Применение процесса деидентификации для снижения рисков нарушения конфиденциальности

К. Ф. Керимов
З. И. Азизова, email: z.i.azizova@mail.ru

Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий

Аннотация. *В данной статье рассматривается процесс деидентификации в качестве одного из способов сохранения ключевой информации из исходного набора персональных данных, а также его применение в целях минимизации возможных рисков, которые необходимо учитывать в процессе реализации методов обезличивания персональных данных.*

Ключевые слова: *персональные данные, деидентификация, риски повторной идентификации, конфиденциальность*.

Введение

Актуальность и востребованность использования больших массивов общедоступных данных вместе с возрастающим интересом и потребностями в повторном применении этих данных в условиях малой стоимости их хранения и обработки приносят сегодня достаточную пользу как обществу в целом, так и отдельным организационным структурам, при этом особое внимание уделяется соблюдению гарантированной защиты прав каждого гражданина на частную жизнь и персональные данные.

Деидентификация является хорошей стратегией для сохранения полезности при использовании персональных данных и для снижения рисков их компрометации впоследствии их публикации. В случае, когда набор данных подвергся процессу деидентификации, и определение принадлежности отдельных данных конкретному субъекту невозможна закон о защите данных теряет свою силу. В этом случае, создание действительно анонимного набора данных из огромного набора персональных данных, с учетом сохранности только требуемой информации становится сложной задачей.

1. Защита персональных данных

Вопросы обеспечения безопасности информационных ресурсов являются важным элементом функционирования организационной

структуры в современных экономических реалиях, что во многом обусловлено ростом числа атак на информационные системы и хранилища данных. С принятием Закона Республики Узбекистан №ЗРУ-547 «О персональных данных» от 02.07.2019 года многочисленные информационные системы, касающиеся сбора, хранения, обработки или передачи идентификационных данных физических лиц, стали подлежать модернизации в строгом соответствии с совершенно новыми требованиями. Реальность исполнения данного закона на практике в полной мере будет зависеть от создания практических инструментов его реализации и четкой формализации требований к защите частной информации.

Нарушения конфиденциальности персональных данных представляют собой значительную угрозу для непрерывной работы организаций, поэтому представляется целесообразным внедрение соответствующей политики и процедуры, регулирующих вопросы управления безопасностью обрабатываемых персональных данных. Такая необходимость определяется количеством и масштабами последствий нарушений безопасности персональных данных, которые происходят сегодня в глобальном масштабе.

В соответствии с требованиями закона №ЗРУ-547 оператор информационной системы при обработке персональных данных обязан принимать необходимые правовые, организационные и технические меры по защите персональных данных от незаконного или случайного доступа к ним, уничтожения, модификации, копирования и распространения персональных данных, а также иных неправомерных действий. Рост объема обрабатываемых данных создает новые вызовы для организаций в отношении надлежащего управления персональными данными, которыми они обладают, и обеспечения их надлежащей защиты.

Приобретенная и собранная информация, которая может быть проанализирована, представляет собой значительную экономическую ценность. В своей работе Джефф Седаяо неоднократно отмечает, что потенциальную ценность процесса деидентификации можно наблюдать в ее применении в качестве стратегии использования открытых данных отдельными лицами или обществом, при этом фактические риски определения субъектов персональных данных снижаются. Если же стоит цель необратимого предотвращения определения субъекта данных и при этом, не исключается возможность использования нескольких методов одновременно, по той причине, что в законодательстве нет предписывающего стандарта, следование которому было бы обязательным условием, необходимо применять анонимизацию [1].

2. Риски нарушения конфиденциальности

Распространение общедоступной информации в глобальной сети в совокупности с более мощным компьютерным оборудованием позволило вновь идентифицировать обезличенные данные. Это говорит о том, что после деидентификации открытые данные могут быть вновь привязаны к конкретному пользователю к которому они относятся.

Де-идентификация предсавляет собой процесс обнаружения квази-идентификаторов, которые прямо или косвенно идентифицируют субъект (или объект), и удаления этих идентификаторов из существующего набора данных [2].

