PrivBayes: Private Data Release via Bayesian Networks

Jun Zhang[1] Graham Cormode[2] Cecilia Procopiuc[3] Divesh Srivastava[3] Xiaokui Xiao[1]

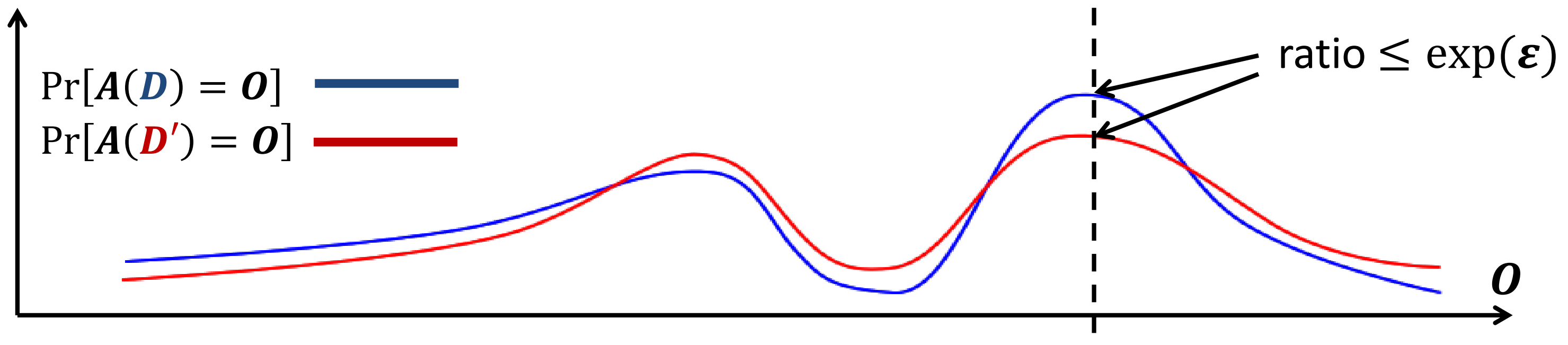
1.Nanyang Technological University
{jzhang027, xkxiao}@ntu.edu.sg

