DOI: https://doi.org/10.30837/IST-2024.P1.055

Ensuring Data Confidentiality During Data Analysis in Information Systems

Mirsakhib Miriiev[1†] та Iryna Buchynska[1†]

[1] *Odessa I.I. Mechnikov National University, 2 Vsevolod Zmiienko Street, Odesa, Ukraine*

Abstract

This paper examines scientific methods for guaranteeing data privacy during analysis in information systems. It explores data protection challenges in the digital age, where information is crucial for all sectors. While data growth and analysis advancements allow studying society and business, storage, processing, and analysis pose privacy threats like unauthorized access and leaks. The paper highlights various scientific methods for data privacy in information systems, focusing on combining data anonymization (removing personal details) with l-differential privacy (adding noise) to ensure high privacy while enabling valuable data analysis.

Keywords

data privacy, information systems, data analysis, data anonymization, l-differential privacy, data protection, information security.

1. Introduction

In the era of the digital age, where data becomes a key resource for all spheres of life, privacy protection becomes a primary task for ensuring the privacy and security of users. The increasing volume of data and the development of analysis technologies allow information to be used to study various aspects of society and business. However, the storage, processing, and analysis of data are associated with serious privacy threats, such as unauthorized access, data leaks, and privacy breaches.

2. Scientific Methods for Ensuring Privacy

To ensure data privacy in modern information systems, it is necessary to apply scientific methods that effectively protect information from potential threats. In this context, scientific methods play an important role in the creation and implementation of data protection strategies that provide not only a high level of privacy but also preserve the ability to usefully process and analyze information for informed decision-making.

The paper will consider various scientific methods for ensuring privacy when analyzing data in information systems. Each of them plays an important role in protecting information, ensuring the preservation of privacy and compliance with security requirements in the modern digital environment. When analyzing data in information systems to ensure privacy, there are various scientific methods that can be used (Figure 1).

Information Systems and Technologies (IST-2024), November 26-28, 2024, Kharkiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

Sahibmiriev4@gmail.com (M. Miriiev); buchinskayaira@gmail.com (I. Buchynska)

0009-0004-7223-1796 (M. Miriiev); 0000-0002-7842-3543 (I. Buchynska)

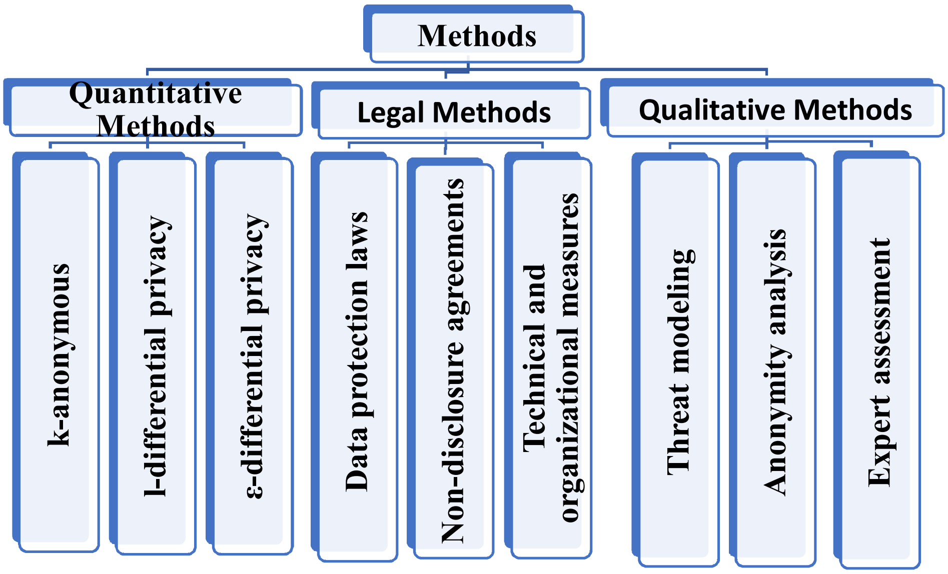

Figure 1: Methods for ensuring privacy when analyzing data in information systems

2.1. Quantitative Methods:

1. Data is considered k-anonymous if each record in the dataset cannot be distinguished from at least k-1 other records based on the values of the attributes used for anonymization.
2. l-differential privacy guarantees that the probability of obtaining any data analysis result for any dataset will be almost the same, regardless of whether a specific record is included in the dataset.
3. ε-differential privacy is similar to l-differential privacy, but uses ε-differential privacy to measure privacy[1].

