Leakage of Dataset Properties in Multi-Party Machine Learning

Wanrong Zhang, Harvard University

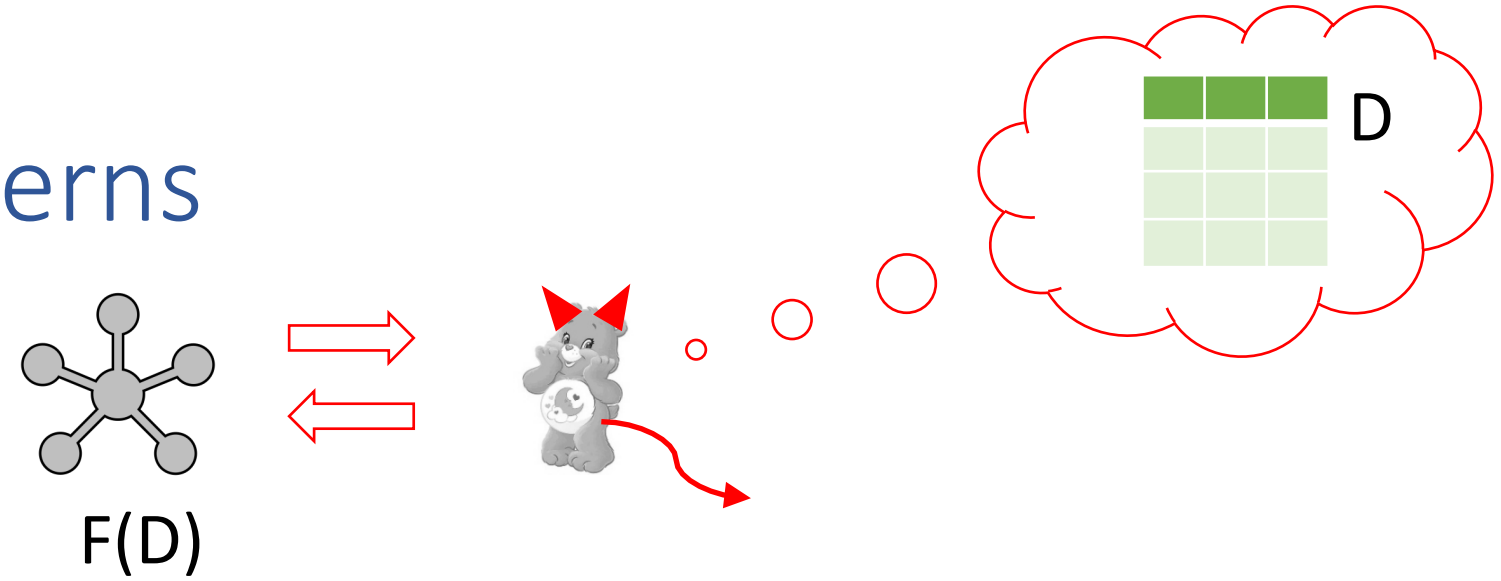
Shruti Tople, Microsoft Research

Olya Ohrimenko, The University of Melbourne

