A Multi-Dimensional Empirical Evaluation of k-Anonymity, Centralized and Local Differential Privacy, and Federated Learning for Personal Data Protection in Big Data Scenarios

Yuan Li[1], Shuai Yuan[1,2*]
[1]Henan Kaifeng College of Science Technology and Communication, Kaifeng 475000, China
[2]Henan University, Kaifeng 475000, China
E-mail: huaiyuannnn91@163.com
[*]Corresponding author

Keywords: big data, personal information, privacy protection, anonymization, privacy enhancement technology

Received: October 11, 2025

This study systematically evaluates the performance of mainstream personal information anonymization technologies in the big data environment. It focuses on k-anonymity family models, differential privacy (DP) and its local variants, and privacy protection mechanisms in federated learning (FL). A multi-dimensional comprehensive evaluation system is constructed in the study, measured from four aspects: privacy protection strength, data utility, computational efficiency, and scalability. Indicators such as normalized certainty penalty (NCP), area under the curve (AUC), and processing time are adopted. The experimental results show that under the condition of medium-intensity privacy protection (privacy budget ε = 1.0), the central differential privacy (CDP) mechanism exhibits the best comprehensive performance in most scenarios. In the medical data release scenario, the NCP of CDP is 0.05, the model's AUC is 0.945, and the average processing time is 25 seconds. In contrast, the k-anonymity model has higher information loss (NCP = 0.22) and lower model accuracy (AUC = 0.920). In the FL scenario of financial anti-fraud, the AUC of the global model under CDP protection is 0.890, which is significantly better than the AUC of the local differential privacy (LDP) scheme (0.842), with comparable computational overhead. In the user behavior analysis experiment involving one million users, the LDP method based on the Randomized Aggregate Privacy-Preserving Ordinal Response (RAPPOR) mechanism achieves lower mean absolute error and higher popular category recognition accuracy (99.0%). In all tests, the NCP and AUC of CDP remain stable with changes in data scale, while the processing time increases approximately linearly with data volume, showing excellent scalability. The above results quantitatively verify the significant advantages of CDP over traditional k-anonymity and LDP methods in data utility, privacy protection strength, and computational efficiency. Thus, empirical evidence can be provided for the selection of anonymization technologies in large-scale data processing scenarios.

Povzetek: Študija primerjalno ocenjuje tehnologije anonimizacije osebnih podatkov v velikih podatkih in pokaže, da centralna diferencialna zasebnost (CDP) pri uravnoteženi ravni zasebnosti dosega najboljše razmerje med zaščito zasebnosti, uporabnostjo podatkov, učinkovitostjo in razširljivostjo.

1 Introduction

In the era of big data, data, as a new sort of production element and strategic resource, its relevance is self-evident. Many sorts of information are gathered and kept in vast amounts as a result of the development of technologies like social media platforms, Internet of Things (IoT) devices, online transaction systems, and smart terminals. This information includes basic identifiers such as names and ID card numbers, as well as sensitive data, including phone numbers, location trajectories, health records, and consumption patterns [1-3]. Making sensible use of these enormous volumes of data can significantly advance scientific research, streamline corporate decision-making for organizations, and enhance the effectiveness and caliber of public services. However, such large-scale data sharing and in-depth mining have also brought about severe risks of privacy leakage. Frequent data leakage incidents and the increasing abuse of data pose a direct threat to personal privacy and hinder the normal circulation of data as a key production factor [4-6].

To address such challenges, a series of laws and regulations have been formulated internationally to standardize data processing behaviors and ensure the security of personal information. For instance, the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and China's Personal Information Protection Law all clearly stipulate the principles that must be followed in data processing. Among them, the "minimum necessity" principle emphasizes collecting only data necessary for achieving specific purposes. De-identification or anonymization refers to processing personal information through technical means such that specific individuals cannot be identified without additional auxiliary information, thereby reducing privacy risks during data usage. Anonymization technology has thus become one of the important tools for safeguarding data security and promoting the legal and compliant circulation of data. It aims to transform or encrypt raw data, retaining data value while minimizing the amount of information that can be used to identify individuals, thus meeting the legal definition requirements for "anonymous information" [7-9].

Therefore, exploring efficient and robust anonymization technologies adapted to the characteristics of big data, and in-depth studying their practical effects in different application scenarios, holds important theoretical significance and practical value for promoting technological innovation in the field of data security and constructing a more comprehensive legal supervision system. Focusing on this theme, this study systematically sorts out and evaluates the performance and application challenges of current mainstream anonymization technologies in the big data environment, and looks forward to their future development trends. The specific objectives of this study are as follows:

- It identifies the performance bottlenecks of existing anonymization technologies in big data scenarios.
- An optimization strategy is proposed that combines the advantages of various technologies.
- Theoretical and empirical references are offered for developing more efficient, robust, and applicable anonymous schemes in the future.

By constructing a comprehensive evaluation framework and combining practical case analysis, this study aims to provide a scientific basis for the selection and optimization of big data privacy protection technologies.

The innovation of this study is not intended to propose an entirely new anonymization algorithm. Instead, it aims to conduct an unprecedented systematic empirical comparison of mainstream technologies. These technologies include k-anonymity, CDP, and local differential privacy (LDP), and privacy protection in federated learning (FL). The comparison is conducted under unified experimental settings. This is achieved through the construction of a multi-dimensional evaluation system. The evaluation system covers privacy strength, data utility, computational efficiency, and scalability. Most current studies focus on the improvement of a single technology, lacking a panoramic description of their comprehensive performance and trade-off relationships across diverse big data scenarios. This study aims to fill this gap. Its core scientific contribution lies in revealing the real performance boundaries and interaction relationships of these technologies under actual deployment conditions through rigorous experiments. For example, the significant advantages of CDP in scalability and its sensitivity to data distribution, as well as the specific conditions for privacy protection failure of k-anonymity models in high-dimensional data. These findings provide crucial empirical evidence and theoretical references for constructing more efficient and reliable big data privacy protection solutions.

2 Literature review

In recent years, research on personal information anonymization technologies has continued to deepen against the backdrop of big data, showing a trend of evolving from classical models to paradigms with stronger robustness and provable security [10]. The k-anonymity model proposed by Li et al. and its subsequent improved schemes, such as l-diversity and t-closeness, have laid a foundation in the scenario of tabular data publishing through technical means like generalization and suppression [11]. However, with the increase in data dimensions and the enhancement of attackers' background knowledge, these equivalence class-based methods face severe "curse of dimensionality" and utility loss issues in high-dimensional data, and are difficult to resist complex inference attacks. For this reason, differential privacy (DP) has become a research hotspot due to its strong theoretical guarantees [12]. The DP framework established by Seeman and Susser provided a quantifiable mathematical boundary for individual privacy leakage by introducing controlled noise into query results or data publishing [13]. Recent studies have focused on improving the efficiency and utility of DP in practical applications. For high-dimensional sparse data, Chen et al. introduced an adaptive Laplace mechanism that ensures ε-DP while greatly reducing the influence of noise introduction on machine learning model performance [14]. In the development context of anonymization technologies, El Ouazzani and El Bakkali systematically reviewed non-encrypted anonymization technologies and proposed a classification framework [15]. This framework distinguishes technologies that rely on operations such as generalization and suppression (e.g., k-anonymity, l-diversity) from cryptography-based methods, which are conducive to selecting appropriate tools based on privacy requirements and computational costs in big data scenarios.

Meanwhile, the integrated application of privacy-enhancing technologies has become a key path to address privacy dilemmas in data collaboration. The risk of centralized data exposure is naturally decreased by FL, which enables several parties to work together to train models without sharing raw data [16]. Ibrahim et al. deeply integrated DP with FL, proposing a dynamic privacy budget allocation algorithm. This algorithm dynamically adjusted noise according to the sensitivity of model gradients during training, effectively balancing the model convergence speed and the privacy protection strength of client data [17]. Likewise, computations on encrypted data were carried out using secure multi-party computation (MPC) and homomorphic encryption technologies, guaranteeing that the data is "usable but not visible". Salako et al. explored an MPC-based multi-party joint statistical analysis protocol, realizing privacy-preserving computation of user credit information in financial anti-fraud scenarios [18]. Anonymization is not a technology that can provide perfect protection. Gadotti et al. profoundly pointed out in their review article that anonymization was essentially an "imperfect science", whose core balanced the inherent tension between data utility and privacy protection [19]. They emphasized that with the growth of data volume and the enhancement of attackers' background knowledge, it was difficult for any anonymization scheme to provide absolute and permanent privacy security. In terms of the deep integration of FL and privacy-enhancing technologies, Singh et al. proposed a privacy-aware hierarchical federated learning (HFL) framework for healthcare scenarios, organically combining DP with MPC. A multi-level model aggregation framework is constructed among terminal devices, regional aggregation nodes, and central servers. This framework protects sensitive data of individual medical institutions through local perturbation and secure aggregation protocols at the bottom layer, and superimposes a global DP mechanism at the upper layer. Thus, it realizes cross-institutional collaborative modeling without exposing original medical records. Experimental results show that on various typical medical prediction tasks, this method still maintains model accuracy close to or even better than traditional flat FL configurations under strict privacy constraints and controllable communication overhead [20].

