H F Labs

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

Оценка рисков в продуктах по обезличиванию

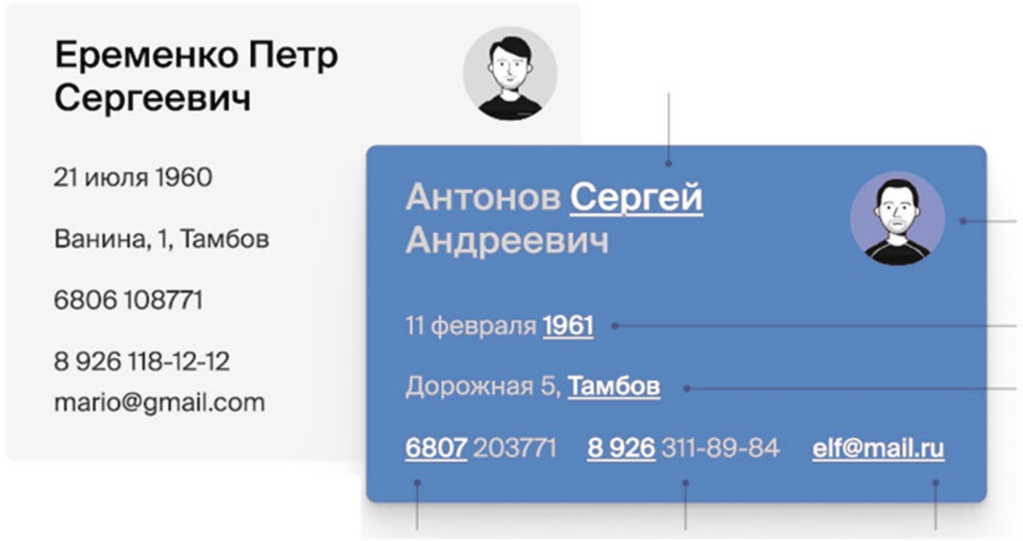
CDI CONF 2024

Марат Тахавиев
Руководитель GR-проектов

МОДЕЛЬ РИСКА АССОЦИАЦИИ БОЛЬШИХ ДАННЫХ

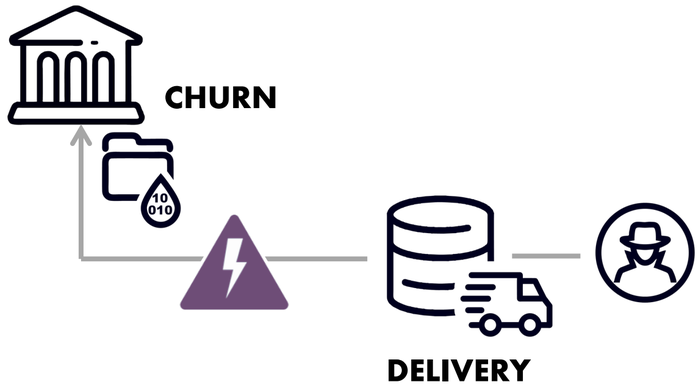
$$R_{total} = \sum_{i=1}^{n} \left(P(R_i) \cdot I(R_i) \cdot \prod_{j=1}^{i-1} P(R_j | R_{j-1}) \right) \Rightarrow P_{total} = P_{\text{контекст}} \times P_{\text{данные}}$$

«МАСКИРОВЩИК»

