CogniCrest

Anonymization Pipeline — Technical Benchmark Report

Fintech · KYC & Transaction Data for Fraud Analytics · synthetic data, generated for demonstration purposes

Important note: This report uses a fully synthetic dataset generated for demonstration purposes only — no real customers, patients, policyholders, or employees are involved, and no real company's data was used. This is a technical benchmark, not a client case study.

KYC & Transaction Data for Fraud Analytics

PROBLEM	A lending platform needs to run fraud models and share transaction data with a BI vendor — but raw KYC and account data can't leave the compliance boundary.
APPROACH	Tokenize identity fields, generalize transaction amounts into bands — enough signal for fraud scoring, without re-identification risk.
GOAL	Analytics keeps working. Compliance keeps a clean audit trail.

Methodology

A synthetic fintech dataset of 450 customer records was generated, following the end-to-end process common to financial data anonymization: ingestion, schema normalization, field classification, policy selection, privacy transformation, risk evaluation, utility evaluation, and a governance gate before delivery. Direct identifiers were tokenized. Quasi-identifiers (age, gender, region, employment type) were generalized using an adaptive suppression ladder until every record achieved k-anonymity of at least 5. Free-text service notes were screened for embedded names and addresses and redacted separately from the structured fields.

Field Classification

Fields are classified into four categories: direct identifiers, quasi-identifiers, sensitive attributes, and operational fields.

Field	Classification
customer_id	Direct identifier
full_name	Direct identifier
email	Direct identifier
phone	Direct identifier
street_address	Direct identifier
age	Quasi-identifier

gender	Quasi-identifier
zip_code	Quasi-identifier
employment_type	Quasi-identifier
risk_tier	Sensitive attribute
fraud_flag	Sensitive attribute
txn_amount	Sensitive attribute
service_note	Sensitive attribute (free text)
credit_score	Operational field
account_tenure_years	Operational field
txn_count_30d	Operational field / target

Text Redaction — Customer Service Notes

Customer service and dispute notes are a common privacy gap in financial data: names, addresses, and case details are often typed directly into free-text fields, bypassing whatever protection was applied to the structured columns. The pipeline screens narrative fields separately and redacts embedded direct identifiers before the note is included in any anonymized release.

BEFORE:

```
Spoke with Michael Chen at 2210 Maple Avenue, Austin regarding disputed transaction of $1,840.00. Risk tier: Medium. Follow-up required: No.
```

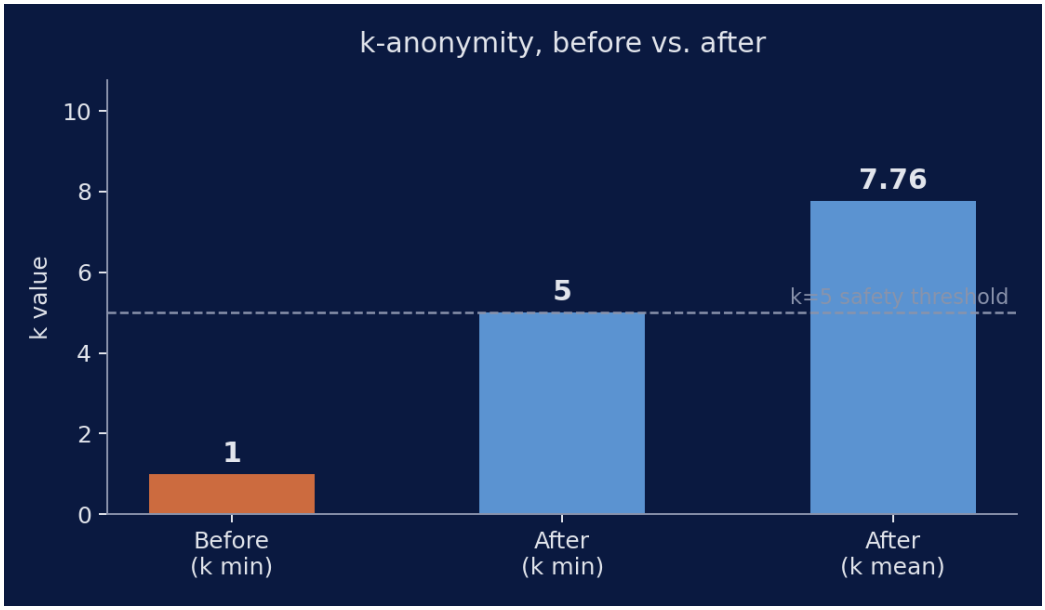
AFTER:

