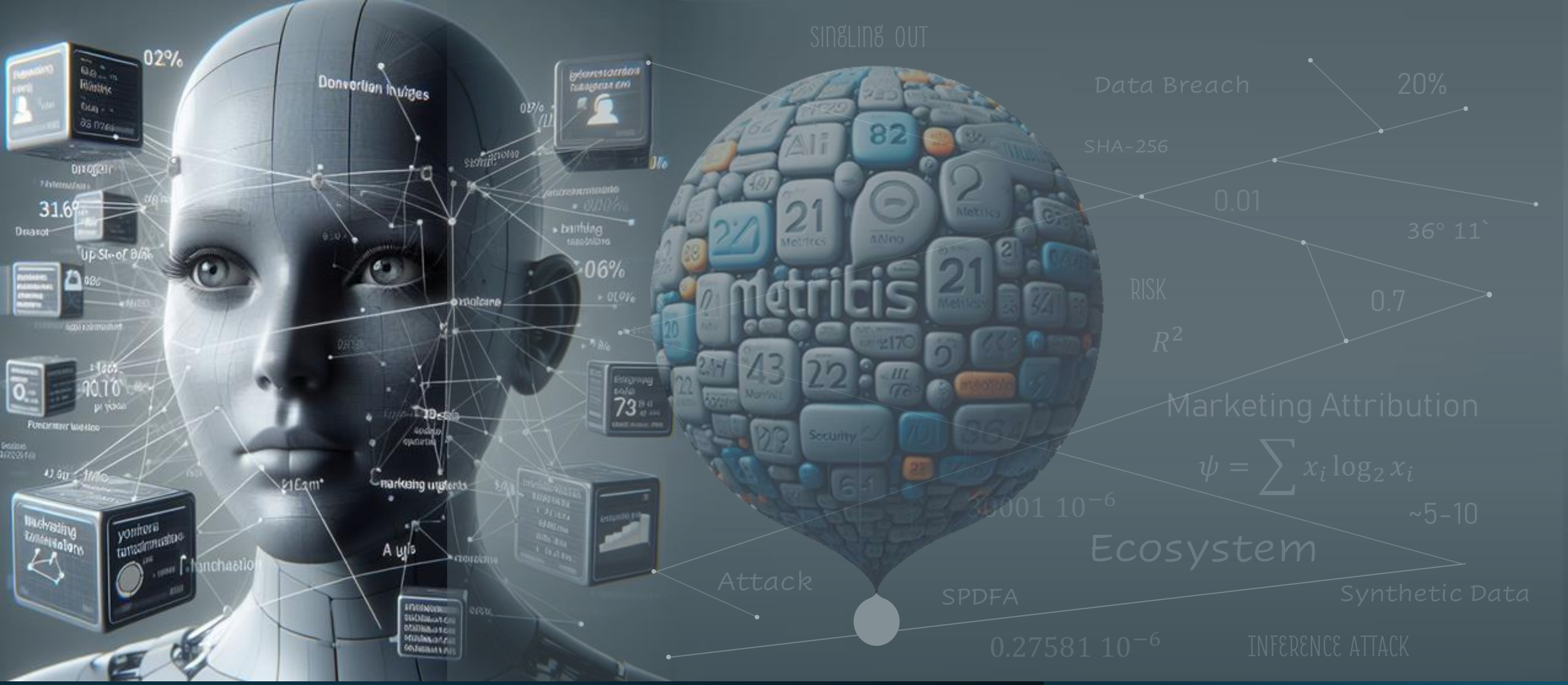


AI-BASED DATA BRIDGE

RE ALM

PRACTICAL CASE:

Privacy centric approach to apply in Marketing Attribution by Financial Organization

2024-02-28

CYBER AI IN FINTECH

According to PS Market Research, the global AI in cyber security reaching

$101.8 billion by 2030

When it comes to integrating Artificial Intelligence in banks for purposes such as forming ecosystems and marketing attribution, there are several ways it can impact key metrics like ARPU (Average Revenue Per User) and CCA (Cost of Customer Acquisition). However, there are also challenges related to privacy and cybersecurity

CHALLENGES

PRIVACY CONCERNS

DATA SECURITY

BIAS AND FAIRNESS

EXPLAINABILITY

RE ALM

DATA-BRIDGES: CATALYZING MICRO-ECOSYSTEMS

2024-02-29

In the landscape of modern business analytics, data-bridges serve as the foundational infrastructure that connects disparate data ecosystems, allowing for seamless and secure integration. These connections are pivotal in micro-ecosystems where collaborative AI models thrive on the synergy of combined data insights, all while maintaining stringent privacy standards.

BUSINESS DRIVEN APPROACH:

Marketing Attribution: In marketing, a data-bridge can link consumer data from multiple touchpoints (online ads, social media, email campaigns) to attribute sales or conversions to the correct marketing stimulus. This enables marketers to accurately measure and optimize the effectiveness of each channel.

Customer Relationship Management (CRM): Data-bridges can sync customer interactions from various platforms (support tickets, chat logs, transaction histories) into a unified CRM system, providing a 360-degree view of the customer for better service and targeted marketing.

Supply Chain Optimization: Data-bridges can connect information from suppliers, logistics, inventory, and sales to create a comprehensive supply chain model. This integration facilitates just-in-time inventory management, demand forecasting, and reduced operational costs.

Financial Services: In finance, data-bridges integrate information from trading platforms, risk management systems, and regulatory databases, allowing for real-time analytics, better risk assessment, and compliance tracking.

A **data-bridge** is a secure and controlled framework that facilitates the transfer and integration of data between independent systems or platforms, often operating within a micro-ecosystem. It ensures data integrity, confidentiality, and compliance with privacy regulations throughout the process.

A **micro-ecosystem** refers to a collaborative environment where various smaller, interconnected entities (such as departments, organizations, or systems) work together to achieve shared goals. Within these ecosystems, data sharing and processing are pivotal for collective intelligence and decision-making.

REALM APPLICATION: MARKETING ATTRIBUTION CASE

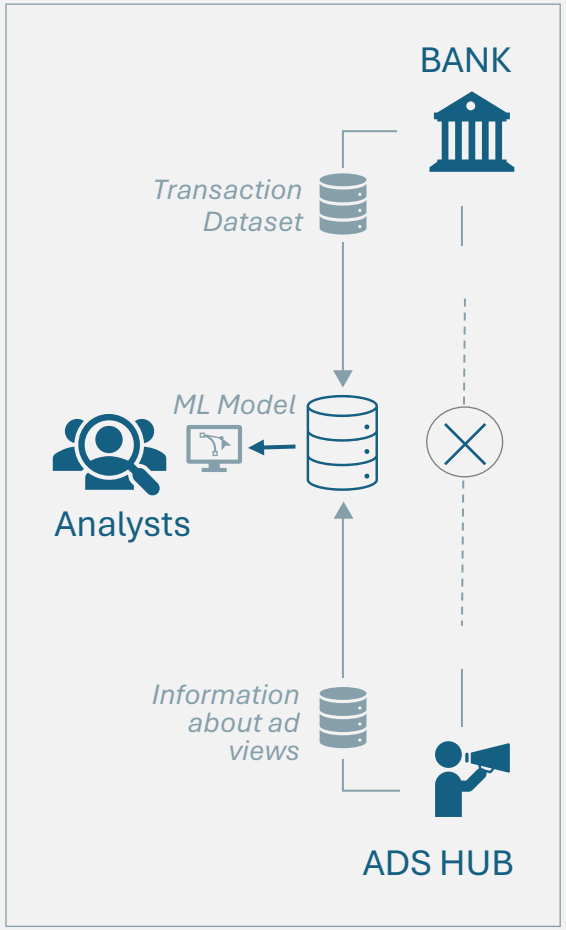
2024-02-28

The DGT REALM team conducted a significant AI security test on behalf of a European bank for a marketing attribution case. The primary goal was to securely merge data on advertising company activities and bank transactions to develop models for studying customer behavior. Due to legal restrictions, banks and advertising platforms (ad hubs) cannot combine their data, as it involves sensitive information

THE PROBLEM: Creating a shared dataset in a legal and confidential manner required a multi-stage process, including the preparation of training synthetic data and the implementation of an AI model to identify vulnerabilities.

CHALLENGES:

- Data Confidentiality: Legal restrictions on merging sensitive data.
- Multi-stage Processing: The need to create shared data in a legal and confidential manner.



SOLUTIONS AND METHODOLOGIES:

Data Integration via ZKP: Utilizing Zero-Knowledge Proof based on the Schnorr algorithm to protect privacy when merging data.

CVPL Algorithm for Vulnerability Detection: Developing a unique Cluster-Vector PCA LINKAGE (CVPL) algorithm to identify threats in large datasets.

Differential Privacy: Applying Laplace noise at the cluster level to generate safe synthetic data.

Advantages and Opportunities:

- **Micro-ecosystems**: Creating safe data bridges for advanced analytical analysis and AI application.
- **Quality and Risk Balance**: Generating synthetic data that balances quality with minimized risks.

WHY PRIVACY IS KEY TO UNLOCKING FINANCIAL AI

2024-02-28

In the digital landscape, merging data from banks and advertising hubs for marketing attribution is crucial yet challenging due to strict privacy regulations. Innovative solutions are needed to ensure data is used effectively while protecting consumer privacy

The REALM platform, supported by the PAMOLO component for data privacy, addresses these challenges by providing a secure environment for data merging, anonymization, and enrichment, ensuring that marketing attribution can be conducted safely and effectively.

Necessity for Advanced Merging Techniques

Data merging, especially using common identifiers like phone numbers, requires advanced techniques such as Private Set Intersection (PSI) or Trusted Execution Environments (TEE) to ensure privacy and compliance, as direct data exchange and exposure of sensitive information are prohibited.

Achieving Anonymization and Balance

Effective anonymization of data is essential before transmission to protect privacy. The challenge lies in maintaining data quality for marketing insights while ensuring absolute security and no traceability to the original data.

Enhancing Data with Learning Models

Learning models that enrich data offer a deeper understanding of consumer behavior. This process must safeguard privacy, underscoring the need for systems that manage data enrichment without compromising security.