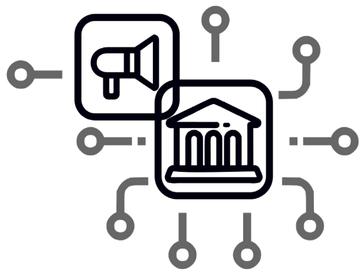


КЕЙСЫ

УТЕЧКА БАНКОВСКОЙ ИНФОРМАЦИИ

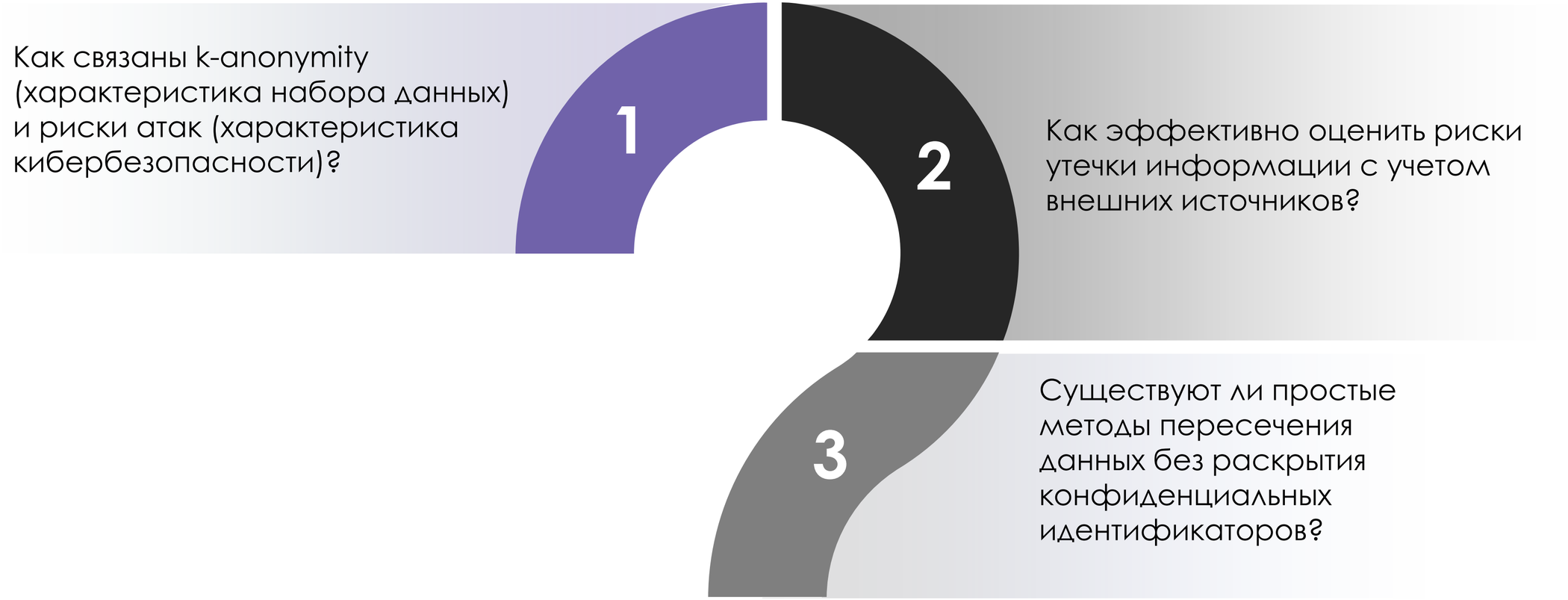


МАРКЕТИНГОВОЕ КАСАНИЕ



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

КЛЮЧЕВЫЕ ВОПРОСЫ



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

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

МЕТОДЫ ЗАЩИТЫ ДАННЫХ

КЛАССЫ МЕТОДОВ ЗАЩИТЫ

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

РИСК-МОДЕЛЬ ПЕРСОНАЛЬНЫХ ДАННЫХ

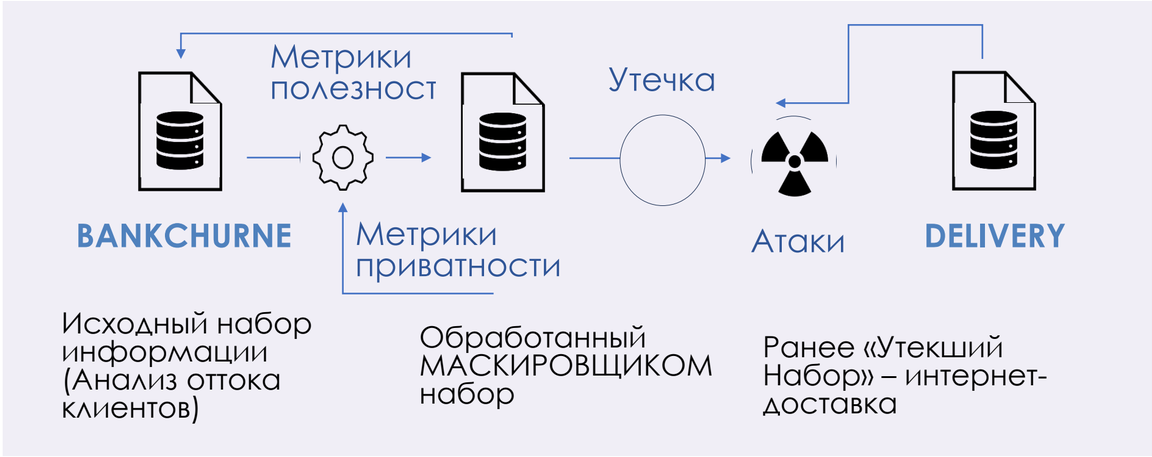
АТАКИ

- **k-Anonymity Model** (l-diversity, t-Closeness)
- **Модель контекстных рисков** Contextual Risk Assessment Model
- **Модель псевдонимизации** Resource-Based Risk Model