Реидентификация обезличенных данных представляет собой серьезные последствия нарушения конфиденциальности. По причине того, что нет строгого регулирования обезличенных данных, они могут быть проданы кому угодно и использованы в любых целях. Без регулирования процесса реидентификации все кто может получить доступ к реидентифицированным данным имеют возможность беспрецедентного доступа к ней. Обезличенные данные обычно реидентифицируются путем объединения двух или более наборов данных, для нахождения пользователя в обоих наборах [3]. При этом отсутствует обязательство сообщать, если данные были реидентифицированы.

Нарушением конфиденциальности является факт раскрытия персональной или конфиденциальной информации неавторизованным субъектам. Обезличенный набор данных может быть реидентифицирован следующими способами: недостаточной анонимизацией, изменением псевдонима или комбинированием наборов данных. Эти методы не являются взаимоисключающими, т.е. все три могут использоваться в совокупности для повторной идентификации субъекта в наборе обезличенных данных.

Пользователи информационных систем или социальных сетей зачастую бывают убеждены в обеспечении защиты персональных данных операторами. А сами операторы данных, в свою очередь, сталкиваются с серьезной проблемой сохранения конфиденциальности пользователей при публикации данных личного характера. С целью защиты они обычно обезличивают данные перед их публикацией при возможном использовании сторонними пользователями. С одной стороны, использование псевдонимизированных или деидентифицированных данных может быть очень полезным для исследователей благодаря детализации на индивидуальном уровне и тому, что псевдонимизированные записи из разных источников можно сравнить без особых затруднений. Однако это также означает, что

существует довольно высокий риск повторной идентификации. С другой стороны, агрегированные данные имеют относительно небольшой риск, в зависимости от уровня детализации, размера выборки и так далее. Такие данные могут быть довольно безопасными, поскольку риск повторной идентификации сравнительно невелик [3].

На рис. 1 приведены три категории нарушения конфиденциальности в сети.

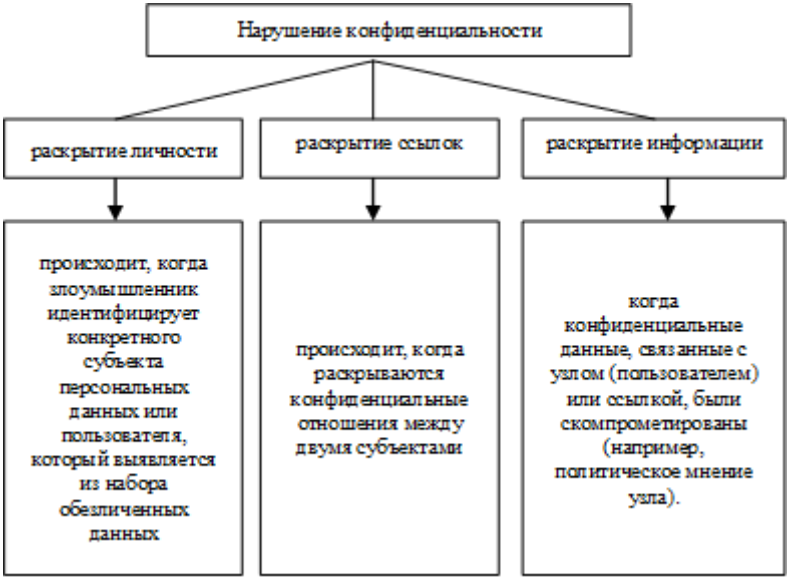

Рис. 1. Способы нарушения конфиденциальности

Применение процесса деидентификации для снижения рисков нарушения конфиденциальности, сопровождается рисками следующего характера: выделение, которое соответствует возможности изолировать некоторые или все записи, которые идентифицируют человека в наборе данных; возможность соединения, то есть способность связывать, по крайней мере, две записи, относящиеся к одному и тому же субъекту данных или группе субъектов данных (либо в одной базе данных, либо в двух разных базах данных); логический вывод, который представляет собой возможность со значительной вероятностью вывести значение атрибута из значений набора других атрибутов.

Предположим, что все данные существуют в диапазоне идентифицируемости, как показано на рис. **Ошибка! Источник ссылки**

не найден., процесс обезличивания перемещает их влево [4]. В левой части расположены данные, которые не связаны с отдельными лицами и поэтому не представляют риска для конфиденциальности. Справа находятся данные, которые связаны с конкретными субъектами. В центральной части находятся данные, которые могут быть связаны только с группами людей, и данные, которые связаны с отдельными лицами, но не могут быть привязаны к другим.

