Algorithms and Incentives in Machine Learning

Princeton CSML Seminar and ECE Korhammer Seminar Series

11/2023

Haifeng Xu

Department of Computer Science
and Data Science Institute

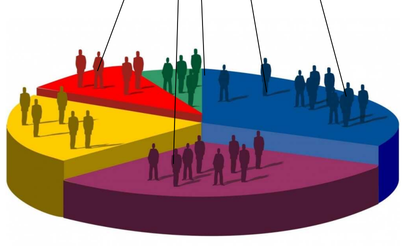
University of Chicago

SIGMA LAB
Strategic IntelliGence in Machine Agents

A classic paradigm of machine learning…

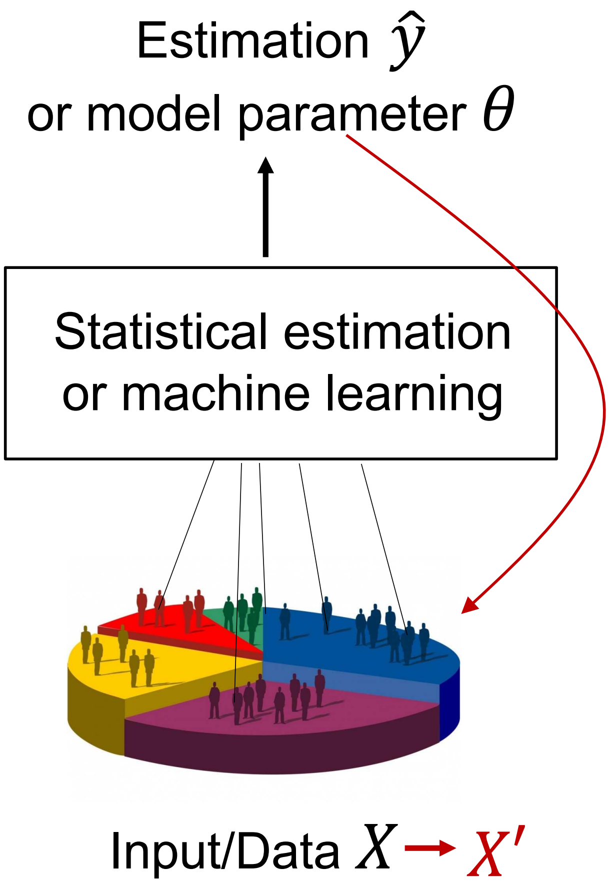
Estimation $\hat{y}$
or model parameter θ

Statistical estimation
or machine learning



Input/Data X

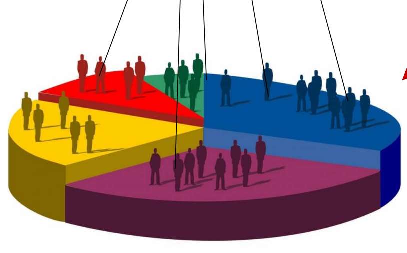
In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data



In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data

Estimation $\hat{y}$
or model parameter θ

Statistical estimation or machine learning



Input/Data $X \rightarrow X'$

✓ Classify loan applicants

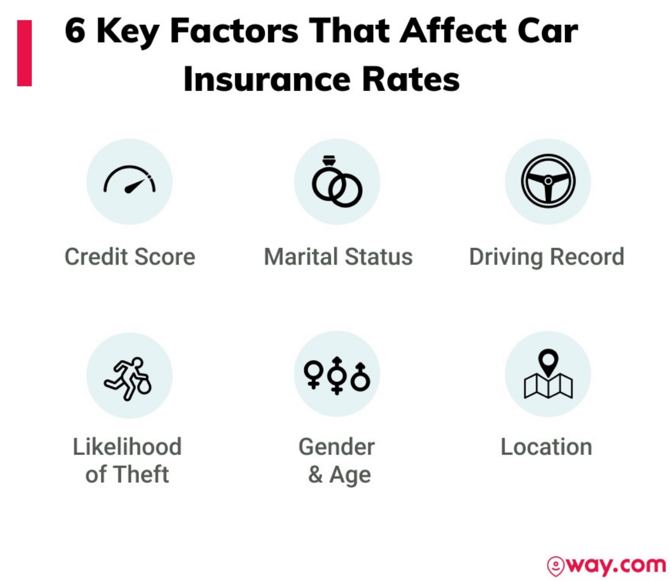
In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data

Estimation $\hat{y}$
or model parameter θ

Statistical estimation or machine learning

Input/Data $X \rightarrow X'$

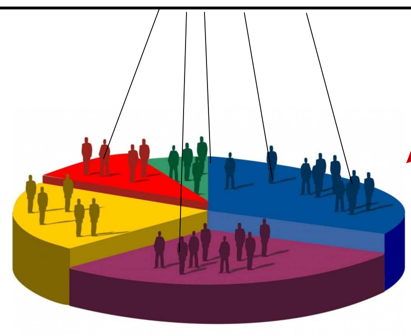
- ✓ Classify loan applicants
- ✓ Estimate insurance rate for applicants



In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data

Estimation $\hat{y}$
or model parameter θ

Statistical estimation
or machine learning



Input/Data $X \rightarrow X'$

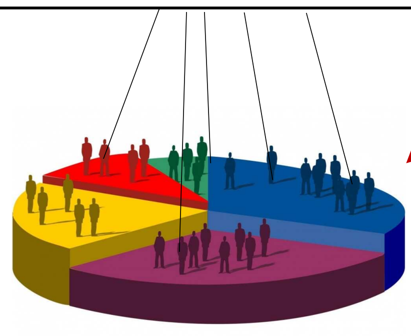
- ✓ Classify loan applicants
- ✓ Estimate insurance rate for applicants
- ✓ Learning to recommend contents

In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data

Estimation $\hat{y}$
or model parameter θ

↑

Statistical estimation or machine learning



Input/Data $X \rightarrow X'$

- ✓ Classify loan applicants
- ✓ Estimate insurance rate for applicants
- ✓ Learning to recommend contents
- ✓ Spam filters

In many applications, learning outcomes affect data providers' welfare → leading to strategically supplied data

Estimation $\hat{y}$
or model parameter θ

Input/Data $X \rightarrow X'$

- ✓ Classify loan applicants
- ✓ Estimate insurance rate for applicants
- ✓ Learning to recommend contents

This Talk

1. Demonstrate why interesting (practically and theoretically)
2. Solutions that blend learning + incentives + algorithms
3. Illustrate their tradeoff and complementarity

Outline

A timely real-world problem

A well-studied classic model

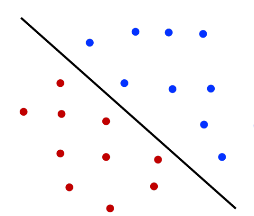
Vignette 1

Elicit truthful information to improve statistical estimation

ICML
International Conference On Machine Learning
NEURAL INFORMATION PROCESSING SYSTEMS
REJECTED

Vignette 2

PAC-Learning in strategic environments



Outline

Joint work with

Jibang Wu
(UChicago, CS)

Yifan Guo
(USTC, Math)

Weijie Su
(UPenn, Wharton)