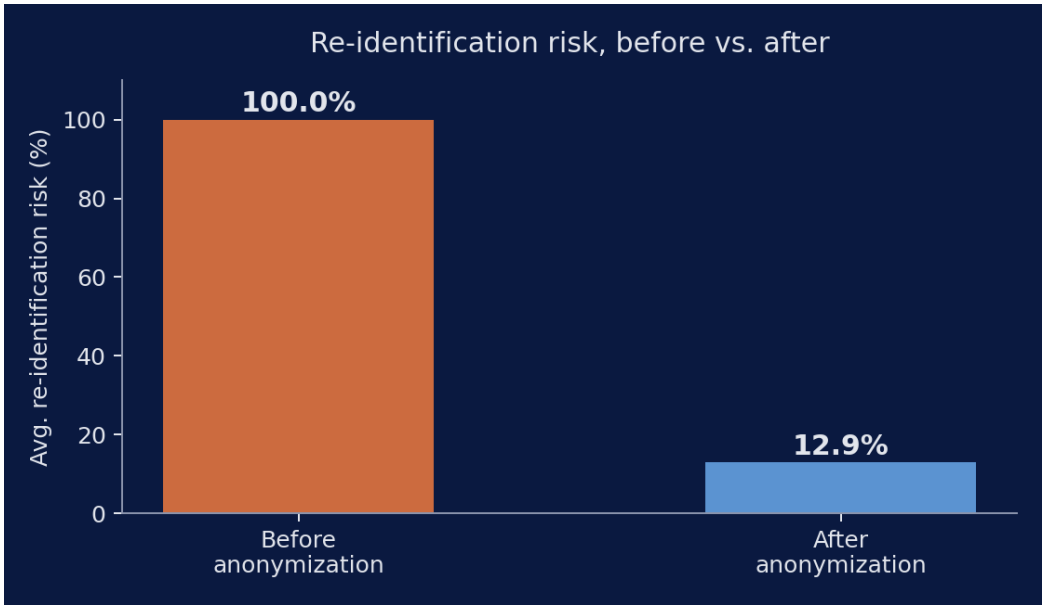
```
Spoke with [NAME REDACTED] at [LOCATION REDACTED], Austin regarding disputed transaction of $1,840.00. Risk tier: Medium. Follow-up required: No.
```

