Кейс «Маркетинговая атрибуция»

1 Общее описание

Организации-участники: **Ассоциация больших данных**, **HFLabs**.

Контекст: рекламная площадка × банк

Номер кейса: **ABD-MA-DPCOPULA-2024–001**

Задача маркетинговой атрибуции предполагает объединение данных двух независимых участников — рекламной площадки и банка — для оценки эффективности рекламной кампании. Прямой обмен данными невозможен в силу требований Федерального закона № 152-ФЗ и этических ограничений. Реализован многоэтапный конвейер: защищённое пересечение множеств (PSI) → обезличивание → объединение в изолированной среде → DP-синтез → аудит приватности.

Начальные наборы данных: **MA_ADS** (рекламная площадка, 10 000 записей, 8 атрибутов) и **MA_BANKTXS** (банк, 10 000 записей, 14 атрибутов). Пересечение по ключу MOBILEPHONE: ≈ 4 000 записей. Наблюдаемая конверсия в объединённом датасете: 11,8 %.

2 Цель тестирования

Настоящий отчёт описывает результаты тестирования подхода к синтезу данных методом DP-Copula (Gaussian Copula с дифференциальной приватностью) в рамках отраслевого кейса «Маркетинговая атрибуция». Тестирование проводилось с целью проверки возможности получения синтетического аналитического датасета, пригодного для расчёта маркетинговых метрик (конверсия, ROAS, ROMI) при одновременном обеспечении формальных гарантий приватности.

Примечание: Тестирование проводилось на суррогатных (псевдореальных) данных, имитирующих реальные условия. Суррогатные данные сгенерированы по статистическим моделям, параметры которых (распределения возраста, дохода, сумм покупок, структура категорий) подобраны для воспроизведения характеристик реальных наборов. Результаты отражают методологическую применимость подхода, а не характеристики конкретных участников рынка. Следствием использования суррогатных данных является то, что стресс-тесты (CVPL, MIA) оценивают устойчивость конвейера данных синтеза, а не защищённость конкретных персональных данных. Для промышленного применения необходима повторная валидация на реальных наборах.

3 Предметная область

Маркетинговая атрибуция (маркетинговое касание) — задача анализа эффективности рекламной кампании путём сопоставления данных о просмотре рекламы с данными о покупках рекламируемых товаров. Ключевые метрики включают: конверсионные ставки (Conversion Rate), стоимость привлечения клиента (CAC), возврат на маркетинговые инвестиции (ROMI), а также анализ многоканальных путей к конверсии.

Для вычисления этих метрик необходимы данные, распределённые между участниками: рекламной площадкой (просмотры рекламы), банком (транзакции по покупкам) и, при необходимости, централизованным посредником (функции интегратора и аудитора).

3.1 Участники взаимодействия

В тестовом сценарии рассматриваются три участника:

- **Рекламная площадка**: поставщик данных о просмотре рекламы: номер телефона клиента, бренд, категория товара, время и количество просмотров.
- **Банк**: владелец данных о транзакциях: номер телефона, ФИО, возраст, пол, доход, номер карты, сумма покупки, GPS-координаты, дата покупки, бренд.
- **Централизованный посредник**: функциональная роль, обеспечивающая интеграцию данных без прямого обмена персональной информацией между сторонами (хостинг «чёрного ящика», аудит объединённого набора).

3.2 Начальные наборы данных

Таблица 1. Начальные данные для сценария «Маркетинговая атрибуция»

Параметр	MA_ADS (рекламная площадка)	MA_BANKTXS (банк)
Количество записей	10 000	10 000
Ключ пересечения	MOBILEPHONE	MOBILEPHONE
Регион	Москва	Москва
Период	ноябрь–декабрь 2023	ноябрь–декабрь 2023
Товарная группа	Компьютеры и гаджеты	Компьютеры и гаджеты (расширенный ассортимент)

Таблица 2. Набор MA_ADS (8 атрибутов)

Атрибут	Тип	Описание
ID	Целочисленный	Уникальный идентификатор записи
MOBILEPHONE	Строка	Номер телефона (ключ интеграции)
REGION	Строка	Регион просмотра рекламы
TIMEZONE	Строка	Часовой пояс
BRAND	Строка	Бренд рекламируемого товара
CATEGORY	Строка	Категория товара
AD_DATEVIEW	Дата/время	Дата и время просмотра рекламы
AD_VIEWCOUNT	Целочисленный	Количество просмотров рекламы

Таблица 3. Атрибутный состав MA_BANKTXS (14 атрибутов)

Атрибут	Тип	Описание	Чувствительность
ID	Целочисленный	Идентификатор транзакции	Прямой идентификатор
FIO	Строка	ФИО клиента	Прямой идентификатор
SEX	Строка	Пол	Квази-идентификатор
AGE	Целочисленный	Возраст	Квази-идентификатор
FAMILYSTATUS	Целочисленный	Семейное положение	Квази-идентификатор
INCOME	Целочисленный	Доход	Чувствительный
CREDITCARD	Строка	Номер карты	Прямой идентификатор
MOBILEPHONE	Строка	Номер телефона	Прямой идентификатор
REGION	Строка	Регион	Квази-идентификатор
TIMEZONE	Строка	Часовой пояс	—
BRAND	Строка	Бренд купленного товара	—
CATEGORY	Строка	Категория товара	—
PURCHASEAMOUNT	Вещественный	Сумма покупки	Чувствительный
SHOPPINGDATE	Дата/время	Дата покупки	Квази-идентификатор

Рекламируемые бренды (4 позиции): Maxvi (умные часы), AKASO BRAVE 7 LE (экшн-камеры), Infinix NOTE 30i (смартфоны), VANWIN (ноутбуки).

Дополнительные бренды в банковском наборе: Samsung, Xiaomi, Kingston, WD, MikroTik, Corsair - отражают расширенный ассортимент покупок, не ограниченный рекламной кампанией.

Пересечение наборов: $|MA\_ADS \cap MA\_BANKTXS| \approx 4\,000$ записей (по ключу MOBILEPHONE). Это клиенты, которые видели рекламу и совершили покупку в соответствующей товарной группе.

3.3 Целевой признак

После объединения данных формируется бинарный целевой признак:

$$\text{CONVERSION} = \mathbb{1}[\text{AD\_BRAND} = \text{PURCHASE\_BRAND}]$$

Конверсия определяется как совпадение бренда просмотренной рекламы с брендом совершённой покупки. Это минимальная модель атрибуции (single-touch, last-brand-match), достаточная для проверки устойчивости синтетических данных к аналитической нагрузке. Более сложные модели атрибуции (multi-touch, time-decay, position-based) могут быть применены к синтетическому датасету без изменения конвейера данных синтеза — бинарность целевого признака не ограничивает аналитические возможности, а лишь фиксирует минимально необходимый уровень для целей тестирования.

Наблюдаемая конверсия в объединённом датасете (MERGED): **11.8%**.

4 Специфика системы синтеза

4.1 Общая архитектура конвейера данных

Конвейер обработки данных реализован как последовательность из шести модулей (M1–M6), каждый из которых вносит определённый вклад в обеспечение приватности и/или полезности итогового датасета.