A timely real-world problem

Vignette 1

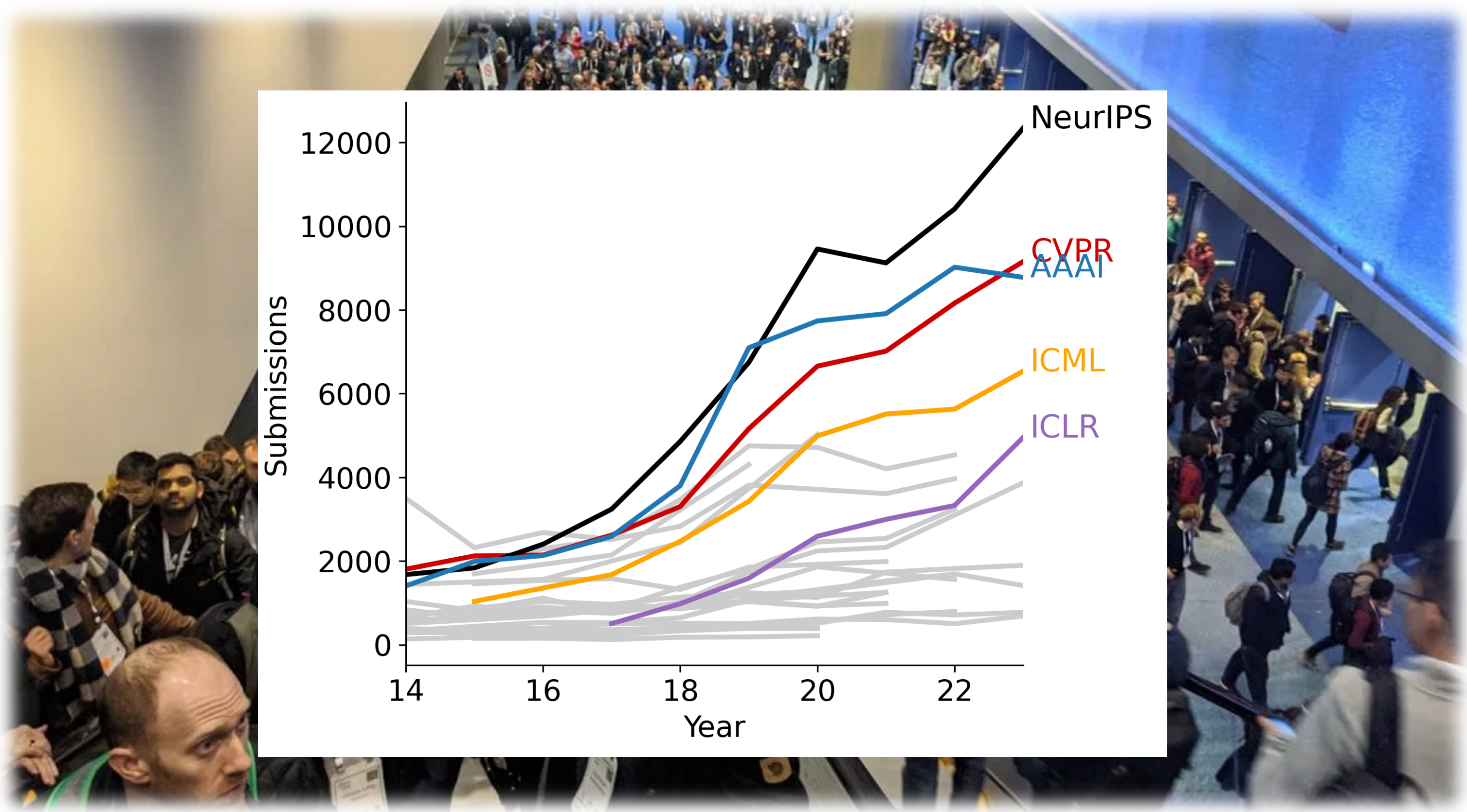
Elicit truthful information to improve statistical estimation

ICML
International Conference
On Machine Learning

NEURAL INFORMATION
PROCESSING SYSTEMS

REJECTED

A Concern of ML Venues – Massive Sizes



SIGMA LAB
Strategic IntelliGence in Machine Agents

Lack of Qualified Reviewers ⇒ Large Noise

- 70% of reviewers in NeurIPS 2016 are PhD students [Shah 2022]
- Nowadays, even many undergrad reviewers

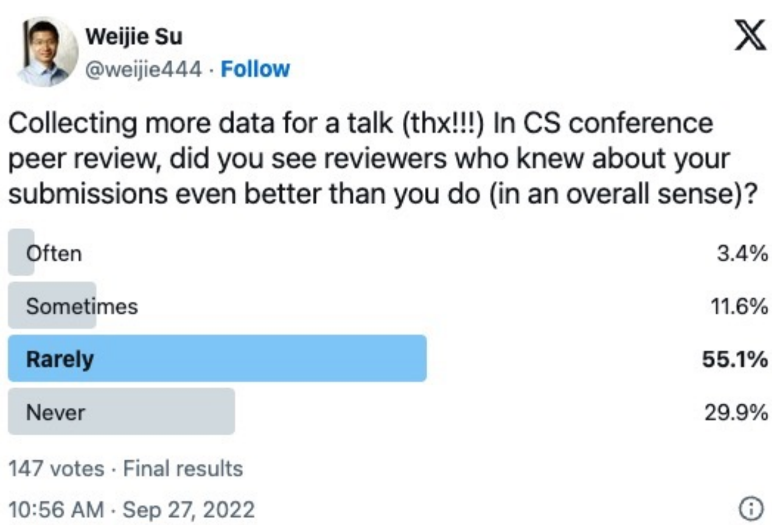
SIGMA LAB
Strategic IntelliGence in Machine Agents

This work tries to develop a workable solution

<u>Core idea</u>: authors' own information about their papers is another source of data for improving paper score estimation

Why?

Authors often have good knowledge about their own papers



However, Challenges Remain

Challenge 1: what information to elicit from authors?

- Cannot be too fine-grained
- Cannot be too coarse neither (then not that useful)

A good compromise: authors' ranking of their papers

Challenge 2: how to guarantee authors will tell truthful information?

Smart strategic agents!

SIGMA LAB
Strategic IntelliGence in Machine Agents

However, Challenges Remain

Challenge 1: what information to elicit from authors?

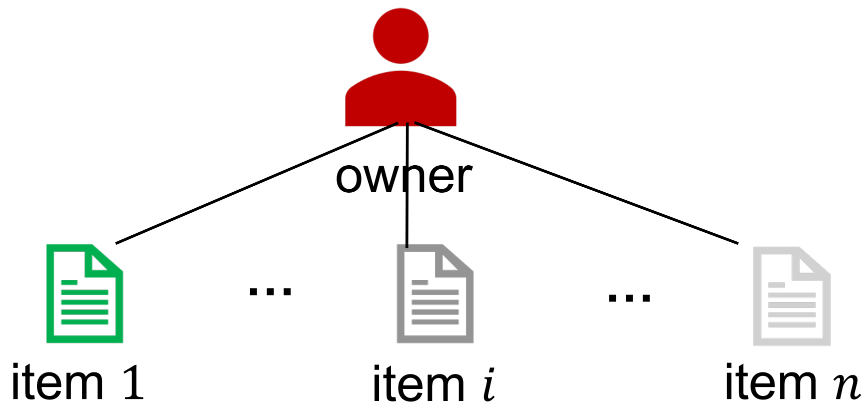
- Cannot be too fine-grained
- Cannot be too coarse neither (then not that useful)

A good compromise: authors' ranking of their papers

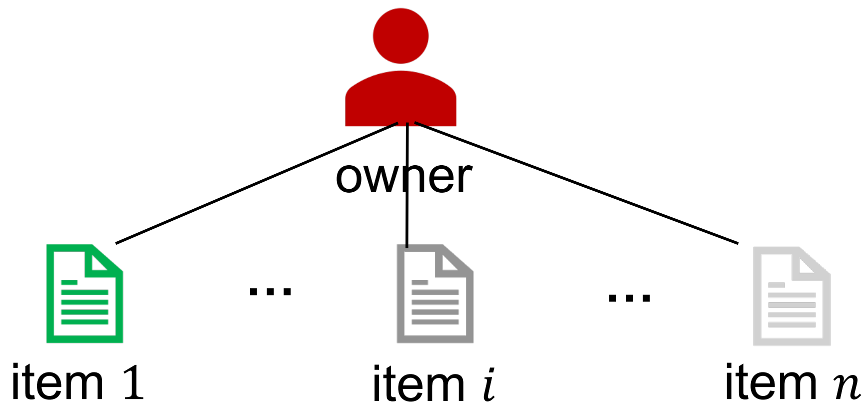
Challenge 2: how to guarantee authors will tell truthful information?