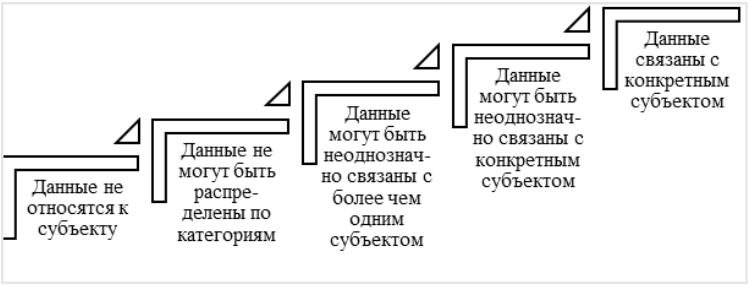

Рис. 2. Повышение риска нарушения конфиденциальности

Процесс обезличивания наблюдается при смещении данных влево с учетом необходимой практической полезности рассматриваемых наборов данных. Это снижает риск распространения обезличенных данных среди широкого круга лиц [5]. Оценка уровня обезличенности производится по специальной модели, в зависимости от риска реидентификации: в случае, когда по конечному набору данных можно определить их владельца, подобным методам обезличивания присваивается коэффициент низшего порядка. Если же установление субъекта персональных данных невозможно без существенных усилий, то коэффициент будет выше.

Применение процесса деидентификации данных как одного из базовых методов защиты конфиденциальности (при условии, что он был выполнен должным образом), риск осуществления повторной идентификации минимизируется. При этом, следует помнить, что выбор наиболее подходящего метода обезличивания зависит от набора данных, степени доступности информации для злоумышленников и типа информации, содержащейся в наборе данных, что является достаточно тривиальной задачей. Комбинированный подход к использованию методов обезличивания позволит довести процесс деперсонализации до необходимого и приемлемого уровня. Обезличенные данные теряют статус персональных данных. Если в какой-либо информационной системе персональных данных отсутствует доступ к программному

обеспечению по реидентификации, алгоритму обезличивания и предполагается работа только с обезличенными данными, например, для статистических выводов или разработки и отладки программного обеспечения, такая «внешняя» база данных перестает быть информационной системой обработки персональных данных с учетом требований к защите персональных данных.

Заключение

Таким образом, при создании системы защиты персональных данных крайне важно учитывать все имеющиеся уязвимости информационной системы персональных данных, а также характеристику возможных объектов нарушения и атак на системы реализованных злоумышленниками, пути получения несанкционированного доступа к системе и способам использования компрометированных данных. Система защиты должна строиться с учетом не только всех известных каналов проникновения, но и с учетом возможности появления преимущественно новых путей реализации угроз безопасности данных.

Литература

1. Sedayao, Jeff B. Making big data, privacy and anonymization work together in the enterprise / Jeff B. Sedayao // IEEE International Congress on Big Data. – 2014. – P. 601-607. [Электронный ресурс] : база данных. – Режим доступа: https://doi.org/10.1109/BigData.Congress.2014.92
2. Health Insurance Portability and Accountability Act (HIPAA). [Электронный ресурс] : база данных. – Режим доступа: https://www.cdc.gov/phlp/publications/topic/hipaa.html
3. Ohm, P. Broken Promises: Responding to the Surprising Failure of Anonymization / P. Ohm // 57 UCLA L. REV. – 2010. – P. 1701-1077.
4. Ganiev, A. A. Understanding of Data De-identification: Issues of Relevance and Problems / A. A. Ganiev, K. F. Kerimov, Z. I. Azizova // 2021 International Conference on Information Science and Communications Technologies (ICISCT), – 2021, P. 1-4. [Электронный ресурс] : база данных. – Режим доступа: doi: 10.1109/ICISCT52966.2021.9670054
5. Barth-Jones, D. The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections / D. Barth-Jones // DCB-J. – 2012. – P. 1-19 [Электронный ресурс] : база данных. – Режим доступа: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076398