу. На рис. 19a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 19b – по обеим осям (при линейном тренде для степенной функции). Эффект «размытия хвоста» не наблюдается.

На рис. 19a виден линейный характер в средней части диапазона, на рис. 19b – в большей части диапазона. По этой причине, а также с учетом семантических особенностей при дальнейшем рассмотрении вместо экспоненциальной функции применяется гамма-функция вида $y(x) = cx^a e^{-bx}$.

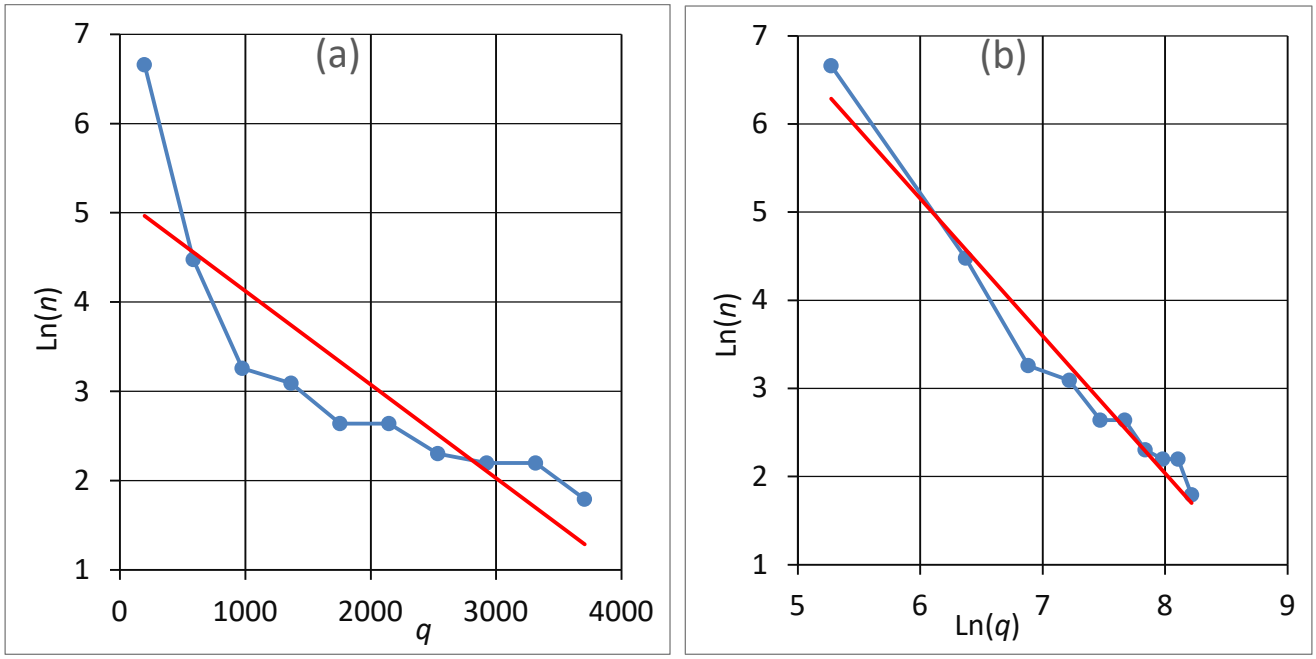

Рис. 19. Диаграмма частот значений *n* атрибута «номер квартиры» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, *q* – количество записей, имеющих одинаковое значение атрибута

В табл. 8 приведены результаты аппроксимации дискретного распределения случайной величины q_6 для атрибута «номер квартиры» (A_6) базы данных B_1 для 10 интервалов гамма-функцией $f_1(x) = 915 \cdot x^{0,07} \cdot e^{-0,0033x}$ и степенной функцией $f_2(x) = 4,42 \cdot 10^6 \cdot x^{-1,658}$, где x – среднее значение интервала на оси количества записей q_6, а значение функций f_1 и f_2 – количество номеров квартир n_6 в интервале.

Таблица 8

Распределение значений номеров квартир в диапазоне 1 … 3900

x	B_1	f_1	f_2
190	760	705,7180184	736,150137
570	98	217,4825485	119,0965982
950	36	64,32037173	51,05921459
1330	22	18,79196643	29,22764656
1710	14	5,457667337	19,26780829
2090	14	1,579436945	13,81456853
2470	10	0,456011252	10,47245057
2850	9	0,131437906	8,260514345

3230	9	0,037837322	6,712465794
3610	6	0,010881692	5,582039776
χ^2	16,919	6645,69906	12,89008919
S_{on}	16,919	391,7682	13,6005
$T(Q)$	1,3581	2,536623	1,272292

В трех нижних строках табл. 8 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что гамма-функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 20 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_6 для атрибута A_6 «номер квартиры» базы B_1 (см. рис. 19b).

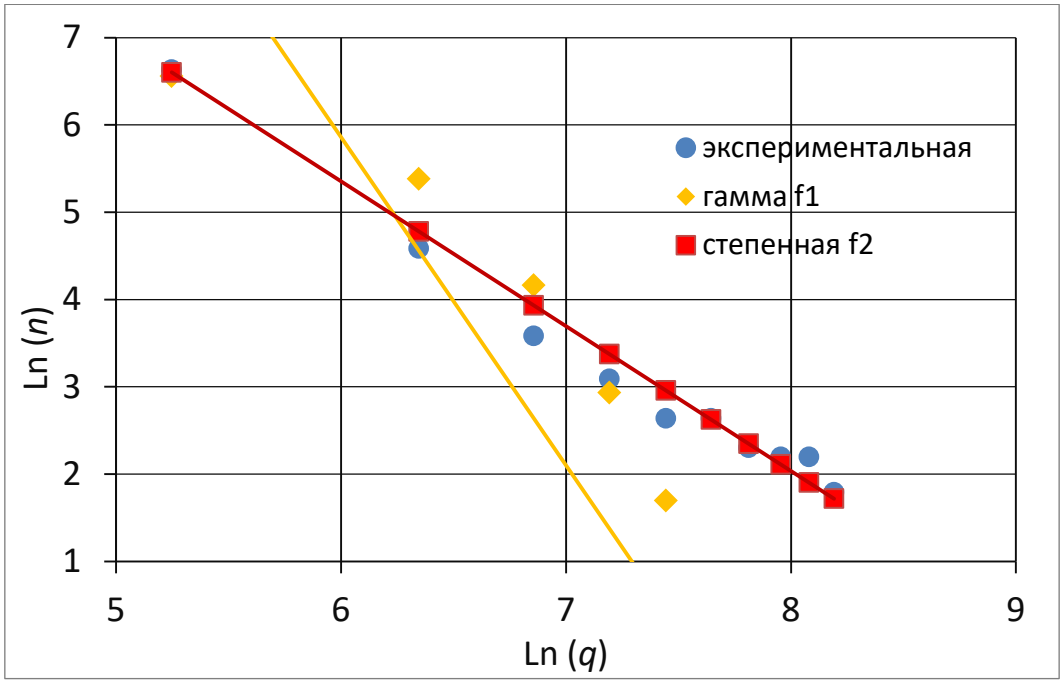

Рис. 20. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для атрибута A_6, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода вид гамма-функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения q_{6W} и $q_{6норм}$. На рис. 21 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_6, $W_{норм}$ и U.

Из рис. 21 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_6 ($U < q_{6W} = 3410$), поэтому атрибут A_6 («номер дома») обезличивать не требуется;

– атрибут A_6 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_6 < W_{норм}$;

– нормативное значение для атрибута A_6 является избыточным относительно возможностей нарушителя ($U < q_{6норм} = 1580$).

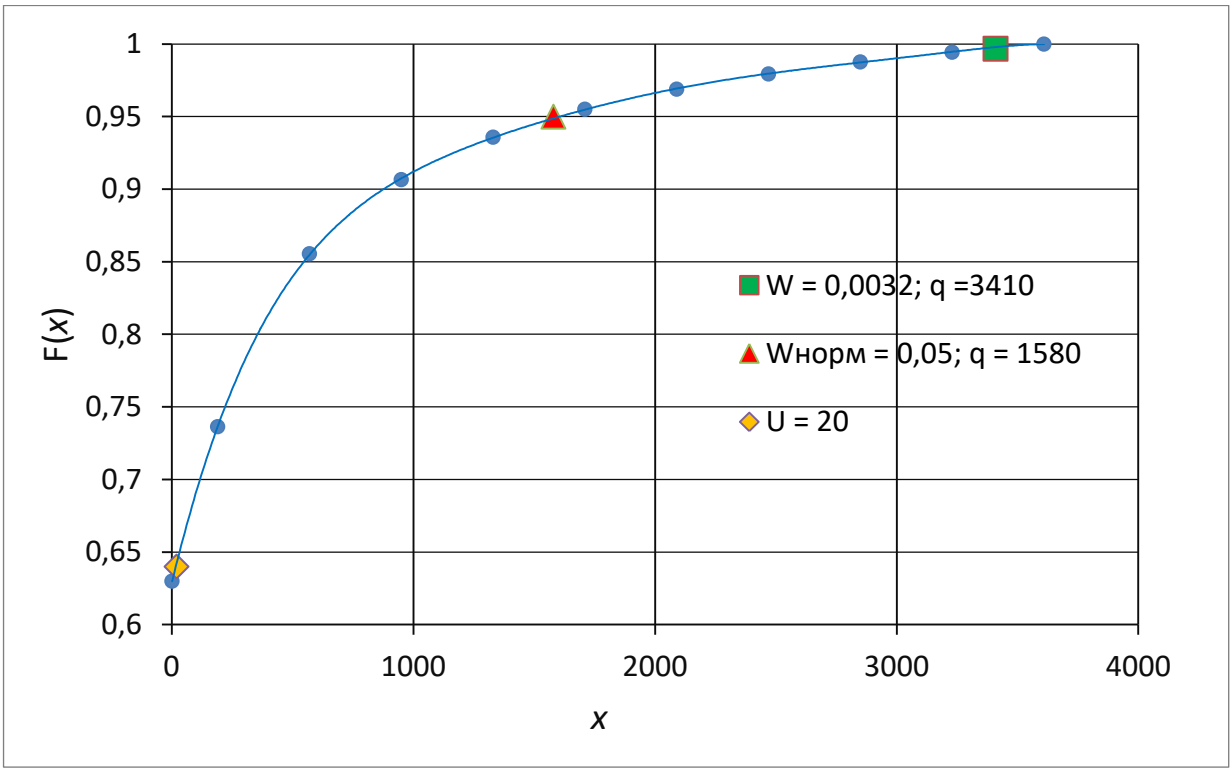

Рис. 21. График распределения вероятности для атрибута «номер квартиры» в сравнении с W_6, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

2.3.7. Распределение характеристик совокупности атрибутов «имя» и «отчество»

Максимально возможное количество вариантов сочетаний «имя+отчество» для базы объемом 310 тыс. записей, при допущении, что все варианты одинаково возможны, равно $Q_{2\cdot 3макс} = Q_2 \cdot Q_3 = 228246$, но это допущение нуждается в проверке. Для проверки был выбран диапазон с наибольшим количеством имен, начинающихся на букву «В» ($V_{2в3} = 45799$ записей) и определено количество различных значений совокупности атрибутов $Q_{2в3} = 2518$ (гипотеза подтвердилась). Количество записей для каждого из этих значений находится в диапазоне от $q_{2в3мин} = 1$ до $q_{2в3макс} = 496$ (количество записей со значением «Виктория Александровна»).

По формуле (7) вычислено значение

$$W_{2в3} = Q_{2в3} / V_{2в3} = 0{,}054979.$$

На рис. 22 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений $n_{2\cdot 3dn}$ в зависимости от $q_{2\cdot 3k}$ для сочетания атрибутов «имя» (A_2) и «отчество» (A_3) базы данных B_1 при $d = 10$ интервалов, где $n_{2\cdot 3d}$ – сумма $n_{2\cdot 3k}$ по каждому интервалу. На рис. 22a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 22b – по обеим осям (при линейном тренде для степенной функции). Эффект «размытия хвоста» не наблюдается.

На рис. 22a виден линейный характер в средней части диапазона, на рис. 22b – в большей части диапазона.

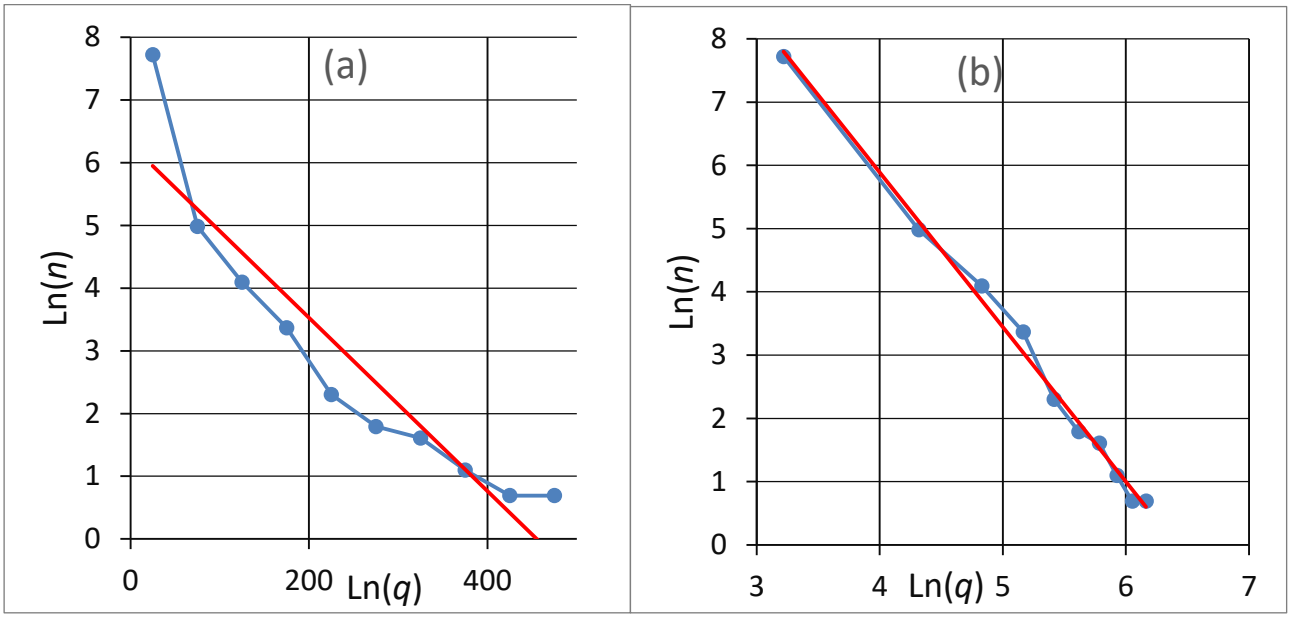

Рис. 22 Диаграмма частот значений *n* сочетания атрибутов «имя» и «отчество» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, *q* – количество записей, имеющих одинаковое значение атрибута, *n* – количество различных значений атрибута

В табл. 9 приведены результаты аппроксимации дискретного распределения случайной величины $q_{2\cdot3}$ для сочетания атрибутов «имя» и «отчество» базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 4340 \cdot e^{-0{,}0026x}$ и степенной функцией $f_2(x) = 6{,}32 \cdot 10^6 \cdot x^{-2{,}443}$, где x – среднее значение интервала на оси количества записей $q_{2\cdot3}$, значение функций f_1 и f_2 – количество фамилий $n_{2\cdot3}$ в интервале.

Таблица 9

Распределение значений пар «имя» и «отчество» в диапазоне 1 … 500

x	B_1	f_1	f_2
25	2255	2265,679	2431,555
75	146	617,4695	166,0645
125	60	168,2801	47,67607
175	29	45,86167	20,95606
225	10	12,49876	11,34148
275	6	3,40631	6,946432
325	5	0,928328	4,618708
375	3	0,252999	3,256066

425	2	0,06895	2,398266
475	2	0,018791	1,827633
χ^2	16,919	785,1822	10,82162
S_{on}	16,919	544,0476	9,699736
$T(Q)$	1,3581	8,434309	0,48902

В трех нижних строках табл. 9 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 23 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений $n_{2\cdot3}$ для сочетания атрибутов A_2 и A_3 «имя» и «отчество» базы B_1 (см. рис. 22b).

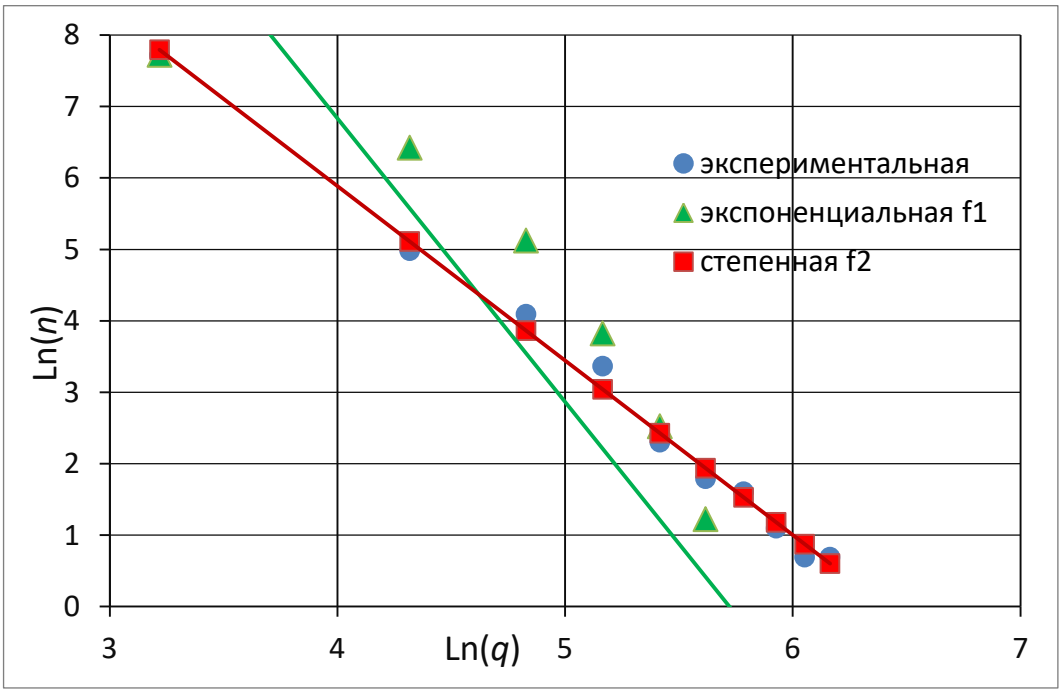

Рис. 23. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для сочетания атрибутов A_2 и A_3,где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения $q_{2в3W}$ и $q_{2в3норм}$. На рис. 24 приведен график функции распределения вероятности $F(x)$ и отмечены значения $W_{2в3}$, $W_{норм}$ и U.

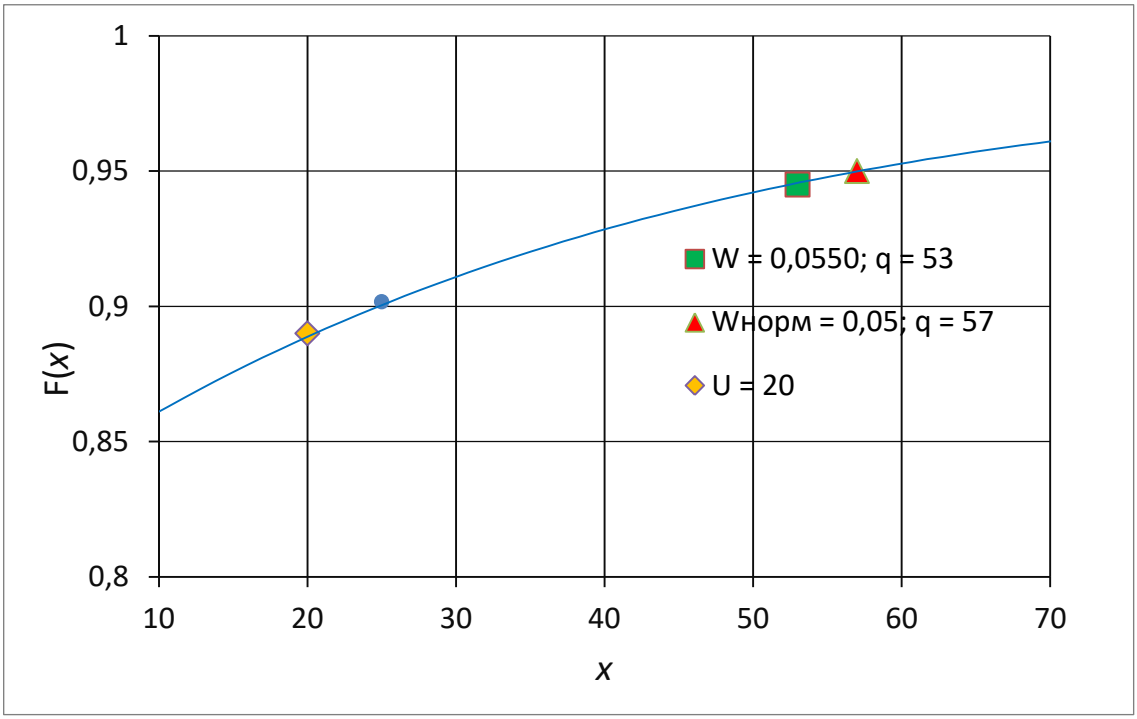

Рис. 24. График распределения вероятности для сочетания атрибутов «имя» и «отчество» в сравнении с $W_{2в3}$, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 24 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным сочетанием атрибутов A_2 и A_3 в базе объема 45799 записей ($U < q_{2в3W} = 53$), поэтому сочетание атрибутов A_2 и A_3 («имя» и «отчество») обезличивать не требуется;

– сочетание атрибутов A_2 и A_3 необходимо обезличивать в соответствии с нормативным значением, поскольку $W_{2в3} > W_{норм}$;

– нормативное значение для сочетания атрибутов A_2 и A_3 является избыточным относительно возможностей нарушителя ($U < q_{2в3норм} = 57$).

Аналогичный расчет был произведен для инициалов (сочетание первой буквы имени и первой буквы отчества). При допущении отсутствия зависимости между атрибутами максимально возможное количество вариантов сочетаний равно 27·27 = 729 (исключены буквы «й», «щ», «ъ», «ы», «ь»), но допущение нуждается в проверке. Экспериментальная проверка показала, что количество различных значений сочетания инициалов $Q_{ИО}$ = 654, то есть допущение было неверным. Количество записей для каждого из этих значений находится в диапазоне от $q_{ИОмин}$ = 1 до $q_{ИОмакс}$ = 8032 (количество записей с инициалами «В.А.»).

По формуле (7) вычислено значение

$$W_{ИО} = Q_{ИО} / V = 0{,}002109.$$

На рис. 25 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений $n_{ИОd}$ в зависимости от $q_{ИОk}$ для сочетания первых символов атрибутов «имя» и «отчество» при d = 10 интервалов, где $n_{ИОd}$ – сумма $n_{ИОk}$ по каждому интервалу. На рис. 25a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальной функции), а на рис. 25b – по обеим осям (при линейном тренде для степенной функции). Эффект «размытия хвоста» не наблюдается.

На рис. 25a виден линейный характер в средней части диапазона, на рис. 25b – в большей части диапазона.

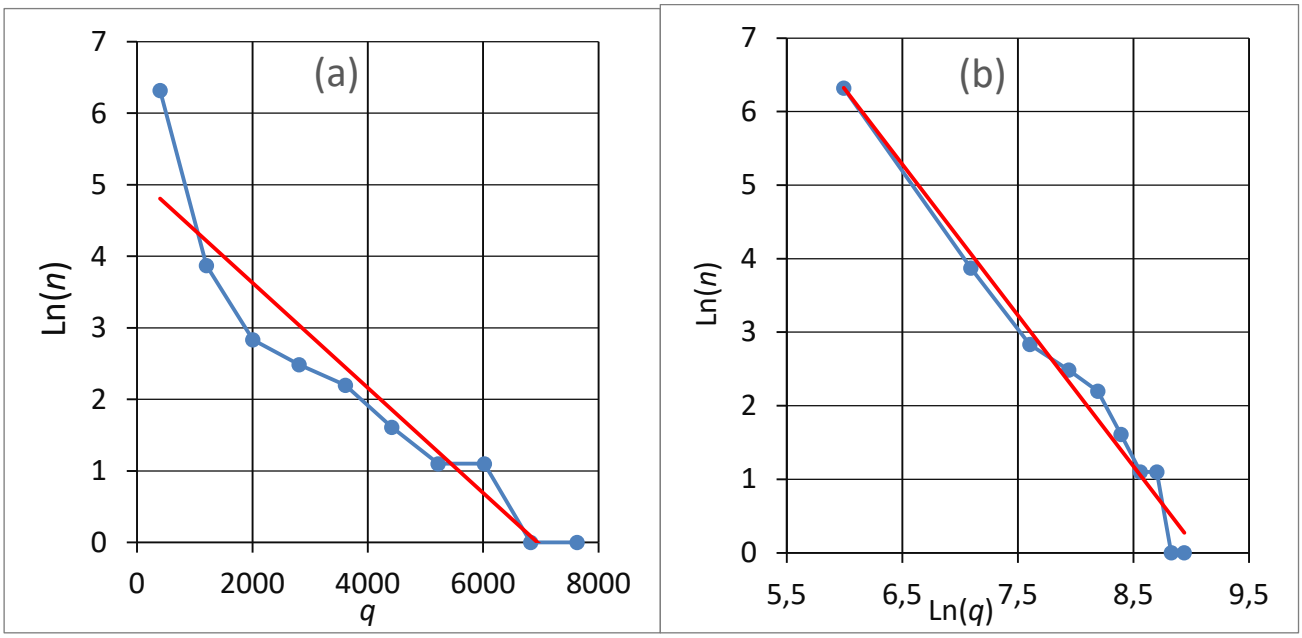

Рис. 25. Диаграмма частот значений *n* сочетания инициалов «имя» и «отчество» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда. *q* – количество записей, имеющих одинаковое значение атрибута, *n* – количество различных значений атрибута

В табл. 10 приведены результаты аппроксимации дискретного распределения случайной величины $q_{ИО}$ для сочетания инициалов «имя» и «отчество» базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 1065 \cdot e^{-0,0017x}$ и степенной функцией $f_2(x) = 1,24 \cdot 10^8 \cdot x^{-2,054}$, где x – среднее значение интервала на оси количества записей $q_{ИО}$, а значение функций f_1 и f_2 – количество фамилий $n_{ИО}$ в интервале.

Таблица 10

Распределение значений наименований улиц в диапазоне 1 … 8000

x	B_1	f_1	f_2
401	555	538,6306	557,467
1204	48	137,5421	58,27347
2007	17	35,12208	20,40071
2810	12	8,968604	10,21962
3613	9	2,290179	6,098414
4416	5	0,584809	4,038201
5219	3	0,149334	2,86519

6022	3	0,038133	2,135454
6825	1	0,009738	1,651318
7628	1	0,002487	1,314033
χ^2	16,919	996,3678	4,90963
S_{on}	16,919	157,8367	4,781549
$T(Q)$	1,3581	2,659076	0,2468

В трех нижних строках табл. 10 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

На рис. 26 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений $n_{ИО}$ для сочетания первых символов атрибутов A_2 и A_3 «имя» и «отчество» базы B_1 (см. рис. 25b).