- Estimation method has to be designed so that information elicitation is aligned with authors' incentives

It Can Work in Idealized Situations! [Su, NeurIPS'21]



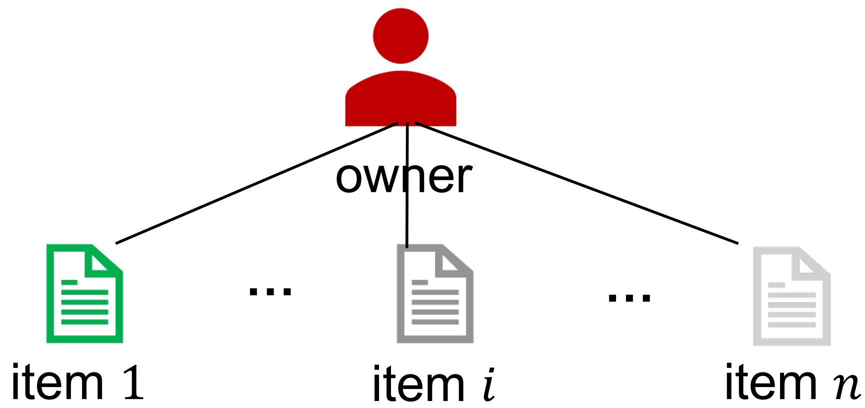
Formal Model



- Ground-truth score: $\boldsymbol{R} = (R_1, \cdots, R_n)$
- Review score: $y_i = R_i + z_i$ (noise)
- Designer's task:
  1. Ask for owner's ranking π of her items
  2. Use π and $\{y_i\}_i$ to compute refined scores $\widehat{\boldsymbol{R}}(\pi, \{y_i\}_i)$
- Owner derives utility $U(\hat{R}_1, \cdots, \hat{R}_n)$ from output scores

The design of $\widehat{\boldsymbol{R}}$ function matters – it may be gamed!

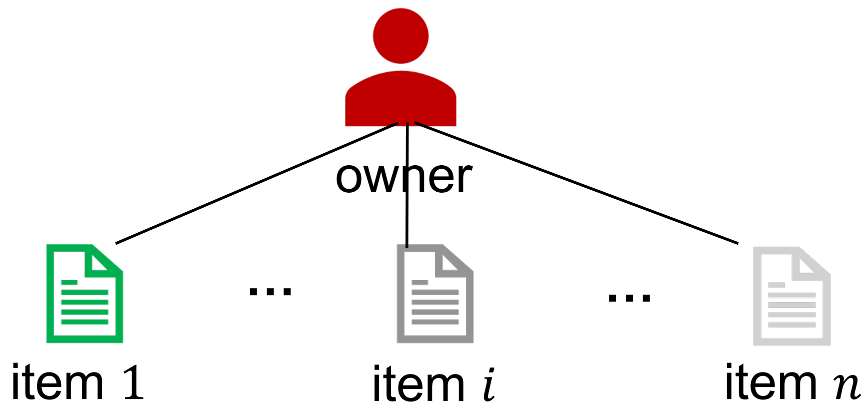
A Simple and Elegant Solution



Isotonic regression

$$\hat{R} = \arg\min_{\mathbf{r}} \quad \|\mathbf{y} - \mathbf{r}\|^2$$
$$\text{s.t.} \quad r_{\pi(1)} \geq r_{\pi(2)} \geq \cdots \geq r_{\pi(n)}$$

A Simple and Elegant Solution



Isotonic regression

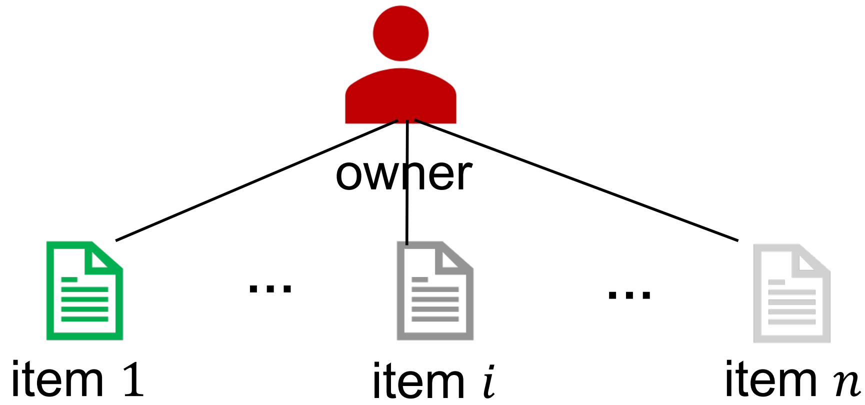
$$\hat{R} = \underset{\mathbf{r}}{\arg\min} \quad \|\mathbf{y} - \mathbf{r}\|^2$$
$$\text{s.t.} \quad r_{\pi(1)} \geq r_{\pi(2)} \geq \cdots \geq r_{\pi(n)}$$

Thm [Su,'21]: Suppose owner's utility function $U(\widehat{\boldsymbol{R}})$ is convex, then isotonic mechanism is truthful.

- Formally, suppose π^* is true ranking of R_i's, then

$$\mathbb{E}_{\text{noisy } \boldsymbol{y}} U\left(\widehat{\boldsymbol{R}}(\pi^*, \boldsymbol{y})\right) \geq \mathbb{E}_{\text{noisy } \boldsymbol{y}} U\left(\widehat{\boldsymbol{R}}(\pi, \boldsymbol{y})\right), \qquad \forall\, \pi$$

A Simple and Elegant Solution

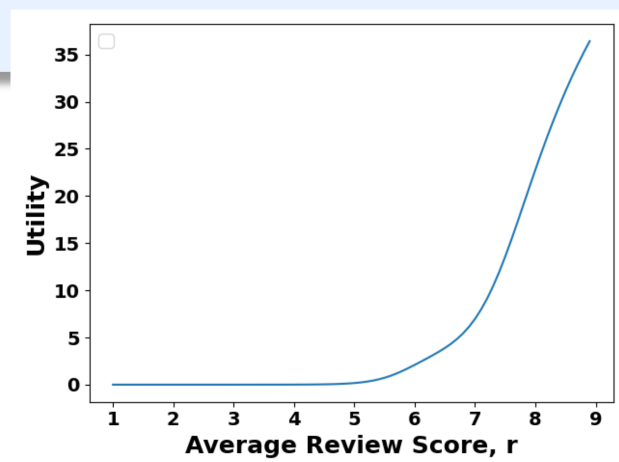


Isotonic regression

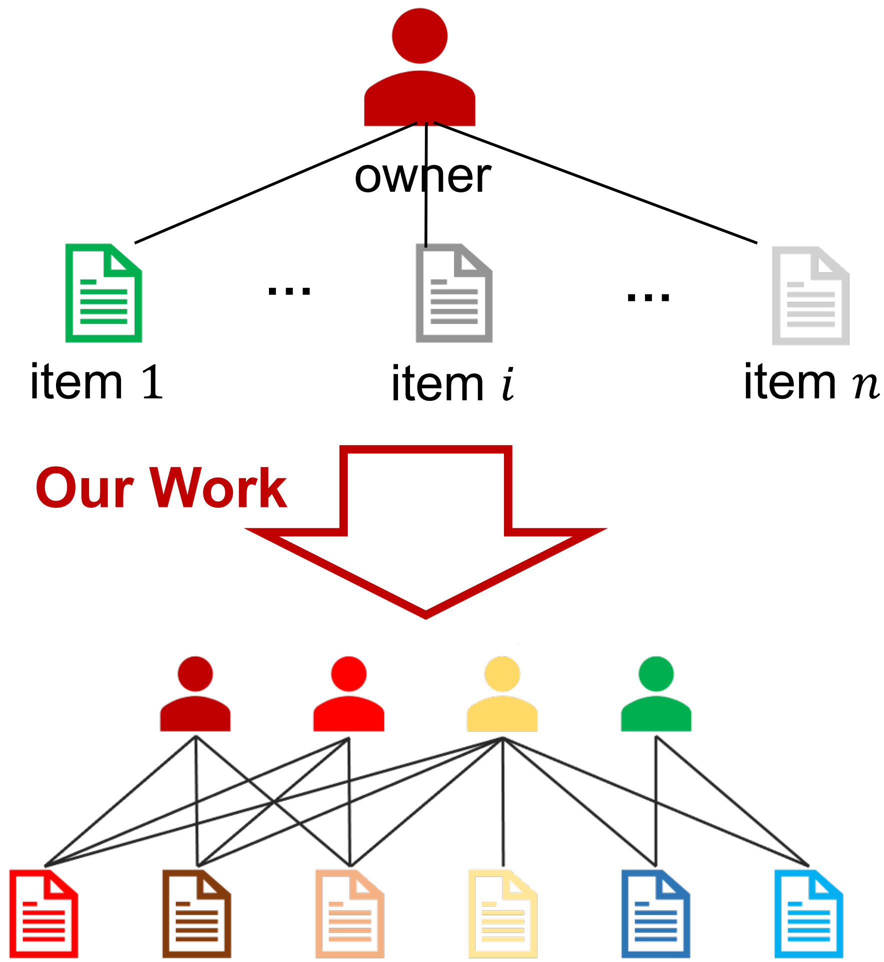
$$\begin{aligned} \hat{R} = \ & \arg\min_{\mathbf{r}} \ \|\mathbf{y} - \mathbf{r}\|^2 \\ \text{s.t.} \ & r_{\pi(1)} \geq r_{\pi(2)} \geq \cdots \geq r_{\pi(n)} \end{aligned}$$

