Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени
первого Президента России Б.Н. Ельцина»

На правах рукописи

Мищенко Евгений Юрьевич

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

2.3.6. Методы и системы защиты информации,
информационная безопасность

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Екатеринбург - 2023

Работа выполнена на кафедре защиты информации Федерального государственного автономного образовательного учреждения высшего образования «Южно-Уральский государственный университет (национальный исследовательский университет)»

Научный руководитель: кандидат технических наук, доцент, СОКОЛОВ Александр Николаевич

Официальные оппоненты: ЗАХАРОВ Александр Анатольевич, доктор технических наук, профессор, ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», г. Екатеринбург, профессор Учебно-научного центра «Информационная безопасность» Института радиоэлектроники и информационных технологий;

БАРАНКОВА Инна Ильинична, доктор технических наук, доцент, ФГБОУ ВО «Магнитогорский государственный технический университет им. Г.И. Носова», г. Магнитогорск, заведующая кафедрой информатики и информационной безопасности;

СПЕВАКОВ Александр Геннадьевич, кандидат технических наук, доцент, ФГАОУ ВО «Московский политехнический университет», г. Москва, доцент кафедры информационной безопасности

Защита диссертации состоится «21» февраля 2023 г. в 11:00 часов на заседании диссертационного совета УрФУ 2.3.12.13 по адресу: 620002, г. Екатеринбург, ул. Мира, 19, ауд. И-420 (зал Ученого совета).

С диссертацией можно ознакомиться в библиотеке и на сайте ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»: https://dissovet2.urfu.ru/mod/data/view.php?d=12&rid=4196

Автореферат разослан «____» _____ 202_ года.

Ученый секретарь
диссертационного совета  Сафиуллин Николай Тахирович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования.

Применение обезличивания персональных данных (ПД) обусловлено необходимостью обработки, хранения и передачи ПД в научных, статистических и прочих целях в форме представления, не допускающей возможности нанесения ущерба физическому лицу (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» (далее – Приказ Роскомнадзора), 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 в информационных системах персональных данных, 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 объема 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 (Y. He, J.F. Naughton) и оценке эффективности обезличивания (P. Kiran, N.P. Kavya).

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

После ввода в действие Приказа Роскомнадзора метод введения идентификаторов рассмотрен в работах А.А. Халафяна и А.А. Кошкарова (разработка, реализация и внедрение, 2015 г.), А.А. Ноздриной и Д.В. Применко (разработка и реализация, 2016 г.). 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.), К.О. Бондаренко, В.А. Козлова (разработка и реализация, 2015 г.), Е.А. Макаровой, Д.Г. Лагерева (разработка и реализация, 2016 г.).

С момента ввода в действие терминологии по обезличиванию ПД Приказом Роскомнадзора стали возможными строгое теоретическое обоснование и оценка эффективности алгоритмов обезличивания ПД, но эти этапы разработки практически не нашли отражения в работах исследователей. Исключение составляют работы И.П. Карповой (методика оценки эффективности обезличивания по методу перемешивания, 2013 г.).

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 позволило 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 и решены 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.

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 (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 (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 (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 исследований, полученные на основе функционального и математического моделирования с применением методов теории вероятностей и математической статистки. Модель идентификации ФЛ разработана на основе частотного метода статистического анализа. Для подтверждения гипотез о характере зависимостей свойств атрибутов ФЛ от их семантики и 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 [1 – 3, 8, 9].

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 [4 – 5].

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 [6 – 7].

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.) [10]; XVI Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2017» (Екатеринбург, 2017 г.) [11]; 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.) [12]; XVIII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2019» (Магнитогорск, 2019 г.) [13]; 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.) [6].

Публикации.

Основные результаты исследования опубликованы в 13 научных работах, 6 из которых – в рецензируемых научных журналах, определенных ВАК РФ и Аттестационным советом УрФУ, включая 1 статью в издании, индексируемом в международной цитатно-аналитической базе Scopus. Имеется 1 патент на полезную модель.

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.

Содержание работы

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, параметры методов 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. Предусмотренные Российскими нормативными актами методы обратимого обезличивания ПД 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 и решена задача создания модели идентификации ФЛ для количественной оценки вероятности идентификации ФЛ по каждому атрибуту и по некоторым совокупностям атрибутов. Для обезличивания предложено использовать атрибуты с наибольшими значениями вероятности идентификации. Поскольку состав обезличиваемых атрибутов зависит и от количества записей базы ПД, по результатам расчета вероятности идентификации эта зависимость определена для каждого атрибута и для некоторых совокупностей атрибутов.

Для первой группы методов основная проблема реализации – оценка эффективности алгоритма искажения и его защищенности. Для решения этой проблемы сформулирована и решена задача разработки модели нарушителя с целью формирования таблицы смещений элементов идентификаторов с использованием достаточного для неавтоматизированной обработки количества записей ФЛ, известных нарушителю и заведомо входящих в обезличенную базу. 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 действий нарушителя, заложенным в модель.

Для второй группы методов основная проблема их реализации – обеспечение защищенной передачи информации между разделенными частями БД. Для решения этой проблемы сформулирована и решена задача реализации функциональной схемы для обезличивания базы ПД методом введения идентификаторов и обеспечения связи между удаленной таблицей перекрестных ссылок и базой обезличенных данных без добавления в процесс обработки данных дополнительных внешних носителей и СКЗИ.

Во второй главе сформулирован критерий необходимости обезличивания ПД ФЛ по атрибутам, основанный на применении модели идентификации ФЛ и модели нарушителя безопасности ПД.

Для решения задачи разработки модели идентификации ФЛ применены методы частотного анализа и параметрической идентификации к эмпирическим последовательностям характеристик атрибутов ФЛ и их сочетаний, с учетом различных объемов БД. В рамках диссертационного исследования разработаны модели идентификации для двух баз ПД объемом 310 тысяч и 330 тысяч записей ФЛ для следующего набора атрибутов: фамилия, имя, отчество, дата рождения, наименование улицы проживания, номер дома, номер квартиры. Кроме этого, построены модели для совокупности атрибутов «имя – отчество», а также для

инициалов (первая буква имени и первая буква отчества). Вероятность идентификации ФЛ по какому-либо атрибуту предложено оценивать по характеристикам функции распределения его значений, которые, в свою очередь, определяются свойствами самого атрибута, например, его семантикой и диапазоном значений.

