Privacy-preserving personal data analysis

Summary

An ever-increasing amount of data is generated by our citizens and used in our daily life every single day. These massive amounts of data can be used to improve digital technologies and develop data-driven innovations that can impact every aspect of peoples lives. However, lack of sharing, accessing to and reusing from multiple organizations hinders the analysis possibilities and hence potential insights from the data. A number of challenges have been recognized such as technical barriers, security, data protection compliance to one or more legal jurisdictions, privacy concerns, and trust issues. The overall aim of this thesis is to develop new privacy-preserving data sharing and analysis techniques that strengthen and extend the (re-)use of personal data while maximally protecting individuals privacy. To achieve this aim, this thesis addressed the research challenges on personal data sharing and use from the perspectives of data organizations, the research community, and individuals (data providers).

This thesis first presents a systematic literature review (**Chapter 2**) on privacy-preserving distributed data mining (PPDDM) techniques which considers the issue of executing data mining algorithms on private, sensitive, and/or confidential data from multiple data organizations while maintaining privacy. This chapter draws an overview of existing PPDDM methods to help researchers better understand the development of this domain and assist practitioners to select suitable solutions for their practical cases. We discussed the highlights and remaining challenges in the field including a lack of standard evaluation criteria for new PPDDM techniques, the ambiguous definition of privacy, and the gap between theoretical solutions and practical applications. Finally, we provided a list of recommendations for future research in the field.

Chapter 3 presents an innovative infrastructure, which supports secure and privacy-preserving analysis of personal health data from multiple independent organizations with different governance policies. Instead of centralizing the data, the infrastructure enables researchers to send data-processing applications to each involved data organization. This chapter describes an optimal solution accounting for scientific, technical, and ethical/legal challenges in a practical use case. **Chapter 4** proves the feasibility of the proposed privacy-preserving infrastructure using real-life patient data from The Maastricht Study and Statistics Netherlands to study the association between Type 2 Diabetes and annual healthcare expenses. We handled challenges that have

not been adequately studied by previous works such as data linkage in vertically partitioned data, privacy definition and measurement corresponding to technical and legal requirements, and the indispensability of ethical-legal support in the development of new privacy-preserving technology. Based on the work in Chapter 3-4, **Chapter 5** solves the limitations of using a third party and decreases the costs of communication and computation. The proposed privacy-preserving generalized linear model is based on a distributed block coordinate descent algorithm to obtain parameter estimates, and appended an extension to compute accurate standard errors without additional communication cost. We critically evaluate the information transfer of our model and prove the security and privacy against data reconstruction.

The motivation of **Chapter 6** comes from the experience of requesting data and building up a data analysis model without accessing the source data using the privacy-preserving infrastructure. Chapter 6 presents DP-CGANS, a conditional GAN model combining differential privacy to generate realistic and privacy-preserving synthetic tabular data that is structurally and statistically similar to the real data. DP-CGANS tackles two outstanding challenges in generating synthetic (tabular) data - 1) capturing the correlations and dependencies between variables in an imbalanced dataset, 2) addressing privacy concerns when training DP-CGANS on sensitive private data using a differential privacy technique. We extensively evaluate DP-CGANS compared with three other state-of-the-art generative models. We demonstrate that DP-CGANS outperforms other comparable models and shows the tradeoff between data utility and privacy in synthetic data generation.

The focus of **Chapter 7** lies on a citizen-centric data platform (TIDAL) which can give individuals ownership of their own data, and includes mechanisms to provide fine-grained access to external parties. Combined with the previous development, the TIDAL integrates a set of components for requesting subsets of RDF (Resource Description Framework) data stored in personal data vaults based on SOcial LInked Data (SOLID) technology and analyzing them in a privacy-preserving infrastructure. We demonstrate the feasibility and efficiency of the TIDAL platform by querying and analyzing personal health data from an increasing number of data pods and variables. This chapter shows platforms such as TIDAL play an increasingly important role to connect citizens, researchers, and data organizations to increase the trust placed by citizens in the processing of their personal data.

Chapter 8 describes the scientific challenges addressed by this thesis in applying theoretical privacy-preserving distributed data mining methods to practical applications such as the data linkage across sources, trusted party in reality, privacy measurement, optimal choice of privacy-preserving methods, and explainability and transparency of the methods. This chapter highlights the generation and use of synthetic data that needs support from a sound legal framework, followed by a discussion on the importance of interdisciplinary collaborations between technical and ethical-legal experts in developing new privacy-preserving technologies. Last but not least, we envision a new personal data paradigm for citizens to take more control over their data access and how their data is processed. We believe the future personal data use and sharing will be in a fully decentralized network. The changes from now to the future require efforts from all the stakeholders such as individuals (data providers), policymakers, researchers and scientists, data organizations, and our society.