Studies on Privacy-Aware Data Trading*

Shuyuan Zheng

Abstract

A data market is a critical component of a vibrant data ecosystem that incentivizes data sharing. In the market, various data products are demanded and circulated to facilitate diverse applications, especially machine learning (ML)-based applications. Specifically, in addition to trading raw datasets, data owners can also choose to sell data products derived from datasets, including (1) queries over their data, (2) trained ML models, and (3) collaborative training services. However, selling these data products poses privacy risks for data owners, making them hesitant to participate in data trading. Therefore, privacy protection should be a crucial incentive for data owners' participation and an essential function of a robust data market. Motivated by this, we conduct the following three studies on privacy-aware data trading.

First, we study query-based data trading with personalized privacy. A query-based data marketplace allows for the purchase of statistical queries over a database. In this context, preventing buyers' arbitrage behavior is vital, as they can combine multiple cheaper queries into a single, more expensive query to exploit price differences. Additionally, privacy protection is a significant concern since private information can potentially be revealed from queries. Previous work has built arbitrage-free query-based data marketplaces using differential privacy (DP), a widely-used privacy protection technique. However, under DP, privacy protection levels for data owners are uniformly applied, neglecting their diverse privacy preferences. Consequently, this study explores ways to ensure arbitrage-free pricing when allowing data owners to personalize their privacy protection levels.

Second, we study model-based data trading with local privacy. In a model marketplace, customers can purchase trained ML models for their specific data analysis tasks. Since model buyers do not have access to the training data, this category of models can alleviate data owners' concerns about losing control

*Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-18-6886, June 2023.

over their data and thus enhance trustworthiness. However, existing privacy-preserving model marketplaces assume that the broker is trusted and authorized to access and control the raw data, which is unrealistic considering that many large corporations have been involved in user data breaches or privacy scandals. In this study, to protect data owners' privacy against both model buyers and the broker, we propose a locally private model marketplace to promote trustworthy model acquisition.

Finally, we study collaborative ML-based data trading with secure contribution evaluation. A collaborative ML marketplace coordinates collaborations among data owners who aim to collaboratively train an ML model by sharing data. For example, banks may pool user data to create a more accurate fraud detection model. Since data owners' datasets may vary widely in size and quality, existing collaborative marketplaces employ the *Shapley value* (SV), a well-justified contribution metric in cooperative game theory, to fairly evaluate their contributions and determine their corresponding rewards as incentives. However, the SV-based contribution evaluation methods in these marketplaces require access to raw data, compromising privacy. Therefore, we investigate secure SV calculation in this study to facilitate secure contribution evaluation in CML.

Achieving privacy-aware data trading is an essential step in building a trustworthy data market, but more is needed. At the end of this thesis, we also look forward to the future of data trading, discussing how to build trust further and facilitate the arrival of the data economy era.

Keywords: data trading, data privacy, data economy, data pricing, data valuation, market design