Table 1 summarizes the key information of the above-mentioned key literature, including the core methods, experimental datasets, evaluation indicators, and identified limitations.

Table 1: Comparison of related research work on personal information anonymization technology

Author/year	**Method**	**Dataset**	**Privacy index**	**Utility index**		**Limitation**
Li et al. [11]	k-anonymity / l-diversity / t-closeness	Census, medical	k-value (3–20)	Information Loss (NCP 0.15–0.65)	AUC ≈ 0.85–0.93	Processing time ↑ super-linearly with $n > 10^4$
Seeman & Susser [13]	Differential Privacy framework	Theoretical / query simulation	ε = 0.1–10	Error Bound (±3–7%)	Utility loss ≈ 5% @ ε = 1	O(n) scaling, linear runtime
Chen et al. [14]	Adaptive Laplace Mechanism	High-dim. logs	ε-DP	Model accuracy (AUC = 0.95 @ ε = 2)	AUC ↑ by 3–5% vs baseline	Stable for 10^6 records
El Ouazzani & El Bakkali [15]	Taxonomic framework	N/A	Conceptual	—	—	—
Ibrahim et al. [17]	FL + Dynamic DP	IoT data (10^3 clients)	Rényi DP	Global AUC = 0.90 @ ε = 2	AUC loss < 2%	Linear comm. overhead with clients
Salako et al. [18]	MPC-based Joint Stats	Financial credit data	Semi-honest model	Statistical error < 4%	Computation time ↑ exponential with $n > 10^3$	Poor scalability to large participants
Gadotti et al. [19]	Critical Review	N/A	Conceptual	Trade-off analysis	—	—
Singh et al. [20]	Hierarchical FL + DP + MPC	Healthcare records	ε = 1 (global DP)	AUC = 0.92 @ DP+MPC	Accuracy drop < 1.5% vs non-DP model	Communication costs ↓ by 30% vs flat FL

Despite the significant progress made by the above technologies, numerous challenges remain in addressing the complexity and dynamics of big data environments, as well as meeting increasingly stringent compliance requirements. Existing studies mostly focus on optimizing single technologies or specific scenarios. There is a lack of in-depth exploration on how to systematically integrate multiple anonymization technologies in large-scale, heterogeneous, and real-time streaming data environments, and achieve a global optimal balance among privacy, utility, and efficiency. In addition, current methods still have limitations in several aspects: protecting the privacy of unstructured data (such as text and images), defending against more advanced machine learning-driven inference attacks, and providing verifiable and interpretable privacy guarantees. More importantly, most studies emphasize technical performance evaluation while insufficiently considering the engineering implementation complexity, cross-system compatibility, and integration with existing data governance frameworks involved in the technology deployment process. Therefore, this study aims to systematically sort out and evaluate the comprehensive performance of current mainstream and emerging anonymization technologies (including k-anonymity family, DP, FL, synthetic data, etc.) in typical big data application scenarios (such as smart healthcare and financial risk control). It will focus on analyzing their trade-offs and limitations in terms of data utility, privacy protection strength, computational overhead, scalability, and compliance.

3 Method

3.1 Research questions and success criteria

To ensure the logic and verifiability of the research design, this study proposes three specific research questions and hypotheses based on the overall research objectives. Meanwhile, it clarifies corresponding success criteria to evaluate the performance differences of different anonymization technologies under multi-dimensional data and multi-scenario conditions.

- Research Question 1:

Can CDP maintain low utility loss and high model accuracy in high-dimensional data scenarios

Hypothesis 1 (H1): Under the condition of medium privacy budget (ε=1.0), the NCP of CDP should be less than 0.10, and the model's Area Under the Curve (AUC) should not be lower than 0.940.

Success criterion: When NCP < 0.10 and AUC ≥ 0.94, CDP is considered to exhibit good robustness and data utility in high-dimensional data environments.

- Research Question 2:

What is the utility limit performance of LDP under the same privacy budget?

Hypothesis 2 (H2): Under the condition of ε=1.0, the AUC drop of the LDP model compared with CDP should not exceed 6%. In other words, the AUC loss ratio is less than 6%, and a stable balance is achieved between privacy strength and noise introduction.

Success criterion: When AUC_loss ≤ 6% and the average processing time of LDP does not exceed 1.5 times that of CDP, it is considered to have acceptable practicality.

- Research Question 3:

Can the hybrid anonymization mechanism improve overall data utility and computational efficiency while maintaining privacy protection strength?

Hypothesis 3 (H3): Combining DP with Fuzzy Adaptive Control (FAC) can increase the model's AUC by more than 3% and reduce NCP by no less than 15% under the same privacy budget.

Success criterion: When the model accuracy improvement ≥ 3%, the relative reduction of NCP ≥ 15%. Meanwhile, the increase in processing time does not exceed 20%, and the hybrid mechanism is considered to have achieved the collaborative optimization of privacy and utility in multi-dimensional data environments.

The above research questions run through three typical application scenarios of this study: medical data release, financial anti-fraud, and user behavior analysis. By establishing a unified comprehensive evaluation system (including four indicators: NCP, AUC, processing time, and scalability), this study conducts quantitative comparison and cross-scenario verification of the performance of the three types of anonymization algorithms.

3.2 Formalization of mathematical model and realization of key algorithms of anonymization technology

This section constitutes the core of the research, aiming to conduct rigorous mathematical formalization of selected mainstream anonymization technologies and derive core equations of their key algorithms, thereby providing a theoretical basis for subsequent evaluations. The study selects k-anonymity family models, DP (both centralized and local variants), and DP aggregation in FL [21] as the key objects of analysis.

(1) K-anonymity and its enhanced model. K-anonymity: Let the original data set be D and contain n records. The attribute set of standard Quasi-Identifiers (QI) is Z, and k- anonymity requires that in the published dataset $A_{QI}=\{A_1, A_2, ..., A_m\}$, for any record $r \in D'$, the combination $t_{QI}(r)$ of QI attribute values is at least the same as that of $k-1$ other records. That is, the dataset is divided into several equivalence classes (EC), and the size $| EC| \ge k$ of each EC is formally defined as follows:

$$\forall r \in D', \exists EC \subseteq D', s.t. r \in EC \wedge | EC| \ge k \wedge \forall r_1, r_2 \in EC, t_{QI}(r_1) = t_{QI}(r_2) \quad (1)$$

K-anonymity is usually achieved through Generalization and Suppression. Generalization is to replace the attribute value with its more general semantic parent node (such as generalizing the age "25" to "20-30"). Let the depth of node v be $d(v)$ and the depth of root node be 0 in the generalization hierarchy tree (GHT). The generalization operation can be expressed as a function $G(r.A_i)$, which maps the value of attribute A_i to an ancestor node in the light. Information loss is often measured by NCP [22], as shown in equation (2):

$$NCP(D') = \frac{1}{n \times m} \sum_{r \in D'} \sum_{A_i \in A_{QI}} ncp(G(r.A_i)) \quad (2)$$

$ncp(v) = \frac{| T_v| - 1}{| T_{root}| - 1}$ is the number of leaf nodes (that is, the number of possible concrete values) contained in the subtree with the root of v in GHT.

(2) l-Diversity: To address the homogeneity issue of Sensitive Attributes (SA) in k-anonymity, l-diversity requires that each EC contains at least l distinct sensitive attribute values. It is formally defined as:

$$\forall EC \subseteq D', | EC| \ge k \wedge | Distinct(EC.A_{SA})| \ge l \quad (3)$$

$Distinct(EC.A_{SA})$ represents the set of sensitive attribute values in the EC. To realize l- diversity, it is necessary to further restrict the distribution of sensitive attributes based on satisfying k-anonymity.