MARKETING ATTRIBUTION

DECODING MARKETING ATTRIBUTION: DATA DYNAMICS

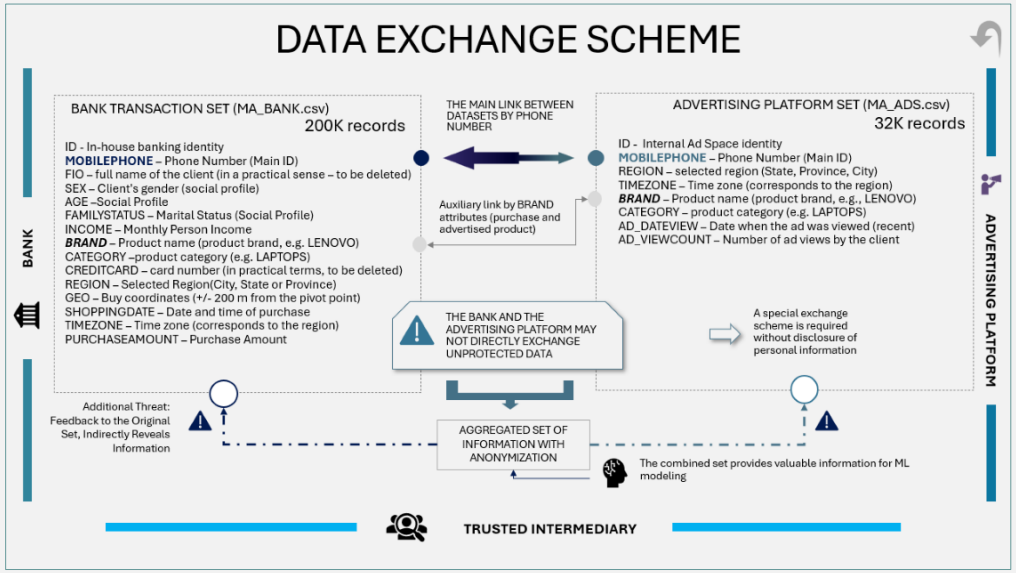
In an era where data is as valuable as currency, the quest to understand the customer journey from ad exposure to conversion has never been more critical. Yet, this quest comes with its fair share of challenges, particularly in the realms of privacy and data security. Marketing attribution stands at the crossroads of these challenges, demanding advanced solutions not for the sake of technology itself, but because the depth of insights and the precision it offers necessitates such sophistication.

The Mechanism of Marketing Attribution: Explains the process of tracking and attributing customer actions back to specific marketing stimuli, utilizing advanced analytics and machine learning models to decipher the impact of various advertising channels.

Benefits for Banks: Highlights enhanced customer insights, improved ROI on marketing spend, and refined targeting strategies, leading to increased customer satisfaction and loyalty.

Advantages for Advertising Hubs: Showcases the ability to demonstrate value to clients through precise impact measurement, fostering stronger partnerships and driving innovation in ad delivery and effectiveness.

Data Exchange Scenarios: Outlines various models for safe and compliant data sharing, including aggregated data insights, anonymized customer behavior patterns, and secure, privacy-preserving techniques like differential privacy and federated learning, to enrich attribution models without compromising individual privacy.



Employing cutting-edge technologies is not merely a choice but a necessity to navigate the delicate balance between leveraging data for actionable insights and safeguarding the privacy of individuals. This case underscores why it's both reasonable and essential to deploy advanced technological frameworks, illuminating the path to not just more effective marketing strategies, but also to a future where data privacy and utility coexist harmoniously.

IDENTIFYING THREATS IN DATA EXCHANGE SCHEME

2024-02-29

As we architect secure frameworks for data exchange between financial institutions and advertising platforms, understanding the myriad of potential threats becomes paramount. The complexity of integrating diverse datasets, such as those in marketing attribution schemes, introduces a broad spectrum of vulnerabilities—from the technology underpinning the integration to the data in transit.

Threat Model Complexity: Discusses the necessity of constructing a comprehensive threat model that considers both internal (insiders) and external (third-party) risks, highlighting the multifaceted nature of potential security breaches.

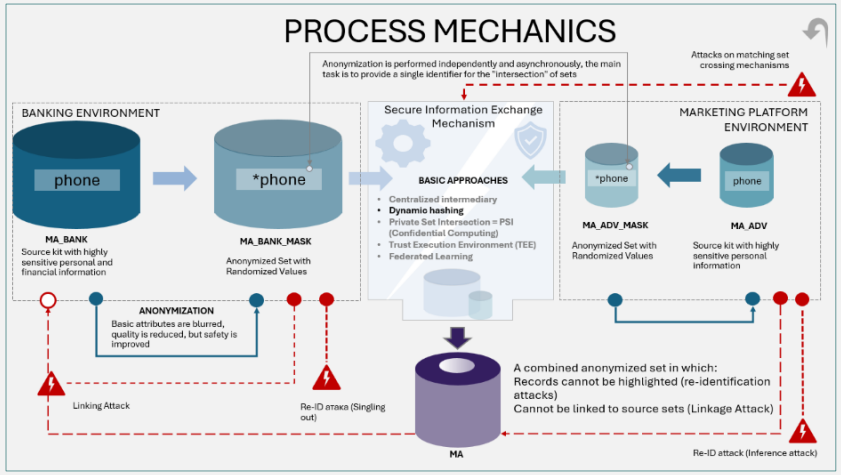
Attack Targets and Mechanisms:

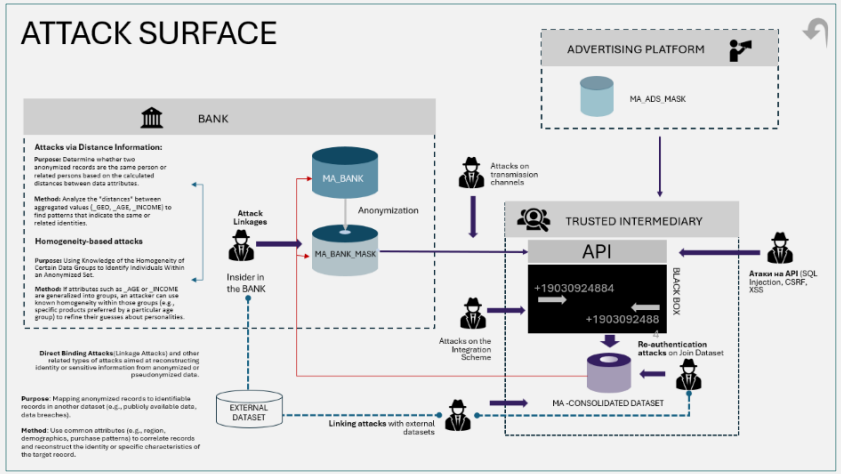
- Outlines the primary targets for attacks, including the integration schemes like PSI (Private Set Intersection) and TEE (Trusted Execution Environment), the merged marketing attribution (MA) dataset, transmission channels, and APIs utilized in data exchange.
- Illuminates the techniques attackers might use, such as exploiting background knowledge, leveraging external information leaks, and engaging in social engineering tactics.

The Back Link Concern: Raises the critical issue of preventing a feedback loop between the combined AI training set and the original banking transaction dataset, emphasizing the challenge of maintaining data integrity and privacy without compromising data utility.

Mitigation Strategies:

- Suggests rigorous testing and preparation of AI learning models to counteract linkage attacks, ensuring the application of differential privacy techniques does not dilute the granularity of data needed for accurate marketing attribution.
- Advocates for continuous evaluation and enhancement of security measures to protect against evolving threats and vulnerabilities, emphasizing the dynamic nature of cybersecurity in the context of financial AI applications.

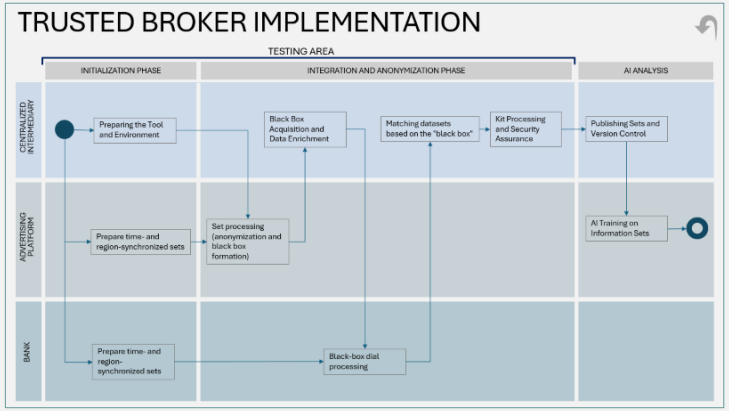


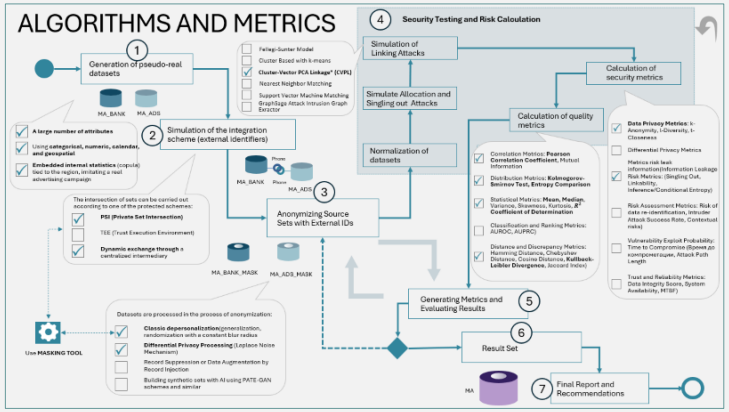


A COMPREHENSIVE TESTING FRAMEWORK

Our comprehensive case testing framework is designed to tackle these challenges head-on, employing a series of innovative tasks aimed at generating pseudo-real data, ensuring dynamic integration, and enhancing data anonymity. This approach not only aims to mitigate the risk of sensitive information leakage but also to refine the balance between data utility and security, paving the way for more reliable and secure marketing attribution practices..

- **Pseudo-Real Data Generation:** Develop synthetic datasets that closely replicate real-world consumer behaviors and marketing dynamics, facilitating risk-free testing environments.
- **Dynamic Centralized Intermediary Integration Scheme:** Explore administrative and agent-based architectural solutions to overcome limitations inherent in traditional data integration methods like PSI and TEE.
- **Anonymization Through Randomization:** Implement randomization techniques to obscure data, transitioning from simple constant blur to sophisticated differential privacy mechanisms using the Laplace distribution.
- **CVPL Attack Simulation:** Conduct simulations of potential attacks on anonymized datasets using Cluster-Vector PCA Linkage to pinpoint vulnerabilities and evaluate the risk of information leakage.
- **Security Metrics Calculation:** Apply advanced metrics such as k-anonymity, l-diversity, and t-proximity to gauge the robustness of data protection measures against potential identification and inference attacks.
- **Quality Score Evaluation:** Utilize statistical and marketing attribution quality metrics to assess the impact of data aggregation on the integrity and reliability of marketing insights.
- **AI Model Training for Vulnerability Detection:** Engage in iterative training of AI models to identify vulnerabilities, fine-tuning anonymization techniques and algorithmic parameters to enhance the security-quality balance.
- **Building Recommendations and Conclusions:** Analyze testing outcomes to formulate strategic recommendations aimed at optimizing the data consolidation process, thereby minimizing information leakage risks and aligning with marketing attribution goals.