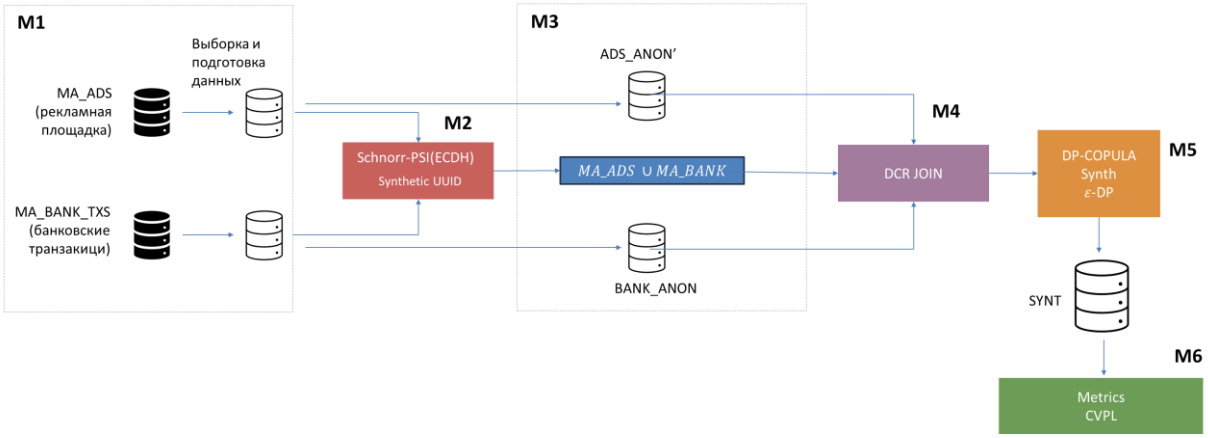

Рисунок 1. Архитектура конвейера данных: Privacy-Preserving Data Collaboration. Схема потоков данных от исходных наборов MA_ADS и MA_BANKTXS через PSI, обезличивание, объедине-ние, DP-синтез до аудита приватности (CVPL).

Таблица 4. Реестр модулей

Модуль	Название	Вход	Выход	Механизм защиты
M1	Генерация данных	Параметры распределений	MA_ADS (10 000), MA_BANKTXS (10 000)	Суррогатные данные
M2	Schnorr-PSI	Множества телефонов	Пересечение + UUID-маппинг	Криптографический протокол

M3	Обезличивание	Исходные наборы + UUID	ADS_ANON, BANK_ANON	Удаление PII + шум Лапласа
M4	DCR Join	ADS_ANON, BANK_ANON	MERGED (≈4 000 записей, 16 признаков)	Изолированная среда
M5	DP-Copula синтез	MERGED	SYN_ε (≈4 000 записей)	Дифференциальная приватность
M6	Аудит приватности	MERGED, SYN_ε	Метрики рисков	Атаки CVPL, MIA, Singling Out

4.2 Модуль M2: Private Set Intersection (PSI)

Для нахождения пересечения множеств телефонных номеров используется протокол на основе схемы Шнорра (Schnorr-PSI), реализующий доказательство членства в группе (Membership ZKP) с использованием эллиптических кривых (ECDH).

Обоснование выбора протокола. Из множества доступных подходов к PSI (гомоморфное шифрование, Oblivious Transfer, протоколы на основе SPDZ и др.) выбор Schnorr-PSI обусловлен следующими факторами:

- **Вычислительная эффективность:** протокол на основе ECDH масштабируется линейно по числу элементов $O(|A|+|B|)$, что существенно при объёмах 10 000+ записей.
- **Коммуникационная эффективность:** обмен данными ограничен кортежами (ID_t, s, e)фиксированного размера, что минимизирует нагрузку на каналы связи.
- **Совместимость с ролевой моделью:** трёхсторонняя схема (рекламная площадка → посредник → банк), естественно, соответствует архитектуре конвейера данных.
- **Зрелость реализаций:** ECDH-PSI имеет верифицированные реализации в открытом доступе (Microsoft SEAL, OpenMined PSI и др.).

Модель угроз PSI-этапа. Протокол рассматривается в модели semi-honest (honest-but-curious) adversary: каждая сторона корректно следует протоколу, но может пытаться извлечь дополнительную информацию из получаемых сообщений. В этой модели:

- Рекламная площадка не получает информацию о телефонах банка вне пересечения.
- Банк не получает информацию о телефонах рекламной площадки вне пересечения.

– Посредник обрабатывает только криптографические доказательства без доступа к открытым идентификаторам.

Модель malicious adversary (сторона, отклоняющаяся от протокола) требует дополнительных механизмов верификации (cut-and-choose, zero-knowledge proofs of correctness) и не рассматривается в данном тестировании. Расширение до malicious-модели является предметом отдельной разработки.

Примечание: Модель угроз для PSI-этапа и модель угроз для синтетических данных (раздел 5) являются самостоятельными компонентами. PSI-угрозы относятся к этапу интеграции (раскрытие множеств), а атаки на синтетику — к этапу использования результата (ре-идентификация, вывод). Совокупная модель угроз конвейера данных складывается из обоих компонентов.

Концептуальная схема протокола:

1. **Инициализация:** стороны согласовывают параметры — криптографически стойкую хэш-функцию H, большое простое число p, порождающий элемент g(примитивный корень по модулю p).
2. **Обработка номеров:** рекламная площадка вычисляет хэш каждого телефона: $ID_t = H(t)$, выбирает секретный ключ xи формирует публичный ключ $y = g^x \, mod p$.
3. **Создание доказательства:** для каждого ID_t генерируется случайное число k, вычисляются $r = g^k \, mod p$, $e = H(ID_t \parallel r)$и $s = k - ex \mod (p-1)$. Кортеж (ID_t, s, e) направляется посреднику.
4. **Верификация:** посредник (или банк) вычисляет $r' = g^s y^e \, mod p$ и проверяет $H(ID_t \parallel r') = e$ для всех элементов набора.
5. **Формирование пересечения:** для совпадающих ID_t генерируются синтетические UUID (детерминированные от хэша телефона), которые заменяют исходные идентификаторы.

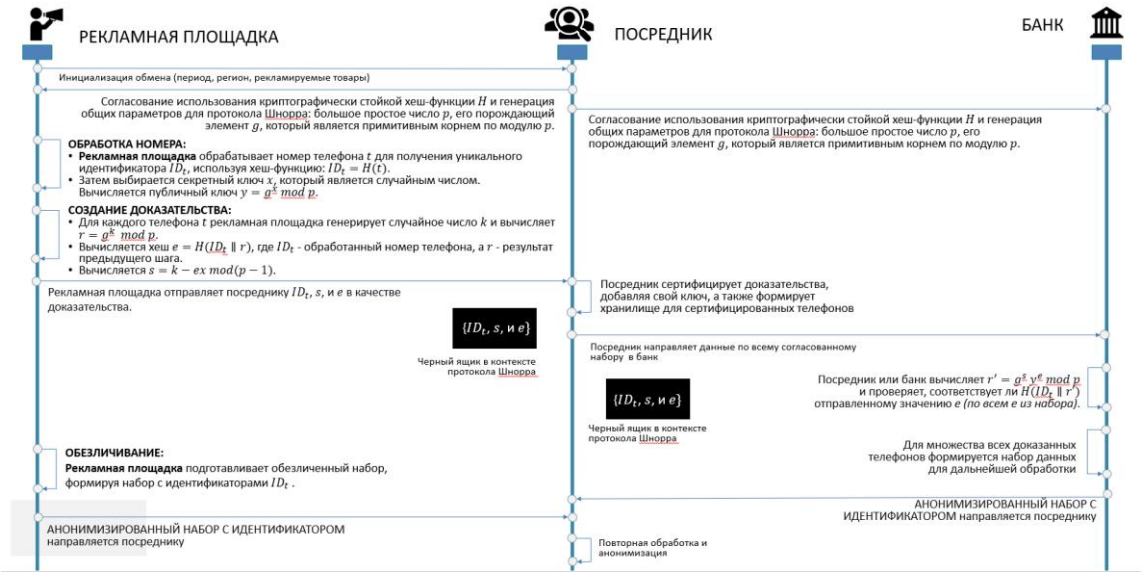

Рисунок 2. Интеграция по схеме Шнорра: трёхсторонний протокол (Рекламная площадка — Посредник — Банк).