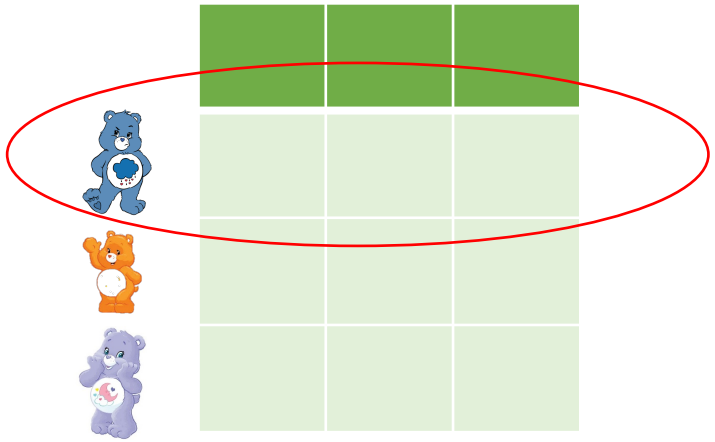
USENIX Security Symposium. 2021

https://arxiv.org/pdf/2006.07267.pdf

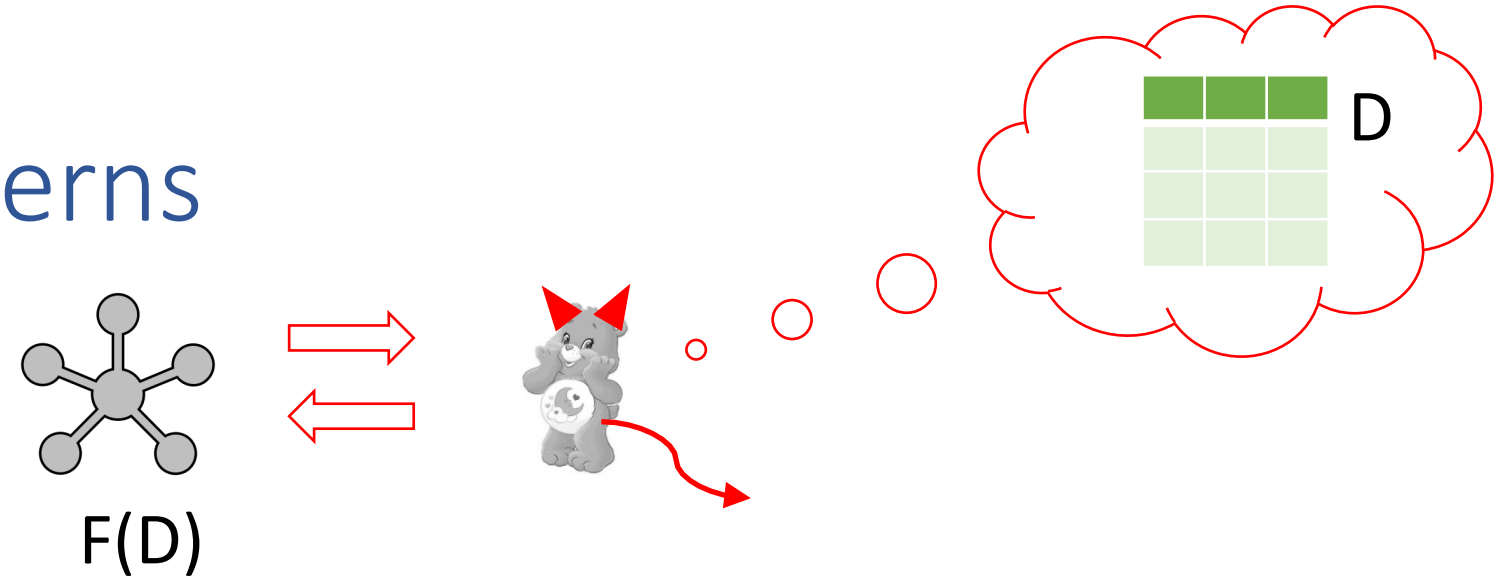
Privacy Concerns



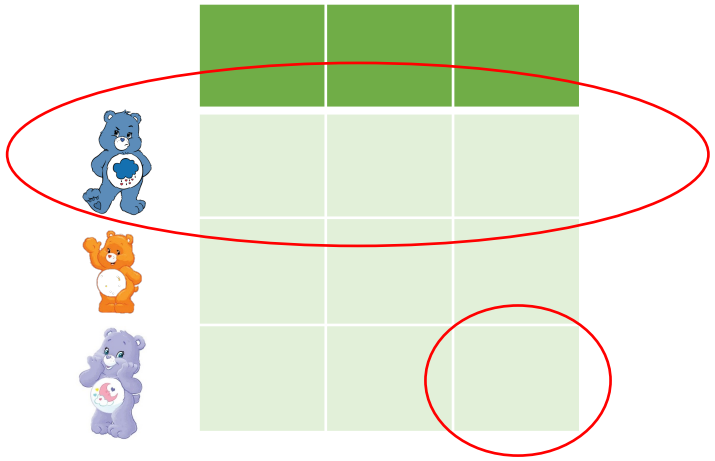
- Membership [Shokri et al, 2017]



