CERIAS

the center for education and research in information assurance and security

Modeling and Integrating Background Knowledge in Data Anonymization

Tiancheng Li, Ninghui Li, Jian Zhang
Purdue University

Background & Problem Statement

► Data Publishing

Age	Sex	Disease
69	M	Emphysema
45	F	Cancer
52	F	Flu
43	F	Gastritis
42	F	Flu
47	F	Cancer
50	M	Flu
56	M	Emphysema
52	M	Gastritis

(a) Original table T

Age	Sex	Disease
[45 – 69]	*	Emphysema
[45 – 69]	*	Cancer
[45 – 69]	*	Flu
[40 – 49]	F	Gastritis
[40 – 49]	F	Flu
[40 – 49]	F	Cancer
[50 – 59]	M	Flu
[50 – 59]	M	Emphysema
[50 – 59]	M	Gastritis

(b) Generalized table T^*

► Example of Background Knowledge

- The prevalence of emphysema is
  - higher for the ≥65 age group
  - higher in males than females

► Inference with Background Knowledge

- Bob is a 69-year-old white male
- Pr(Bob has emphysema) = 1/3.
- With background knowledge, this probability becomes much larger.

► Challenges & Our Solutions

- How to model background knowledge?
  - What if background knowledge is incorrect?
  - Has to be consistent with the original data.
- How to compute adversarial belief change?
  - Bayesian inference
- How to measure privacy?
  - Distance between the prior belief and the posterior belief.

Methodology & Kernel Estimation

► Objective & Methodology

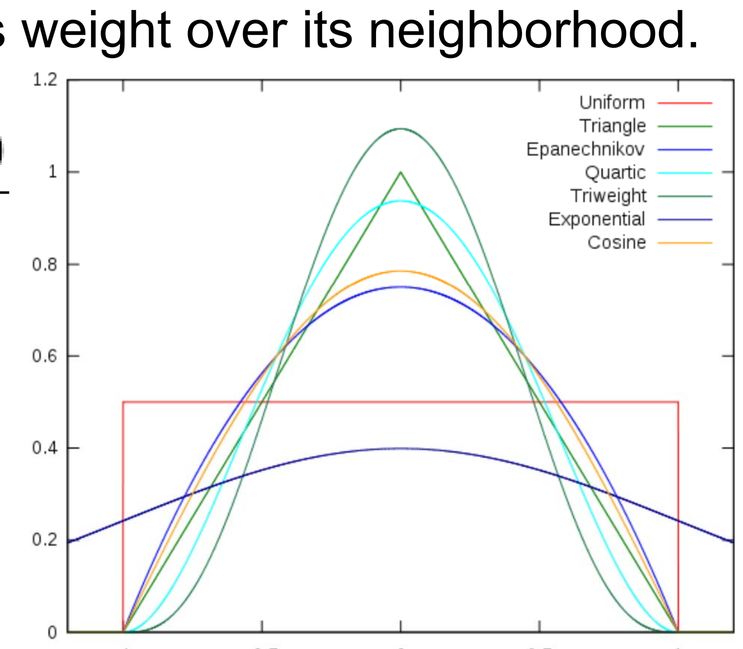
- Model consistent background knowledge: P(SA|QI)
- The original data can be viewed as samples from the distribution.
- The problem becomes inferring a distribution from samples.

► Kernel Regression Estimation

- Each record *r* is a point (r[QI], r[SA])
- Find the function P_{pri}: D[QI]→Dis[SA] best-fits these data points.
- Intuition: each point spreads its weight over its neighborhood.

$$\hat{P}_{pri}(q) = \frac{\sum_{t_j \in T} P(t_j) K(q - t_j[A_1])}{\sum_{t_j \in T} K(q - t_j[A_1])}$$

- Kernel function *K* determines the shape of the bumps.
- Bandwidth b determines the width of the bumps.