Thm [Su,'21]: Suppose owner's utility function $U(\widehat{\boldsymbol{R}})$ is convex, then isotonic mechanism is truthful.

- Convex utility captures the high-risk-high-reward nature of research
  - Empirically justified with ICLR'22 data



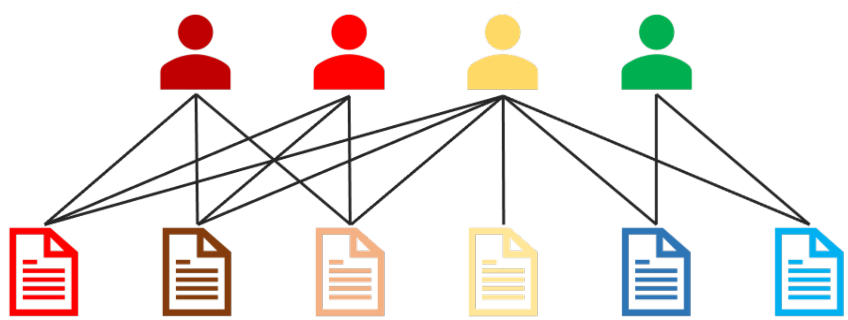
Address More Realistic Peer Review Setups



Address More Realistic Peer Review Setups

Main Question: Can we still elicit truthful information from owners to improve review score estimation?

Ans: Yes, though to some extent



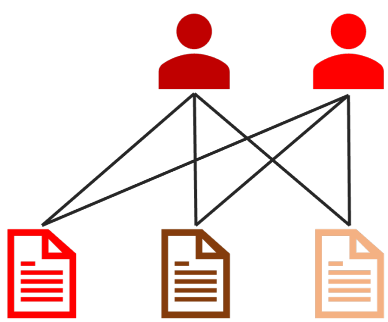
Address More Realistic Peer Review Setups

Main Question: Can we still elicit truthful information from owners to improve review score estimation?

Ans: Yes, though to some extent

Our approaches have two steps:

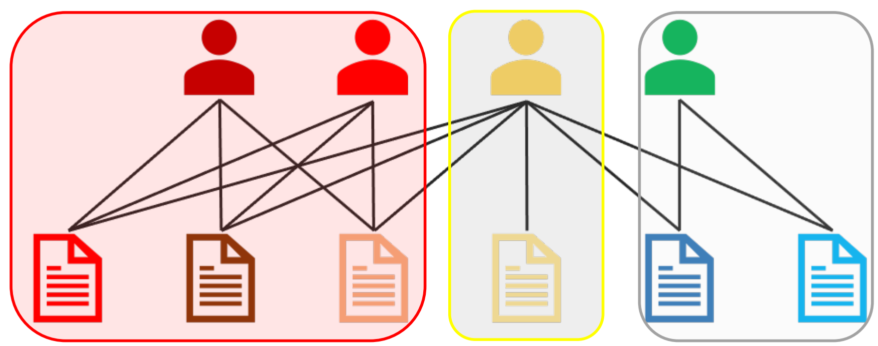
Step 1



Complete ownership

(Statistics + mechanism design)

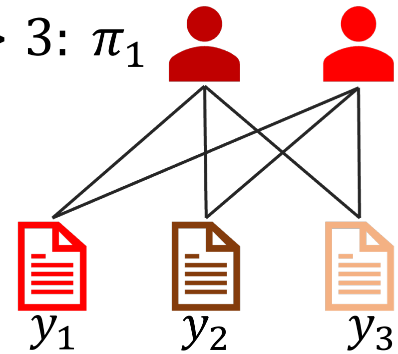
Step 2



Partition general ownership into blocks of complete ownerships (algorithm design)

Step 1: the Complete Ownership Situation

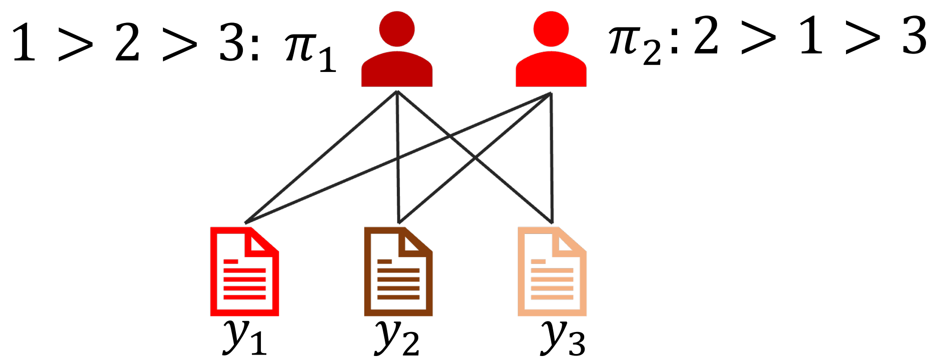
$1 > 2 > 3$: π_1 π_2: $2 > 1 > 3$



Model: the same, except all owners hold ranking information

Goal: elicit information from all owners to refine score estimation

Step 1: the Complete Ownership Situation



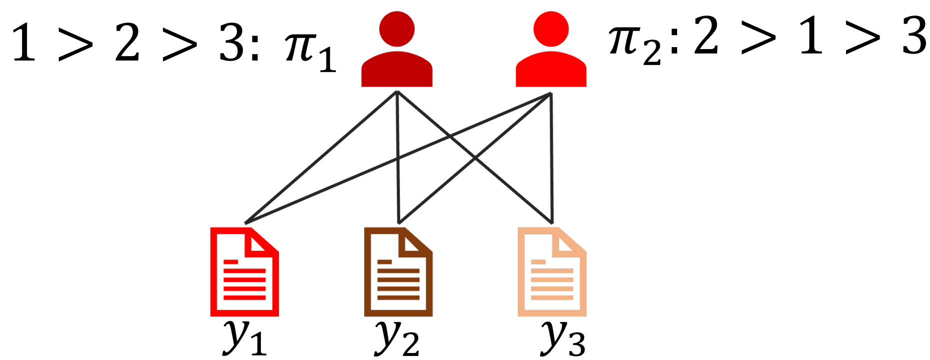
Suppose we get ranking π_j from every owner j, what's the most natural way to calculate estimated score?