2.2. Qualitative Methods:

1. Threat modeling involves identifying potential privacy threats and assessing the likelihood and impact of each threat.
2. Anonymity analysis involves analyzing an anonymized dataset to determine whether it is possible to identify records from that dataset.
3. Expert assessment involves involving cybersecurity and privacy experts to assess the risks associated with data analysis.

2.3. Legal Methods:

1. Data protection laws define the rules and regulations that govern the collection, use, and storage of personal data.
2. Non-disclosure agreements (NDAs) oblige parties not to disclose confidential information.
3. Technical and organizational measures include data encryption, access control, and security policies.

3. Combination of Methods

These methods can be used individually or in combination to maximize privacy when analyzing data in information systems. Combining different data protection methods can provide a more effective level of privacy.

Here are some possible combinations:

3.1. Data anonymization with l-differential privacy.

This combination of methods can be used to ensure data privacy when analyzing data that is first anonymized and then l-differential privacy is used to quantitatively assess privacy.

3.2. Threat modeling with k-anonymity.

This combination can be used to identify potential privacy threats associated with data analysis, where k-anonymity is then used to anonymize data to reduce the risk of these threats.

3.3. Data protection laws with data encryption.

The combination of methods can be used to ensure compliance with data protection laws. Data is encrypted to protect it from unauthorized access.

4. Combination of Data Anonymization and l-Differential Privacy

Among the combinations described above, the combination of data anonymization and l-differential privacy is chosen for further examination. Initially, data anonymization is applied to the dataset, which involves removing or replacing personally identifiable information (PII).[2]

Next, considering the anonymized data, l-differential privacy is employed for data analysis. It utilizes algorithms that guarantee that even with minor changes in the input data, the analysis results remain unaltered.

The combination of data anonymization and l-differential privacy can be implemented as follows:

D - represents the dataset.

A - represents the anonymization function that transforms D into an anonymized dataset D'.

Data anonymization can be expressed as:

$$D' = A(D) \tag{1}$$

Let Q represent a database query.

D1 and D2 represent two datasets that differ by only one record.

A(Q, D) represents the response to query Q on database D.

l-differential privacy mandates: for all Q and D1, D2 differing by only one record:

$$\mathrm{Pr}[\mathrm{A}(Q, D1) \in S] \leq e^{l} \times Pr[Q\mathrm{A}(Q, D2) \in S] \tag{2}$$

where Pr[] is the probability that the query Q's result on database D will be within a specific set S.

Formula [2] implies that the probability of obtaining a data analysis result for dataset D1 is almost identical to that for dataset D2, even if they differ by only one record.

Consequently, the combination of data anonymization and l-differential privacy safeguards data privacy by ensuring anonymity and maintaining nearly equal probabilities for data analysis outcomes even with minor dataset changes.

Thus, the combination of data anonymization and l-differential privacy allows protecting the confidentiality of data by ensuring anonymity and nearly equal probability of data analysis results even with small changes in the dataset.

The diagram in Figure 2 illustrates various data protection methods and their interrelationships.

The combination of data anonymization and l-differential privacy simplifies the identification of individuals in the dataset and ensures guaranteed confidentiality during data analysis.