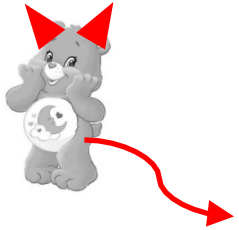
Privacy Concerns

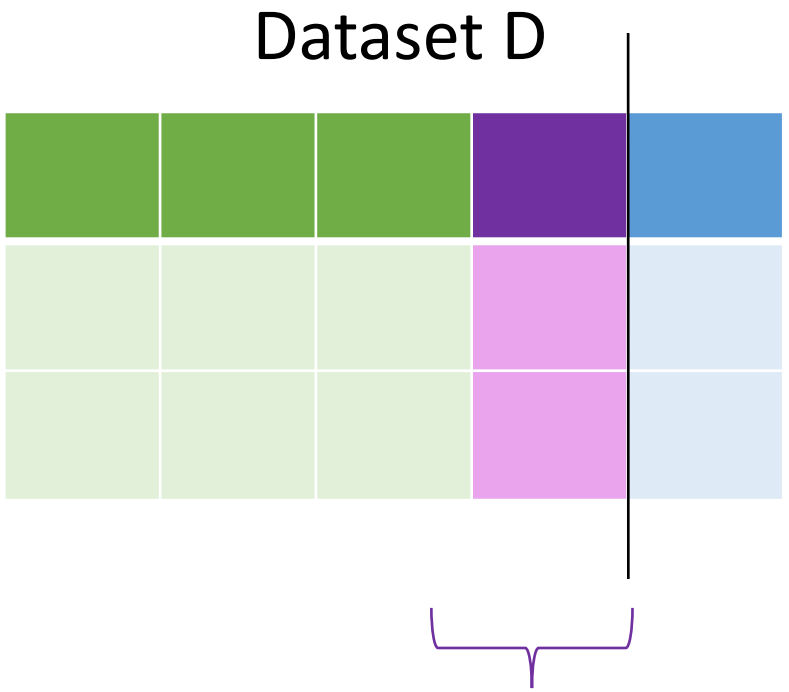


- Membership [Shokri et al, 2017]
- Individual Attributes [Fredrikson et al, 2015]



Dataset-level Privacy





Examples:

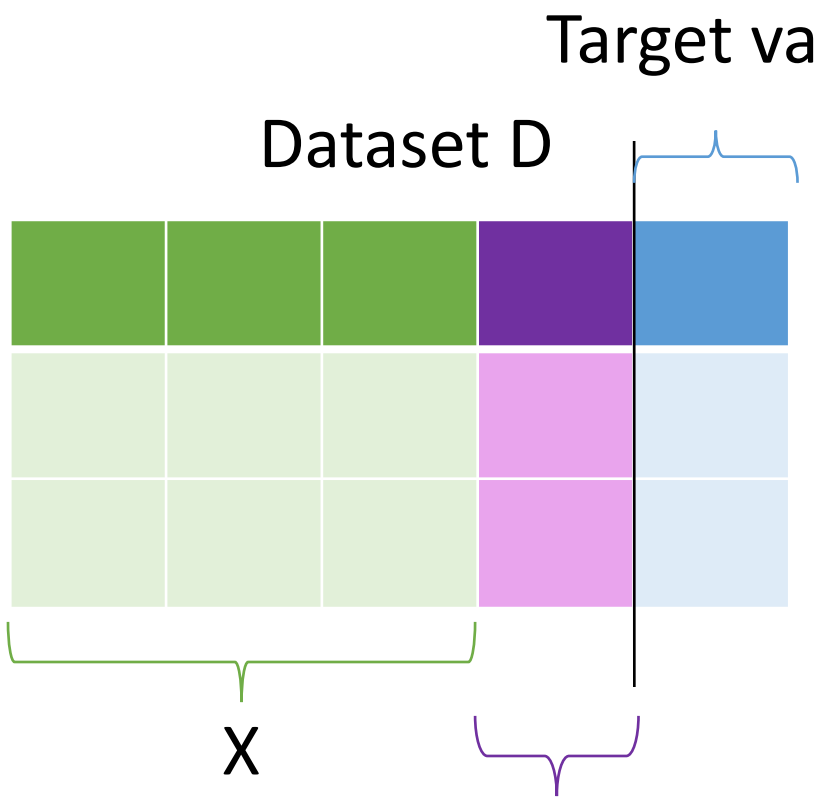
Hospital releasing information about patients' stays:
- Can gender and race distribution of the patients be leaked?

