АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

ТЕСТИРОВАНИЕ МЕТОДОВ ЗАЩИТЫ ИНФОРМАЦИИ

ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ ДЛЯ ТЕСТ-КЕЙСОВ
«УТЕЧКА БАНКОВСКОЙ ИНФОРМАЦИИ» И «МАРКЕТИНГОВОЕ КАСАНИЕ»

H F Labs

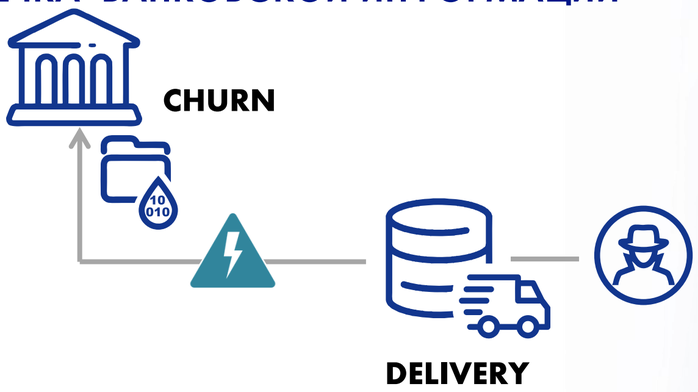
ПРЕДПОСЫЛКИ

Ассоциация Больших Данных совместно с **HFLabs** провела серию численных экспериментов по отработке риск-методики в реальных задачах обработки данных с включениями персональной информации.

HF LABS владеет востребованным продуктом «**МАСКИРОВЩИК**», предназначенным для маскирования данных и внесение шума на основе заготовленных справочников и собственной библиотеки алгоритмов.

В рамках тестирования специалисты Ассоциации подготовили тестовые наборы данных, которые затем были обработаны МАСКИРОВЩИКОМ и протестированы в рамках риск-модели Ассоциации. Такое тестирование включало в себя построение поверхностей атак, оценку риска различными способами, оценку качества, проведение итераций по балансировке качества и защищенности.

КЕЙС А:
УТЕЧКА БАНКОВСКОЙ ИНФОРМАЦИИ



КЕЙС В:
МАРКЕТИНГОВОЕ КАСАНИЕ

Целями тестирования были проверка модели риска данных с учетом работы МАСКИРОВЩИКА и разработка эталонных методов защиты конфиденциальной информации в рамках выделенных сценариев взаимодействия.

КЛЮЧЕВЫЕ ВОПРОСЫ

1. Как связаны k-anonymity (характеристика набора данных) и риски атак (характеристика кибербезопасности)?
2. Как эффективно оценить риски утечки информации с учетом внешних источников?
3. Существуют ли простые методы пересечения данных без раскрытия конфиденциальных идентификаторов?
4. Как различные виды защиты влияют на обобщенную модель информационной утечки при работе с несколькими источниками данных?

РИСК-МЕТОДИКА АССОЦИАЦИИ БОЛЬШИХ ДАННЫХ

КАСКАДНАЯ МОДЕЛЬ РИСКА

$$R_{total} = \sum_{i=1}^{n}\left(P(R_i)\cdot I(R_i)\cdot\prod_{j=1}^{i-1}P(R_j|R_{j-1})\right) \Rightarrow P_{total} = P_{\text{контекст}} \times P_{\text{данные}}$$

Модель риска данных существенно зависит от класса методов защиты информации. Ассоциация построила фреймворк, выделяющий 6 классов для методов защиты, позволяющих рассматривать широкий спектр задач защиты конфиденциальности данных

КЛАССЫ МЕТОДОВ ЗАЩИТЫ

МЕТОДЫ ЗАЩИТЫ ДАННЫХ

1. Организационные и оперативно-технические меры
2. Криптографические методы защиты
3. Методы псевдонимизации (хэши, замены)
4. Обезличивание (анонимизация: рандомизация, подавление, агрегация)
5. Конфиденциальные вычисления (SMPC, пересечения наборов)
6. Статистические методы и машинное обеспечение (дифференциальная приватность, синтетические данные)

РИСК-МОДЕЛЬ ПЕРСОНАЛЬНЫХ ДАННЫХ

Модель коммуникации | Модель угроз | Модель воздействия

АТАКИ

k-Anonymity Model
(l-diversity, t-Closeness)

$$P_{\text{риски данных}} = \frac{k}{\langle \tilde{E} \rangle}$$

Модель псевдонимизации
Resource-Based Risk Model

$$P_{\text{риски данных}} = \frac{k}{\langle r|R \rangle}$$

Модель информационной утечки
Composite Risk Model for Data Leakage

$$P_{data} = w_1 \times P_{snglout} + w_2 \times P_{link} + w_3 \times P_{inf}$$

- **Выделение** Singling Out: $P_{diff} = 1 - e^{-\epsilon}$
- **Связывание** Linkage: $P_{link} = \frac{\sum_{i=1}^{S}\gamma_i}{N}$
- **Вывод** Inference: $P_{inf} = 1 - \frac{H(D')}{H(D)}$

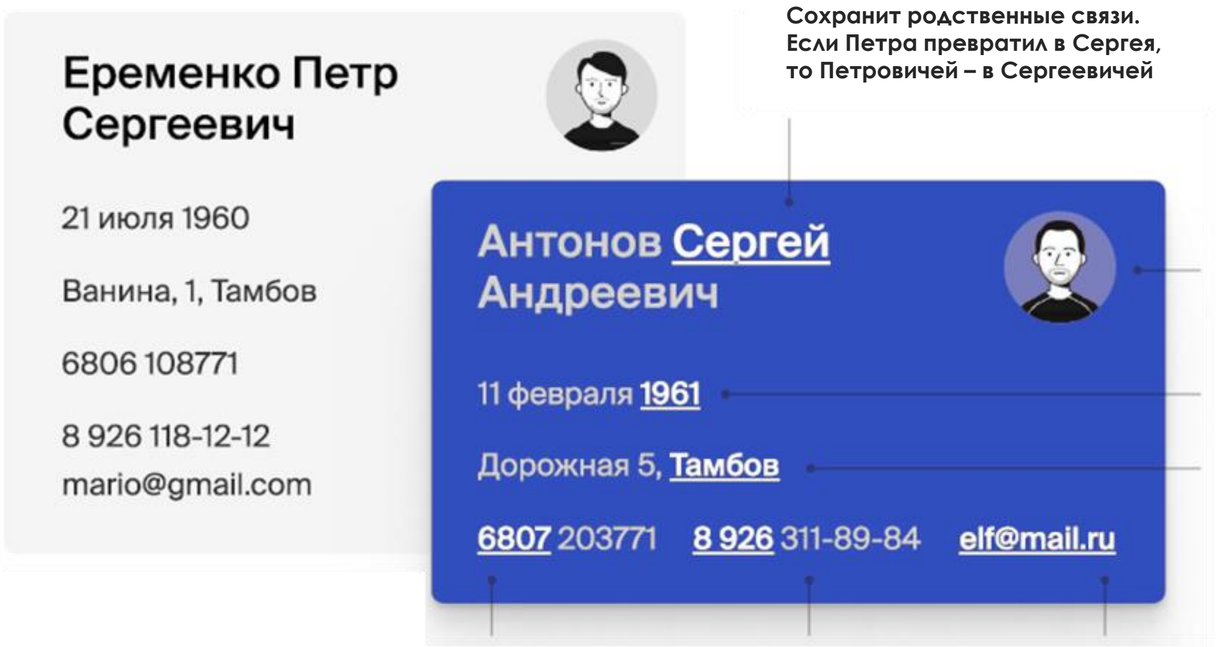
Модель контекстных рисков
Contextual Risk Assessment Model

$$P_{\text{контекстные риски}} = \frac{\sum_{j=1}^{n}\omega_j K_j}{\sum_{j=1}^{n}\omega_j}$$

«МАСКИРОВЩИК». ПРОДУКТ ДЛЯ ОБЕЗЛИЧИВАНИЯ ДАННЫХ ОТ HFLABS

Физлица

КАК РАБОТАЕТ «МАСКИРОВЩИК»



Сохранит родственные связи. Если Петра превратил в Сергея, то Петровичей – в Сергеевичей

Сохранит гендерный баланс

Сохранит социально-демографические группы

В адресе сохранит страну, город и район

Серию паспорта привяжет к году рождения, чтобы она прошла форматно-логический контроль

Оставит оператора и страну для номера телефона

Принимает домены электронных почт: рабочие, личные, одноразовые такими и останутся

H F Labs

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

«МАСКИРОВЩИК». ПРОДУКТ ДЛЯ ОБЕЗЛИЧИВАНИЯ ДАННЫХ ОТ HFLABS

Полезно и безопасно

- Безопасно обезличивает данные, путем случайных замен с учетом гибко настраиваемых правил.
- Сохранение структуры, семантической значимости и смысла.
  Валидные данные, которые можно продолжать использовать в аналитике, тестировании, построении моделей.
- Консистентность при последовательном или параллельном обезличивании нескольких стендов. Для качества интеграционного тестирования.
- 19 лет опыт в качестве данных.
  Раньше только стандартизировали, теперь и маскируем.
  Под капотом те же уникальные алгоритмы.
  Понимаем какие характеристики качества данных действительно важны.

