**New Media User Privacy Protection Mechanism Based on Differential Privacy and Data Anonymization**

Yuting Xu 1, a, Yanghe Liu 1, b, and Mingshan Hou 1, c

1School of Digital Art and Design, Chengdu Neusoft University, Chengdu, China;

a xuyuting@nsu.edu.cn, b liuyanghe@nsu.edu.cn, c houmingshan@nsu.edu.cn

**Abstract.** With the rapid development of new media, user privacy issues have become increasingly important. New media platforms, such as TikTok, Twitter, and WhatsApp, process vast amounts of user data daily, including personal information, behavioral data, and social relationships. The extensive collection and use of this data present numerous privacy protection challenges. Many users are not fully aware of how their data is collected and used when using new media platforms, lacking informed consent regarding data collection. Consequently, relying on users to take proactive privacy measures to prevent the disclosure of critical information is difficult to achieve. To address this issue, this paper proposes a privacy protection mechanism that combines differential privacy and data anonymization. The mechanism protects query results through differential privacy techniques and hides user identity information using data anonymization techniques, thereby ensuring user privacy during data analysis and processing. This paper conducts necessary experimental analysis on the proposed method, and the results demonstrate its usability and effectiveness.

**Keywords:** New Media, User Privacy, Data Anonymization, Differential Privacy.

1. **Introduction**

With the rapid development of new media technologies, the speed, interactivity, and coverage of information dissemination have significantly improved. This has made the acquisition, dissemination, and utilization of personal information more convenient[1]. However, this convenience has also exacerbated the risk of personal privacy breaches[2]. In recent years, the risk of privacy breaches has continued to rise, driven not only by substantial financial incentives but also by the lack of adequate legal and regulatory frameworks.

In 2018, Facebook was involved in a user data breach incident, where the personal information of 87 million users was misused. Through a data analytics company named Cambridge Analytica, Facebook users' personal data was improperly exploited for political propaganda and election activities. This breach sparked a global outcry and brought privacy protection issues to the forefront, dealing a severe blow to Facebook[3].

This paper explores an innovative privacy protection mechanism that combines differential privacy and data anonymization. Differential privacy, a robust privacy protection tool, protects user privacy by adding random noise to query results, making it impossible for attackers to infer specific individual information from the query results. During data analysis and processing, differential privacy ensures that changes in an individual's data within the dataset do not significantly impact the query results, thereby preventing the disclosure of sensitive information. However, differential privacy primarily focuses on the privacy protection of query results and is relatively weak in protecting user identity information. Therefore, this paper introduces data anonymization techniques to further enhance user privacy protection. Data anonymization removes or replaces direct identifiers in the data, such as names and identification numbers, making it impossible to directly identify users. This technique can be applied at various stages of data collection, storage, and processing, effectively protecting user identity information from being disclosed.

The privacy protection mechanism proposed in this paper combines differential privacy and data anonymization techniques to form a comprehensive privacy protection system. During the data analysis process, data anonymization is first used to preprocess the raw data, hiding user identity information. Then, before returning the query results to the user, differential privacy is applied to...
add random noise to the query results, further protecting user privacy. In this way, even if an attacker obtains the query results, they cannot infer the user’s identity information or sensitive data.

2. Related Work

From the research content, user privacy protection in the new media environment encompasses multiple levels, including legal, technical, and managerial aspects. On the legal level, regulations such as the European Union’s General Data Protection Regulation (GDPR) mandate that any organization or individual collecting, using, processing, or storing personal data of EU residents must comply with strict data protection principles. GDPR imposes fines of up to 4% of global annual turnover or 20 million euros (whichever is higher) for non-compliant companies[4]. These regulations and policies clearly outline requirements for data collection, usage, storage, and transmission, with corresponding enforcement and supervisory bodies. In technical research, the focus is on developing efficient data encryption and anonymization techniques to protect user privacy from being illegally accessed and exploited. These technical measures have improved the security of user data to some extent but also face challenges, such as balancing data utilization and privacy protection[5]. Unlike traditional information dissemination platforms, new media platforms typically have more complex privacy settings, which users often find difficult to understand and configure appropriately. This can lead to unnecessary public or shared exposure of users’ private information[6-7]. For example, social media platforms often require users to provide a substantial amount of personal information, such as name, gender, age, and geographic location. If this information is leaked, it can pose threats to the users’ personal and financial security. Additionally, with the continuous development of big data and artificial intelligence technologies, social media platforms have increasingly powerful capabilities to process and analyze user data, which further increases the risk of privacy breaches. According to research, most users lack sufficient security and privacy awareness when using new media platforms. This can lead to unintentional disclosure of their private information, thereby increasing the risk of privacy breaches[8]. Therefore, enhancing users’ security and privacy awareness is also an important direction for current research.

In summary, the study of user privacy issues on new media platforms shows multiple areas of focus and challenges. Future research needs to continue exploring solutions from legal, technical, and managerial perspectives to better protect user privacy and security.

3. Technical Solution

The privacy protection mechanism proposed in this paper primarily consists of two parts: data anonymization and differential privacy protection.