Insurance company providing quotes:
- Can income distribution of its customers be leaked?

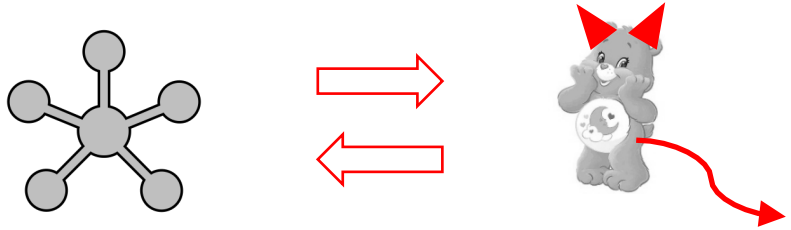
Pharmaceutical company releasing information about a new drug:
- Can proportion of chemicals be leaked?

Advertising:
- Can an ad reveal most purchased product of the company?

Dataset-level Privacy



Release f(D)



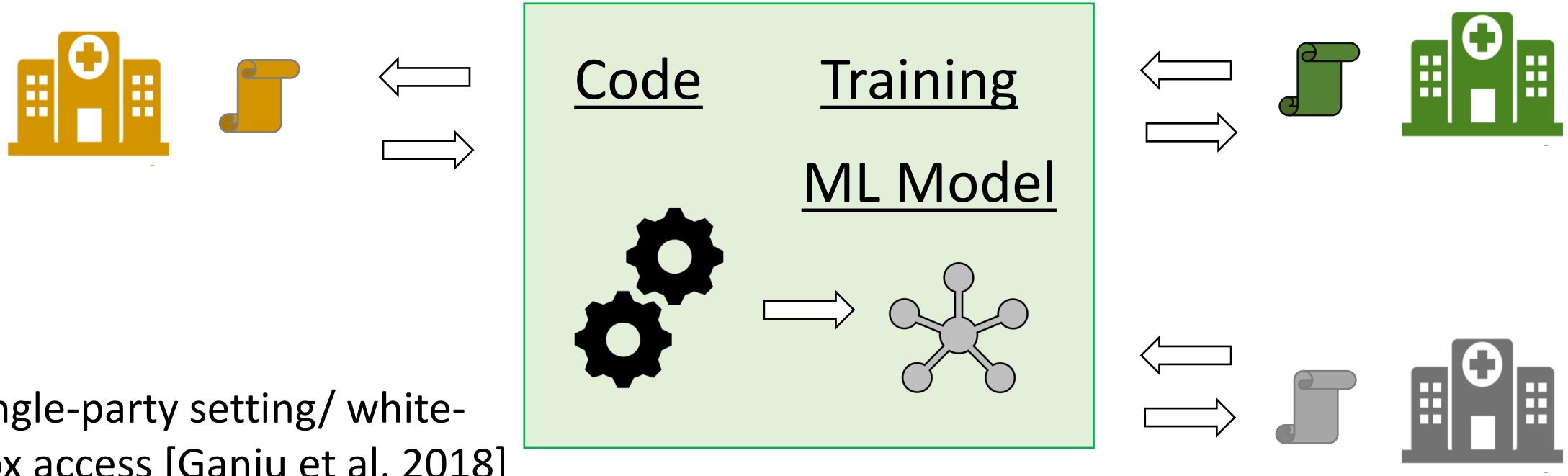
What does f(D) reveal about A?