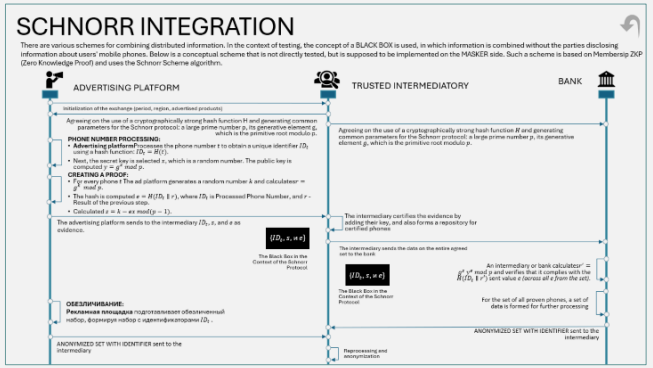
SECURITY ARCHITECTURE

2024-02-29

In the pursuit of robust data protection, particularly in the sensitive interplay between financial services and advertising, deploying comprehensive security schemes is not just a regulatory mandate but a strategic imperative. These schemes are designed to not only uphold a high level of data reliability, as confirmed by stringent metrics, but also to ensure that data quality remains uncompromised.

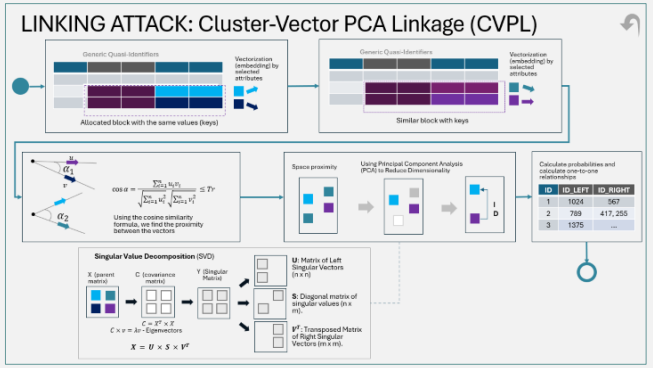
Integration Scheme Using ZKP by Schnorr:

- **Confidential Verification**: Leverage Schnorr's ZKP to prove membership of a phone number in a dataset without exposing the actual number or any associated sensitive details.
- **Dual Anonymization**: Employ ZKP for intermediary verification between banks and advertising platforms, thereby safeguarding personal data and ensuring adherence to privacy laws.
- **Breach Risk Reduction**: Utilize the inherent properties of ZKP to enable data verification that significantly reduces the possibility of user information leaks.



Anonymization Scheme:

- **Variable Radius Randomization**: Apply a variable blur radius to data attributes like age, location, and transaction details to obstruct user identification.
- **Differential Privacy Implementation**: Inject controlled noise into datasets to achieve differential privacy, thereby preventing individual identification, even in the presence of auxiliary information.
- **Suppression of Identifiers**: Eliminate or obfuscate direct identifiers to avoid the direct tracing of data to individuals.



Employing CVPL Attack Model for Testing:

- **Identifying Vulnerabilities**: Use the CVPL attack model to probe the anonymization measures, gauging the resilience of datasets against re-identification attacks.
- **Enhancing Data Defense**: Analyze CVPL attack outcomes to pinpoint and fortify weaknesses within anonymization and differential privacy frameworks.
- **Trust in Data Integration**: Showcase resistance to CVPL attacks to foster confidence in the data management and protective measures implemented for marketing attribution purposes.

AI Model Training for Attack Recognition:

- **Adversarial Network Training**: With insights into potential cluster vulnerabilities and weights, train an adversarial network to recognize and respond to simulated attack patterns and attack graph utilization.

MARKETING ATTRIBUTION

THE QUALITY MODEL

2024-02-29

Solving the problem of integrating data between different sources for marketing attribution requires a careful balance between ensuring data security and maintaining high quality insights. A high level of safety inevitably leads to a loss of quality, and a balancing scheme is used to find a balance.

Estimating Entropy Change: Comparison of the entropy of each column before and after anonymization to assess the degree of data loss or retention of information.

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 P(x,y)$$

Here, $p(x,y)$ is the joint probability that $X = x, Y = y$

Correlation Analysis: Comparison of Pearson correlation coefficients for pairs of variables before and after anonymization to assess how the anonymization process affects the structural relationships between variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Application of the Kolmogorov-Smirnov test: Use a test to compare the distributions of data before and after anonymization to test how much anonymization changes the statistical properties of the data.

$$D_{n,m} = \sup_x \left|F_{1,n}(x) - F_{2,m}(x)\right|$$

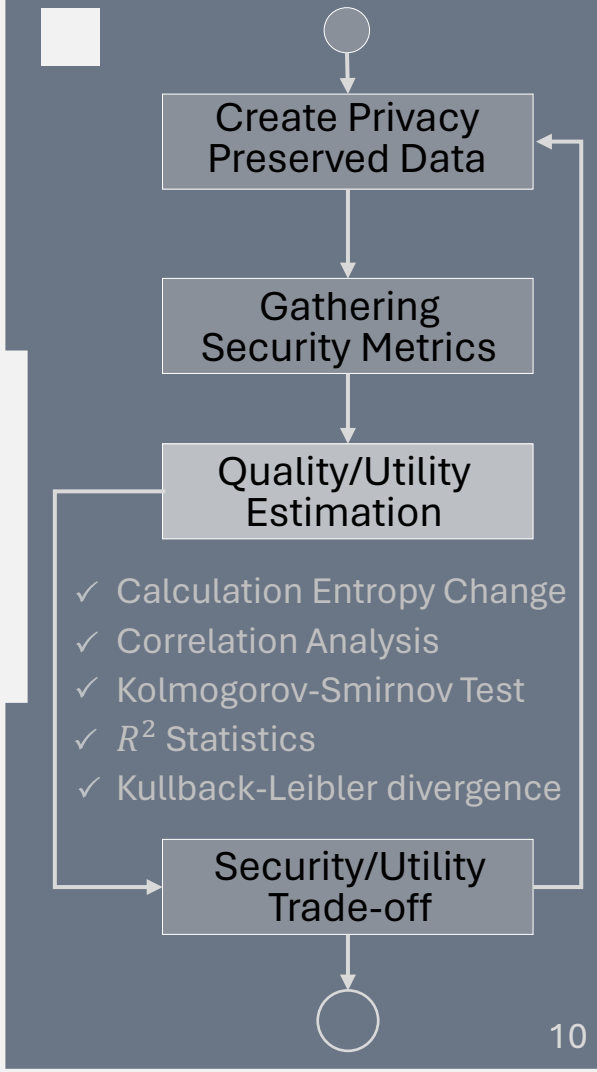
Calculating R^2 statistics (R-squared): The R^2 metric, also known as the coefficient of determination, explains the variation in the dependent variable in the context of comparing tables before the transformation (the original is MA_BANK) and after the transformation. The value R^2 ranges from 0 to 1. The closer R^2 is to 1, the better the model fits the data

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \text{In here: } ss_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \Big| \quad SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i - Observed value of the dependent variable for the i-th observation
$\hat{y}_i$ - Predicted (calculated) value of the dependent variable
$\overline{y}$ - mean value for the dependent variable

Kullback-Leibler divergence (KL divergence): It is a measure of the difference between two probability distributions. It shows how different one distribution is from another.

$$[KL(p||q) = \sum_i p_i \cdot \log(\frac{p_i}{q_i})]$$



RESULTS AND CONCLUSIONS

Our rigorous modeling of attack scenarios and quality-risk assessments has led to significant advancements in securing marketing attribution data. By simulating realistic data scenarios and refining protection measures, we've achieved a delicate balance between data utility and security. .

Pseudo-Real Data Generation: Successfully created pseudo-real datasets with embedded statistics for robust hypothesis testing, addressing dual-layer security challenges such as data integration and coupling attacks.

Primary Data Processing with MASKER:

- Utilized MASKER for data randomization, achieving high data quality with a KL divergence of ~0.2 and preserving Pearson correlation at ~0.9.
- Ensured data stability with only a maximum of 5% of records vulnerable to allocation attacks, demonstrating robust initial security.

Advanced Attack Technique Application (CVPL):

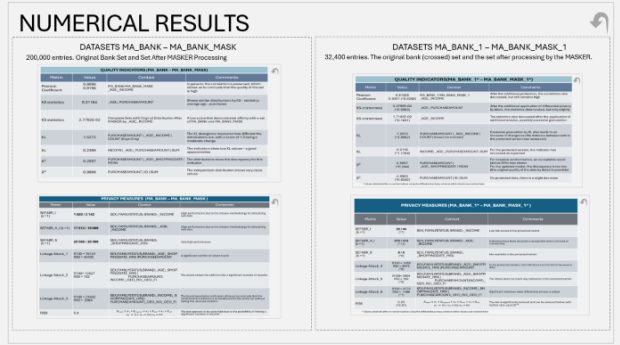
- Implemented sophisticated CVPL attack techniques to uncover feedback relationships between datasets post-randomization, effectively reducing the risk level to R=0.4 in a complex risk model.
- Highlighted the importance of non-trivial attack simulations as a means to rigorously test and improve defense systems.

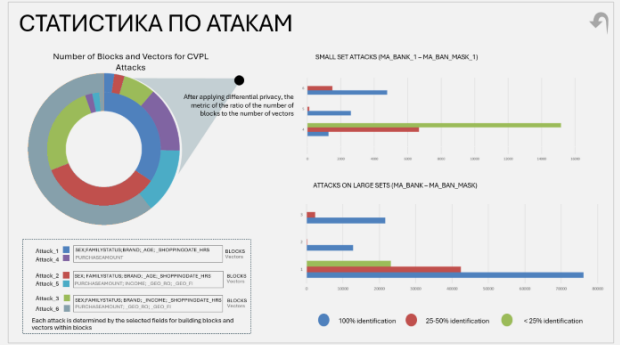
Re-application of Protective Measures:

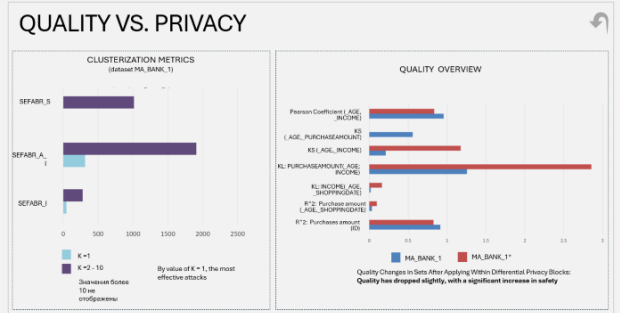
- Enhanced security further by applying differential privacy to specific clusters, reducing risk levels to 0.01 or lower with an increase in Laplace noise.
- Quality metrics remained largely stable, with only a slight increase in KL divergence, suggesting its potential as an indirect risk indicator for these operations.

Development of the Secure Integration Method: Established a secure integration method incorporating the Zero-Knowledge Proof (ZKP) mechanism via the Schnorr algorithm and a centralized intermediary model, with a keen eye on contextual risks.

- **Refinement Potential:** Identified opportunities for refining results by adjusting parameters and environmental constraints, indicating a path forward for continuous improvement in data security protocols.