2. University of Warwick
g.cormode@warwick.ac.uk

3. AT&T Labs -- Research
{magda, divesh}@research.att.com

1. Private Data Release

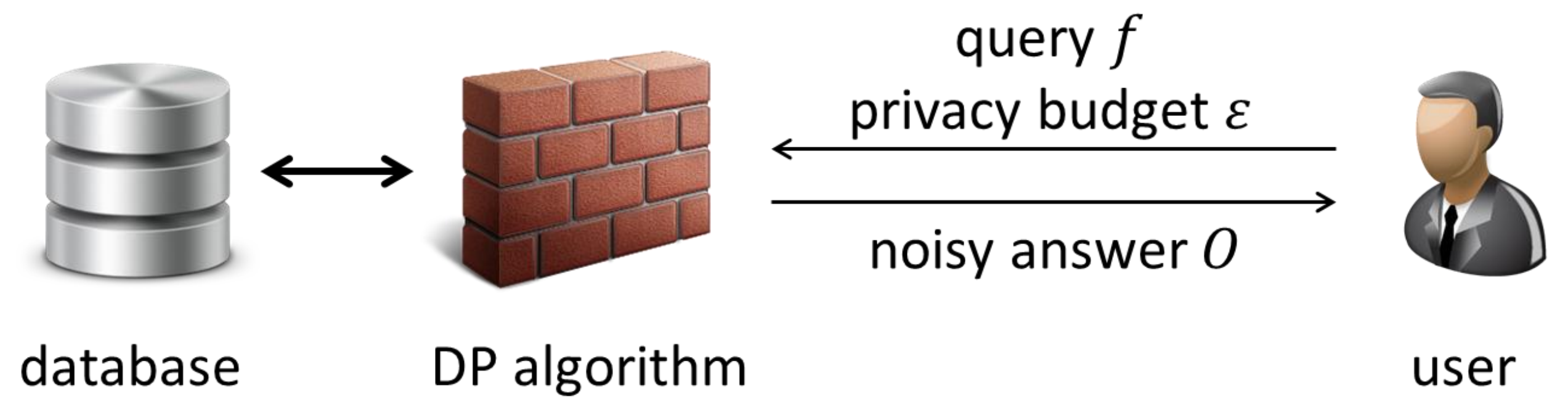
❑ Differential privacy



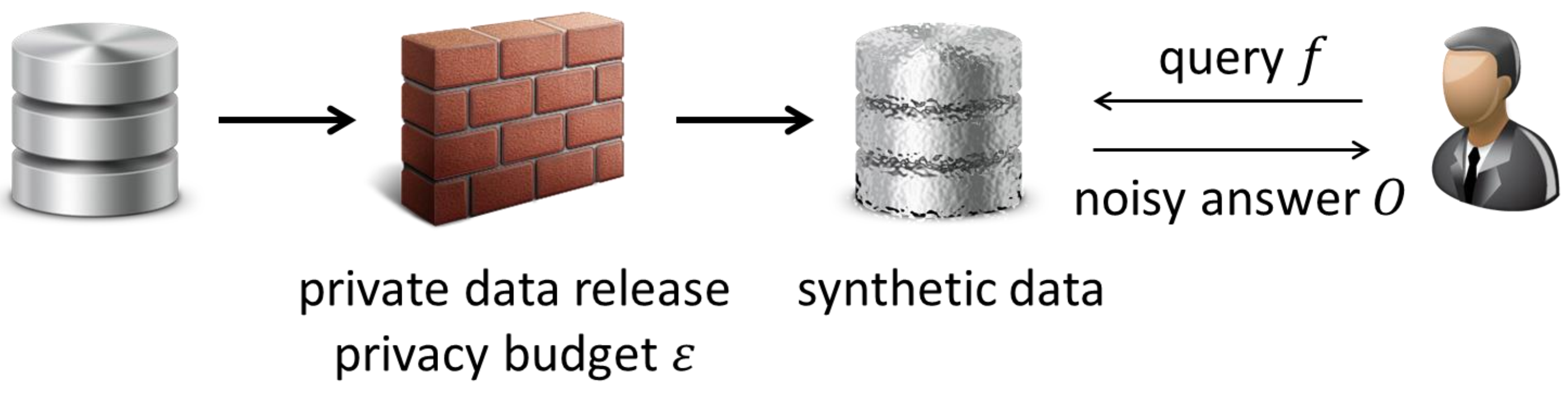
where D and D' are neighboring databases that differ by **at most one** tuple

$$\exp(-\varepsilon) \le \frac{\Pr[A(D)=O]}{\Pr[A(D')=O]} \le \exp(\varepsilon)$$

❑ Interactive mode



❑ Non-interactive mode (synthetic data release)