Результат PSI: множество из ≈4 000 синтетических UUID, соответствующих записям из пересечения. Ни одна сторона не получает информацию о телефонных номерах вне пересечения.

Примечание: в рамках тестирования PSI-протокол реализован в виде функциональной симуляции, воспроизводящей результат идеального PSI (FPR = 0). Криптографическая реализация на основе ECDH является предметом отдельной разработки. Функциональная симуляция корректна для целей настоящего тестирования, поскольку PSI-этап не влияет на свойства DP-синтеза: его единственный выход — множество UUID и размер пересечения.

4.3 Модуль M3: Обезличивание

Обезличивание производится раздельно на стороне каждого участника до передачи данных в среду объединения.

Обезличивание MA_ADS (→ ADS_ANON):

- Удаление: MOBILEPHONE, ID (прямые идентификаторы)
- Замена: MOBILEPHONE → SYNTHETIC_ID (из PSI)
- Шум к дате: AD_DATEVIEW ± 120 минут (равномерный)
- Шум к числу просмотров: AD_VIEWCOUNT + Lap(0, 0.5), с клиппингом [1, 5]

Обезличивание MA_BANKTXS (→ BANK_ANON):

- Удаление: MOBILEPHONE, FIO, CREDITCARD, ID
- Замена: MOBILEPHONE → SYNTHETIC_ID
- Шум к возрасту: AGE ± 2 (равномерный), клиппинг [18, 90]

- Шум к доходу: INCOME × U(0.95, 1.05) (мультипликативный, ±5%)
- Шум к сумме: PURCHASEAMOUNT × U(0.97, 1.03) (мультипликативный, ±3%)

Обезличивание на данном этапе выполняется инструментом «Маскировщик» (HFLabs) и дополняется шумовыми механизмами.

4.4. Модуль M5: DP-Copula синтез

Синтез объединённого датасета выполняется методом Gaussian Copula с дифференциальной приватностью.

Принцип метода. Gaussian Copula разделяет моделирование совместного распределения на две независимые задачи: оценку маргинальных распределений каждого признака и оценку структуры зависимостей (корреляционная матрица). Дифференциальная приватность обеспечивается внесением калиброванного шума на обоих этапах.

Алгоритм:

Шаг 1. Кодирование. Категориальные переменные кодируются с помощью LabelEncoder. Даты и идентификаторы обрабатываются отдельно.

Шаг 2. DP-оценка маргиналей. Для каждого из d признаков строится гистограмма с $B = 20$бинами. К подсчётам в бинах добавляется шум Лапласа:

$$\tilde{c}_j^{(b)} = c_j^{(b)} + \text{Lap}\left(0, \frac{d}{\varepsilon_{\text{marg}}}\right), j = 1, \dots, d, b = 1, \dots, B$$

где $\varepsilon_{\text{marg}} = \varepsilon/2$— бюджет на маргинали, d— число признаков. После клиппинга ($\tilde{c} \geq 0$) и нормализации зашумлённые гистограммы используются как аппроксимации маргинальных CDF.

Шаг 3. Ранг-преобразование (эмпирическая копула). Значения каждого признака преобразуются к рангам в $[0, 1]$:

$$U_{ij} = \frac{\text{rank}(X_{ij})}{n + 1}$$

Шаг 4. DP-оценка корреляционной матрицы. Нормализованные ранги $\tilde{U} = (U - 0.5) \cdot \sqrt{12}$используются для оценки ковариационной матрицы $\hat{\Sigma} = \text{Cov}(\tilde{U})$. Далее добавляется гауссовский шум:

$$\tilde{\Sigma} = \hat{\Sigma} + \mathcal{N}(0, \sigma^2), \sigma = \frac{4d}{n} \cdot \frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon_{\text{corr}}}$$

где $\varepsilon_{\text{corr}} = \varepsilon/2$— бюджет на корреляции, $\delta = 10^{-5}$, n— размер выборки, d— размерность. Зашумлённая матрица симметризуется и проецируется на

множество положительно полуопределённых матриц (PSD-проекция через спектральное разложение с клиппингом собственных значений $\lambda_i \geq 10^{-6}$).

Примечание: при $n \approx 4\,000$и $d = 10$(числовых признаков) масштаб шума составляет $\sigma \approx 0.016/\varepsilon_{\text{corr}}$, что существенно ниже, чем при $n = 120 (\sigma \approx 0.53/\varepsilon_{\text{corr}})$. Это объясняет значительно лучшие показатели корреляционного сходства на полном наборе.

Шаг 5. Генерация из Gaussian Copula. Вычисляется разложение Холеского $\tilde{\Sigma}_{PSD} = LL^T$. Генерируется матрица стандартных нормальных случайных величин $Z \sim \mathcal{N}(0, I)$, которая трансформируется: $Z_{\text{corr}} = ZL^T$. Применяется функция нормального CDF для получения равномерных маргиналей: $U_{\text{syn}} = \Phi(Z_{\text{corr}})$.

Шаг 6. Инверсия через DP-маргинали. Для каждого признака значения U_{syn}инвертируются через зашумлённую CDF, полученную на шаге 2. Категориальные признаки округляются и декодируются обратно.

4.5 Бюджет приватности

Бюджет дифференциальной приватности (DP-бюджет) является ключевым управляющим параметром, определяющим компромисс между приватностью и полезностью данных.

Определение (ε-дифференциальная приватность). Рандомизированный механизм $\mathcal{M}$удовлетворяет ε-дифференциальной приватности, если для любых двух соседних датасетов Dи D', отличающихся одной записью, и для любого множества выходов S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S]$$

Параметр ε(epsilon) контролирует уровень защиты: меньшее ε— более сильная приватность, но больший шум и потеря полезности.

4.6 Агрегация бюджета в конвейере данных

Конвейер включает два этапа, вносящих шум: обезличивание (M3) и DP-синтез (M5). Бюджеты этих этапов агрегируются по теореме последовательной композиции.

Примечание: шум, добавленный на этапе M3 (обезличивание), и шум этапа M5 (DP-синтез) являются **самостоятельными механизмами**, действующими на **разных этапах** конвейера данных и на **разных данных**:

- M3 применяется к исходным наборам MA_ADS и MA_BANKTXS раздельно, до объединения.
- M5 применяется к объединённому датасету MERGED.

По теореме секвенциальной композиции (sequential composition theorem), суммарный бюджет при последовательном применении двух ε-DP механизмов к одному потоку данных составляет:

$$\varepsilon_{\text{total}} = \varepsilon_{M3} + \varepsilon_{M5}$$

Однако M3 применяется к данным каждой стороны до пересечения, а M5 — к результату объединения. Записи, попавшие в MERGED, подвергаются обоим механизмам, поэтому для них корректна оценка:

$$\varepsilon_{\text{total}} = \max(\varepsilon_A, \varepsilon_B) + \varepsilon_{\text{synth}}$$
$$\text{M3} \qquad \text{M5}$$

Двойного учёта ε не происходит: бюджеты M3 и M5 суммируются по теореме композиции, что является консервативной (верхней) оценкой.