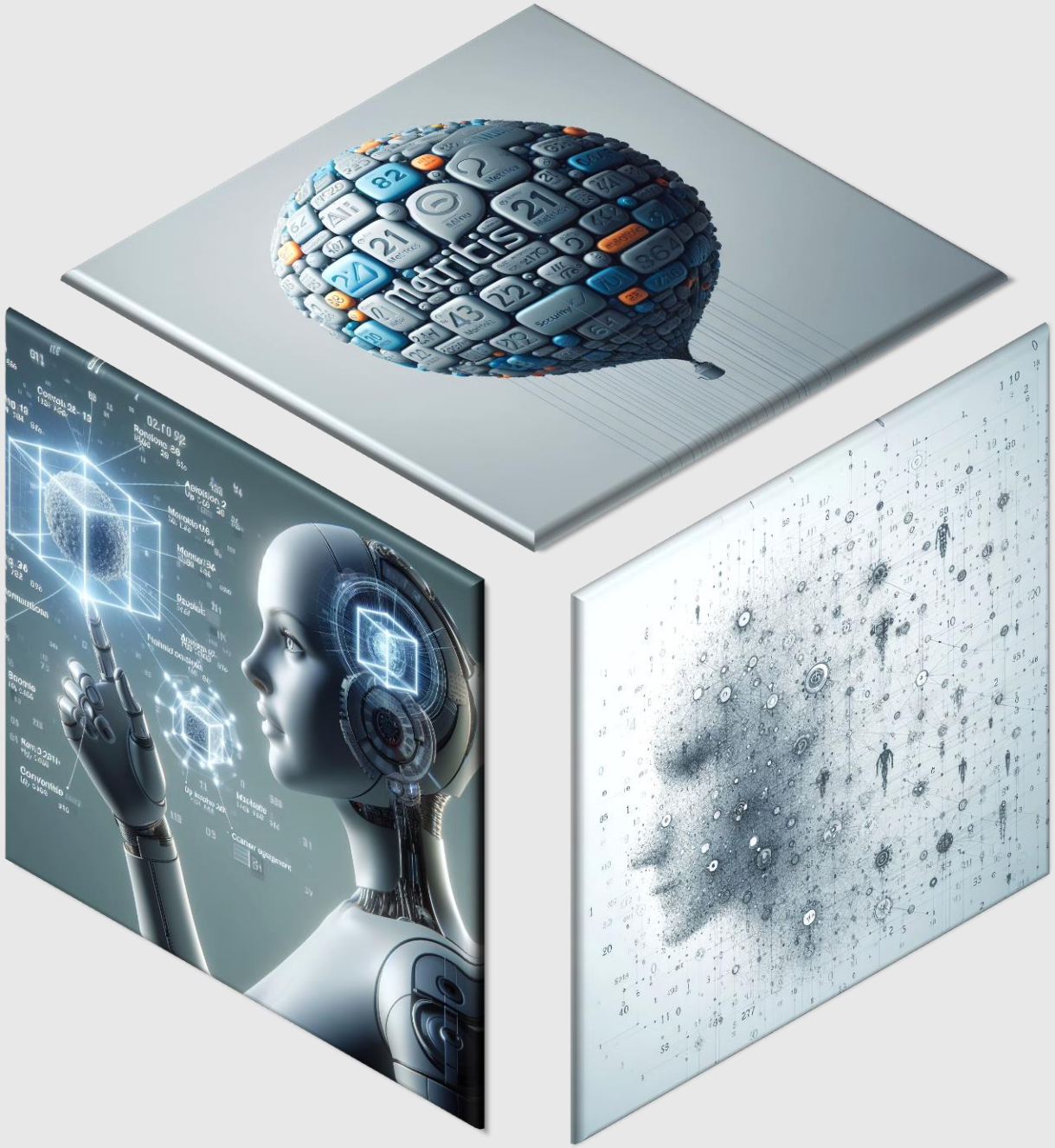


PROJECT SUMMARY

We have embarked on a journey to redefine security in marketing attribution, bridging the gap between data utility and confidentiality. Through a series of strategic innovations and rigorous testing, we have fortified our data processing against sophisticated threats, ensuring the integrity of our insights and the trust of our users.

- **Data Security Milestones**: Generated synthetic data mirroring real behavior, reinforcing dual-layer security for integration and attack resilience.
- **Refined Attack Defenses**: Employed CVPL techniques, enhancing defense mechanisms against sophisticated data breaches and reducing risk levels.
- **Innovative Data Integration**: Crafted a secure data exchange framework using ZKP by Schnorr, tailored for micro-ecosystem interoperability and privacy compliance.
- **Strategic Enhancements**: Applied differential privacy selectively, fortifying data anonymity while preserving the analytical integrity essential for micro-ecosystems.
- **Forward Path**: Continuous refinement guided by stringent security and quality metrics, poised for the evolution of secure micro-ecosystems.

THANK YOU



APPENDIX

MARKETING ATTRIBUTION

RE ALM

DATA EXCHANGE SCHEME

BANK

ADVERTISING PLATFORM

BANK TRANSACTION SET (MA_BANK.csv)

200K records

ID - In-house banking identity
MOBILEPHONE – Phone Number (Main ID)
FIO – full name of the client (in a practical sense – to be deleted)
SEX – Client's gender (social profile)
AGE –Social Profile
FAMILYSTATUS – Marital Status (Social Profile)
INCOME – Monthly Person Income
BRAND – Product name (product brand, e.g. LENOVO)
CATEGORY –product category (e.g. LAPTOPS)
CREDITCARD – card number (in practical terms, to be deleted)
REGION – Selected Region(City, State or Province)
GEO – Buy coordinates (+/- 200 m from the pivot point)
SHOPPINGDATE – Date and time of purchase
TIMEZONE – Time zone (corresponds to the region)
PURCHASEAMOUNT – Purchase Amount

THE MAIN LINK BETWEEN DATASETS BY PHONE NUMBER

Auxiliary link by BRAND attributes (purchase and advertised product)

ADVERTISING PLATFORM SET (MA_ADS.csv)

32K records

ID - Internal Ad Space identity
MOBILEPHONE – Phone Number (Main ID)
REGION – selected region (State, Province, City)
TIMEZONE – Time zone (corresponds to the region)
BRAND – Product name (product brand, e.g., LENOVO)
CATEGORY – product category (e.g. LAPTOPS)
AD_DATEVIEW – Date when the ad was viewed (recent)
AD_VIEWCOUNT – Number of ad views by the client

THE BANK AND THE ADVERTISING PLATFORM MAY NOT DIRECTLY EXCHANGE UNPROTECTED DATA

A special exchange scheme is required without disclosure of personal information

Additional Threat: Feedback to the Original Set, Indirectly Reveals Information

AGGREGATED SET OF INFORMATION WITH ANONYMIZATION

The combined set provides valuable information for ML modeling

TRUSTED INTERMEDIARY

PROCESS MECHANICS

Anonymization is performed independently and asynchronously, the main task is to provide a single identifier for the "intersection" of sets

Attacks on matching set crossing mechanisms

BANKING ENVIRONMENT

phone

MA_BANK

Source kit with highly sensitive personal and financial information

*phone

MA_BANK_MASK

Anonymized Set with Randomized Values

Secure Information Exchange Mechanism

BASIC APPROACHES

- Centralized intermediary
- **Dynamic hashing**
- Private Set Intersection = PSI (Confidential Computing)
- Trust Execution Environment (TEE)
- Federated Learning

MARKETING PLATFORM ENVIRONMENT

*phone

MA_ADV_MASK

Anonymized Set with Randomized Values

phone

MA_ADV

Source kit with highly sensitive personal information

ANONYMIZATION

Basic attributes are blurred, quality is reduced, but safety is improved

Linking Attack

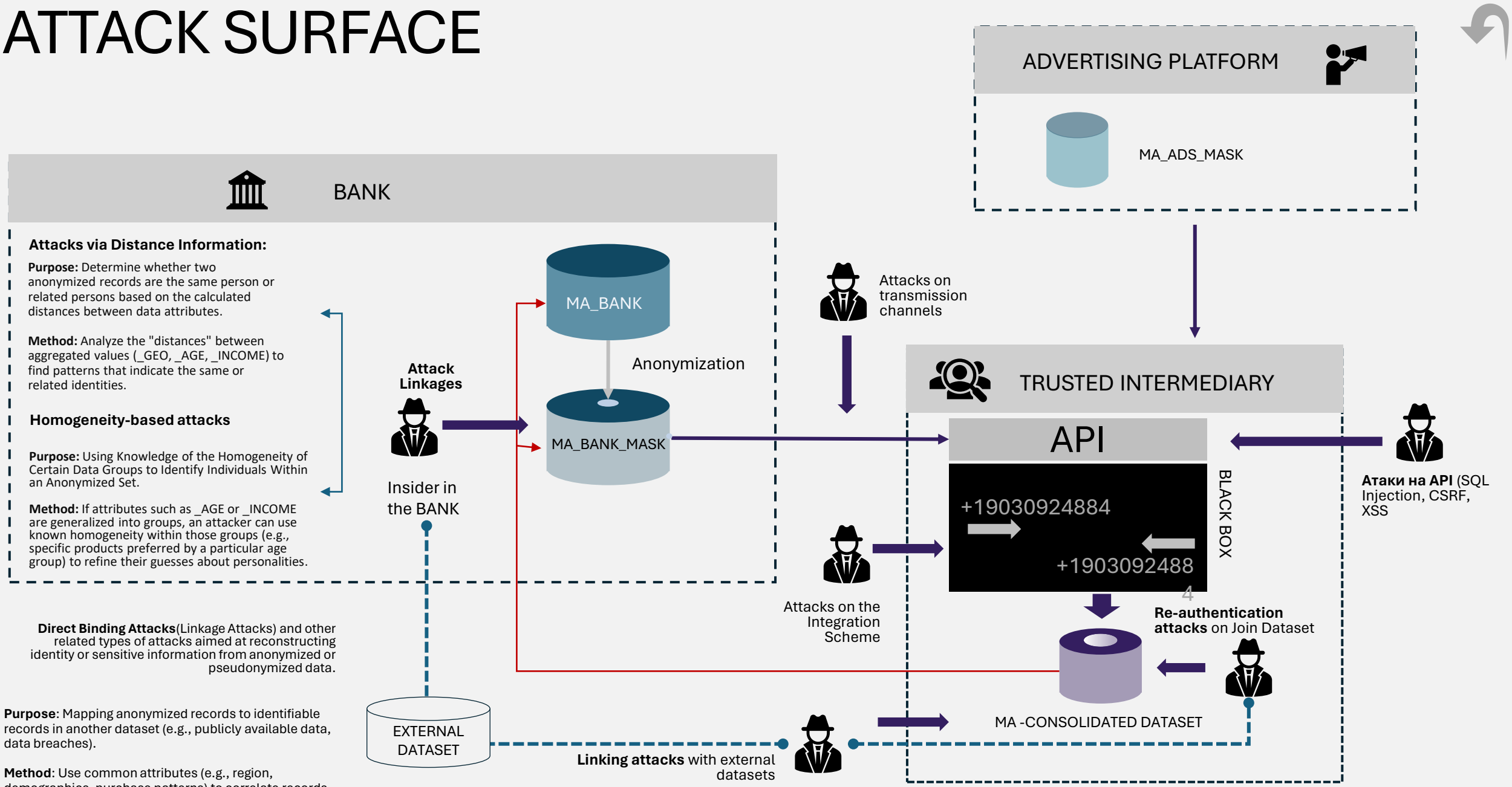
Re-ID атака (Singling out)