3.1 L-diversity-based Data Anonymization Technique for New Media User Data

L-diversity is a data anonymization technique designed to enhance the protection of individual privacy within datasets[9-10]. It builds upon k-anonymity by addressing some of its shortcomings, particularly the issues of homogeneity and background knowledge attacks.

New media user data has certain unique characteristics, hence the following main modifications in applying L-diversity in this context:

1. Identification and Classification of Quasi-Identifiers and Sensitive Attributes**: For new media user data, quasi-identifiers may include user ID, age, gender, and location, while sensitive attributes might include browsing history, user preferences, or interaction records.

2. Generalization and Suppression of Quasi-Identifiers**: This involves transforming specific quasi-identifiers to meet k-anonymity requirements. For example, replacing specific ages with age ranges (generalization) and deleting certain records (suppression) to create equivalence classes that satisfy k-anonymity.
(3) Ensuring Diversity of Sensitive Attributes within Each Equivalence Class**: This can be achieved by shuffling sensitive attributes within the same class, introducing synthetic data to increase diversity, or splitting or merging classes to achieve the required level of diversity.

By incorporating these modifications, the L-diversity-based data anonymization technique can effectively protect new media users' privacy by ensuring that sensitive attributes remain diverse within equivalence classes, thereby mitigating risks of privacy breaches due to homogeneity or background knowledge attacks.

The main steps of the algorithm proposed in this paper are as follows:

**Step 1: Data Collection and Preprocessing**

**Data Collection**: Gather user data, including quasi-identifiers (e.g., user ID, age, gender, location) and sensitive attributes (e.g., browsing history).

**Preprocessing**: Clean the data, handle missing values, and deal with outliers.

**Step 2: Identification, Generalization, and Suppression using k-Anonymity**

**Quasi-Identifier Recognition**: During data preprocessing, all identifier columns are removed or desensitized. Despite this, attackers might still identify individuals through combinations of quasi-identifier attribute values (e.g., ZIP code, birthday, gender). Therefore, further processing of quasi-identifiers is necessary. k-Anonymity, a measure of data release safety, requires that at least k records in the published data are indistinguishable by their quasi-identifier values. This step involves recognizing quasi-identifier columns in the dataset, which contain attribute values that could potentially identify individuals.

**Generalization**: After identifying quasi-identifiers, the next phase is generalization. Generalization transforms data from specific and individual forms to general and broader forms. In k-anonymity, generalization typically involves desensitizing quasi-identifier columns so that multiple records share the same quasi-identifier attribute values, forming equivalence classes. The goal of generalization is to reduce the specificity of individual information, thereby increasing the anonymity of the dataset.

**Suppression**: Finally, suppression is performed. In k-anonymity, suppression usually involves deleting or hiding certain sensitive information to further protect individual privacy. Suppression can occur after generalization to ensure that the dataset does not contain any sensitive information that can directly identify individuals. By following these three steps (quasi-identifier recognition, generalization, and suppression), k-anonymity effectively protects individual privacy while maintaining the usability and security of the dataset.

**Step 3: Ensuring L-Diversity**

**Checking Sensitive Attribute Diversity**: Examine the diversity of sensitive attributes within each equivalence class. This involves ensuring that sensitive attributes within each class are sufficiently diverse.

**Enhancing Diversity**: If necessary, further generalize, shuffle sensitive attribute values, or introduce synthetic data to ensure L-diversity. This step ensures that even within each equivalence class formed by k-anonymity, the sensitive attributes remain diverse enough to prevent re-identification of individuals based on these attributes.

By integrating these steps, the proposed algorithm ensures a robust privacy protection mechanism through the combination of k-anonymity and L-diversity techniques.

**3.2 Privacy Protection Technology for New Media Users Based on Differential Privacy**

Differential Privacy[11] is a crucial cryptographic technique aimed at maximizing the accuracy of statistical database queries while minimizing the chances of identifying individual records. It achieves this by introducing random perturbations during data processing, making it difficult to pinpoint specific individual data, thereby effectively mitigating potential data leakage risks. In the context of new media environments, this technology can be applied to various statistical and analytical tasks involving user data, such as user behavior analysis and content recommendation.
This paper proposes an implementation scheme for privacy protection technology for new media users based on differential privacy.

(1) Differential privacy noise addition

Commonly used noise types in differential privacy techniques include Laplace noise and Gaussian noise. The choice of noise type depends on the specific application scenario and characteristics of the data. For Laplace noise, the noise scale (scale parameter \( b \)) is typically calculated based on the privacy budget \( \varepsilon \) and sensitivity \( \Delta f \) of the function. The formula is as follows:

\[
b = \frac{\Delta f}{\varepsilon}
\]

The privacy budget \( \varepsilon \) determines the strength of privacy protection. A smaller \( \varepsilon \) value indicates stronger privacy protection, but it may reduce the availability of data. The sensitivity \( \Delta f \) of data reflects how changes in the data affect query results. To calculate the noise scale using the privacy loss function of differential privacy, given the privacy budget \( \varepsilon \) and sensitivity \( \Delta f \), this paper adopts the Laplace mechanism\(^{[12]}\). This calculation process typically involves mathematical formulas and derivations. For the specific application in this paper, we use a commonly used approximate formula:

\[
\sigma = \frac{\Delta f}{\varepsilon} \sqrt{2 \ln (1.25/\delta)}
\]

To add the computed noise in query results or data processing, this can be achieved by directly adding noise to the original data or inserting noise at a specific step in the query result process. The amount of noise added should correspond to the computed noise scale.