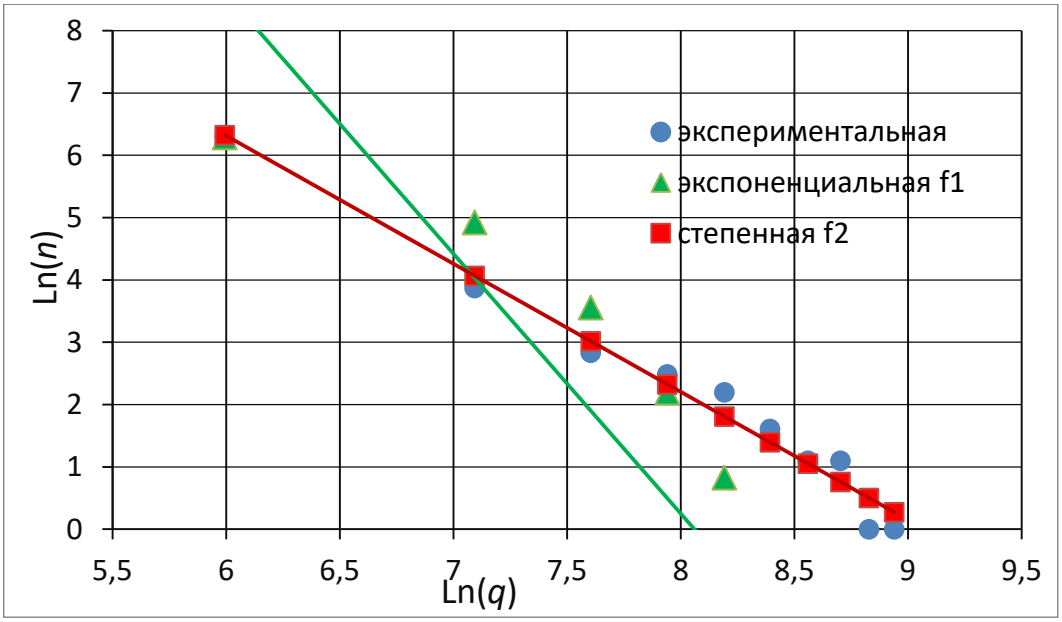

Рис. 26. Функции f_1 и f_2 в сравнении с дискретной экспериментальной последовательностью для первых символов атрибутов A_2 и A_3,где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

Из рис. 26 видно, что линия тренда $n_{ИО}$ совпадает с графиком f_2.

В результате проверки по критериям первого рода экспоненциальный вид функций отвергнут, а степенной вид – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

В результате решения уравнения (10) получены значения $q_{ИОW}$ и $q_{ИОнорм}$. На рис. 27 приведен график функции распределения вероятности $F(x)$ и отмечены значения $W_{ИО}$, $W_{норм}$ и U.

Из рис. 27 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным сочетанием атрибутов A_2 и A_3 ($U < q_{ИОW} = 6790$), поэтому сочетание атрибутов A_2 и A_3 («инициалы») обезличивать не требуется;

– сочетание атрибутов A_2 и A_3 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_{ИО} > W_{норм}$;

– нормативное значение для атрибута сочетания атрибутов A_2 и A_3 является избыточным относительно возможностей нарушителя ($U < q_{ИОнорм} = 1690$).

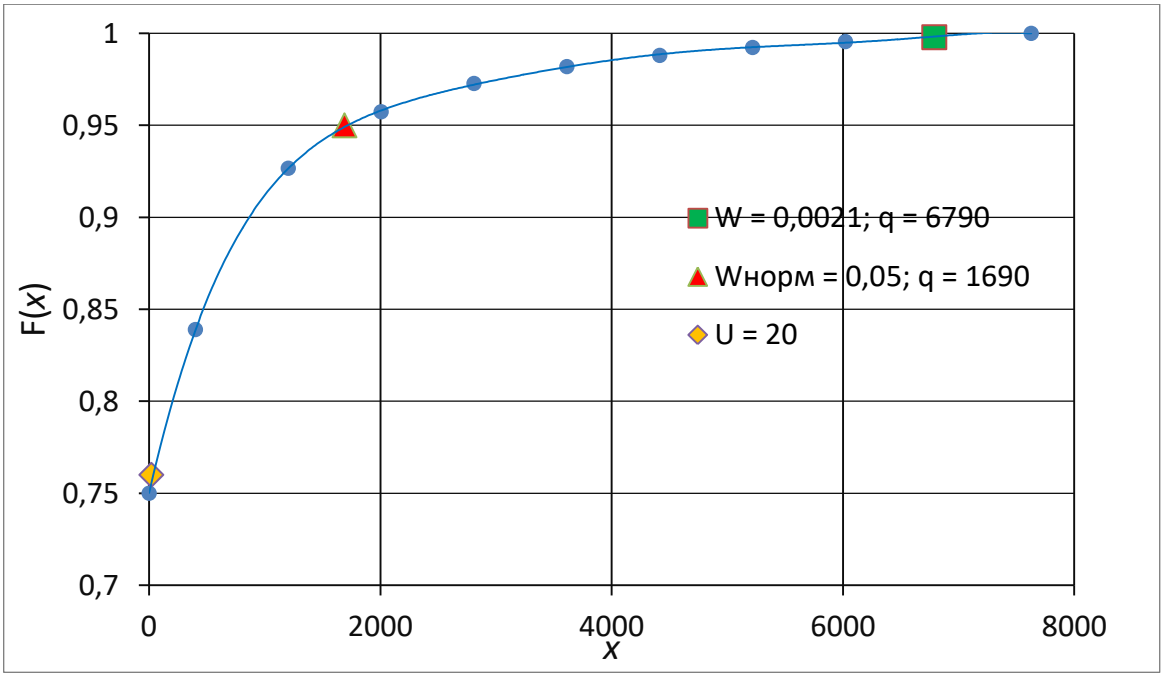

Рис. 27. График распределения вероятности для сочетания атрибутов «имя» и «отчество» в сравнении с $W_{ИО}$, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

2.3.8. Распределение характеристик атрибута «дата рождения»

Атрибут «дата рождения» (A_8) базы данных B_2 (V = 329115 записей) рассматривался следующим образом. Определено количество различных значений атрибута Q_8 = 35885. Количество записей для каждого из этих значений находится в диапазоне от $q_{8мин}$ = 1 до $q_{8макс}$ = 147 (количество записей со значением «1950-01-01»).

По формуле (7) вычислено значение

$$W_8 = Q_8 / V = 0{,}109012.$$

На рис. 28 в логарифмическом масштабе по обеим координатам представлена диаграмма частот значений n_{8d} в зависимости от q_{8k} для атрибута «дата рождения» базы данных B_2 для d = 12 интервалов, где $n_{ИОd}$ – сумма $n_{ИОk}$ по каждому интервалу. На рис. 28a использован логарифмический масштаб по одной оси (при линейном тренде для экспоненциальная функция), а на рис. 28b – по обеим осям (при линейном тренде для степенной функции).

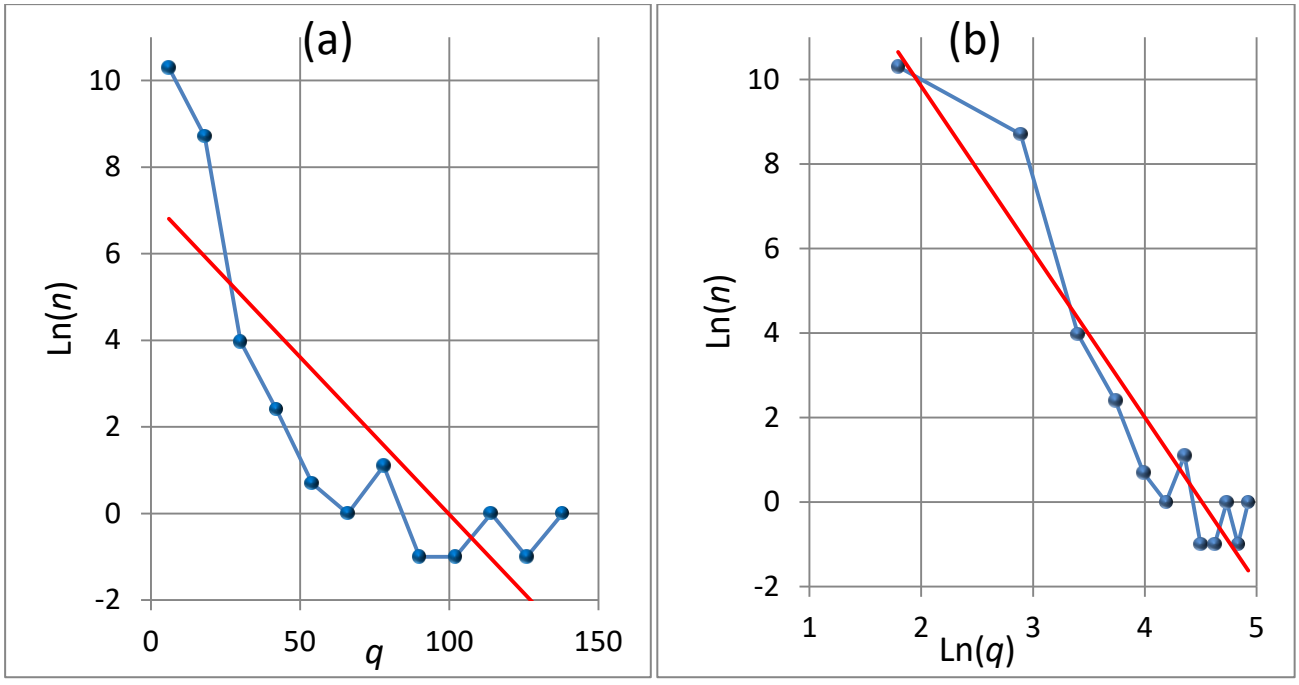

Рис. 28. Диаграмма частот значений n атрибута «дата рождения» в логарифмическом масштабе, a – по одной оси, b – по обеим осям, где красным обозначена линия тренда, q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

На рис. 28a виден линейный характер в средней части диапазона, на рис. 28b – в большей части диапазона, при этом четко виден эффект «размытия хвоста», поэтому принято решение об изменении размера интервалов (Табл.11).

В табл. 11 приведены результаты аппроксимации дискретного распределения случайной величины q_8 для атрибута «дата рождения» базы данных B_2 для 11 интервалов экспоненциальной функцией $f_1(x) = 21800 \cdot e^{-0,121x}$ и степенной функцией $f_2(x) = 3,99 \cdot 10^4 x^{-0,98}$ и гамма-функцией $f_3(x) = 3100 \cdot x^{2,27} \cdot e^{-0,39x}$, где x – среднее значение интервала на оси количества записей q_8, а значение функций f_1, f_2 и f_3 – количество дат рождения n_8 в интервале.

Таблица 11

Распределение значений наименований улиц в диапазоне 1 … 100

x	B_2	f_1	f_2	f_3
2,5	9372	16109,51304	16275,55	9359,4083
7,5	16242	8796,993099	5545,705	16122,76
12,5	7431	4803,812963	3361,592	7314,1157
17,5	2166	2623,239409	2417,35	2233,5631
22,5	520	1432,483956	1889,635	562,18847
27,5	111	782,2428551	1552,282	126,13587
32,5	26	427,1628186	1317,866	26,221309
37,5	6	233,2626913	1145,424	5,1624502
42,5	2	127,3787904	1013,201	0,9758286
47,5	1	69,55829993	908,5672	0,1787108
98,5	0,4	0,145317431	444,5793	0,0394211
χ^2	18,307	12410,39	35595,89	17,56688053
S_{on}	18,307	12768,83	33624,49	13,40005
$T(Q)$	1,3581	36,7033	41,48699	0,699699

В трех нижних строках табл. 11 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_2) [85] и рассчитанные для f_1, f_2 и f_3 по формулам (11) – (13) при заданном уровне

значимости α = 0,05. Необходимо отметить, что параметры функций f_1, f_2 и f_3 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная фугкция f_1 и степенная функция f_2 не отвечают ни одному из критериев согласия, а гамма-функция f_3 отвечает всем трем.

На рис. 29 в логарифмическом масштабе по обеим координатам приведены графики функций f_1 и f_2 в сравнении с дискретной последовательностью экспериментальных значений n_8 для атрибута A_8 «дата рождения» базы B_2 (см. рис. 28b).

Из рис. 29 видно, что линия тренда n_8 наиболее близка к графику f_3.

В результате проверки по критериям первого рода экспоненциальный вид и степенной вид функций отвергнут, а вид гамма-функций – не отвергнут и в итоге принят вследствие отсутствия альтернативы для сравнения по формуле (14).

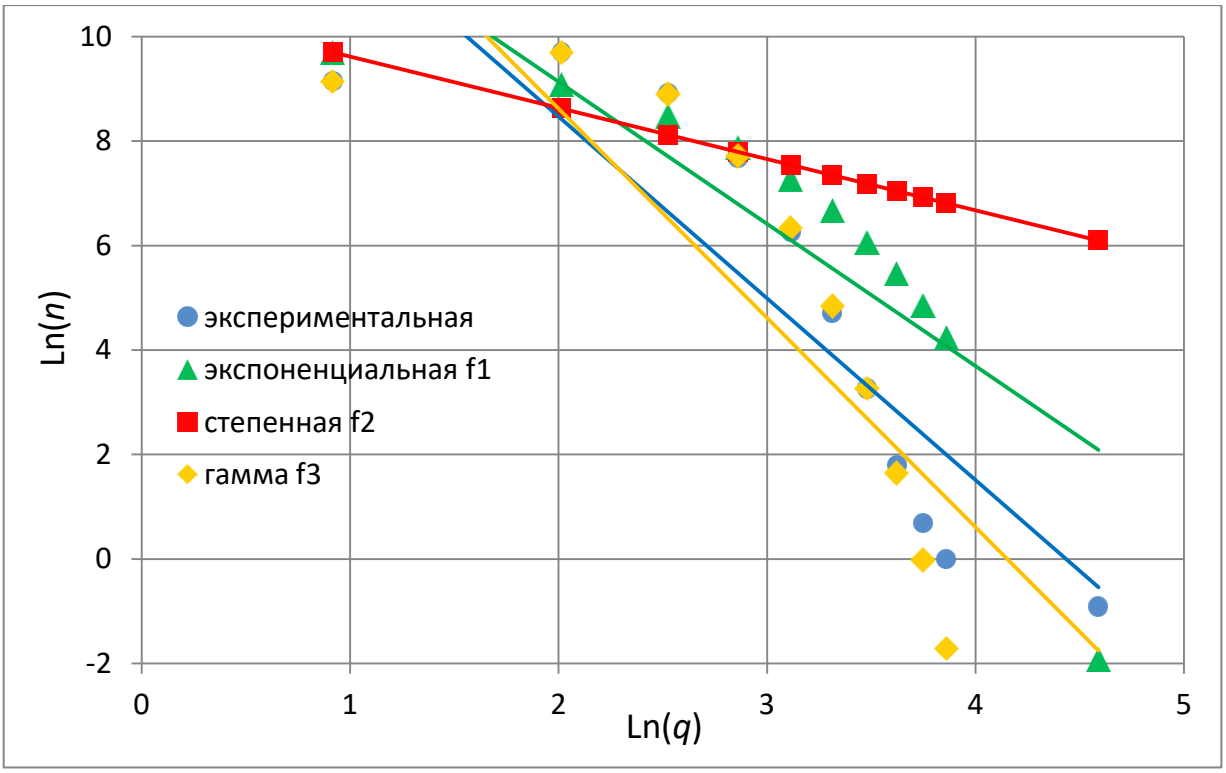

Рис. 29. Функции f_1, f_2 и f_3 в сравнении с дискретной экспериментальной последовательностью для атрибута A_8 базы B_2, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате решения уравнения (10) получены значения q_{8W} и $q_{8норм}$. На рис. 30 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_8, $W_{норм}$ и U.

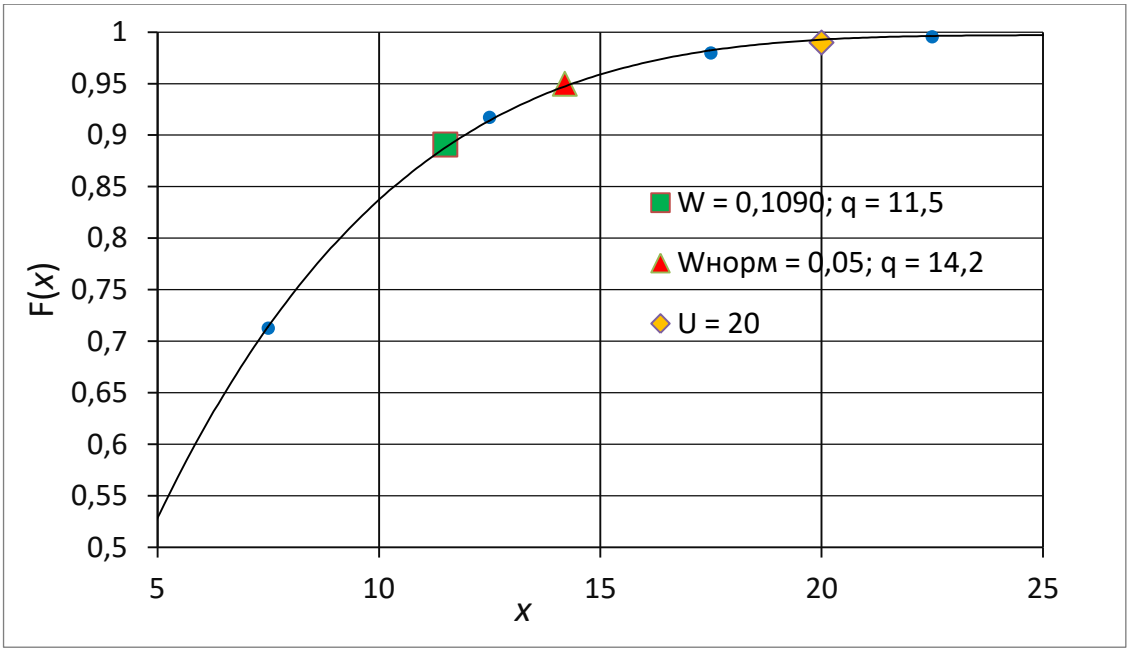

Рис. 30. График распределения вероятности для атрибута «дата рождения» в сравнении с W_8, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 30 можно сделать следующие выводы:

– возможности нарушителя позволяют ему успешно работать с необезличенным атрибутом A_8 ($U > q_{8W} = 11{,}5$), поэтому атрибут A_8 («дата рождения») необходимо обезличивать;

– атрибут A_8 необходимо обезличивать в соответствии с нормативным значением, поскольку $W_8 > W_{норм}$;

– нормативное значение для атрибута A_8 является недостаточным относительно возможностей нарушителя ($U > q_{8норм} = 14{,}2$), возможна опасная ситуация в случае $W_{норм} > W_8$.

Результаты расчетов вероятности идентификации W для различных атрибутов сведены в табл. 12.

Из таблицы видно, что для атрибутов «имя» и «отчество» в отдельности обезличивание является необязательным, а для их совокупности – обязательным.

Таблица 12

Вероятность идентификации *W* для различных атрибутов

Атрибут	**Кол-во значений Q**	**Вероятность идентификации W**
Фамилия	45099	0,145419
Имя	654	0,002434
Отчество	349	0,001116
Улица	890	0,002863
Номер дома	731	0,002357
Номер квартиры	978	0,003154
Имя + Отчество	2518	0,054979
ИО	654	0,002109
Дата рождения	35885	0,109012

Расчеты для прочих атрибутов-идентификаторов ФЛ (место рождения и др.), а также прочих сочетаний атрибутов не вошли в рамки представленной работы и являются предметом дальнейших исследований.

2.4. Зависимость вероятности идентификации от количества записей базы данных

Согласно алгоритму создания модели идентификации было сделано $b_1 = 9$ дополнительных случайных выборок (по географическим признакам) из базы данных B_1 и $b_2 = 6$ дополнительных случайных выборок из базы данных B_2. В процессе исследований для различного количества записей базы решались две задачи:

1) определение влияния количества записей базы V_b на вид диаграммы частот значений n_{jd} в зависимости от q_{jk} и параметры аппроксимирующей функции для различных атрибутов;
2) определение вида зависимости вероятности идентификации W_j по атрибуту в целом от количества записей базы V и параметры аппроксимирующей функции для различных атрибутов.

Результаты расчетов показали, что с уменьшением количества записей базы V для всех атрибутов значение Q_j уменьшилось, но значение W_j при этом увеличилось.

Полученные результаты по каждому атрибуту приведены ниже.

2.4.1. Зависимость параметров атрибута «фамилия» от количества записей базы данных

Для атрибута «фамилия» диаграмма частот значений n_d в зависимости от q_k на всех рассмотренных объемах V_b сохраняет степенной вид, как это показано в табл. 2 (атрибут A_1 в объеме V_{max} = 310132 записи базы данных B_1) и табл. 3 (атрибут A_7 в объеме V_{10} = 87551 записи базы данных B_2). Построена диаграмма частот W_1 в диапазоне от V_{min} = 1666 записей до V_{max} = 310132 записей. На рис. 31 приводится диаграмма частот значений W_1 в зависимости от V в обычном (рис. 31a) и логарифмическом (рис. 31b) масштабе по обеим осям. На рис. 31b виден линейный характер зависимости в большей части диапазона.

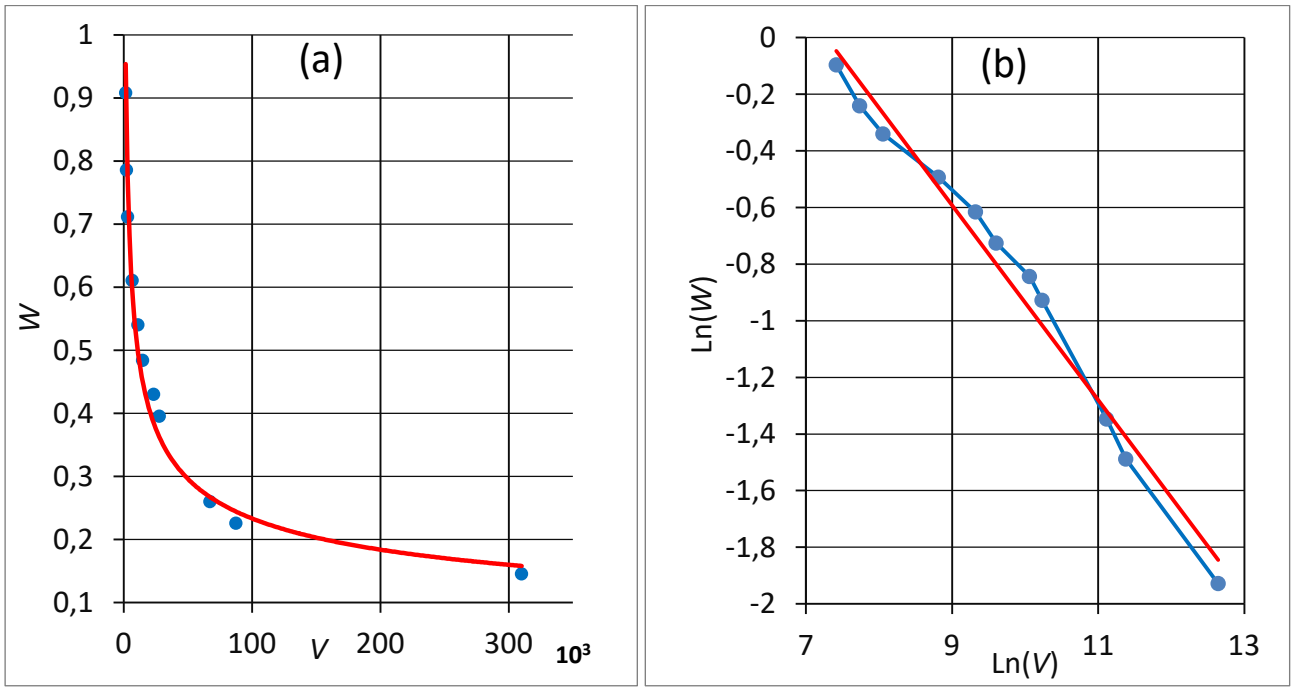

Рис. 31. Диаграмма зависимости вероятности идентификации по атрибуту «фамилия» в обычном (a) и логарифмическом (b) масштабе координат, где красным обозначена линия тренда. V – количество записей базы, W – вероятность идентификации

В табл. 13 приведены результаты аппроксимации дискретного распределения значений W_1 в зависимости от V для атрибута «фамилия» базы данных B_1 для 11 значений V экспоненциальной функцией $f_1(x) = 0{,}6632 \cdot e^{-0{,}000007x}$, степенной функцией $f_2(x) = 12{,}2435 \cdot x^{-0{,}344}$ и логарифмической функцией $f_3(x) = -0{,}136 \cdot \ln(x) + 1{,}8597$, где x – значение на оси количества записей базы V, а значение функций f_1, f_2 и f_3 – значение критерия W_1.

Таблица 13

Распределение значений вероятности идентификации в диапазоне 1666 … 310132

x	W_1	f_1	f_2	f_3
310132	0,145418725	0,07565266	0,158057698	0,140013552
87551	0,225731288	0,359322245	0,244213605	0,312023161
67308	0,259983954	0,414022617	0,267333102	0,347783324
27853	0,395397264	0,545719771	0,362136013	0,46778135
23429	0,430150668	0,56288402	0,384337169	0,49130474
14847	0,484003502	0,597735055	0,449638988	0,553344778
11194	0,540289441	0,613216837	0,495514917	0,591753885
6733	0,610723303	0,632667877	0,590206399	0,660890452
3155	0,711568938	0,648713781	0,766040791	0,763982847
2293	0,785870039	0,652639953	0,85492631	0,807384186
1666	0,908163265	0,655510685	0,954226511	0,850827408
χ^2	18,307	0,41264	0,030192	0,054443
S_{on}	18,307	0,39313	0,029903	0,05515
$T(Q)$	1,3581	0,23205	0,06650	0,07910

В трех нижних строках табл. 13 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке W_1) [85] и рассчитанные для f_1, f_2 и f_3 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Из таблицы видно, что и экспоненциальная функция f_1, и степенная функция f_2, и логарифмическая функция f_3 отвечают всем трем

критериям согласия. Причем все три функции отвечают всем критериям согласия даже при уровне значимости α = 0,5.

На рис. 32 в логарифмическом масштабе по обеим координатам приведены графики функций f_1, f_2 и f_3 в сравнении с дискретной последовательностью экспериментальных значений для критерия W_1 атрибута A_1 «фамилия» базы B_1 (см. рис. 31b).

Из рис. 32 видно, что линия тренда W_1 совпадает с графиком f_2.

В результате проверки по критериям первого рода все три вида функций не были отвергнуты, в связи с этим по формуле (14) для функций f_1, f_2 и f_3 были рассчитаны значения дисперсии ошибок. Сравнение показало, что значение дисперсии $s_2^2 = 7865423180$ для степенной функции f_2 является минимальным, дисперсия логарифмической функции s_3^2 и экспоненциальной функции s_1^2 превышают s_2^2, соответственно, в 1,34 и в 3,23 раза. В результате выбрана степенная зависимость W_1 от V.

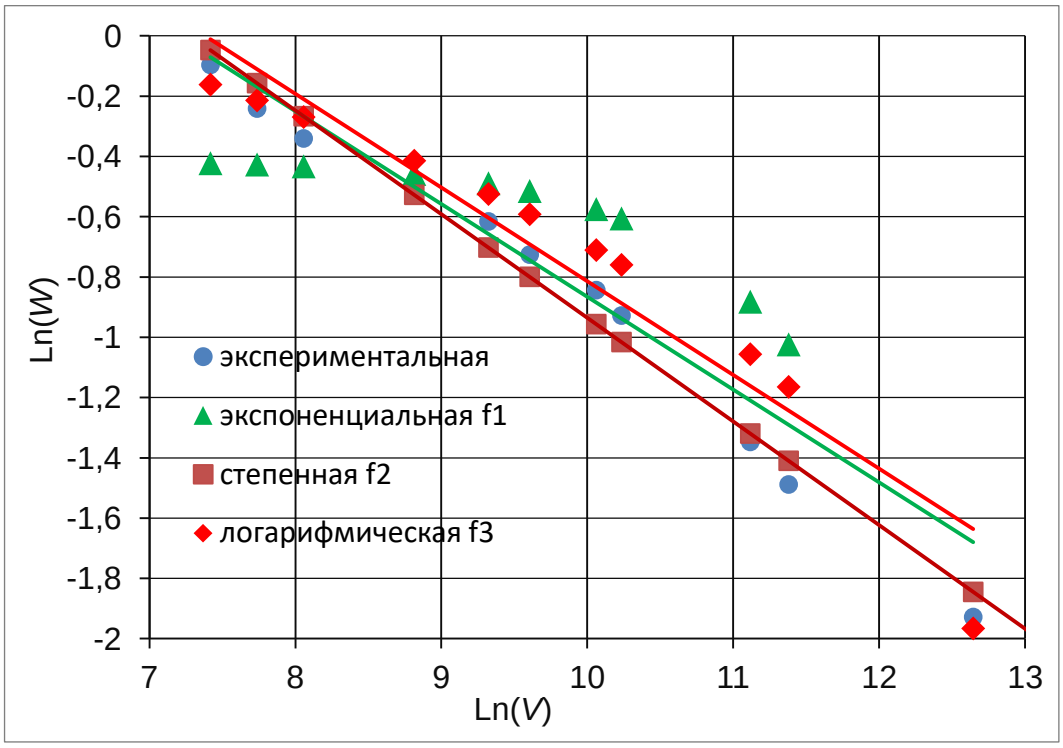

Рис. 32. Функции f_1, f_2 и f_3 в сравнении с экспериментальной дискретной последовательностью для W_1 атрибута A_1 базы B_1, где V – количество записей базы, W – вероятность идентификации

На рис. 33 приведен график степенной функции зависимости W_1 от V с целью экстраполяции поведения этой функции до значений V_{ext} = 1млн записей относительно $W_{норм}$.