Удобно

- Горизонтальное и вертикальное масштабирование
- Возможность использования как отдельных сервисов, так и встроенного ETL
- Можно маскировать срезы данных по условию. В том числе инкремент.
- Поддержка различных источников и типов данных
- Возможность настройки алгоритмов маскирования в зависимости от бизнес сценария. В том числе кастомизация под уникальный запрос.
- Совместимость с другими системами посредством API

КЕЙС 1: УТЕЧКА БАНКОВСКОЙ ИНФОРМАЦИИ

Цель кейса - протестировать АТАКУ СВЯЗЫВАНИЯ обезличенного банковского набора с ранее утекшим набором интернет-доставки

ОСНОВНЫЕ ПОДХОДЫ:

CHURN 10 000

DELIVERY 130 000

1

2

МАСКИРОВЩИК →

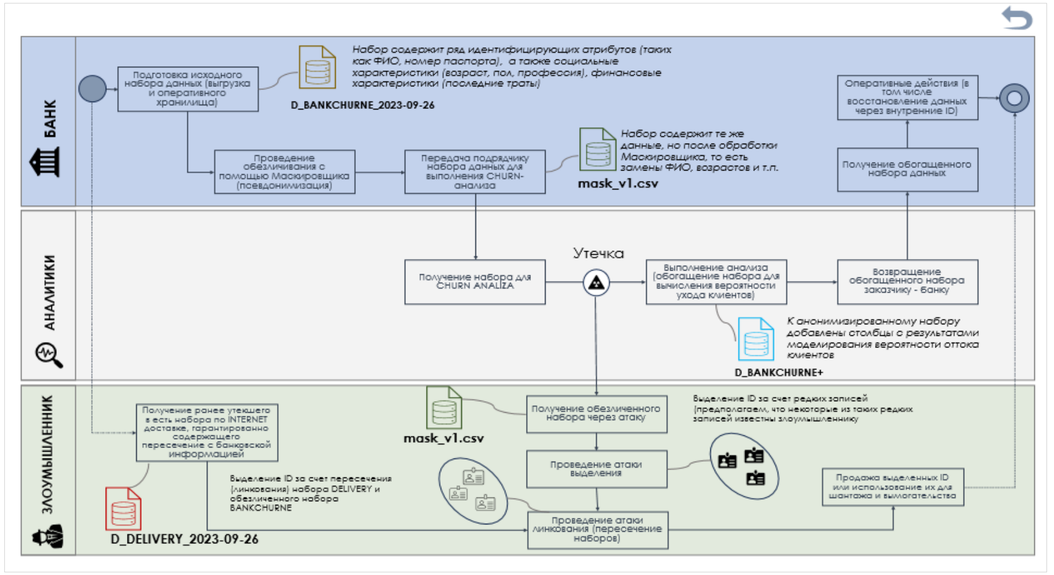
Fellegi-Sunter: $P_{posterior_{match}} = \frac{\lambda_m}{\lambda_m + (1-\lambda)u}$

K-Anonymity

$K_{Pearson}$

R^2

3

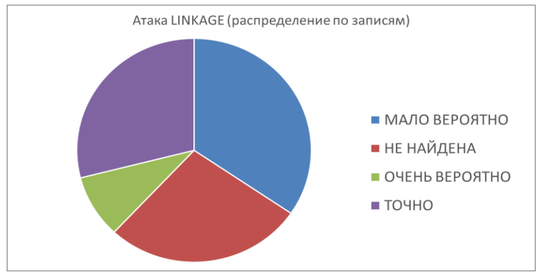


МЕТОДЫ:

- Обезличивание данных с помощью продукта "МАСКИРОВЩИК".
- Симуляция атак по методу Fellegi-Sunter для вероятностного анализа риска.
- Анализ качества через корреляцию Пирсона и метрики R^2
- Оценка риска с использованием различных метрик (доверительный интервал Уилсона).

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ:

- Значительное **снижение риска повторной идентификации** данных после обезличивания.
- **Высокая корреляция** между исходными и обезличенными данными, что подтверждает их полезность для аналитики.
- Успешная **адаптация методики для различных сценариев атак** и защиты данных.



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

КЕЙС 2: МАРКЕТИНГОВОЕ КАСАНИЕ

Цель кейса - моделирование конфиденциального объединения данных рекламной площадки и набора с банковскими транзакциями для оценки эффективности рекламной кампании.

ОСНОВНЫЕ ПОДХОДЫ:

1

БАНК

ТРАНЗАКЦИИ
100 000

РЕКЛАМНАЯ ПЛОЩАДКА

ПРОСМОТРЫ
2 000

2

ШИНОРР

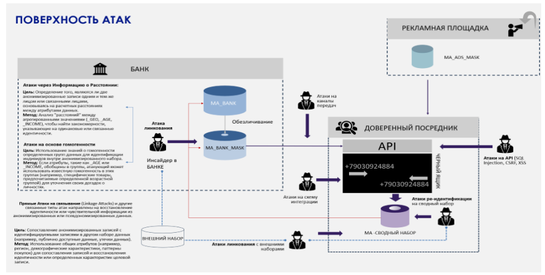
3

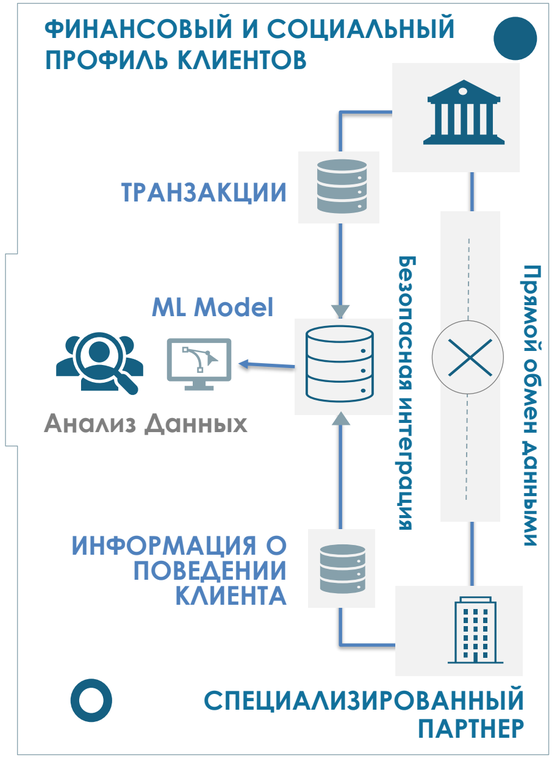
ИНФОРМАЦИОННАЯ УТЕЧКА
(Выделение, Линкование, Вывод)

$K_{Pearson}$
KS-статистика
R^2

CVPL: кластерно – векторная атака

$$sim(c_i, c_j) = \frac{1}{|c_i| \cdot |c_j|} \sum_{x_p \in c_i} \sum_{x_q \in c_j} dist(x_p, x_q)$$



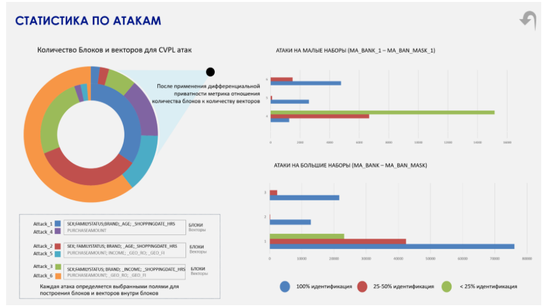


МЕТОДЫ:

- Обезличивание данных с помощью продукта "МАСКИРОВЩИК".
- Симуляция атак по методу CVPL.
- Анализ качества через корреляцию Пирсона, статистику Колмогорова-Смирнова и метрики R^2
- Оценка риска с использованием модели «УТЕЧКА ИНФОРМАЦИИ»

ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ:

- Успешное **объединение** данных рекламной площадки и банка **без утечки** персональной информации.
- Обучение модели на обезличенных данных показало **высокую точность** прогнозирования конверсии.
- Значительное снижение рисков утечки данных при сохранении их полезности для маркетингового анализа (**дифференциальная конфиденциальность**).
- **CVPL** позволяет оценить максимальную вероятность атаки с использованием **любых доступных наборов данных**



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

АЛГОРИТМЫ И ПОДХОДЫ

В рамках численных экспериментов были использованы оригинальные подходы и алгоритмы:

ИНТЕГРАЦИЯ ПО МЕТОДУ ШНОРРА

- Использование протоколов доказательства с нулевым разглашением (Zero-Knowledge Proof, ZKP) для конфиденциального объединения данных.
- Протокол Шнорра обеспечивает доказательство принадлежности данных без раскрытия самих данных. Обеспечивает высокий уровень конфиденциальности и минимизацию риска утечки данных при объединении информации.
- Альтернатива более тяжелому методу PSI (Private Set Intersection), обеспечивая высокую безопасность при меньших ресурсных затратах.

CVPL (CLUSTER-VECTOR PCA LINKAGE)

- Использование кластеризации данных и метода главных компонент (PCA) для уменьшения размерности данных.
- Векторизация данных для создания кластеров с последующим сравнением векторов внутри каждого кластера.
- Симуляция атак на обезличенные данные для оценки устойчивости к повторной идентификации.
- Позволяет выявить потенциальные уязвимости в схемах обезличивания и оптимизировать методы защиты

РИСК-МОДЕЛЬ "УТЕЧКА ИНФОРМАЦИИ"

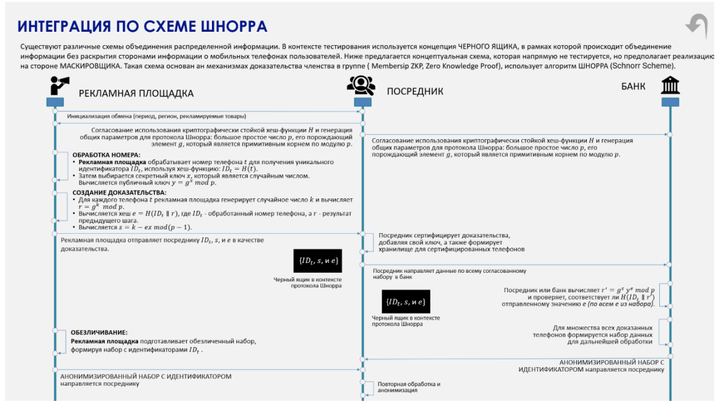
- Оценка риска утечки информации при работе с методами защиты конфиденциальных данных. Расширяет классическую модель k-anonymity и дает возможность применения для работы со связанными наборами (например, синтетическими данными).