(2) Privacy-preserving query execution

Step 1: Receive Query Request

The system receives query requests from authorized users such as data analysts, researchers, etc. These requests may involve statistical analysis, pattern recognition, and other operations on new media user data.

Step 2: Validate Query Request

Before executing the query, the system needs to validate the query request to ensure it comes from authorized users and complies with privacy policies and legal requirements.

Step 3: Execute Differential Privacy Query

Execute the query request using a differential privacy dataset that has been augmented with noise. This process typically involves data retrieval, computation, aggregation, etc.

First, determine the sensitivity of the query. The sensitivity varies depending on the type of query. For example, for a counting query, the sensitivity is typically 1, as adding or removing one user can affect the query result by at most 1 unit. For a sum query, the sensitivity is the difference between the maximum and minimum values of a single element in the dataset.

\[
S(f) = \max_{D,D'} |f(D) - f(D')|
\]

In this context, \( S(f) \) is the sensitivity of query \( f \), where \( D \) and \( D' \) are adjacent datasets.

Select the privacy budget parameter \( \varepsilon \). A smaller value strengthens privacy protection but also increases noise. Based on the query result and sensitivity, add Laplace noise to ensure differential privacy:

\[
\text{Lap}(x|\lambda) = \frac{1}{2} \exp \left( -\frac{|x|}{\lambda} \right)
\]

Where, the noise scale parameter \( \lambda \) is
The differentially private version of the query result is:

\[ \tilde{f} = f(D) + \text{Lap}\left(\frac{S(f)}{\varepsilon}\right) \]

Where \( \text{Lap} \) denotes noise drawn from a Laplace distribution.

**Step 4: Return Privacy-Preserving Query Results**

Return the privacy-preserving query results obtained from executing the differentially private query to the requester. These results have been processed with noise, making it impossible to directly identify or track any individual user information.

**Step 5: Monitoring and Auditing**

Monitor and audit the entire privacy-preserving query process to ensure compliance with differential privacy requirements and prevent any privacy breaches. This can be achieved through logging, security audits, and other methods.

## 4. Evaluation and Analysis

The dataset used in this paper's experiments is based on the Adult Data Set[13], aiming to validate the performance of the proposed algorithm. Due to the potential randomness in experimental results, each experiment in this paper was conducted 20 times. Boxplot graphs were used to represent the results, providing a visual depiction of the data distribution, including measures of central tendency and variability across multiple runs.

### 4.1 Analysis of Information Loss

This experiment involves testing datasets of various sizes to analyze the impact of dataset size on the ratio of information loss incurred by the algorithm. The resulting information loss reflects the effect of anonymization on data availability. In the experiments, datasets of sizes 2K, 4K, 8K, 16K, and 32K were randomly selected from the overall dataset for analysis. The study aimed to compare and analyze how the size of the processed dataset affects the ratio of information loss in the algorithm.

As shown in Figure 1 of the experimental results, it is observed that as the volume of data in the anonymized datasets increases, the proportion of information loss generated by the algorithm decreases consistently. This trend can be attributed to the increased stability in the clustering phase.
as the dataset size grows. When the clustering results become more stable, the generalization phase incurs lower information loss during data processing. Therefore, the stability of clustering results directly influences the ratio of information loss incurred during anonymization processing.

Overall, the algorithm proposed in this paper achieves a maximum information loss of no more than 10%. This indicates that it meets the basic requirements of reducing information quality for privacy protection in new media users.

4.2 Analysis of Delay

In this experiment, anonymization processing was conducted on datasets of sizes 2K, 4K, 8K, 16K, and 32K. Five datasets of different sizes were randomly selected from the total experimental dataset to analyze the time required for the algorithm to anonymize different dataset sizes.

As shown in Figure 2 of the experimental results, the execution time of the algorithm increases as the size of the dataset being processed increases. This is because larger datasets require processing more data records during execution, leading to longer anonymization processing times. According to the experimental data, the execution time does not exceed 100ms for the 32K dataset size, which is generally sufficient for most application scenarios. Furthermore, the overall time consumption exhibits a linear growth trend with dataset size, indicating good performance scalability of the algorithm with respect to dataset size.

5. Summary

This paper proposes a novel privacy protection mechanism for new media users by integrating differential privacy and data anonymization. By anonymizing user data and applying differential privacy protections, the mechanism ensures the privacy security of users during data analysis and processing. Experimental results indicate that the proposed method performs well in terms of information loss and algorithm latency.

Future research can focus on further optimizing algorithm performance to enhance privacy protection effectiveness. Additionally, practical implementations will be necessary to validate the feasibility and effectiveness of this mechanism in real-world applications.

References