В качестве случайной величины рассмотрено значение q_j – количество записей, найденных нарушителем в БД, содержащих любое искомое значение j-го атрибута, при этом 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 базе.

Если ПД конкретного ФЛ содержатся в записи $L_{ФЛ}$, то вероятность идентификации ФЛ по значению A_{jk} определяется как:

$$p(A_{jk} \in L_{ФЛ}) = 1/q_{jk}.$$

Если в БД существует n_{jk} групп записей j-го атрибута, по q_{jk} записей в каждой (например, 2 группы по 100 фамилий «Иванов» и «Петров»), то вероятность получения такого количества записей при поиске определяется по формуле:

$$p(q_j = q_{jk}) = n_{jk}/Q_j,$$

а вероятность идентификации по j-му атрибуту в целом определяется как

$$W_j = Q_j / V, \qquad (1)$$

где V – количество записей БД (объем базы).

Результат поиска ПД в обезличенной базе 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 обезличенной БД;
2) нарушитель знает структуру исходной базы ПД;
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);
5) нарушитель знает некоторые ПД искомого ФЛ, хочет узнать другие идентификаторы и прочие ПД этого лица;
6) нарушитель знает, что в БД есть полные ПД определенного количества известных ему ФЛ, при этом прочие (неискаженные) данные он может найти;
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;
8) нарушитель не может однозначно выбрать искомое 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.

Таким образом, параметрами модели идентификации являются:

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

– количество найденных нарушителем записей U, целесообразное для дальнейшей обработки («компетенция нарушителя», п.8.);

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

Предложенный подход позволил сформулировать критерий необходимости обезличивания по атрибуту ФЛ:

обезличивание ФЛ по j-му атрибуту 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:

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

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

$W_{норм}$ – допустимое значение вероятности идентификации ФЛ по любому атрибуту (нормативное значение, заданное в перспективе регулирующим органом);

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

U – показатель компетенции нарушителя (нормативное значение, заданное в перспективе регулирующим органом);

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

$$U > q_j,$$

где q_j – количество записей, найденных нарушителем в БД, содержащих искомое значение j-го атрибута.

Таким образом, критерий необходимости обезличивания по j-му атрибуту применяется по следующему алгоритму:

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

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

2) определить количество различных значений атрибута Q_j;

3) вычислить W_j – вероятность идентификации ФЛ по j-му атрибуту (1);

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, как решение уравнения

$$W_j = \int_{q_{jW}}^{\infty} p_j(x)dx, \quad (2)$$

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 объема базы данных V;

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

11) произвести действия (1) – (10) для рассмотренных выше атрибутов для БД с другим количеством записей V_b, где b – номер базы/части базы с количеством записей, отличным от V;

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

Точность определения значения $W_{\text{норм}}$ зависит от точности аппроксимации величины q_j. В диссертационном исследовании выдвинуты и проверены по критериям согласия гипотезы о степенных и гамма- зависимостях функции плотности вероятности случайной величины, аппроксимирующей q_j.

Для атрибутов текстовой семантики с неограниченным набором значений (например, «фамилия», «имя» и т.п.) в качестве аппроксимирующих рассмотрены 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 (например, «номер дома», «дата рождения») с возможным наличием максимума дополнительно рассмотрена гамма-функция вида $y(x) = ax^b e^{-cx}$.

Для проверки согласия экспериментальной и теоретической зависимостей для нескольких коэффициентов экспоненциальной и степенной функции $y(x)$ использованы три критерия согласия первого рода:

1) критерий хи-квадрат χ^2;
2) критерий отношения правдоподобия S_{on};
3) критерий Колмогорова $T(q_{jk})$.

Все три критерия проверялись при заданном уровне значимости $\alpha = 0{,}05$. Далее для устранения ошибок второго рода все функции, прошедшие проверку первого рода, подвергались проверке по методу распознавания зависимостей на основе обратного отображения А.Н. Тырсина. В результате сравнения выбрана функция $y(x)$ с минимальной дисперсией ошибок s^2, вычисляемой как

$$s^2 = \sum_{k=1}^{dj} (Y(q_{jk}) - q_{jk})^2, \quad (3)$$

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

Для атрибута A_1 «фамилия» (V = 310132 записей ФЛ) определено количество различных значений Q_1 = 45099, для каждого из которых определено количество записей q_{1k} в диапазоне от $q_{1мин}$ = 1 до $q_{1макс}$ = 1892.

По формуле (1) вычислено значение W_1 = 0,1454.

Произведена аппроксимация дискретного распределения случайной величины q_1 для атрибута A_1 БД 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 n_1 фамилий в интервале.

Рассчитаны значения критериев согласия χ^2, отношения правдоподобия и Колмогорова для функций f_1 и f_2 при заданном уровне значимости $\alpha = 0,05$. Расчеты показали, что экспоненциальная функция f_1 не отвечает ни одному критерию согласия, а степенная функция f_2 отвечает всем трем.

На рис. 1 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 ПД.