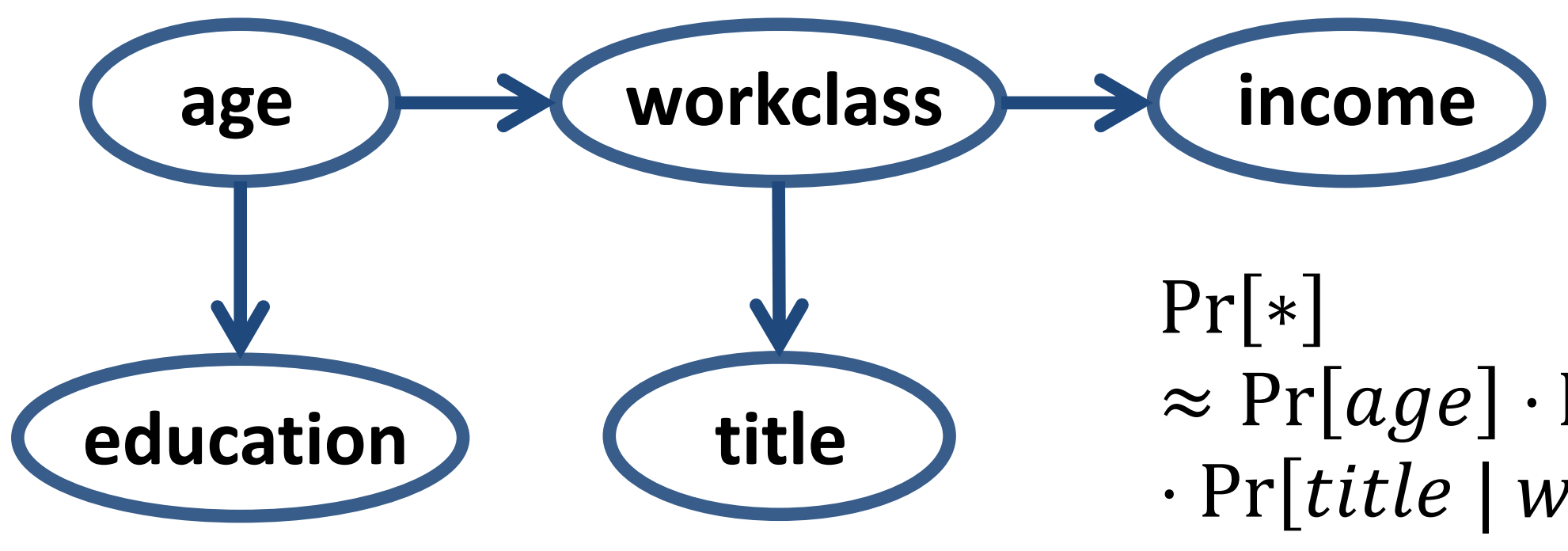


Reusability: only access sensitive data once Generality: support most queries

However, the tuple distribution has a huge domain (exponential to dimension), which leads to high computational cost and low signal-to-noise ratio.

