Privacy Preserving Data Mining

Hiromi Arai (The University of Tokyo)

Data privacy is a fundamental problem in today's big data era. There are huge social benefits in collection and manipulation of data from people: access to genomic data enables personalized medicine and helps research progress; aggregated mobility data can help to build better transportation systems. However, the wide availability of personal data raises privacy concerns.

It is important to protect the privacy of individuals or organizations that contributed their data. Data holders need to make sure that such sensitive information kept private. Then, they need to know privacy risks exist and reduce them. To achieve these requirements, we focus on privacy preserving data mining technologies.

We develop a method for quantifying genomic privacy. The aggregates such as statistics do not contain private data explicitly. However, recent research has shown that sharing the aggregates of private data, such as personal genomic data, may compromise participant privacy. To quantify privacy risk in the aggregates, we introduce a reverse-engineering method to measure privacy. Unlike previous methods that have been designed to measure the total privacy loss, our method gives an individual privacy measure for each private attribute. We investigate genetic privacy in a scenario where an adversary aims to infer individual's genetic information from his/her genetic testing results. We measure the privacy loss and show the trade-off between privacy and utility in the release of genetic testing results. We also demonstrate that genomic data can reveal sensitive information about individuals even in the aggregates.

We also develop a privacy-preserving search protocol using cryptographic techniques. Information retrieval is essential to the functioning of our communities. We focus on the case of searching for similar compounds in databases for drug development. Since a query compound is an important starting point for the new drug, a query holder usually uses the database in-house. However, when the database holder also wants to output no information except for the search results due to the privacy reasons, we cannot utilize the database. In order to overcome this dilemma, we developed a cryptographic protocol that enables database searching while keeping both the query and database private. Our protocol is successfully built only on an additive homomorphic cryptosystem, which is computationally efficient compared with versatile techniques. The proposed method, easily scales, may help to accurate making full use of distributed private information.
Method and protocol presented here offer the tools to enhance privacy. Though our work focuses on the technical aspects related to privacy preservation, a

similar aspect comes from the policy side. For example, our privacy measure could be used in combination with controlled access. Our cryptographic protocol enables to use data with the less amount of data access. We believe that our works will encourage the use of private data.