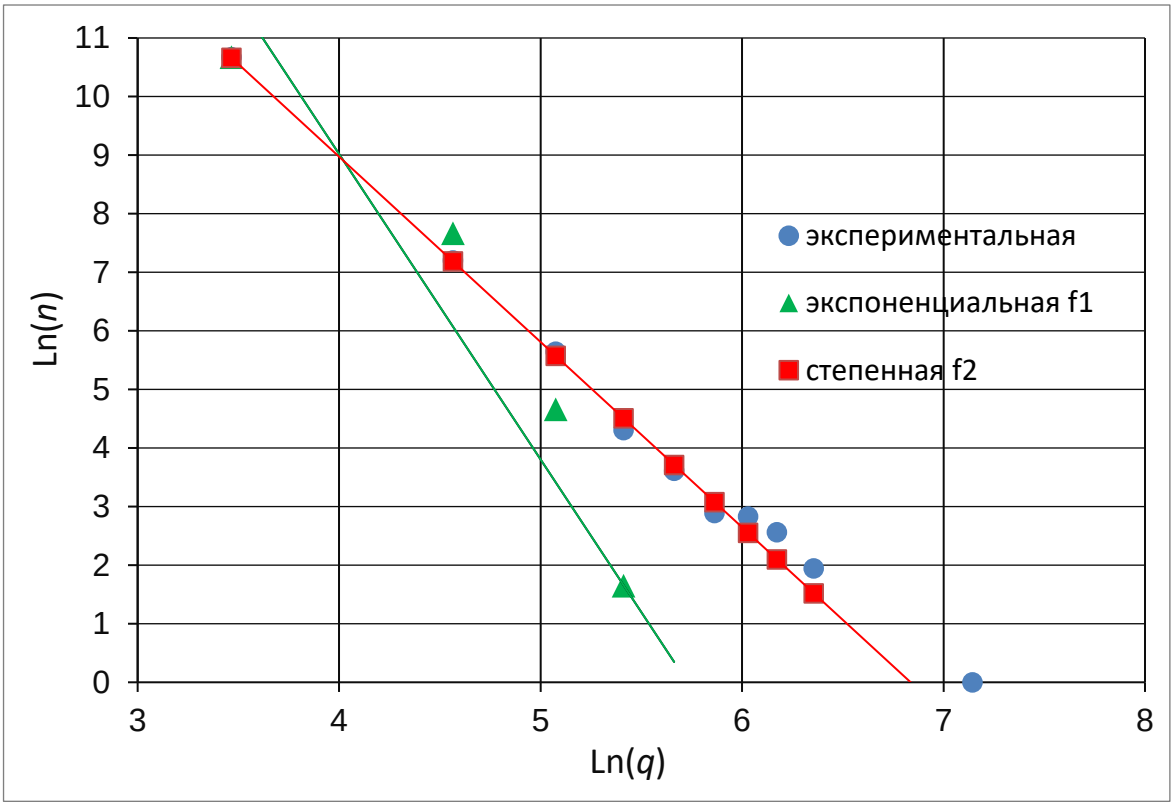

Рис. 1. Функции f_1 и f_2 в сравнении с экспериментальной дискретной последовательностью для атрибута A_1, где q – количество записей, имеющих одинаковое значение атрибута, n – количество различных значений атрибута

В результате проверки по критериям первого рода экспоненциальный вид аппроксимирующих функций отвергнут, в итоге принят степенной вид из-за отсутствия альтернативы для сравнения по критерию согласия второго рода.

В результате решения уравнения (2) получены значения q_{1W} и $q_{1норм}$. На рис. 2 приведен график функции распределения вероятности $F(x)$ и показаны значения W_1, $W_{норм}$ и U.

Из рис. 2 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 ($W_1 > W_{норм}$);

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

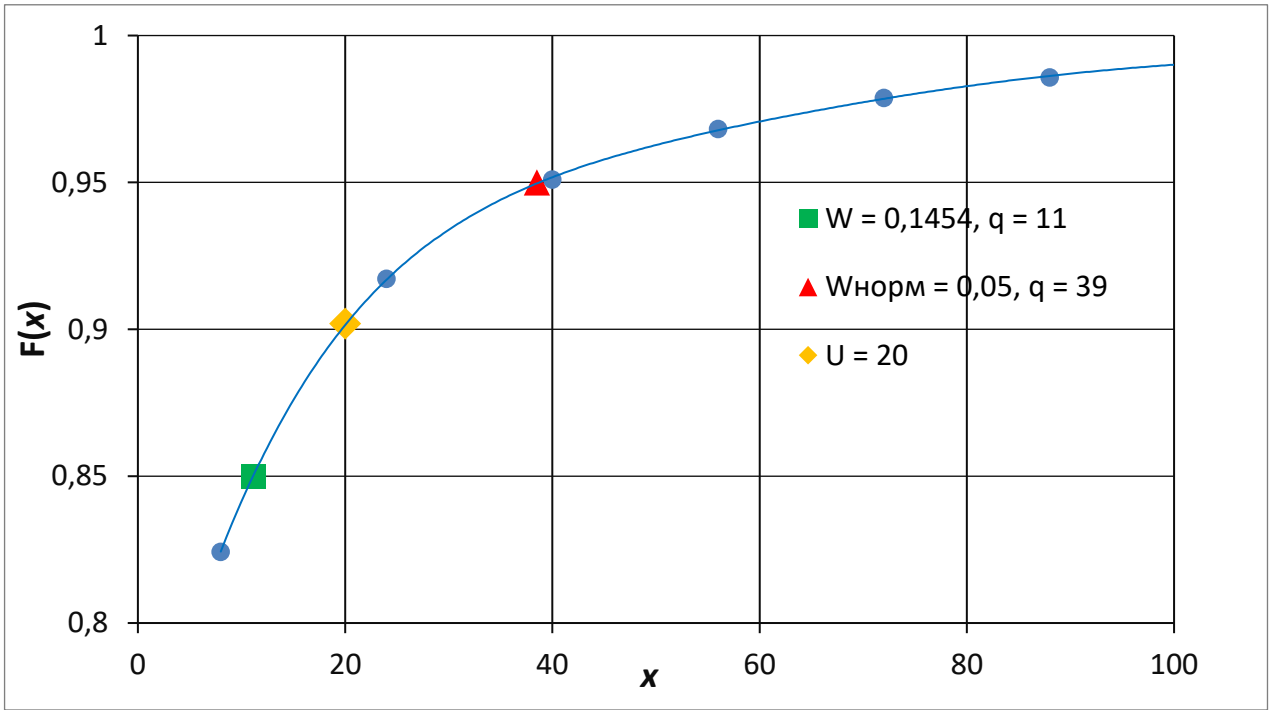

Рис. 2. Распределение вероятности для атрибута «фамилия» в сравнении с W_1, $W_{норм}$ и U, где x – количество записей, имеющих одинаковое значение атрибута, $F(x)$ – функция распределения вероятности

Для определения зависимости вероятности идентификации от количества записей БД для атрибута A_1 была построена диаграмма частот W_1 в диапазоне V = [1666, 310132] записей. На рис. 3 приведена диаграмма частот значений W_1 в зависимости от V в обычном (рис. 3a) и логарифмическом (рис. 3b) масштабе координат. На рис. 3b заметен линейный характер функции на большей части диапазона.