MA

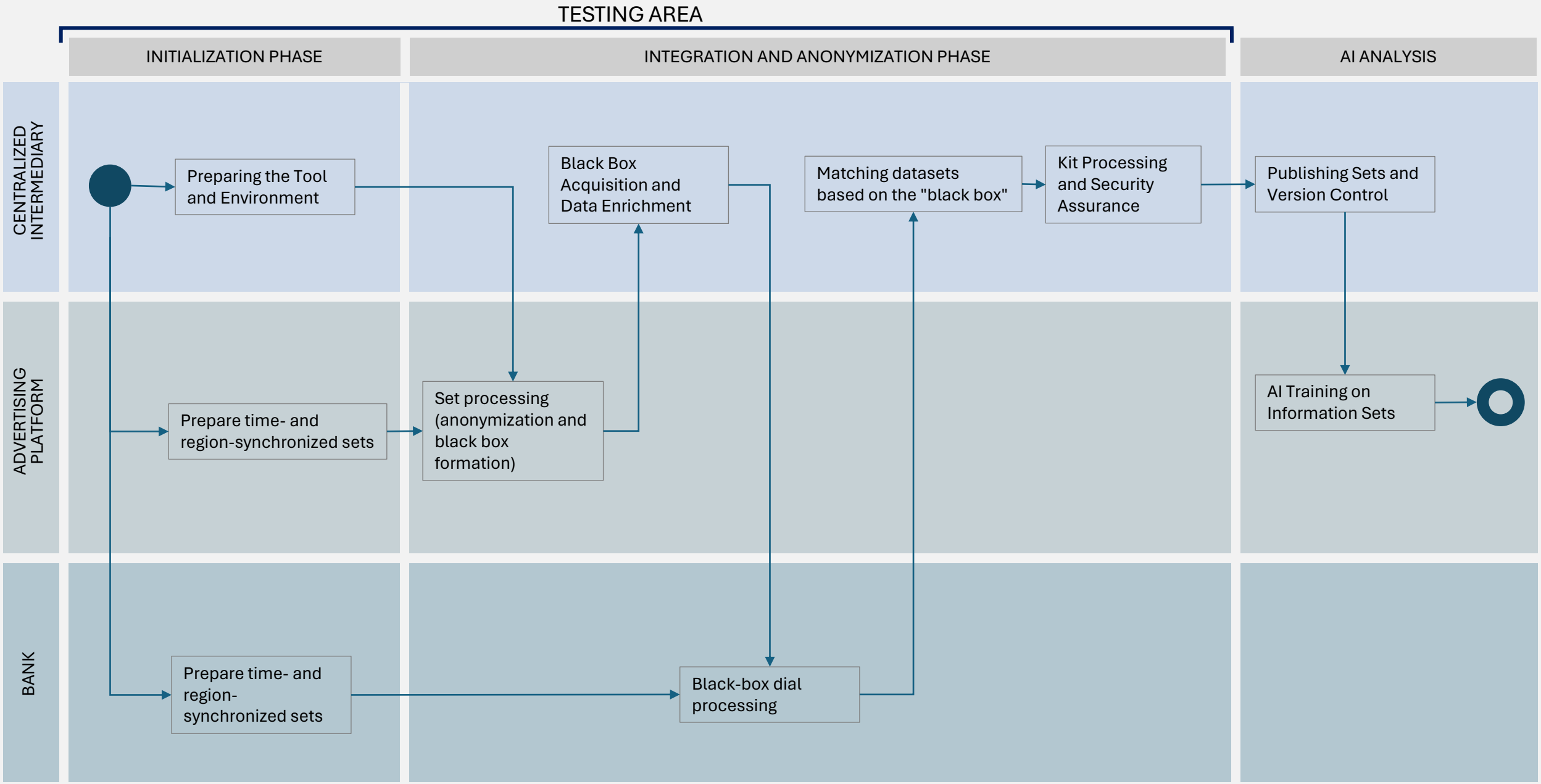
A combined anonymized set in which:
Records cannot be highlighted (re-identification attacks)
Cannot be linked to source sets (Linkage Attack)

Re-ID attack (Inference attack)