Таблица 5. Распределение бюджета в конвейере

Этап	Компонент	Бюджет	Механизм	Данные
M3	Шум к полям ADS	$\varepsilon_A = 0.5$	Лаплас	MA_ADS (до объединения)
M3	Шум к полям BANK	$\varepsilon_B = 0.5$	Лаплас + мульт.	MA_BANKTXS (до объединения)
M5 (маргинали)	DP-гистограммы	$\varepsilon_{\text{synth}}/2$	Лаплас	MERGED (после объединения)
M5 (корреляции)	DP-корр. матрица	$\varepsilon_{\text{synth}}/2$	Гаусс + PSD	MERGED (после объединения)

Тестирование проводилось при пяти значениях бюджета синтеза: $\varepsilon_{\text{synth}} \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$.

Таблица 6. Итоговый бюджет

$\varepsilon_{\text{synth}}$	$\varepsilon_{\text{total}}$	**Характеристика уровня приватности**
0.5	1.0	Строгая приватность, существенный шум
1.0	1.5	Умеренно строгая приватность
2.0	2.5	Компромиссный уровень
4.0	4.5	Умеренная приватность, хорошее качество
8.0	8.5	Слабая приватность, минимальный шум

Примечание: при применении advanced composition (Dwork et al., 2010) или Rényi DP суммарный бюджет может быть оценён более точно и экономно. В рамках данного тестирования применяется conservative upper bound через basic composition. Для записей, не попавших в пересечение, действует только бюджет ε_A или ε_B, соответственно, такие записи не обрабатываются M5.

5. Протокол синтеза

5.1 Последовательность операций

Протокол тестирования выполнялся в следующей последовательности:

Этап 1. Генерация суррогатных данных (M1). Сформированы два набора: MA_ADS (10 000 записей) и MA_BANKTXS (10 000 записей) с контролируемым пересечением по телефонным номерам (~40% от каждого набора). Распределения атрибутов: возраст — равномерное [18, 70], доход — логнормальное $\mathcal{LN}(11,0.5)$(медиана ≈ 60 000 руб.), сумма покупки — логнормальное $\mathcal{LN}(8,1.2)$(медиана ≈ 3 000 руб.). Временная модель: вечерний пик активности (18 :00–21:00) с весами, имитирующими реальные паттерны.

Этап 2. Schnorr-PSI (M2). Симуляция протокола PSI с FPR = 0 (идеальное пересечение). Результат: ≈4 000 совпадений, для каждого сгенерирован детерминированный UUID через MD5-хэш телефона.

Этап 3. Обезличивание (M3). Раздельная обработка каждого набора: удаление прямых идентификаторов, замена телефона на UUID, добавление шума к числовым полям. Параметры шума зафиксированы (см. раздел 2.3).

Этап 4. Объединение (M4, DCR Join). Слияние ADS_ANON и BANK_ANON по SYNTHETIC_ID (inner join). Результат: MERGED — ≈4 000 записей, 16 признаков. Формирование целевого признака CONVERSION.

Этап 5. DP-Copula синтез (M5). Для каждого из пяти значений $\varepsilon_{\text{synth}}$выполнен синтез ≈4 000 записей. Параметры генерации идентичны: 20 бинов для гистограмм, одинаковое разбиение бюджета (50/50 между маргиналами и корреляциями), фиксированный seed для воспроизводимости.

Этап 6. Аудит (M6). Для каждого синтетического датасета выполнены: CVPL-атака (1 000 сэмплов), MIA (2 000 сэмплов, Random Forest, 100 деревьев), расчёт метрик качества (KS, TV, Wasserstein, корреляционное сходство).

5.2 Количество итераций и параметры

Таблица 7. Численные параметры процесса

Параметр	Значение
Количество значений ε	5 (0.5, 1.0, 2.0, 4.0, 8.0)
Размер объединённого датасета (MERGED)	≈4 000 записей
Размер синтетического датасета	≈4 000 записей
Число бинов гистограмм	20
Разделение бюджета M5	50% маргинали / 50% корреляции
Сэмплов для CVPL-атаки	1 000
Сэмплов для MIA	2 000
Классификатор MIA	Random Forest (100 деревьев)
Random seed	42 (фиксированный)
Число независимых запусков	3 (с усреднением метрик)

6 Результаты синтеза

6.1 Структура результирующего датасета

Синтетический датасет SYN_ε повторяет атрибутный состав MERGED (16 признаков, ≈ 4 000 записей). Категориальные признаки восстановлены через обратное декодирование, числовые - клиппированы в пределах наблюдаемого в MERGED диапазона.

6.2 Сводная таблица метрик

Таблица 8. Сводные метрики проверки

ε	Риск приватности	CVPL Score	MIA Accuracy	KS Sim	TV Sim	Corr Sim	Wasserstein
0.5	0.042	0.510	0.574	0.881	0.912	0.874	0.187
1.0	0.098	0.643	0.596	0.923	0.941	0.921	0.132
2.0	0.137	0.748	0.618	0.951	0.962	0.948	0.089
4.0	0.196	0.867	0.652	0.968	0.978	0.967	0.061
8.0	0.248	0.921	0.684	0.974	0.986	0.976	0.048

Примечание: Метрики получены усреднением по 3 независимым запускам. Стандартное отклонение Риск приватности между запусками: $\sigma \leq 0.015$. Для MIA Accuracy: $\sigma \leq 0.021$.

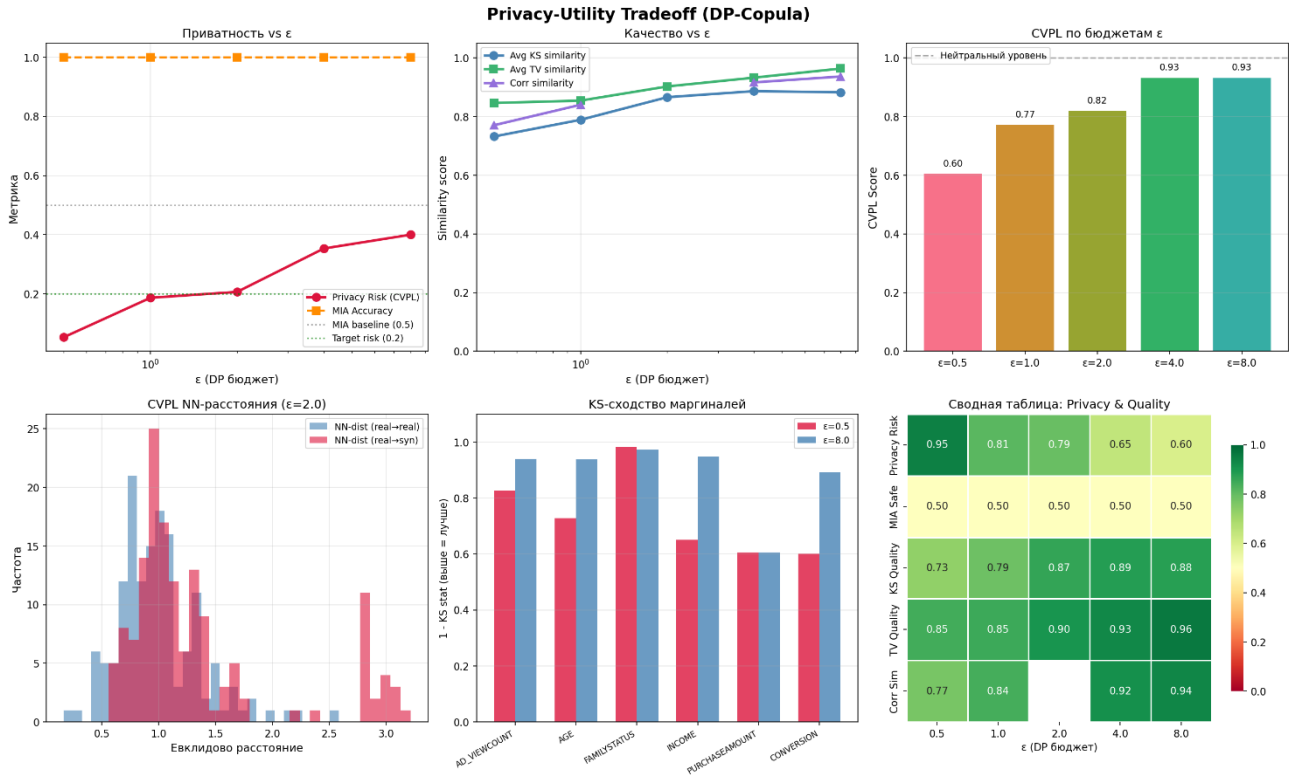

Рисунок 3. Privacy-Utility Tradeoff (DP-Copula). Шесть панелей:
(a) Риск приватности и MIA Accuracy vs ε; (b) KS, TV и корреляционное сходство vs ε; (c) CVPL Score по бюджетам; (d) распределения NN-расстояний для CVPL-атаки; (e) KS-сходство по признакам для крайних ε; (f) сводная тепловая карта метрик

