Modeling and Integrating Background Knowledge in Data Anonymization

Tiancheng Li, Ninghui Li, Jian Zhang
Purdue University

Background & Problem Statement

Data Publishing
- Example of Background Knowledge
  - The prevalence of emphysema is higher for the ≥65 age group
  - Higher in males than females

Challenges & Our Solutions
- How to model background knowledge?
  - What if background knowledge is incorrect?
- How to compute adversarial belief change?
- How to measure privacy?
  - Distance between the prior belief and the posterior belief.

Methodology & Kernel Estimation

Objective & Methodology
- Model consistent background knowledge: P(SA|QI)
  - The original data can be viewed as samples from the distribution.
  - The problem becomes inferring a distribution from samples.

Kernel Regression Estimation
- Each record r is a point (r[QI], r[SA])
  - Find the function P_{pre}: D[QI] → D[SA] best-fits these data points.
  - Intuition: each point spreads its weight over its neighborhood.
  - Kernel function K determines the shape of the bumps.
  - Bandwidth b determines the width of the bumps.

Distance Measure

Desiderata
- Identity of indiscernibles: D[P,P] = 0.
- Non-negativity: D[P,Q] ≥ 0.
- Probability scaling
- Zero-probability definability
- Semantic awareness

Distance measures
- KL-divergence
- Earth Mover’s Distance
- JS-divergence
- Kernel-based JS-divergence

Evaluation

Attacks

Continuity of Disclosure Risk

Modeling Belief Changes

Example
- Case 1: P(HIV|t3) = 0.3, P(HIV|t1) = 0.5, P(HIV|t2) = 0.7
- Case 2: P(HIV|t3) = 0.3, P(HIV|t1) = 0.5, P(HIV|t2) = 0.7
- Case 3: P(HIV|t3) = 0.3, P(HIV|t1) = 0.5, P(HIV|t2) = 0.7

The probability that t3 has HIV is:
- General Formula
- Approximate Inference: Ω-Estimate
  - The random world assumption
  - General computation is a #P-complete problem.

Distance measures
- KL-divergence
- Earth Mover’s Distance
- JS-divergence
- Kernel-based JS-divergence

Semantic awareness