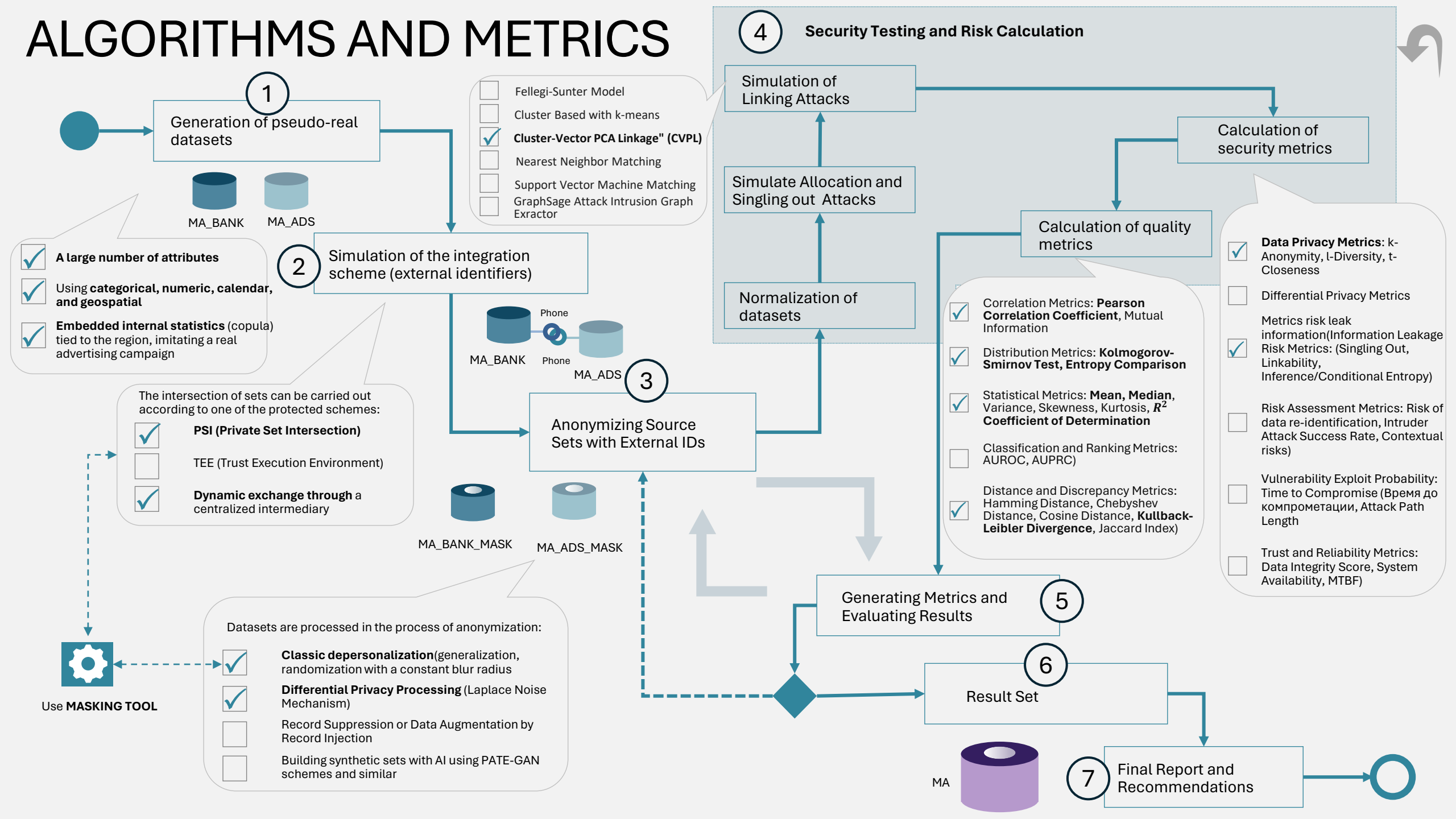
ATTACK SURFACE



TRUSTED BROKER IMPLEMENTATION

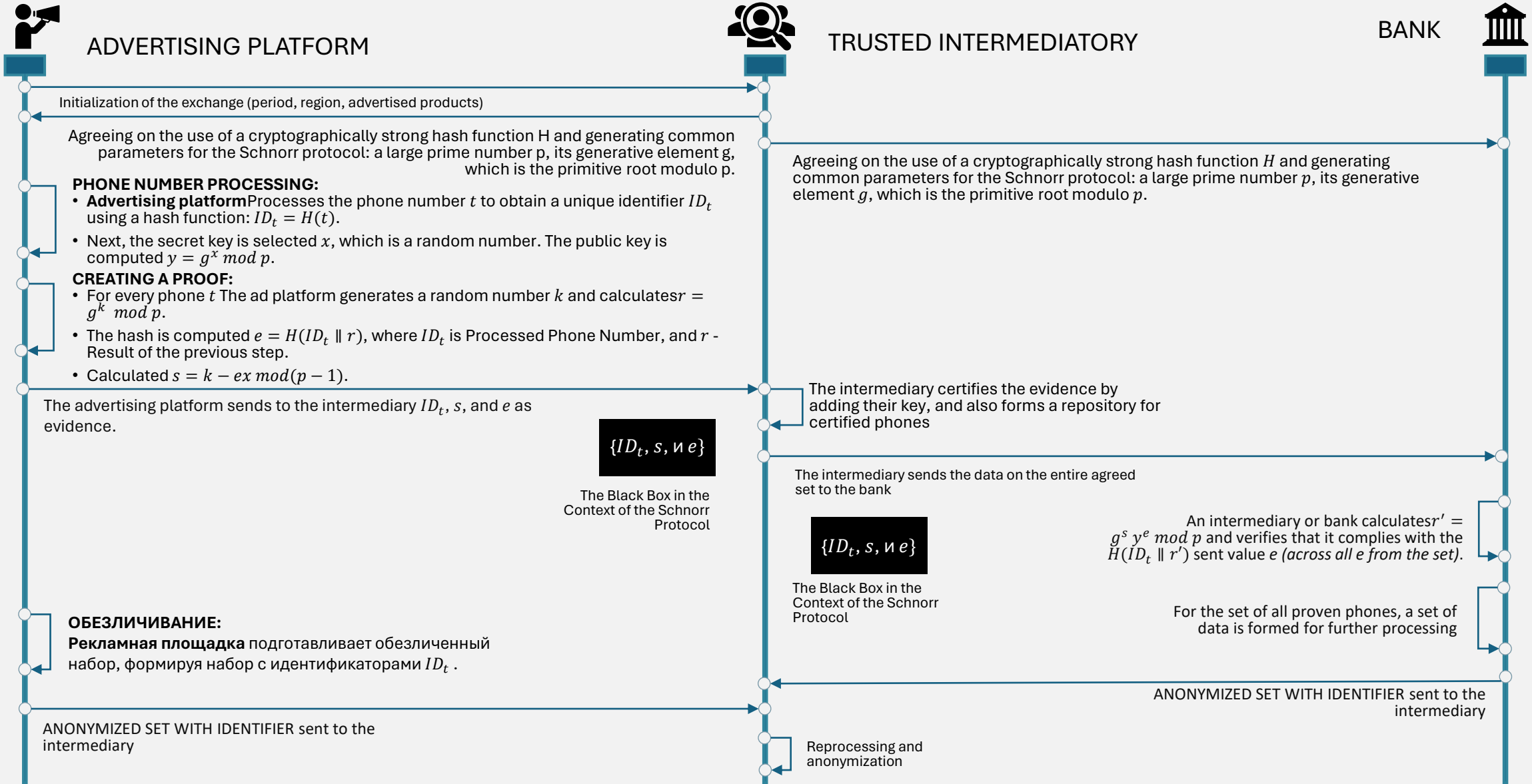


ALGORITHMS AND METRICS

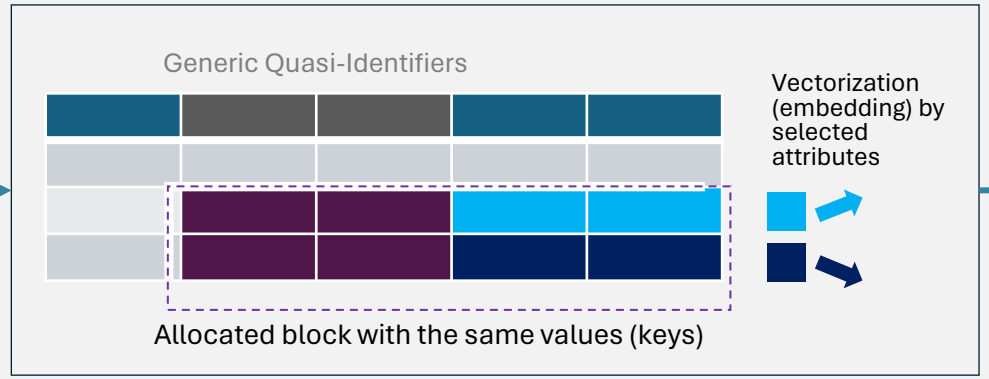


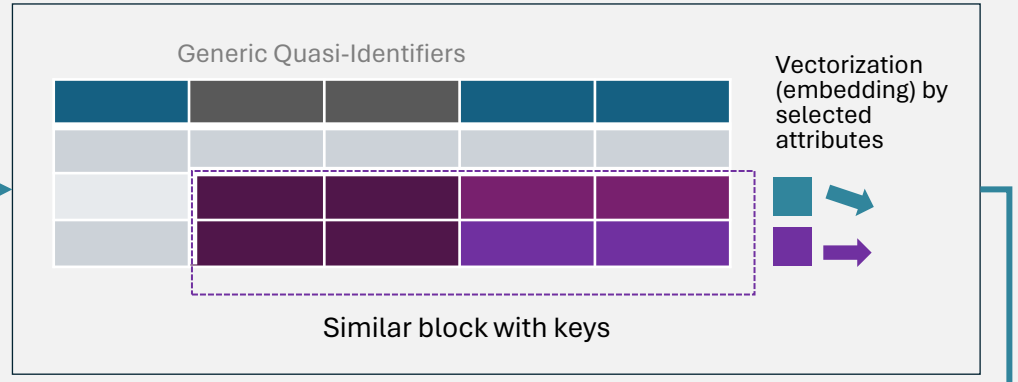
SCHNORR INTEGRATION

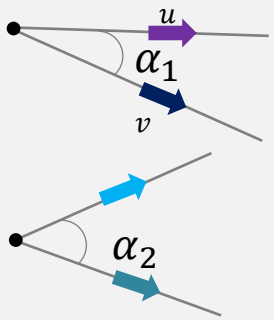
There are various schemes for combining distributed information. In the context of testing, the concept of a BLACK BOX is used, in which information is combined without the parties disclosing information about users' mobile phones. Below is a conceptual scheme that is not directly tested, but is supposed to be implemented on the MASKER side. Such a scheme is based on Membersip ZKP (Zero Knowledge Proof) and uses the Schnorr Scheme algorithm.



LINKING ATTACK: Cluster-Vector PCA Linkage (CVPL)

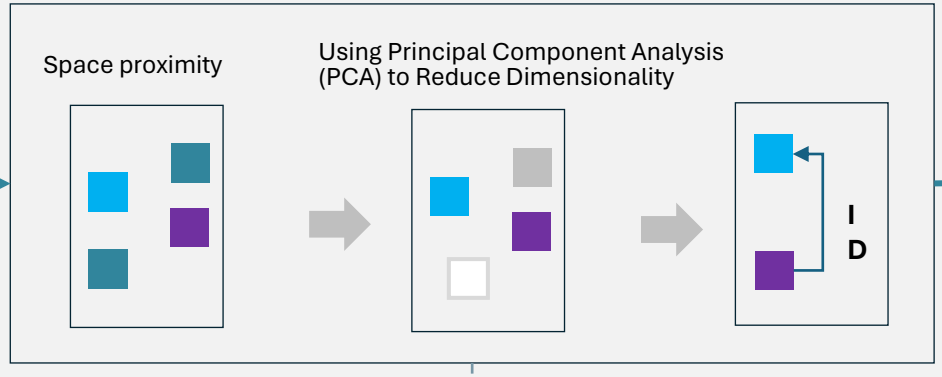






$$\cos\alpha = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \leq Tr$$

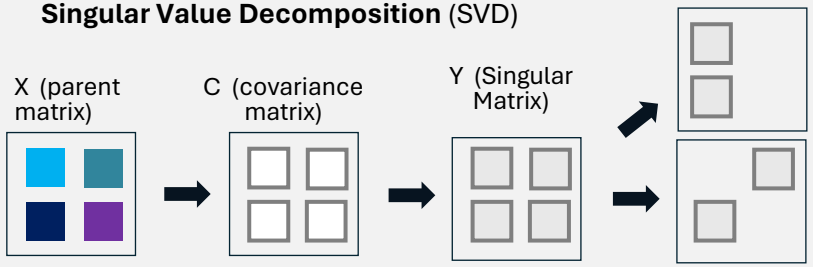
Using the cosine similarity formula, we find the proximity between the vectors



Calculate probabilities and calculate one-to-one relationships

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

Singular Value Decomposition (SVD)



$C = X^T \times X$

$C \times v = \lambda v$ - Eigenvectors

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

U: Matrix of Left Singular Vectors (n x n)

S: Diagonal matrix of singular values (n x m).

$\boldsymbol{V^T}$: Transposed Matrix of Right Singular Vectors (m x m).

NUMERICAL RESULTS

DATASETS MA_BANK – MA_BANK_MASK

200,000 entries. Original Bank Set and Set After MASKER Processing

QUALITY INDICATORS (MA_BANK – MA_BANK_MASK)

Metric	Value	Context	Comments
Pearson Coefficient	0.9086 0.9196	MA_BANK/MA_BANK_MASK _AGE;_INCOME	In general, the correlation is preserved, which allows us to conclude that the quality of the set is high
KS statistics	0.01164	_AGE;_PURCHASEAMOUNT	Shows similar distributions by KS – statistics average age – purchases
KS statistics	2.7702E-02	Complete Sets with Original Distribution After MASKER by _AGE;_INCOME	A low score that demonstrates affinity with a set of MA_BANK and MA_BANK_MASK
KL	1.5272	PURCHASEAMOUNT | _AGE;_INCOME | COUNT (Keys Only)	The KL divergence measures how different the distributions are, with a score of 1.5 being a moderate change
KL	0.2399	INCOME| _AGE;_PURCHASEAMOUNT | SUM	The indicators show low KL values – a good approximation
R^2	0.2597	PURCHASEAMOUNT | _AGE;_SHOPPINGDATE | MEAN	The distributions show the discrepancy for this indicator
R^2	0.9998	PURCHASEAMOUNT | ID | SUM	The independent distribution shows very close values

PRIVACY MEASURES (MA_BANK – MA_BANK_MASK)

Metric	Value	Context	Comments
SEFABR_I (k =1)	**1 889 / 2 142**	SEX; FAMILYSTATUS; BRAND; _INCOME	High performance due to the chosen methodology for simulating real data
SEFABR_A_I (k =1)	**17 812 / 18 609**	SEX; FAMILYSTATUS; BRAND; _AGE; _INCOME	High performance due to the chosen methodology for simulating real data
SEFABR_S (k =1)	**20 566 / 20 500**	SEX; FAMILYSTATUS; BRAND; _SHOPPINGDATE_HRS	Very high performance
Linkage Attack_1	R100 = 76141 R50 = 42435	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS |PURCHASEAMOUNT	A significant number of values found
Linkage Attack_2	R100= 12827 R50 = 192	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS | PURCHASEAMOUNT; INCOME;_GEO_RO;_GEO_FI	The attack shows the ability to link a significant number of records
Linkage Attack_3	R100 = 21645 R50 = 2384	SEX;FAMILYSTATUS;BRAND;_INCOME;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;_GEO_RO;_GEO_FI	The found association indicators allow us to conclude that the randomization scheme is unreliable within the entire set without taking into account clusters
RISK	0.4	$R_{total} = w_1 \times R_{englout} + w_2 \times R_{link} + w_3 \times R_{inf}$ $w_1 = 0.1; w_2 = 0.6; w_3 = 0.3$	The risk appears to be quite high due to the possibility of linking a significant number of records

DATASETS MA_BANK_1 – MA_BANK_MASK_1

32,400 entries. The original bank (crossed) set and the set after processing by the MASKER.