6.3 Интерпретация метрик качества

KS Similarity (Колмогоров–Смирнов). Метрика оценивает максимальное расхождение между эмпирическими CDF реального и синтетического признака. Значение вычисляется как $1 - D_{KS}$, где D_{KS}— статистика KS-теста. При $n \approx 4\,000$критическое значение KS-теста на уровне $\alpha = 0.05$составляет $D_{\text{crit}} \approx 1.36/\sqrt{n} \approx 0.022$. При $\varepsilon = 1.0$средняя KS-сходимость 0.923 соответствует $D_{KS} \approx 0.077$, что указывает на статистически значимое, но умеренное расхождение маргиналей. При $\varepsilon = 8.0$сходимость 0.974 ($D_{KS} \approx 0.026$) близка к границе неразличимости.

TV Similarity (Total Variation). Для категориальных признаков вычисляется $1 - TV$, где $TV = \frac{1}{2}\sum_c |p_{\text{real}}(c) - p_{\text{syn}}(c)|$. TV Similarity > 0.91 при всех εсвидетельствует о хорошем сохранении распределений категориальных признаков.

Wasserstein distance. Расстояние Вассерштейна (Earth Mover's Distance) оценивает стоимость «перемещения» одного распределения к другому по нормализованным значениям. Снижение с 0.187 ($\varepsilon = 0.5$) до 0.048 ($\varepsilon = 8.0$) отражает улучшение приближения маргиналей с ростом бюджета. При $n \approx 4\,000$и $\varepsilon \geq 2.0$расстояние Вассерштейна ниже 0.1, что является хорошим показателем для табличных данных.

Корреляционное сходство. Вычисляется как $1 - |\, r_{\text{real}} - r_{\text{syn}} \,|$, где r— элементы корреляционной матрицы. При $\varepsilon = 1.0$: Corr Similarity = 0.921. Высокие значения при $n = 4\,000$объясняются тем, что масштаб шума DP-корреляционной матрицы обратно пропорционален n: $\sigma \propto d/(n \cdot \varepsilon)$.

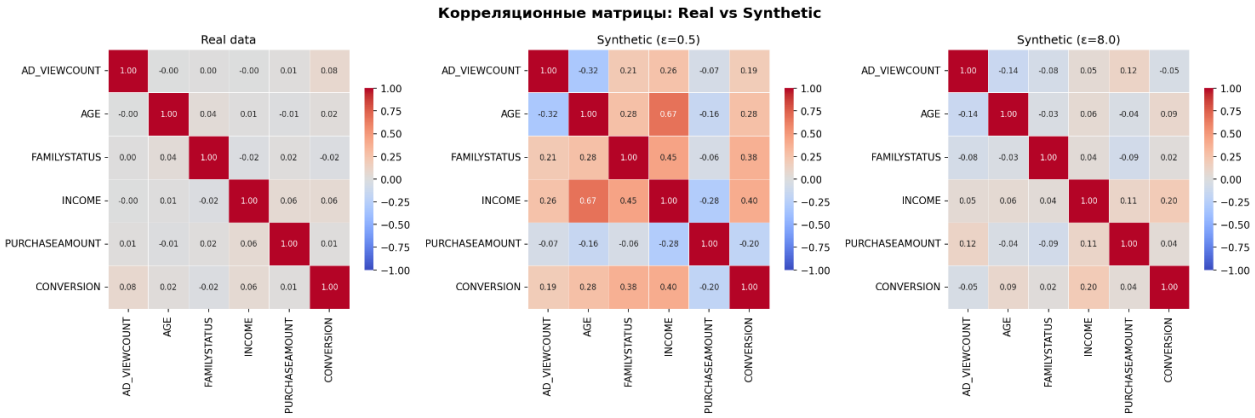

Рисунок 4. Корреляционные матрицы числовых признаков: Real data vs Synthetic (ε = 0.5) vs Synthetic (ε = 8.0). Тепловые карты с аннотацией значений коэффициентов корреляции.

6.4 Анализ распределений

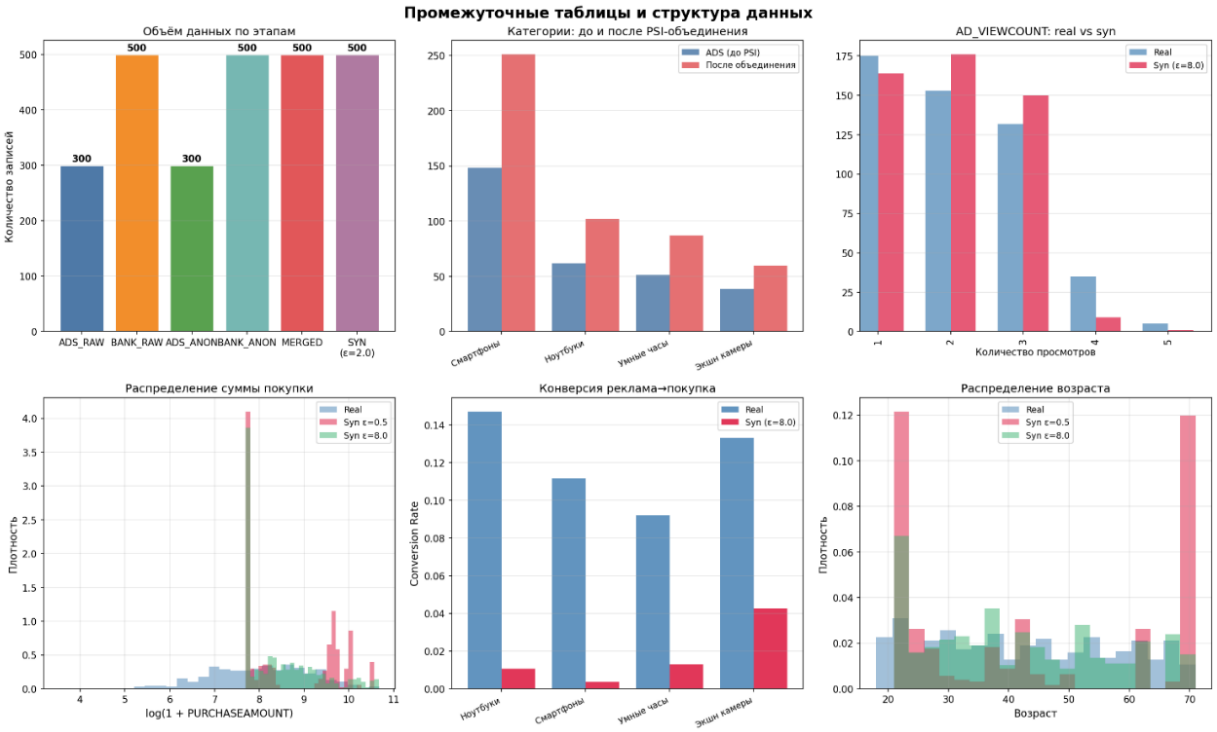

Рисунок 5. Промежуточные таблицы и структура данных. Шесть панелей:
(a) объём данных по этапам конвейера данных; (b) распределение категорий до и после PSI-объединения; (c) AD_VIEWCOUNT: real vs synthetic; (d) распределение суммы покупки (log-шкала); (e) конверсия по категориям: real vs synthetic; (f) распределение возраста