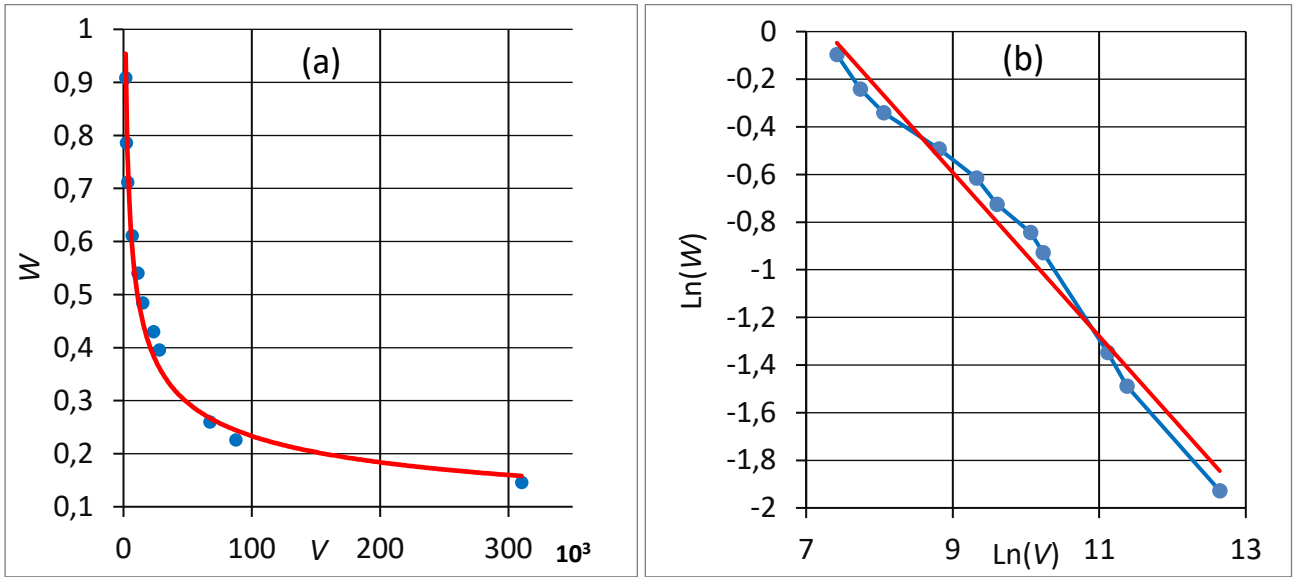

Рис. 3. Диаграмма зависимости вероятности идентификации по атрибуту «фамилия» в обычном (a) и логарифмическом (b) масштабе координат от объема БД, где красным обозначена линия тренда, V – количество записей БД, W – вероятность идентификации

На рис. 4 в логарифмическом масштабе по обеим координатам приведены графики функций $f_1(x) = 0{,}6632 \cdot e^{-0{,}000007x}$, $f_2(x) = 12{,}2435 \cdot x^{-0{,}344}$ и $f_3(x) = -0{,}136 \cdot \ln(x) + 1{,}8597$ в сравнении с дискретной последовательностью экспериментальных значений для критерия W_1 атрибута A_1, где x – количество записей базы V.

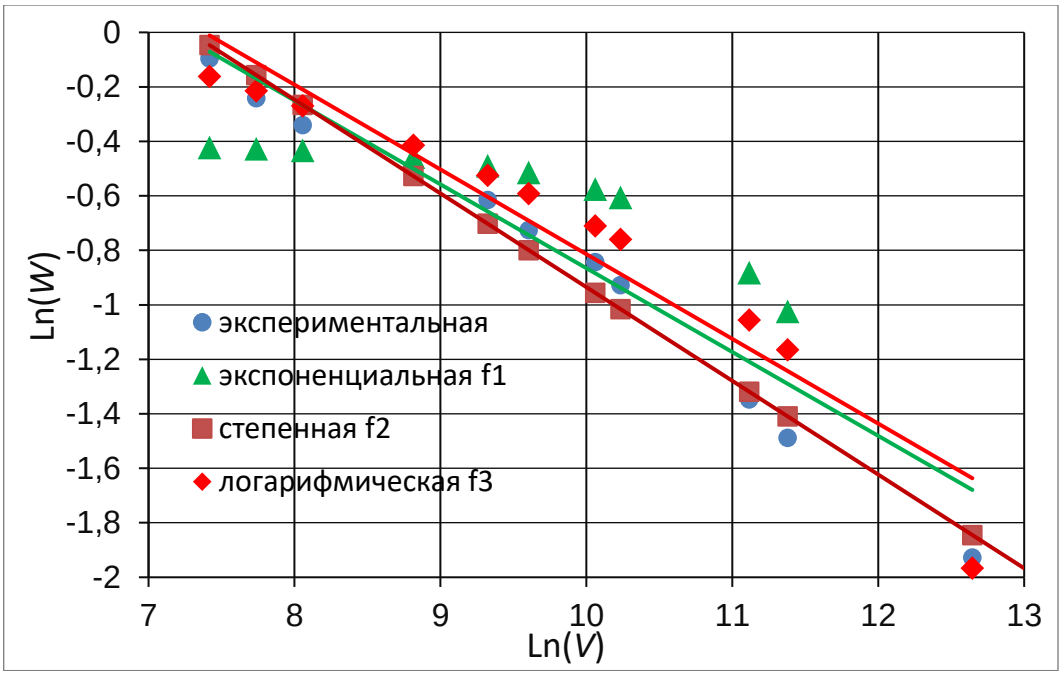

Рис. 4. Функции f_1, f_2 и f_3 в сравнении с экспериментальной дискретной последовательностью для W_1 атрибута A_1 базы B_1, где V – количество записей БД, W – вероятность идентификации

В результате проверки по критериям первого рода при заданном уровне значимости $\alpha = 0{,}05$ все три вида функций не были отвергнуты. Проверка по критерию второго рода (3) показала, что значение дисперсии s_2^2 для степенной функции f_2 является минимальным. В результате для аппроксимации была выбрана степенная зависимость $W_1(V)$.