QUALITY INDICATORS(MA_BANK_1* – MA_BANK_MASK_1*)

Metric	Value	Context	Comments
Pearson Coefficient	0.91205 0.9301 (*0.8245)	MA_BANK_1/MA_BANK_MASK_1 _AGE;_INCOME	After the additional protection, the correlation rate decreased, but still remains high
KS-статистика	3.3750E-02 (*0.0963)	_AGE;_PURCHASEAMOUNT	After the additional application of differential privacy by block, the statistics deteriorated, but only slightly
KS-статистика	1.7102E-02 (*0.1621)	_AGE;_INCOME	The statistics also decreased after the application of additional erosion, possibly excessive granulation
KL	1.2573 (*2.8562)	PURCHASEAMOUNT | _AGE;_INCOME | COUNT (только по ключам)	Excessive granulation by KL also leads to an increase in divergence (the distance between sets in the protected version has increased)
KL	0.2116 (*1.1754)	INCOME| _AGE;_PURCHASEAMOUNT | SUM	For the protected version, the indicator has increased as expected
R^2	0.5597 (*0.244)	PURCHASEAMOUNT | _AGE;_SHOPPINGDATE | MEAN	For baseline randomization, an acceptable result (above 50%) was shown For the updated model, the discrepancy is too low (the original quality of the data by dates is possible)
R^2	0.9553 (*0.8342)	PURCHASEAMOUNT | ID | SUM	On protected data, there is a slight decrease

* Values obtained after re-randomization using the differential privacy scheme within blocks are marked here

PRIVACY MEASURES (MA_BANK_1* – MA_BANK_MASK_1*)

Metric	Value	Context	Comments
SEFABR_I (k =1)	**29 / 46** (*7)	SEX; FAMILYSTATUS; BRAND; _INCOME	Low risk scores in the protected variant
SEFABR_A_I (k =1)	**273 / 313** (*11)	SEX; FAMILYSTATUS; BRAND; _AGE; _INCOME	Indicators have been reduced to acceptable due to the lack of connectivity
SEFABR_S (k =1)	**0 / 0** (*0)	SEX; FAMILYSTATUS; BRAND; _SHOPPINGDATE_HRS	Not available in the protected variant
Linkage Attack_4	R100 = 1269 R50 = 6675 (*6)	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS |PURCHASEAMOUNT	In the protected version, low indicators are not tied to the source data
Linkage Attack_5	R100= 2604 R50 = 192 (*0)	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;INCOME;_GEO_RO;_GEO_FI	The attack does not reach any indicators in the protected variant
Linkage Attack_6	R100 = 4772 R50 = 1498 (*1)	SEX;FAMILYSTATUS;BRAND;_INCOME;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;_GEO_RO;_GEO_FI	Significant reduction when differential privacy is added
RISK	0.23 (*0.01)	$R_{total} = w_1 \times R_{englout} + w_2 \times R_{link} + w_3 \times R_{inf}$ $w_1 = 0.1; w_2 = 0.6; w_3 = 0.3$	The risk is significantly reduced and can be reduced further with further runs: up to 10^{-4}

* Values obtained after re-randomization using the differential privacy scheme within blocks are marked here

QUALITY INDICATORS (MA_BANK – MA_BANK_MASK)			
Metric	Value	Context	Comments
Pearson Coefficient	0.9086 0.9196	MA_BANK/MA_BANK_MASK _AGE;_INCOME	In general, the correlation is preserved, which allows us to conclude that the quality of the set is high
KS statistics	0.01164	_AGE;_PURCHASEAMOUNT	Shows similar distributions by KS – statistics average age – purchases
KS statistics	2.7702E-02	Complete Sets with Original Distribution After MASKER by _AGE;_INCOME	A low score that demonstrates affinity with a set of MA_BANK and MA_BANK_MASK
KL	1.5272	PURCHASEAMOUNT | _AGE;_INCOME | COUNT (Keys Only)	The KL divergence measures how different the distributions are, with a score of 1.5 being a moderate change
KL	0.2399	INCOME| _AGE;_PURCHASEAMOUNT | SUM	The indicators show low KL values – a good approximation
R^2	0.2597	PURCHASEAMOUNT | _AGE;_SHOPPINGDATE | MEAN	The distributions show the discrepancy for this indicator
R^2	0.9998	PURCHASEAMOUNT | ID | SUM	The independent distribution shows very close values

PRIVACY MEASURES (MA_BANK – MA_BANK_MASK)

Metric	Value	Context	Comments
SEFABR_I (k =1)	**1 889 / 2 142**	SEX; FAMILYSTATUS; BRAND; _INCOME	High performance due to the chosen methodology for simulating real data
SEFABR_A_I (k =1)	**17 812 / 18 609**	SEX; FAMILYSTATUS; BRAND; _AGE; _INCOME	High performance due to the chosen methodology for simulating real data
SEFABR_S (k =1)	**20 566 / 20 500**	SEX; FAMILYSTATUS; BRAND; _SHOPPINGDATE_HRS	Very high performance
Linkage Attack_1	R100 = 76141 R50 = 42435	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS |PURCHASEAMOUNT	A significant number of values found
Linkage Attack_2	R100= 12827 R50 = 192	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS | PURCHASEAMOUNT; INCOME;_GEO_RO;_GEO_FI	The attack shows the ability to link a significant number of records
Linkage Attack_3	R100 = 21645 R50 = 2384	SEX;FAMILYSTATUS;BRAND;_INCOME;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;_GEO_RO;_GEO_FI	The found association indicators allow us to conclude that the randomization scheme is unreliable within the entire set without taking into account clusters
RISK	0.4	$R_{total} = w_1 \times R_{snglout} + w_2 \times R_{link} + w_3 \times R_{inf}$ $w_1 = 0.1 ; w_2 = 0.6; w_3 = 0.3$	The risk appears to be quite high due to the possibility of linking a significant number of records

QUALITY INDICATORS(MA_BANK_1* – MA_BANK_MASK_1*)

Metric	Value	Context	Comments
Pearson Coefficient	0.91205 0.9301 (*0.8245)	MA_BANK_1/MA_BANK_MASK_1 _AGE;_INCOME	After the additional protection, the correlation rate decreased, but still remains high
KS-статистика	3.3750E-02 (*0.0963)	_AGE;_PURCHASEAMOUNT	After the additional application of differential privacy by block, the statistics deteriorated, but only slightly
KS-статистика	1.7102E-02 (*0.1621)	_AGE;_INCOME	The statistics also decreased after the application of additional erosion, possibly excessive granulation
KL	1.2573 (*2.8562)	PURCHASEAMOUNT | _AGE;_INCOME | COUNT (только по ключам)	Excessive granulation by KL also leads to an increase in divergence (the distance between sets in the protected version has increased)
KL	0.2116 (*1.1754)	INCOME| _AGE;_PURCHASEAMOUNT | SUM	For the protected version, the indicator has increased as expected
R^2	0.5597 (*0.244)	PURCHASEAMOUNT | _AGE;_SHOPPINGDATE | MEAN	For baseline randomization, an acceptable result (above 50%) was shown For the updated model, the discrepancy is too low (the original quality of the data by dates is possible)
R^2	0.9553 (*0.8342)	PURCHASEAMOUNT | ID | SUM	On protected data, there is a slight decrease

* Values obtained after re-randomization using the differential privacy scheme within blocks are marked here

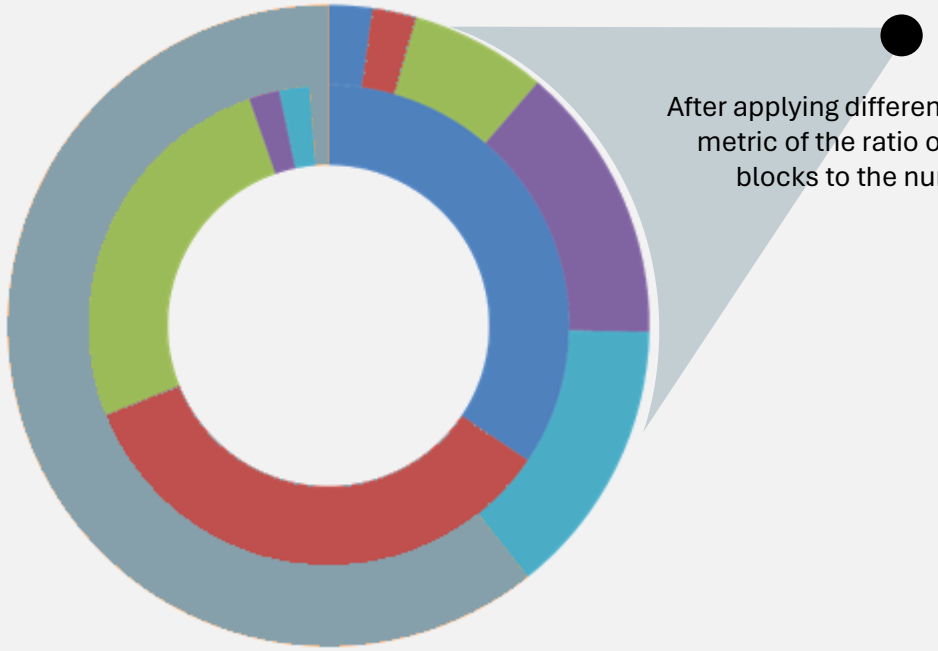
PRIVACY MEASURES (MA_BANK_1* – MA_BANK_MASK_1*)

Metric	Value	Context	Comments
SEFABR_I (k =1)	**29 / 46** (*7)	SEX; FAMILYSTATUS; BRAND; _INCOME	Low risk scores in the protected variant
SEFABR_A_I (k =1)	**273 / 313** (*11)	SEX; FAMILYSTATUS; BRAND; _AGE; _INCOME	Indicators have been reduced to acceptable due to the lack of connectivity
SEFABR_S (k =1)	**0 / 0** (*0)	SEX; FAMILYSTATUS; BRAND; _SHOPPINGDATE_HRS	Not available in the protected variant
Linkage Attack_4	R100 = 1269 R50 = 6675 (*6)	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS |PURCHASEAMOUNT	In the protected version, low indicators are not tied to the source data
Linkage Attack_5	R100= 2604 R50 = 192 (*0)	SEX;FAMILYSTATUS;BRAND;_AGE;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;INCOME;_GEO_RO;_GEO_FI	The attack does not reach any indicators in the protected variant
Linkage Attack_6	R100 = 4772 R50 = 1498 (*1)	SEX;FAMILYSTATUS;BRAND;_INCOME;_SHOPPINGDATE_HRS | PURCHASEAMOUNT;_GEO_RO;_GEO_FI	Significant reduction when differential privacy is added
RISK	0.23 (*0.01)	$R_{total} = w_1 \times R_{snglout} + w_2 \times R_{link} + w_3 \times R_{inf}$ $w_1 = 0.1 ; w_2 = 0.6; w_3 = 0.3$	The risk is significantly reduced and can be reduced further with further runs: up to10^{-4}

* Values obtained after re-randomization using the differential privacy scheme within blocks are marked here

СТАТИСТИКА ПО АТАКАМ

Number of Blocks and Vectors for CVPL Attacks

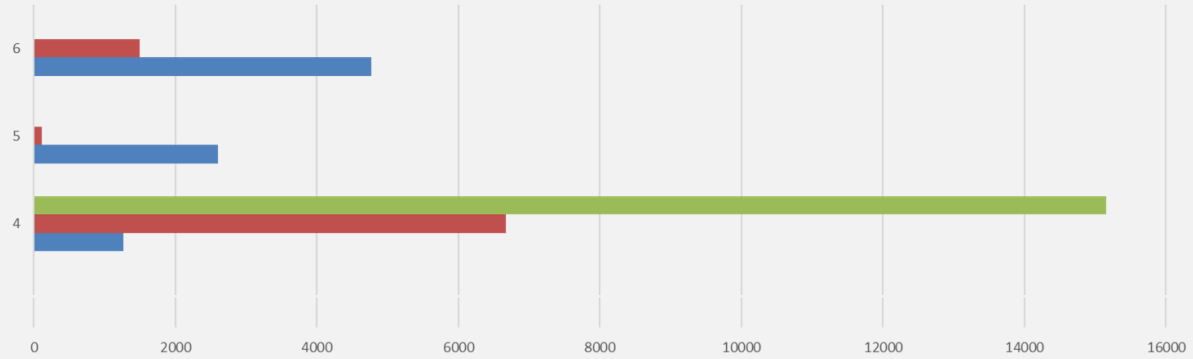


After applying differential privacy, the metric of the ratio of the number of blocks to the number of vectors