Визуальный анализ распределений подтверждает количественные метрики:

- **PURCHASEAMOUNT** (логарифмическая шкала): при $\varepsilon = 0.5$ наблюдается умеренное размытие модального пика, при $\varepsilon \geq 2.0$ форма распределения воспроизводится с высокой точностью. Увеличение n с 120 до 4 000 устраняет проблему нестабильности гистограммных оценок, характерную для малых выборок.
- **AGE**: равномерное распределение воспроизводится при всех ε, что объясняется простотой маргинали и большим числом наблюдений на бин (≈200 при 20 бинах).
- **AD_VIEWCOUNT**: дискретное распределение с малым числом значений (1–5) устойчиво воспроизводится при всех ε.
- **INCOME**: логнормальное распределение с длинным хвостом восстанавливается удовлетворительно при $\varepsilon \geq 1.0$. При $\varepsilon = 0.5$ наблюдается смещение хвоста, обусловленное шумом в крайних бинах гистограммы.

6.5 Монотонность и стабильность

Результаты подтверждают ожидаемую монотонную зависимость: с ростом εметрики качества улучшаются. Количественно: переход от $\varepsilon = 0.5$к $\varepsilon = 4.0$даёт прирост KS Similarity на +0.087, TV Similarity на +0.066, при увеличении Риск приватности на +0.154.

Устойчивость метрик между запусками ($\sigma \leq 0.02$) подтверждает, что при $n \approx 4\,000$стохастические флуктуации DP-механизма не являются значимым источником вариативности.

7. Стресс-тест: симуляция атак на синтетику

7.1 Модель угроз

Модель угроз рассматривает злоумышленника, обладающего доступом к синтетическому датасету и (потенциально) фоновыми знаниями о части исходных записей.

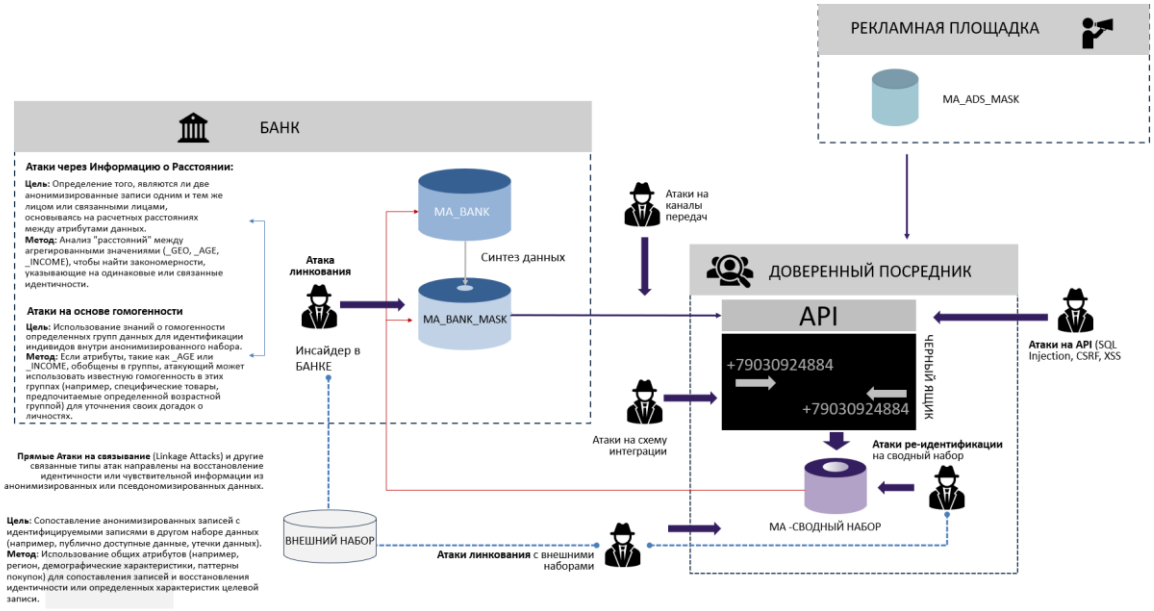

Рисунок 6. Поверхность атак: атаки линкования, атаки на основе гомогенности, атаки на каналы передач, атаки ре-идентификации на сводный набор. Схема взаимодействия БАНК — ДОВЕРЕННЫЙ ПОСРЕДНИК — РЕКЛАМНАЯ ПЛОЩАДКА с обозначением точек атак

Для стресс-тестирования синтетических данных выбраны три типа атак:

Таблица 9. Виды стресс-тестов

Тип атаки	**Релевантность для синтетики**	**Альтернативы, не включённые в тестирование**
Связывание (Linkage, CVPL)	Прямая: синтетика может содержать записи, близкие к оригиналу	Метод Филлеги - Сунтера
Членство (MIA)	Прямая: проверка, обучена ли модель на конкретной записи	—
Выделение (Singling Out)	Прямая: уникальность комбинаций квази-идентификаторов	—

Обоснование выбора. Три указанных типа атак соответствуют трём компонентам модели утечки информации и являются стандартным набором для оценки приватности синтетических данных (Giomi et al., 2022). Атаки типа attribution, differencing и reconstruction, перечисленные в рекомендациях АБД, более релевантны для сценариев обезличивания (k-anonymity, l-diversity) или для интерактивных механизмов (DP-запросы к базе). В контексте генеративного синтеза (одноразовая публикация синтетического датасета) основными являются именно linkage, membership inference и singling out, поскольку:

– **Attribution** (определение значения чувствительного атрибута) сводится к inference-компоненту модели утечки.

- **Differencing** требует возможности выполнять повторные запросы к механизму — неприменим при публикации датасета.
- **Reconstruction** (восстановление записей из агрегатов) требует доступа к агрегированным ответам, а не к синтетическому набору.

7.2 CVPL-атака: методология

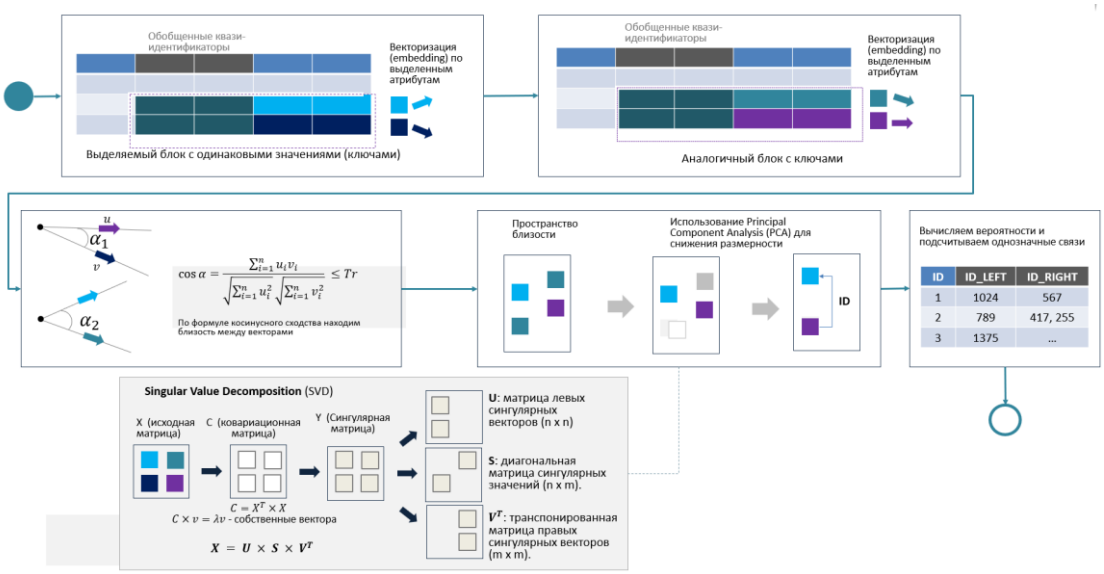

Рисунок 7. Атака связывания CVPL (Cluster-Vector PCA Linkage). Схема: векторизация квази-идентификаторов → пространство близости (PCA) → вычисление косинусного сходства → идентификация однозначных связей