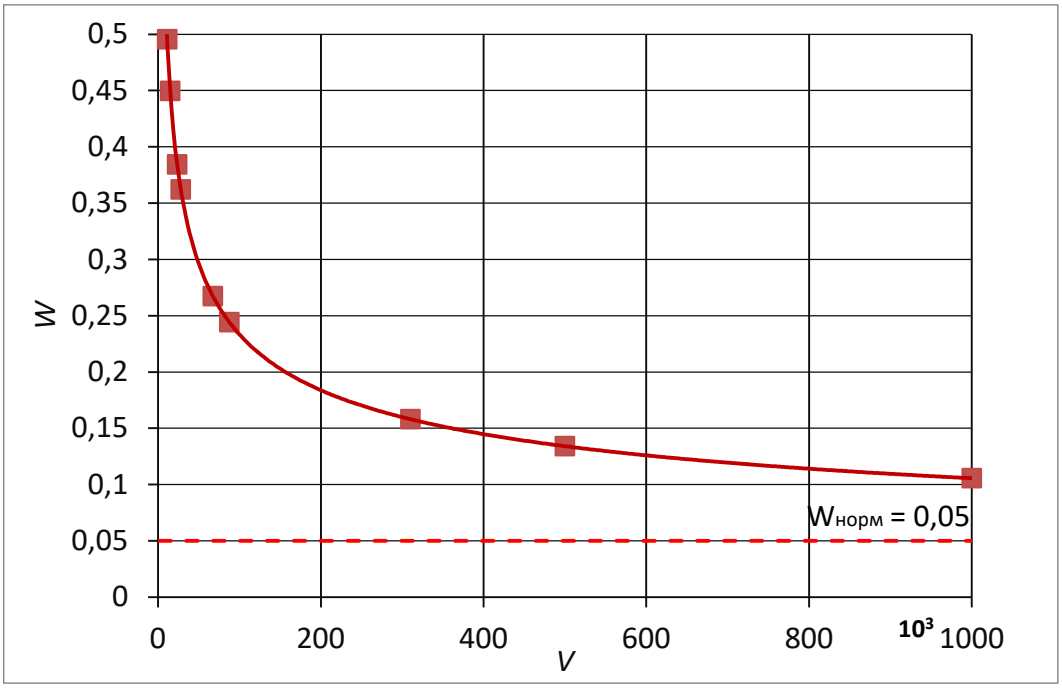

Рис. 33. Экстраполяция степенной функции f_2 до значений V = 1000000 записей в сравнении с $W_{норм}$, где V – количество записей базы, W – вероятность идентификации

Из рис. 33 можно сделать вывод, что для БД объема 1 миллион записей обезличивание по атрибуту «фамилия» является обязательным.

2.4.2. Зависимость параметров атрибута «имя» от количества записей базы данных

Для атрибута «имя» диаграмма частот значений n_d в зависимости от q_k на всех рассмотренных объемах V_b базы B_1 сохраняет степенной вид, как это показано в табл. 4 (атрибут A_2 в объеме V_{max} = 310132 записи базы данных B_1) и табл. 14 (атрибут A_2 в объеме V_8 = 27853 записей базы данных B_1).

Количество различных значений в объеме 27853 записей Q_2 = 564.

По формуле (7) вычислено значение

$$W_2 = Q_2 / V_8 = 0{,}020249165.$$

В табл. 14 приведены результаты аппроксимации дискретного распределения случайной величины q_2 для атрибута «имя» (A_2) в объеме 27853

записей базы данных B_1 для 10 интервалов экспоненциальной функцией $f_1(x) = 722 \cdot e^{-0,0121x}$ и степенной функцией $f_2(x) = 4,17 \cdot 10^5 \cdot x^{-1,884}$, где x – среднее значение интервала на оси количества записей q_2, а значение функций f_1 и f_2 – количество фамилий n_2 в интервале.

Таблица 14

Распределение значений имен в диапазоне 1 … 743

x	B_1	f_1	f_2
37	463	461,4275894	462,6776783
111	42	188,4674172	58,39586606
185	17	76,97842124	22,30586449
259	12	31,44138878	11,83351703
333	9	12,84205252	7,370308387
407	7	5,245261717	5,050038963
481	6	2,142396664	3,686457844
555	3	0,875049466	2,815286616
629	3	0,357408869	2,223887326
703	2	0,145981575	1,803461035
χ^2	16,919	274,6636659	8,601541266
S_{on}	16,919	148,3673662	8,638412392
$T(Q)$	1,3581	5,445329235	0,490799176

В трех нижних строках табл. 14 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_1) [85] и рассчитанные для f_1 и f_2 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0,05$. Необходимо отметить, что параметры функций f_1 и f_2 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 не отвечает ни одному из критериев согласия, а степенная функция f_2 отвечает всем трем.

В результате решения уравнения (10) получены значения q_{2W} и $q_{2норм}$. На рис. 34 приведен график функции распределения вероятности $F(x)$ в объеме $V_8 = 27853$ записей базы данных и отмечены значения W_2, $W_{норм}$ и U.

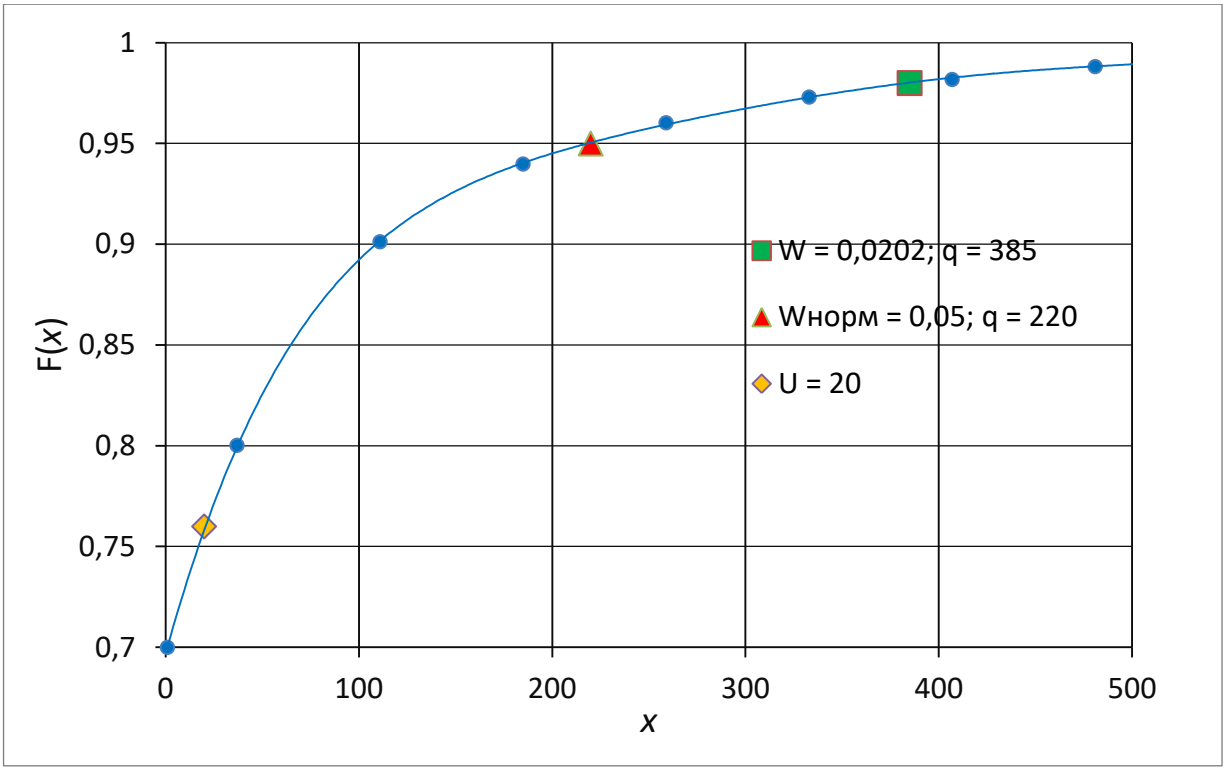

Рис. 34. График распределения вероятности для атрибута «имя» в сравнении с W_2, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Из рис. 34 можно сделать следующие выводы:

– возможности нарушителя не позволяют ему успешно работать с необезличенным атрибутом A_2 ($U < q_{2W} = 385$), поэтому атрибут A_2 («имя») обезличивать не требуется;

– атрибут A_2 не нужно обезличивать в соответствии с нормативным значением, поскольку $W_2 < W_{норм}$;

– нормативное значение для атрибута A_2 является избыточным относительно возможностей нарушителя ($U < q_{2норм} = 220$).

Для определения зависимости W_2 от количества записей БД построена диаграмма частот W_2 в диапазоне от $V_{min} = 1666$ записей до $V_{max} = 310132$ записей. На рис. 35 приводится диаграмма частот значений W_2 от V в обычном (рис. 35а) и

логарифмическом (рис. 35b) масштабе по обеим осям. На рис. 35b виден линейный характер в большей части диапазона.

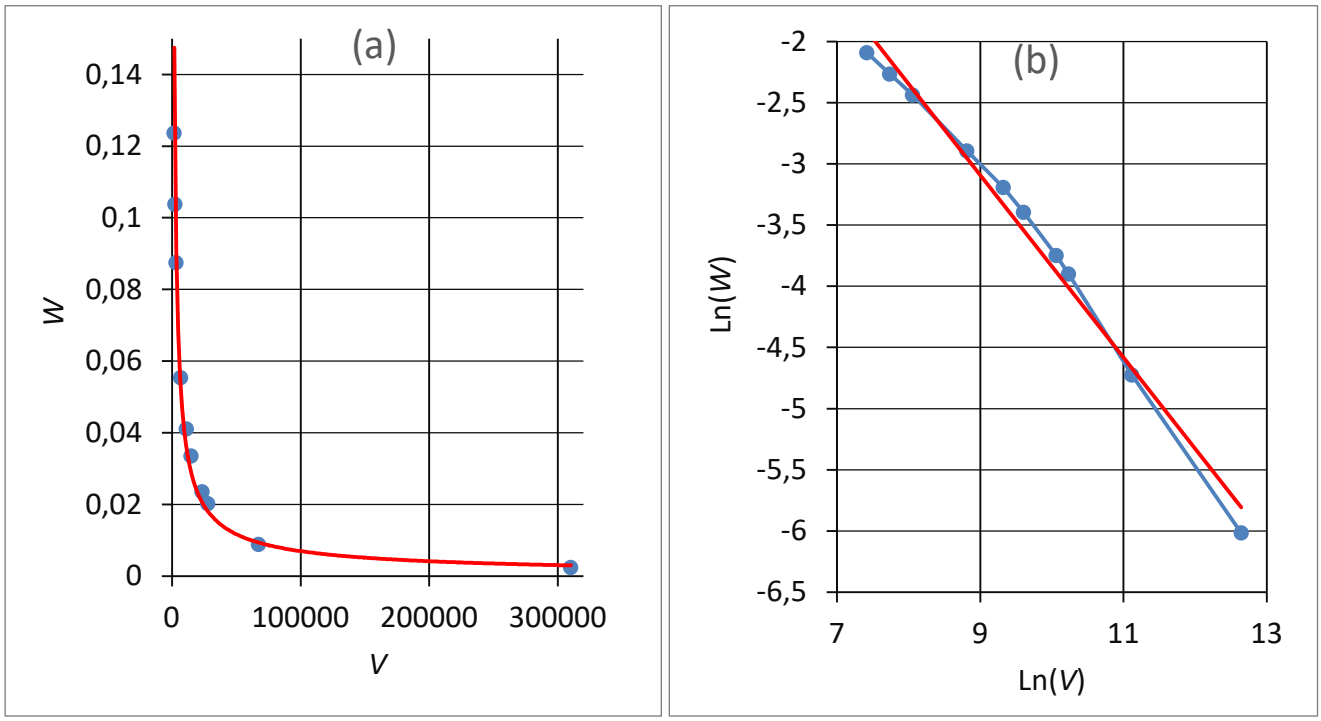

Рис. 35. Диаграмма зависимости вероятности идентификации по атрибуту «имя» в обычном (a) и логарифмическом (b) масштабе координата, где красным обозначена линия тренда, V – количество записей базы, W – вероятность идентификации

В табл. 15 приведены результаты аппроксимации дискретного распределения значений W_2 в зависимости от V для атрибута «имя» базы данных B_1 для 10 значений V экспоненциальной функцией $f_1(x) = 0{,}074 \cdot e^{-0{,}00002x}$, степенной функцией $f_2(x) = 37{,}074 \cdot x^{-0{,}745}$ и логарифмической функцией $f_3(x) = -0{,}024 \cdot \ln(x) + 0{,}307$, где x – значение на оси количества записей базы V, а значение функций f_1, f_2 и f_3 – значение критерия W_2.

Таблица 15

Распределение значений вероятности идентификации в диапазоне 1666 … 310132

x	W_2	f_1	f_2	f_3
310132	0,002440928	0,000149794	0,003005186	0,003526231
67308	0,008869674	0,019257587	0,00937907	0,040191175
27853	0,020249165	0,042393926	0,018098385	0,061367297
23429	0,023560545	0,04631589	0,020587466	0,065518484

x	W_2	f_1	f_2	f_3
14847	0,03354213	0,054988556	0,028920021	0,076466726
11194	0,041093443	0,059156418	0,035692371	0,083244803
6733	0,055398782	0,064676964	0,052125955	0,095445374
3155	0,08748019	0,069474869	0,091688511	0,113638149
2293	0,103794156	0,070683	0,116296977	0,121297209
1666	0,12364946	0,071574946	0,147544007	0,12896366
χ^2	16,919	0,136281	0,007277	0,069443
S_{on}	16,919	0,110295	0,007119	0,073112
$T(Q)$	1,3581	0,14508	0,035076	0,11945

В трех нижних строках табл. 15 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке W_2) [85] и рассчитанные для f_1, f_2 и f_3 по формулам (15 – 17) при заданном уровне значимости $\alpha = 0{,}05$. Из таблицы видно, что и экспоненциальная функция f_1, и степенная функция f_2, и логарифмическая функция f_3 отвечают всем трем критериям согласия. Причем все три функции отвечают всем критериям согласия даже при уровне значимости $\alpha = 0{,}5$.

На рис. 36 в логарифмическом масштабе по обеим координатам приведены графики функций f_1, f_2 и f_3 в сравнении с дискретной последовательностью экспериментальных значений для критерия W_2 атрибута A_2 «имя» базы B_1 (см. рис. 35b).

Из рис. 36 видно, что линия тренда W_2 совпадает с графиком f_2.

В результате проверки по критериям первого рода все три вида функций не были отвергнуты, в связи с этим по формуле (14) для функций f_1, f_2 и f_3 были рассчитаны значения дисперсии ошибок. Сравнение показало, что значение дисперсии $s_2^2 = 10041284942$ для степенной функции f_2 является минимальным, дисперсия логарифмической функции s_3^2 и экспоненциальной функции s_1^2 превышают s_2^2 соответственно в 7,03 и в 2,56 раза. В результате выбрана степенная зависимость W_2 от V.

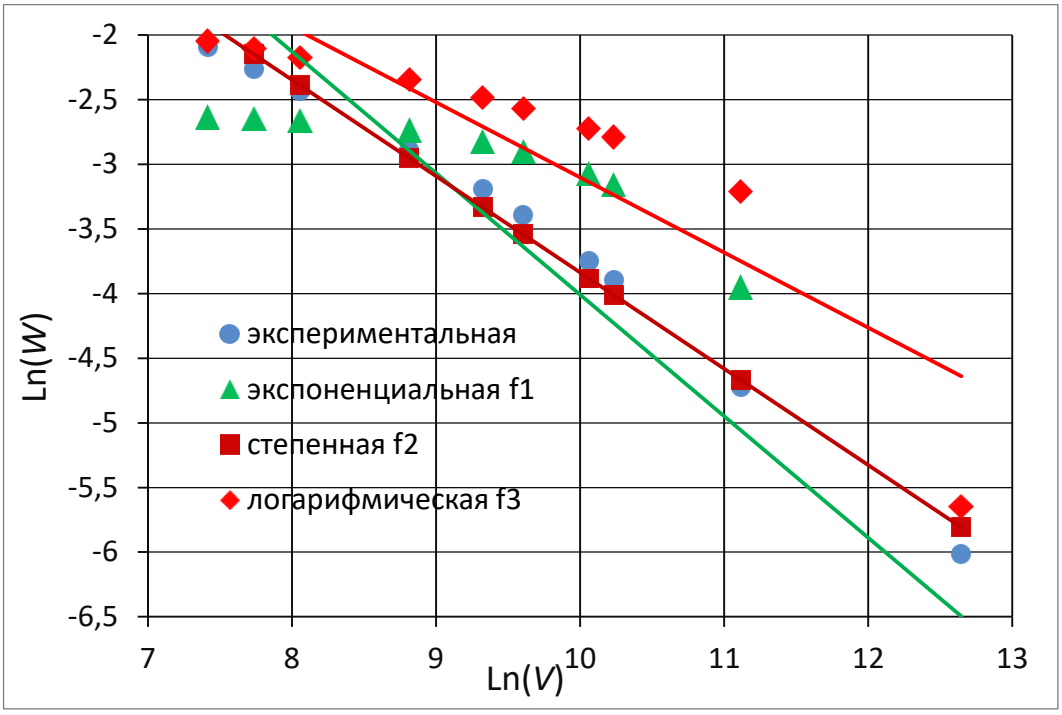

Рис. 36. Функции f_1, f_2 и f_3 в сравнении с дискретной экспериментальной последовательностью для W_2 атрибута A_2 базы B_1, где V – количество записей базы, W – вероятность идентификации

На рис. 37 приведен график степенной функции зависимости W_2 от V в обычных координатах.

Из рис. 37 можно сделать вывод, что для базы объемом менее 6700 записей обезличивание по атрибуту «имя» становится обязательным.

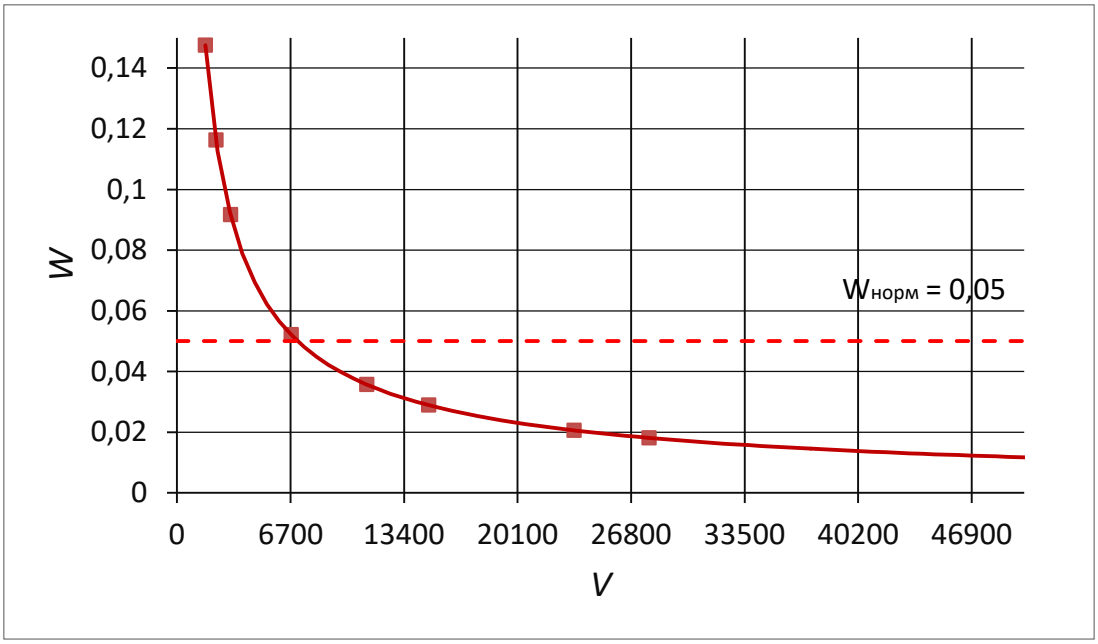

Рис. 37. График степенной функции f_2 в сравнении с $W_{норм}$, где V – количество записей базы, W – вероятность идентификации

2.4.3. Зависимость параметров атрибута «дата рождения» от количества записей базы данных

Для атрибута «дата рождения» диаграмма частот значений n_d в зависимости от q_k на всех рассмотренных объемах V_b базы B_2 сохраняет вид гамма-функции, как это показано в табл. 11 (атрибут A_8 в объеме $V_{\max}$ = 329115 записи базы данных B_2) и табл. 16 (атрибут A_8 в объеме V_5 = 65315 записей базы данных B_2).

Количество различных значений в объеме 65315 записей Q_8 = 22537.

По формуле (7) вычислено значение

$$W_8 = Q_8 / V_5 = 0{,}343523.$$

В табл. 16 приведены результаты аппроксимации дискретного распределения для атрибута «дата рождения» (A_8) в объеме 65315 записей базы данных B_2 для 7 интервалов экспоненциальной функцией $f_1(x) = 22125 \cdot e^{-0{,}43x}$, степенной функцией $f_2(x) = 23118 \cdot x^{-1{,}576}$ и гамма-функцией $f_3(x) = 23350 \cdot x^{1{,}75} \cdot e^{-0{,}944x}$, где x – среднее значение интервала на оси количества записей q_8, а значение функций f_1, f_2 и f_3 – количество дат рождения n_8 в интервале.

Таблица 16

Распределение значений дат рождения в диапазоне 1 … 40

x	B_2	f_1	f_2	f_3
1,5	11586	11608,15874	12201,98355	11520,879
3,5	7510	4912,132626	3209,955572	7682,4014
5,5	2554	2078,628271	1574,485483	2564,8658
7,5	703	879,5966674	965,7255204	668,09645
9,5	163	372,2119574	665,3617909	152,95028
11,5	31	157,5059869	492,3687296	32,345004
13,5	5	66,65056141	382,423208	6,4821952
15,5	2	28,20399035	307,6008301	1,2495985
27,5	0,5	0,161938843	124,6125932	0,0310959
χ^2	15,507	1355,277	13388,54	14,48364996
S_{on}	15,507	1473,787	7293,935	9,148087
$T(Q)$	1,3581	9,57225	16,12324	0,68719

В трех нижних строках табл. 16 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке B_2) [85] и рассчитанные для f_1, f_2 и f_3 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Необходимо отметить, что параметры функций f_1, f_2 и f_3 подобраны так, чтобы хотя бы два из трех критериев согласия были минимальными. Из таблицы видно, что экспоненциальная функция f_1 и степенная функция f_2 не отвечают ни одному из критериев согласия, а гамма-функция f_3 отвечает всем трем.

В результате решения уравнения (10) получены значения q_{8W} и $q_{8\text{норм}}$. На рис. 38 приведен график функции распределения вероятности $F(x)$ и отмечены значения W_8, $W_{\text{норм}}$ и U.

Из рис. 38 можно сделать следующие выводы:

– возможности нарушителя позволяют ему успешно работать с необезличенным атрибутом A_8 ($U > q_{8W} = 2$), поэтому атрибут A_8 («дата рождения») необходимо обезличивать;

– атрибут A_8 необходимо обезличивать в соответствии с нормативным значением, поскольку $W_8 > W_{\text{норм}}$;

– нормативное значение для атрибута A_8 является недостаточным относительно возможностей нарушителя ($U > q_{8\text{норм}} = 5$), возможна опасная ситуация в случае $W_{\text{норм}} > W_8$.

Для определения зависимости W_8 от количества записей БД построена диаграмма распределения W_8 в диапазоне от $V_{min} = 5486$ записей до $V_{max} = 329115$ записей.

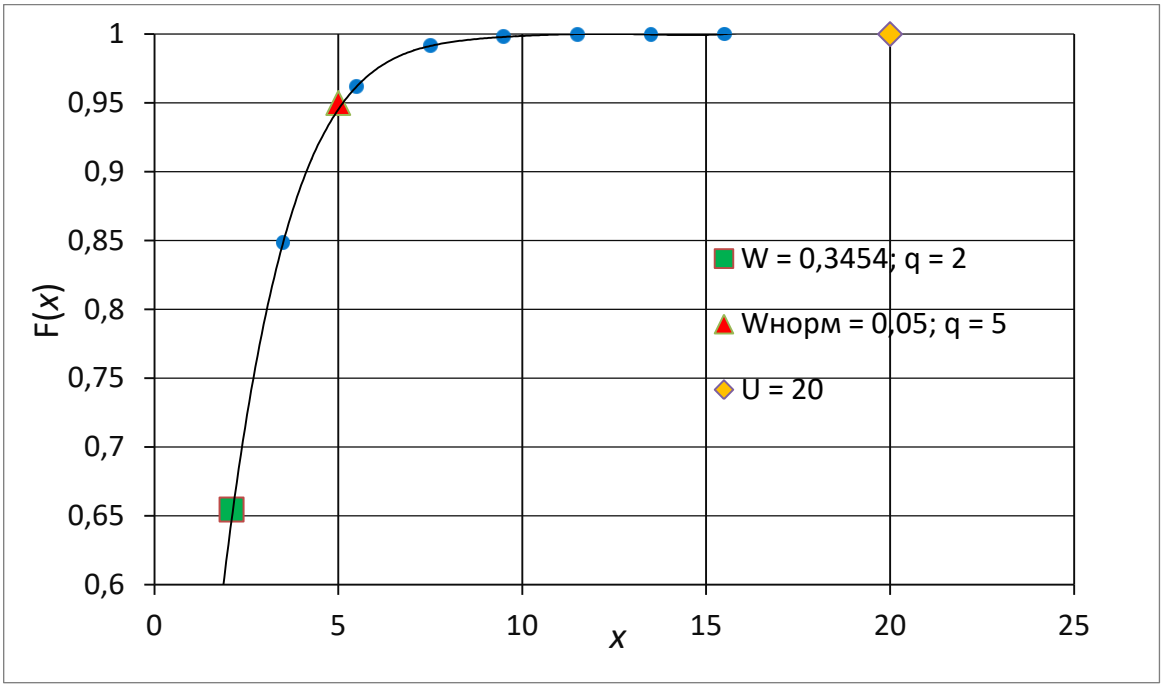

Рис. 38. График распределения вероятности для атрибута «дата рождения» в сравнении с W_8, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

На рис. 39 приводится диаграмма частот значений W_8 в зависимости от V в обычном (рис. 39a) и логарифмическом (рис. 39b) масштабе по обеим осям. На рис. 39b виден линейный характер в правой части диапазона.