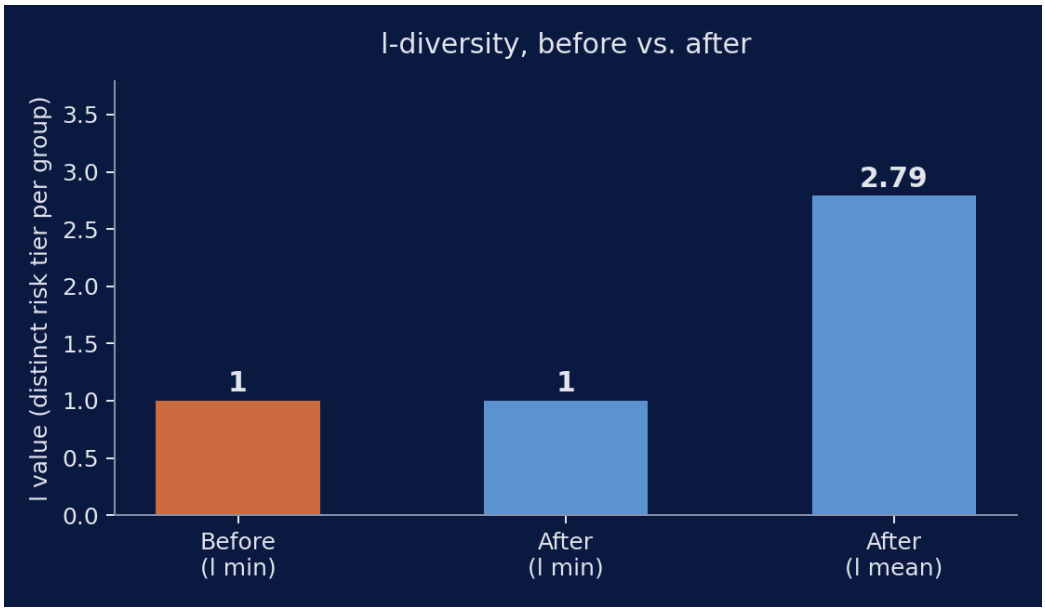
Note that city-level location ("Austin") was deliberately retained — it's already generalized to region elsewhere in the pipeline and provides useful signal for fraud analytics, while the specific street address (a direct identifier on its own) is removed.

Anonymity Metrics

Two metrics are reported together deliberately. k-anonymity alone can be satisfied while still leaking sensitive information: if every record in a group shares the same sensitive value, an attacker who narrows a target to that group learns the answer with certainty even without picking out the exact record. l-diversity checks for that specific gap by measuring how many distinct risk tier values exist inside each k-anonymous group.







Metric	Before	After
Minimum k (k-min)	1	5
Mean k (k-mean)	1.0	7.76
Average re-identification risk	100.0%	12.9%
Minimum l-diversity (l-min)	1	1
Mean l-diversity (l-mean)	1.0	2.79
Groups with l=1 (attribute disclosure risk remains)	100.0%	3.4%

Honest finding: 3.4% of equivalence classes still have l=1 after anonymization — meaning k-anonymity was satisfied for those groups but attribute disclosure risk on risk tier was not fully eliminated. This is a known, real limitation of k-anonymity alone, not a bug: it's exactly why both metrics are measured and reported together rather than stopping at k-anonymity and calling the dataset safe. A production release gate would flag these specific residual groups for additional generalization or exclusion before delivery to a recipient sensitive to attribute disclosure.

Utility Metrics

Coverage Retention

100% of input records are present in the anonymized output. Where a record's quasi-identifier combination couldn't reach k-anonymity through normal generalization, the pipeline generalized its fields further rather than deleting the row — no record is ever silently dropped.

KPI Fidelity

Transaction amounts are heavily right-skewed (true mean $1,227 driven up by a small number of large transactions; median is far lower). Two reconstruction methods were tested against this specific challenge:

Method	Reconstructed Mean	Error %
Naive band midpoint (no aggregate published)	$1,469.04	19.71%
Published bucket mean (aggregation release)	$1,227.21	0.0%

Honest finding: guessing from the band midpoint alone overestimates the true average transaction amount by 19.71% — even quantile-based bins leave some distortion because the top bin still spans a long tail of rare, large transactions. This is exactly why generalized row-level data (used for fraud-model training) and aggregate analytics releases (used for portfolio and BI reporting) are treated as separate outputs with separate techniques, not one dataset trying to serve both. Publishing the true per-bucket average alongside the generalized label — standard practice for aggregate reporting releases — reconstructs the KPI exactly, with no row-level values exposed.

CogniCrest — Privacy engineering for regulated data. This benchmark was run on synthetic data for demonstration purposes. Real client engagements are scoped individually; results will vary by dataset size, field composition, and regulatory requirements.