When A is dropped during training

When A has low correlation with target variable Y

Multi-Party Data Analysis

- Multi-party data analysis where several parties combine data and perform analysis on shared data

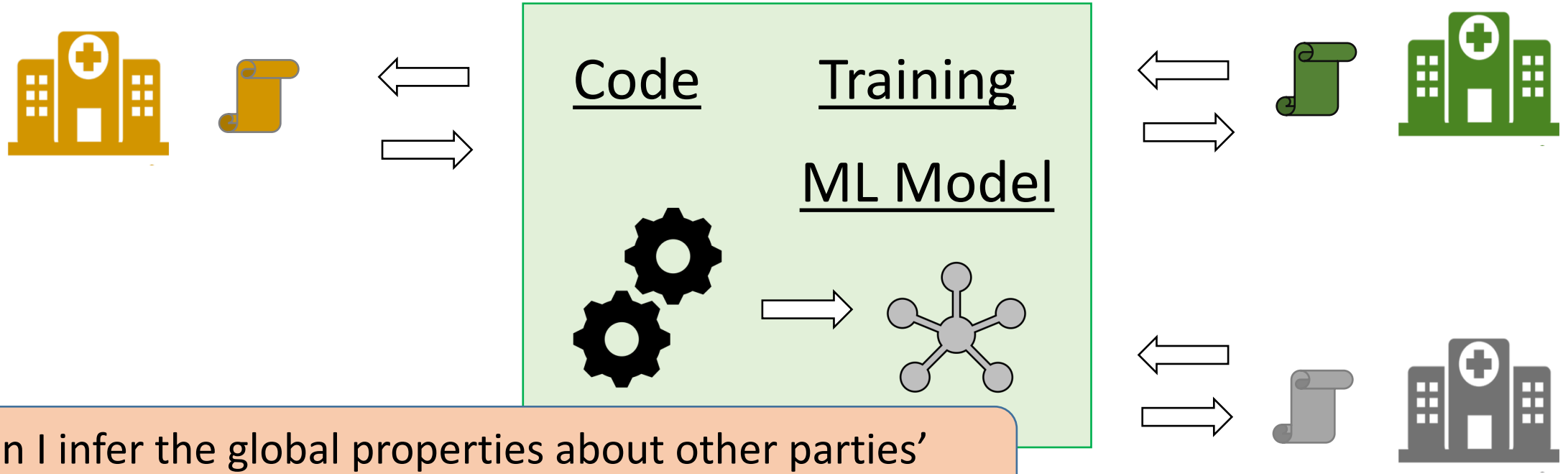


Single-party setting/ white-box access [Ganju et al. 2018]