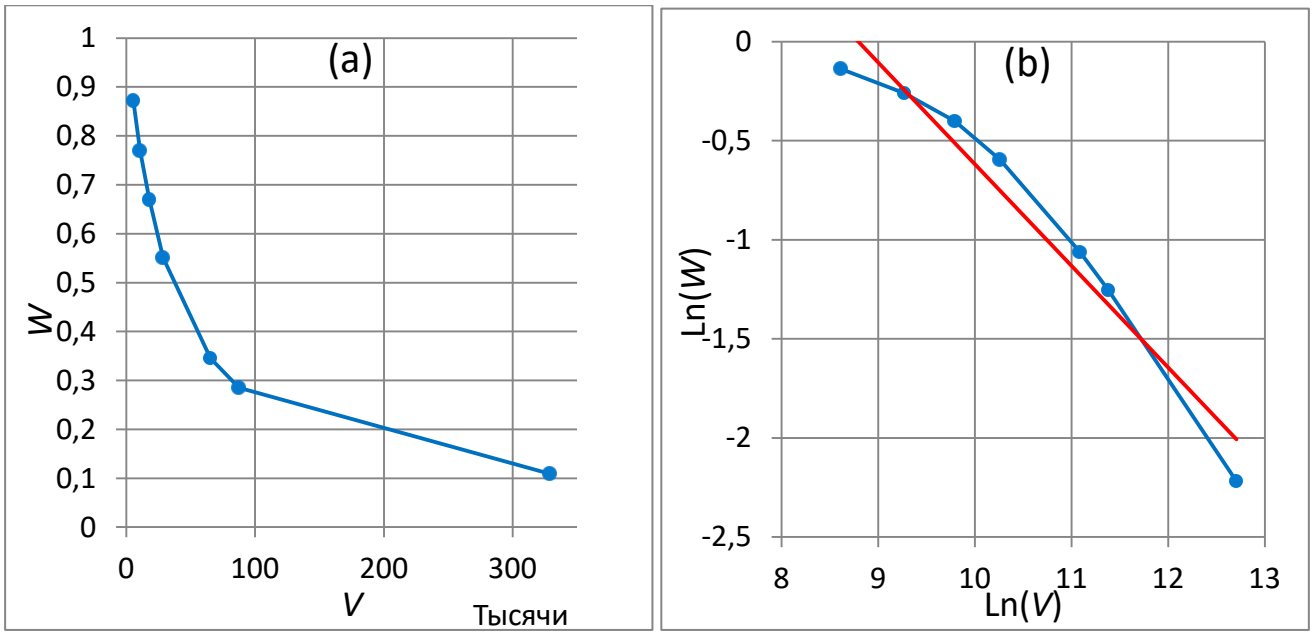

Рис. 39. Диаграмма зависимости вероятности идентификации по атрибуту «дата рождения» в обычном (a) и логарифмическом (b) масштабе координат, где красным обозначена линия тренда, где V – количество записей базы, W – вероятность идентификации

В табл. 17 приведены результаты аппроксимации дискретного распределения значений W_8 в зависимости от V для атрибута «дата рождения» базы данных B_2 для 7 значений V гамма-функцией $f_1(x) = 0{,}815 \cdot x^{0{,}012} \cdot e^{-0{,}000014x}$, степенной функцией $f_2(x) = 99{,}1 \cdot x^{-0{,}535}$ и логарифмической функцией $f_3(x) = -0{,}198 \cdot \ln(x) + 2{,}74$, где x – значение на оси количества записей базы V, а f_1, f_2 и f_3 – значение критерия W_8.

Таблица 17

Распределение значений вероятности идентификации в диапазоне 5486 … 329115

x	W_8	f_1	f_2	f_3
329115	0,10881607	0,009469	0,11073697	0,22457582
87551	0,28530799	0,2742483	0,2248868	0,4867646
65315	0,34544898	0,3730908	0,26305208	0,54477855
28569	0,55105184	0,6179101	0,40942071	0,70850466
17865	0,66996921	0,7137732	0,52632262	0,80146144
10637	0,77032998	0,7848835	0,69458501	0,90412543
5486	0,87130879	0,8368991	0,9898555	1,03522898
χ^2	12,592	0,068289	0,094673	0,080513
S_{on}	12,592	0,31838567	0,096396	0,086306
$T(Q)$	1,3581	0,054506923	0,127107	-0,06493

В трех нижних строках табл. 17 приведены значения критериев согласия χ^2, отношения правдоподобия и Колмогорова: справочные (в колонке W_8) [85] и рассчитанные для f_1, f_2 и f_3 по формулам (11) – (13) при заданном уровне значимости $\alpha = 0{,}05$. Из таблицы видно, что и гамма-функция f_1, и степенная функция f_2, и логарифмическая функция f_3 отвечают всем трем критериям согласия. Причем все три функции отвечают всем критериям согласия даже при уровне значимости $\alpha = 0{,}5$.

На рис. 40 в логарифмическом масштабе по обеим координатам приведены графики функций f_1, f_2 и f_3 в сравнении с дискретной последовательностью экспериментальных значений для критерия W_8 атрибута A_8 «дата рождения» базы B_2 (см. рис. 39b).

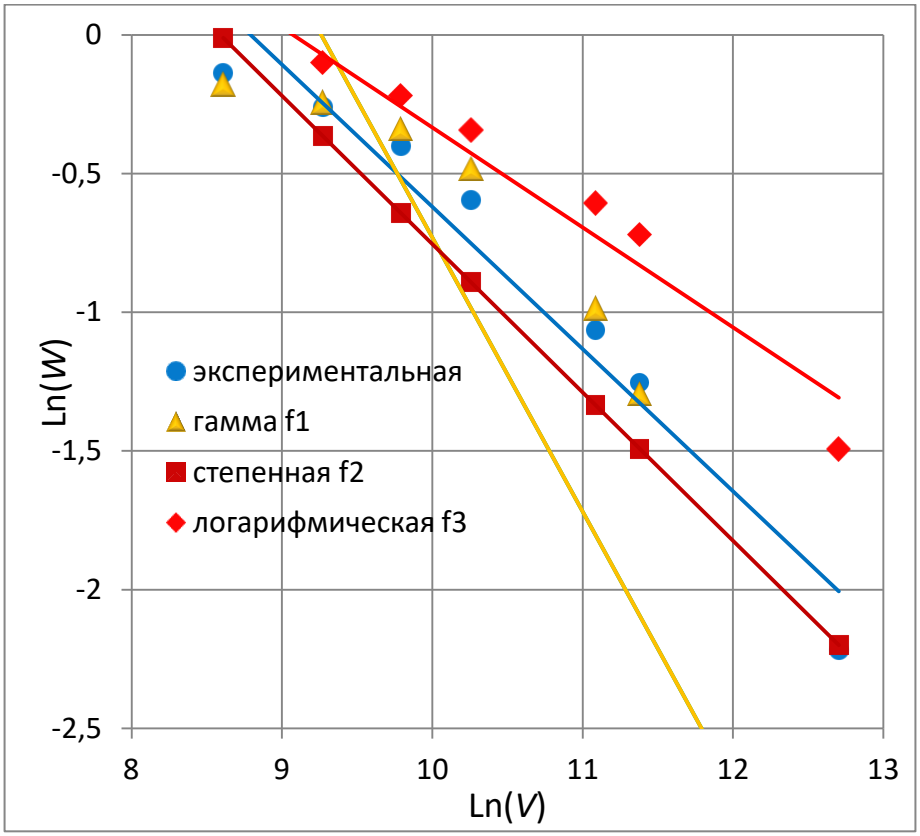

Рис. 40. Функции f_1, f_2 и f_3 в сравнении с дискретной экспериментальной последовательностью для W_8 атрибута A_8 базы B_2, где V – количество записей базы, W – вероятность идентификации

Из рис. 40 видно, что линия тренда W_8 наиболее близка к линии тренда f_2.

В результате проверки по критериям первого рода все три вида функций не были отвергнуты, в связи с этим по формуле (14) для функций f_1, f_2 и f_3 были рассчитаны значения дисперсии ошибок. Сравнение показало, что значение дисперсии s_2^2 = 1983375582 для степенной функции f_2 является минимальным, дисперсия логарифмической функции s_3^2 и гамма-функции s_1^2 превышают s_2^2 соответственно в 53,83 и в 15,53 раза. В результате выбрана степенная зависимость W_8 от V.

На рис. 41 приведен график степенной функции зависимости W_8 от V с целью экстраполяции поведения этой функции до значений V = 1млн записей относительно $W_{норм}$.

Из рис. 41 можно сделать вывод, что для базы объема 1 миллион записей обезличивание по атрибуту «дата рождения» является обязательным.

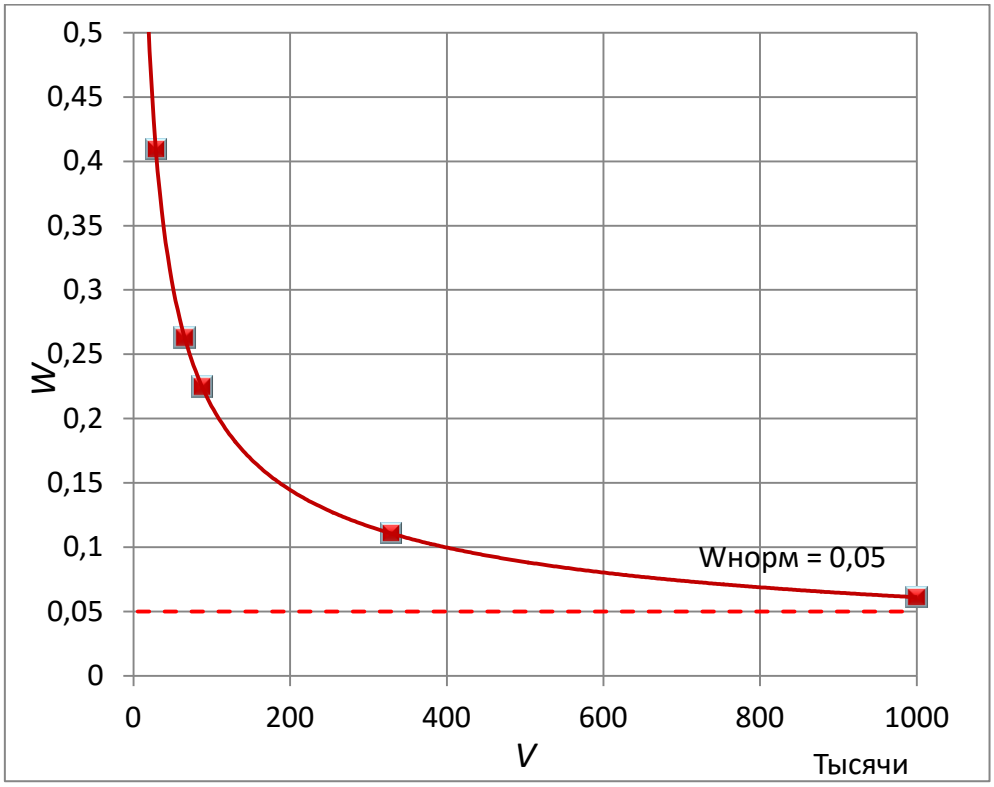

Рис. 41. Экстраполяция степенной функции f_2 до значений V = 1млн записей в сравнении с $W_{норм}$

Аналогичные исследования по прочим атрибутам для различных объемов баз данных не вошли в рамки представленной работы. Можно предположить, что зависимость таких атрибутов, как «отчество», «наименование улицы» и подобные им будет аналогична атрибуту «имя», что может быть уточнено во время дальнейших исследований.

2.5. Выводы по главе 2

В соответствии с поставленной в разделе 1.5 задачей выполнены количественные оценки вероятности идентификации ФЛ по отдельным атрибутам и их совокупностям и определена зависимость значения вероятности идентификации произвольного ФЛ по атрибуту в целом от количества записей базы ПД.

Для вычисления вероятности идентификации по атрибуту в целом по БД независимо от конкретного значения атрибута предложено использовать отношение количества различных значений идентификаторов или их совокупностей к количеству записей БД. С целью обоснования нормативного значения вероятности идентификации в зависимости от вида распределения

значений атрибутов сформулирована и экспериментально доказана гипотеза о степенном характере распределения значений для всех изучаемых атрибутов, кроме атрибута «дата рождения», для которого сформулирована и доказана гипотеза о гамма-распределении.

Полученные результаты позволяют сделать вывод о необходимости обезличивания ПД по конкретным атрибутам или их сочетаниям с учетом количества записей базы ПД. Основываясь на результатах, полученных для отдельной выборки из базы ПД, возможно масштабирование решения об обезличивании для всей БД.

Глава 3. Оценка эффективности обезличивания персональных данных на основе модели нарушителя

Задачей этой главы является разработка методики оценки эффективности искажающих методов обезличивания ПД, которая должна учитывать параметры алгоритмов обезличивания и возможности нарушителя, заложенные в модель нарушителя. В соответствии с разработанной методикой для каждого исследованного алгоритма, с учетом его параметров и на основании количества циклов алгоритма деобезличивания делается вывод об эффективности алгоритма обезличивания. Для части алгоритмов сделаны рекомендации о способах повышения эффективности обезличивания ПД.

3.1. Условия применения алгоритмов искажающих методов обезличивания

В методах обезличивания, основанных на внесении искажений в содержимое базы ПД, секретом является алгоритм искажения. Секретом будем называть дополнительную информацию, необходимую для деобезличивания в соответствии с [1]. Поскольку идентификаторы не удалены из базы, теоретически их можно восстановить простым перебором или с помощью вычислений. Согласно [2], можно выделить два принципа реализации искажающих алгоритмов:

1) в методе изменения состава или семантики производятся искажения внутри одной строки (при этом многие операции с данными не требуют предварительного деобезличивания). Использование различных таблиц смещения для разных записей базы большого объема представляется нецелесообразным с точки зрения производительности, поскольку в этом случае сортировка базы без деобезличивания будет невозможна. Поэтому в этом исследовании рассмотрено применение одной и той же таблицы смещения или формулы смещения для всех записей БД. Кроме того, в представленном исследовании в качестве искажающего воздействия рассматривается только смещение символов, но не их замена в соответствии с таблицей подстановки;

2) в методе перемешивания производятся искажения внутри группы строк (все операции с данными требуют предварительного деобезличивания

группы строк). В соответствии с [20], реализация метода перемешивания с несегментированным подходом является нецелесообразной для баз большого объема по причине необходимости деобезличивания всей базы при внесении любых изменений. В представленном исследовании рассматривается только сегментированный подход, но без применения криптографических методов, которые в значительной мере увеличивают сложность любых методов обезличивания.

При применении любых искажающих методов перед проведением процедуры обезличивания все атрибуты, независимо от их синтаксиса, необходимо перевести в текстовый (символьный) вид с целью унификации представления данных.

3.2. Модель нарушителя для искажающих методов обезличивания

Общая для различных методов модель нарушителя, описанная в разделе 2.1, предполагает уточнение параметров G (количество известных нарушителю записей о ФЛ в базе), U (максимальное количество записей, полученных в результате деобезличивания, при котором поиск считается эффективным) и алгоритма деобезличивания, зависящего от метода обезличивания. При этом должна быть рассчитана количественная оценка возможности достижения заданного результата U при заданных параметрах G и алгоритма деобезличивания.

Для обеспечения достоверности модели исследована экспериментальная база ПД B_1 объемом 310 тыс. ФЛ, где в качестве идентифицирующих атрибутов использовались фамилия A_1 (15 символов), имя A_2 (10 символов), отчество A_3 (15 символов), наименование улицы проживания A_4 (25 символов), номер дома A_5 (5 символов) и номер квартиры A_6 (3 символа), в сумме составляющие строку длиной 73 символа. Прочие атрибуты, не являющиеся идентификаторами, суммарно были обозначены как A_z. В качестве алгоритмов искажения рассматривались следующие алгоритмы:

– перестановка символов внутри полной строки идентификаторов;

– перестановка битов внутри строки идентификатора;

– перемешивание полей внутри группы из 256 записей базы с сохранением структуры строки;

– перемешивание символов внутри группы из 256 записей базы с сохранением места расположения в строке.

В качестве допущений были заданы следующие:

– формула/таблица смещения принималась произвольной, но одинаковой для всех строк (для метода изменения состава) или сегментов записей (для метода перемешивания);

– искажению подвергнуты атрибуты $A_1 – A_6$;

– максимальное количество известных нарушителю записей в базе $G = 5$. Это значение параметра является предварительным и может быть уточнено в результате применения алгоритма деобезличивания;

– максимальное количество записей, полученных в результате деобезличивания, при котором поиск считается эффективным – $U = 20$. Это означает, что нарушитель не может за приемлемое время однозначно выбрать определяемое ФЛ из более чем 20 вариантов, которые он получил в результате поиска по имеющимся элементам идентификаторов.

Алгоритм действий нарушителя:

1) нарушитель находит в базе вручную или автоматизированным путем множество записей B_1, содержащих прочие (неискаженные) данные A_{z1} первого из известных ему лиц, при этом одинаковые прочие данные могут принадлежать нескольким лицам

$$B_1 = \{L_1, L_2, \ldots L_k\},$$

где k – количество записей в группе ($1 \leq k \leq 20$);

2) нарушитель выбирает из множества B_1 те записи, которые содержат искаженные атрибуты $\{A_1, A_2, \ldots A_6\}$ первого известного ему лица. Выбор производится по составу символов искаженной строки или составу полей в сегменте из 511 записей (по 255 записи в обе стороны от найденной). В идеальном случае должна быть выбрана одна запись, но если записей будет несколько (но

менее 20), нужно использовать вторую известную запись и т.д. Если количество выбранных записей более 20, то модель нарушителя не применима;

3) нарушитель по символам/полям известного ему ФЛ составляет таблицу смещений (*T)* символов в строке или полей в сегменте

$$T = \{t_1, t_2, \ldots, t_N\},$$

где t_N – смещение элемента относительно начала строки/группы записей, N – количество перемещаемых элементов.

В таблицу заносятся смещения по первому найденному совпадению значения символа/поля. Эти смещения будут абсолютно точными для неповторяющихся элементов и будут сомнительными для повторяющихся символов/полей;

4) нарушитель использует для устранения неточностей таблицы T данные второго известного ему лица по описанному выше алгоритму. Если неточности в таблице останутся, необходимо использовать третью запись и т.д.;

5) если нарушитель получил точную и полную таблицу смещений, позволяющую получить однозначные данные о любом лице в БД, при использовании части или всех известных ему записей ФЛ, модель нарушителя считается полностью эффективной;

6) если таблица смещений, полученная нарушителем с использованием всех известных ему записей ФЛ, позволяет при поиске данных о любом лице в базе получить набор записей, но не более чем из 20 лиц, то модель нарушителя считается эффективной условно;

7) если таблица смещений, полученная нарушителем при использовании всех известных ему записей, позволяет получить данные о любом ФЛ только в виде набора записей более чем из 20 ФЛ, то модель нарушителя считается неэффективной.

Далее описаны особенности алгоритмов деобезличивания в зависимости от алгоритма искажения.

3.2.1. Алгоритм деобезличивания при перемещении символов внутри строки

В процессе обезличивания исходная таблица смещений символов строки идентификаторов $S_0 = \{1, 2, \ldots, N\}$ преобразована в таблицу $S = \{s_1, s_2, \ldots, s_N\}$. Для искажаемых идентификаторов $A_1 - A_6$ суммарное количество символов в строке $N = 73$.

Сначала нарушитель должен найти в обезличенной базе запись с прочими (необезличенными) данными известного лица A_z, затем ему необходимо убедиться, что эта запись принадлежит известному лицу. Возможность совпадения прочих данных у разных лиц существует и зависит от контекста и объема прочих данных. Но поскольку перемещение символов идентификаторов производится в одной строке, их суммарный состав сохраняется. Вероятность полного совпадения суммарных составов символов у разных лиц определяется выражением

$$p_{sum} = p_1 \cdot p_2 \cdot \ldots \cdot p_N,$$

где p_N – вероятность наличия в базе (в полях искажаемых атрибутов) символа, стоящего на месте n в строке идентификаторов. Данные вероятности определяются в результате анализа частотного распределения символов в соответствующих полях всех записей базы данных.

После нахождения единственной записи первого известного лица по его прочим данным производится процедура посимвольного сравнения искаженной строки идентификаторов $A_1 = \{A_{11}, A_{12}, \ldots, A_{1N}\}$ с их реальной строкой. При определении таблицы смещений символов $T = \{t_1, t_2, \ldots, t_N\}$ в перемешанной строке длиной N путем посимвольного сравнения со значением известной строки проблему представляют совпадающие (повторяющиеся) символы. Например, если символы A_{1i} и A_{1j} совпадают, то найденные смещения t_i и t_j могут быть ошибочными. С целью определения общего количества m неоднозначных мест в строке проводится статистический анализ частотного распределения, т.е. определяется, какие символы встречаются в общей строке идентификаторов не менее двух раз.

Поскольку длины идентификаторов в структуре базы задаются строго, а значения каждого атрибута дополняются пробелами, то в строке идентификаторов больше всего будет повторяющихся пробелов. Условно их можно считать незначимыми и игнорировать в расчетах числа совпадений, т.к. пробелы не различимы. Для определения наиболее вероятного количества пробелов в полях атрибутов проводится частотный статистический анализ.

Оценивая возможность деобезличивания при использовании нарушителем второй записи, необходимо учесть, что:

− вторая строка независима от первой, поэтому в ней также ожидается m неоднозначных мест, но они необязательно совпадут с неоднозначными местами, возникшими после сравнения первой записи;

− минимальное количество совпадений неоднозначных мест двух строк равно 0. Это возможно при условии, что удвоенное количество неоднозначных мест меньше длины строки:

$$2m < N; \tag{15}$$

при этом полное несовпадение сразу заполняет таблицу смещений, и задача считается решенной;

− если одному неоднозначному месту первой записи соответствует точное место второй записи, то количество неоднозначностей уменьшается на одно место;

− повторяющиеся символы могут отличаться друг от друга, даже если место одного повторяющегося символа в первой строке совпадет с местом другого повторяющегося символа во второй строке, поэтому все повторяющиеся символы условно считаются одним неоднозначным символом.

Данный алгоритм проиллюстрирован на рис. 42, где звездочками указаны точно установленные места, а стрелками указаны места символов, для которых возможность ошибочного определения смещения сохраняется. Для упрощения отмечены положения только двух часто повторяющихся букв, причем все буквы, показанные в каждой записи по отдельности, имеют неоднозначное

местоположение, но при совмещении двух записей исключаются неоднозначности местоположения некоторых из них (они стрелками не помечены).

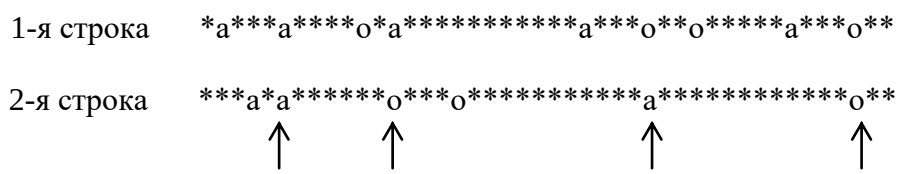

Рис. 42. Варианты неоднозначного определения смещения символов

Из рис. 42 видно, что при использовании первой строки из 44 символов не удалось однозначно установить 9 положений для двух букв. Если бы символы в строке не были связаны в слова, количество неоднозначных вариантов их размещения определялось бы значением 9! = 362880.

При применении второй строки независимо от первой, количество неоднозначных вариантов размещения двух букв определялось бы значением 6! = 720, но если строки применить последовательно, то количество неоднозначных вариантов снижается до 4! = 24. Поскольку неоднозначные символы не совпадают друг с другом, некоторые из них перестанут быть повторяющимися, и реальное количество неоднозначностей уменьшится в большей степени. Например, это произошло бы, если в результате применения двух строк из 4-х мест совпадения было бы три одинаковых буквы и одна одиночная. В этом случае количество неоднозначных вариантов снижается до 3! = 6.

Поскольку символы в строке связаны в слова, принято допущение, что взаимное расположение различных букв любого языка не является равновероятным. Выдвинута гипотеза, что взаимное влияние символов уменьшает количество неоднозначных вариантов их расположения в строке. Для оценки влияния этого фактора проведен анализ взаимного расположения пар символов в идентификаторах.

Алгоритм поиска смещений символов в строке искаженной базы для количества известных нарушителю записей $G = 3$ приведен на рис. 43. Пример процедуры поиска смещений символов приведен в приложении Г.

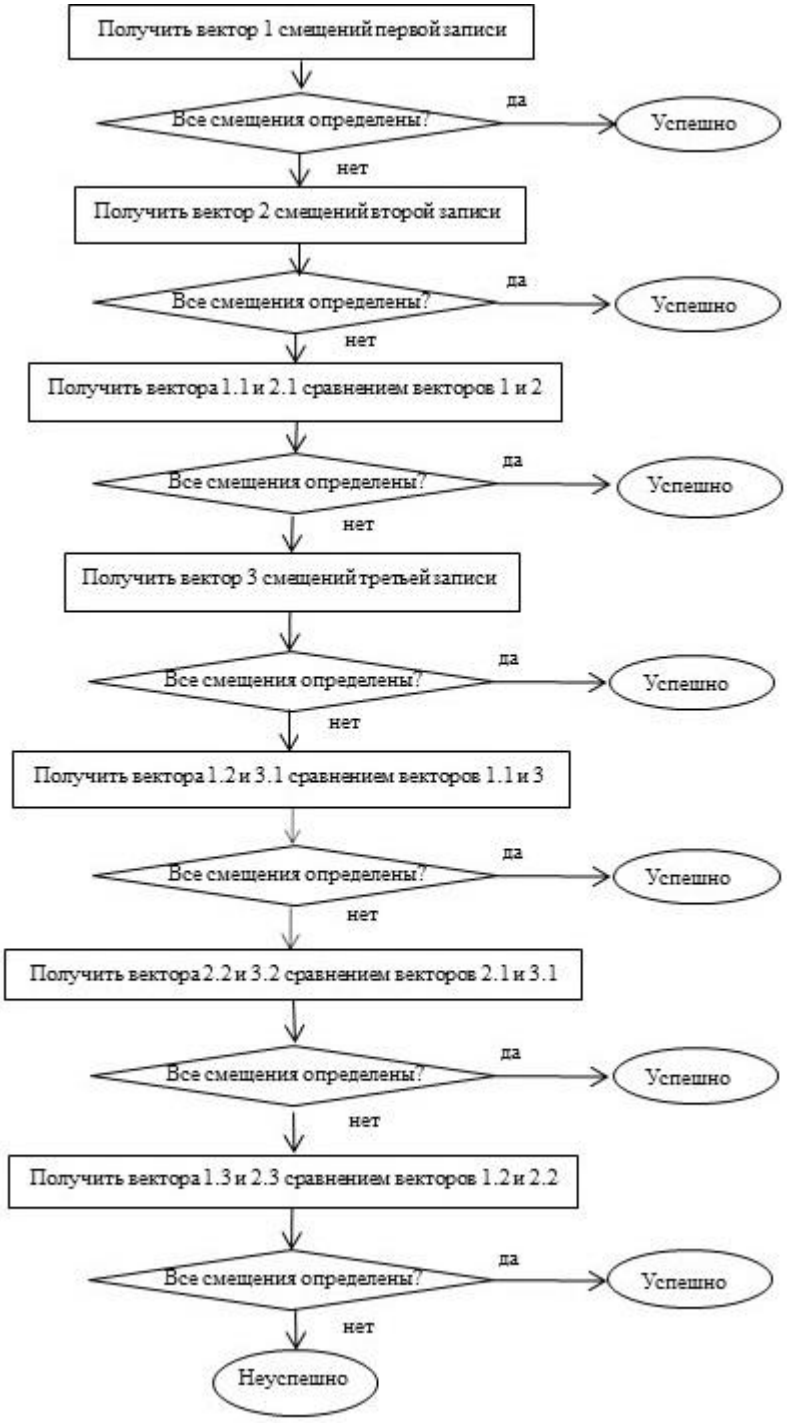

Рис. 43. Алгоритм поиска смещений символов в строке для $G = 3$

Для расчета наиболее вероятного количества совпадений применена формула определения математического ожидания:

$$C_0 = \mathrm{M}[F_v(c)], \tag{16}$$

где M – математическое ожидание, $F_v(c)$ – функция зависимости количества вариантов совпадений от количества совпадений c. Сама функция $F_v(c)$ является дискретной и представляет собой произведение количества различных вариантов расположения c неоднозначных символов в множестве символов строки идентификаторов (N) и количества различных вариантов расположения остальных ($m - c$) неоднозначных символов в оставшейся части множества символов строки идентификаторов ($N - c$), т.е.

$$F_v(c) = R_{Nc} \cdot R_{(N-c)(m-c)} = \frac{N!}{c!(m-c)!(N-m)!}, \quad (17)$$

где m – количество неоднозначных мест в одной строке, $R_{ab} = a! / b!(a - b)!$ – количество различных вариантов расположения b символов в строке длиной a [89 – 91]. Функция $F_v(c)$ является симметричной, ее максимальное значение соответствует центру диапазона значений $C_0 = m/2$, но при большом диапазоне m наибольшая часть вариантов находится в интервале значений $[C_0 - \sigma, C_0 + \sigma]$, где σ – среднеквадратичное отклонение, рассчитанное в соответствии с [88].