2. Private Bayesian Network

❑ Approximate full distribution by low dimensional ones

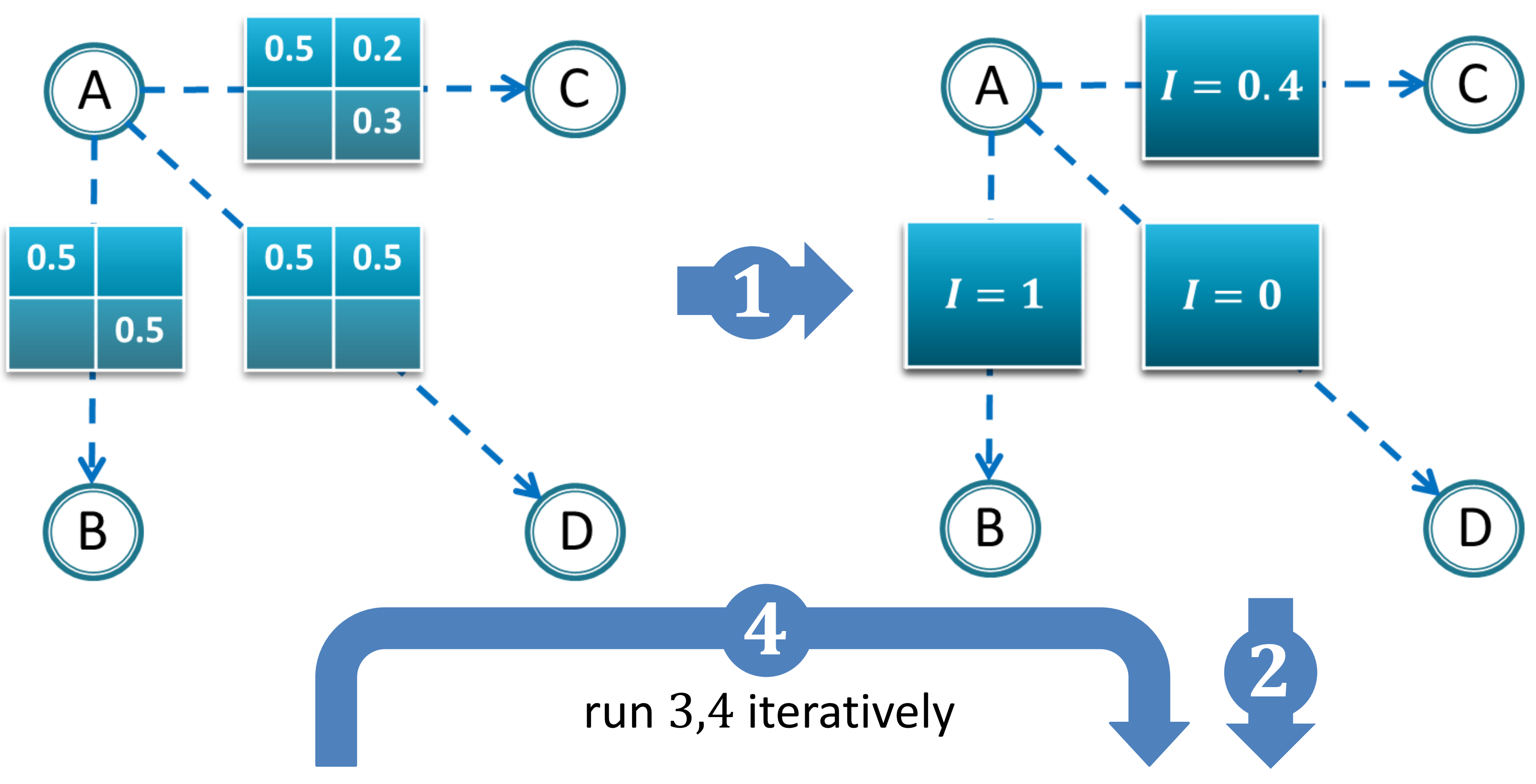


$$\Pr[*] \approx \Pr[age]\cdot\Pr[work \mid age]\cdot\Pr[edu \mid age]\cdot\Pr[title \mid work]\cdot\Pr[income \mid work]$$

❑ Build a Bayesian network

The quality of a Bayesian network is measured by sum of mutual information I of its edges. I is defined as

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} \Pr[x,y] \log\left(\frac{\Pr[x,y]}{\Pr[x]\Pr[y]}\right).$$



EDGE SELECTION

NoPrivacy:

Select the next tree edge with maximum I.

Private selection:

Use exponential mechanism[1] to select: $q(edge) = I(edge)$;

$\Delta(I) = O(\log n / n), n = |D|$.

[1] Frank McSherry and Kunal Talwar. "Mechanism design via differential privacy." FOCS'07

3. Function F: Linear vs. Logarithmic

❑ Drawback of Function I

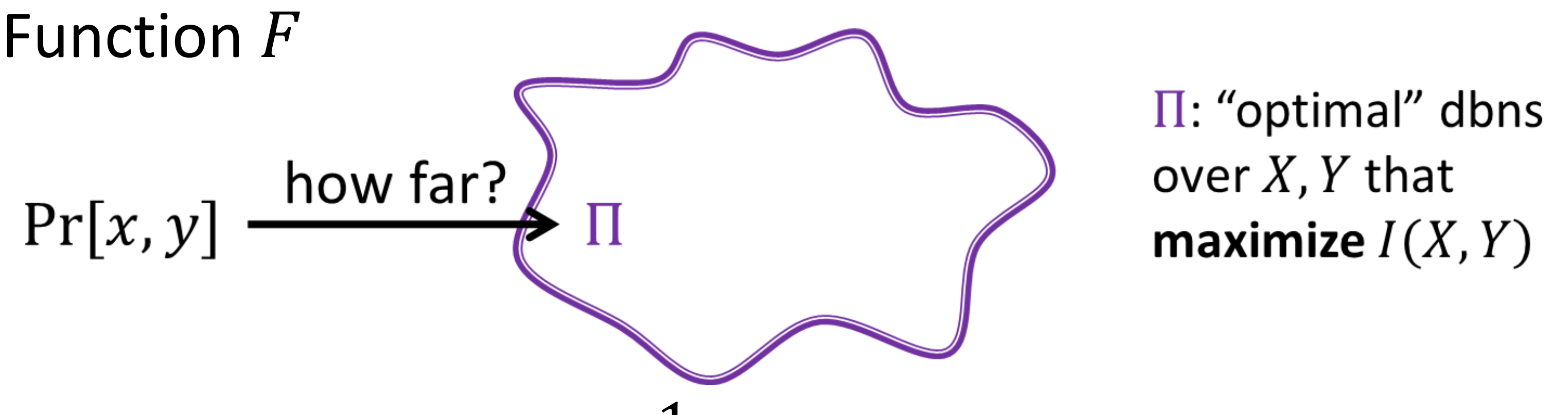
range (info)	$O(1)$	sensitivity (noise)	$O(\log n / n)$

Problem: low info-noise ratio

Solution: design a new function F that (i) has a higher info-noise ratio; (ii) has a strong positive correlation with I

Idea: define F to agree with I at maximum values and interpolate linearly in-between