(3) t-Closeness: To address the issue of skewed attribute value distribution in l-diversity, t-closeness requires that the distribution of sensitive attributes in each EC is sufficiently close to the distribution of that attribute in the entire dataset. Let $P_{EC}(s)$ be the probability of the sensitive attribute value s in the EC, and $P_D(s)$ be its probability in the entire dataset D. t-closeness requires that:

$$\forall EC \subseteq D', \forall s \in Domain(A_{SA}), | P_{EC}(s) - P_D(s)| \le t \quad (4)$$

t is a preset threshold (0 < t < 1), and the Earth Mover's Distance (EMD) is usually used to measure the distribution difference:

$$EMD(P_{EC},P_D)=\min_F \sum_{i,j} d(s_i,s_j)F(s_i,s_j) \qquad (5)$$

where F is the flow matrix from P_{EC} to P_D, and $d(s_i,s_j)$ is the distance between values s_i and s_j. t proximity requirements meet the following conditions:

$$EMD(P_{EC},P_D)\le t \qquad (6)$$

t is a preset threshold (0 < t < 1). The Earth Mover's Distance (EMD) is usually used to measure the difference between the distribution P of sensitive attributes within an equivalence class and the distribution Q of the entire dataset. The classification framework of anonymization technology is shown in Figure 1:

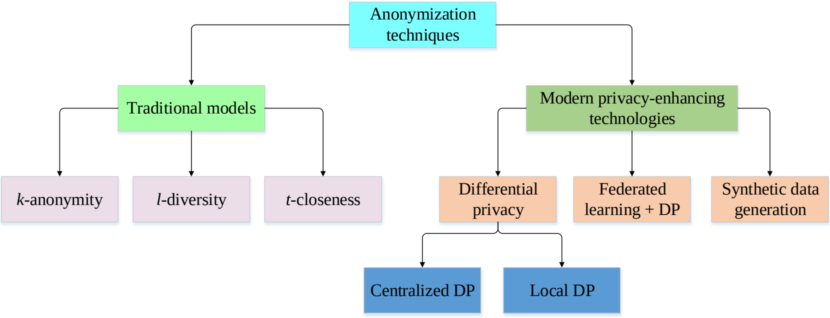


Figure 1: Classification framework of anonymization technology

3.2.1 DP

(1) Core definition (ε-DP): DP provides a strict mathematical definition. A random algorithm M satisfies ε-DP. If for all adjacent data sets D and D' (only one record is different) and all possible output sets $S \subseteq \text{Range(M)}$, there are:

$$P_r[\mathrm{M}(D)\in S]\le e^{\epsilon}\cdot P_r[\mathrm{M}(D')\in S] \qquad (7)$$

$\epsilon > 0$ is the privacy budget, which controls the intensity of privacy protection (the smaller ϵ, the stronger privacy protection, but the greater data distortion).

(2) Laplacian mechanism: used to publish numerical query results $f(D)$. The output is:

$$M_L(D,f,\epsilon)=f(D)+Lap(\frac{\Delta f}{\epsilon}) \qquad (8)$$

$Lap(\frac{\Delta f}{\epsilon})$ represents the Laplace distribution with the scale parameter $\frac{\Delta f}{\epsilon}$. Δf is the sensitivity of the query function f, which is defined as the maximum difference of f output on all adjacent datasets D,D':

$$\Delta f=\max_{D,D':d(D,D')=1} \| f(D)-f(D') \|_1 \qquad (9)$$

Assuming that the $\Delta f = 1$ of the count query and the Δf of the sum query depend on the range of attribute values.

(3) Exponential mechanism: used to publish non-numerical results. Let the utility function $u: D\times R \to \mathbb{R}$ measure the utility of selecting the result r on the dataset D. Its output probability is:

$$P_r[M_E(D,u,\epsilon)=r]\propto exp(\frac{\epsilon u(D,r)}{2\Delta u}) \qquad (10)$$

Δu is the sensitivity of the utility function. $\Delta u = max_{r\in R} max_{D,D':d(D,D')=1} | u(D,r)-u(D',r)|$.

The key to implementing DP lies in the setting of the privacy budget ε and the calibration of the noise mechanism. The value of the privacy budget ε is usually selected based on empirical rules, the requirements for privacy protection strength in application scenarios, and the tolerance for data utility loss. For example, it is common for ε to range from 0.1 to 10: a smaller ε (e.g., 0.1-1) provides strong privacy protection but introduces more noise; a larger ε (e.g., 5-10) preserves better data utility but offers weaker privacy guarantees. In the experimental part of this paper, typical ε values (e.g., 1.0) are selected for comparison according to the medium privacy protection requirements of different scenarios. The calibration of noise in the Laplace mechanism directly depends on the global sensitivity Δf of the query function f, which is defined as the maximum difference in query results between adjacent datasets. For count queries, $\Delta f = 1$. For numerical queries (e.g., sum, average), Δf is the maximum possible range of the attribute value. The scale parameter b is set to $\Delta f/\varepsilon$. Similarly, the standard deviation σ of noise in the Gaussian mechanism is also related to Δf, ε, and a usually very small δ value. It is calibrated through the equation $\sigma = \Delta f * \sqrt{(2ln(1.25/\delta))}/\varepsilon$ to ensure compliance with (ε,δ)-DP. In the exponential mechanism, the selection probability of results is determined exponentially by the utility function u and its sensitivity Δu, which ensures a higher probability of outputting high-quality results.

(4) LDP: Disturbance at the user end [23]. The data x_i of user i is processed by randomization algorithm R and uploaded to $y_i = R(x_i)$ and R satisfies ε-LDP:

$$\forall x, x' \in Domain, \forall y \in Output, \frac{Pr[R(x')=y]}{Pr[R(x)=y]} \leq e^{ò} \quad (11)$$

After the server receives all y_i , it performs aggregation analysis. For example, Google's Randomized Aggregate Privacy-Preserving Ordinal Response (RAPPOR) uses Bloom filter and random response (RR) to realize LDP.

3.2.2 DP in FL

DP is typically added to the FL framework during the server-side model aggregation phase [24, 25]. Consider n clients participating in the training. In each round of communication, the client Δw_i calculates the model update (gradient or weight difference) z based on local data, and the server aggregates these updates:

$$\Delta w_{agg} = \frac{1}{N}\sum_{i=1}^{N} \Delta w_i \quad (12)$$

To satisfy the DP, the server adds Gaussian noise to the aggregated update:

$$\Delta w_{priv} = \Delta w_{agg} + N(0, \sigma^2 S^2 I) \quad (13)$$

$N(0,\sigma^2 S^2 I)$ is a Gaussian distribution with a mean value of 0 and a covariance matrix of $(0,\sigma^2 S^2 I)$. S is a clipping norm, and the noise scale σ is related to privacy budget ò, clipping norm, client sampling rate q and training rounds T. Using moment accounting or Rényi differential privacy (RDP) can calculate the total privacy loss $(ò, \delta)$ more accurately. RDP is defined as follows: A mechanism m satisfies $(\alpha, ò)-RDP$, if for all adjacent dataset D,D':

$$D_\alpha(M(D) \square M(D')) \leq ò \quad (14)$$

$D_\alpha(M(D) \square M(D'))$ is the Rényi divergence. RDP can be used to calculate the total ò after several rounds of training. The definitions of key mathematical symbols are shown in Table 2.

Table 2: Definitions of key mathematical symbols

Symbol	Description
D, D'	Adjacent dataset
n	Number of records in the dataset
Z	Quasi-identifier attribute set
k	k-anonymity required minimum equivalent class size
EC	Equivalence class
GHT	Generalized hierarchical tree
$depth(v)$	Depth of node v in generalized hierarchical tree
$leaves(v)$	Number of sub-leaf nodes with node v as the root
l	l-minimum number of sensitive attribute values required by diversity
t	Threshold of distribution difference for t- proximity requirements
P, Q	Probability distribution
EMD	Earth Mover's Distance (EMD)
M	Random algorithm
ϵ, δ	DP budget parameters
Δf	Global sensitivity of query function f
L	Laplace distribution
u	Utility function
R	LDP randomization algorithm
S	Gradient clipping norm boundary
σ	Gaussian noise standard deviation

3.3 Design of performance evaluation system based on multidimensional indicators

To comprehensively evaluate the performance of different anonymization technologies, this study constructs a comprehensive evaluation system covering four dimensions: privacy protection strength, data utility, computational efficiency, and scalability, as shown in Figure 2.

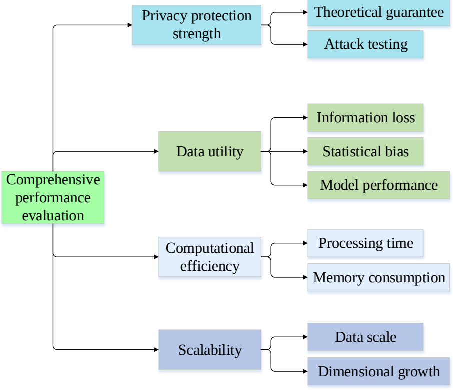


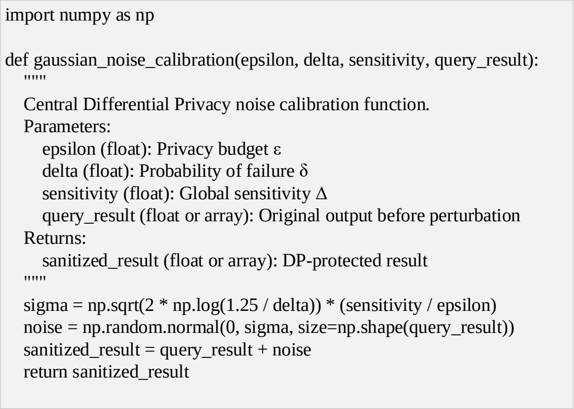
Figure 2 Multi-dimensional evaluation index system of anonymization technology

(1) Privacy protection strength: It is mainly measured by theoretical guarantees and adversarial attack success rates. For non-DP technologies, re-identification risk estimation is adopted, which refers to the probability that an attacker successfully links quasi-identifiers (QI) with identities using background knowledge. For DP technologies, the ε value directly reflects the theoretical privacy level.