3.2.2. Алгоритм деобезличивания при перемещении битов внутри идентификатора

Поскольку реализации алгоритмов обезличивания путем перемещения битов и перемещения символов в строке идентификаторов аналогичны, то аналогичны и соответствующие им алгоритмы деобезличивания.

Применяя первую известную запись, нарушитель по прочим данным известного ему ФЛ получит группу записей, каждая из которых предположительно содержит искаженную строку идентификаторов этого ФЛ. Но уже на этапе выбора из полученной группы записей одной записи с идентификаторами, принадлежащими известному ФЛ, осуществить этот выбор по суммарному набору символов не получится, так как искаженная строка является битовой. Поэтому нарушителю придется осуществлять выбор по суммарному набору битов.

При разработке последующих этапов алгоритма деобезличивания предварительно принята гипотеза о невозможности деобезличивания для случая перемещения битов в объединенной строке идентификаторов. Для проверки гипотезы были приняты следующие допущения:

– объем вычислений ограничивается размерами одного идентификатора. При этом если расчеты вероятности ошибок в объеме одного идентификатора покажут эффективность алгоритма деобезличивания, то область расчетов будет расширена. Если же расчеты вероятности ошибок в объеме одного идентификатора покажут неэффективность алгоритма, то использование в расчетах всей строки идентификаторов заведомо будет неэффективным;

– в структуре идентификатора не учитываются пробелы и коды страниц различных языков. При этом если расчеты без учета этих структурных элементов покажут эффективность алгоритма деобезличивания, то область расчетов будет расширена за счет включения структурных элементов. Если же расчеты без учета этих структурных элементов покажут неэффективность алгоритма, то их включение заведомо будет неэффективным.

Таким образом, количество перемещаемых элементов в строке одного атрибута вычисляется как

$$N = 8 \cdot n_s,$$

где n_s – количество значимых (без пробелов) символов в атрибуте.

В рамках указанных условий применен тот же подход, что при перемешивании символов, но дополнительно учтено следующее:

– все биты 1 и 0 являются значимыми, т.е. теоретически неоднозначными могут быть все N мест в строке (биты значимых символов идентификатора), и предыдущий подход в стандартном виде применить нельзя (при $m = N$ по формуле (15) получаем $2N > N$);

– тот же подход, что при перемешивании символов, используется сначала для учета совпадений битов 1 друг с другом, затем для учета совпадений битов 0

друг с другом, затем – совпадений 1 и 0, в итоге количества вариантов соответствующих сочетаний перемножаются, и рассчитывается полная функция $F_v(c)$ и ее максимум;

– поскольку существуют только биты 1 и 0, то дополнительный анализ встречаемости символов не проводится и никакие дополнительные условия уменьшения количества совпадений не действуют.

3.2.3. Алгоритм деобезличивания при перемешивании полей внутри группы записей

Поскольку сегментированный подход подразумевает перемешивание идентификаторов в группе из 256 записей, то, применяя первую известную запись, нарушителю сначала надо найти прочую информацию A_z в базе и далее – убедиться, что эта информация принадлежит именно этому ФЛ. Алгоритм поиска приведен на рис. 44.

Нарушителю, приняв в качестве точки отсчета порядковый номер найденной записи (его наличие необходимо во всех алгоритмах перемешивания), надо отступить в обе стороны от него по 255 записей и в этой области из 511 записей найти один из идентификаторов известного лица. Искать нужно наиболее редко встречающийся идентификатор, обозначим его A_i. Если первый выбранный идентификатор не будет найден в этой области, необходимо искать в базе следующее вхождение прочей информации. Теоретически в полученной области может быть найдено одно и даже больше значений искомого идентификатора. Если прочие данные принадлежат ФЛ из первой известной записи, то одно из найденных значений идентификатора точно принадлежит определяемому ФЛ.

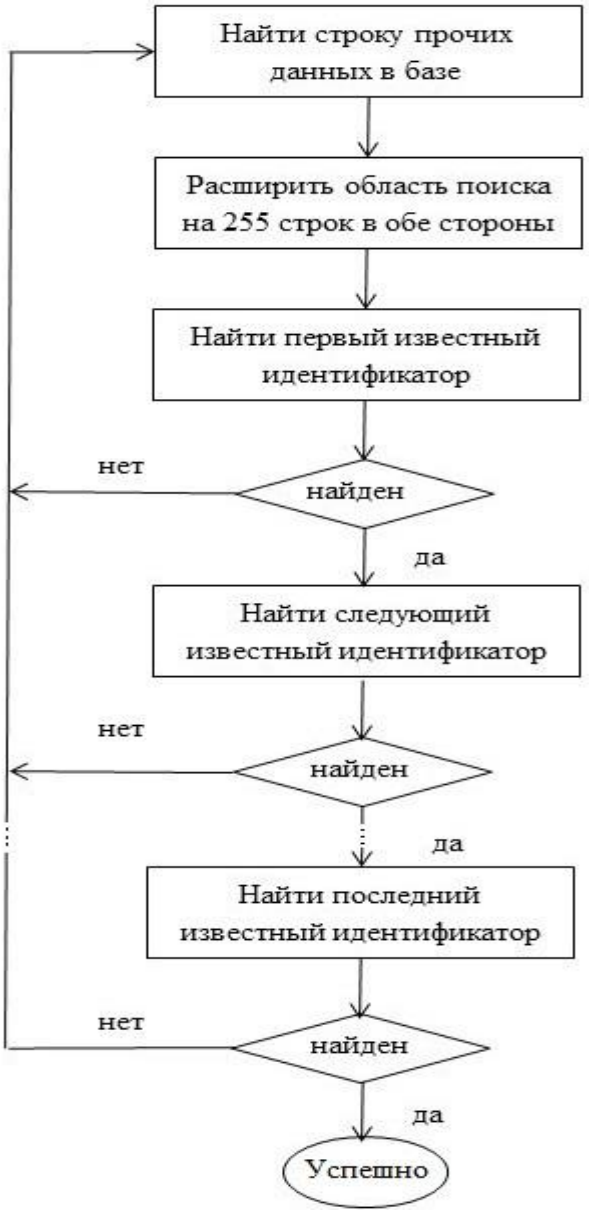

Рис. 44. Алгоритм поиска идентификаторов в группе записей

Частоту попадания нужного значения идентификатора A_{jk} в полученную область можно определить по формуле:

$$p_{jk1} = q_{jk} \cdot 511 / V, \quad (18)$$

где q_{jk} – количество записей с значением идентификатора A_{jk} в базе, V – количество записей базы. Частоту второго вхождения значения идентификатора (при условии вхождения первого) в эту же область можно оценить по формуле:

$$p_{jk2} = (q_{jk} - 1) \cdot 510 / V, \quad (19)$$

Строго говоря, значения, рассчитанные по (18) и (19), могут быть больше 1, т.е. они показывают вероятное количество вхождений одинаковых значений

идентификаторов в полученную область записей. Необходимо рассчитать количество вхождений для всех идентификаторов, выбрать по наименьшему значению наиболее редкий идентификатор в качестве первого для начала поиска, а также определить наиболее вероятное количество повторных вхождений для всех идентификаторов в целом, чтобы оценить необходимость использования последующих известных нарушителю записей.

Если принять найденные прочие данные A_z (а значит и полученную область из 511 записей) за истинно соответствующие известному ФЛ, то, аналогично атрибуту A_j, необходимо в полученной области найти смещения остальных идентификаторов. Успех поиска зависит от возможности повтора известных значений идентификаторов в этой области. Если хотя бы один из идентификаторов будет не найден в полученной области записей, значит, прочие данные не принадлежат известному ФЛ из первой записи, и необходимо искать следующее вхождение прочих данных в базе.

С другой стороны, если все идентификаторы будут найдены в области 511 записей, необходимо с учетом наличия повторных вхождений выбрать те смещения, которые поместятся в границы 256 записей, внутри которых производится реальное перемешивание. При сужении объема группы записей количество повторных вхождений в среднем уменьшится вдвое, но не исчезнет совсем. Это приводит к необходимости использовать вторую известную нарушителю запись. Со второй записью производятся те же операции, что и с первой, в результате нужно сравнить смещения в двух суженных группах по 256 записей. Вероятность того, что совпадут по два смещения для одного идентификатора в двух группах, равна 1/256.

3.2.4. Алгоритм деобезличивания при перемешивании символов внутри группы записей

Действия нарушителя по данному алгоритму аналогичны его действиям при перемешивании полей в группе записей. Используя прочие данные первой известной записи, нарушитель найдет группу записей с одинаковыми прочими

данными. Далее нарушителю нужно по наиболее редкому идентификатору выбрать одну из записей группы и убедиться, что все идентификаторы этой записи принадлежат ФЛ из первой известной записи. Но алгоритм поиска (рис. 44) осложняется тем, что после выбора наиболее редкого идентификатора в заданной области требуется найти не сам идентификатор, а символы, из которых он состоит. По условиям алгоритма символы не изменяют своего положения в строке, поэтому область поиска первого идентификатора можно представить в виде таблицы символов с количеством полей, равным длине идентификатора с пробелами, и количеством строк, равным 511. Если полученную таблицу транспонировать, получим строки длиной 511 символов, в каждой из которых нужно найти один символ идентификатора: в первой строке − первый символ, во второй − второй символ и т.д.

Из-за стандартной длины поля и различного количества значимых символов последними символами в идентификаторе в большинстве случаев будут пробелы, что значительно сокращает поиск смещений для коротких значений, но впоследствии возникнет неопределенность для длинных значений. Для оценки этого фактора произведен анализ количества значимых символов в идентификаторах. Дальнейшая процедура поиска смещений аналогична процедуре, описанной для алгоритма перемешивания символов внутри строки, но с учетом того, что в одной строке ищется только один символ.

В строке из 511 символов будут множественные повторные вхождения значимых символов (без учета пробелов), что значительно повышает вероятность ошибки позиционирования. Данный фактор сохраняет вероятность ошибки позиционирования символа даже после нахождения смещений для всех символов идентификатора. Для оценки вероятности ошибки в определении смещения произведен частотный анализ символов в каждом идентификаторе.

Если в результате определения смещений для всех символов первого идентификатора нарушителем будет принято решение о том, что прочие данные принадлежат лицу из первой записи, нужно проводить процедуру определения

смещений для всех остальных идентификаторов − для каждого идентификатора в отдельности.

Таким образом, всю группу идентификаторов (длиной 73 символа) можно представить в виде транспонированной таблицы из 73 записей по 511 символов, и в целом задача определения таблицы смещений 73 символов может быть представлена в виде последовательности из 73 задач поиска смещения символов в строке (по одному символу в каждой из 73 строк). При этом вторая и последующие известные нарушителю записи будут применяться к другим группам записей.

3.2.5. Результаты применения модели нарушителя при перемещении символов внутри строки

При обнаружении первого вхождения прочих данных известного лица нарушитель сравнивает суммарный состав символов строки идентификаторов, включая пробелы. Длина идентификатора (количество символов за исключением пробелов) является случайной величиной. Для каждого идентификатора экспериментальной базы путем разделения диапазона на n = 10 равных интервалов построена диаграмма частоты количества значимых символов и произведен расчет математического ожидания M и дисперсии σ для дискретных последовательностей в соответствии с [88] по формулам

$$M = \sum_i c_i \cdot p(c_i) \,/\, \sum_i p(c_i),$$

$$\sigma^2 = \sum_i (c_i - M)^2 \,/\, n,$$

где $i = 1, \ldots, n$ – количество интервалов распределения.

На рис. 45 – 48 приведены диаграммы частоты количества значимых символов $p(c)$ для текстовых атрибутов.

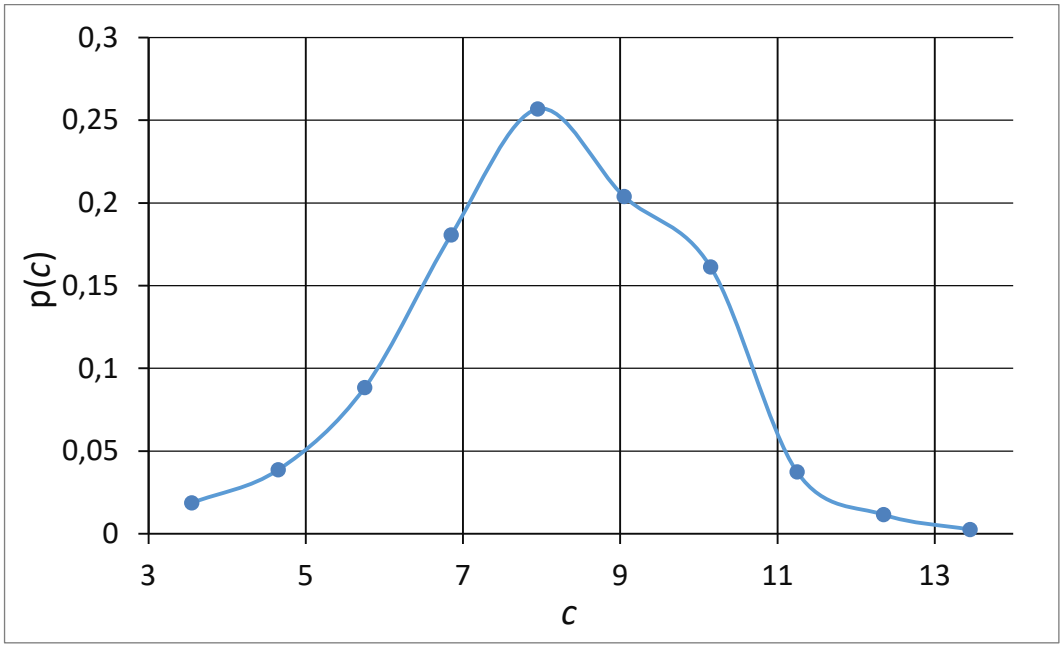

Рис. 45. Диаграмма частоты количества значимых символов *c* для идентификатора «фамилия». $M = 8{,}11$; $\sigma^2 = 3{,}14$

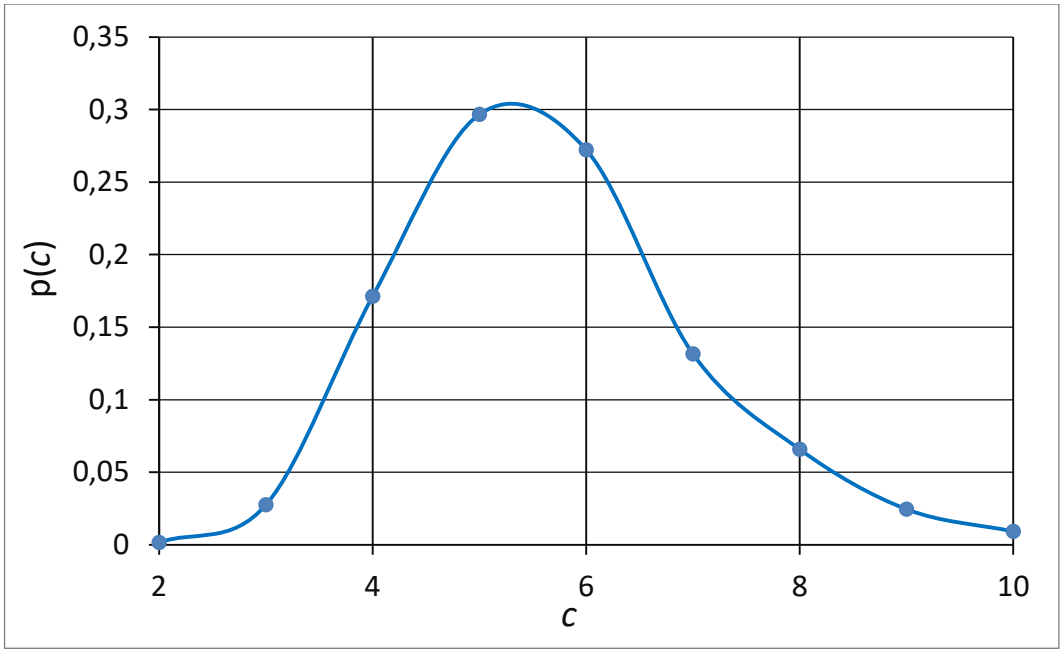

Рис. 46. Диаграмма частоты количества значимых символов *c* для идентификатора «имя». $M = 5{,}64$; $\sigma^2 = 0{,}62$

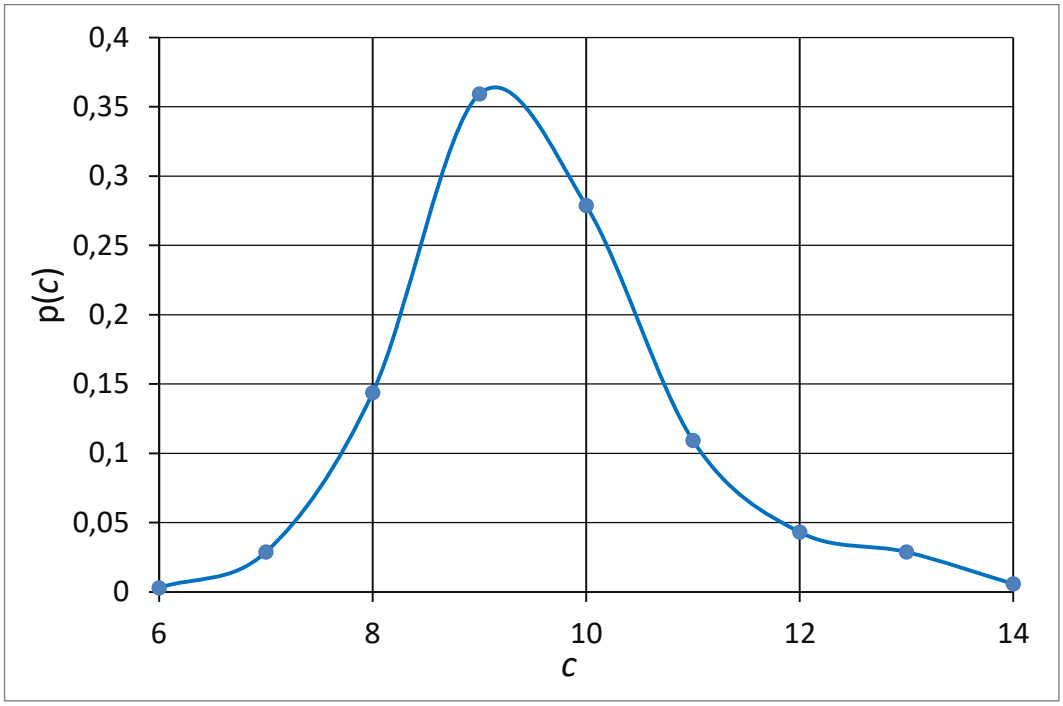

Рис. 47. Диаграмма частоты количества значимых символов *c* для идентификатора «отчество».

$M = 9{,}56$; $\sigma^2 = 1{,}92$

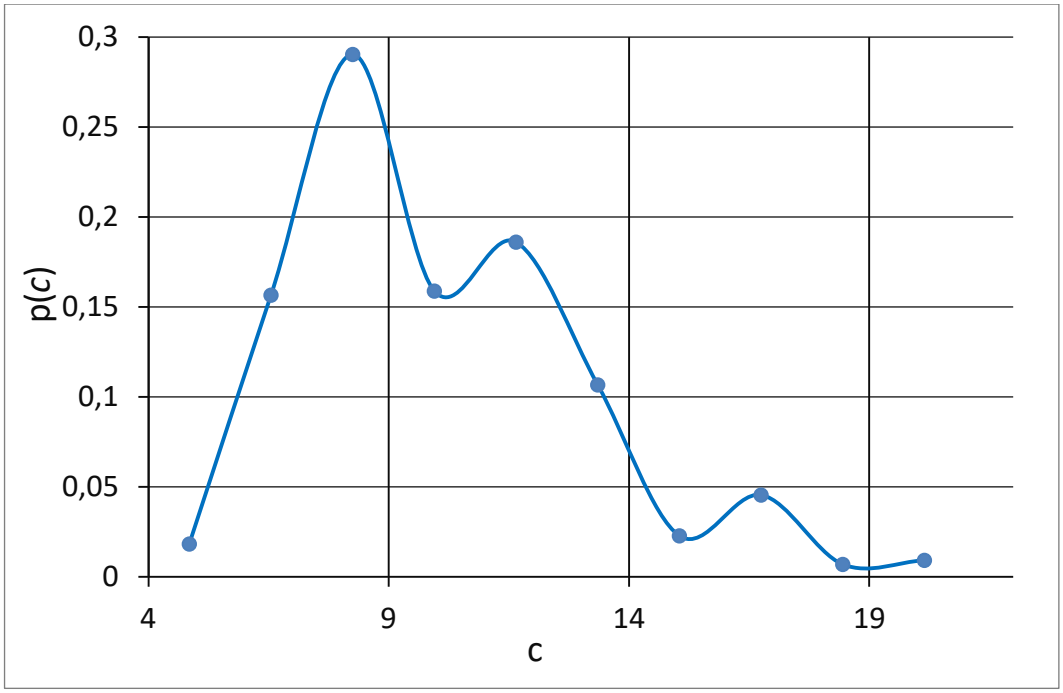

Рис. 48. Диаграмма частоты количества значимых символов *c* для идентификатора «наименование улицы». $M = 10{,}08$, $\sigma^2 = 8{,}99$

Математические ожидания количества значимых символов для различных идентификаторов приведены в табл. 18.

Таблица 18

Количество значимых символов и пробелов для различных идентификаторов

Идентификатор	**Длина в структуре**	**Значимые символы**	**Пробелы**
Фамилия	15	8	7
Имя	10	6	4
Отчество	15	10	5
Улица	25	10	15
Номер дома	5	3	2
Номер квартиры	3	2	1
Общая строка	73	39	34

Для определения наиболее вероятного количества повторений различных символов во всех идентификаторах экспериментальной базы B_1 был произведен анализ частотного распределения символов, что позволило оценить вероятность наличия в строке конкретного символа и суммарное количество повторений символов в общей строке идентификаторов. Результаты расчетов для всех идентификаторов приведены в табл. 19.

Таблица 19

Расчет наиболее вероятного количества повторений символов в различных идентификаторах

Поле/ буква	**Фамилия**	**Имя**	**Отчество**	**Улица**	**Дом**	**Квартира**	**Сумма**	**Доля в базе**	**Кол-во в строке**
Всего по базе	2352995	1968415	3065020	3099888	773930	651797	11912045		39
а	263158	374292	435597	325842	61945		1460834	0,122	4,78
б	46780	8297	8198	69123	15745		148143	0,012	0,48
в	231820	77210	400549	154023	5088		868690	0,073	2,84
г	38945	46631	63000	74064	1141		223781	0,018	0,73
д	44663	60628	82763	68000	426		256480	0,021	0,84
е	165936	128209	269895	246497	104		810641	0,068	2,65
ж	9964	6155	197	14166			30482	0,002	0,10
з	27387	9733	3175	16699			56994	0,005	0,18

Поле/ буква	Фамилия	Имя	Отчество	Улица	Дом	Квартира	Сумма	Доля в базе	Кол-во в строке
и	168959	212871	324120	167020			872970	0,073	2,85
й	13606	41052	657	56512			111827	0,009	0,36
к	151781	49270	80083	212099			493233	0,041	1,619
л	97749	179709	154952	144814			577224	0,048	1,88
м	62851	49507	70297	97914			280569	0,023	0,92
н	177712	185104	332411	162208			857435	0,072	2,80
о	271550	57221	286837	332390			947998	0,079	3,10
п	43406	5007	23879	78537			150829	0,012	0,49
р	122622	117608	147263	212324			599817	0,050	1,96
с	78460	76327	93074	179208			427069	0,035	1,39
т	62401	101595	59618	88273			311887	0,026	1,02
у	65610	10293	6959	44981			127843	0,010	0,41
ф	13971	5156	11533	118			30778	0,002	0,10
х	29140	3491	10940	28000			71571	0,006	0,23
ц	17773	98	0	25117			42988	0,003	0,14
ч	27341	1494	115438	45416			189689	0,016	0,62
ш	35831	958	11786	28196			76771	0,006	0,25
щ	4865	0	0	88			4953	0,000	0,01
ъ	25	0	0	0			25	0,000	0
ы	17486	112	0	42350			59948	0,005	0,19
ь	21770	56848	56969	38398			173985	0,014	0,56
э	462	1966	891	7384			10703	0,001	0,03
ю	9125	19741	6400	12059			47325	0,004	0,15
я	29846	81832	7539	81021			200238	0,016	0,65
дефис				277	84881		85158	0,007	0,27
1				3159	125250	125170	253579	0,021	0,83
2				5761	82013	86921	174695	0,014	0,57
3				2128	80850	73011	155989	0,013	0,51
4				6413	56711	64796	127920	0,011	0,41
5				9280	53529	61202	124011	0,010	0,40

Поле/ буква	Фамилия	Имя	Отчество	Улица	Дом	Квартира	Сумма	Доля в базе	Кол-во в строке
6				2223	50546	55101	107870	0,009	0,35
7				0	38850	51778	90628	0,007	0,29
8				242	41645	46558	88445	0,008	0,28
9				107	36752	44346	81205	0,007	0,26
0				17457	38454	42914	98825	0,008	0,32

На рис. 49 приведена диаграмма частоты повторений символов для общей строки идентификаторов в соответствии с табл. 19.

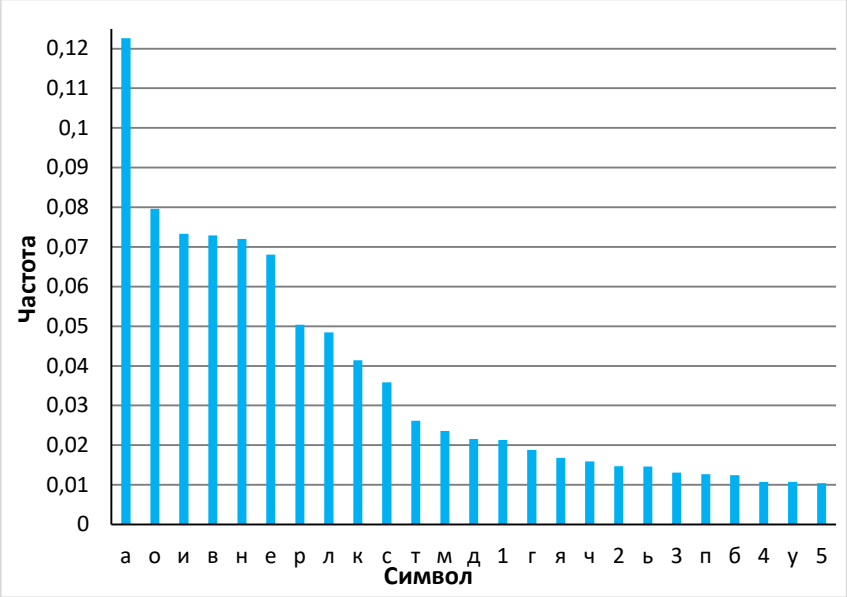

Рис. 49. Частота повторений символов в общей строке идентификаторов

В строке идентификаторов длиной $N = 73$ символа наиболее часто повторяются следующие символы: пробел – 34 раза, «а» – 5 раз, «в», «е», «и», «н», «о» – по 3 раза, «к», «л», «р» – по 2 раза. Не повторяются – 13 символов. Вероятность $p_{\text{сумм}}$ совпадения суммарного состава 73 повторяющихся символов у двух разных лиц не более вероятности совпадения 13 неповторяющихся символов, равной обратному количеству вариантов сочетаний по 13 из 73,

$$p_{\text{сумм}} < 1{,}16 \cdot 10^{-14},$$

то есть эта вероятность пренебрежимо мала, а значит, нарушитель однозначно найдет в обезличенной базе запись известного ФЛ по его прочим данным.