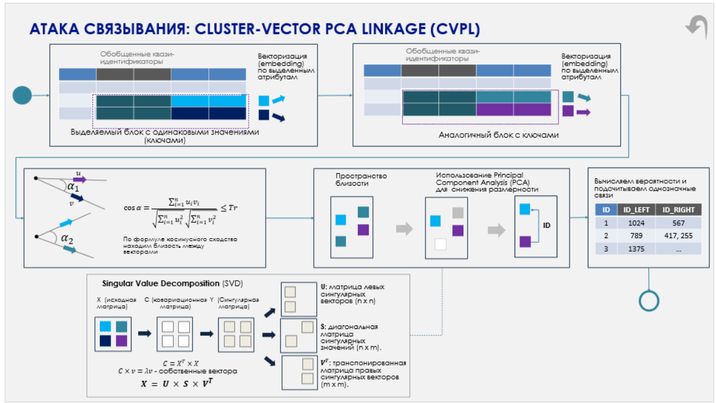
- Три ключевых аспекта рисков (три составляющих):
  - **Выделение** (Singling Out): Вероятность уникальной идентификации записи.
  - **Связывание** (Linkability): Возможность связывания записей из разных наборов данных.
  - **Вывод** (Inference): Возможность угадывания неизвестных атрибутов.

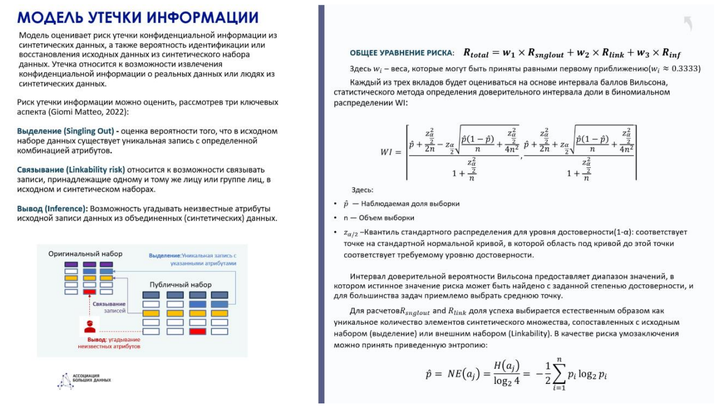
- Использование доверительных интервалов для оценки рисков на основе статистического метода Уилсона.

Преимущества:

- Полный охват различных аспектов риска утечки информации.
- Применение модели для различных сценариев анализа и защиты данных.







ПОКАЗАТЕЛИ НА ИСХОДНЫХ АЛГОРИТМАХ ДО РАСШИРЕНИЯ ФУНКЦИОНАЛА

Качество и польза

Pearson Coefficient	0.9	Уровень корреляции высокий, набор данных после обезличивания сохраняет свои характеристики
KS-статистика 1	0.01	Минимальные различия между распределениями
KS-статистика 2	0.27	Низкий показатель, демонстрирующий близость наборов
KL-дивергенция 1	2.77	Умеренное изменение
KL-дивергенция 2	0.23	Очень близкое распределение
R^2 детерминация	0.55	Приемлемый результат (выше 50%)
R^2 детерминация	0.99	Очень точное соответствие данных

Безопасность

Уровень комплексной модели риска R_{total}	**0.4**	**Относительная устойчивость**

Данные достаточно защищены, но есть потенциал для улучшения

Без учета контекста обработки

АДАПТАЦИЯ ПАРАМЕТРОВ МАСКИРОВАНИЯ С УЧЕТОМ РЕКОМЕНДАЦИЙ

Схемы обезличивания:

ПОДАВЛЕНИЕ ПРЯМЫХ ИДЕНТИФИКАТОРОВ

Удаление или замена прямых идентификаторов, для предотвращения непосредственной идентификации лиц

Основной сценарий

РАНДОМИЗАЦИЯ И РАЗМЫТИЕ С ПЕРЕМЕННЫМ РАДИУСОМ

Усложняет идентификацию отдельных пользователей

Применялось для дат рождения

ДИФФЕРЕНЦИАЛЬНАЯ ПРИВАТНОСТЬ

Квази-идентификаторы групп, введение шума. Предотвращает идентификацию даже при наличии дополнительной информации

Не применялось ранее

Расширен функционал сценариев маскирования:

1. **Размытие для бОльшего набора типов данных.**

2. **Дифференциальная приватность**

   Замены через справочники с квази-идентификаторами групп, объединяющие данные с учетом приоритетных характеристик

Понимаем влияние выбранных сценариев на баланс качества и безопасности

Можем рекомендовать оптимальный выбор модели маскирования для конкретного бизнес сценария, с учетом существующих контекстных рисков и среды

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ПОКАЗАТЕЛИ ПОСЛЕ ПРИМЕНЕНИЯ РЕКОМЕНДАЦИЙ

Качество и польза

Pearson Coefficient	0.8	- 0.1	Уровень корреляции снизился, но все еще остается высоким
KS-статистика 1	0.09	+ 0.08	Статистика ухудшилась, но незначительно
KS-статистика 2	1.7	+ 1.43	Увеличение разницы между распределениями
KL-дивергенция 1	2.85	+ 0.08	Расстояние между наборами увеличилось
KL-дивергенция 2	1.17	+ 0.94	Для защищенного варианта показатель ожидаемо вырос
R^2 детерминация	0.24	- 0.31	Для обновленной модели низкое значение
R^2 детерминация	0.83	- 0.16	На защищенных данных незначительное снижение

Безопасность

Уровень комплексной модели риска R_{total}	**0.01**	**Риск значительно снижен**

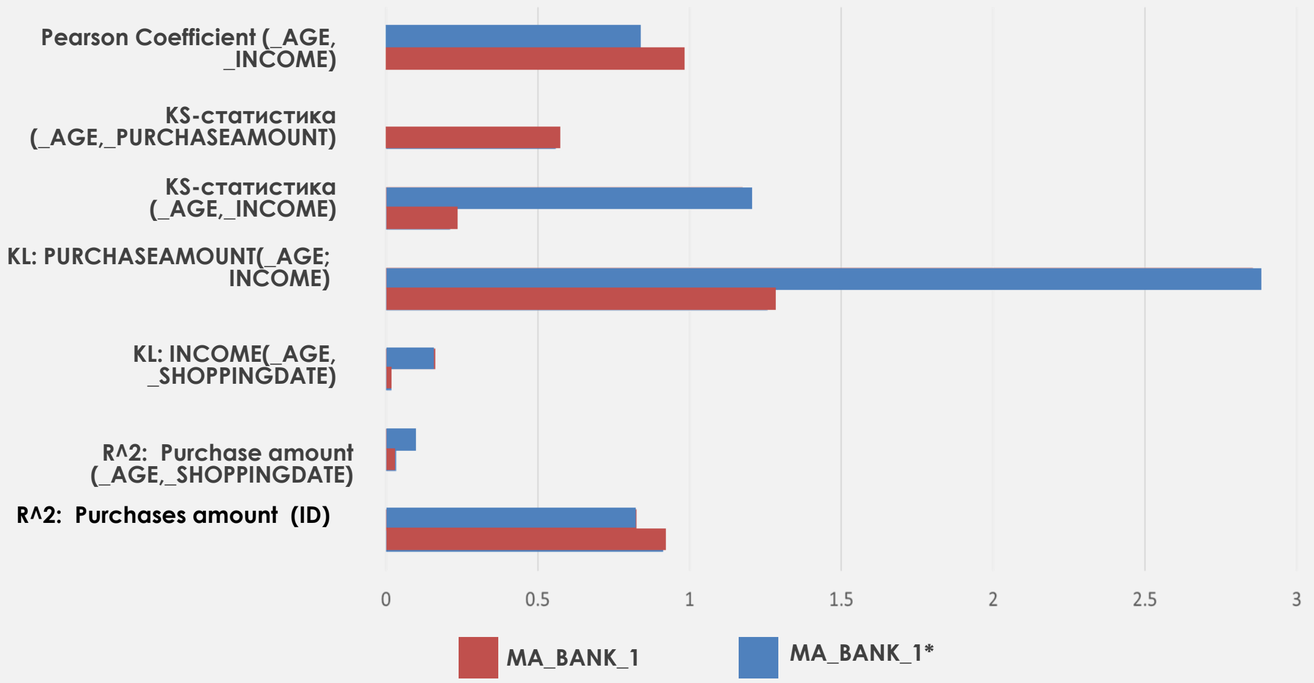
С учетом продвинутых техник атак,
при теоретическом пороге риска не более 0.1