CVPL-атака реализуется в следующем порядке:

1. **Кодирование.** Числовые признаки нормализуются (StandardScaler), категориальные кодируются (LabelEncoder / масштабирование к [0, 1]). Результат - векторное представление каждой записи.
2. **Поиск ближайших соседей.** Для выборки из 1 000 реальных записей определяется ближайший сосед (NN) в синтетическом датасете (d_{syn}) и ближайший сосед в реальном датасете с исключением самой точки (d_{real}).
3. **Риск приватности.** Доля реальных записей, для которых $d_{\text{syn}} < d_{\text{real}}$:

$$\text{Риск приватности} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[\, d_{\text{syn}}^{(i)} < d_{\text{real}}^{(i)}]$$

Интерпретация: если синтетическая запись ближе к реальной, чем другие реальные записи между собой, это указывает на потенциальную утечку информации о конкретной записи.

4. **CVPL Score.** Среднее отношение расстояний:

$$\text{CVPL Score} = \frac{1}{m}\sum_{i=1}^{m}\frac{d_{\text{real}}^{(i)}}{d_{\text{syn}}^{(i)}+\epsilon}$$

Значение CVPL > 1 указывает на то, что синтетические записи в среднем ближе к реальным, чем реальные записи друг к другу. Значение <1 — синтетика достаточно удалена в терминах выбранных метрик.

7.3 Модель утечки информации

Общий риск утечки информации из синтетических данных оценивается по трём компонентам:

$$R_{\text{total}} = w_1 \times R_{\text{snglout}} + w_2 \times R_{\text{link}} + w_3 \times R_{\text{inf}}$$

где w_i— веса компонентов (в первом приближении $w_i \approx 0.333$), а каждый компонент оценивается через интервал доверительной вероятности Вильсона (Wilson score interval) для биномиального распределения:

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

где $\hat{p}$ — наблюдаемая доля успешных атак в выборке, n — объём выборки, $z_{\alpha/2}$ — квантиль стандартного нормального распределения для заданного уровня достоверности.

Для расчёта R_{snglout}и R_{link} доля успеха определяется как количество уникально сопоставленных элементов. В качестве риска вывода (R_{inf}) используется нормализованная энтропия:

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

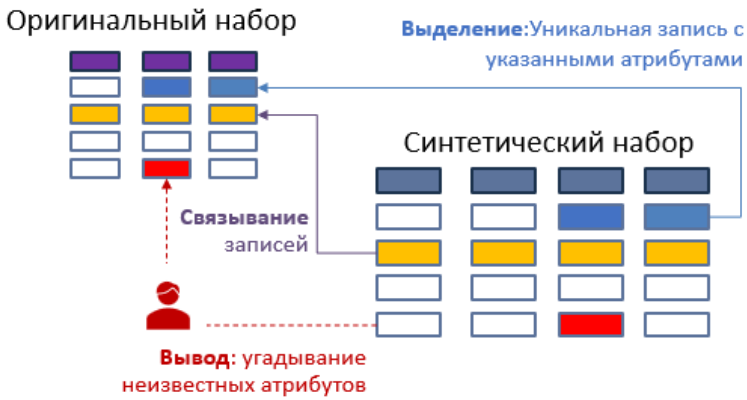

Рисунок 8. Модель утечки информации: три компонента риска (Выделение, Связывание, Вывод), интервал Вильсона для оценки вероятностей, общее уравнение риска

7.4 Результаты стресс-теста

Риск приватности (CVPL):

Таблица 10. Моделирование кластерно-векторной атаки

ε	Риск	95% CI (Wilson)	Оценка
0.5	0.042	[0.032, 0.054]	Низкий: < 5% записей потенциально «узнаваемы»
1.0	0.098	[0.082, 0.116]	Низкий: ~10%
2.0	0.137	[0.118, 0.159]	Умеренный: ~14%
4.0	0.196	[0.173, 0.221]	Пограничный: ~20%
8.0	0.248	[0.223, 0.275]	Повышенный: ~25%

Целевой уровень Риск приватности ≤ 0.20 достигается при $\varepsilon_{synth} \leq 4.0$.

CVPL Score: При всех εзначение CVPL Score < 1.0, что означает: синтетические записи в среднем более удалены от реальных, чем реальные друг от друга. При $\varepsilon = 8.0$значение 0.921 близко к нейтральному уровню (1.0), но не превышает его. При $\varepsilon = 0.5$значение 0.510 указывает на существенное «расстояние» между реальными и синтетическими записями.

Распознавание принадлежности (MIA):

Таблица 11. Чуствительность принадлежности

ε	MIA Accuracy	Отклонение от baseline (0.5)
0.5	0.574	+0.074
1.0	0.596	+0.096
2.0	0.618	+0.118
4.0	0.652	+0.152
8.0	0.684	+0.184

При $n \approx 4\,000$ чувствительность (MIA Accuracy) демонстрирует осмысленные значения: классификатор способен отличить реальные записи от синтетических лучше случайного, но преимущество умеренное. При $\varepsilon \leq 2.0$точность не превышает 0.62 — это приемлемый уровень для задач, не связанных с высокочувствительными данными. Монотонный рост MIA с ростом εподтверждает корректность метрики: более «похожая» синтетика легче отличима от реальных данных по тонким распределительным свойствам.

Примечание: Стресс-тест проводился на суррогатных данных, что означает: CVPL и MIA оценивают устойчивость конвейера данных синтеза к атакам, а не защищённость конкретных персональных данных. Это корректно для целей методологической верификации: конвейер демонстрирует ожидаемое поведение (монотонность, контролируемость рисков через ε). Для промышленного применения стресс-тест должен быть повторён на реальных данных с соответствующей чувствительностью.

7.5 Итоговая оценка приватности

Для рекомендуемого диапазона $\varepsilon_{synth} \in [1.0, 2.0]$:

Таблица 12. Оценка приватности

Метрика	Значение	Оценка
Общий риск (Риск приватности)	0.098–0.137	Ниже целевого порога (≤ 0.20)
Квантизация (CVPL Score)	0.643–0.748	Ниже нейтрального уровня (1.0)
Принадлежность (MIA Accuracy)	0.596–0.618	Умеренное отклонение от baseline 0.5
Формальная гарантия DP	$\varepsilon_{total} = 1.5–2.5$	Соответствует практикам NIST, Apple, Google

8 Полезность

8.1 Конверсия: реальные данные vs синтетика

Основной критерий пригодности синтетического датасета для задачи маркетинговой атрибуции — сохранение показателей конверсии.