Пробелы при посимвольном сравнении игнорируются, смещение 13-ти одиночных символов определяется однозначно, следовательно, в формируемой таблице смещений останется $m = 39 - 13 = 26$ неоднозначных (сомнительных) мест.

По формуле (17) построена диаграмма зависимости количества вариантов совпадений F_v от количества совпадений c, которая приведена на рис. 50.

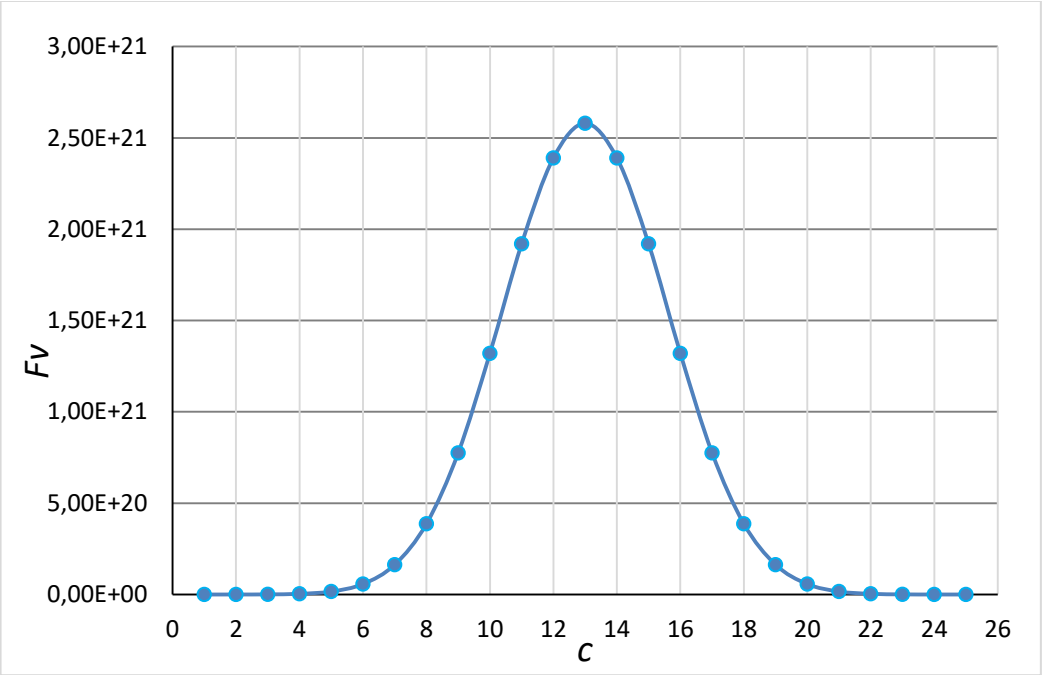

Рис. 50. Диаграмма зависимости количества вариантов совпадений F_v от количества совпадений c. $N = 73$, $m = 26$

Из рис. 50 видно, что $F_v(c)$ имеет явный максимум при значении $c = 13$, которое по (16) и является математическим ожиданием C_0, при этом среднеквадратичное отклонение $\sigma = 2{,}5$, т.е. наибольшая часть вариантов находится в диапазоне от 11 до 15. При таком значении c применение третьей известной записи ФЛ нарушителем необходимо. Поскольку ошибочно определенные символы отличаются друг от друга, их повторяемость уменьшается пропорционально их частотности и при сужении диапазона становится меньше двух (например, для символов «к», «л», «р»). При расчетах учитывалось, что к ошибке позиционирования приводит не только совпадение одинаковых символов

двух строк, но и совпадение мест различных повторяющихся символов. Но не все совпадения различных букв возможны семантически по нормам русского языка. Наиболее вероятны ошибки позиционирования при совпадении места в парах типа «гласная-гласная», либо «согласная-согласная». Повторяющихся символов должно стать меньше, что подтверждено экспериментом: количество совпадений сократилось до 7.

Кроме того, соседнюю с однозначно определенным символом позицию не может занимать любой произвольный символ. Для учета этого фактора был проведен анализ сочетаемости символов (в том числе пробелов) для текстовых идентификаторов (всех, кроме номера дома и квартиры), который показал, что 44% различных сочетаний в сумме составляют 99% от количества всех сочетаний, то есть большая часть сочетаний является маловероятными и даже невозможными. На рис. 51 приведена диаграмма частоты сочетаний символов в строке идентификаторов (150 наиболее частых сочетаний в сумме составляют 83%).

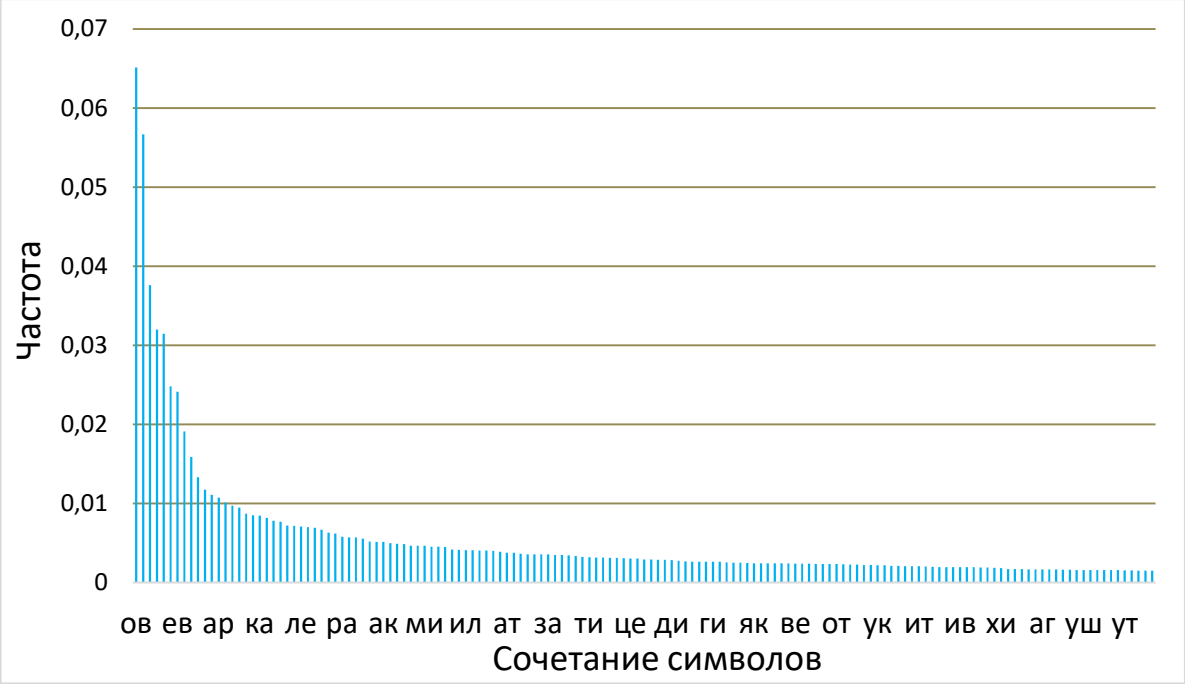

Рис. 51. Частота сочетаний символов в идентификаторах A_1 – A_4. Знак подчеркивания означает сочетание с пробелом, т.е. символ – последний в идентификаторе

В результате анализа сочетаемости символов ожидалось уменьшение количества совпадений до $c = 2$, что подтверждено экспериментом. При таком

значении c использование третьей известной записи ФЛ нарушителем может оказаться необязательным. Расчеты показали, что даже без учета сочетаемости символов использование третьей известной записи снижает количество совпадений до 4 (количество вариантов – до 24, что близко к U =20 в модели нарушителя), то есть практически приводит к решению задачи с точки зрения нарушителя.

3.2.6. Результаты применения модели нарушителя при перемещении битов внутри идентификатора

Для проверки гипотезы об эффективности модели нарушителя при перемещении битов внутри идентификатора был выбран атрибут «фамилия». В соответствии с табл. 18 в атрибуте «фамилия» наиболее вероятно содержится 8 значимых символов, что без учета кодовых страниц дает 64 бита. Расчеты частотности битов показали наиболее вероятное распределение − 33 бита «1» и 31 бит «0», но для упрощения использовалось симметричное распределение − по 32 бита. Расчет проводился по отличной от символьной строки схеме − сначала для битов «1», затем для битов «0», затем для их сочетаний. Поскольку все функции $F_v(c)$ симметричны, их максимум (соответствующий математическому ожиданию количества совпадений C_0) находится в середине диапазона, а количество вариантов на этот факт не влияет. Сужение диапазона количества совпадений при переходе к следующей известной нарушителю записи ФЛ хоть и происходит, но достаточно медленно. Эксперимент показал, что для первой записи значение C_0 = 35, для второй C_0 = 32, для третьей C_0 = 28, что показывает неэффективность модели нарушителя по причине большого количества необходимых нарушителю записей G даже для одного идентификатора. Следовательно, работа алгоритма в случае полной строки идентификаторов также неэффективна.

Задача оценки эффективности алгоритма обезличивания ПД может быть поставлена в альтернативном варианте: определить количество известных нарушителю записей достаточное для формирования удовлетворительной

таблицы смещения битов. Выдвинута гипотеза, что при наличии у нарушителя известных записей ФЛ в количестве гораздо большем, чем 5 (для больших баз с объемом 1млн. записей ФЛ и более это вполне возможно), задача построения таблицы смещений будет решена с достаточной точностью. Так как в ручном режиме произвести такой объем вычислений практически невозможно, то необходимо разработать специальное программное обеспечение, что выходит за рамки принятой модели нарушителя.

3.2.7. Результаты применения модели нарушителя при перемешивании полей внутри группы записей

По формулам (18) и (19) было рассчитано наиболее вероятное количество вхождений для всех исследованных идентификаторов. Результаты расчетов приведены в табл. 20.

Таблица 20

Количество вхождений в группу 511 записей для различных идентификаторов

Идентификатор	**Случайное вхождение**	**Повторное вхождение**
Фамилия	0,0132	0,0115
Имя	1,185	1,181
Отчество	2,371	2,366
Улица	1,059	1,057
Номер дома	0,727	0,725
Номер квартиры	0,709	0,707
Сумма	6,064	6,047

Из таблицы видно, что наиболее редкие вхождения наблюдаются у атрибута «фамилия», следовательно, его и нужно принять в качестве первого идентификатора для подтверждения принадлежности прочих данных первому известному ФЛ. С вероятностью 98,7% прочие данные не могут совпасть у ФЛ с одинаковыми фамилиями. Строка «сумма» в табл. 20 показывает, что наиболее вероятное количество повторных вхождений для всей группы идентификаторов равно 6. Теоретически 6 различных повторных вхождений порождает 6! = 720

ошибочных вариантов смещений, но нахождение уже первого смещения для фамилии, например на расстоянии 100 записей от прочих данных, сразу сужает область поиска до 511 − 100 = 411 записей. Это снижает суммарное количество вхождений остальных идентификаторов до 4, а количество ошибочных вариантов снижается до 4! = 24. Таким образом, диапазон группы записей должен уменьшиться до 256, а количество ошибочных вариантов до 3! = 6.

Применение второй записи нарушителем становится необходимым, аналогичные операции поиска в ней приведут к аналогичным результатам. Но при совместном рассмотрении этих результатов вероятность ошибки оценивается как вероятность совпадения смещений идентификаторов двух пар независимых лиц, то есть 1/256. Таким образом, использование второй записи однозначно приведет к решению задачи.

3.2.8 Результаты применения модели нарушителя при перемешивании символов внутри группы записей

Для применения модели нарушителя при перемешивании символов внутри строки уже был произведен анализ количества символов (см. табл. 19) и частотный анализ символов для каждого идентификатора. Аналогично перемешиванию полей в группе записей, для подтверждения принадлежности прочих данных был выбран идентификатор «фамилия». Результаты частотного анализа символов для этого идентификатора приведены на рис. 52.

Для использования подхода, примененного при перемешивании символов в строке, необходимо соблюдение условия $2m < N$, где $N = 511$, а значение m зависит от структуры таблицы и определяется форматом полей идентификаторов в базе.

Например, поле фамилии имеет длину 15 символов, которые в повернутой таблице займут 15 строк. Несмотря на то, что в экспериментальной базе наиболее вероятная длина фамилии − 8 символов, в 8-й строке этой таблицы наиболее вероятно наличие 231 значимого символа и 280 пробелов. Т.е. в каждой отдельно взятой строке вероятное количество пробелов, а также совпадений символов m

будет различным. Например, в первых трех строках символ «а» повторится 62 раза, символ «б» − 7 раз, и наиболее вероятна ситуация, когда все 511 мест будут заполнены различными, но повторяющимися символами. Расчеты показывают, что условие $2m < N$ соблюдается, только начиная с 7-й строки. Для остальных идентификаторов это условие также соблюдается, начиная не с первой строки.

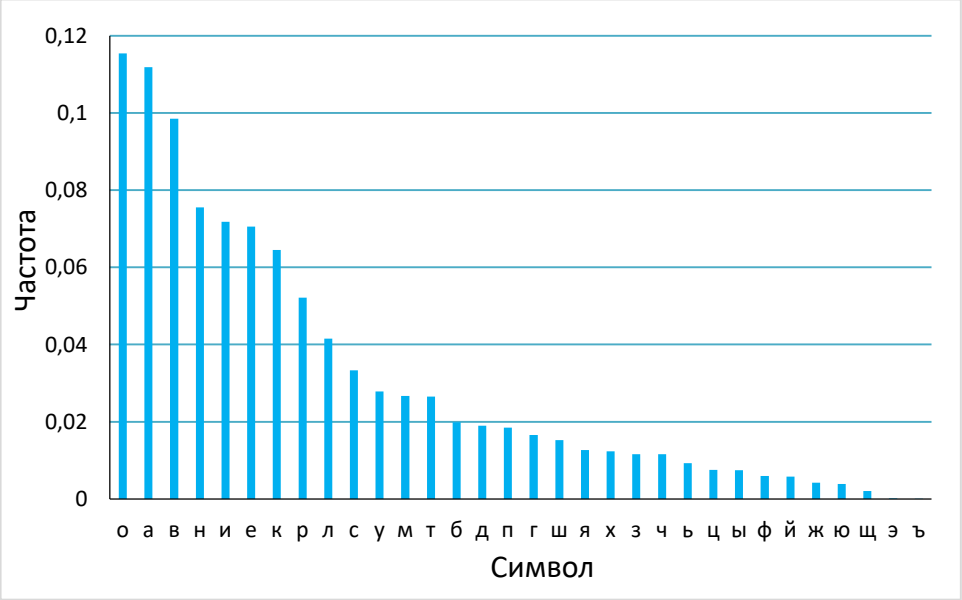

Рис. 52. Частота символов в идентификаторе «фамилия»

Таким образом, первые 6 строк фамилии должны рассчитываться по методу, примененному для перемешивания битов в строке. Поскольку длина строки символов в данном случае гораздо больше, чем использовалось в расчетах для перемешивания битов, то в худшем случае процедура деобезличивания может прерваться уже на начальном этапе − подтверждения принадлежности прочих данных известному лицу. Следовательно, можно сделать вывод о неэффективности модели нарушителя для рассмотренного варианта метода перемешивания.

3.3. Выводы по главе 3

В соответствии с поставленной в разделе 1.5 задачей сформирована таблица смещений элементов идентификаторов на основе уточненных параметров модели нарушителя, в том числе, параметра целесообразности действий нарушителя.

Произведен расчет вариантов искажения ПД для нескольких алгоритмов для искажающих методов обезличивания, что позволило оценить возможность восстановления этих алгоритмов в зависимости от их сложности и, тем самым, предложить рекомендации по применению рассмотренных методов обезличивания.

В результате применения модели нарушителя сделаны следующие выводы:

- методы обезличивания, использующие перестановку полей идентификаторов в группе записей базы, не являются эффективными;

- методы, использующие перестановку символов, эффективны только при перемешивании между различными записями, но не эффективны в пределах одной записи;

- методы, использующие перестановку битов, обладают максимальной эффективностью.

Глава 4. Функциональная схема реализации обезличивания методом введения идентификаторов

Задачей этой главы является реализация функциональной схемы взаимодействия частей базы ПД, обезличенной методом введения идентификаторов. Функциональная схема должна учитывать регламент процесса обработки ПД, параметры идентификатора связи и возможности нарушителя, заложенные в модель нарушителя. В соответствии с предлагаемой схемой модернизируется структура базы данных, интерфейс ввода-вывода ПД, после чего делается вывод об эффективности алгоритма обезличивания.

4.1. Алгоритм метода введения идентификаторов

В отличие от искажающих методов обезличивания метод введения идентификаторов позволяет проводить обработку обезличенных данных без предварительного деобезличивания. Согласно [2], метод введения идентификаторов состоит в том, что единая база (B) разделяется на две базы:

– таблица перекрестных ссылок, в которой некий набор идентифицирующих атрибутов однозначно сопоставляется с неким абстрактным идентификатором, который может быть либо атрибутом базы (номер паспорта, ИНН и т.п.), либо искусственно сгенерированным кодом, причем количество записей этой таблицы равен количеству ФЛ;

– база прочих данных, в которой некоему абстрактному идентификатору однозначно сопоставляется несколько записей с прочими данными – не значимыми с точки зрения идентификации, но определяющими суть обработки.

В процессе разделения базы данных B необходимо решить три проблемы:

– выбрать атрибуты (идентификаторы) для включения в таблицу перекрестных ссылок;

– выбрать или сгенерировать связующий идентификатор;

– обеспечить связь между двумя базами.

В рамках предложенной функциональной схемы в качестве идентификатора связи используется бумажный носитель – бланк рецепта на получение лекарств.

Такой идентификатор является частью технологического процесса и не вносит дополнительных сложностей и затрат при обработке данных, обеспечивая безопасность обработки обезличенных данных на всех этапах. Более подробно реализация метода изложена в разделах 4.3 и 4.4.

4.2. Описание информационной системы до внедрения обезличивания

В качестве объекта исследования выбрана ИСПДн льготного лекарственного обеспечения Челябинской области, принадлежащая оператору АО «Областной аптечный склад», состоящая из центра обработки данных (ЦОД) на сервере и 103 аптечных пунктов. Общий объем базы B_2 составлял около 330 тысяч записей ПД ФЛ, а минимальный объем базы аптечного пункта составлял 4300 записей.

До модернизации технологический процесс обработки ПД предполагал хранение полных ПД пациентов в базах данных под управлением СУБД Microsoft SQL Server 2012, как в ЦОД, так и локально во всех аптечных пунктах. Периодически происходил двусторонний обмен данными между ЦОД и каждым аптечным пунктом с использованием незащищенного канала связи.

Схема взаимодействия субъектов и объектов технологического процесса до модернизации приведена на рис. 53.

Рис. 53. Схема взаимодействия баз данных до модернизации

Базы данных аптечных пунктов, в отличие от базы в ЦОД, не были защищены, угрозы хищения при хранении ПД, перехвата при передаче ПД, прочие угрозы несанкционированного доступа при обработке ПД в аптечных пунктах были актуальными. Ввод информации о выданных физическим лицам рецептах в базу данных ЦОД осуществлялся лечебно-профилактическими учреждениями (ЛПУ) по защищенному каналу. Ввод информации о выдаче лекарств по рецепту в базу данных аптечных пунктов осуществлялся с использованием прикладного программного обеспечения (ПО) собственной разработки путем считывания информации в виде штрих-кода и других текстовых реквизитов с бумажного носителя (рецепт). Форма рецепта утверждена Приказом Минздрава России от 20.12.2012 № 1175н [92] (в настоящее время также действует Приказ Минздрава России от 14.01.2019 №4н [93]). Пример заполнения формы рецепта на бумажном носителе приведен на рис. 54.

Рис. 54. Пример заполнения рецепта на бумажном носителе

Синхронизация данных базы аптечного пункта с базой ЦОД осуществлялась с помощью идентификатора связи, который формировался следующим образом:

– вводились фамилия, имя, отчество пациента из рецепта;

– определялся централизованный идентификатор (сгенерированный в ЦОД);

– дополнительно использовался СНИЛС, считываемый с штрих-кода рецепта.

При вводе-выводе ПД с помощью прикладного ПО в аптечном пункте возникал риск несанкционированного доступа к чувствительным данным со стороны злоумышленника. Форма заполнения рецепта в рамках прикладного ПО приведена на рис. 55.

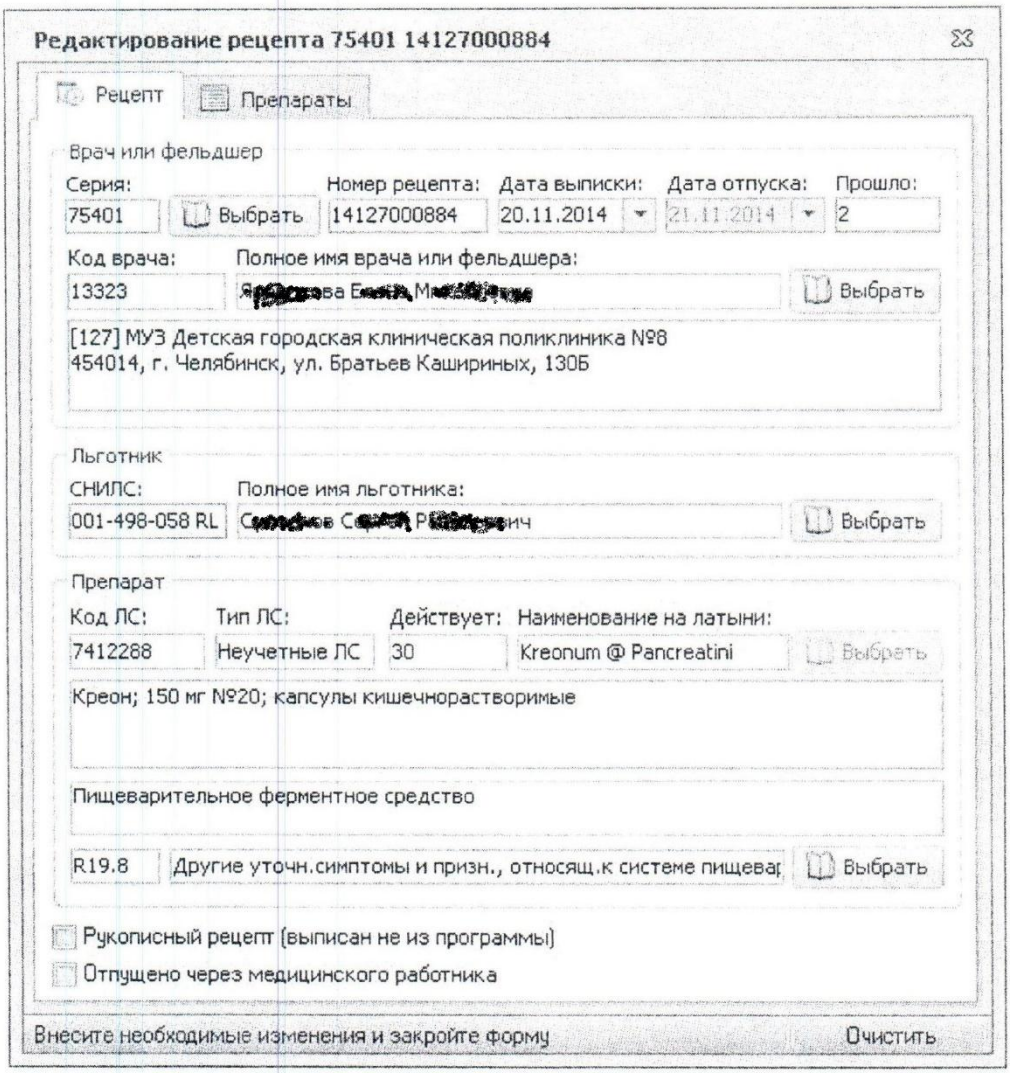

Рис. 55. Форма заполнения рецепта в прикладном ПО

Структура хранения данных показана в табл. 21. Примеры хранения ПД в базе данных до внедрения обезличивания приведены в приложении А (рис. 59 – 61).

Таблица 21

Структура базы данных льготного лекарственного обеспечения до модернизации

Таблица данных	**Описание**	**Чувствительные данные**	**Прочие данные**	**Идентификатор**
People	Перечень пациентов	PeopleFullName (фамилия A_1, имя A_2, отчество A_3 пациента), PeopleBirthday (дата рождения A_8 пациента)	ActiveDate, PreferentialCategory	PeopleInsuranceNumber, (СНИЛС)
Doctor	Перечень врачей	DoctorFullName (фамилия A_9, имя A_{10}, отчество A_{11}врача)	DoctorActive	HospitalID (код ЛПУ), DoctorID (код врача)
DrugStore	Перечень аптек	DrugStoreDirector (фамилия, имя отчество директора	DrugStoreName, DrugStorePhone	DrugStoreID (код аптеки)
Recipe	Перечень рецептов	InvoiceManager, InvoiceComment (фамилия, инициалы работника аптеки)	RecipeDate, DrugCode, PeopleInsuranceNumber, DrugStoreID, HospitalID, DoctorID, RecipeStatus	RecipeCode (номер рецепта)

В качестве идентифицирующих атрибутов использовались:

– фамилия A_1 (15 символов), имя A_2 (10 символов), отчество A_3 (15 символов) пациента, полученные в результате расформирования единого поля PeopleFullName (полное имя пациента);

– дата рождения пациента A_8 (10 символов), поле PeopleBirthday;

– фамилия A_9 (15 символов), имя A_{10} (10 символов), отчество A_{11} (15 символов) доктора, полученные в результате расформирования единого поля DoctorFullName (полное имя доктора).

В качестве возможных рассматривались следующие алгоритмы обезличивания:

– выделение в таблицу идентификаторов атрибутов A_1, A_8, A_9 с сокращением атрибутов A_2, A_3, A_{10}, A_{11} до инициалов (для визуального контроля);

– выделение в таблицу идентификаторов всех перечисленных атрибутов A_1, A_8, A_9, A_2, A_3, A_{10}, A_{11} без исключения.

4.3. Модель нарушителя для метода введения идентификаторов. Реализация схемы взаимодействия разделенных частей базы данных

Общая для различных методов модель нарушителя, описанная в разделе 2.1, основанная на наличии в базе ПД известных нарушителю ФЛ, предполагает уточнение параметров G (количество известных нарушителю записей о ФЛ), U (максимальное количество записей для дальнейшей обработки, полученных в результате деобезличивания) и алгоритма деобезличивания, зависящего от метода обезличивания. При этом должна быть оценена возможность успеха (найдено менее, чем U записей) при заданных параметрах G и алгоритма деобезличивания.

Максимальное количество известных нарушителю записей в базе $G = 5$.

Максимальное количество записей, полученных в результате деобезличивания, при котором поиск считается эффективным, $U = 20$.

Технологический процесс обработки ПД предполагает размещение обезличенной базы данных локально во всех аптечных пунктах, а таблицы перекрестных ссылок – на сервере в ЦОД.

Алгоритм действий нарушителя – следующий:

– из-за отсутствия в обезличенной базе каких-либо идентифицирующих атрибутов (при выделении в таблицу всех идентификаторов) использование ПД лиц, известных нарушителю, позволяет ему ограничить зону поиска идентификаторами аптек известного населенного пункта до минимального количества пациентов аптечного пункта – 4 тыс. записей;

– при наличии в базе инициалов пациентов и докторов задача нарушителя сводится к получению менее U записей с известными инициалами.

На этапе выбора идентификаторов из базы данных аптечного пункта сначала были устранены чувствительные данные, которые не являются

необходимыми для ведения перечня рецептов, а именно, поля DrugStoreDirector, InvoiceManager, InvoiceComment. Далее, с помощью методики, описанной в главе 2, произведена оценка необходимости обезличивания по атрибутам «полное имя пациента», «полное имя врача», а также по атрибуту «дата рождения».