The Weighted Isotonic Mechanism

1. Elicit π_j from every j
2. Run isotonic regression to find $\widehat{\boldsymbol{R}}^{(j)} = \mathrm{Isotonic}(\pi_j, \boldsymbol{y})$
3. Output weighted combination $\widehat{\boldsymbol{R}} = \sum_j \alpha_j \widehat{\boldsymbol{R}}^{(j)}$

$\{\alpha_j\}$ can be arbitrary

Step 1: the Complete Ownership Situation



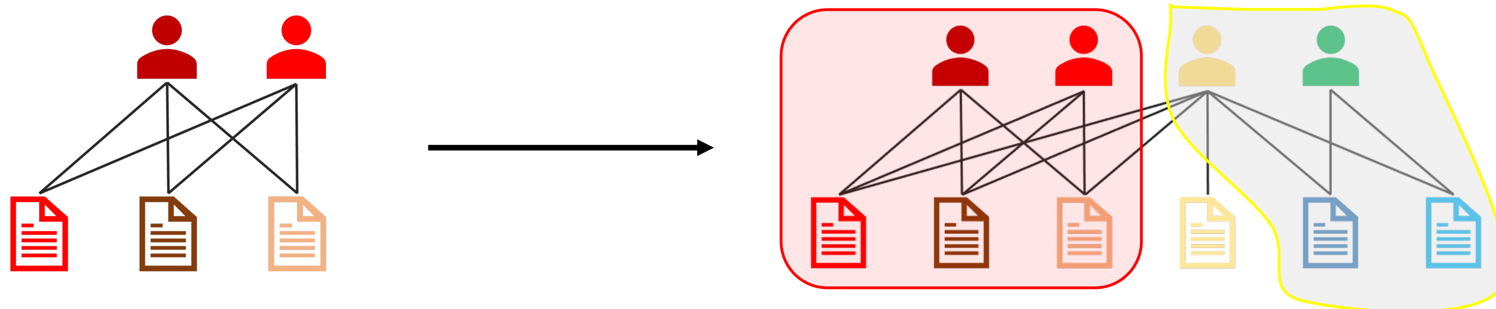
Theorem [WXGS'23]. Under weighted isotonic mechanism and convex utility, every owner reports truthful ranking is a Nash equilibrium (NE)

Moreover, this NE is payoff-dominant – everyone simultaneously achieves highest possible utility among all possible NEs.

- Strong evidence of truthful behaviors
- Generalizes truthful behavior in previous single-agent optimization [Su'21] to truthful behaviors in multi-agent strategic gaming
- Proof uses a new technique *majorization*

An Isotonic Mechanism for Overlapping Ownership. Jibang Wu, Haifeng Xu, Yifan Guo, Weijie Su

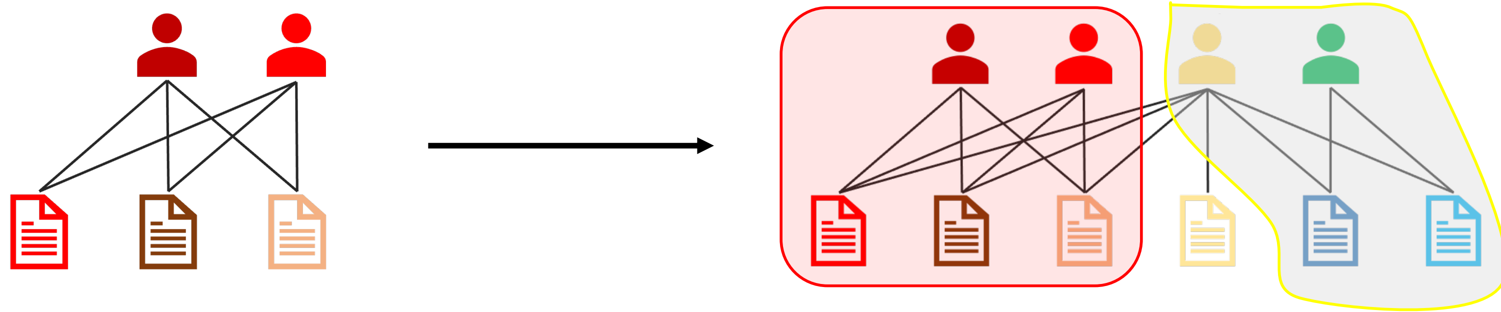
Step 2: Generalizing to Overlapping Ownership



Ideally, we want to elicit ever j's (partial) ranking π_j for all her own items, and design a way to aggregate them $\hat{\boldsymbol{R}}(\pi_1, \cdots, \pi_m; \boldsymbol{y})$

- Design such a statistical estimation $\hat{\boldsymbol{R}}$ seems quite challenging …
- We resort to algorithmic approach – use partition to create independence
  1. Partition ownership graph into blocks, each as a complete ownership
  2. Run previous truthful mechanism independently for each block

Step 2: Generalizing to Overlapping Ownership



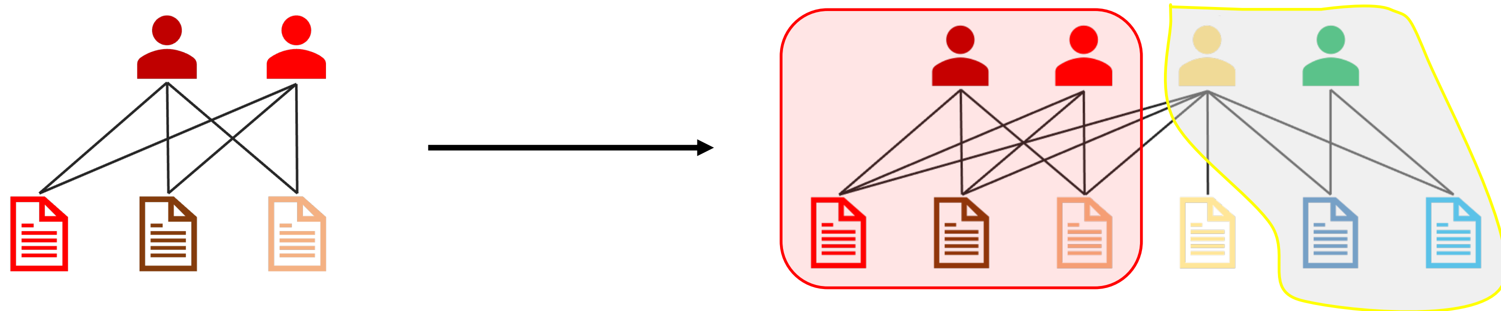
Any partition will lead to truthful equilibrium

Question is which partition gives the “best” score estimation?

- Difficult to statistically quantify how good an estimation is
- However, intuitively, the larger a block is, the better

If $l = l_1 + l_2$, then $w(l) \geq w(l_1) + w(l_2)$

Step 2: Generalizing to Overlapping Ownership



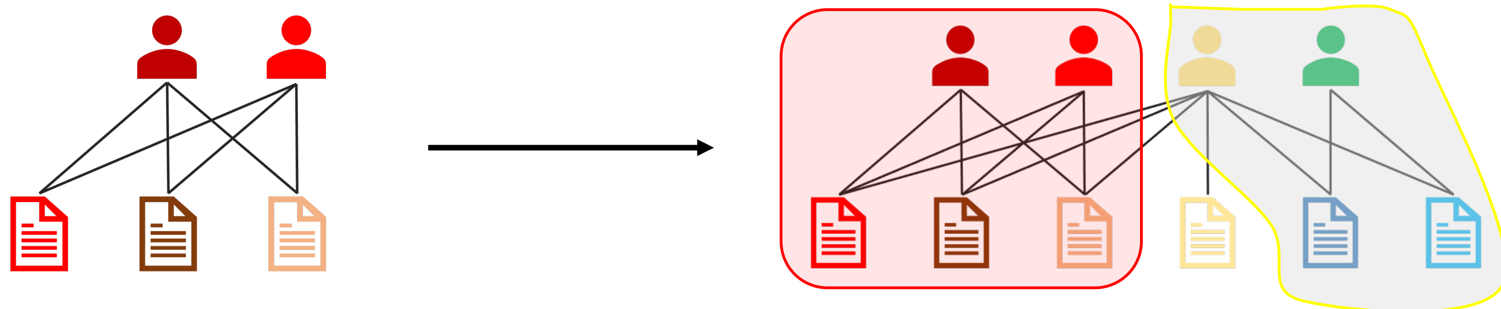
Any partition will lead to truthful equilibrium

Question is which partition gives the "best" score estimation?