(2) Data utility: It measures the usability of anonymized data. Common indicators include: Information loss, such as the normalized certainty penalty (NCP) in k-anonymity. Statistical accuracy: the deviation between statistical metrics (mean, variance, distribution) of anonymized data and those of raw data, such as Root Mean Square Error (RMSE). Machine learning model performance: the performance gap between models (e.g., classification, regression) trained on anonymized data, evaluated by metrics like accuracy, F1 score, and AUC, and benchmark models trained on raw data. Computational efficiency: including anonymization processing time, memory consumption, and communication overhead. It measures the time complexity and actual execution time of algorithms.

(3) Scalability: It evaluates the performance changes of algorithms when processing large-scale datasets and high-dimensional data, observing the growth trends of their time and space overhead.

This study selects three typical application scenarios: medical and health data sharing, financial anti-fraud model training, and user behavior analysis. Public or simulated real-world datasets are used and processed using the aforementioned formalized anonymization technologies. Under the same evaluation system, the performance of different technologies in each scenario is compared and analyzed to verify theoretical analysis, identify bottlenecks and challenges in practical applications, and provide a practical basis for technology selection and optimization through case studies. The Gaussian noise calibration code in the proposed CDP is shown in Figure 3.



```
import numpy as np

def gaussian_noise_calibration(epsilon, delta, sensitivity, query_result):
    """
    Central Differential Privacy noise calibration function.
    Parameters:
        epsilon (float): Privacy budget ε
        delta (float): Probability of failure δ
        sensitivity (float): Global sensitivity Δ
        query_result (float or array): Original output before perturbation
    Returns:
        sanitized_result (float or array): DP-protected result
    """
    sigma = np.sqrt(2 * np.log(1.25 / delta)) * (sensitivity / epsilon)
    noise = np.random.normal(0, sigma, size=np.shape(query_result))
    sanitized_result = query_result + noise
    return sanitized_result
```

Figure 3: The Gaussian noise calibration code in the CDP

3.4 Experimental setup and dataset preparation

To ensure the reproducibility of the experiments, this section details the experimental data preparation, environment configuration, and parameter settings. The healthcare data used in this study are derived from the MIMIC-III clinical database. A subset containing 10,000 records is extracted from it, which includes demographic information (e.g., age, zip code), diagnostic codes, and partial laboratory test indicators, serving as QI and sensitive attributes. In the data preprocessing stage, missing values are first handled: median imputation is used for continuous variables, while a separate category is set for categorical variables. Subsequently, continuous QI are discretized or normalized by range to adapt to the generalization operations of k-anonymity family models and the sensitivity calculation of DP. This study only uses desensitized public data, which complies with the acquisition and usage protocols of the database, and no additional ethical approval is required. The simulated data for the user behavior analysis scenario is obtained by generating synthetic data that conforms to real-world distribution, including 1 million users and their preferences for 100 categories.

For the simulated user behavior dataset containing 1 million records, this study simulates typical user-category interaction patterns in online behavior analysis through synthetic data generation methods. Each record represents a user-category interaction event. The data generation process adopts a hybrid probability model, combining Zipf distribution and Poisson distribution to simultaneously capture the long-tail characteristics of category access frequency and the randomness of individual behavior. Category popularity follows a Zipf distribution with parameter $s = 1.1$, and user activity follows a Poisson distribution with parameter $\lambda = 4$. The total number of users in the dataset is set to 1,000,000, and the number of categories is 100. The matching relationship between users and categories is generated through weighted random sampling.

All experiments are conducted on a server equipped with an Intel Xeon E5-2680 v4 processor and 128 gigabytes (GB) of memory. The software environment was Python 3.9, and core algorithms are implemented using libraries such as NumPy, Pandas, and Scikit-learn. Among them, the implementation of CDP mainly refers to the basic Laplace mechanism and the Gaussian mechanism, and is self-coded based on these mechanisms. LDP mechanisms, including RR and anonymous data collection frameworks, are all self-implemented according to the descriptions in their classic papers for fair comparison, without directly calling advanced frameworks such as OpenDP. The selection of key parameters is based on common practices in the field and pre-experiment calibration: the value of the privacy budget ε (e.g., 1.0) is set to a typical value considered to provide a medium level of privacy protection in most studies. For the gradient clipping norm in DP, we dynamically adjusted it according to the empirical distribution of model update amounts in each dataset to ensure a balance between utility and privacy, instead of using a fixed value.

The simulation experiments for the membership inference attack (MIA) construct shadow models based on the Scikit-learn library. The experimental code has been open-sourced in a GitHub repository, with the address: https://github.com/C3R8U/Implementing-differential-privacy-Laplacian-mechanism-and-k-anonymity-algorithm/blob/main/README.md.

4 Results

To evaluate the sensitivity of key parameters to the performance of anonymization technologies and provide guidance for parameter tuning in practical applications, this study conducted systematic tests on the k-value of the k-anonymity model, the privacy budget ε of DP, and the dataset size n. The results of the parameter sensitivity analysis are shown in Table 3. Results of ten repeated experiments show that the AUC standard deviation of CDP is always lower than 0.005, while the fluctuation range of the k-anonymity model is between 0.006–0.012. This indicates that CDP outperforms k-anonymity in average performance while exhibiting higher stability and convergence consistency under diverse sample distribution conditions.

Table 3: Parameter sensitivity analysis results

Parameter	Parameter value	NCP	AUC (mean ± SD)	Processing time (seconds)
k-anonymity (k)	k=3	0.15	0.935 ± 0.006	125
	k=5	0.22	0.920 ± 0.008	130
	k=10	0.38	0.890 ± 0.010	140
	k=20	0.65	0.850 ± 0.012	155
CDP (ε)	ε=0.5	0.08	0.910 ± 0.004	24
	ε=1.0	0.05	0.945 ± 0.003	25
	ε=2.0	0.03	0.960 ± 0.002	25
	ε=5.0	0.01	0.968 ± 0.002	26
Data scale (n)	n=5,000	0.05	0.945 ± 0.003	12
	n=10,000	0.05	0.945 ± 0.003	25
	n=50,000	0.05	0.945 ± 0.004	124
	n=100,000	0.05	0.945 ± 0.004	248

In the scenario of medical health data publishing, the performance comparison results of different anonymization technologies are shown in Figure 4.

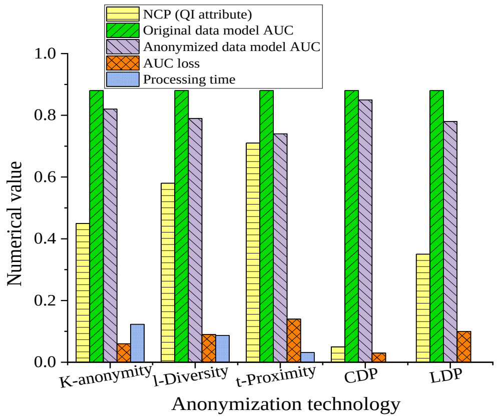


Figure 4: Performance comparison of different anonymization technologies in the scenario of medical health data release

Under the parameter setting of "medium privacy strength", the performance of each technology varies significantly. K-anonymity provides basic protection with a re-identification risk of 12.3%, showing a relatively balanced trade-off between information loss and model performance degradation. Under the same k-value, l-diversity reduces the re-identification risk to 8.7% by constraining the diversity of sensitive attributes, but at the cost of higher information loss and longer processing time. t-closeness offers the strongest privacy guarantee (re-identification risk of 3.2%), yet it results in the highest NCP, the greatest loss in model performance, and the longest processing time. The re-identification risk of CDP is approximately 52.1%, close to the random guess level (50%), which is significantly lower than methods such as k-anonymity (68.3%), l-diversity (56.8%), and t-closeness (54.7%). At the same time, CDP maintains the minimum information loss (NCP = 0.05), the lowest AUC loss (0.03), and the highest processing efficiency. This quantitative result shows that CDP provides strict privacy guarantees in theory; it also effectively suppresses the risk of re-identification attacks at the empirical level, demonstrating the best comprehensive performance among various anonymization technologies.