Threat Model

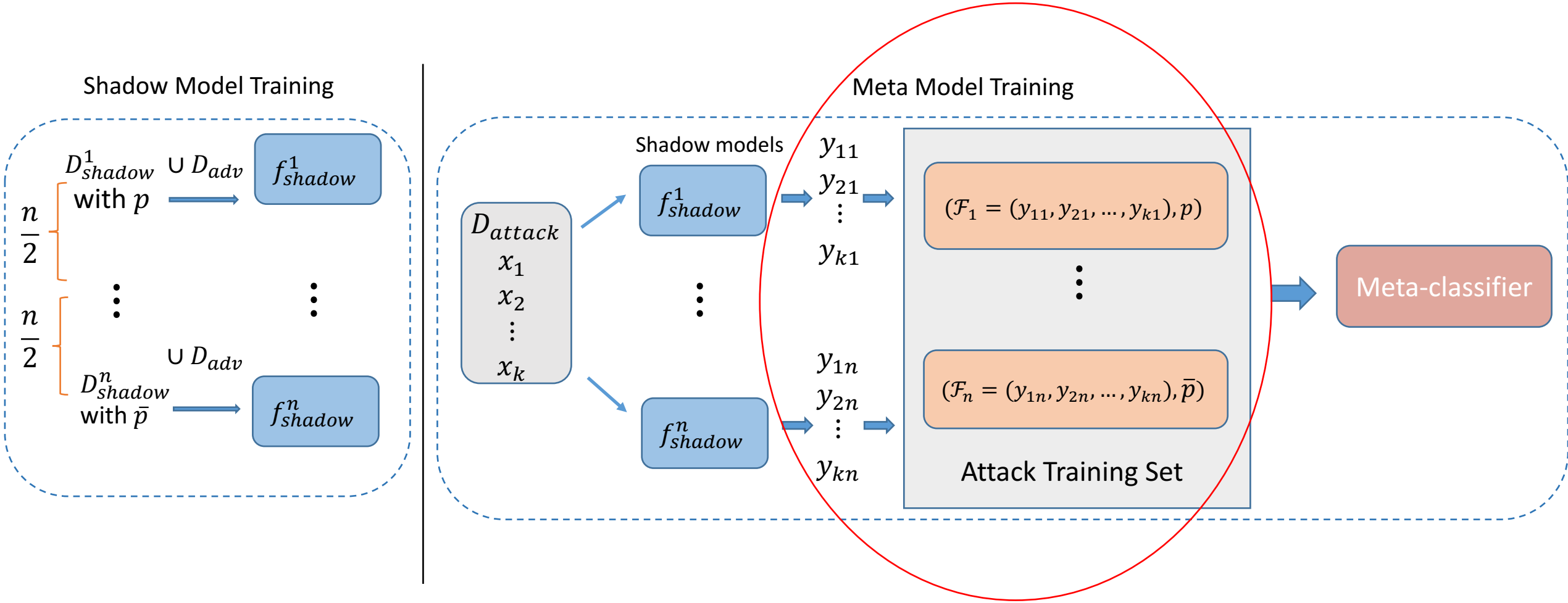
Honest-but-curious

Black-box access

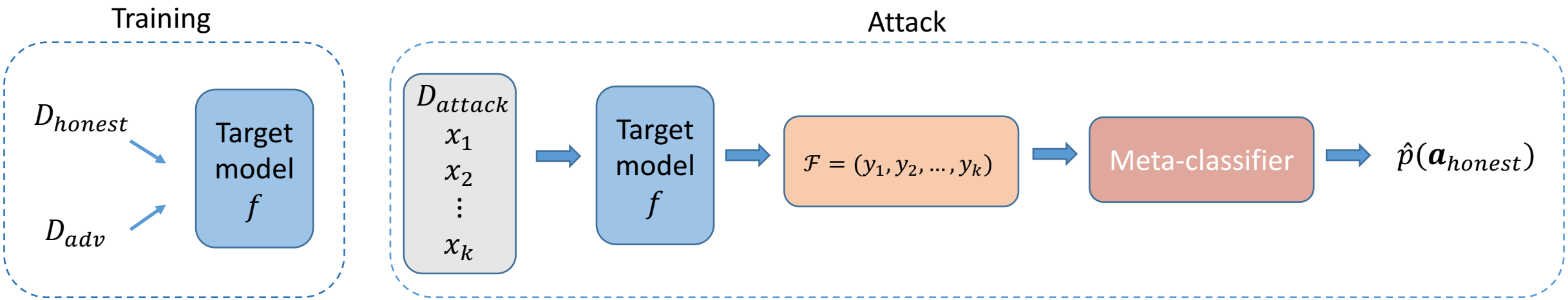


Can I infer the global properties about other parties' sensitive attribute A?