Modeling Belief Changes

► Example

tuple
t_1
t_2
t_3

disease
none
none
HIV

(a) A group of three tuples

t_1	t_2	t_3
$P(HIV\|t_1) = .05$	$P(HIV\|t_2) = .05$	$P(HIV\|t_3) = .3$
$P(none\|t_1) = .95$	$P(none\|t_2) = .95$	$P(none\|t_3) = .7$

(b) The adversary's prior belief table

	t_1	t_2	t_3
Case 1	none	none	HIV
Case 2	none	HIV	none
Case 3	HIV	none	none

(c) The three possible cases

$$P(Case\ 1) \propto p_1 = P(none|t_1) \times P(none|t_2) \times P(HIV|t_3) = 0.95 \times 0.95 \times 0.3 = 0.271$$

$$P(Case\ 2) \propto p_2 = P(none|t_1) \times P(HIV|t_2) \times P(none|t_3) = 0.95 \times 0.05 \times 0.7 = 0.033$$

$$P(Case\ 3) \propto p_3 = P(HIV|t_1) \times P(none|t_2) \times P(none|t_3) = 0.95 \times 0.05 \times 0.7 = 0.033$$

$$P(Case\ 1) = \frac{p_1}{p_1 + p_2 + p_3} = 0.8$$

The probability that t_3 has HIV is:

$$P(Case\ 1) \times 1 + P(Case\ 2) \times 0 + P(Case\ 3) \times 0 = P(Case\ 1) = 0.8$$

► General Formula

- General computation is a #P-complete problem.

► Approximate Inference: Ω-Estimate

- The random world assumption

$$\Omega(HIV|t_3) = \frac{1 \times \frac{0.3}{0.4}}{1 \times \frac{0.3}{0.4} + 2 \times \frac{0.7}{2.5}} = 0.57$$

Distance Measure

► Desiderata

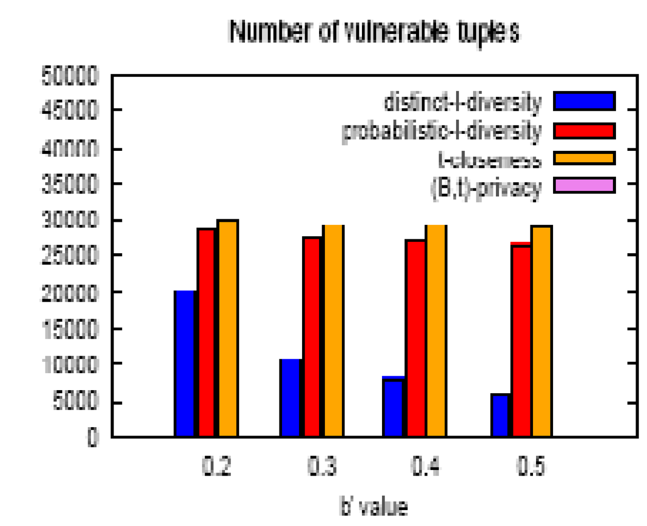
- Identity of indiscernibles: D[P,P] = 0.
- Non-negativity: D[P,Q] ≥ 0.
- Probability scaling
- Zero-probability definability
- Semantic awareness

► Distance measures

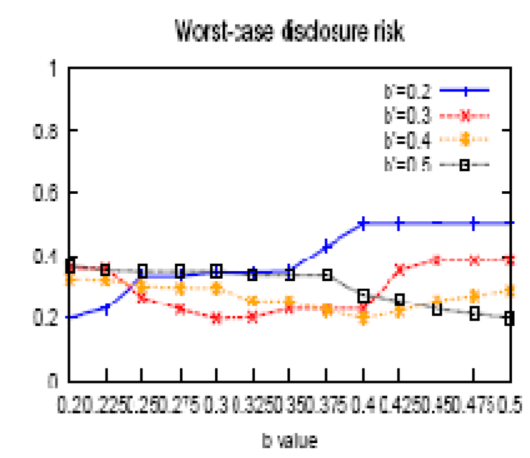
- KL-divergence
- Earth Mover's Distance
- JS-divergence
- Kernel-based JS-divergence

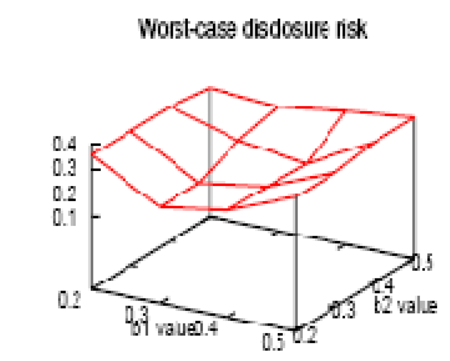
Evaluation

► Attacks



► Continuity of Disclosure Risk





PURDUE UNIVERSITY

CERIAS®

Discovery Park
e-Enterprise Center