В третьей главе предложена модель нарушителя, реализующая алгоритм деобезличивания, и методика оценки эффективности обезличивания для методов обезличивания, основанных на искажении ПД.

Для решения задачи разработки модели нарушителя применены методы комбинаторного анализа в алгоритмах деобезличивания. В качестве алгоритмов искажения рассмотрены следующие алгоритмы:

– перестановка символов внутри полной строки идентификаторов;

– перестановка битов внутри строки идентификатора;

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

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

Общая для различных методов модель нарушителя, предложенная в главе 2, 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 выполнена количественная оценка возможности достижения заданного результата U при заданных параметрах G и алгоритма деобезличивания.

Для обеспечения достоверности модели использована база ПД, содержащая 310 тыс. записей ФЛ, где в качестве идентификаторов использованы 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 обозначены как A_z.

В качестве заданных приняты следующие условия:

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

– искажению подвергаются атрибуты $A_1 – A_6$;

– максимальное количество известных нарушителю записей в базе $G = 5$;

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

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 ФЛ, при этом одинаковые прочие данные могут принадлежать нескольким ФЛ:

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

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 ФЛ. Выбор производится по составу символов искаженной строки или составу полей в сегменте из 511 записей (по 255 записи в обе стороны от найденной). Если записей

будет более 1 (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 известным ему символам/атрибутам 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 ФЛ 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 ФЛ в БД, при использовании части или всех известных ему записей, модель нарушителя считается полностью эффективной;

6) если таблица смещений, полученная нарушителем с использованием всех известных ему записей, позволяет получить данные о любом ФЛ в БД в виде набора записей не более чем из 20 ФЛ, то модель нарушителя считается эффективной условно;

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

При перестановке символов внутри полной строки идентификаторов определено наиболее вероятное количество повторений различных символов во всех идентификаторах БД. Для этого произведен анализ частотного распределения символов, что позволило определить вероятность наличия в строке конкретного символа и суммарное количество повторений символов в общей строке идентификаторов.

В строке идентификаторов A_1 – A_6 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.

Для оценки возможности деобезличивания при использовании нарушителем второй записи учитывалось, что:

− полное несовпадение неоднозначных мест 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;

− если неоднозначное место одного символа в первой строке совпадет с неоднозначным местом другого символа во второй строке, то количество

неоднозначностей сохраняется (все повторяющиеся символы условно считаются одним неоднозначным символом).

Процесс сравнения для двух записей проиллюстрирован на рис. 5, где звездочками указаны точно установленные места, а стрелками указаны сохранившиеся неоднозначные места символов. При сравнении двух записей неоднозначности в местоположении некоторых символов были исключены (они стрелками не помечены).

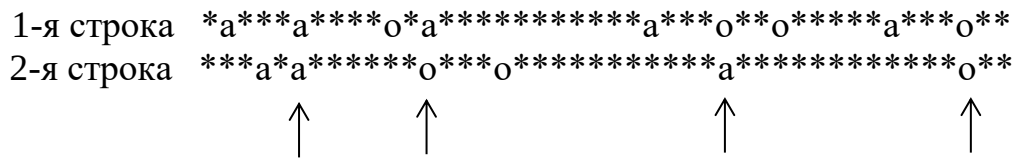

Рис. 5. Варианты неоднозначного определения смещения символов

Из рис. 5 видно, что при использовании первой строки не удалось однозначно установить 9 положений для двух букв. При применении второй строки независимо от первой, количество неоднозначных мест – 6, но если строки применить последовательно, то количество неоднозначных вариантов снижается до 4! = 24.

Алгоритм поиска смещений символов в строке искаженной базы для количества известных нарушителю записей $G = 3$ приведен на рис. 6.

Для расчета наиболее вероятного количества совпадений неоднозначностей в двух записях была применена формула:

$$C_0 = M[F_v(c)],$$

где M – математическое ожидание;

$F_v(c)$ – функция зависимости количества вариантов совпадений от количества совпадений c.

Функция $F_v(c)$ является дискретной и представляет собой произведение количества различных вариантов расположения c неоднозначных символов в множестве символов строки идентификаторов N на количество различных вариантов расположения остальных $(m - c)$ неоднозначных символов в оставшейся части множества символов строки идентификаторов $(N - c)$:

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

где m – количество неоднозначных мест в одной строке;

$R_{ab} = a! / b!(a - b)!$ – количество различных вариантов расположения b символов в строке длиной a.

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.