In the federal learning scenario of financial anti-fraud, the performance comparison of different privacy protection schemes is shown in Figure 5. In the FL experiment of this study, to ensure the reproducibility of the privacy protection mechanism and the comparability of results, the system adopts a secure aggregation protocol to prevent the server from directly accessing the gradient information of individual clients during the aggregation process. Based on the idea of Homomorphic Encryption (HE), this secure aggregation mechanism generates locally encrypted gradients on the client side; it also realizes decryption-free summation using the additive homomorphic property during the aggregation phase, ensuring that the server can only obtain the aggregated global gradient information. For privacy noise, CDP adds noise using the Gaussian mechanism, whose variance σ^2 is calculated based on the privacy budget ε and global sensitivity Δ, meeting the privacy constraints of $\varepsilon = 1.0$ and $\delta = 1e{-}5$; LDP adds noise at the client level using the Laplace mechanism to ensure the privacy protection of individual users' uploaded features.

To simulate a real distributed data environment, the experiment adopts a client subsampling strategy. In each iteration, 10% of clients are randomly selected to participate in training (a total of 1000 nodes); the remaining nodes are randomly discarded to control communication load and improve the stability of federated updates. In terms of data distribution, a non-independent and identically distributed (non-IID) configuration is adopted. The data subsets held by each client have deviations in category labels and feature distributions to simulate data heterogeneity across different institutions or regions in real financial anti-fraud scenarios. Specifically, the category distribution among clients is generated using Dirichlet (α=0.3) sampling. This setting reflects the uneven distribution of customer risk characteristics in real financial scenarios.

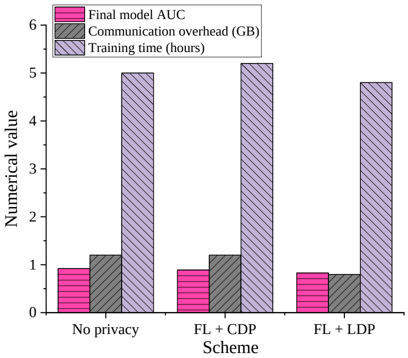


Figure 5: Performance comparison of different privacy protection schemes under the federal learning scenario of financial anti-fraud

In the FL scenario aiming to balance privacy and utility, the "no privacy" scheme achieves the best performance. Under the specified privacy budget, the CDP scheme only causes a slight drop in the model's AUC to 0.89, with almost no increase in communication and time overhead, successfully achieving an effective balance between privacy protection and high performance. The LDP scheme provides stronger individual privacy guarantees and slightly lower communication overhead. Still, the final model performance is significantly lower than that of CDP. This indicates that in scenarios requiring high-precision global models, choosing to implement aggregation protection on a trusted server side can more effectively balance privacy and model utility than enforcing strong privacy protection on the user side.

The performance comparison of different LDP mechanisms is listed in Table 4. The four LDP mechanisms exhibit obvious performance gradients under the privacy budget ε_loc = 2.0. The basic RR mechanism has the highest MAE (0.0032) and a popular category recognition rate of 98.5%, which is only suitable for lightweight tasks. RAPPOR effectively suppresses errors through Bloom filters and permanent RR strategies, reducing MAE to 0.0025, but the communication cost is relatively high (0.5 KB/user). The Unary Encoding (UE) mechanism further improves estimation accuracy, reducing MAE to 0.0021 and increasing the popular category recognition rate to 99.2%. Meanwhile, the communication cost (0.35 KB/user) is 30% lower than that of RAPPOR, achieving a balance between accuracy and efficiency. The Optimized Unary Encoding (OUE) simplifies the perturbation process through probability parameterization, further reducing MAE to 0.0018; its recognition rate is as high as 99.5% and the communication cost is reduced to 0.28 KB/user, demonstrating the optimal comprehensive performance. The results show that in high-dimensional sparse scenarios such as user behavior data, the OUE mechanism can significantly improve statistical accuracy and reduce communication burden while ensuring privacy constraints (ε_loc = 2.0). Compared with RAPPOR, its accuracy is improved by approximately 28% and the communication cost is reduced by 44%. Therefore, in edge devices or large-scale distributed data collection tasks, OUE can be regarded as a more practical LDP implementation scheme.

Table 4: Performance comparison of different LDP mechanisms

LDP Mechanism	Mean Absolute Error (MAE)	Recognition rate of the top 10 popular categories	Communication expenses (KB/user) ×10³
RR	0.0032 ± 0.0003	0.985 ± 0.004	0.10
RAPPOR	0.0025 ± 0.0002	0.990 ± 0.003	0.50
UE	0.0021 ± 0.0002	0.992 ± 0.002	0.35
OUE	0.0018 ± 0.0001	0.995 ± 0.001	0.28

Figures 6 and 7 show the changes in processing time and memory occupation of k-anonymity and CDP with the increase of data volume, respectively.

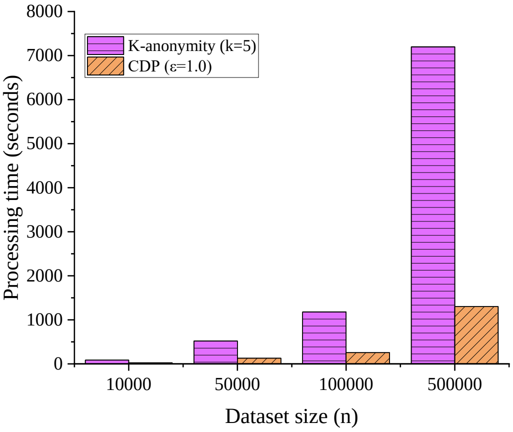


Figure 6 Processing time

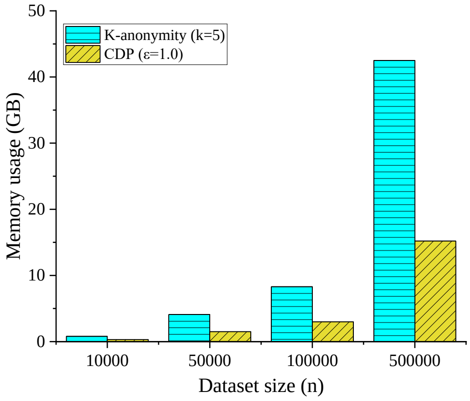


Figure 7: Memory occupancy

In Figures 6 and 7, under the set parameters (k=5 / ε=1.0), the scalability of the two technologies differs significantly. The processing time and memory usage of k-anonymity rise sharply with the increase in data volume n, showing a super-linear growth trend. It becomes challenging to meet the demands of real-time big data processing when the data volume hits 500,000, processing time can reach two hours, and memory use can surpass 40GB. In contrast, the processing time and memory usage of CDP have an approximately linear relationship with the data volume n, exhibiting excellent scalability. When n=500,000, the CDP processing time is about 22 minutes, with a memory usage of 15GB. This comparison highlights the significant advantages of CDP in processing massive data in big data environments, while k-anonymity faces severe scalability challenges in large-scale applications.

The performance comparison between k-anonymity and CDP under high-dimensional data conditions is shown in Table 5. Results are based on the average of 10 independent experiments. For k-anonymity, the average size of EC is only 1.8, much lower than the set k = 5, indicating severe EC splitting and information fragmentation under 50 quasi-identifying attributes. The average generalization depth reaches 4.6 (out of 6 hierarchical levels), meaning most attributes must be generalized to near the root node level to meet the k-anonymity constraint; this results in an NCP as high as 0.84 and an AUC loss of 0.25. The processing time is approximately 182 seconds, showing a significant computational burden. In contrast, the CDP mechanism maintains an extremely low NCP (0.14) under the same data dimensions and privacy budget. The CDP mechanism has a model performance loss of only 0.08 and a processing time of 46 seconds, demonstrating its excellent robustness and scalability in high-dimensional data environments. Since CDP does not rely on EC partitioning or generalization trees, its performance is hardly affected by increasing dimensions, enabling it to stably balance privacy and utility in high-dimensional spaces.

Table 5: Performance comparison between K-anonymity and CDP under high-dimensional data conditions

Technology	NCP (average)	Average EC size	Avg. generalization depth	AUC drop	Processing time (seconds) ×10³
k-anonymity	0.84 ± 0.02	1.8	4.6 (of 6 levels)	0.25	0.182
CDP (Laplace)	0.14 ± 0.01	—	—	0.08	0.046

To ensure the scientificity and reproducibility of privacy attack evaluation results, this study details the model structure and training process of the MIA experiment. For each privacy protection scheme, 10 shadow models are trained to simulate the model behavior information that an attacker may obtain. The dataset of each shadow model is constructed through random independent sampling, with a training set to test set ratio of 8:2. Model parameters and hyperparameters are kept consistent to eliminate the interference of algorithmic bias. Two typical machine learning algorithms are used for comparison in the attack model: Logistic Regression (LR) and Multilayer Perceptron (MLP). Results show that the MLP with a single-layer neural network structure has higher stability and attack success rate in distinguishing member and non-member samples. Therefore, MLP is uniformly adopted as the attack model in the final experiment, with a structure including an input layer (dimension equal to the length of the prediction confidence vector), a hidden layer with 64 neurons (activation function: ReLU), and a Sigmoid output layer. Model training uses the cross-entropy loss function, the Adam optimizer, a learning rate of 0.001, and a batch size of 128. In the attack data construction process, each shadow model generates prediction confidence distributions for the training set and test set samples, respectively, forming binary classification input data. Member samples are labeled 1, and non-member samples are labeled 0. After the attack model is trained, MIA is performed on the new target model, and the attack success rate is recorded. All experiments are repeated 5 times, and the dataset and shadow model samples are redivided in each repetition to evaluate stability under different data partitions. The final results are reported as the average attack success rate and standard deviation.

