Privacy-preserving personal data analysis

Impact Paragraph

Digital technologies have advanced rapidly and applied broadly in our society and affect everyone's life. By using digital technologies, our citizens generate a massive amount of personal data every single day. These distributed personal data are collected and used to improve digital technologies and enhance data-driven innovations. The potential values and benefits of sharing and (re-)use of distributed personal data in a responsible manner are significant for our society and the scientific community. However, these data are collected and maintained by different independent organizations. Sharing personal data across multiple organizations faces challenges from technical barriers, security and privacy concerns, legal restrictions, and trust issues. Moreover, citizens, whose data have been collected and used, highly value their data rights and privacy. However, our citizens currently have very limited control over their own data. Technical tools and standards are lacking to facilitate citizens to make their own decision for their data and shift data control from the data organizations to individual data providers.

The overall goal of this thesis is to develop new privacy-preserving data sharing and analysis techniques so as to enable new possibilities for (re-)use of personal data while maximally protecting individual privacy. To achieve this, this thesis makes contributions of interest to three key stakeholders:

1. Data organizations: we developed a secure infrastructure that can combine and analyze personal data from multiple sources without revealing sensitive private information.
2. Scientific community: 1) we developed and applied privacy-preserving distributed data mining methods to analyze vertically partitioned data with and without a third party; 2) built a synthetic data generator to simulate the personal data so that researchers can have an insight into data before the lengthy data request process or build-up analysis model without accessing the source data.
3. Individuals: we designed a novel citizen-controlled technology that enables individuals to access and control their personal data and monitor the (re)use of their data.

1 Scientific Impact

The highlighted scientific contribution of this thesis is creating and experimenting new data paradigms for sharing and using personal data with respect to privacy from the organizational to the individual levels. Among data organizations, we proposed a new infrastructure to transfer the analysis models to vertically partitioned data. It is a scalable and secure solution to analyze personal data across multiple sources. Significantly, it unlocks research questions that could not be answered before due to the restrictions on data access and privacy concerns. Unlike other theoretical methods, our infrastructure has been successfully implemented and tested in practice using a large size of real-life data with the support of an ethical-legal framework. We demonstrated the feasibility of our infrastructure by studying the association between diabetes and annual healthcare costs from a Dutch cohort.

The second new data paradigm presented in this thesis is for researchers to use synthetic data to design accurate analysis algorithms without accessing the source data. Our generative model (DP-CGANS) creates realistic and privacy-preserving synthetic tabular data that are structurally and statistically similar to the source data. DP-CGANS tackles two remaining scientific challenges in generating synthetic (tabular) data - 1) capturing the correlations and dependencies between variables in an imbalanced dataset, 2) addressing privacy concerns when training on sensitive private data using a differential privacy technique. We prove DP-CGANS outperforms other state-of-the-art generative models in extensive experiments.

Another innovation lies in the TIDAL citizen-centric data platform, which makes it easier for individuals to store and access their personal data using personal data vault technologies and provide direct consent to health-related research using SOLID (SOcial LInked Data) and Personal Health Train architecture. TIDAL integrates vocabulary services and standards to 1) structure digital consents to meet the requirements of GDPR and 2) address a scientific challenge in improving the interoperability of personal data use. We believe TIDAL is a start to shift the control and use of personal data from a centralized system to a decentralized network.

The datasets, experiments, algorithms, and intermediate and final results in this thesis are all uploaded to public data or code repositories with descriptive documentation following FAIR principles (Findable, Accessible, Interoperable, Reusable). The accessible links to these materials are provided in each chapter. The manuscripts in this thesis are or will be published in open-access scientific journals or conference proceedings. The FAIR data, open-source code, and open-access manuscripts ensure the works in this thesis are reproducible for other researchers.

2 Social Impact

Advancing privacy-preserving data sharing and analysis techniques is a key to achieving responsible use of personal data. The privacy-preserving infrastructure that we developed to securely share data between organizations uncovers more potential use of personal data to improve public and social services, deliver timely healthcare treatments, and other potential benefits to society. This infrastructure protects individual data rights and privacy, which may increase confidence and trust from the data providers (e.g., citizens) in data organizations and how their data is being used by and between organizations.

The generation and use of synthetic data uncover the possibility of mining the value of the data even when the data are inaccessible or unavailable. Like the digital twin can accurately reflect a physical object and simulate its life cycle, our synthetic data generator can generate realistic synthetic personal data that can be used to build and test the analysis models as a replacement for real data. We found that the higher the quality of synthetic data we generate, the more data privacy is sacrificed. This may accelerate research projects which suffer from data access issues. However, it opens new challenges and discussions to the public and our society on the proper generation and responsible use of synthetic personal data.

The citizen-centric data platform (TIDAL) gives individual citizens fine-grained access to their personal data and provides digital consent to use their data for health research. Citizens can monitor and control the whole life cycle of their data including the access, storage, and analysis. TIDAL connects citizens, researchers, and data organizations and facilitates citizens to contribute to health research in a simple way that will improve our society. TIDAL shifts data stewardship and access control from organizations to individuals and encourages citizens to take more responsibility for managing their own data. We believe that TIDAL can start a completely new personal data paradigm that can gain more trust placed by citizens and the transparency of the processing of personal data.