- **Модель информационной утечки** Composite Risk Model for Data Leakage
  - **Выделение** Singling Out
  - **Связывание** Linkage
  - **Вывод** Inference

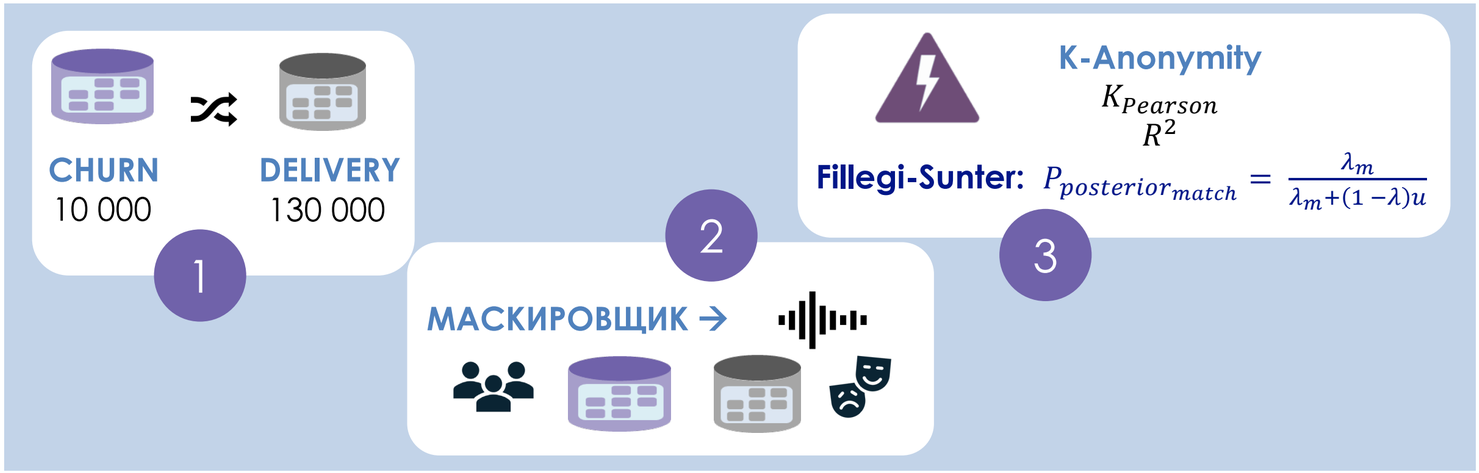
АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

КЕЙС А: УТЕЧКА БАНКОВСКОЙ ИНФОРМАЦИИ

ЦЕЛЬ КЕЙСА – протестировать АТАКУ СВЯЗЫВАНИЯ обезличенного банковского набора с ранее утекшим набором интернет-доставки



ПРИМЕНЕННЫЕ ПОДХОДЫ:



КЕЙС В: МАРКЕТИНГОВОЕ КАСАНИЕ

ЦЕЛЬ КЕЙСА – моделирование конфиденциального объединения данных рекламной площадки и набора с банковскими транзакциями для оценки эффективности рекламной кампании.