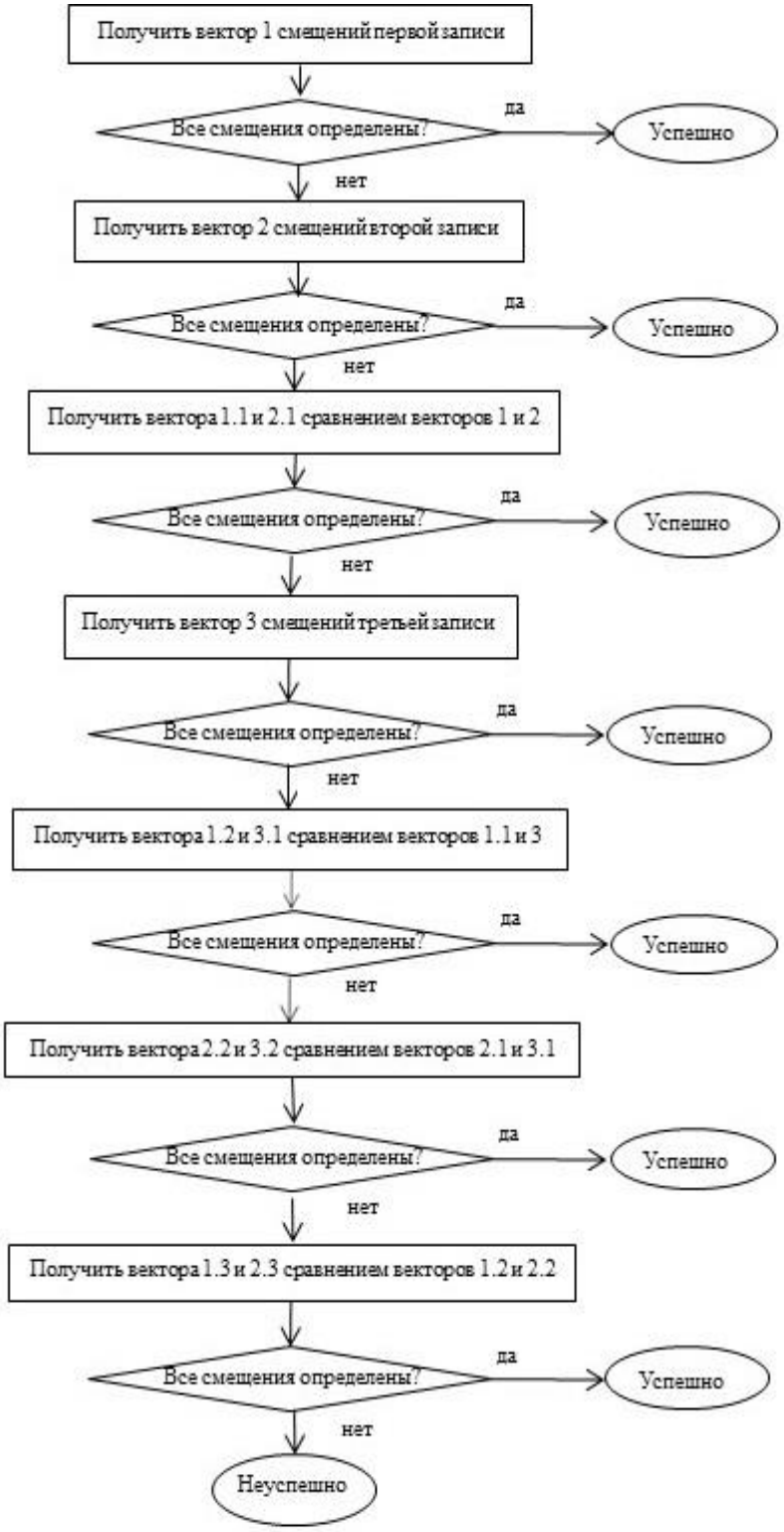

Рис. 6. Алгоритм поиска нарушителем смещений символов в строке для $G = 3$

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

По результатам расчетов построена диаграмма зависимости количества вариантов совпадений F_v от количества совпадений c, которая приведена на рис. 7.

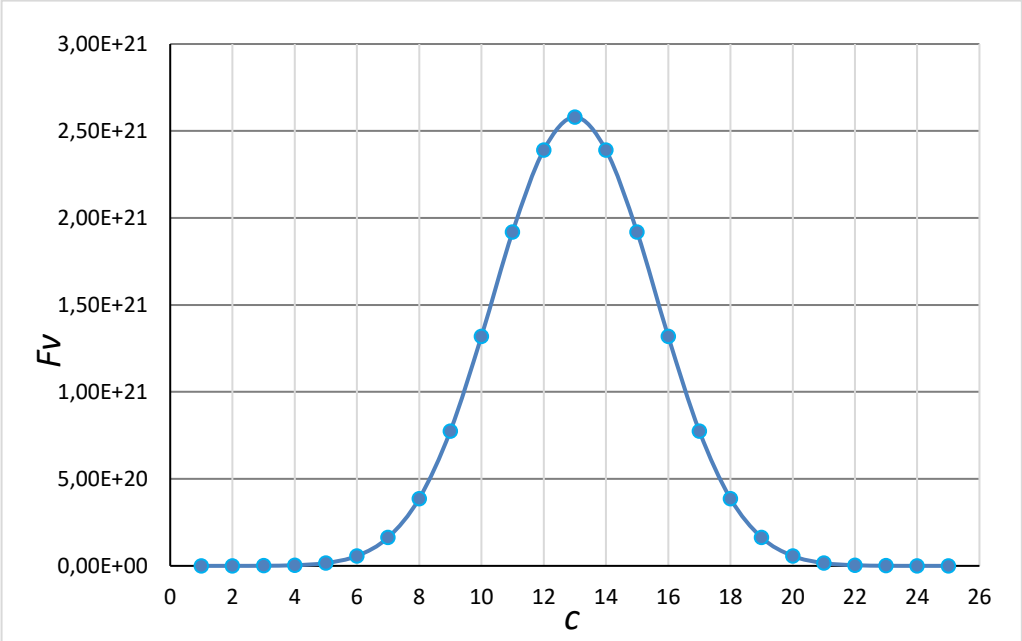

Рис. 7. Диаграмма зависимости количества вариантов совпадений F_v от количества совпадений c, где $N = 73$, $m = 26$

Из рис. 7 видно, что $F_v(c)$ имеет максимум при $C_0 = 13$, которое и является математическим ожиданием, а 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 нарушителем неизбежно. Поскольку ошибочные символы неодинаковые, их повторяемость уменьшается пропорционально их частотности и при сужении диапазона становится меньше двух (для символов «к», «л», «р»). Расчеты показали, что даже без учета сочетаемости символов использование третьей известной записи снизит количество совпадений до 4 (количество вариантов – до 24), то есть приведет к решению задачи с точки зрения количества записей U в модели нарушителя.

Аналогичные расчеты произведены для алгоритмов, реализующих перестановку битов внутри строки идентификатора, перемешивание полей внутри группы из 256 записей базы, перемешивание символов внутри группы из 256 записей БД. В результате расчетов 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.

Разработанная методика оценки эффективности опирается на количество итераций в работе алгоритмов деобезличивания и позволяет сформулировать рекомендации для нормативных значений вероятности идентификации для обезличиваемых идентификаторов. Процесс деобезличивания признается эффективным, а примененный метод обезличивания, напротив, неэффективным,

если нарушителю удалось раскрыть алгоритм обезличивания за количество циклов деобезличивания, меньшее или равное *G*.