Как уже было показано в разделе 2.3.7, при размере базы данных 310 тыс. записей для сочетания атрибутов «инициалы» вероятность идентификации $W_{ИО} = 0{,}002109$. В то же время в разделе 2.4.2 показано, что для такого же количества записей базы вероятность идентификации по атрибуту «имя» составляет $W_2 = 0{,}002434$, а при уменьшении объема до 4 тыс. записей увеличивается до $W_2 = 0{,}075$. При экстраполяции этого изменения количества записей для такого же атрибута «инициалы» на изменение объема базы с 329 тыс. записей до 4 тыс. записей, был сделан предварительный вывод о том, что значение вероятности идентификации явно превысит нормативное $W_{норм} = 0{,}05$. Полученный предварительный вывод был подтвержден экспериментом, в процессе которого количество найденных записей $Q_{ИО}$ с известными инициалами в объеме 4 тыс. записей не превысило 8. В результате было сделано заключение о необходимости обезличивания сочетания атрибутов «инициалы».

При принятии решения о необходимости обезличивания для аптечных пунктов для расчетов использовано минимальное количество записей пациентов в одном аптечном пункте – 4 тыс. Исследование позволило сделать следующие выводы:

– атрибуты «полное имя пациента», «полное имя врача» и «дата рождения» обязательны для обезличивания в ЦОД и аптечном пункте;

– сокращение атрибутов «имя» и «отчество» до инициалов не является достаточным для обезличивания в аптечном пункте.

Таким образом, в таблицу перекрестных ссылок были перенесены следующие атрибуты: PeopleFullName, PeopleBirthday, DoctorFullName и соответствующие идентификаторы связи с базой рецептов PeopleInsuranceNumber, HospitalID, DoctorID. Полученная таблица перекрестных ссылок размещена в

центре обработки данных на сервере, за пределами аптечного пункта, и стала недоступна для злоумышленника.

На этапе выбора/генерации связующего идентификатора было принято решение использовать существующий атрибут СНИЛС (поле PeopleInsuranceNumber), так как он в наибольшей степени соответствует требованиям, изложенным в работе [74] (является однозначным, всеобщим и абстрактным).

Технологический процесс и соответствующее программное обеспечение обработки рецептов были модернизированы путем автоматизированного считывания идентификаторов с бумажного носителя (по штрих-коду). Центр обработки данных на сервере не модернизировался. Полный двусторонний обмен данными с центром обработки данных на сервере был заменен на односторонний обмен обезличенными данными от аптечного пункта к серверу. При этом на рабочих местах во время сеансов вместо чувствительных ПД использовались абстрактные идентификаторы.

На этапе обеспечения связи между таблицей перекрестных ссылок и базой прочих данных использована полезная модель «Система взаимодействия разделенных баз персональных данных информационной системы» [64]. Схема взаимодействия субъектов и объектов технологического процесса после модернизации приведена на рис. 56. Патент на полезную модель приведен в приложении Б (рис. 62).

Согласно внедренной схеме, таблица перекрестных ссылок содержится в ЦОД и через межсетевой экран заполняется в ЛПУ, где пациенту выдается рецепт. База данных отпуска рецептов физически также расположена в ЦОД. Но таблица перекрестных ссылок защищена, а БД с прочими данными является обезличенной и может находиться в открытом (для чтения) доступе. Эта БД является компиляцией баз данных аптечных пунктов (База аптеки), которые также являются открытыми (для чтения). Доступ на запись к базам аптек имеют только сотрудники аптечных пунктов.

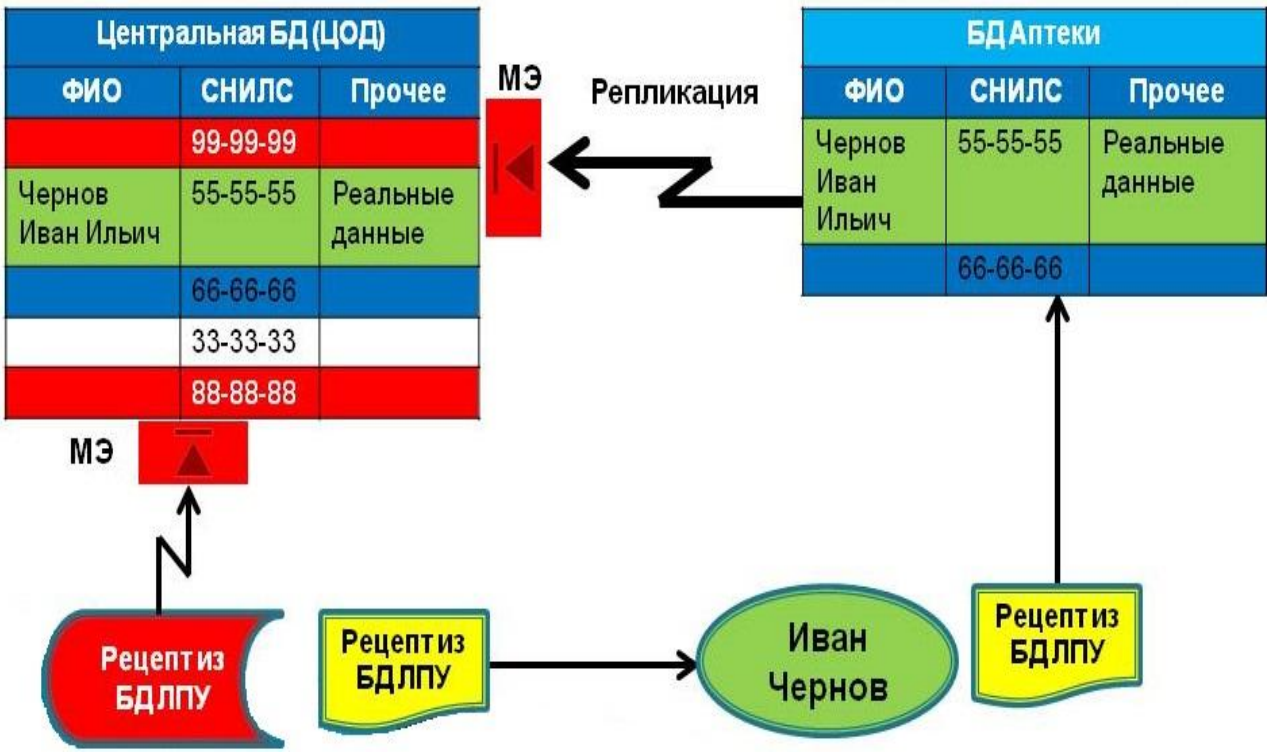

Рис. 56. Схема взаимодействия разделенных баз данных после модернизации

В качестве внешнего носителя идентификатора СНИЛС использован бумажный рецепт. Преимуществами такого подхода являются:

– возможность автоматизированного считывания по штрих-коду с целью исключения ошибок ввода информации;

– малый срок действия рецепта (до 30 дней) и факт его изъятия при получении лекарства, что значительно снижает риск утечки ПД и уязвимость бумажного носителя от внешних факторов во время использования;

– бумажный рецепт предусмотрен технологическим процессом, что исключает дополнительные затраты на его изготовление и считывание;

– регламентированный приказом [93] переход к технологии электронных рецептов с использованием в качестве носителя средств мобильной связи не противоречит используемому алгоритму.

Программное обеспечение аптечных пунктов было модернизировано для исключения использования идентификаторов в обезличенной базе. Полный двусторонний обмен данными аптечных пунктов с центром обработки данных на сервере был заменен на односторонний обмен обезличенными данными от

аптечного пункта к серверу. Частное техническое задание на модернизацию ведомственной автоматизированной информационной системы «Льготная аптека» и Акт о внедрении результатов диссертационного исследования приведены в приложении В (рис. 63 – 66).

4.4. Результаты внедрения схемы обезличивания

Новая структура базы данных показана в табл. 22.

Таблица 22

Структура базы данных льготного лекарственного обеспечения после модернизации

Таблица данных	**Описание**	**Чувствительные данные**	**Прочие данные**	**Идентификатор**
PeopleCR	Таблица перекрестных ссылок пациентов	PeopleFullName, PeopleBirthday		PeopleInsuranceNumber,
People	Перечень пациентов		ActiveDate, PreferentialCategory	PeopleInsuranceNumber,
DoctorCR	*Таблица перекрестн*	DoctorFullName		HospitalID, DoctorID
Doctor	Перечень врачей		DoctorActive	HospitalID, DoctorID
DrugStore	Перечень аптек		DrugStoreName, DrugStorePhone	DrugStoreID
Recipe	Перечень рецептов		RecipeDate, DrugCode, PeopleInsuranceNumber, DrugStoreID, HospitalID, DoctorID, RecipeStatus, InvoiceComment	RecipeCode

Форма заполнения рецепта приобрела вид, показанный на рис. 57.

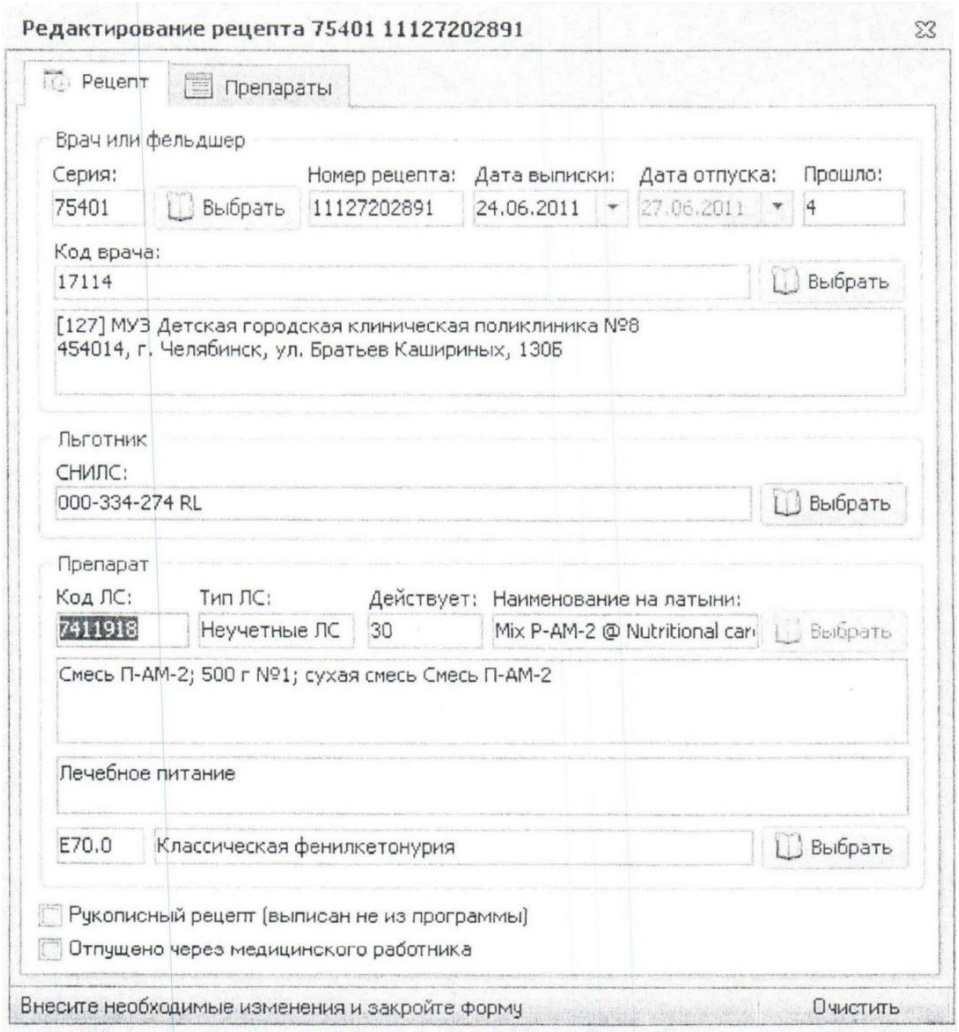

Рис. 57. Форма заполнения рецепта после модернизации. ПД пациента и врача не вводятся

В соответствии с моделью нарушителя при расчете показателей вероятности идентификации W в соответствии с (7) для всех идентификаторов были приняты следующие допущения:

– значения полей «полное имя пациента» и «полное имя врача» являются уникальными ($W = 1$ – однозначная идентификация);

– объем БД ЦОД – 329 тысяч записей пациентов;

– минимальный объем БД аптечного пункта – 4 тысячи записей пациентов;

– объем справочной БД врачей в ЦОД – 20,5 тысяч записей;

– минимальный объем справочной БД врачей в аптечном пункте – 150 записей.

На рис. 58 приведены значения W для различных атрибутов, рассчитанные до модернизации (синий цвет) и после нее (красный цвет) для ЦОД (рис. 58a) и для аптечной БД (рис. 58b).

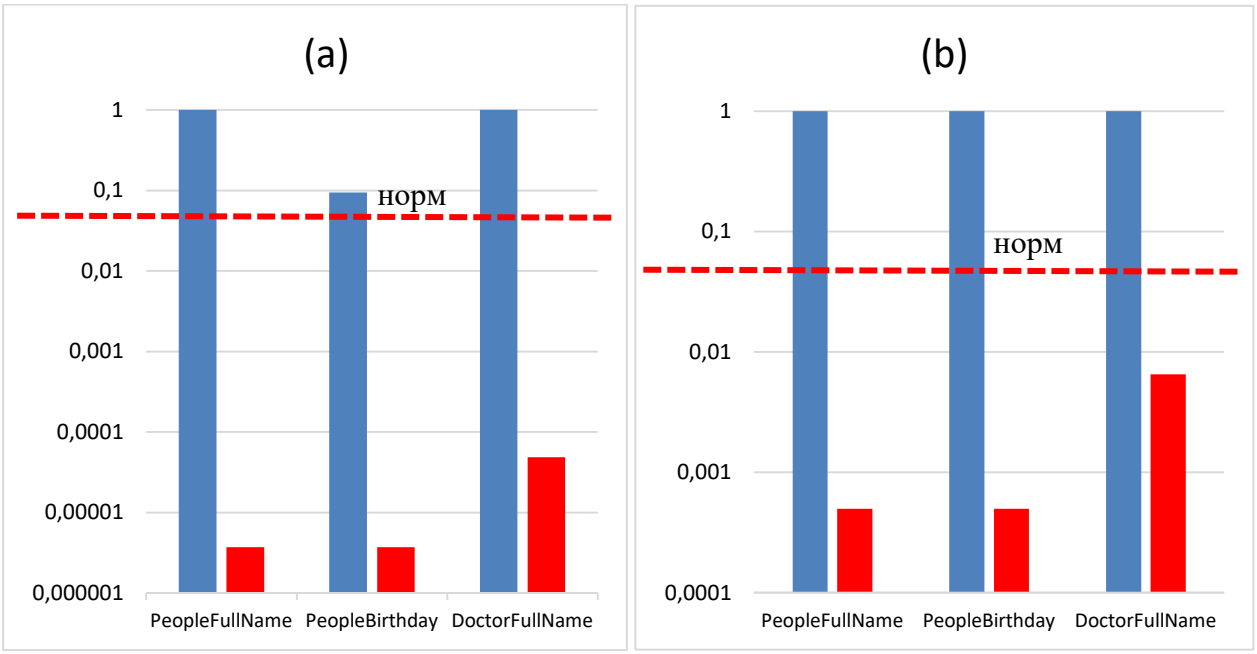

Рис. 58. Изменение вероятности идентификации пациента и врача. Линия «норм» соответствует $W = 0{,}05$

Принципиальным отличием внедренной схемы от внедрения [7], в котором на основе запатентованной полезной модели в качестве носителя для идентификатора используется сим-карта, является использование внешнего бумажного носителя. Преимуществом сим-карты является бо́льшая надежность хранения, чем у бумажного носителя. Но, при этом, во-первых, у бумажного рецепта очень короткий срок действия (до 30 дней, после чего пациенту выдается новый рецепт); во-вторых, для сим-карты требуется дополнительное оборудование для хранения и считывания (например, мобильный телефон), что приводит к увеличению затрат не только на стороне аптечного пункта, но и на стороне лечебного учреждения, выдающего рецепт. Использование сим-карты имеет преимущество только в тех системах, где не предусмотрено использование стандартных технологических носителей, поэтому внедрение этой технологии в сфере здравоохранения представляется сомнительным.

Экономический эффект при использовании предложенной схемы был достигнут за счет экономии затрат на СЗИ на каждом рабочем месте аптечного пункта. Для обеспечения СЗИ 103 рабочих мест в аптечных пунктах потребовалось бы около 1,9 млн. руб. Расходы на модернизацию технологического процесса и программного обеспечения составили около 0,5 млн.

руб. Таким образом, был получен экономический эффект 1,4 млн. руб. Кроме того, предложенная модель предполагает практически неограниченное масштабирование по количеству аптечных пунктов.

4.5. Выводы по главе 4

В соответствии с поставленной в разделе 1.5 задачей база ПД обезличена методом введения идентификаторов. При этом была обеспечена связь между удаленной таблицей перекрестных ссылок и базой обезличенных данных, расположенных в разных контролируемых зонах, без добавления в процесс обработки данных дополнительных внешних носителей и СКЗИ. Для решения задачи в ИСПДн «Льготная аптека» предложена защищенная патентом полезная модель на базе внешнего идентификатора в виде бумажного рецепта со штрих-кодом, содержащим идентификатор пациента СНИЛС.

В соответствии с условиями использования модели модернизированы структура базы данных и программное обеспечение. Внедрение выполнено в 103 аптечных пунктах, объем базы ПД – 329 тыс. записей о пациентах. Получен значительный экономический эффект.

Заключение

В процессе работы были получены следующие результаты и сделаны следующие выводы:

1. Проведен анализ современного состояния разработанности и реализации методов обезличивания ПД. Определены принципы работы методов обезличивания в соответствии с нормативной базой РФ, параметры методов и возникающие при их реализации проблемы.

2. Разработана математическая модель идентификации ФЛ, что позволило сформулировать количественный критерий необходимости обезличивания, который определяется как характеристика функции распределения параметров отдельных идентификаторов или их сочетаний, в частности:

– установлен степенной вид распределения всех исследованных идентификаторов и некоторых их сочетаний, кроме идентификатора «дата рождения», для которого установлена зависимость типа гамма-распределения;

– установлен степенной характер зависимости вероятности идентификации от количества записей базы ПД для всех атрибутов, что позволяет, основываясь на результатах, полученных для выборки из базы ПД, масштабировать решение об обезличивании для всей БД.

3. Разработана функциональная модель нарушителя для контроля эффективности искажающих методов обезличивания в зависимости от сложности алгоритма искажения, в частности:

– разработаны алгоритмы восстановления (деобезличивания) искаженных идентификаторов без применения специального программного обеспечения;

– предложена методика оценки эффективности искажающих методов обезличивания на основе модели нарушителя.

4. Предложено решение проблемы передачи информации между таблицей идентификаторов и базой обезличенных данных на основе полезной модели с

использованием внешнего идентификатора. Решение внедрено в сфере здравоохранения, получен заметный экономический эффект.

Перспективы дальнейшей разработки темы исследования заключаются в следующем:

1. Создание нормативной базы показателей вероятности идентификации по любым сочетаниям атрибутов для любого количества записей БД.

2. Использование методики для оценки эффективности искажающих методов обезличивания, основанных на алгоритмах, не рассмотренных в работе.

3. Разработка программного обеспечения для определения количества известных нарушителю записей, достаточного для формирования удовлетворительной таблицы смещения битов в строке идентификаторов.

Список сокращений

БД	база данных
ИСПДн	информационная система персональных данных
ЛПУ	лечебно-профилактическое учреждение
МЭ	Межсетевой экран
ПД	персональные данные
ПО	программное обеспечение
СЗИ	средство защиты информации
СКЗИ	средство криптографической защиты информации
СУБД	система управления базой данных
ФЛ	физическое лицо
ЦОД	центр обработки данных

Список терминов

Атрибуты:	Отдельные данные ФЛ, содержащиеся в базе ПД
Деобезличивание:	Метод обработки обезличенных данных, в результате которой возможно определить принадлежность этих данных ФЛ
Значимый атрибут:	Атрибут, идентификация по которому наиболее вероятна
Идентификатор:	Атрибут (набор атрибутов), который используется для идентификации
Идентификация по атрибуту (набору атрибутов):	Сравнение искомого значения атрибута (набора атрибутов) с рассматриваемым значением в базе данных
Искажающие методы обезличивания:	Методы, основанные на совместном хранении в свободном доступе значимых (предварительно измененных) и прочих (не измененных) атрибутов ФЛ
Обезличивание:	Вид обработки ПД, в результате которой невозможно без дополнительной информации определить принадлежность этих ПД физическому лицу
Объем базы данных	Количество записей базы данных
Разделяющие методы обезличивания:	Методы, основанные на раздельном хранении значимых (в защищенном виде) и прочих (в свободном доступе) атрибутов ФЛ
Таблица смещения:	Таблица, в которой каждому элементу соответствует величина смещения относительно начала списка элементов
Таблица подстановки:	Таблица, в которой каждому элементу соответствует измененное значение этого элемента

Формула смещения:	Выражение, параметром которого является начальное смещение элемента относительно начала списка, а результатом выполнения является величина конечное смещение элемента относительно начала списка элементов
Чувствительный атрибут:	То же, что и «Значимый атрибут»

Список литературы

1. О персональных данных [Электронный ресурс]. Федеральный закон от 27.07.2006 № 152-ФЗ. Доступ из справочной системы «КонсультантПлюс» (дата обращения: 10.05.2019).

2. Об утверждении требований и методов по обезличиванию персональных данных [Электронный ресурс] : Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 05.09.2013 № 996 Режим доступа: http://54.rkn.gov.ru/ protection/acts/p13580/ (дата обращения: 14.05.2019).

3. M. Terrovitis, N. Mamoulis, P. Kalnis Privacy-preserving anonymization of set-valued data // Proceedings of the VLDB Endowment. Aug. 2008. Volume 1. Issue 1. pp. 115–125.

4. Y. He, J. F. Naughton Anonymization of SetValued Data via TopDown, Local Generalization // Proceedings of the VLDB Endowment. Aug. 2009. Volume 2. Issue 1. pp. 934–945.

5. Kiran P., Kavya N. P. SW-SDF Based Personal Privacy with QIDB-Anonymization Method // (IJACSA) International Journal of Advanced Computer Science and Applications. 2012. Vol. 3. No.8. Pages 60–66.

6. Кучин И.Ю. Обработка баз данных с персонифицированной информацией для задач обезличивания и поиска закономерностей // Диссертация ктн. 2012. 132 с.

7. Пат. RU 121 618 Патент на полезную модель. Система идентификации субъекта персональных данных по обезличенным данным / Е.С. Волокитина (RU). – № 2011139879/08; заявл. 30.09.2011.

8. Волокитина Е.С. Метод и алгоритмы гарантированного обезличивания и реидентификации субъекта персональных данных в автоматизированных информационных системах // Диссертация ктн. – Санкт-Петербург: Издательство Санкт-Петербургского национального исследовательского ун-та информационных технологий, механики и оптики, 2013. 183 с.

9. Денисов М.И., Чехонин К.А. Защита персональных данных в информационной системе медицинского учреждения методом обезличивания // Научно-техническое и экономическое сотрудничество стран АТР в XXI веке. 2013. Том: 1. С. 229–232.

10. Кошкаров А.А., Халафян А.А. Система управления базами данных льготного лекарственного обеспечения в Краснодарском крае с использованием облачных технологий // Политематический сетевой электронный научный журнал кубанского государственного аграрного университета. 2015. № 109 (05). С. 451–467.

11. Ноздрина А.А., Применко Д.В. Метод обезличивания персональных данных, основанный на введении идентификаторов и хешировании // Молодежь и новые информационные технологии. Всероссийская научно-практическая конференция молодых ученых. 2016. № 1. С. 64–67.

12. Воронин В.В., Нехай Н.Л. Защита персональных данных в информационных системах методом обезличивания // Информационные технологии XXI века: сборник научных трудов. 2017. С. 479–483.

13. Бондаренко К.О., Козлов В.А. Универсальный быстродействующий алгоритм процедур обезличивания данных // Известия ЮФУ. Технические науки. 2015. № 11 (172). С. 130–142.

14. Макарова Е.А., Лагерев Д.Г. Применение методов обезличивания персональных данных для обеспечения защиты конфиденциальной информации в медицинских организациях // Сборник трудов конференции «Молодые ученые - ускорению научно-технического прогресса в XXI веке» Брянск. 2016. С. 280–285.

15. Трифонова Ю.В., Жаринов Р.Ф. Возможности обезличивания персональных данных в системах, использующих реляционные базы данных // Доклады ТУСУРа. 2014. № 2 (32). С. 188–194.

16. Ажмухамедов И.М., Демина Р.Ю., Сафаров И.В. Системный подход к обеспечению конфиденциальности обезличенных персональных данных в

учреждениях здравоохранения [Электронный ресурс] // Современные проблемы науки и образования: [2015]. Режим доступа: http://www.science-education.ru/ru/article/view?id=18610/, свободный (дата обращения: 16.11.2018).

17. Методические рекомендации по исполнению приказа Роскомнадзора от 5 сентября 2013 г. № 996 «Об утверждении требований и методов по обезличиванию персональных данных» [Электронный ресурс] : Руководящий документ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 13.12.2013. Режим доступа: https://rkn.gov.ru/docs/Xerox_Phaser_3200MFP_20131216122746.pdf, (дата обращения: 14.05.2019).

18. Карпова И.П. О реализации метода обезличивания персональных данных // Вестник компьютерных и информационных технологий. 2013. № 6. С. 56–60.

19. Мищенко Е.Ю. Обезличивание персональных данных как способ снижения затрат на создание системы защиты информации // XVI Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых "Безопасность информационного пространства - 2017". Сборник трудов. 2018. С. 199–203.

20. E. McCallister, T. Grance, K. Scarfone Guide to protecting the confidentiality of personally identifiable information (PII) // National Institute of Standards and Technology Special Publication 800-122. Apr. 2010. 59 pages.

21. C. Graham Anonymisation: managing data protection risk code of practice // Information Commissioner’s Office. Nov. 2012. 106 pages.

22. G. S. Nelson Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification // SAS Global Users Group Leadership Development and Nominations Committee. Apr. 2015. 23 pages.

23. Y.-R. Lee, Y.-C. Chung, J.-S. Kim, H.-K. Park Personal Health Information De-identified Performing Methods in Big Data Environments // (IJSEIA) International

Journal of Software Engineering and Its Applications. 2016. Vol. 10. No. 8. pp. 127–138.

24. B. Shehu, Sh. Ahmetaj, M. Aranitasi, A. Xhuvani Protection of Personal Data in Information Systems // International Journal of Computer Science Issues. 2013. Vol. 10. Issue 4. No 2. pp. 78–81.

25. T. Kowshiga, T. Saranya, T. Jayasudha, Prof. M. Sowmiya and Prof. S. Balamurugan Studies on Protecting Privacy of Anonymized Medical Data // International Journal of Innovative Research in Science, Engineering and Technology. 2015. Vol. 4. Issue 2. pp. 711–715.

26. Y. Morisawa, Sh. Matsune: NESTGate-Realizing Personal Data Protection with k-Anonymization Technology // Fujitsu scientific & technical journal. 2016. Vol. 50. No 3. pp. 37–42.

27. B. Zhou, J. Pei, W.S. Luk A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data // ACM SIGKDD Explorations Newsletter Volume 10 Issue 2, December 2008 Pages 12–22.

28. N. Ganz Data Anonymization and its Effect on Personal Privacy // An honors thesis presented to the School of Business, University at Albany, State University Of New York. May. 2015. 22 pages.

29. M. Kayaalp, Modes of De-identification // Proceedings of AMIA Annual Symposium. 2017. pp. 1044–1050.

30. K. Gwan-Hyung, L. Joon-Yun and O. Am-Suk Fusion of Medical IT and Big Data // Korea Computer and Information Society. 2013. vol. 21. no. 2, pp. 17–26.

31. Y.-C. Chung De-identification Policy of Personal Information and Tasks on Healthcare Big Data // Health and Welfare Forum. 2015. vol. 227. pp. 50–60.

32. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction // J Am Med Inform Assoc. 2013;20(1):84–94.