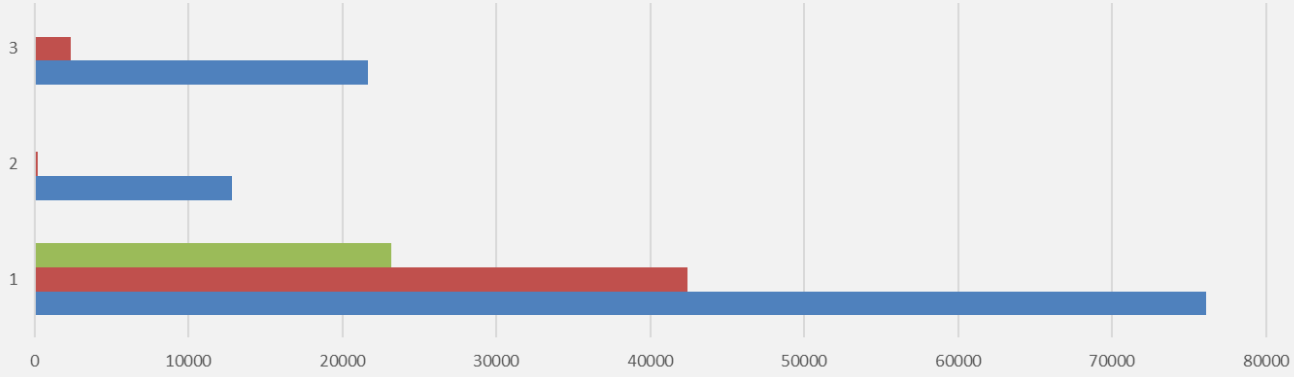
Attack_1 Attack_4	SEX;FAMILYSTATUS;BRAND;_AGE; _SHOPPINGDATE_HRS PURCHASEAMOUNT	BLOCKS Vectors
Attack_2 Attack_5	SEX; FAMILYSTATUS; BRAND; _AGE;_SHOPPINGDATE_HRS PURCHASEAMOUNT; INCOME; _GEO_RO; _GEO_FI	BLOCKS Vectors
Attack_3 Attack_6	SEX;FAMILYSTATUS; BRAND; _INCOME; _SHOPPINGDATE_HRS PURCHASEAMOUNT;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

Each attack is determined by the selected fields for building blocks and vectors within blocks

SMALL SET ATTACKS (MA_BANK_1 – MA_BAN_MASK_1)



ATTACKS ON LARGE SETS (MA_BANK – MA_BAN_MASK)

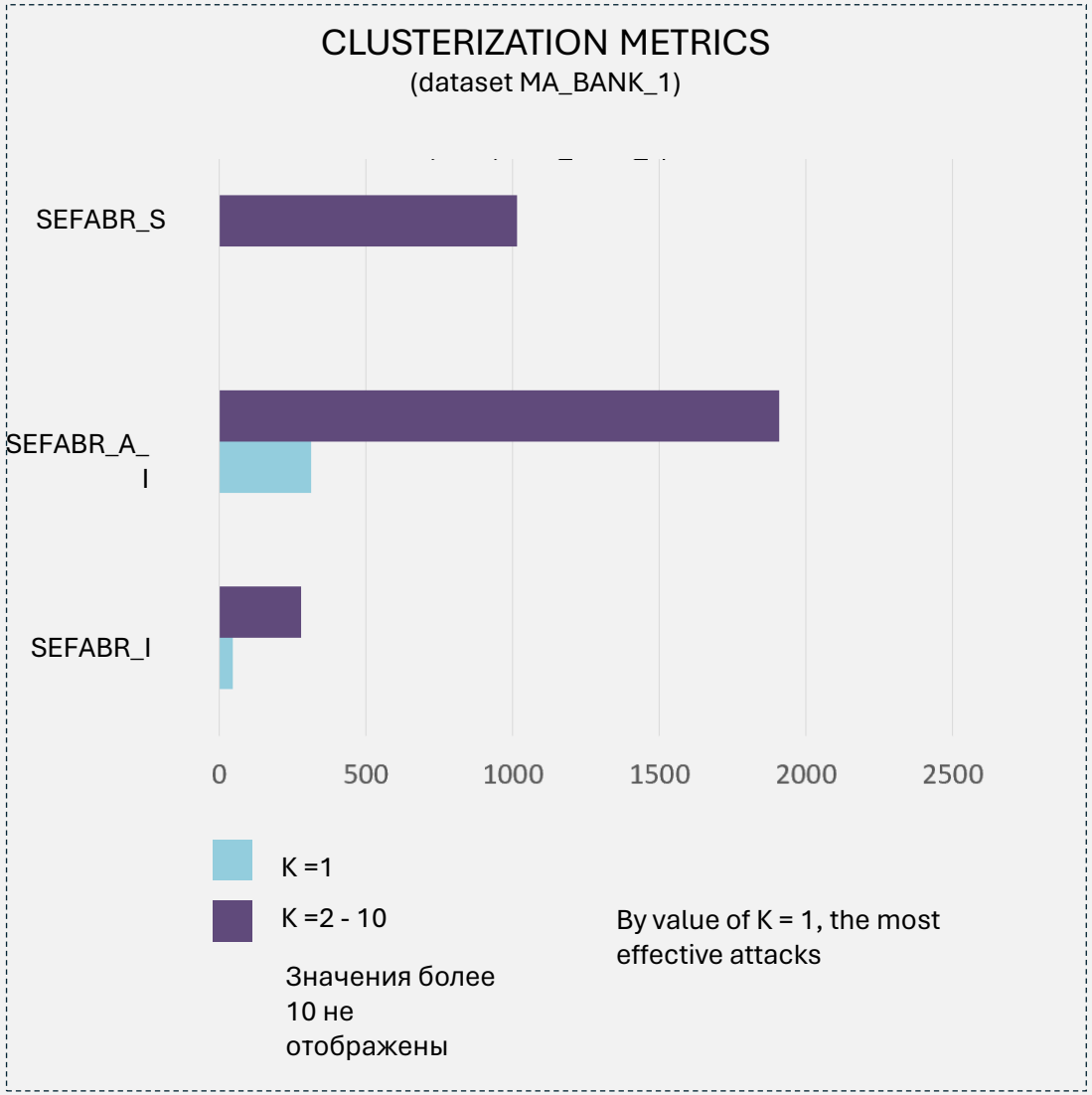


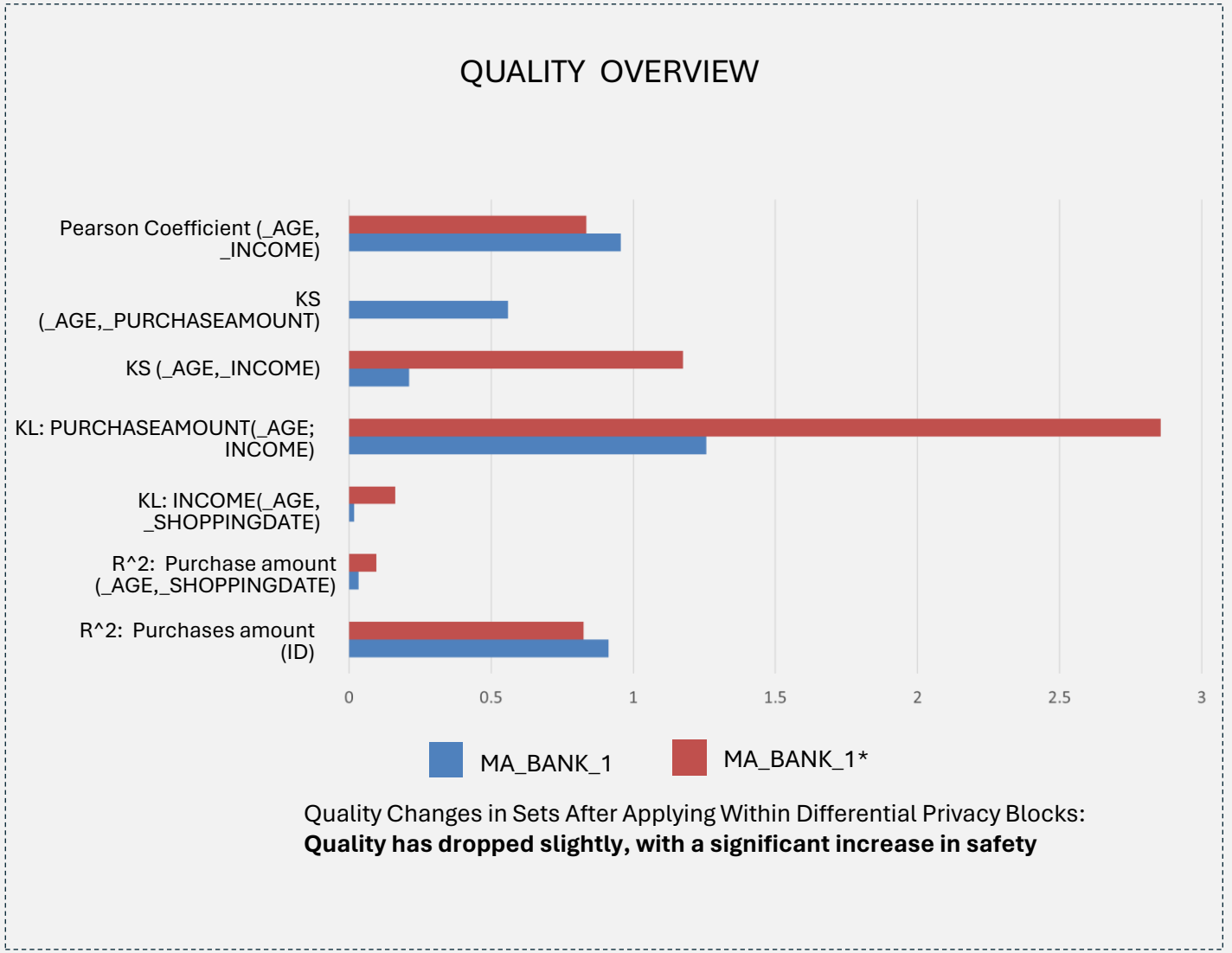
100% identification

25-50% identification

< 25% identification

QUALITY VS. PRIVACY





Quality Changes in Sets After Applying Within Differential Privacy Blocks:
Quality has dropped slightly, with a significant increase in safety