- Difficult to statistically quantify how good an estimation is
- Formally, suppose block sizes are $l_1, l_2, \cdots, l_k$,
  partition wellness = $w(l_1) + w(l_2) + \cdots + w(l_k)$ for some convex w

If $l = l_1 + l_2$, then $w(l) \geq w(l_1) + w(l_2)$

Step 2: Generalizing to Overlapping Ownership



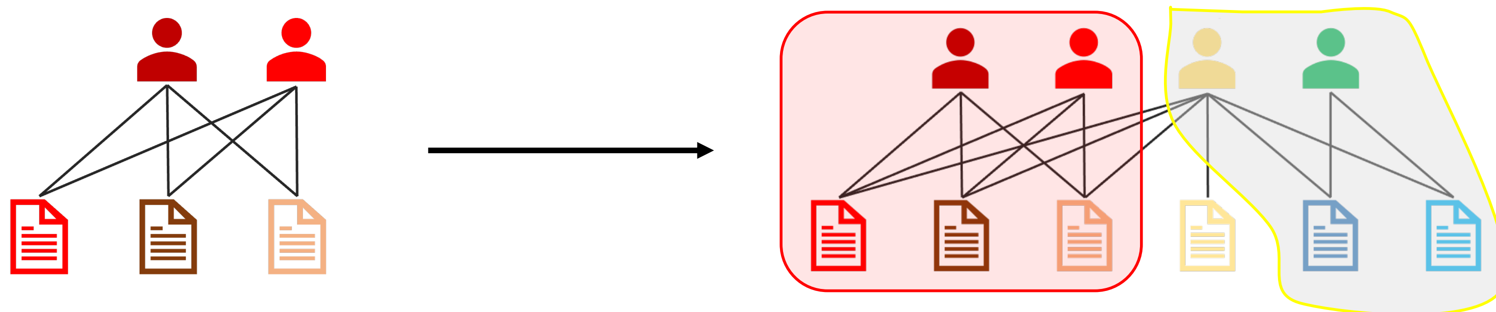
Any partition will lead to truthful equilibrium

Question is which partition gives the "best" score estimation?

- Difficult to statistically quantify how good an estimation is
- Formally, suppose block sizes are $l_1, l_2, \cdots, l_k$,
  partition wellness = $w(l_1) + w(l_2) + \cdots + w(l_k)$ for some convex w
- What is w ? Impossible to know… → will resort to robust analysis

SIGMA LAB
Strategic IntelliGence in Machine Agents

Step 2: Generalizing to Overlapping Ownership



<u>Partition Optimization</u>

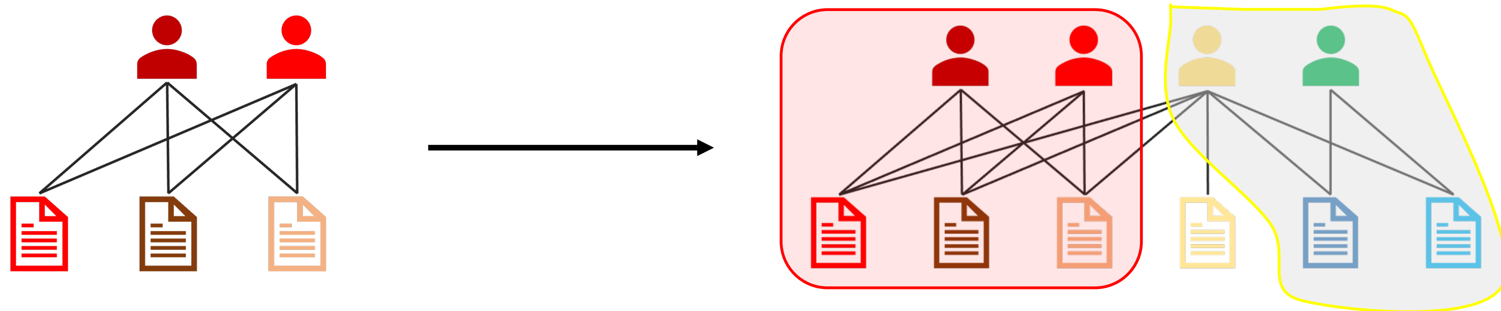
$$\text{maximize}_{l_1,l_2,\cdots}\ [w(l_1) + w(l_2) + \cdots + w(l_k)]$$

$$\text{subject to}\quad \text{each block has} \geq k \text{ owners } (k\text{-strongness})$$

Challenges

- Have to solve this problem "blindly" without knowing w

Step 2: Generalizing to Overlapping Ownership



<u>Partition Optimization</u>

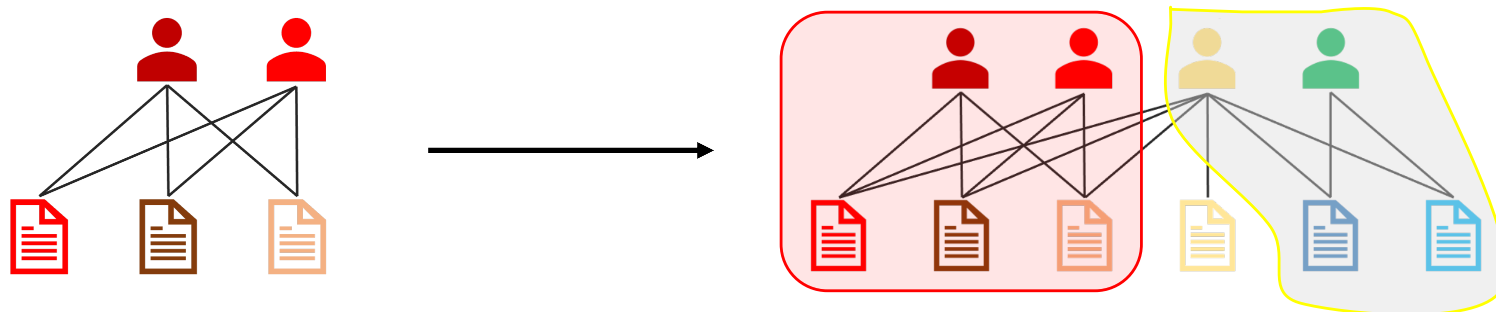
$$\text{maximize}_{l_1, l_2, \cdots} \; [\,?\,(l_1) + \,?\,(l_2) + \cdots + \,?\,(l_k)]$$

$$\text{subject to} \quad \text{each block has} \geq k \text{ owners } (k\text{-strongness})$$

Challenges

- Have to solve this problem "blindly” without knowing w
- Provably NP-hard even for w as simple as $w(l) = \max\{l - 1, 0\}$

Step 2: Generalizing to Overlapping Ownership



Partition Optimization

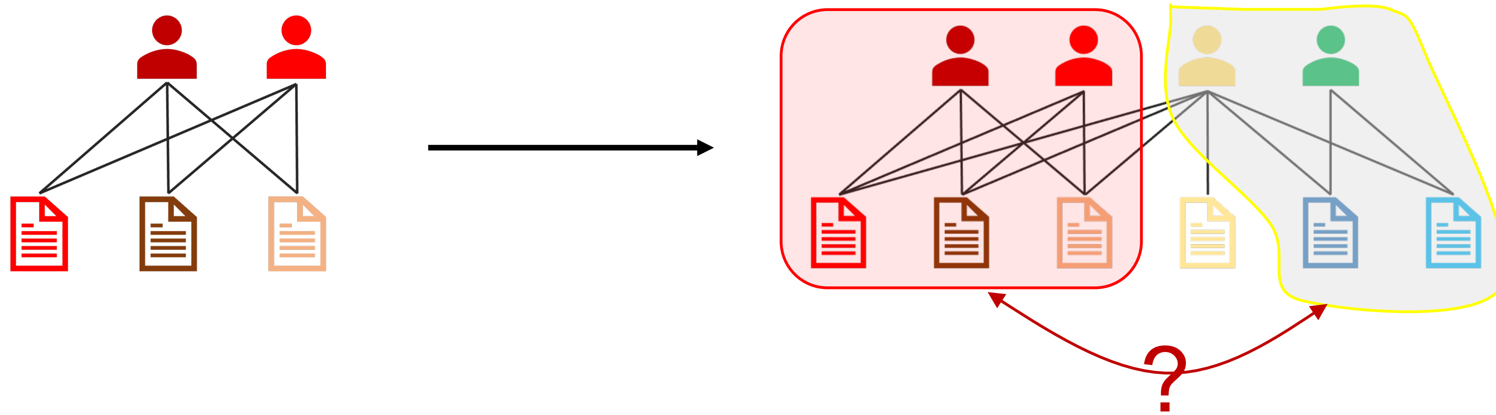
$$\text{maximize}_{l_1, l_2, \cdots} \ [\,?\,(l_1) + \,?\,(l_2) + \cdots + \,?\,(l_k)]$$

subject to each block has $\geq k$ owners (k-strongness)