ПРИМЕНЕННЫЕ ПОДХОДЫ:

ИНФОРМАЦИОННАЯ УТЕЧКА (Выделение, Линкование, Вывод)

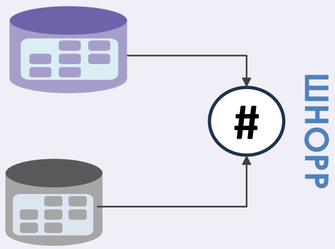
$K_{Pearson}$

KS-статистика

R^2

CVPL: кластерно – векторная атака

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



АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ПОКАЗАТЕЛИ ДО И ПОСЛЕ ПРИМЕНЕНИЯ РЕКОМЕНДАЦИЙ

КАЧЕСТВО И ПОЛЬЗА

Pearson Coefficient	0.9	0.8	- 0.1	Уровень корреляции снизился, но все еще остается высоким
KS-статистика 1	0.01	0.09	+ 0.08	Статистика ухудшилась, но незначительно
KS-статистика 2	0.27	1.7	+ 1.43	Увеличение разницы между распределениями
KL-дивергенция 1	2.77	2.85	+ 0.08	Расстояние между наборами увеличилось
KL-дивергенция 2	0.23	1.17	+ 0.94	Для защищенного варианта показатель ожидаемо вырос
R^2 детерминация	0.55	0.24	- 0.31	Для обновленной модели низкое значение
R^2 детерминация	0.99	0.83	- 0.16	На защищенных данных незначительное снижение

БЕЗОПАСНОСТЬ

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

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

РЕЗУЛЬТАТЫ КЕЙСА

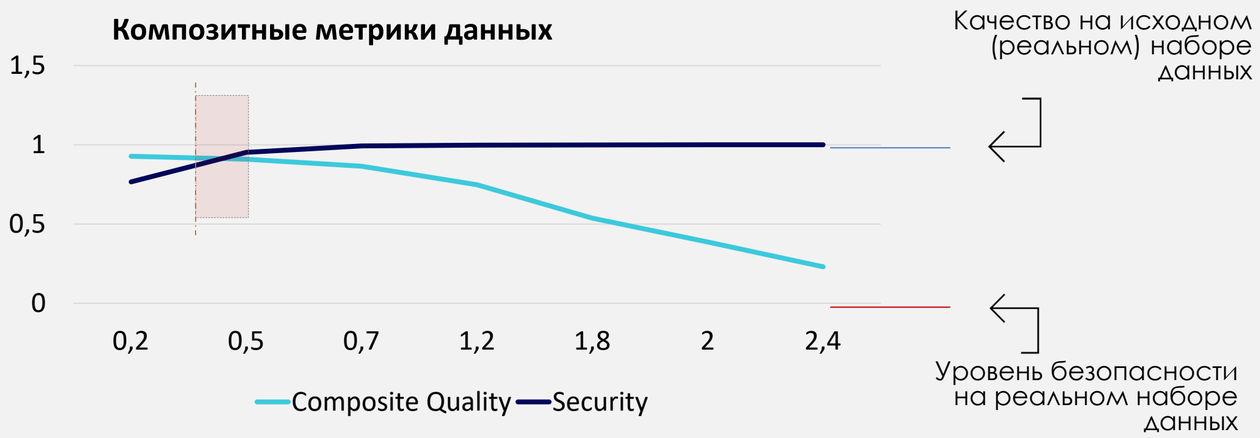
Применение мер защиты незначительно снизило качество данных, как по метрикам, так и по целевым бизнес-характеристикам. Ожидаемое соотношение – при увеличении шума снижаются метрики риска, включая риски атак (растет степень защиты), но также снижается качество)

Бизнес-метрики:

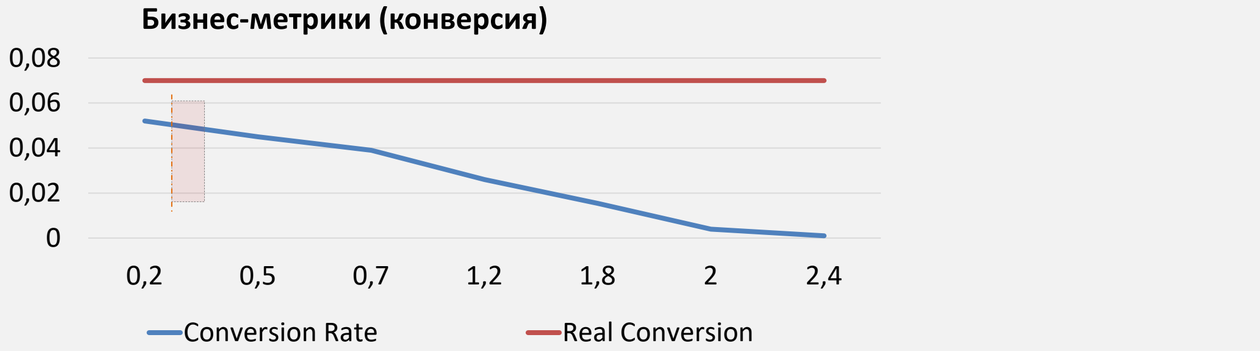
$$Conversion\ Rate\ = \left(\frac{Number\ of\ Conversions}{Total\ Number\ of\ Interactions}\right) \times 100\%$$