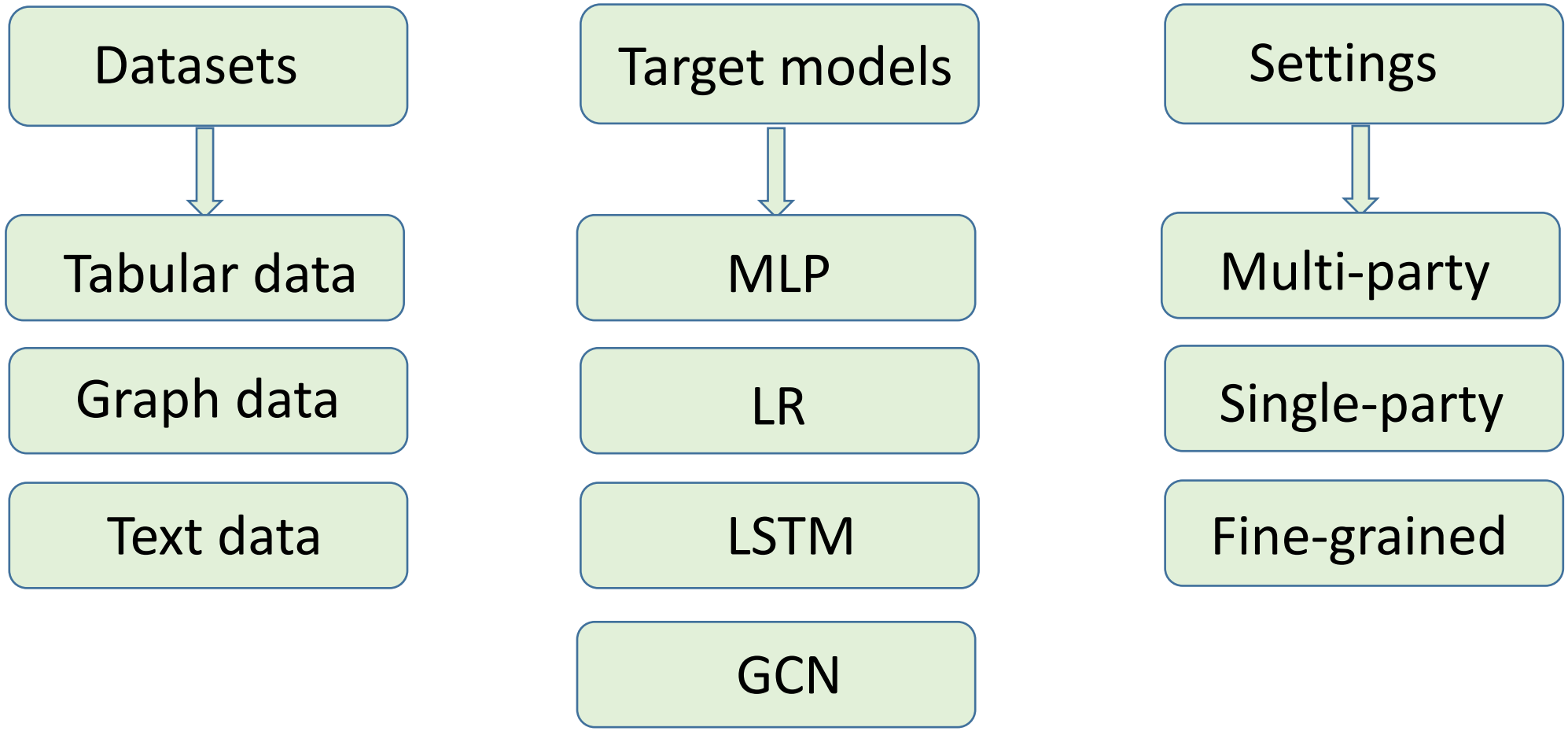
Our attack


Shadow Model Training
D^1_{shadow} with p
$\cup D_{adv}$
f^1_{shadow}
$\frac{n}{2}$
$\frac{n}{2}$
$\cup D_{adv}$
D^n_{shadow} with $\bar{p}$
f^n_{shadow}
Meta Model Training
Shadow models
D_{attack}
x_1
x_2
x_k
f^1_{shadow}
f^n_{shadow}
y_{11}
y_{21}
y_{k1}
y_{1n}
y_{2n}
y_{kn}
$(\mathcal{F}_1 = (y_{11}, y_{21}, \dots, y_{k1}), p)$
$(\mathcal{F}_n = (y_{1n}, y_{2n}, \dots, y_{kn}), \bar{p})$
Attack Training Set
Meta-classifier