Thm [WXGS'23]. A simple greedy algorithm outputs a partition that is *simultaneously* a $c(w) = \inf_{l \geq 2} \frac{w(l)}{l \cdot w'(l)}$ approximation for every convex w

- When $w(l) = l^{\alpha} \to c(w) = 1/\alpha$, and this ratio is tight for every monomial
- The algorithm simply greedily pick the largest next block

Step 2: Generalizing to Overlapping Ownership



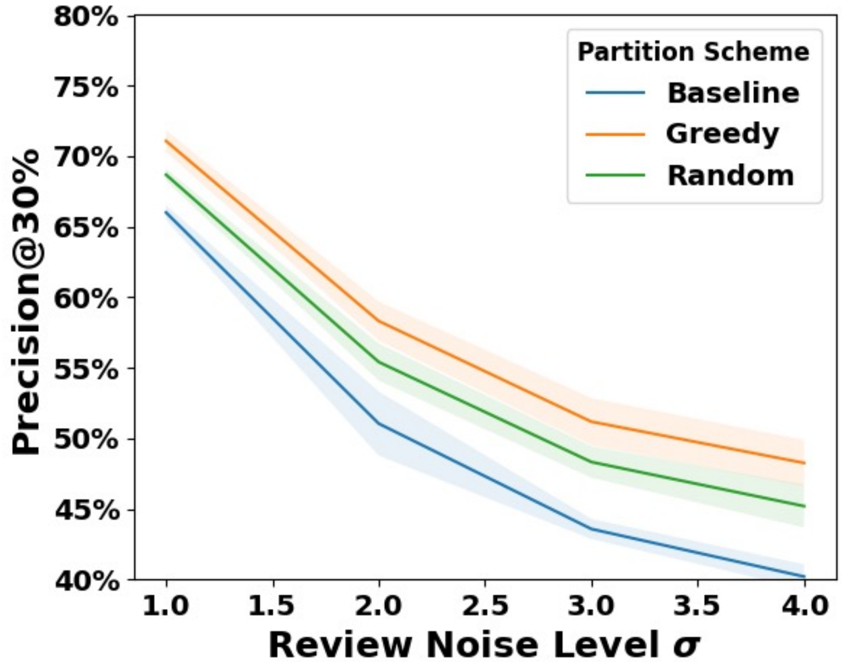
A potential criticism: partition gives up rankings for papers across partitions

- ➢ Indeed, but we show that any truthful mechanism has to be partition-based

There is fundamental tradeoff between
incentive constraints *vs* statistic efficiency

Empirical Evaluation

- ICLR 2021–2023 dataset with review score $\boldsymbol{y}$ and authorship graph
- Synthesized component: group-truth score, simulated as $\boldsymbol{R} = \boldsymbol{y} + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(0, \sigma)$



Precision on acceptance (top 30%)

Outline

A timely real-world problem

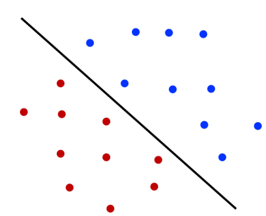
A well-studied classic model

Vignette 1

Elicit truthful information to improve statistical estimation

Vignette 2

PAC-Learning in strategic environments



Outline

Joint work with

Ravi Sundaram
(Northeastern, CS)

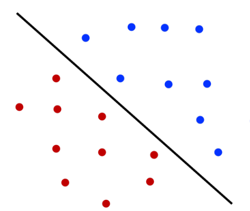
Anil Vullikanti
(UVA, CS)

Fan Yao
(UChicago, CS)