❑ Function F

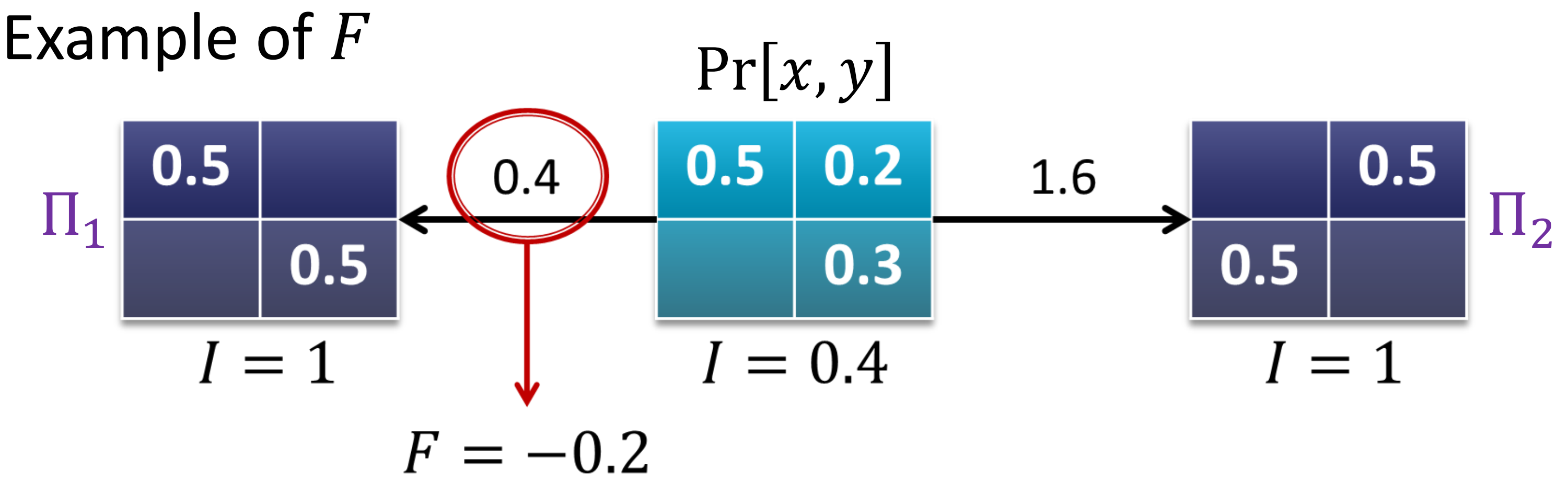


Π: "optimal" dbns over X, Y that **maximize** $I(X,Y)$

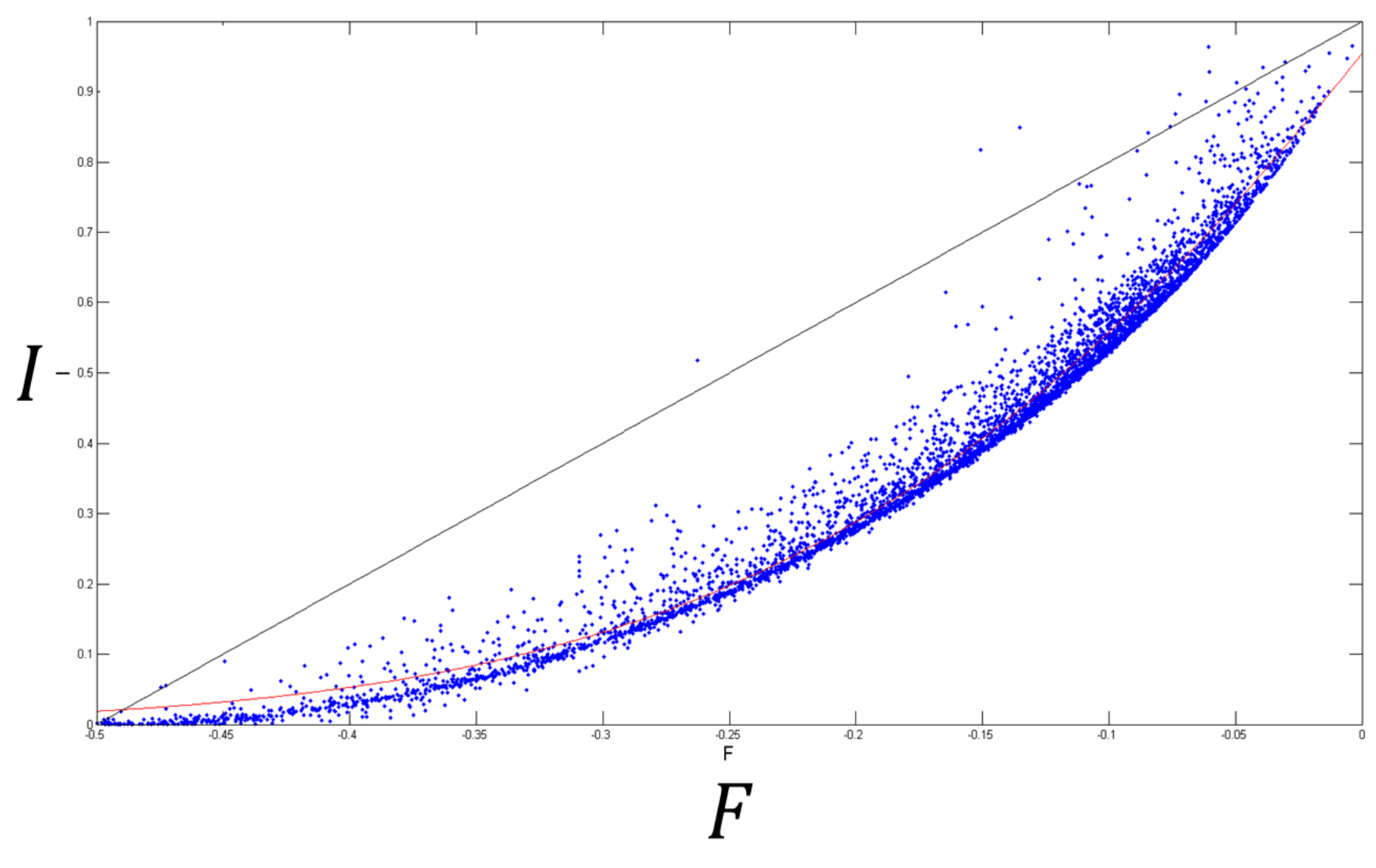
$$F(X,Y) = -\frac{1}{2}\min_{\Pi:optimal} \|\Pr[x,y] - \Pi\|_1$$