Без учета контекста обработки

ИЗМЕНЕНИЯ КАЧЕСТВА ПОСЛЕ ПРИМЕНЕНИЯ ДИФФЕРЕНЦИАЛЬНОЙ ПРИВАТНОСТИ

Качество незначительно упало при существенном росте безопасности

КАЧЕСТВО (ПОЛЕЗНОСТЬ) ПРОТИВ ПРИВАТНОСТИ



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

РЕЗУЛЬТАТЫ КЕЙСА

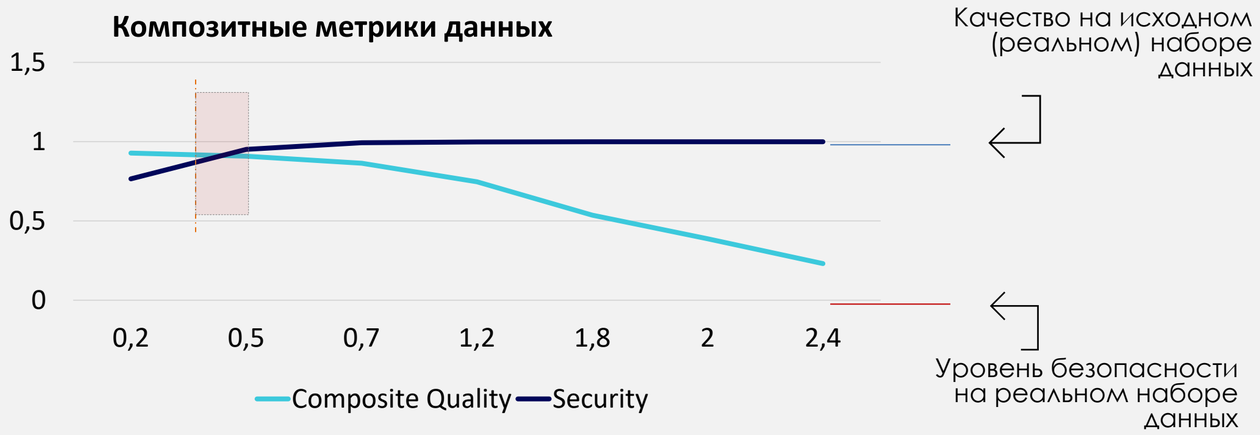
Применение мер защиты незначительно снизило качество данных, как по метрикам, так и по целевым бизнес-характеристикам. Ожидаемое соотношение – при увеличении шума снижаются метрики риска, включая риски атак (растет степень защиты), но также снижается качество)

Бизнес-метрики:

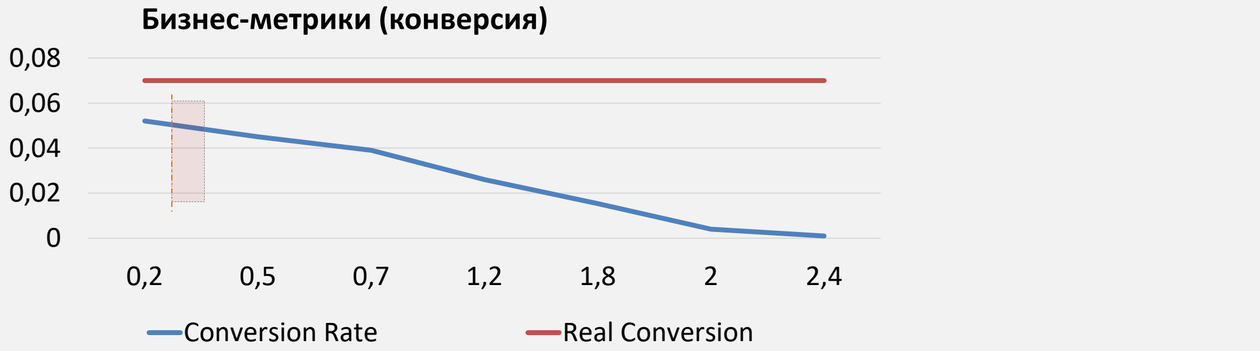
$$Conversion\ Rate\ = \left(\frac{Number\ of\ Conversions}{Total\ Number\ of\ Interactions}\right) \times 100\%$$

	КАЧЕСТВО ДАННЫХ	БЕЗОПАСНОСТЬ	БИЗНЕС-МЕТРИКА
ПРЯМАЯ ОБРАБОТКА	100%	0%	7%
ЗАЩИЩЕННАЯ ОБРАБОТКА	86%	99.98%	5%

Поведение композитных метрик и бизнес-значения при изменении шума



Конверсия на реальном наборе составляет 7% и падает при размывании данных



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

РАСШИРЕНИЕ ФУНКЦИОНАЛЬНЫХ ВОЗМОЖНОСТЕЙ «МАСКИРОВЩИКА»

ИСХОДНО:

Алгоритмы, сохраняющие больше смысловых характеристик.

Стандартизация и умный анализ, сохранение качества и распределения.

Данные максимально полезны для аналитики.

Дополнительные алгоритмы для гибкой настройки соотношения качества и безопасности:

УПРОЩЕННЫЕ АЛГОРИТМЫ БЕЗ СТАНДАРТИЗАЦИИ
Усложняют обратную идентификацию.
Сохраняют меньше характеристик либо только формат.

Увеличение скорости от 30% до 2000% относительно исходных в зависимости от типа данных

АЛГОРИТМЫ С РАСШИРЕНИЕМ СХЕМЫ ОБЕЗЛИЧИВАНИЯ
(дифференциальная приватность и прочее).
Сохраняют меньше детальных характеристик для аналитики.

Снижение скорости в пределах 7% относительно исходных

ПЛАНЫ РАЗВИТИЯ

1

Преднастроенные рекомендации по выбору набора алгоритмов маскирования

Возможность выбора и настройки схемы обработки

2

Поиск чувствительной информации и маскирование данных в файлах и свободном тексте

Учитываем контекст сценария использования данных:

РАСПОЛОЖЕНИЕ
- во внутреннем контуре/ выкладываются вовне

ДОСТУП
- только сотрудники / внешние подрядчики

СЦЕНАРИЙ РАБОТЫ С ДАННЫМИ
- Обмен данными со сторонними организациями
- Отправка данных в chatGpt и другие внешние системы
- Интеграционное тестирование внутри контура

УРОВЕНЬ СОХРАНЕНИЯ КАЧЕСТВА И СМЫСЛА ДЛЯ КОНКРЕТНОГО ТИПА ДАННЫХ
- Выделение групповых характеристик, приоритетных для аналитики
- Выделение данных, где достаточно сохранить только формат

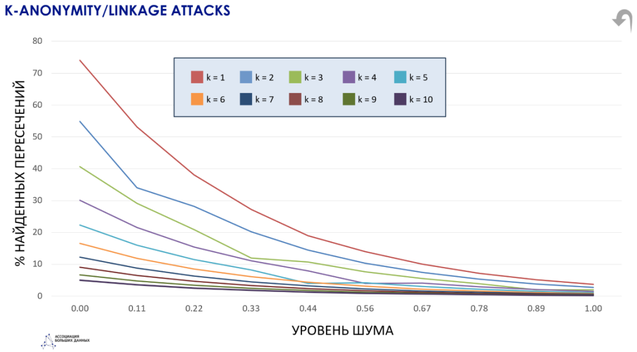
МАСКИРОВАНИЕ ДОКУМЕНТОВ РАЗЛИЧНЫХ ТИПОВ С СОХРАНЕНИЕМ ФОРМАТИРОВАНИЯ

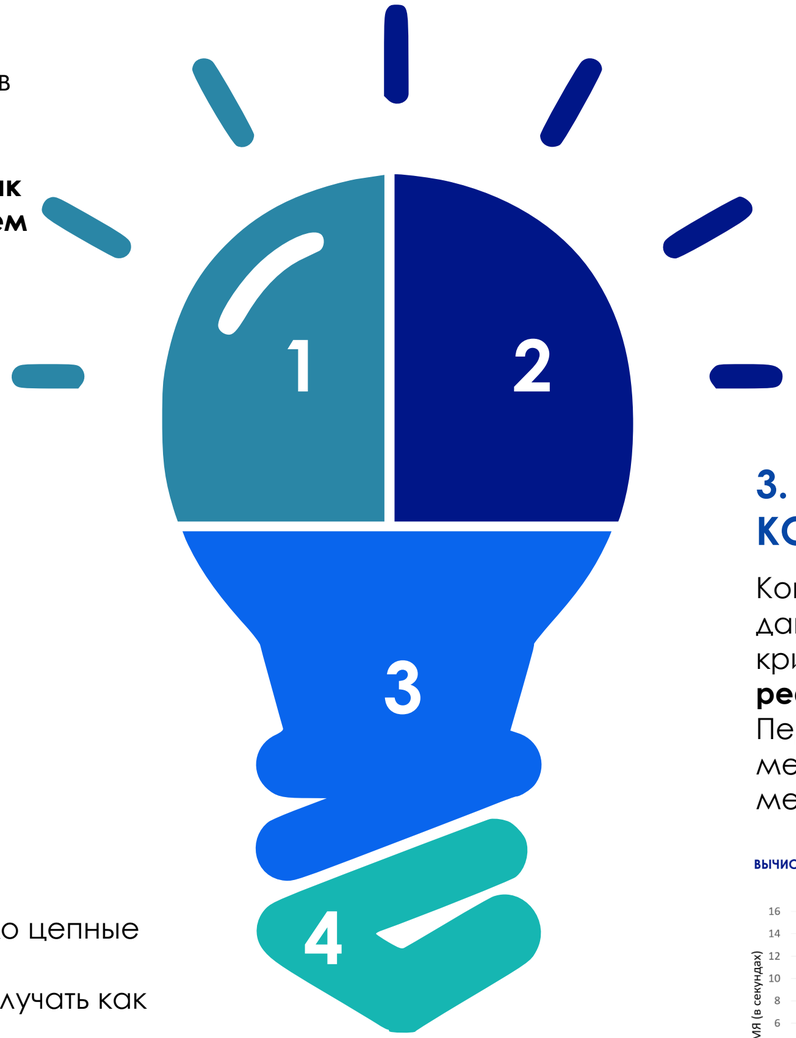
СЦЕНАРИИ РАБОТЫ С CHATGPT, СОХРАНЯЮЩИЕ КОНТЕКСТ ПРИ ЗАМЕНАХ В ДИАЛОГЕ

ОТВЕТЫ НА ПОСТАВЛЕННЫЕ ВОПРОСЫ

1. K-ANONYMITY / LINKAGE SUCCESS

Модель k-anonymity может успешно применяться в случае единственного источника данных. С возрастанием k вероятность повторной идентификации падает, однако, **возможность атак связывания хотя и уменьшается, но медленнее чем рост k**.
ВЫВОД: В сложных случаях взаимодействия рекомендуется моделировать атаки связывания и включать их в оценку рисков





2. ОЦЕНКА РИСКОВ ПЕРЕСЕЧЕНИЯ

Оценка рисков пересечения данных с внешними источниками может производиться на основе сравнения защищенного набора информации и исходного набора: такое сравнение дает **оценку сверху (максимальный риск) по отношению к “неизвестным утечкам”**

3. СЛОЖНОСТЬ КОНФИДЕНЦИАЛЬНЫХ ПЕРЕСЕЧЕНИЙ

Конфиденциальные методы вычислений (такие как PS3I) дают высокую надежность за счет использования сложных криптографических техник. В тоже время они **сложны в реализации** и тяжело обобщаются на многих участников. Перспективным направлением является использование методов “с нулевым доказательством” (ZKP), таких как метод Шнорра

4. РИСК МОДЕЛЬ:

Модель ИНФОРМАЦИОННОЙ УТЕЧКИ представляет обобщенную риск-модель, учитывающую не только цепные (каскадные) эффекты, но и кумулятивные угрозы. Применение доверенных интервалов позволяет получать как оценки сверху, так и снизу для рисков работы с конфиденциальной информацией.

$$P_{data} = w_1 \cdot \hat{P}_{singl} + w_2 \cdot \hat{P}_{linkage} + w_3 \cdot \hat{P}_{inher}$$

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

1. **ЭФФЕКТИВНОСТЬ РИСК-МОДЕЛИ**: Риск-модель доказала свою работоспособность. Для увеличения точности прогнозирования и оценки публикуемых данных необходимо использовать симуляции атак, которые углубляют понимание рисков повторной идентификации за счет учета рисков выделения и связывания.