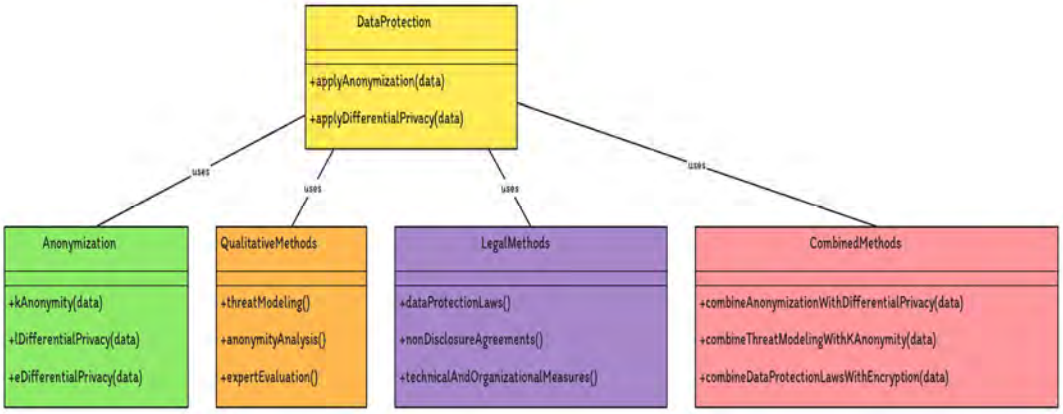
DataProtection
+applyAnonymization(data)
+applyDifferentialPrivacy(data)
uses
uses
uses
uses
Anonymization
+kAnonymity(data)
+lDifferentialPrivacy(data)
+eDifferentialPrivacy(data)
QualitativeMethods
+threatModeling()
+anonymityAnalysis()
+expertEvaluation()
LegalMethods
+dataProtectionLaws()
+nonDisclosureAgreements()
+technicalAndOrganizationalMeasures()
CombinedMethods
+combineAnonymizationWithDifferentialPrivacy(data)
+combineThreatModelingWithKAnonymity(data)
+combineDataProtectionLawsWithEncryption(data)

Figure 2: Data Protection Methods

Anonymization transforms data to complicate or prevent the identification of individuals, while l-differential privacy ensures that even with some information about other records, the ability to exclude the presence of a specific record from the response to a query is maintained at a high level. For example, suppose it's necessary to analyze a dataset of patients to determine if there is a correlation between age and heart rate frequency. Data anonymization can be used to hide the names and addresses of patients. Then, l-differential privacy can be applied to add noise to the age data, making it more difficult to identify specific patients. This will allow us to conduct the analysis without compromising the confidentiality of patients.

The advantages of combining data anonymization and l-differential privacy include: increased level of confidentiality, which can ensure a higher level of data confidentiality than either of them separately; quantitative assessment of confidentiality providing the ability to quantitatively evaluate the level of data confidentiality; flexibility in using them together with various types of data and analytical methods. As for the disadvantages of the combination: the combination of these two methods can be a challenging task; information loss may lead to loss of information, which can affect the results of data analysis; decreased data utility may result in decreased usefulness of data for analysis.

5. Conclusion

Scientific methods for ensuring privacy when analyzing data in information systems are crucial tools for guaranteeing information privacy and security in the digital world. These methods play a vital role in creating and implementing data protection strategies that ensure not only a high level of privacy but also the ability to process and analyze information effectively for informed decision-making. Understanding and effectively implementing these methods helps safeguard information and maintain user trust in digital technologies. Combining methods can provide a more robust level of data protection. The combination of data anonymization and l-differential privacy offers an effective and flexible approach to ensuring data privacy during analysis. It allows you to strike a balance between protecting personal information and obtaining valuable analytical results.

References

[1] M Ros Martín, Impact evaluation clustering-based k-anonymity for recommendations, 2017.

[2] Kazukuni Kobara, Cyber Physical Security for Industrial Control Systems and IoT, IEICE Transactions on Information and Systems, 2016, Volume E99.D, Issue 4, Pages 787-795, DOI: https://doi.org/10.1587/transinf.2015ICI0001