Approaches to distributed privacy protecting data mining

Bartosz Przydatek

CMU

Introduction

- Data Mining and Privacy Protection → conflicting goals

Introduction

- Data Mining and Privacy Protection $\longrightarrow$ conflicting goals
- Conflict Resolution: Inference Control
- Inference Control Techniques
  - Controlled Release
  - Input/Output Perturbation
  - Query Restriction & Auditing
  - ...

Introduction

- Data Mining and Privacy Protection $\longrightarrow$ conflicting goals
- Conflict Resolution: Inference Control
- Inference Control Techniques
  - Controlled Release
  - Input/Output Perturbation
  - Query Restriction & Auditing
  - . . .
- Distributed data mining – old and new challenges

Outline

- What does “privacy” mean?
- Perturbation techniques
- Secure Multi-Party Computation (MPC)
- Privacy by secure MPC
- Conclusions

What Does "Privacy" Mean?

- Intuitively, it seems to be clear . . .
- Exact vs. partial disclosure

What Does “Privacy” Mean?

- Intuitively, it seems to be clear . . .
- Exact vs. partial disclosure
- Quantifying Privacy
  - Interval width for a confidence level [Agrawal, Srikant 2000]
  - Information theoretic approach [Agrawal, Aggarwal 2001]
  - Game theoretic approach [Kleinberg, Papadimitriou, Raghavan 2001]

What Does "Privacy" Mean?

- Intuitively, it seems to be clear . . .
- Exact vs. partial disclosure
- Quantifying Privacy
  - Interval width for a confidence level [Agrawal, Srikant 2000]
  - Information theoretic approach [Agrawal, Aggarwal 2001]
  - Game theoretic approach [Kleinberg, Papadimitriou, Raghavan 2001]
- Privacy in secure multi-party computation

Privacy by Perturbation

- Studied extensively in the context of single databases
- Can be applied in distributed setting

Privacy by Perturbation

- Studied extensively in the context of single databases
- Can be applied in distributed setting
- Various techniques
  - randomized input distortion
  - output perturbation
  - k-anonymity [Sweeney '98]

Problems with Perturbations

- Bias, precision & consistency
- Can be computationally challenging
- Outlier removal & “blurring” the data $\longrightarrow$ detection of anomalies?
- Combining multiple versions of data released for different purposes

Secure Multi-Party Computation

- Introduced by Yao in 1982, inspired by "coin-flipping" (Blum) and "mental poker" (Shamir, Rivest, Adleman)
- m parties $P_1, \ldots, P_m$ want to compute $f(x_1, \ldots, x_m)$, where x_i is a private input of P_i, without revealing more than necessary $\ldots$

Secure Multi-Party Computation

- Introduced by Yao in 1982, inspired by “coin-flipping” (Blum) and “mental poker” (Shamir, Rivest, Adleman)
- m parties $P_1, \ldots, P_m$ want to compute $f(x_1, \ldots, x_m)$, where x_i is a private input of P_i, without revealing more than necessary $\ldots$
- $\ldots$ i.e., simulation of a trusted party!

Secure Multi-Party Computation

- Introduced by Yao in 1982, inspired by “coin-flipping” (Blum) and “mental poker” (Shamir, Rivest, Adleman)
- m parties $P_1, \ldots, P_m$ want to compute $f(x_1, \ldots, x_m)$, where x_i is a private input of P_i, without revealing more than necessary $\ldots$
- $\ldots$ i.e., simulation of a trusted party!
- A very general and powerful tool, various models
- Efficient completeness results: [Yao’86] (2-party), [GMW’87] (crypt.) and [BGW+CCD’88] (uncond.)

Privacy by Multi-Party Computation

- MPC “creates” a trusted party!

Privacy by Multi-Party Computation

- MPC “creates” a trusted party!
- Problems:
  - Efficiency $\rightarrow$ communication complexity

Privacy by Multi-Party Computation

- MPC "creates" a trusted party!
- Problems:
  - Efficiency $\rightarrow$ communication complexity
  - Does it really solve the privacy problem?

Efficient MPC Solutions

- Efficient special purpose protocols
  - Learning decision trees [Lindell, Pinkas 2000]

Efficient MPC Solutions

- Efficient special purpose protocols
  - Learning decision trees [Lindell, Pinkas 2000]
- Private approximations
  - Introduced by [FIMNSW 2000]
  - A tradeoff between privacy and approximability [Halevi, Krauthgamer, Kushilevitz, Nissim, 2001]
  - Some functions cannot be computed with low communication (set equality vs. set disjointness)

Efficient MPC Solutions

- Efficient special purpose protocols
  - Learning decision trees [Lindell, Pinkas 2000]
- Private approximations
  - Introduced by [FIMNSW 2000]
  - A tradeoff between privacy and approximability [Halevi, Krauthgamer, Kushilevitz, Nissim, 2001]
  - Some functions cannot be computed with low communication (set equality vs. set disjointness)
- A different approach to MPC?

Which queries preserve privacy?

Which queries preserve privacy?

- Query restriction
  - query-set-size, query-set-overlap
  - query auditing
  - partitioning

Which queries preserve privacy?

- Query restriction
  - query-set-size, query-set-overlap
  - query auditing
  - partitioning
- Query auditing
  - efficient in simple cases
  - a NP-hard problem in general
    [Kleinberg, Papadimitriou, Raghavan 2001]

Conclusions

- “Privacy” means . . .
- Various approaches, problem dependent
- Probably no “the best” single solution
- Still a lot of work to be done