2. **ЗАЩИТА ДАННЫХ МАСКИРОВЩИКОМ**: Продукт HF Labs "МАСКИРОВЩИК" успешно применен для обезличивания данных, сохраняя их значимость для аналитики. Обезличенные данные показали высокую устойчивость к атакам и минимальный риск повторной идентификации.

3. **МЕТОДИКА БЕЗОПАСНОЙ ИНТЕГРАЦИИ**: Протокол Шнорра продемонстрировал высокую эффективность в конфиденциальном объединении данных. Использование схем PSI и TEE обеспечило высокий уровень безопасности при интеграции данных.

4. **УЧЕТ КОНТЕКСТНЫХ РИСКОВ**: Модель безопасности не может существовать в отрыве от контекстных рисков. Дополнительные факторы, такие как композитные риски, могут быть учтены с помощью аппроксимации полиномиальными функциями, что позволит учитывать большее количество косвенных влияний на безопасность, чем это возможно при прямой факторизации риска.

ОСНОВНЫЕ ВЫВОДЫ

Риски обработки клиентских данных могут (и должны) быть измерены для **конкретного бизнес-кейса**

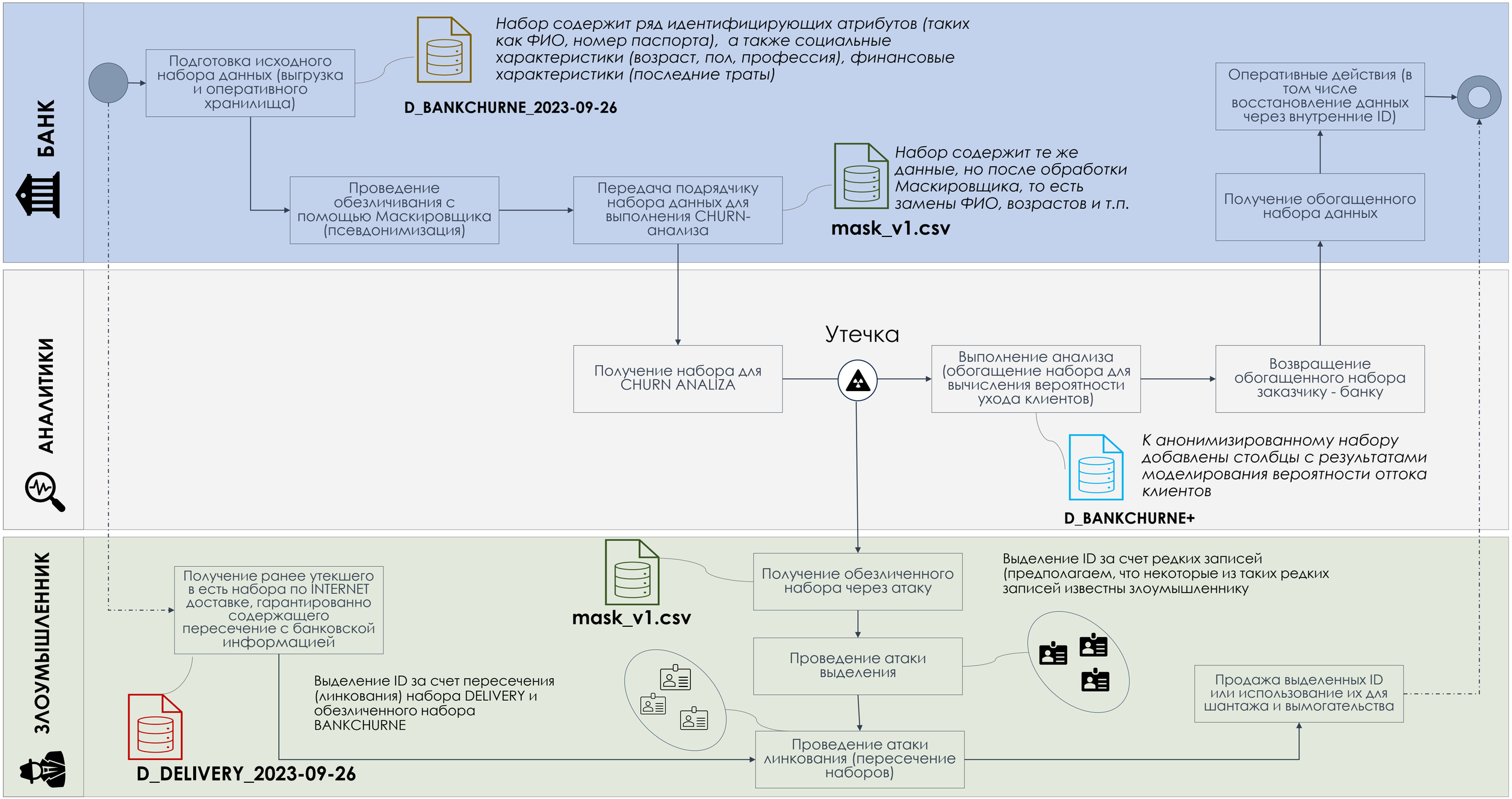
Существуют техники и технологии снижения риска реидентификации **до околонулевых значений** даже **при использовании дополнительной информации**

Использование технологий повышения конфиденциальности лежит **в «серой» зоне** нормативного регулирования, не успевающего за их развитием

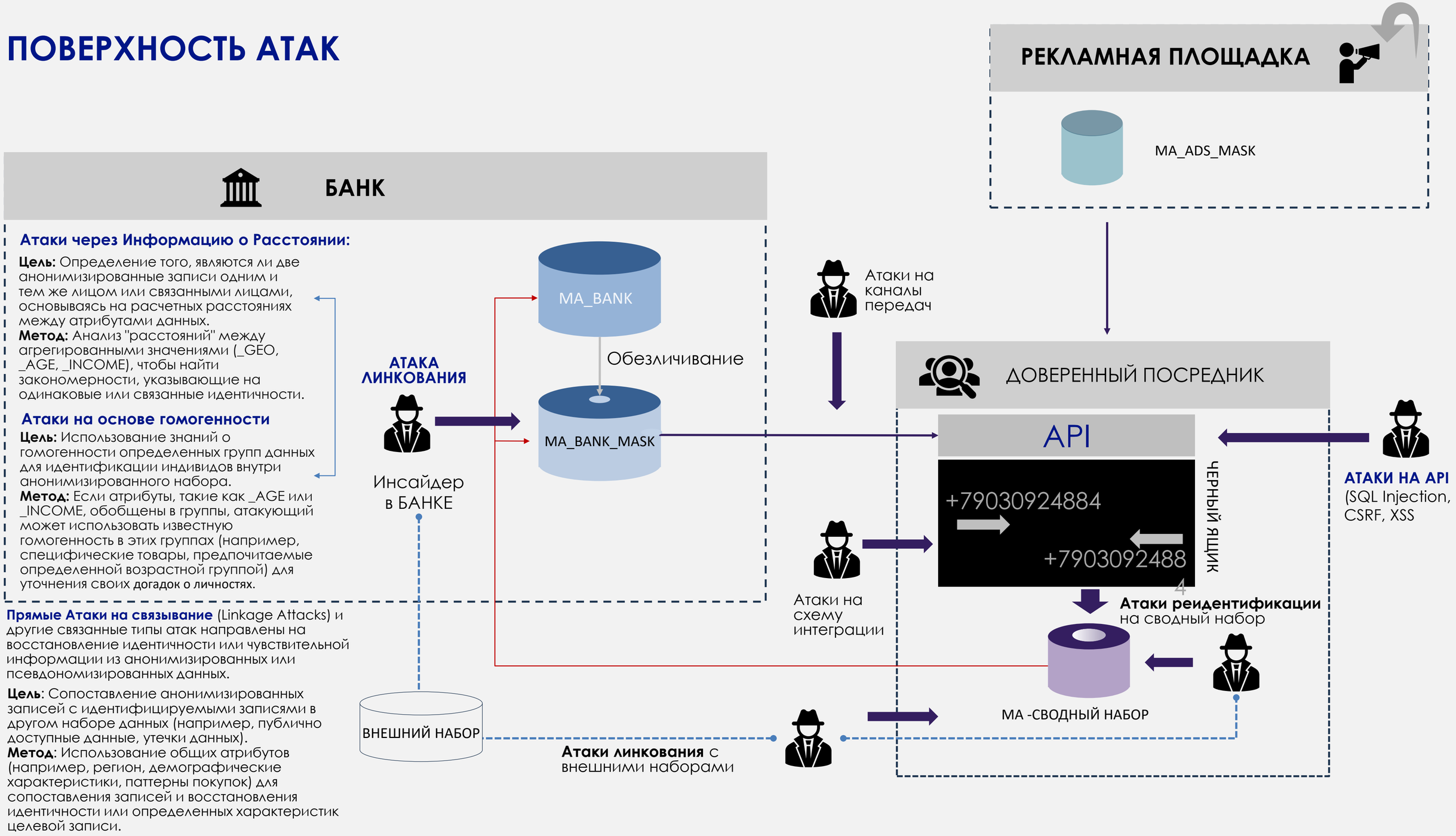
Закрепление модели оценки рисков будет способствовать **быстрому внедрение технологий на данных** при сохранении должного уровня конфиденциальности

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

СПАСИБО ЗА ВНИМАНИЕ!