В четвертой главе предложена функциональная схема передачи информации между частями базы ПД, разделенными методом введения идентификаторов.

Для решения задачи разработки функциональной модели передачи информации между обезличенной частью ПД и таблицей перекрестных ссылок применен внешний бумажный носитель, содержащий идентификатор ФЛ в виде штрих-кода, как это показано на рис. 8. В основу схемы положена полезная модель, разработанная Д.Н. Ивановым и автором представленного исследования, на которую получен патент [7] в 2010 году.

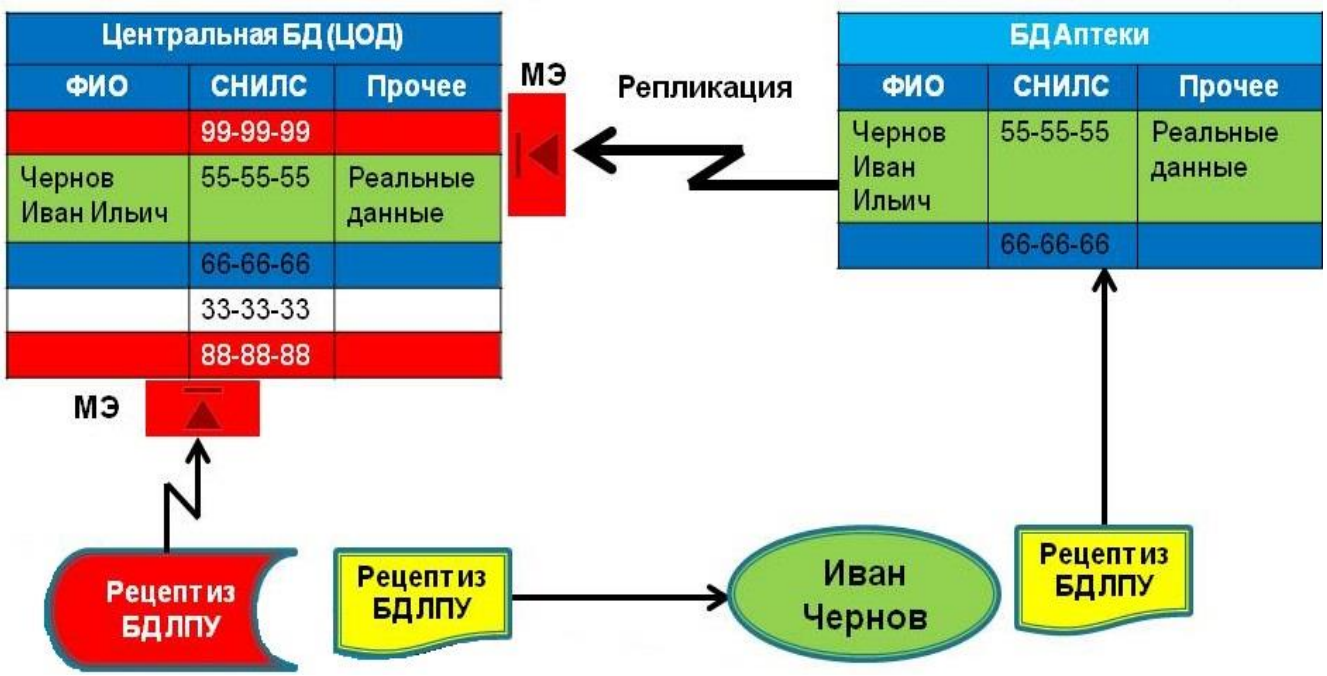

Рис. 8. Схема передачи информации между разделенными базами данных после модернизации

Представленная схема внедрена на предприятиях сферы здравоохранения в 2017 году.

В качестве объекта для внедрения выбрана база ПД 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 ПД – около 329 тысяч записей ФЛ, а минимальный объем базы аптечного пункта – 4 тыс. записей. По результатам внедрения оценена возможность успеха нарушителя (найдено менее, чем *U* записей) при заданных параметрах *G* и алгоритма деобезличивания.

В соответствии с моделью нарушителя в качестве граничных значений при расчете показателей вероятности идентификации *W* для всех идентификаторов были приняты следующие:

– максимальное количество записей, полученных в результате деобезличивания, при котором поиск считается эффективным $U = 20$;

– максимальное количество известных нарушителю записей в базе $G = 5$;

– максимальная граница W для атрибутов пациента и врача была определена как 1 (однозначная идентификация);

– минимальная граница W атрибутов пациента для базы в ЦОД была определена как $3.0 \cdot 10^{-6}$, исходя из 329 тысяч записей пациентов;

– минимальная граница W атрибутов пациента для базы в аптечном пункте была определена как $2{,}5 \cdot 10^{-4}$ исходя из 4 тысяч записей пациентов (минимум);

– минимальная граница W атрибутов врача для базы в ЦОД была определена как $4.9 \cdot 10^{-5}$ исходя из 20,5 тысяч записей врачей;

– минимальная граница W атрибутов врача для базы в аптечном пункте была определена как $6{,}7 \cdot 10^{-3}$ исходя из 150 записей врачей (минимум).

На рис. 9 показаны значения W для различных атрибутов, рассчитанные до модернизации (синий цвет) и после нее (красный цвет) для ЦОД (рис. 9a) и для аптечной БД (рис. 9b). Экономический эффект при использовании предложенной схемы составил 1,4 млн. руб.