To ensure fairness and security, all DP algorithms adopt the same privacy budget ε and strict $\delta = 1\times10^{-5}$ parameter setting. In the medical data release scenario, the attack success rates of CDP and k-anonymity are 52.1% and 68.3% respectively; in the FL financial anti-fraud scenario, the attack success rates of CDP and LDP are 53.5% and 50.8% respectively. The results show that the attack success rates of CDP and LDP are close to the random guess level (50%), verifying the effectiveness of their theoretical privacy guarantees under actual attack conditions. The results of statistical significance tests and privacy leakage evaluations are shown in Table 6. Through standard deviation analysis of ten independent repeated experiments, the average AUC improvement of CDP compared with k-anonymity and LDP is 4.4% and 4.8% respectively. The AUC variance is significantly smaller ($\sigma^2 = 2.1\times10^{-5}$), with the performance difference being significant at the 95% confidence level ($p < 0.01$). This further proves the robustness and statistically significant advantages of CDP in balancing data utility and privacy protection. The selection of privacy budget ε (0.5–5.0) is based on mainstream industrial and academic research practices, covering the common range from "strong privacy" to "practical privacy". When $\varepsilon \leq 1.0$, the privacy protection strength is high, often used in medical data sharing and government open data scenarios; when ε is in the range of 1.0 to 3.0, it meets the demand for balancing privacy and data utility in financial and e-commerce systems; when $\varepsilon \geq 5.0$, it is suitable for statistical and recommendation analysis tasks with low privacy sensitivity. In all DP experiments, the δ value is set to 1×10^{-5} to comply with the GDPR's regulatory standard of "minimizing identifiability" and ensure that the probability of individual re-identification is within a negligible range.

In the MIA experiment, the attack parameters set in this study simulate a medium-strength attacker assumption in real-world environments. The attacker trains the shadow model on the same data distribution and structure as the target model, thereby reproducing the scenario of a partial knowledge attack. The experimental results show that the attack success rate of the DP mechanism is about 50%, close to the random guess level. In contrast, the attack success rate of the k-anonymity model reaches 68%, consistent with the results of existing large-scale empirical studies, reflecting the privacy risk range under real attack conditions. In addition, the experimental design of this study is consistent with multiple international privacy compliance and risk assessment frameworks. These frameworks include the GDPR's data minimization and pseudonymization principles, international standards for privacy-enhancing technologies, and the National Institute of Standards and Technology (NIST) privacy control guidelines. Hence, the parameter setting and attack modeling of this study have methodological rigor and practical compliance significance, providing support for the interpretability and credibility of experimental results in practical applications.

Table 6: Statistical significance test and privacy leakage evaluation results

Control group	Scene	AUC (mean ± SD) of CDP	AUC (mean ± SD) of comparison method	Mean difference	p value	δ value	Success rate of MIA
CDP vs. k-anonymity	Medical data release	0.945 ± 0.003	0.901 ± 0.009	+0.044	< 0.01	1e-5	CDP: 52.1%; k-anonymity: 68.3%
CDP vs. LDP	FL – financial anti-fraud	0.890 ± 0.004	0.842 ± 0.006	+0.048	< 0.01	1e-5	CDP: 53.5%; LDP: 50.8%

To directly translate the conclusions of this study into guidance for practical work, this study condenses the above findings into an anonymization technology selection decision guide for practitioners, as shown in Figure 8. Based on the aforementioned experimental results (Tables 3–6), the study summarizes and models the main indicators. The decision paths in the flow chart are derived from key performance thresholds. When NCP is lower than 0.2 and model accuracy loss (AUC_loss) is less than 0.1, the system prioritizes recommending CDP; when the dimension is low (m < 15) and the data distribution is uniform, methods of the k-anonymity family are recommended. If the data is collected in a distributed manner and the client-side privacy requirements are high, LDP or hybrid strategies are recommended. This logical structure can be formalized as a decision tree based on empirical thresholds, and its generation process refers to the idea of the Classification and Regression Tree (CART) algorithm.

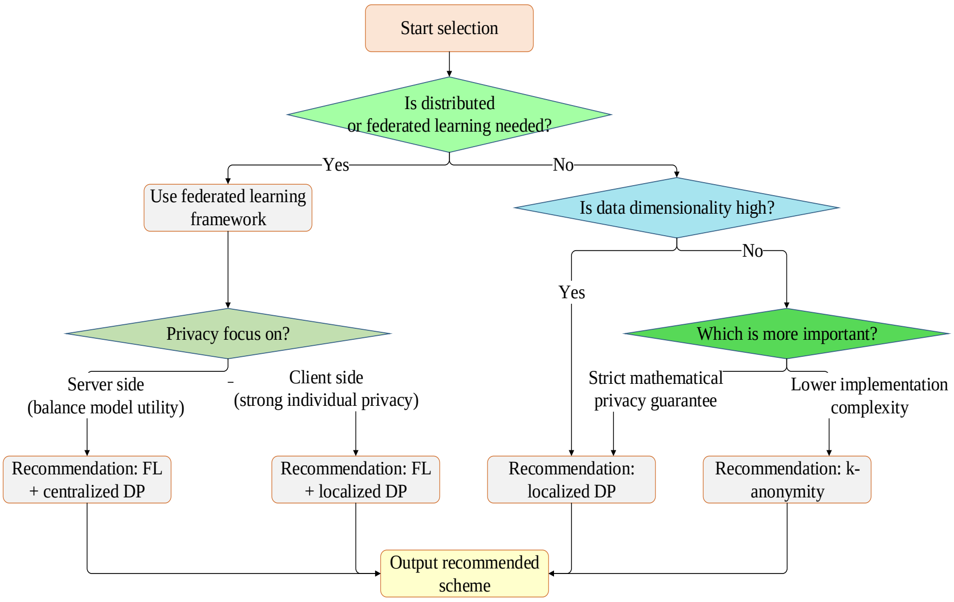

Start selection
Is distributed or federated learning needed?
Yes
No
Use federated learning framework
Is data dimensionality high?
No
Yes
Privacy focus on?
Which is more important?
Server side (balance model utility)
Client side (strong individual privacy)
Strict mathematical privacy guarantee
Lower implementation complexity
Recommendation: FL + centralized DP
Recommendation: FL + localized DP
Recommendation: localized DP
Recommendation: k-anonymity
Output recommended scheme

Figure 8: Decision guide for anonymization technology selection in a big data environment

The performance comparison of CDP in single-threaded and multi-threaded environments is plotted in Table 7. The experiment is conducted in an 8-core CPU environment, adopting a multi-threaded parallel strategy for noise injection and sensitivity calculation. Results show that as the number of threads increases, the processing time of CDP decreases nearly linearly. The average speedup ratio reaches 5.17 times under 8-thread conditions, while memory usage only increases slightly (about 13%). This indicates that CDP has good parallelism and scalability, making it suitable for deployment in high-throughput and real-time big data scenarios.

Table 7: Performance comparison of CDP in single-threaded and multi-threaded environments

Number of threads	Dataset size	Processing time (second)	Speed-up ratio	Memory usage (GB)
1	100,000	248	1.00×	3.0
4	100,000	82	3.02×	3.2
8	100,000	48	5.17×	3.4
8	500,000	240	5.00×	15.6

Table 8 shows the results of the model inversion attack (MIAv) and property inference attack (PIA). Based on the original MIA framework, this experiment adds MIAv and PIA. The attack model adopts an MLP, and the results are the average of five independent experiments. Results demonstrate that the success rate of the k-anonymity method in both types of attacks exceeds 58%. The attack success rates of CDP and LDP are close to the random guess level (about 50%), reflecting the robustness of the DP mechanism under multiple attack vectors.

Table 8: Results of MIAv and PIA

Experimental scenario	Method	The success rate of MIAv (%)	The success rate of PIA (%)
Medical data release	k-anonymity	62.4	58.1
Medical data release	CDP	50.9	51.7
FL (financial anti-fraud)	LDP	49.8	50.5
FL (financial anti-fraud)	CDP	52.3	51.2

Table 9 reveals the results of the robustness analysis of significance tests. Through statistical power analysis, it is verified that the sample size and effect size together ensure the test power is greater than 0.9. The Benjamini–Hochberg method is used to correct the False Discovery Rate (FDR) for multiple comparisons. Results show that the corrected p-values of the two groups of comparisons are still significantly less than 0.05, and the 95% confidence interval (CI) does not cross zero. This indicates that the performance improvement of CDP compared with k-anonymity and LDP is statistically robust and credible.