БАНК
Подготовка исходного набора данных (выгрузка и оперативного хранилища)
Набор содержит ряд идентифицирующих атрибутов (таких как ФИО, номер паспорта), а также социальные характеристики (возраст, пол, профессия), финансовые характеристики (последние траты)
D_BANKCHURNE_2023-09-26
Проведение обезличивания с помощью Маскировщика (псевдонимизация)
Передача подрядчику набора данных для выполнения CHURN-анализа
Набор содержит те же данные, но после обработки Маскировщика, то есть замены ФИО, возрастов и т.п.
mask_v1.csv
Оперативные действия (в том числе восстановление данных через внутренние ID)
Получение обогащенного набора данных
АНАЛИТИКИ
Получение набора для CHURN ANALIZA
Утечка
Выполнение анализа (обогащение набора для вычисления вероятности ухода клиентов)
Возвращение обогащенного набора заказчику - банку
К анонимизированному набору добавлены столбцы с результатами моделирования вероятности оттока клиентов
D_BANKCHURNE+
ЗЛОУМЫШЛЕННИК
Получение ранее утекшего в есть набора по INTERNET доставке, гарантированно содержащего пересечение с банковской информацией
mask_v1.csv
Получение обезличенного набора через атаку
Выделение ID за счет редких записей (предполагаем, что некоторые из таких редких записей известны злоумышленнику
Проведение атаки выделения
Выделение ID за счет пересечения (линкования) набора DELIVERY и обезличенного набора BANKCHURNE
Проведение атаки линкования (пересечение наборов)
Продажа выделенных ID или использование их для шантажа и вымогательства
D_DELIVERY_2023-09-26

ПОВЕРХНОСТЬ АТАК



АЛГОРИТМЫ И МЕТРИКИ

1

ГЕНЕРАЦИЯ ПСЕВДОРЕАЛЬНЫХ НАБОРОВ ДАННЫХ

MA_BANK

MA_ADS

- ☑ **Большое количество атрибутов**
- ☑ Использование **категориальных, числовых, календарных и гео**
- ☑ **Заложенная внутренняя статистика (копула)**, привязанная к региону, имитирующая реальную рекламную компанию

2

СИМУЛЯЦИЯ ИНТЕГРАЦИОННОЙ СХЕМЫ (ВНЕШНИЕ ИДЕНТИФИКАТОРЫ)

- ☐ Fellegi-Sunter Model
- ☐ Cluster Based with k-means
- ☑ **Cluster-Vector PCA Linkage" (CVPL)**
- ☐ Nearest Neighbor Matching
- ☐ Support Vector Machine Matching
- ☐ Deep Learning Based Matching

MA_BANK — Phone — Phone — MA_ADS

3

ОБЕЗЛИЧИВАНИЕ ИСХОДНЫХ НАБОРОВ С ВНЕШНИМИ ИДЕНТИФИКАТОРАМИ

Пересечение наборов может производиться по одной из защищенных схем:

- ☐ PSI (Private Set Intersection)
- ☐ TEE (Trust Execution Environment)
- ☑ Динамический обмен через централизованного посредника

MA_BANK_MASK

MA_ADS_MASK

Наборы данных в процессе обезличивания проходят обработку:

- ☑ **Классическое обезличивание** (обобщение, рандомизация с постоянным радиусом размытия
- ☐ Обработка по схемам с дифференциальной приватностью (лапласовский механизм шума)
- ☐ Подавление записей или аугментация данных за счет инъекций записей
- ☐ Построение синтетических наборов с AI по схемам PATE-GAN и подобным

Использование **МАСКИРОВЩИКА**

4

ТЕСТИРОВАНИЕ БЕЗОПАСНОСТИ И РАСЧЕТ РИСКОВ

Симуляция атак связывания

Симуляция атак выделения

Нормирование наборов информации

Расчет метрик безопасности

Расчет метрик качества

- ☑ Корреляционные метрики (Correlation Metrics: **Pearson Correlation Coefficient**, Mutual Information)
- ☑ Метрики распределения (Distribution Metrics: **Kolmogorov-Smirnov Test, Entropy Comparison**)
- ☑ Статистические метрики (Statistical Metrics: **Mean, Median**, Variance, Skewness, Kurtosis, R^2 **Coefficient of Determination**)
- ☐ Классификационные и ранжировочные метрики (Classification and Ranking Metrics: AUROC, AUPRC)
- ☑ Метрики расстояния и различия (Distance and Discrepancy Metrics: Hamming Distance, Chebyshev Distance, Cosine Distance, **Kullback-Leibler Divergence**, Jaccard Index)

- ☑ Метрики приватности данных (Data Privacy Metrics: k-Anonymity, l-Diversity, t-Closeness)
- ☐ Метрики дифференциальной приватности (Differential Privacy)
- ☑ Метрики риска утечки информации (Information Leakage Risk Metrics: (Singling Out, Linkability, Inference/Conditional Entropy)
- ☐ Метрики оценки риска (Risk Assessment Metrics: Риск повторной идентификации данных, Intruder Attack Success Rate, Контекстные риски)
- ☐ Vulnerability Exploit Probability (Вероятность эксплуатации уязвимости: Time to Compromise (Время до компрометации), Attack Path Length - Длина пути атаки)
- ☐ Метрики доверия и надежности (Trust and Reliability Metrics: Data Integrity Score, System Availability, MTBF)

5

ФОРМИРОВАНИЕ МЕТРИК И ОЦЕНКА РЕЗУЛЬТАТОВ

6

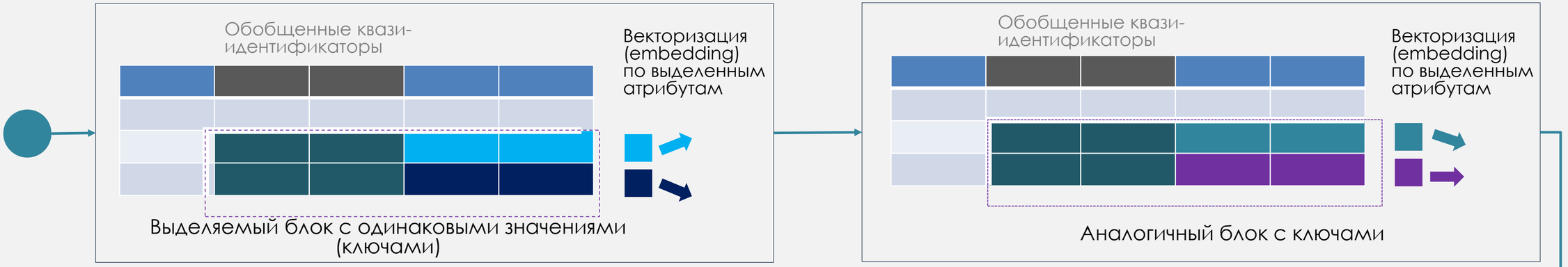
РЕЗУЛЬТИРУЮЩИЙ НАБОР

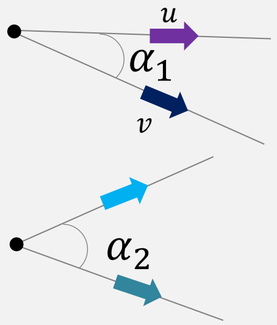
MA

7

ЗАВЕРШАЮЩИЙ ОТЧЕТ И РЕКОМЕНДАЦИИ

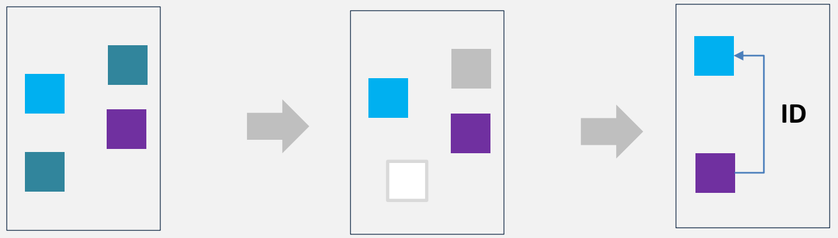
АТАКА СВЯЗЫВАНИЯ: CLUSTER-VECTOR PCA LINKAGE (CVPL)





$$\cos\alpha = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \leq Tr$$

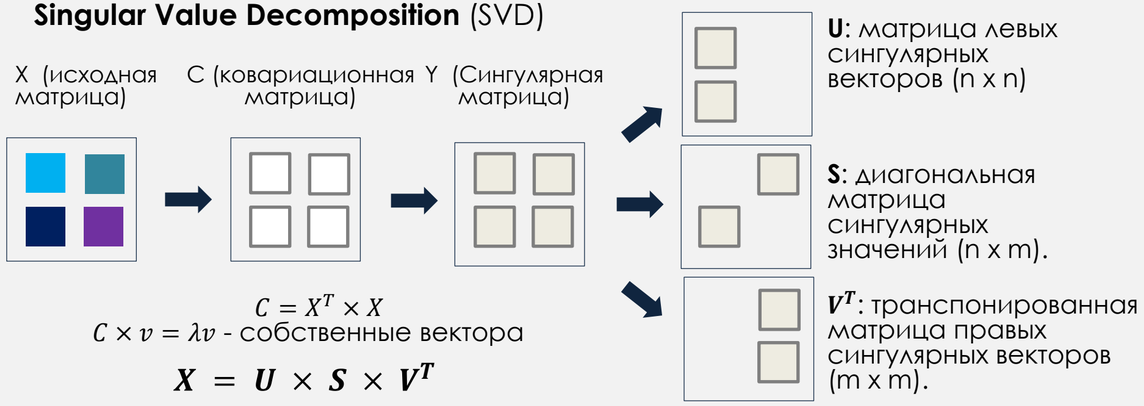
По формуле косинусного сходства находим близость между векторами