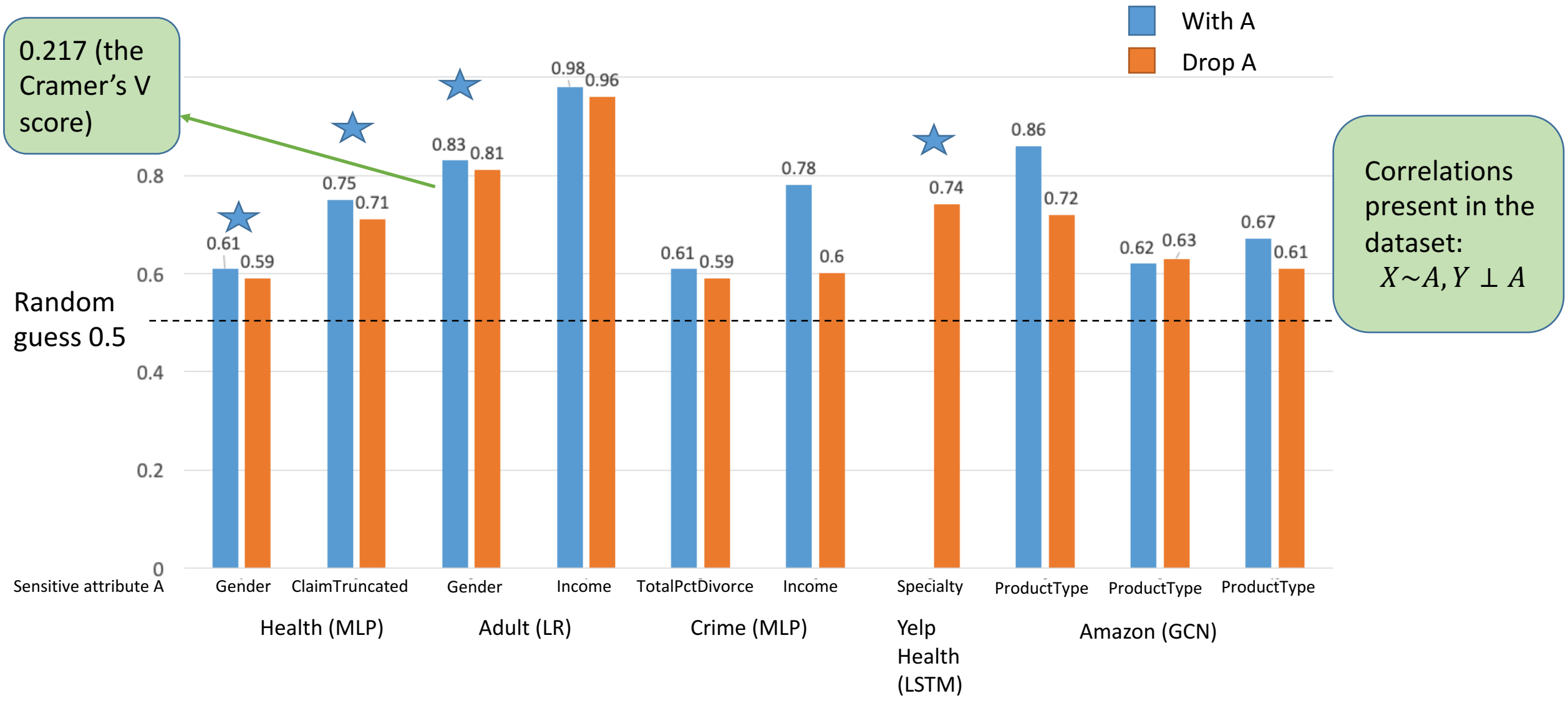
Our attack


Training
D_{honest}
D_{adv}
Target model f
Attack
D_{attack}
x_1
x_2
$\vdots$
x_k
Target model f
$\mathcal{F} = (y_1, y_2, \dots, y_k)$
Meta-classifier
$\hat{p}(\boldsymbol{a}_{honest})$

Experimental Setup



Attack Results



Conclusions

Leakage of sensitive dataset properties is possible even when

- the sensitive attribute column is dropped during training
- the sensitive attribute has low or no correlation with the final task.

Open Question

How to protect dataset-level property leakage since secure computation and differential privacy are not directly applicable to protect leakage of population-level properties.

Thanks!

Contact info:

Wanrong Zhang, wanrongz@gatech.edu

Shruti Tople, Shruti.Tople@microsoft.com

Olya Ohrimenko, oohrimenko@unimelb.edu.au