Abstract:

Data sharing is a central aspect of modern biomedical research. It is accompanied by significant privacy concerns and often data needs to be protected from re-identification. With methods of de-identification datasets can be transformed in such a way that it becomes extremely difficult to link their records to identified individuals. The most important challenge in this process is to find an adequate balance between an increase in privacy and a decrease in data quality. Accurately measuring the risk of re-identification in a specific data sharing scenario is an important aspect of data de-identification. Overestimation of risks will significantly deteriorate data quality, while underestimation will leave data prone to attacks on privacy. Several models have been proposed for measuring risks, but there is a lack of generic methods for risk-based data de-identification. The aim of the work described in this article was to bridge this gap and to show how the quality of de-identified datasets can be improved by using risk models to tailor the process of de-identification to a concrete context. We implemented a generic de-identification process and several models for measuring re-identification risks into the ARX de-identification tool for biomedical data. By integrating the methods into an existing framework, we were able to automatically transform datasets in such a way that information loss is minimized while it is ensured that re-identification risks meet a user-defined threshold. We performed
an extensive experimental evaluation to analyze the impact of using different risk models and assumptions about the goals and the background knowledge of an attacker on the quality of de-identified data. The results of our experiments show that data quality can be improved significantly by using risk models for data de-identification. On a scale where 100 % represents the original input dataset and 0 % represents a dataset from which all information has been removed, the loss of information content could be reduced by up to 10% when protecting datasets against strong adversaries and by up to 24 % when protecting datasets against weaker adversaries. The methods studied in this article are well suited for protecting sensitive biomedical data and our implementation is available as open-source software. Our results can be used by data custodians to increase the information content of de-identified data by tailoring the process to a specific data sharing scenario. Improving data quality is important for fostering the adoption of de-identification methods in biomedical research.