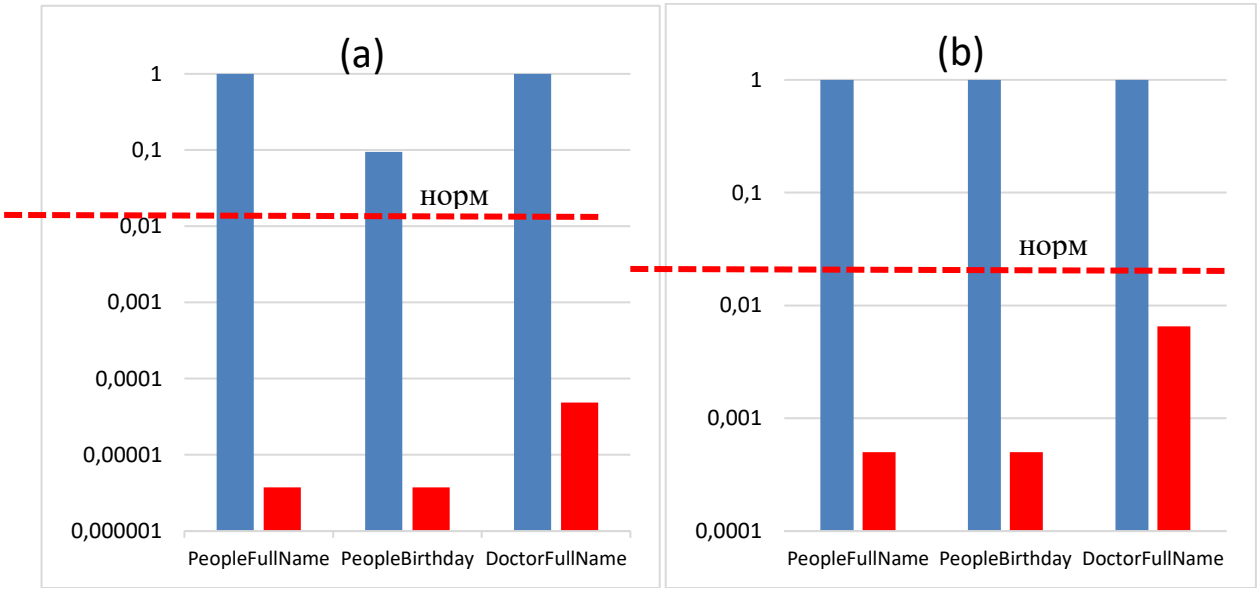

Рис. 9. Вероятность идентификации пациента (a) и врача (b) до (синий цвет) и после внедрения (красный цвет), где линия «норм» соответствует $W = 0{,}05$

Основные результаты исследования:

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 при любых объемах баз ПД.
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.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в рецензируемых научных журналах и изданиях, определенных ВАК РФ и Аттестационным советом УрФУ:

1. **Mishchenko E.Y.** Model of Identification of a Person in Databases of Various Sizes / **E.Y. Mishchenko,** A.N. Sokolov // 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). – 2021. – pp. 0407–0410. (0,44 п.л. / 0,22 п.л.) (Scopus)

2. **Мищенко Е.Ю.** Определение эффективности обезличивания персональных данных с использованием модели нарушителя / **Е.Ю. Мищенко**, А.Н. Соколов // Вестник УрФО. Безопасность в информационной сфере. – 2020. – № 2 (36). – С. 34–42. (0,74 п.л. / 0,37 п.л.)

3. **Мищенко Е.Ю.** Алгоритмы реализации методов обезличивания персональных данных в распределенных информационных системах / **Е.Ю. Мищенко**, А.Н. Соколов // Доклады Томского Государственного Университета Систем Управления и Радиоэлектроники. – 2019. – Т. 22. № 1. – С. 66–70. (0,64 п.л. / 0,32 п.л.)

4. **Мищенко Е.Ю.** Количественный анализ процедуры обезличивания персональных данных. Метод перемешивания / **Е.Ю. Мищенко**, А.Н. Соколов

// Вестник УрФО. Безопасность в информационной сфере. – 2016. – № 3(21). – С. 30–37. (0,76 п.л. / 0,38 п.л.)

5. **Мищенко Е.Ю.** Количественный анализ процедуры обезличивания персональных данных. Метод изменения состава или семантики / **Е.Ю. Мищенко**, А.Н. Соколов // Вестник УрФО. Безопасность в информационной сфере. – 2016. – № 1(19). – С. 30–38. (0,84 п.л. / 0,42 п.л.)

6. **Мищенко Е.Ю.** Количественный анализ процедуры обезличивания персональных данных. Метод введения идентификаторов / **Е.Ю. Мищенко**, А.Н. Соколов // Вестник Южно-Уральского государственного университета. Компьютерные технологии, управление, радиоэлектроника. – 2015. – № 3(15). – С. 18–25. (0,76 п.л. / 0,38 п.л.)

Патент:

7. Иванов Д.Н. Патент 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.

Другие публикации:

8. **Мищенко Е.Ю.** Вероятность идентификации в базе персональных данных: выбор идентифицирующих атрибутов / **Е.Ю. Мищенко** // XVIII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2019». Сборник трудов. – 2019. – С. 206–209. (0,18 п.л. / 0,18 п.л.)

9. **Мищенко Е.Ю.** Модель нарушителя в системах обезличенных персональных данных / **Е.Ю. Мищенко**, А.Н. Соколов // XVII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2018». Сборник трудов. – 2018. – С. 124–128. (0,24 п.л. / 0,12 п.л.)

10. **Мищенко Е.Ю.** Обезличивание персональных данных как способ снижения затрат на создание системы защиты информации / **Е.Ю. Мищенко** // XVI Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых «Безопасность информационного пространства – 2017». Сборник трудов. – 2018. – С. 199–203. (0,24 п.л. / 0,24 п.л.)

11. **Мищенко Е.Ю.** Количественные критерии идентификации физического лица при обезличивании персональных данных / **Е.Ю. Мищенко**, А.Н. Соколов // Вестник УрФО. Безопасность в информационной сфере. – 2014. № 1(11). С. 27–33. (0,54 п.л. / 0,27 п.л.)

12. **Мищенко Е.Ю.** Обезличивание персональных данных / **Е.Ю. Мищенко**, А.Н. Соколов // Актуальные проблемы автоматизации и управления. Труды научно-практической конференции. – 2013. – С.356–359. (0,22 п.л. / 0,11 п.л.)

13. **Мищенко Е.Ю.** Обезличивание персональных данных: термины и определения / **Е.Ю. Мищенко**, А.Н. Соколов // Вестник УрФО. Безопасность в информационной сфере. – 2013. – № 1(7). – С. 10–13. (0,26 п.л. / 0,13 п.л.)