range (info)	$O(1)$	sensitivity (noise)	$O(1/n)$

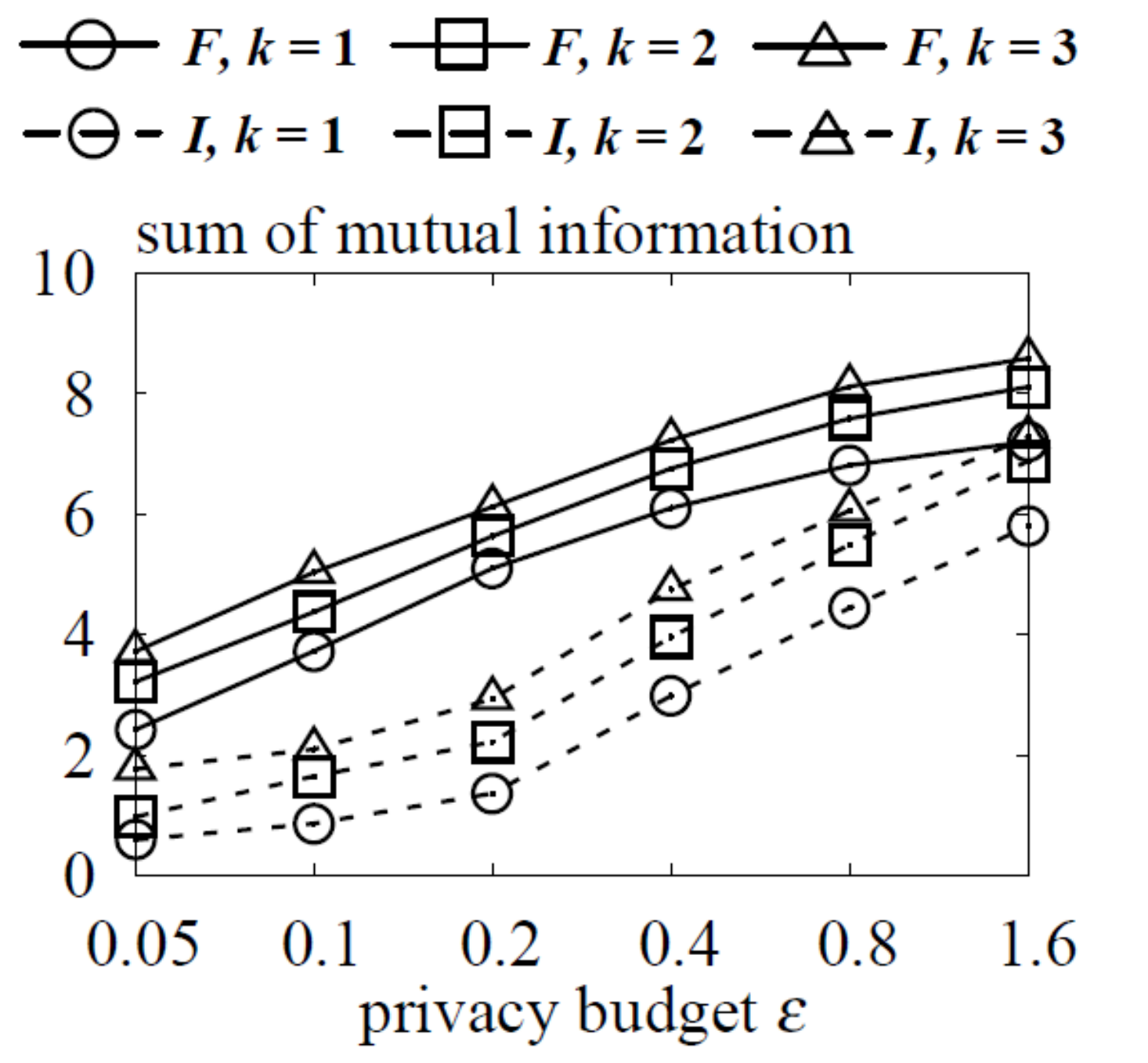
❑ Example of F



❑ F vs. I



F and I of random distributions
correlation coefficient $r = 0.9472$

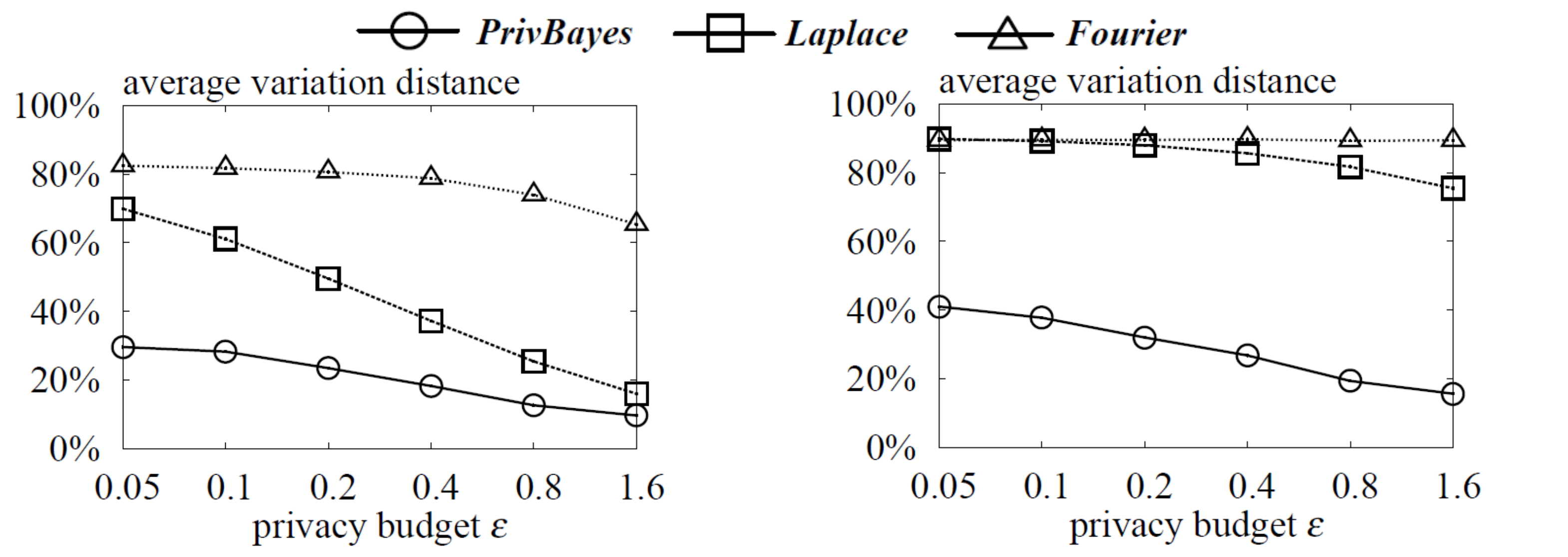


quality of private Bayesian network using F or I