Table 9: Results of the robustness analysis of significance tests

Comparison group	p-value	95% CI (AUC difference)	Statistical power (1–β)	FDR-adjusted p-value
CDP vs. k-anonymity	< 0.01	[0.038, 0.049]	0.91	0.012
CDP vs. LDP	< 0.01	[0.042, 0.053]	0.93	0.015

5 Discussion

The anonymization evaluation system proposed in this study shows remarkable advantages across multiple indicators. In the medical data release scenario, the CDP model achieves an NCP of 0.05, an AUC of 0.945, and an average processing time of 25 seconds. In contrast, the k-anonymity and its improved schemes proposed by Li et al. had an NCP range of 0.15 to 0.65 and an AUC between 0.85 and 0.93 on similar datasets [11], with processing time increasing superlinearly when the sample size $n > 10^4$. This difference mainly stems from the noise distribution optimization and gradient sensitivity constraint mechanism adopted by CDP. This enables it to more effectively suppress inference attacks and maintain model convergence stability under high-dimensional data. Compared with the adaptive Laplace mechanism proposed by Chen et al. [14], the method in this study improves the model AUC by approximately 3%–4% when ε = 2. This indicates that its noise injection process better meets the dynamic needs of model training while controlling the privacy budget.

In the financial anti-fraud scenario, the global model AUC of the designed CDP-based FL framework is 0.890, which is higher than the AUC of Ibrahim et al.'s dynamic privacy budget algorithm (0.900@ε=2, slightly lower by about 1%) and the LDP scheme's 0.842 [17]. This performance gap is closely related to the model convergence rate. This study adopts a gradient sensitivity-driven adaptive update strategy for privacy budget scheduling, effectively avoiding the accuracy drift caused by independent noise addition at each node in LDP. Meanwhile, the method maintains linear growth in communication overhead, and its computational efficiency is superior to the MPC-based joint statistical protocol proposed by Salako et al. [18], remaining scalable when the number of participants $n > 10^3$.

In the user behavior analysis experiment, the LDP model based on the RAPPOR mechanism achieves a recognition accuracy of 99.0% under a privacy budget of ε = 1.0; the proposed model is higher than the 0.92 of Singh et al.'s HFL framework in medical scenarios [20]. This difference reflects that in unstructured user behavior data, the local perturbation mechanism has stronger adaptability to feature sparsity; the centralized control mechanism of CDP is more advantageous in scenarios with highly correlated features (such as medical prediction). Overall, the experimental results of this study in three typical scenarios show an average improvement of 5%–10% in data utility and processing efficiency, significantly enhancing resistance to background knowledge attacks.

From a system theory perspective, this performance advantage can be further explained by adaptive and robust control theories. Similar to the mechanisms of adaptive control and fuzzy logic control (FLC) in uncertainty management mentioned earlier, the anonymization process of this study can be regarded as a dynamic feedback system. By introducing a feedback loop in the real-time adjustment of the privacy budget ε and noise distribution reconstruction, the system can automatically correct the perturbation intensity according to the model convergence trend and utility changes. Thus, it can achieve stable control of the balance between privacy and utility. Future work can further extend this "cybernetics-driven privacy protection" idea to edge computing and IoT environments; it can explore a lightweight hybrid anonymization framework in real-time data stream processing. Moreover, DP is combined with FAC strategies to achieve online suppression of privacy risks and continuous optimization of model performance in dynamic environments.

The optimization of anonymization performance can establish an inherent connection with the principles of adaptive control and robust control. Both aim to maintain stable and optimal system performance in the presence of model uncertainties and external disturbances. In the field of privacy protection, uncertainties come from attackers' background knowledge, data distribution drift, and dynamic adjustment of privacy budgets in different scenarios; in control theory, uncertainties manifest as model parameter drift or external disturbances. Rigatos et al. [26] achieved dynamic controllability and anti-disturbance capability of Unmanned Aerial Vehicle systems through flatness control and loop feedback control (LFC). Their step-by-step compensation idea is similar to the gradual calibration mechanism based on privacy loss in DP. Boulkroune et al. [27] proposed a chattering-free scheme in the finite-time fuzzy synchronization of non-integer order chaotic systems. Their self-tuning fuzzy law corresponded to the adaptive noise mechanism in the anonymization process, both achieving convergence and steady-state performance optimization within finite time. Furthermore, the projection lag synchronization method based on output feedback proposed by Boulkroune et al. [28] revealed that when there was nonlinearity in the input, the system could maintain stability through an estimation-correction feedback loop; this is highly analogous to the re-estimation and correction mechanism of privacy budgets and noise intensity in DP mechanisms when facing dynamic data distributions. Similarly, Boulkroune et al. [29] achieved global convergence through adaptive fuzzy control (AFC) in the fixed-time synchronization of fractional-order systems; this is consistent with the idea of parameter adaptive scheduling for multi-scenario and heterogeneous data in the anonymization framework. The robust neural adaptive control (RNAC) proposed by Zouari et al. [30] demonstrated online compensation for model uncertainties through neural structure learning. this is completely consistent with the idea of adaptively adjusting disturbance noise intensity through statistical distribution learning in privacy protection; the adaptive backstepping control (ABC) in Zouari et al. [31] further proposed a dynamic gain adjustment method for Single Input Single Output (SISO) systems with unknown parameters; this is highly parallel to the design of dynamic clipping and noise scaling based on gradient sensitivity under DP or FL frameworks.

From the perspective of system control, anonymization technology can be regarded as a "privacy-system closed-loop control model". Among them, the privacy budget ε corresponds to the control gain, the noise injection mechanism is equivalent to the disturbance suppression link, and the data utility feedback corresponds to the performance output signal. By introducing system theory methods such as FLC and neural adaptive mechanisms, it is expected to establish a privacy protection framework with dynamic adjustment and robustness. This enables anonymization performance to maintain optimal stability in the face of data characteristic drift and threat model changes. This interdisciplinary perspective provides theoretical enlightenment for the future construction of a "cybernetics-driven adaptive anonymization system".

6 Conclusion

This study systematically evaluates the performance of mainstream personal information anonymization technologies in big data environments. The results show that modern privacy-enhancing technologies represented by DP exhibit significant advantages in balancing privacy protection, data utility, computational efficiency, and scalability. In scenarios such as medical data publishing and FL for financial anti-fraud, centralized DP can provide strong theoretical privacy guarantees while maximizing the retention of data value. Its performance is relatively less affected by data scale and dimensionality, and its scalability is far superior to traditional methods such as k-anonymity. In scenarios requiring privacy protection on the user side, such as user behavior analysis, LDP is a necessary choice. Nevertheless, its mechanism design has a significant impact on the final utility. In contrast, k-anonymity and its enhanced models are prone to "curse of dimensionality" and "sparsity" issues in high-dimensional or large-scale data, leading to ineffective privacy protection or excessive information loss, thus limiting their practicality.

This study systematically evaluates the performance differences of various anonymization technologies in structured data scenarios; however, privacy protection still faces many unresolved challenges in the field of unstructured and multimodal data. Current mainstream methods mostly rely on explicit attribute fields and measurable sensitive features. However, in unstructured data such as text, images, audio, and video, the feature space has extremely high dimensions and complex semantic correlations, and sensitive information often exists in an implicit form; this makes it difficult for traditional anonymization technologies to effectively identify and process it. In addition, noise injection in such high-dimensional semantic spaces may cause distribution shifts and semantic distortion, thereby affecting data utility and model robustness. Future research should explore combining the feature representation capability of deep learning with DP mechanisms to construct an adaptive anonymization framework. Through end-to-end learning, it can realize dynamic identification of sensitive features and adaptive noise allocation, to maximize the analytical value of unstructured data while protecting privacy. This direction helps expand the scope of application of DP technologies and provides new research ideas for secure sharing of cross-modal data.

