（ 続紙 1 ）

京都大学	博士（情報学）	氏名	鄭　舒元（Shuyuan Zheng）
論文題目	Studies on Privacy-Aware Data Trading （プライバシーを考慮したデータ取引に関する研究）		

（論文内容の要旨）

A data market is a key 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 can result in privacy loss 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. For this reason, three studies on privacy-aware data trading are proposed in this dissertation.

Chapter 1 introduces the background of data trading and the motivation for the research on privacy-aware data trading. The chapter also provides an overview of the three studies on privacy-aware data trading and the main contributions.

Chapter 2 provides an overview of the research related to data trading. This chapter categorizes the existing work in the literature into four categories, namely dataset-based, query-based, model-based, and collaborative ML (CML)-based data trading. The chapter introduces them by category and discuss this work's position within the literature.

Chapter 3 proposes the first study of the thesis on 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 behaviors 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 facilitated 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, the author’s study explored ways to ensure arbitrage-free pricing while allowing data owners to personalize their privacy protection levels.

Chapter 4 discusses model-based data trading with local privacy. In a model marketplace, customers can directly 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 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. To protect data owners' privacy not only against model buyers but also against the broker, this study proposed a locally private model marketplace to promote trustworthy model acquisition.

Chapter 5 discusses CML-based data trading with secure contribution evaluation. A CML marketplace coordinates collaborations among data owners who aim to collaboratively train an ML model by sharing data. For example, banks may pool their user data together 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, which compromises privacy. Hence, this study investigates secure SV calculation to facilitate secure contribution evaluation in CML-based data trading.

Chapter 6 summarizes the preceding three studies and provide a vision for the future of data trading. The author discusses several key factors that drive the large-scale application of data trading in real life, and based on this, we envision potential future research directions.

（続紙　２　）

（論文審査の結果の要旨）

個人や組織が持つデータの共有を誘引する手段としてデータの市場化がある．データ市場では，生のデータセットだけではなく，機械学習に基づくアプリケーションなど多様なアプリケーションの下でのデータセットから派生したデータ製品の販売など様々な形態のデータの取引が考えられる．また，対象とするデータがパーソナルデータを含む場合は，取引に際してプライバシー保護がデータ取引への参加誘引のために必要になる．

本論文は，プライバシーを考慮したデータ取引に関する次の三つの課題に取り組んだ研究成果をまとめたものである。(1) データ所有者がデータセットに対する問合せ結果を販売する場合の価格設定．(2) データ所有者が学習済み機械学習モデルを販売する場合のプライバシー保護．(3) 協調的な機械学習サービスにおける参加者の貢献度の安全な評価法．具体的には、これら三つの課題の各課題について以下の成果を上げている。

第一に，差分プライバシーによりデータを保護しながら，問合せに基づくデータ取引を実現する問題に取り組み，従来の研究より現実的な想定としてデータ所有者によってプライバシー保護選好が異なる場合を仮定し，裁定取引のない価格設定を保証する方法を提案した．

第二に，買い手が特定のデータ分析タスクのために訓練された学習済み機械学習モデルを直接購入する機械学習モデル市場において，データ所有者のプライバシー保護を強化する手法を提案した．提案した市場機構では，市場ブローカが連合学習のサーバの役割を果たす場合に，データ所有者が局所差分プライバシーを用いる事により，機械学習モデルの購入者のみならず市場ブローカに対してもプライバシーを保護できる．そのために，各データ所有者が異なるプライバシー保護選好をもつ場合に，連合学習における最適な勾配集約方法を考案した．

第三に，複数の組織による協調的な機械学習であるサイロ横断型機械学習を行う場合に，各所有者の貢献に応じた報酬を与えることによる市場への参加誘引性を確保するために，貢献度をシャープレイ値で安全に評価する手法に取り組んでいる．サーバが所有者の生データをアクセスする事なくシャープレイ値を計算するために準同型暗号を用いるが，暗号化されたデータと機械学習モデルパラメータの行列計算は計算速度の点で実用性に問題がある．そこで，本論文では，高速化するために機械学習モデルを準同型暗号により暗号化し，データは加法的秘密分散により保護するハイブリッド型データ保護スキームを提案し，その有効性を実験により確認した．

以上、本論文は、プライバシーを考慮したデータ取引における三つの重要な課題に対する解決手法をまとめたもので、学術上、および、実際上、寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和５年７月３１日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、本論文は、京都大学学位規程第１４条第２項に該当するものと判断し、公表に際しては、当面の間当該論文の全文に代えてその内容を要約したものとすることを認める。