4. Experiments

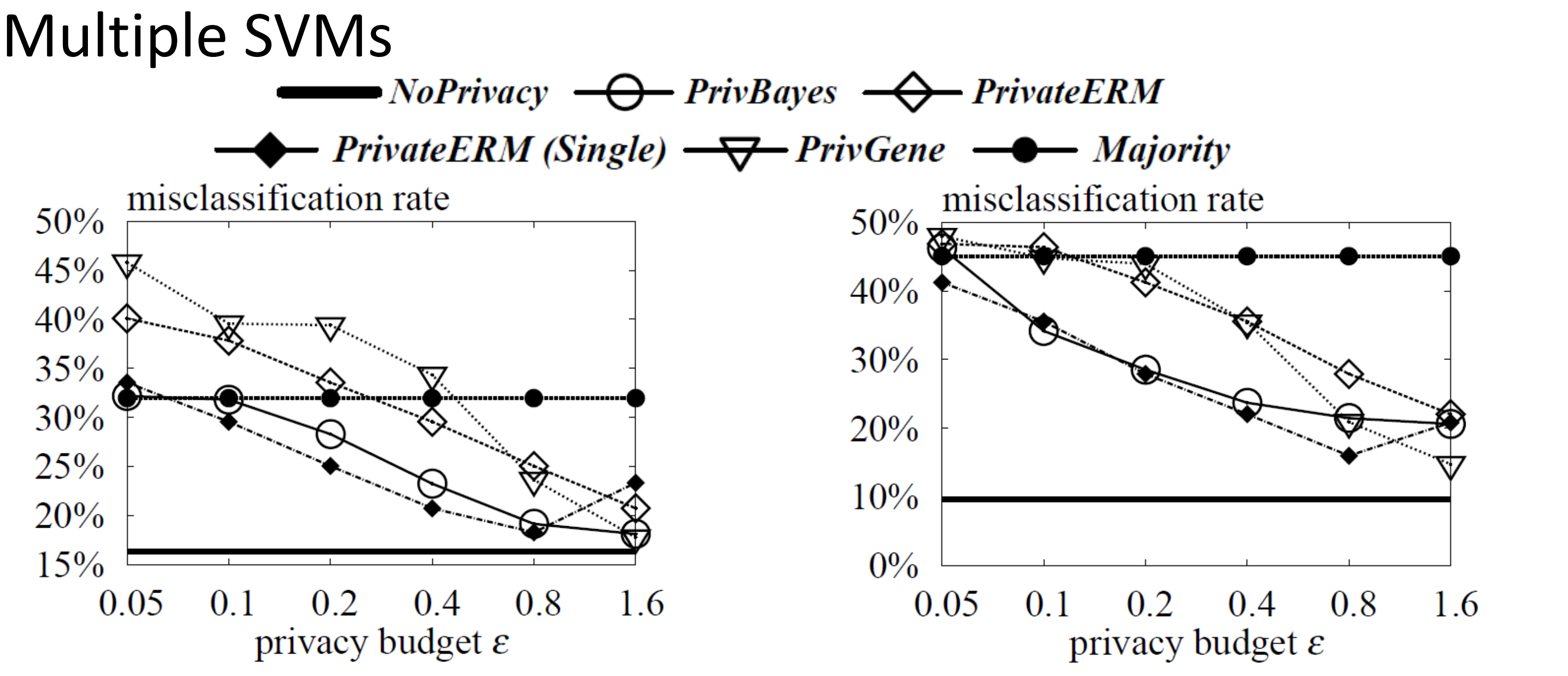
We apply PrivBayes to generate one synthetic dataset, to answer a set of counting and SVM training queries simultaneously.

❑ Counting queries



(a) Adult, Q_2 (b) Adult, Q_3

❑ Multiple SVMs



(a) Adult, Y = gender (b) Adult, Y = education