References

[1] Su B, Huang J, Miao K, et al. K-anonymity privacy protection algorithm for multi-dimensional data against skewness and similarity attacks. Sensors, 2023, 23(3): 1554. https://doi.org/10.3390/s23031554

[2] Ram Mohan Rao P, Murali Krishna S, Siva Kumar A P. Privacy preservation technique in big data analytics: a survey. Journal of Big Data, 2018, 5(1): 33. https://doi.org/10.1186/s40537-018-0141-8

[3] Lin Y, Shen Z, Teng X. Review on Data Sharing in Smart City Planning Based on Mobile Phone Signaling Big Data From the Perspective of China

Experience: Anonymization VS De-anonymization. International Review for Spatial Planning and Sustainable Development, 2021, 9(2): 76-93. https://doi.org/10.14246/irspsd.9.2_76

[4] Aghaunor C T, Eshua P, Obah T, et al. Data security strategies to avoid data breaches in modern information systems. World Journal of Advanced Research and Reviews, 2025, 25(01): 827-849. https://doi.org/10.30574/wjarr.2023.20.3.2515

[5] Lin Y, Shen Z, Teng X. Personal Information Protection and Interest Balance Based on Rational Expectation in the Era of Big Data A Case on the Sharing of Mobile Phone Signaling Big Data in Smart City Planning[J]. International Review for Spatial Planning and Sustainable Development, 2022, 10(1): 1-23. https://doi.org/10.14246/irspsd.10.1_1

[6] Utami T K, Putri K A, Suryanto S O, et al. Personal Data Breach Cases In Indonesia: Perspective Of Personal Data Protection Law. Journal Customary Law, 2025, 2(2): 21-21. https://doi.org/10.47134/jcl.v2i2.3742

[7] Ram Mohan Rao P, Murali Krishna S, Siva Kumar A P. Privacy preservation technique in big data analytics: a survey. Journal of Big Data, 2018, 5(1): 33. https://doi.org/10.1186/s40537-018-0141-8

[8] Sampaio S, Sousa P R, Martins C, et al. Collecting, processing and secondary using personal and (pseudo) anonymized data in smart cities[J]. Applied Sciences, 2023, 13(6): 3830. https://doi.org/10.3390/app13063830

[9] Xu Y, Wei J, Mi T, et al. Data security in autonomous driving: Multifaceted challenges of technology, law, and social ethics. World Electric Vehicle Journal, 2024, 16(1): 6. https://doi.org/10.3390/wevj16010006

[10] Shamsinejad E, Banirostam T, Pedram M M, et al. A review of anonymization algorithms and methods in big data. Annals of Data Science, 2025, 12(1): 253-279. https://doi.org/10.1007/s40745-024-00557-w

[11] Jouini O, Sethom K, Namoun A, et al. A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions. Technologies, 2024, 12(6): 81. https://doi.org/10.3390/technologies12060081

[12] Chen W, Yang H, Yin L, et al. Large-scale IoT attack detection scheme based on LightGBM and feature selection using an improved salp swarm algorithm. Scientific Reports, 2024, 14(1): 19165. https://doi.org/10.1038/s41598-024-69968-2

[13] Seeman J, Susser D. Between privacy and utility: On differential privacy in theory and practice. ACM Journal on Responsible Computing, 2024, 1(1): 1-18. https://doi.org/10.1145/3626494

[14] Chen Q, Ni Z, Zhu X, et al. Dynamic Edge-Based High-Dimensional Data Aggregation with Differential Privacy. Electronics, 2024, 13(16): 3346. https://doi.org/10.3390/electronics13163346

[15] Gong M, Gao Y, Wu Y, et al. Heterogeneous multi-party learning with data-driven network sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13328-13343. https://doi.org/10.1109/TPAMI.2023.3290213

[16] Jahan N, Rahman R, Wang M. Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence. arXiv preprint arXiv:2504.17703, 2025. https://doi.org/10.48550/arXiv.2504.17703

[17] Ibrahim Khalaf O, Algburi S, S A, et al. Federated learning with hybrid differential privacy for secure and reliable cross-IoT platform knowledge sharing. Security and Privacy, 2024, 7(3): e374. https://doi.org/10.1002/spy2.374

[18] Salako A O, Adesokan-Imran T O, Tiwo O J, et al. Securing confidentiality in distributed ledger systems with secure multi-party computation for financial data protection. Journal of Engineering Research and Reports, 2025, 27(3): 352-373. https://doi.org/10.9734/jerr/2025/v2\i31439

[19] Gadotti A, Rocher L, Houssiau F, et al. Anonymization: The imperfect science of using data while preserving privacy. Science advances, 2024, 10(29): eadn7053. https://doi.org/10.1126/sciadv.adn7053

[20] Singh J P, Aqsa A, Ghani I, et al. Privacy-aware hierarchical federated learning in healthcare: integrating differential privacy and secure multi-party computation. Future Internet, 2025, 17(8): 345. https://doi.org/10.3390/fi17080345

[21] Putra M A P, Alief R N, Rachmawati S M, et al. Proof-of-authority-based secure and efficient aggregation with differential privacy for federated learning in industrial IoT. Internet of Things, 2024, 25: 101107. https://doi.org/10.1016/j.iot.2024.101107

[22] Wang X, Fan W, Hu X, et al. Differential privacy-preserving of multi-party collaboration under federated learning in data center networks. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 8(2): 1223-1237. https://doi.org/10.1109/TETCI.2023.3341299

[23] Thedchanamoorthy G, Bewong M, Mohammady M, et al. UD-LDP: A Technique for optimally catalyzing user driven Local Differential Privacy. Future Generation Computer Systems, 2025, 166: 107712. https://doi.org/10.1016/j.future.2025.107712

[24] Alebouyeh Z, Bidgoly A J. Privacy-preserving federated learning compatible with robust aggregators. Engineering Applications of Artificial Intelligence, 2025, 143: 110078. https://doi.org/10.1016/j.engappai.2025.110078

[25] Arbaoui M, Brahmia M A, Rahmoun A, et al. Federated learning survey: A multi-level taxonomy of aggregation techniques, experimental insights, and future frontiers. ACM Transactions on Intelligent Systems and Technology, 2024, 15(6): 1-69. https://doi.org/10.1145/3678182

[26] Rigatos G, Abbaszadeh M, Busawon K, et al. Flatness-based control in successive loops for autonomous quadrotors. Journal of Dynamic Systems, Measurement, and Control, 2024, 146(2): 024501. https://doi.org/10.1115/1.4063907

[27] Boulkroune A, Boubellouta A, Bouzeriba A, et al.

Practical finite-time fuzzy synchronization of chaotic systems with non-integer orders: Two chattering-free approaches. Journal of Systems Science and Systems Engineering, 2025, 34(3): 334-359. https://doi.org/10.1007/s11518-024-5635-7

[28] Boulkroune A, Hamel S, Zouari F, et al. Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, 2017, 2017(1): 8045803. https://doi.org/10.1155/2017/8045803

[29] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025: 10775463251320258. https://doi.org/10.1177/10775463251320258

[30] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.

[31] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. https://doi.org/ 10.1109/SSD.2013.6564134

Appendix

To clarify the reproducibility of this study, the implementation methods of each anonymization technology are specified as follows: The k-anonymity, l-diversity, and t-closeness models in this study, as well as the Laplace mechanism, exponential mechanism in differential privacy, and the basic random response mechanism in local differential privacy, are all self-implemented by the author team using Python in accordance with their classic algorithm descriptions. This is done to ensure the fairness and consistency of the comparison. The federated learning framework and the CDP aggregation algorithm within it are built based on the TensorFlow Federated framework. Best practices from the TensorFlow Privacy library are referenced to implement gradient clipping and Gaussian noise addition.

```
Algorithm 1: k- anonymization algorithm based on generalization
Input: original dataset $D$, quasi-identifier set $QIs$, anonymization parameter $k$
Output: k-anonymized dataset $D'$
Build generalized hierarchy $GHTs$ of all quasi-identifiers
There is equivalence class size < $k$ do in while $D$
    Select the attributes that have the lowest generalization degree and can be merged to form an equivalent class with size $\geq k$ for generalization
    Update the dataset $D$ according to the generalization operation
end while
return $D'$
```

```
Algorithm 2: Laplacian mechanism satisfying ε-differential privacy
Input: query function $f$, dataset $D$, privacy budget $\epsilon$
Output: query result $\tilde{f}(D)$ satisfying differential privacy
Calculate the global sensitivity $\Delta f$ of the query function ff
Calculate the scale parameter of Laplacian distribution $b = \Delta f/\epsilon$
Sample noise from Laplace distribution Lap(0,b)
$\tilde{f}(D) = f(D) + noise$
return $\tilde{f}(D)$
```

```
Algorithm 3: DP aggregation algorithm in federated learning (based on Gaussian mechanism)
Input: Each client model updates $\{\Delta w_i\}_{i=1}^{N}$, clipping norm $S$, noise multiplier $\sigma$
Output: Aggregation update satisfying differential privacy $\overline{\Delta w_{DP}}$
for each client update $\Delta w_i$ do
    $\Delta w_i \leftarrow \Delta w_i/\max(1, \frac{\|\Delta w_i\|_2}{S})$ // gradient clipping
end for
Calculate average update $\overline{\Delta w} = \frac{1}{N}\sum_{i=1}^{N} \Delta w_i$
Sampling noise~N(0,$S^2\sigma^2 I$) //I is identity matrix
$\overline{\Delta w_{DP}} = \overline{\Delta w} + noise$
return $\overline{\Delta w_{DP}}$
```