	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

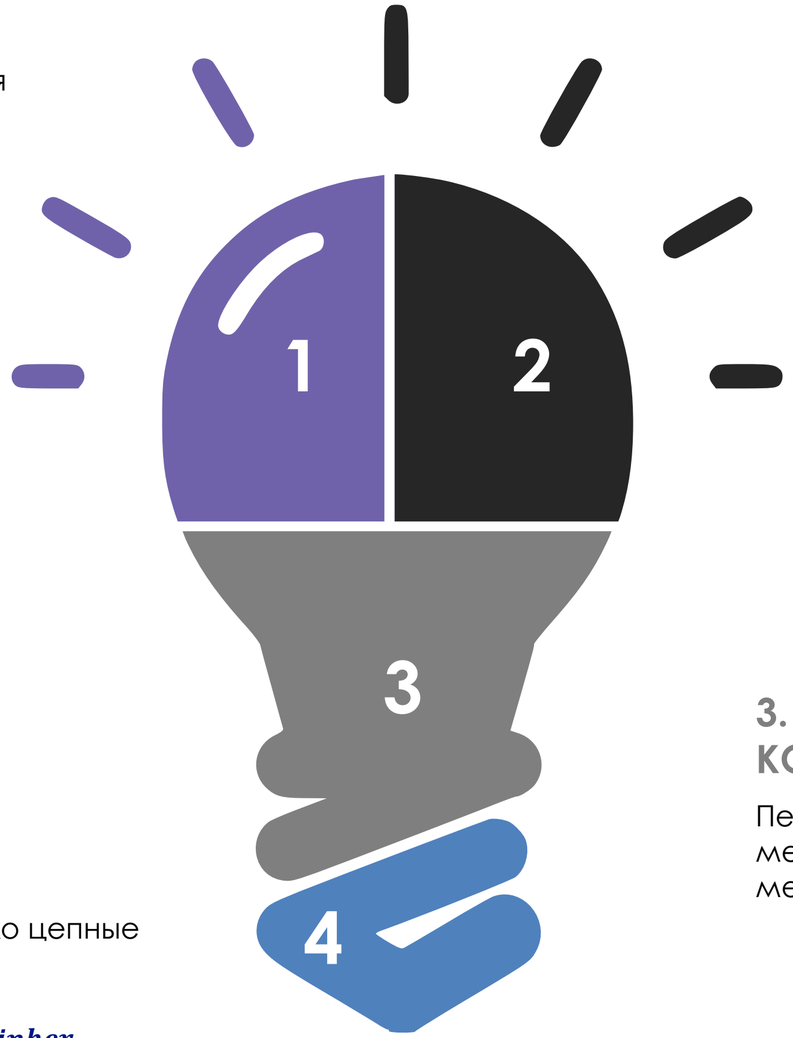


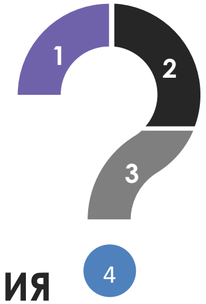
АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

ОТВЕТЫ НА ПОСТАВЛЕННЫЕ ВОПРОСЫ

1. K-ANONYMITY / LINKAGE SUCCESS

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

3. СЛОЖНОСТЬ КОНФИДЕНЦИАЛЬНЫХ ПЕРЕСЕЧЕНИЙ

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.

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

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

H F Labs

АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

Спасибо за внимание!
Вопросы?

CDI CONF 2024

Марат Тахавиев
m.takhaviev@rubda.ru