Вычисляем вероятности и подсчитываем однозначные связи

ID	ID_LEFT	ID_RIGHT
1	1024	567
2	789	417, 255
3	1375	…

Singular Value Decomposition (SVD)



$C = X^T \times X$

$C \times v = \lambda v$ - собственные вектора

$$\boldsymbol{X = U \times S \times V^T}$$

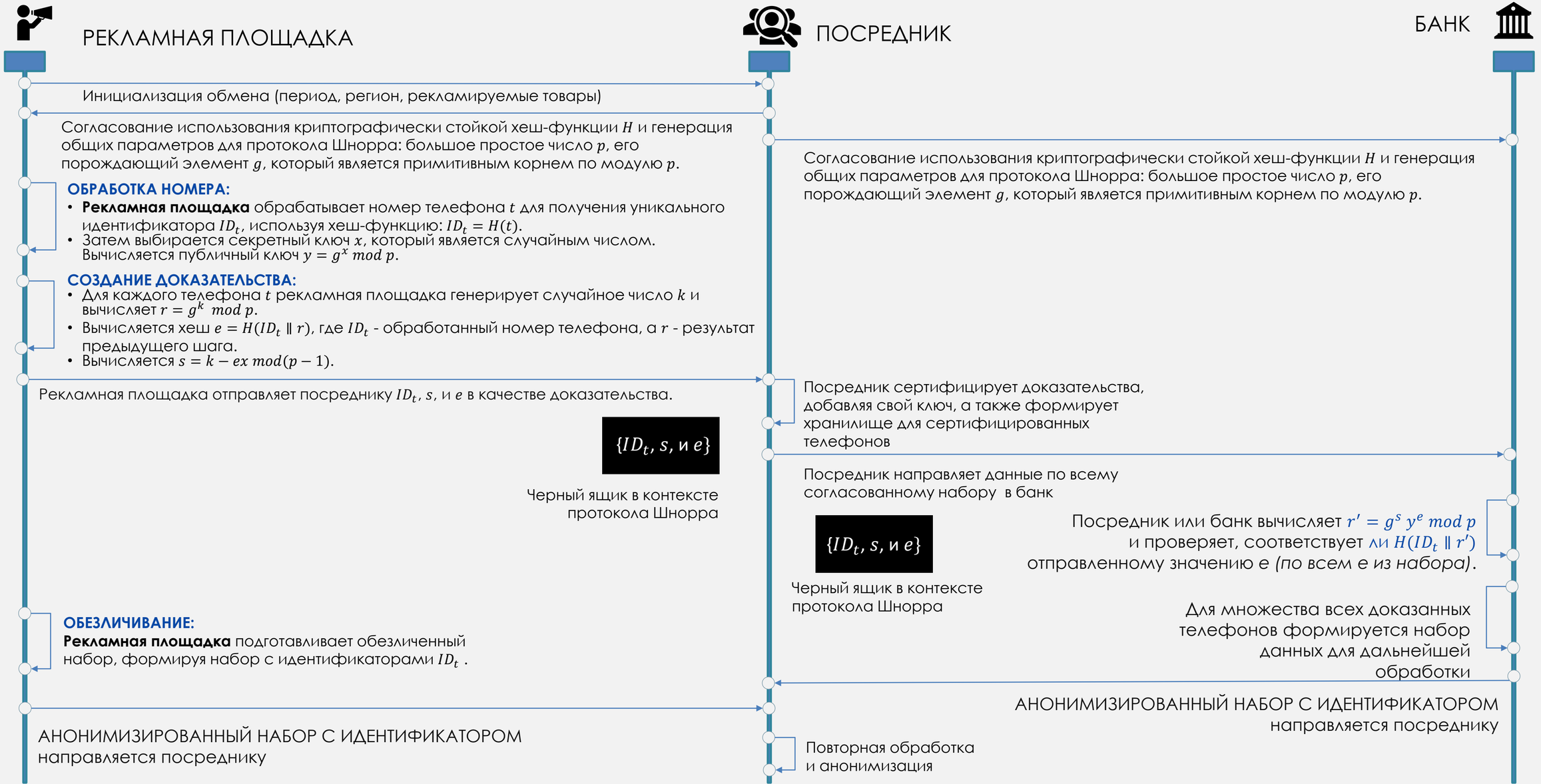
U: матрица левых сингулярных векторов (n x n)

S: диагональная матрица сингулярных значений (n x m).

$\boldsymbol{V^T}$: транспонированная матрица правых сингулярных векторов (m x m).

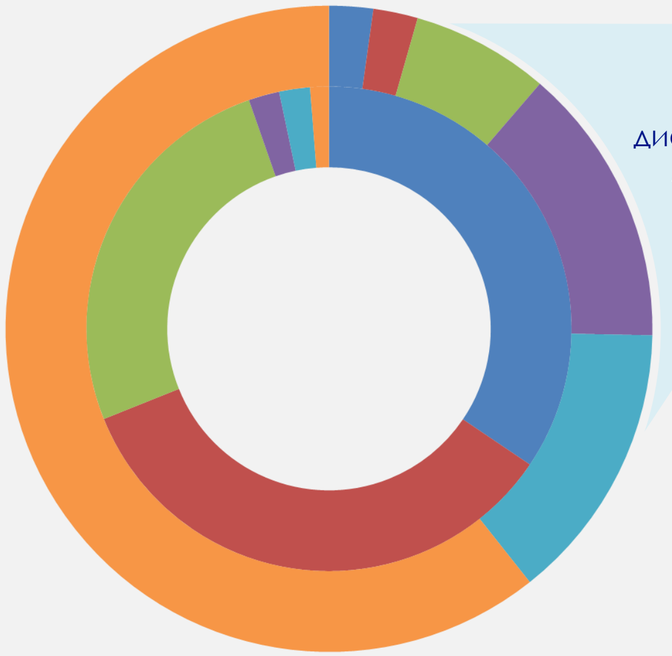
ИНТЕГРАЦИЯ ПО СХЕМЕ ШНОРРА

Существуют различные схемы объединения распределенной информации. В контексте тестирования используется концепция ЧЕРНОГО ЯЩИКА, в рамках которой происходит объединение информации без раскрытия сторонами информации о мобильных телефонах пользователей. Ниже предлагается концептуальная схема, которая напрямую не тестируется, но предполагает реализацию на стороне МАСКИРОВЩИКА. Такая схема основан ан механизмах доказательства членства в группе (Membersip ZKP, Zero Knowledge Proof), использует алгоритм ШНОРРА (Schnorr Scheme).



СТАТИСТИКА ПО АТАКАМ

Количество Блоков и векторов для CVPL атак

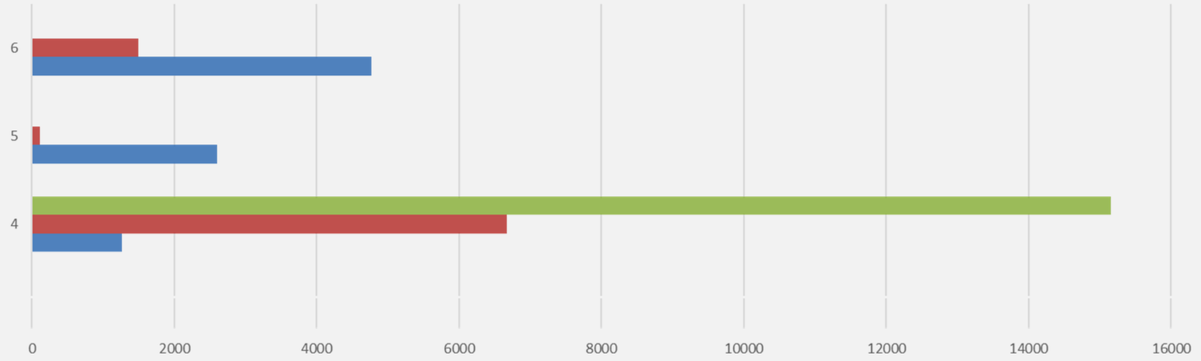


После применения дифференциальной приватности метрика отношения количества блоков к количеству векторов

