A Thesis for the Degree of Ph.D. in Engineering

Advances in Formal Anonymization for Practical Applications

— Efficient and Secure Techniques for Structured and Unstructured Data —

September 2025

Graduate School of Science and Technology

Keio University

JING JIA

別表６
（３）

Keio University

Thesis Abstract

No.

Registration Number	■ “KOU” □ “OTSU” No. *Office use only	Name	JIA, JING

Thesis Title

Advances in Formal Anonymization for Practical Applications
— Efficient and Secure Techniques for Structured and Unstructured Data —

Thesis Summary

Modern information systems and applications increasingly rely on large-scale data collection and sharing to provide intelligent services. However, data sharing carries significant privacy risks that can seriously compromise individual privacy. When sensitive information is not properly managed or used for unauthorized purposes, privacy issues such as leakage of users' personal data will arise, potentially discouraging users from sharing their data and thereby hindering the advancement of data-driven research.

This thesis presents three novel anonymization approaches that address the fundamental challenge of enabling secure data analysis while protecting sensitive information from untrusted entities. The proposed approaches focus on practical anonymization solutions that satisfy data owners' privacy requirements while preserving data utility for research and analytical purposes.

The first approach develops a central differential privacy scheme that allows for data analysis without revealing the private information of a user. Particularly, when computing the statistics of the original dataset, it cannot be determined if the original dataset contains a specific individual's data. It generates the perturbed dataset in the encrypted domain using homomorphic encryption. The unmodified data of each user is not exposed to the data curator; therefore, data users do not need a trusted data curator. Compared with existing models in which significant amounts of noise are locally added before sending data to the data curator, the proposed scheme adds considerably less noise while maintaining the same privacy requirements, resulting in enhanced data utility.

The second approach presents a privacy-preserving hierarchical k-anonymization framework over encrypted data. The framework divides the computational domain into two different domains: local and global, where different secure computational methods are conducted flexibly. Homomorphic encryption is conducted in the local domain with limited computational power, whereas secret sharing is conducted in the global domain with sufficient computational subjects. The hierarchical structure is achieved by sending the anonymized results of local domains to global domains, which can achieve higher-level anonymization. The experimental results show that connecting two domains can accelerate the anonymization process by more than two times, with the total execution time reduced by 61.7% while information loss increased by only 5.6%.

The third approach introduces a flexible two-stage k-anonymization framework for unstructured clinical records that adapts to various language models. The framework combines natural language processing

methods and privacy-preserving techniques, where the first stage extracts sensitive entities from narrative clinical records according to identifiers predefined by existing privacy rules, and the second stage generates anonymized data that satisfies k-anonymity. Fine-tuned Bidirectional Encoder Representations from Transformer (BERT) models and prompt-driven Large Language Models (LLMs) were developed and customized in this framework. Experimental results demonstrate that the framework achieves high F1-scores of over 90% across multiple entity types, and the two-stage structure allows for dynamic adjustment of entity categories and anonymization strategies to comply with various privacy regulations.

In summary, these contributions advance the field of privacy-preserving data anonymization by providing practical, deployable solutions that bridge the gap between theoretical privacy guarantees and real-world implementation requirements. The proposed approaches demonstrate that effective anonymization can be achieved without sacrificing data utility, enabling secure data sharing across diverse application domains including smart cities and healthcare systems.