Privacy and Security in Distributed Data Markets

Daniel Alabi, Sainyam Galhotra, Shagufta Mehnaz, Zeyu Song, Eugene Wu

SIGMOD 2025 Tutorial

Part 5: Open Problems

Privacy Challenges Are Everywhere!

Query specification
Privacy risks

Data Market/Data Discovery

Query interface
Latency, Scalability
Privacy risks

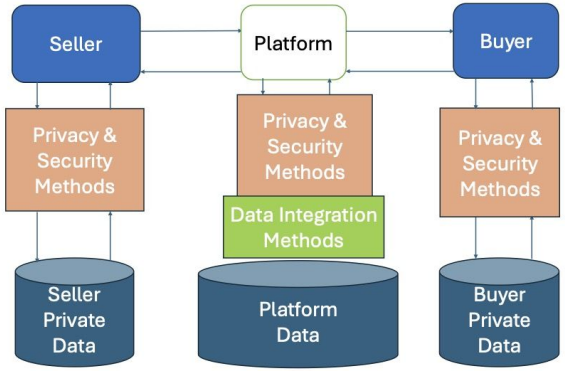
Stats | Stats | Stats

Data | Data | Data

Data acquisition
Data preparation
Privacy risks

Protect Information in Data Markets

1. Protect buyers from *malicious* sellers
2. Protect sellers from *malicious* buyers
3. Prevent *unauthorized* users from accessing:
   a. Seller private data
   b. Buyer private data
   c. Platform private data
4. Prevent manipulation of data acquisition mechanisms:
   a. Data discovery
   b. Data valuation
   c. Data negotiation
   d. Data delivery



Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections.*

We need more attacks for illustrative and motivational purposes.

Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections.*

We need more methods to protect against attacks.

Research Questions for Legal Considerations

- Can we cryptographically enforce legal policies?
- What counts as legally sufficient anonymization?
- Consent revocation in distributed systems?

Data Ownership and Stewardship

- Ambiguity in data and model ownership
- Data Controller vs. Data Processor roles
- Tension between legal rights and cryptographic control

An Agentic Web is a Data Market

Agent-friendly protocols like MCP sidestep web UIs completely

- No GUI, no user, just APIs and automation
- *“The web is a series of databases”* - Sundar Pichai on Decoder Podcast

In an agentic world, every “website” is a database API + business logic…

- Arrow’s paradox? Pricing? Privacy? Security? Discovery? Market structure?

More Future Directions

Our investigation into data marketplaces reveals critical challenges for building secure, decentralized AI systems.

1. The Attack Surface Has Shifted.

The primary vulnerability is not just the model, but the marketplace's economic and selection mechanisms.

2. Standard Metrics are Deceptive.

High model accuracy and low cost can mask catastrophic security failures and unfair economic outcomes.

3. Similarity-Based Defenses are Not a Silver Bullet.

They are fundamentally vulnerable to mimicry attacks and struggle most in the realistic, heterogeneous environments they are designed for.

Path Forward: Building a Robust Data Economy

To build truly secure and equitable marketplaces, future work must move beyond simple similarity checks. We need to focus on:

- **Orthogonal Trust Signals**: Integrating seller reputation, transaction history, and data provenance to make more holistic trust decisions.
- **Multi-Stage Filtering**: Designing a defense-in-depth pipeline that combines anomaly detection, similarity checks, and impact analysis.
- **Incentive-Compatible Mechanisms**: Creating reward and selection systems that are provably resilient to strategic manipulation and fairly compensate true value.

Funding Acknowledgements & Questions