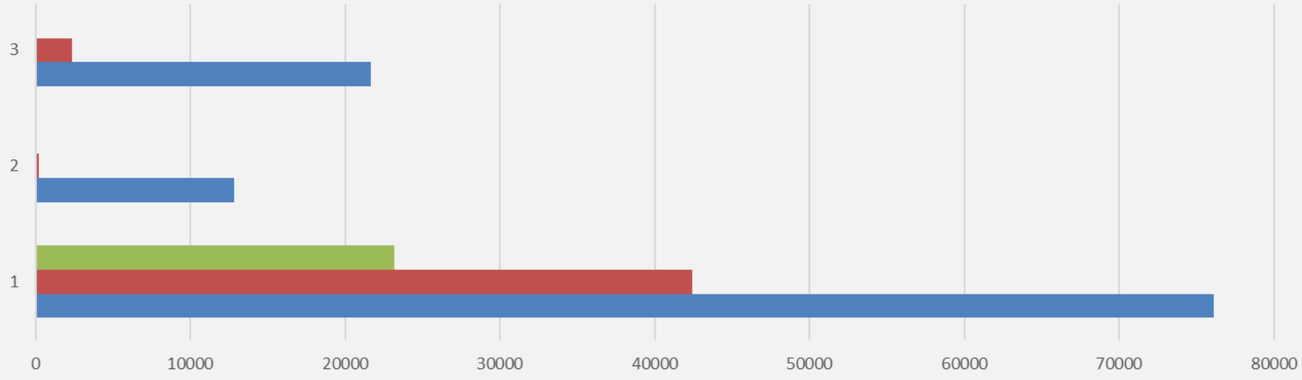
Attack_1 Attack_4	SEX;FAMILYSTATUS;BRAND;_AGE; _SHOPPINGDATE_HRS PURCHASEAMOUNT	БЛОКИ Векторы
Attack_2 Attack_5	SEX; FAMILYSTATUS; BRAND; _AGE;_SHOPPINGDATE_HRS PURCHASEAMOUNT; INCOME; _GEO_RO; _GEO_FI	БЛОКИ Векторы
Attack_3 Attack_6	SEX;FAMILYSTATUS; BRAND; _INCOME; _SHOPPINGDATE_HRS PURCHASEAMOUNT; _GEO_RO; _GEO_FI	БЛОКИ Векторы

Каждая атака определяется выбранными полями для построения блоков и векторов внутри блоков

АТАКИ НА МАЛЫЕ НАБОРЫ (MA_BANK_1 – MA_BAN_MASK_1)



АТАКИ НА БОЛЬШИЕ НАБОРЫ (MA_BANK – MA_BAN_MASK)



100% идентификация

25-50% идентификация

< 25% идентификация

МОДЕЛЬ УТЕЧКИ ИНФОРМАЦИИ

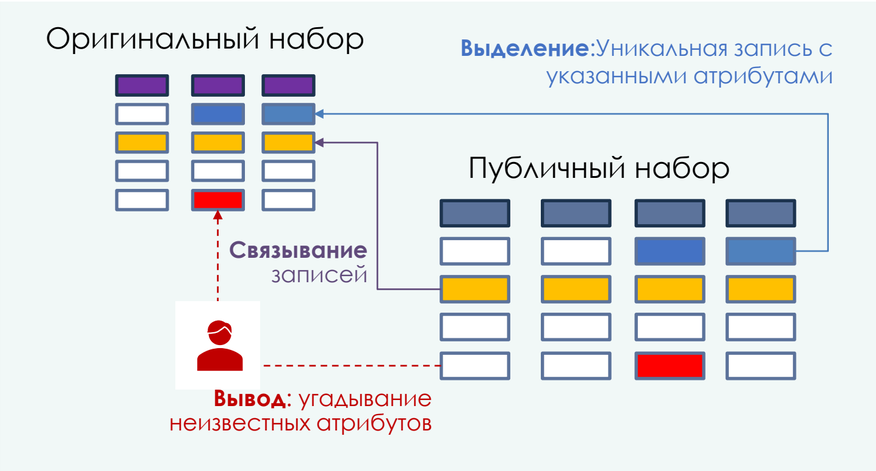
Модель оценивает риск утечки конфиденциальной информации из синтетических данных, а также вероятность идентификации или восстановления исходных данных из синтетического набора данных. Утечка относится к возможности извлечения конфиденциальной информации о реальных данных или людях из синтетических данных.

Риск утечки информации можно оценить, рассмотрев три ключевых аспекта (Giomi Matteo, 2022):

Выделение (Singling Out) - оценка вероятности того, что в исходном наборе данных существует уникальная запись с определенной комбинацией атрибутов.

Связывание (Linkability risk) относится к возможности связывать записи, принадлежащие одному и тому же лицу или группе лиц, в исходном и синтетическом наборах.

Вывод (Inference): Возможность угадывать неизвестные атрибуты исходной записи данных из объединенных (синтетических) данных.



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ОБЩЕЕ УРАВНЕНИЕ РИСКА: $R_{total} = w_1 \times R_{snglout} + w_2 \times R_{link} + w_3 \times R_{inf}$

Здесь w_i – веса, которые могут быть приняты равными первому приближению($w_i \approx 0.3333$)

Каждый из трех вкладов будет оцениваться на основе интервала баллов Вильсона, статистического метода определения доверительного интервала доли в биномиальном распределении WI:

$$WI = \left[\frac{\hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2n} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{z_{\frac{\alpha}{2}}^2}{n}}, \frac{\hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2n} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{z_{\frac{\alpha}{2}}^2}{n}} \right]$$

Здесь:

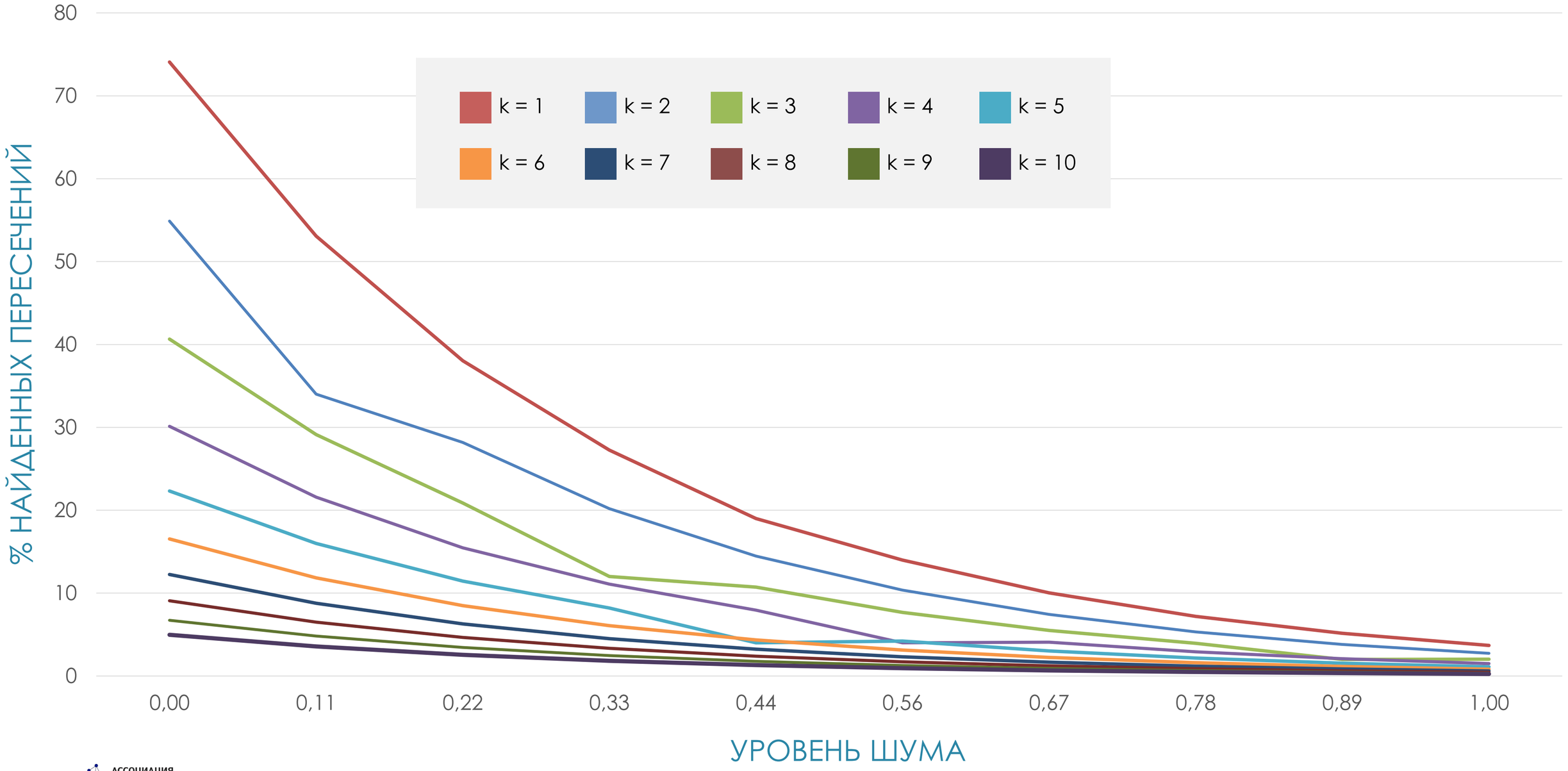
- $\hat{p}$ — Наблюдаемая доля выборки
- n — Объем выборки
- $z_{\alpha/2}$ – Квантиль стандартного распределения для уровня достоверности(1-a): соответствует точке на стандартной нормальной кривой, в которой область под кривой до этой точки соответствует требуемому уровню достоверности.

Интервал доверительной вероятности Вильсона предоставляет диапазон значений, в котором истинное значение риска может быть найдено с заданной степенью достоверности, и для большинства задач приемлемо выбрать среднюю точку.

Для расчетов $R_{snglout}$ and R_{link} доля успеха выбирается естественным образом как уникальное количество элементов синтетического множества, сопоставленных с исходным набором (выделение) или внешним набором (Linkability). В качестве риска умозаключения можно принять приведенную энтропию:

$$\hat{p} = NE(a_j) = \frac{H(a_j)}{\log_2 4} = -\frac{1}{2}\sum_{i=1}^{n} p_i \log_2 p_i$$

K-ANONYMITY/LINKAGE ATTACKS



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ ПО ШНОРРУ

Intel Core i7-9750 CPU 32Gb RAM

Время генерации ключей: $O(n^2)$
- Время генерации ключей на одной стороне (Client).
- Время генерации ключей на другой стороне (Server).

Время обмена ключами: $O(n)$
- Время, затраченное на передачу данных между сторонами (сетевое время).

Время вычисления пересечений: $O(n^2)$
- Время, затраченное на вычисление пересечений на одной стороне.
- Время, затраченное на вычисление пересечений на другой стороне.