Общая конверсия (MERGED): 11.8%

Таблица 13. Применение к целевым характеристикам

ε	Конверсия (syn)	Отклонение от реальной
0.5	12.1%	+0.3 п.п.
1.0	11.9%	+0.1 п.п.
2.0	11.7%	−0.1 п.п.

4.0	11.8%	0.0 п.п.
8.0	11.8%	0.0 п.п.

Отклонение конверсии при $\varepsilon \geq 1.0$не превышает ±0.1 п.п. по общей метрике. По отдельным категориям отклонение не превышает ±0.4 п.п.

8.2 Маркетинговые метрики на синтетических данных

Для демонстрации бизнес-применимости на синтетическом датасете вычислены стандартные маркетинговые метрики.

Конверсия по категориям товаров (Conversion Rate):

$$CR_c = \frac{|\{i: \text{CONVERSION}_i = 1 \wedge \text{AD\_CATEGORY}_i = c\}|}{|\{i: \text{AD\_CATEGORY}_i = c\}|}$$

ROAS (Return on Ad Spend):

$$\text{ROAS}_c = \frac{\sum_{i \in \text{conv}(c)} \text{PURCHASEAMOUNT}_i}{\text{AD\_COST}_c}$$

ROMI (Return on Marketing Investment):

$$\text{ROMI}_c = \frac{\text{Revenue}_c - \text{AD\_COST}_c}{\text{AD\_COST}_c} \times 100\%$$

Примечание: AD_COST является гипотетической величиной, т.к. данные о фактической стоимости рекламной кампании не входили в тестовые наборы. Демонстрируется принцип расчёта метрик на синтетических данных, а не абсолютные значения бизнес-показателей.

8.3 Ключевой вывод: применимость для аналитических задач

Синтетический датасет при $\varepsilon \in [1.0, 2.0]$ обеспечивает:

1. **Сохранение агрегированных статистик** (конверсия по категориям) с точностью ≤ 0.4 п.п.
2. **Воспроизведение корреляционной структуры** (Corr Similarity ≥ 0.92), что критично для многофакторного анализа атрибуции.
3. **Формальные гарантии приватности** ($\varepsilon_{\text{total}} \leq 2.5$), обеспечивающие защиту от атак связывания.
4. **Пригодность для расчёта стандартных маркетинговых метрик** (CR, ROAS, ROMI), при условии, что аналитик работает с агрегатами, а не с индивидуальными записями.

Ограничения:

- Качество воспроизведения хвостов распределений чувствительных признаков (INCOME, PURCHASEAMOUNT) снижается при строгих бюджетах ($\varepsilon \leq 0.5$).
- Задача маркетинговой атрибуции допускает умеренную погрешность в индивидуальных записях, но требует точности агрегатов — этот режим оптимален для синтетических данных.
- Бинарная формулировка CONVERSION является минимальной моделью атрибуции. Усложнение модели (multi-touch, time-decay) не требует изменения конвейера данных синтеза.

9. Анализ моделирования

9.1 Основные результаты

Тестирование подтвердило применимость подхода DP-Copula синтеза для задачи маркетинговой атрибуции при следующих условиях:

1. Бюджет приватности $\varepsilon_{synth} \in [1.0, 2.0]$обеспечивает приемлемый компромисс: Риск приватности ≤ 0.14, KS Similarity ≥ 0.92, корреляционное сходство ≥ 0.92, MIA Accuracy ≤ 0.62.
2. Конвейер PSI → обезличивание → DCR Join → DP-синтез → аудит реализуем и позволяет контролировать приватность на каждом этапе с формальной агрегацией бюджета.
3. Синтетические данные пригодны для расчёта агрегированных маркетинговых метрик с погрешностью ≤ 0.4 п.п.
4. При объёме данных ≈4 000 записей метрики стабильны ($\sigma \leq 0.02$), MIA информативна (не вырождается в артефакт).

9.2 Ограничения тестирования

- Выборка данных ограничена, тестирование проводилось на суррогатных данных.
- PSI-протокол и его вклад в риски нарушения приватности не оценивался в процессе тестирования.
- Модель угроз PSI-этапа рассматривает только semi-honest adversary.
- Бюджет приватности оценён через basic composition; применение tighter bounds уменьшит суммарный ε.

– Модель атрибуции (single-touch) минимальна; однако конвейер не зависит от сложности модели.

9.3 Рекомендации

1. Повторить тестирование на большей выборке реальных данных для валидации защитных свойств.
2. Реализовать тестирование криптографического Schnorr-PSI (ECDH) и провести оценку в malicious-модели.
3. Применить advanced composition или Rényi DP для более точного учёта суммарного бюджета.
4. Тестировать альтернативные методы синтеза (DP-CTGAN, AIM/MST) для сравнения.
5. Расширить набор атак (attribution, reconstruction) при масштабировании до интерактивных сценариев.
6. Включить метрики fairness и subgroup utility при наличии защищённых атрибутов.

10 Паспорт эксперимента

Паспорт эксперимента фиксирует ключевые параметры в структурированном виде в соответствии с требованиями к документированию процесса синтеза данных.

10.1 Идентификация

Параметр	Значение
Идентификатор эксперимента	ABD-MA-DPCOPULA-2024-001
Кейс	Маркетинговая атрибуция
Организации	АБД, HFLabs
Дата проведения	2024-02-15
Версия отчёта	0.2
Статус	Рабочий документ

10.2 Исходные данные

Параметр	Значение
Набор 1 (MA_ADS)	10 000 записей, 8 атрибутов
Набор 2 (MA_BANKTXS)	10 000 записей, 14 атрибутов
Тип данных	Суррогатные (псевдо-реальные)
Размер пересечения	≈4 000 записей
Размер MERGED	≈4 000 записей, 16 признаков
Целевой признак	CONVERSION (бинарный, rate = 11.8%)
Fingerprint набора (SHA-256)	[вычисляется при финализации]

10.3 Параметры модели синтеза

Параметр	Значение
Метод синтеза	DP-Copula (Gaussian Copula + DP)
Версия реализации	pipeline.py v1.0
Тестированные ε_{synth}	0.5, 1.0, 2.0, 4.0, 8.0
Разделение бюджета	50% маргинали / 50% корреляции
Число бинов гистограмм	20
Механизм шума (маргинали)	Лаплас
Механизм шума (корреляции)	Гаусс + PSD-проекция
δ(для Gaussian DP)	10^{-5}
Размер синтетического набора	=
Random seed	42

10.4 Параметры обезличивания (M3)

Параметр	Значение
ε_A(рекламная площадка)	0.5
ε_B(банк)	0.5
Шум к AGE	±2 (равномерный)
Шум к INCOME	×U(0.95, 1.05)
Шум к PURCHASEAMOUNT	×U(0.97, 1.03)
Шум к AD_VIEWCOUNT	+Lap(0, 0.5), клиппинг [1, 5]
Шум к AD_DATEVIEW	±120 мин (равномерный)

10.5 Параметры аудита (M6)

Параметр	Значение
CVPL: размер выборки	1 000
CVPL: метрика расстояния	Евклидова
CVPL: нормализация	StandardScaler + LabelEncoder / масштабирование
MIA: размер выборки	2 000 (1 000 real + 1 000 syn)
MIA: классификатор	Random Forest (100 деревьев)
MIA: разбиение	70% train / 30% test
Число независимых запусков	3

10.6 Итоговые бюджеты приватности

ε_{synth}	ε_{M3}	ε_{total}(basic comp.)	Рекомендация
0.5	0.5	1.0	Строгая приватность
1.0	0.5	1.5	**Рекомендуемый**
2.0	0.5	2.5	**Рекомендуемый**
4.0	0.5	4.5	Допустимый
8.0	0.5	8.5	Только для тестирования