33. Мавринская Т.В., Лошкарёв А.В., Чуракова Е.Н. Обезличивание персональных данных и технологии "больших данных" (BIGDATA) // Интерактивная наука. 2017. № 6 (16). С. 78–80.

34. Столбов А.П. Обезличивание персональных данных в здравоохранении // Сибирский вестник медицинской информатики и информатизации здравоохранения. 2018. № 1–2. С. 13–23.

35. Антошечкин А.В. Анализ возможностей применения биометрических технологий для реализации процедур обезличивания персональных данных // Вестник Астраханского Государственного Технического Университета: Управление, вычислительная техника и информатика. 2018. №1. С. 27-36.

36. Геращенко О.М., Капралова Н.Н. Защита персональных данных в информационных системах методом обезличивания // Уголовно- исполнительная система сегодня: взаимодействие науки и практики. Материалы юбилейной XX Всероссийской научно-практической конференции. 2020. С. 255–257.

37. Мухаметьева Е.С., Захаров А.Б. Обезличивание как метод учета обучающихся в информационных системах образовательных организаций в случае отказа от обработки персональных данных // Научно-методическое обеспечение оценки качества образования. 2020. № 1 (9) С. 96–100.

38. Столбов А.П. О стандартизации методов псевдонимизации персональных данных в здравоохранении // Проблемы стандартизации в здравоохранении. 2017. № 9–10. С. 25-36.

39. Солдатова В.И. Защита персональных данных в условиях применения цифровых технологий // Lex russica (русский закон). 2020. № 2 (159). С. 33–43.

40. Солдатова В.И. Проблемы защиты персональных данных в условиях применения цифровых технологий // Право и экономика. 2019. № 12 (382). С. 24–34.

41. Климко Е.И. Проблема регулирования пользовательских данных // Актуальные проблемы современного права: соотношение публичных и частных

начал. Сборник научно-практических статей V международной научно-практической конференции (симпозиума). 2021. С. 141–145.

42. Сухарева Е.Р. Генетическая информация как разновидность биометрических данных в условиях цифровизации медицинской сферы // Право и образование. 2021. № 7. С. 94–100.

43. Писковский В.О., Грушо А.А., Забежайло М.И., Николаев А.В., Сенчило В.В., Тимонина Е.Е. Архитектуры безопасности в системах цифровой экономики // International journal of open information technologies. 2020. Т. 8. № 9. С. 48-52.

44. Докучаев В.А., Маклачкова В.В., Статьев В.Ю. Цифровизация субъекта персональных данных // T-Comm: Телекоммуникации и транспорт. 2020. Т. 14. № 6. С. 27–32.

45. Newman M. E. J. Power laws, Pareto distributions and Zipf's law // Contemporary Physics. 2005. No. 46. pp. 323–351.

46. Clauset A., Shalizi C. R., Newman M. E. J. Power-law distributions in empirical data // SIAM Rev.,2009. 51(4), 661–703.

47. Куркина Е.С., Куретова Е.Д. Общие закономерности распределения городов по численности населения // Прикладная математика и информатика. Труды факультета ВМК МГУ имени М.В. Ломоносова. 2011. С. 37–57.

48. Андреев В.В. Территориальное распределение населения в Российской Федерации // Экономика региона. 2017. Т. 13, вып. 3. С. 803 – 811.

49. Ибрагимова А.И., Ибрагимов М.Я., Исмагилов И.И. Процессы урбанизации в современном Татарстане // Казанский экономический вестник. Региональная экономика. 2019. №1(39). С. 57–63.

50. Манаева И. В. Распределение городов в федеральных округах России. Тестирование закона Ципфа // Экономика региона. 2019. Т. 15, вып. 1. С. 84–98.

51. Веневцева Ю.Л., Мельников А.Х., Казидаева Е.Н. Практическая хронобиология: дата рождения и возможные болезни (обзор литературы) // Вестник новых медицинских технологий. 2020. Т. 27. С. 20–29.

52. Кочуров М.Г. Влияние даты рождения на личностные особенности // Психологические науки. Международный научно-исследовательский журнал. 2019. № 1 (79). Часть 2. Январь. С. 54–61.

53. Березкин В.Г., Буляница А.Л. О некоторых демографических характеристиках членов российской академии наук в XX веке // Успехи геронтологии. 2007. Т. 20. № 1. С. 29–39.

54. Вишнякова О.А. Лавров Д.Н. Формат обмена данными в системе сбора и обработки биометрических образцов // Информационные ресурсы в образовании: материалы Международной научно-практической конференции. 2013. С. 146–149.

55. Корнеев М.Б., Тэвина А.В., Журилов Н.В. Электронные рецепты: Возможности, перспективы и проблемы // Медицинское право: теория и практика 2019. Том 5. №1 (9). С. 123–129.

56. Раузина С.Е., Шелгунов В.А., Зарубина Т.В. Проблемы и перспективы системы "Электронный рецепт" в России. Систематический обзор // Социальные аспекты здоровья населения [Электронный ресурс] 2020. 66(5):8. URL: http://vestnik.mednet.ru/content/view/1201/ 30/lang,ru/ (дата обращения: 14.06.2019).

57. Секретов М.В., Ахметов Б.С., Сериков И.В., Сауанова К.Т. Защита персональных данных больных социально значимыми заболеваниями биометрическим обезличиванием электронных историй болезни // Труды международного симпозиума «Надежность и качество». 2012. Том: 2. С. 289–290.

58. Спеваков А.Г., Калуцкий И.В., Никулин Д.А., Шумайлова В.А. Обезличивание персональных данных при обработке в автоматизированных информационных системах // Телекоммуникации. 2016. №10. С. 16–20.

59. Разуваев В.А., Бурков С.М., Савин С.З. Методы защиты персональной информации при обработке медицинских данных // Методы компьютерной диагностики в биологии и медицине – 2019: Сборник статей Всероссийской школы-семинара, посвященной 110-летию Саратовского государственного университета имени Н.Г. Чернышевского. 2019. С. 149–152.

60. Захаров А.А., Оленников Е.А., Паюсова Т.И., Зулькарнеев И.Р., Овчаренко Д.И. Оптимизация затрат на защиту персональной информации в распределенных медицинских системах // Инновационное развитие экономики. 2017. № 3 (39). С. 244-250.

61. Спеваков А.Г., Плугатарев А.В. Программа для формирования уникальной последовательности, используемой в задачах обезличивания персональных данных // Свидетельство о государственной регистрации программы для ЭВМ RU 2016661169. Патентное ведомство: Россия 2016. Номер заявки: 2016618382. Дата регистрации: 01.08.2016. Дата публикации: 03.10.2016.

62. Спеваков А.Г., Плугатарев А.В. Быстродействующее устройство формирования уникальной последовательности, используемой при обезличивании персональных данных // Патент на изобретение. RU 2665899 C1 Патентное ведомство: Россия Год публикации: 2018. Номер заявки: 2016145614. Дата регистрации: 22.11.2016. Дата публикации: 04.09.2018.

63. Гулов В.П., Иванов А.И., Язов Ю.К., Корнеев О.В. Перспектива нейросетевой защиты облачных сервисов через биометрическое обезличивание персональной информации на примере медицинских электронных историй болезни (краткий обзор литературы) // Вестник новых медицинских технологий. 2017. Т. 24. № 2.

64. Пат. RU 103 414 U1, МПК G06F 17/40 (2006.01). Система взаимодействия разделенных баз персональных данных информационной системы / Д.Н. Иванов (RU), Е.Ю. Мищенко (RU). – № 2010149391/08; заявл. 02.12.2010; опубл. 10.04.2011. Бюл. № 10. 2 с.

65. Серышев А.С., Кротов А.Д., Ефанова Н.В. Разработка приложения для обезличивания персональных данных // Цифровизация экономики: направления, методы, инструменты. Сборник материалов III всероссийской научно-практической конференции. 2021. С. 294–297.

66. Куракин А. С. Алгоритм деперсонализации персональных данных // Научно-технический вестник информационных технологий, механики и оптики. 2012. №6 (82). С. 130–135.

67. Капралова Н. Н. Перспективы применения метода перемешивания при обезличивании персональных данных // Сборник трудов конференции «Саморазвивающаяся среда технического вуза: Научные исследования и экспериментальные разработки». 2016. С. 42–45.

68. Шередин Р.В. Защита персональных данных в информационных системах методом обезличивания // Диссертация ктн. Московский государственный институт электроники и математики. Москва, 2011. 138 с.

69. Трофимов В.С., Минакова Н.Н. Система обезличивания персональных данных на основе метода генерации случайных перестановок // Измерение, контроль, информатизация. Материалы XIX международной научно-технической конференции. 2018. С. 155–159.

70. Bondarech E.A. Modification of the algorithm mixing of personal data // Ученые заметки ТОГУ(Хабаровск). 2015. Т. 6. № 2. С. 282–288.

71. ГОСТ Р 34.10-2018 от 01.06.2019. Информационная технология. Криптографическая защита информации. Процессы формирования и проверки электронной цифровой подписи [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200161706/, свободный (дата обращения: 21.09.2021).

72. Куимов В. А. Яковлев А. В. Метод обезличивания персональных данных на основе хеш-ссылок доступа к записям // Информация и безопасность. 2014. Том: 17. № 4. С. 586–591.

73. Мищенко Е.Ю., Соколов А.Н. Алгоритмы реализации методов обезличивания персональных данных в распределенных информационных системах // Доклады Томского Государственного Университета Систем Управления и Радиоэлектроники. 2019. Т. 22. № 1. С. 66–70.

74. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод введения идентификаторов // Вестник Южно-Уральского государственного университета. Серия:

Компьютерные технологии, управление, радиоэлектроника. 2015. № 3(15). С. 18–25.

75. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод изменения состава или семантики // Вестник УрФО. Безопасность в информационной сфере. 2016. № 1 (19). С. 30–38.

76. Мищенко Е.Ю., Соколов А.Н. Количественный анализ процедуры обезличивания персональных данных. Метод перемешивания // Вестник УрФО. Безопасность в информационной сфере. 2016. № 3 (21). С. 30–37.

77. Мищенко Е.Ю., Соколов А.Н. Модель нарушителя в системах обезличенных персональных данных // XVII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2018». Сборник трудов. – 2018. – С. 124–128.

78. Мищенко Е.Ю., Соколов А.Н. Определение эффективности обезличивания персональных данных с использованием модели нарушителя // Вестник УРФО. Безопасность в информационной сфере. 2020. – № 2 (36). С. 34–42.

79. Поромошкин, А.А., Баранкова И.И. Разработка модели нарушителя для медицинского учреждения // Идеи и проекты молодежи России: материалы II Всероссийской научно-практической конференции. 2019. – С. 108-113.

80. Мищенко Е.Ю., Соколов А.Н. Обезличивание персональных данных: термины и определения // Вестник УрФО. Безопасность в информационной сфере. 2013. № 1(7). С. 10–13.

81. Мищенко Е.Ю., Соколов А.Н. Обезличивание персональных данных // Актуальные проблемы автоматизации и управления. Труды научно-практической конференции. 2013. С.356–359.

82. Мищенко Е.Ю., Соколов А.Н. Количественные критерии идентификации физического лица при обезличивании персональных данных // Вестник УрФО. Безопасность в информационной сфере. 2014. № 1(11). С. 27–33.

83. Мищенко Е.Ю. Вероятность идентификации в базе персональных данных: выбор идентифицирующих атрибутов // XVIII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2019». Сборник трудов. – 2019. – С. 206–209.

84. Mishchenko E. Y., Sokolov A. N. Model of Identification of a Person in Databases of Various Sizes // 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). 2021, pp. 0407–0410.

85. Дударев О.К., Кустицкая Т.А., Овчинникова Е.В. Математическая статистика. Методические указания. Красноярск.: Сиб. гос. аэрокосмич. ун-т, 2016. 156 с.

86. Денисов В.И., Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика.

Правила проверки согласия опытного распределения с теоретическим.

Методические рекомендации. Часть I. Критерии типа χ^2 . – Новосибирск:

НГТУ, 1998. 139 с.

87. Тырсин А.Н., Серебрянский С.М. Распознавание зависимостей на основе обратного отображения // Информатика и ее применения. 2016. т. 10. Вып. 2. С. 58 – 64

88. Жуковский М.Е. Родионов И.В., Шабанов Д.А. Введение в математическую статистику. М.: МФТИ, 2017. 109 с.

89. Виленкин Н.Я. Глава III. Комбинаторика кортежей и множеств. Размещения с повторениями. М.: Наука, 1975. 208 с.

90. Рыбников К.А. Введение в комбинаторный анализ. М.: Изд-во Моск. ун-та, 1985. 308 с.

91. Ширяев А.Н., Эрлих И.Г., Яськов П.А. Вероятность в теоремах и задачах (с доказательствами и решениями). М.: МЦНМО, 2013. 648 с.

92. Об утверждении порядка назначения и выписывания лекарственных препаратов, а также форм рецептурных бланков на лекарственные препараты,

порядка оформления указанных бланков, их учета и хранения [Электронный ресурс]: Приказ Минздрава России от 20.12.2012 № 1175н (ред. от 31.10.2017). Доступ из справочно-правовой системы «КонсультантПлюс».

93. Об утверждении порядка назначения лекарственных препаратов, форм рецептурных бланков на лекарственные препараты, порядка оформления указанных бланков, их учета и хранения [Электронный ресурс]: Приказ Министерства Здравоохранения Российской Федерации от 14.01.2019 №4н. Доступ из справочно-правовой системы «КонсультантПлюс».

Приложение А

Пример заполнения базы данных льготного лекарственного обеспечения до модернизации

marker	peopleInsuranceNumber	PeopleFullName	категория льготы	пол	PeopleBirthday
5245584	000-000-001 01	Щарипова Дилноза Шоанваровна	9	Ж	05.02.1998
5246401	000-000-002 02	Щарипова Дилноза Шоанваровна	9	Ж	06.02.1998
4975473	000-116-595 RL	Ларин Виктор Матвеевич	37	М	18.01.1953
3856394	000-116-599 RL	Солоненко Ольга Петровна	27	Ж	12.10.1955
5106414	000-116-609 RL	Шелудько Ирина Андреевна	32	Ж	03.12.1972
5106417	000-117-509 RL	Требенкова Александра Федоровна	47	Ж	25.04.1942
4975472	000-117-514 RL	Гузь Валентина Марковна	47	Ж	16.01.1942
5106418	000-117-878 RL	Селиванова Екатерина Валерьевна	40	Ж	20.12.1988
5106419	000-117-879 RL	Губина Александра Дмитриевна	47	Ж	10.05.1947
5106420	000-117-889 RL	Краснова Татьяна Александровна	27	Ж	13.08.1966

Рис. 59. Таблица People

DoctorID	HospitalID	DoctorFullName	DoctorActive	DoctorActiveBegin	DoctorActiveEnd
1	234	Динмухаметова Лилия Дамировна	1	2005-01-01 00:00:00.000	2012-08-17 00:00:00.000
2	234	Манаков Владимир Глебович	1	2005-01-01 00:00:00.000	NULL
3	234	Хамидуллин Марат Равилевич	1	2005-01-01 00:00:00.000	NULL
4	234	Плаксина Ирина Петровна	1	2005-01-01 00:00:00.000	NULL
5	234	Фахретдинова Юлия Вадимовна	1	2005-01-01 00:00:00.000	2013-04-05 00:00:00.000
6	234	Базуленко Татьяна Юрьевна	1	2005-01-01 00:00:00.000	NULL
7	234	Кулуева Фарида Рифовна	1	2005-01-01 00:00:00.000	2013-04-05 00:00:00.000
8	234	Султанова Галина Маратовна	1	2005-01-01 00:00:00.000	2013-04-05 00:00:00.000
9	234	Кокурина Татьяна Владимировна	1	2005-01-01 00:00:00.000	NULL
10	234	Якупова Мастура Сунагатовна	1	2005-01-01 00:00:00.000	NULL

Рис. 60. Таблица Doctor

invoiceManager	InvoiceComment				
	1695@148223+апрос возврата "Нет в заявке"	70915.6800	.0000	70915.6800	27.03.2014 15:44
		1274.3400	.0000	1274.3400	07.02.2013 17:47
1@ZAV	Запрос возврата	.0000	764.6800	764.6800	20.08.2012 17:11
Рябиничева Л. В.	Аптека № 4 ***	1100.5500	.0000	1100.5500	10.11.2014 16:27
User@APT19-ZAV	Перемещение	.0000	830.6000	830.6000	12.08.2011 14:12
Рябиничева Л. В.	ДКП 8, Низамовой РИ	3203.5200	.0000	3203.5200	27.05.2013 15:14
Истомина А. А.	1695@14196	7001.6300	.0000	7001.6300	24.05.2012 13:40
		4940.3200	.0000	4940.3200	23.12.2011 15:18
Рябиничева Л. В.	+++СРОЧНО!! ДГКП 8, Фоменков КП	20347.5000	.0000	20347.5000	17.09.2014 15:27
Рябиничева Л. В.	Аптека № 4 ***	578.7200	.0000	578.7200	24.02.2014 16:35

Рис. 61. Таблица Invoice

Приложение Б

Патент на полезную модель


РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ПОЛЕЗНУЮ МОДЕЛЬ

№ 103414

СИСТЕМА ВЗАИМОДЕЙСТВИЯ РАЗДЕЛЕННЫХ БАЗ ПЕРСОНАЛЬНЫХ ДАННЫХ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Патентообладатель(ли): *ООО "Стратегия безопасности" (RU)*

Автор(ы): *Иванов Денис Николаевич (RU), Мищенко Евгений Юрьевич (RU)*

Заявка № 2010149391

Приоритет полезной модели 02 декабря 2010 г.

Зарегистрировано в Государственном реестре полезных моделей Российской Федерации *10 апреля 2011 г.*

Срок действия патента истекает 02 декабря 2020 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов

Рис. 62. Патент на полезную модель

Приложение В

Техническое задание на модернизацию структуры данных и программного обеспечения базы льготного лекарственного обеспечения

Частное техническое задание на модернизацию ведомственной автоматизированной информационной системы «Льготная аптека»

Челябинск, 2016 г.

Рис. 63. Частное техническое задание на модернизацию ПО в ИСПДн «Льготная аптека». Титул

СОГЛАСОВАНО:

Директор департамента информационных технологий — О.В. Руднева

Директор ООО «Стратегия безопасности» — Д.Н. Иванов

Заместитель директора ООО «Стратегия безопасности» — Е.Ю. Мищенко

Старший инженер-программист отдела сопровождения информационных систем — П.В. Калинин

Ведущий специалист по защите информации отдела системного администрирования — А.С. Пономарев

Директор департамента по обеспечению государственных программ, лицензионных требований и качества — И.И. Юсько

Рис. 64. Частное техническое задание на модернизацию ПО в ИСПДн «Льготная аптека».

Лист согласований

5 Состав и содержание работ по модернизации Системы

№ п/п	Наименование изменения
1	Исключить из таблицы Doctor фамилию, имя, отчество врача, заменив их на код врача. В таблице должны отображаться код ЛПУ, код врача. Изменение должны быть внесено в процедуру синхронизации БД серверной части Системы и БД на стороне аптек/аптечных пунктов.
2	Исключить из таблицы DrugStore фамилию, имя, отчество заведующего аптекой/аптечного пункта. Изменение должны быть внесено в процедуру синхронизации БД серверной части Системы и БД на стороне аптек/аптечных пунктов.
3	Исключить из таблицы Invoice фамилию, имя, отчество менеджера заказа, льготника. В качестве замены фамилии льготника в поле «комментарии к заказу» может использоваться RL код для регионального льготника и СНИЛС для федерального.
4	Исключить из таблицы People фамилию, имя, отчество льготника. Процедура поиска лекарственного средства в БД на стороне аптек/аптечных должна осуществляться по RL коду для регионального льготника, либо СНИЛС для федерального льготника. Изменение должны быть внесено в процедуру синхронизации БД серверной части Системы и БД на стороне аптек/аптечных пунктов.
5	Заблокировать возможность ввода нового льготника в аптеках/аптечных пунктах. Изменение должны быть внесено в процедуру синхронизации БД серверной части Системы и БД на стороне аптек/аптечных пунктов.

Рис. 65. Частное техническое задание на модернизацию ПО в ИСПДн «Льготная аптека». Состав работ

АКЦИОНЕРНОЕ ОБЩЕСТВО «Областной аптечный склад»
(АО «ОАС»)
Северо-Крымская ул., д. 20, офис 210, г. Челябинск,
Челябинская область, 454106
ИНН/КПП 7451344670/744801001, ОГРН 1127451015592
тел./факс (351) 268-93-57/(351) 268-93-58, e-mail: oac@oac74.ru

«16» октября 2017г.

Акт

о внедрении результатов диссертационного исследования

Настоящий акт подтверждает, что в период 2016-2017 гг. в систему защиты персональных данных АО «Областной аптечный склад» внедрен метод обезличивания персональных данных, разработанный Мищенко Евгением Юрьевичем в рамках диссертационного исследования «Моделирование процессов обезличивания персональных данных и оценка эффективности используемых методов на основе модели нарушителя».

Результаты применения указанного метода использовались при модернизации системы защиты персональных данных «Льготная аптека», обрабатываемых в 103 подразделениях организации (аптечных пунктах), расположенных в г. Челябинске и Челябинской области. На основании экспертиз и расчетов, произведенных Е.Ю. Мищенко, было разработано «Частное техническое задание на модернизацию ведомственной автоматизированной информационной системы «Льготная аптека», в соответствии с которым была модернизирована структура хранения данных и внесены изменения в процесс их обработки. Использование метода обезличивания персональных данных обусловлено Приказом Роскомнадзора №996 от 05.09.2013 г. «Об утверждении требований и методов по обезличиванию персональных данных».

В качестве преимуществ метода, предложенного Е.Ю. Мищенко, следует отметить следующие:

– характеризуется высокой точностью оценки уязвимости персональных данных;

– обладает высокой эффективностью защиты персональных данных;

– не требует значительных финансовых и временных затрат.

Использование метода позволило сократить расходы на создание системы защиты персональных данных «Льготная аптека» по сравнению с применением стандартных средств и мер защиты информации более, чем в 15 раз. Экономический эффект от внедрения составил более 1,5 млн руб.

Директор департамента ИТ
АО «Областной аптечный склад» Матвеев П.В.

Рис. 66. Акт внедрения полезной модели в ИСПДн «Льготная аптека»

Приложение Г

Пример выполнения алгоритма поиска смещений при перемешивании символов внутри строки

Первая строка идентификаторов (73 символа), известная нарушителю:

1)«мищенко_____евгений___юрьевич_____салавата_юлаева_____29___61_».

Из 39 значимых символов есть следующие повторения: «и» – 3 раза, «е» – 5 раз, «н» – 2 раза, «в» – 4 раза, «ю» – 2 раза, «а» – 6 раз, «л» – 2 раза. Итого потенциальных ошибок – 24. Та же строка, найденная в обезличенной базе по прочим данным:

1)«юлаева_____29___61_салавата_евгений___мищенко_____юрьевич_____».

То есть вектор смещений, используемый в базе и не известный нарушителю, имеет следующий вид (цифра означает позицию, куда сместился символ с этого места):

«44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,34,35,36,37,38,39,40,41,42,43,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,25,26,27,28,29,30,31,32,33,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24»

В результате посимвольного сравнения первой записи у нарушителя определился следующий вектор:

1)«44,39,46,4,38,49,50,_,_,_,_,_,_,_,_,_,34,5,36,37,48,45,40,_,_,_,1,60,61,47,29,64,65,_,_,_,_,_,_,_,_,_,25,3,27,6,35,26,31,28,_,59,2,30,62,63,32,_,_,_,_,_,_,_,_,_,_,17,18,_,_,_,22,23,_»

Итого из 39 значимых символов 15 смещений определено однозначно (голубой цвет) и 24 неоднозначно (красный и зеленый). Необходимо использовать вторую запись. Вторая строка идентификаторов, известная нарушителю:

2)«синянская_____римма_____ивановна_____набережная_____14___84_».

Из 36 значимых символов есть следующие повторения: «и» – 3 раза, «е» – 2 раза, «н» – 6 раз, «в» – 2 раза, «с» – 2 раза, «а» – 6 раз, «я» – 2 раза, «р» – 2 раза, «м» – 2 раза. Итого потенциальных ошибок – 27. Та же строка, найденная в обезличенной базе по прочим данным:

2)«я_____14___84_набережнаримма_____синянская_____ивановна_____».

В результате посимвольного сравнения второй записи у нарушителя определился следующий вектор:

2)«44,35,46,1,32,49,50,26,47,_,_,_,_,_,_,_,29,45,36,37,33,_,40,_,_,_,59,60,61,48,63,64,65,38,_,_,_,_,_,_,_,25,51,27,28,34,30,31,62,66,52,2,_,_,_,_,_,_,_,_,_,_,_,_,_,_,17,18,_,_,_,22,23,_»

Итого из 36 значимых символов 9 смещений определено однозначно и 29 неоднозначно. При совмещении векторов первой и второй записей во втором векторе дополнительно определилось 11 символов, в первом дополнительно определилось 5 символов (зеленый цвет), в результате в первом осталось 19 неоднозначностей, во втором – 18. Необходимо использовать третью запись. Третья строка идентификаторов, известная нарушителю:

3)«малышев_____владимир__сергеевич_____бр.кашириных_____228__134».

Из 42 значимых символов есть следующие повторения: «е» – 4 раза, «и» – 5 раз, «в» – 3 раза, «р» – 3 раза, «а» – 3 раза, «л» – 2 раза, «ш» – 2 раза, «м» – 2 раза, «ы» – 2 раза. Итого потенциальных ошибок – 24. Та же строка, найденная в обезличенной базе по прочим данным:

3)«ных_____228__134бр.каширивладимир__малышев_____сергеевич_____»

В результате посимвольного сравнения третьей записи у нарушителя определился следующий вектор:

3)«44,29,46,47,30,49,50,_,_,_,_,_,_,_,_,34,35,36,37,33,39,40,26,_,_,59,60,61,62,63,64,65,38,67,_,_,_,_,_,_,25,32,27,28,45,48,31,41,66,1,2,3,_,_,_,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24»

Итого из 42 значимых символов 18 смещений определено однозначно и 24 неоднозначно. При совмещении векторов третьей и второй записей во втором векторе дополнительно определилось 13 символов (осталось 10 неоднозначностей):

2)«44,45,46,47,48,49,50,26,52,_,_,_,_,_,_,29,35,36,37,33,39,40,_,_,_,59,60,61,62,63,64,65,38,67,_,_,_,_,_,_,25,51,27,28,34,30,31,32,66,1,2,3,_,_,_,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24»

При совмещении векторов третьей и первой записей в первом векторе дополнительно определилось 13 символов (осталось 10 неоднозначностей):

1)«44,45,46,47,38,49,50,_,_,_,_,_,_,_,_,34,35,36,37,48,39,40,_,_,_,59,60,61,62,63,64,65,_,_,_,_,_,_,_,_,25,3,27,28,35,26,31,28,_,1,2,3,62,63,32,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24».

При совмещении векторов с третьей в третьем дополнительно определилось 7 символов (осталось 4 ошибки):

3)«44,45,46,47,48,49,50,_,_,_,_,_,_,_,_,34,35,36,37,38,39,40,26,_,_,59,60,61,62,63,64,65,33,67,_,_,_,_,_,_,25,32,27,28,29,30,31,32,66,1,2,3,_,_,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24»

При обратном совмещении векторов третьей и второй записей во втором дополнительно определилось 6 символов (осталось 4 ошибки):

2)«44,45,46,47,48,49,50,26,52,_,_,_,_,_,_,34,35,36,37,38,_,40,_,_,_,59,60,61,62,63,64,65,33,67,_,_,_,_,_,_,25,51,27,28,29,30,31,32,66,1,2,3,_,_,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24»

При обратном совмещении векторов третьей и первой записей в первом дополнительно определилось 8 символов (осталось 2 ошибки):

1)«44,45,46,47,48,49,50,_,_,_,_,_,_,_,_,34,35,36,37,38,39,40,_,_,_,59,60,61,62,63,64,65,_,_,_,_,_,_,_,_,25,26,27,28,29,30,31,32,_,1,2,3,4,5,6,_,_,_,_,_,_,_,_,_,_,17,18,19,_,_,22,23,24».