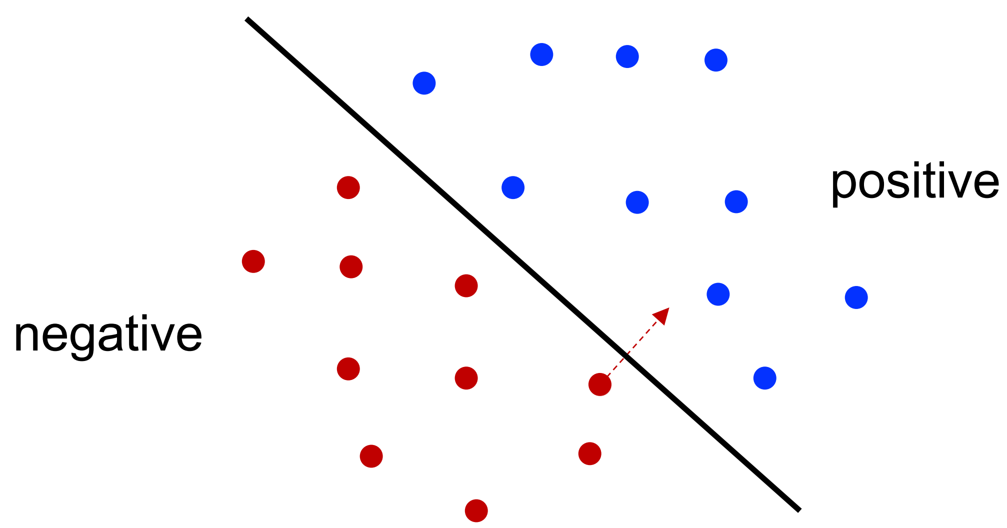
A well-studied classic model

Vignette 2

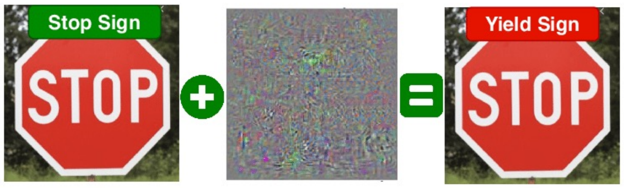
PAC-Learning in strategic environments



Classification

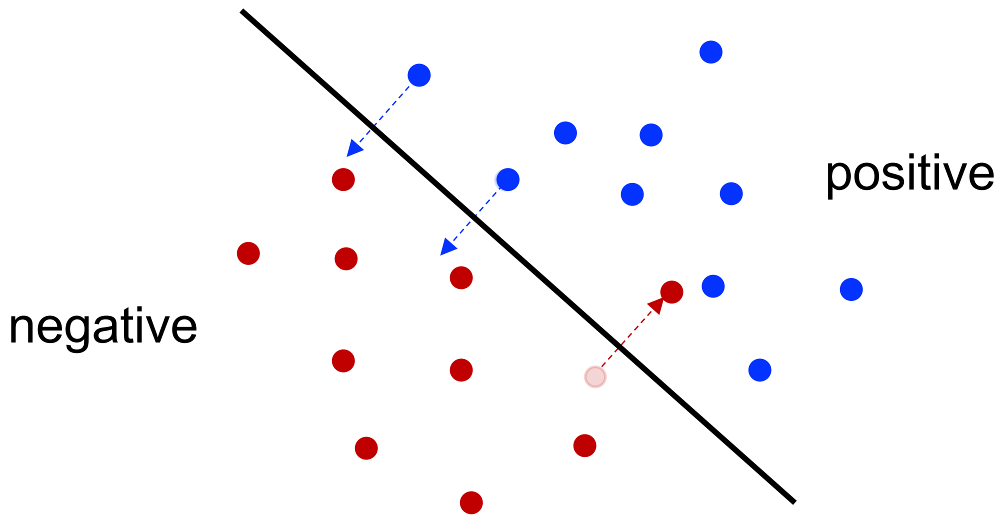


Data points' features may be manipulated

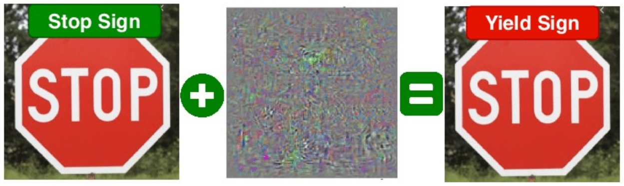


Adversarial attack

[Goodfellow et al.’15]
[Eykholt et al.’18]
[Cullina et al.’18]
……



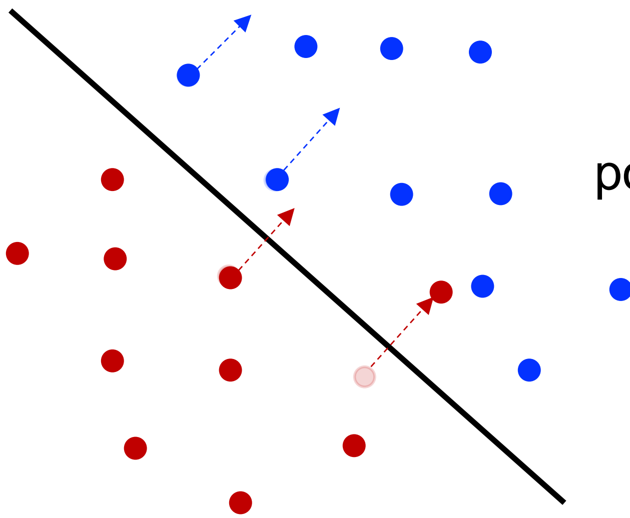
Data points’ features may be manipulated



Adversarial attack

Strategic manipulation

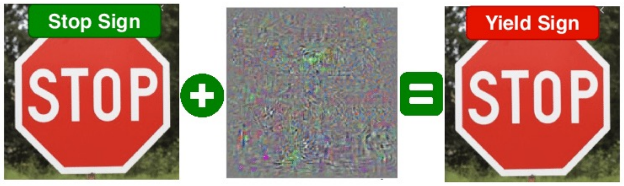
[Hardt et al.’16]
[Hu et al.’19]
[Ghalme et al.’21]
……



positive

negative

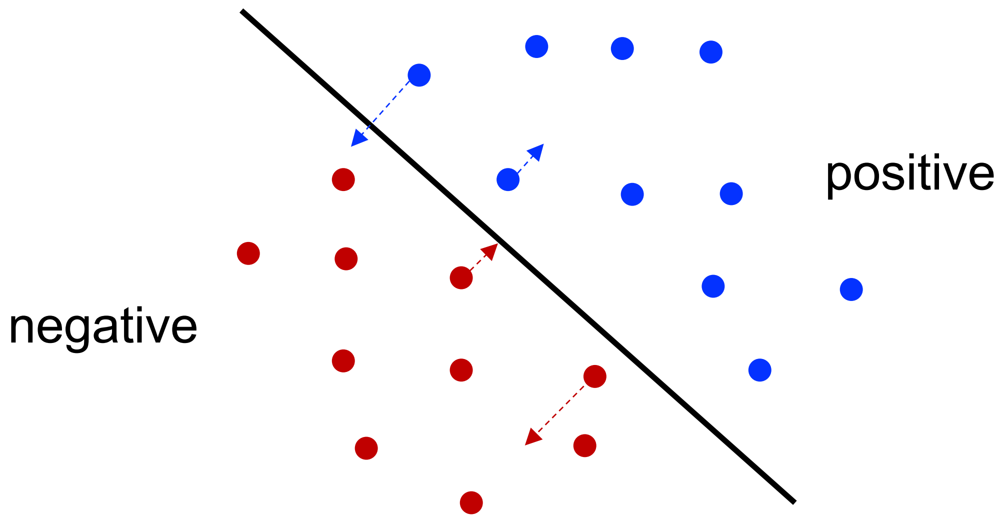
Data points’ features may be manipulated



Adversarial attack

[SVXY, JMLR’23]

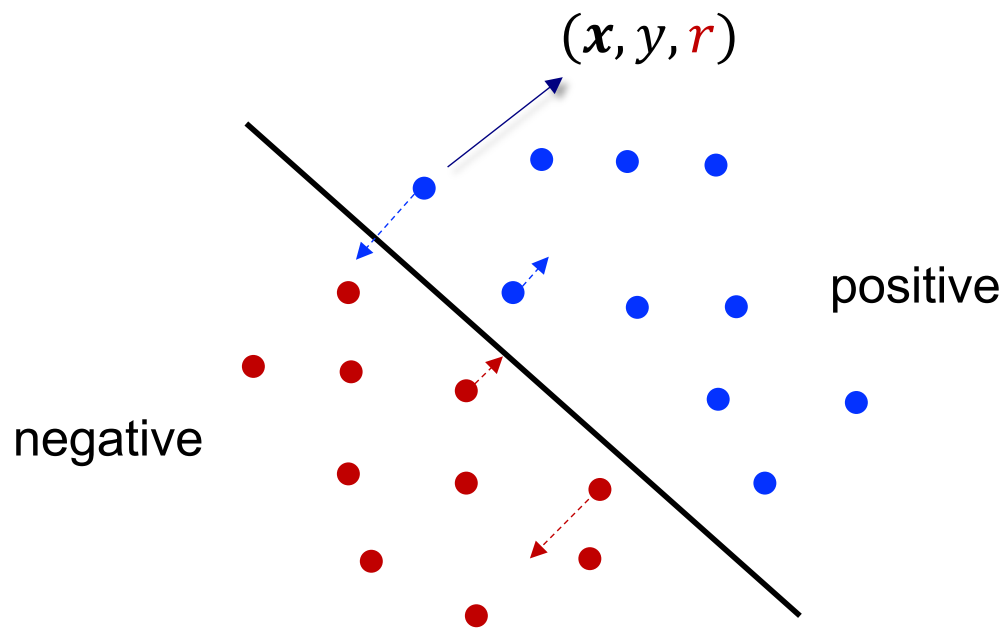
Strategic manipulation



Data points’ features may be manipulated

A Unified Model of Strategic Classification

- Each data point is an economic agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive

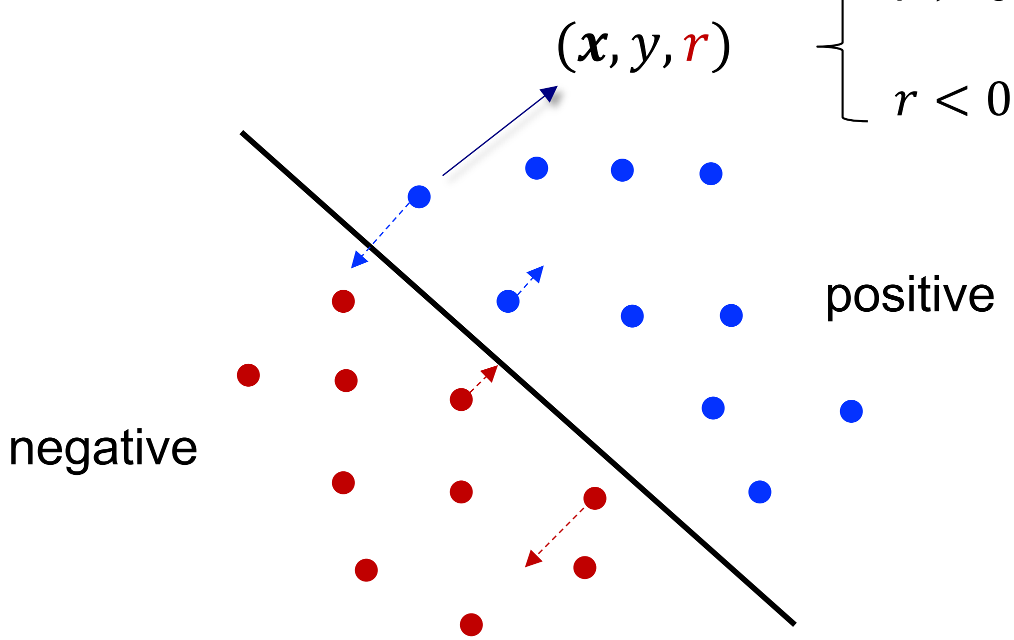


PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

- Each data point is an economic agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive

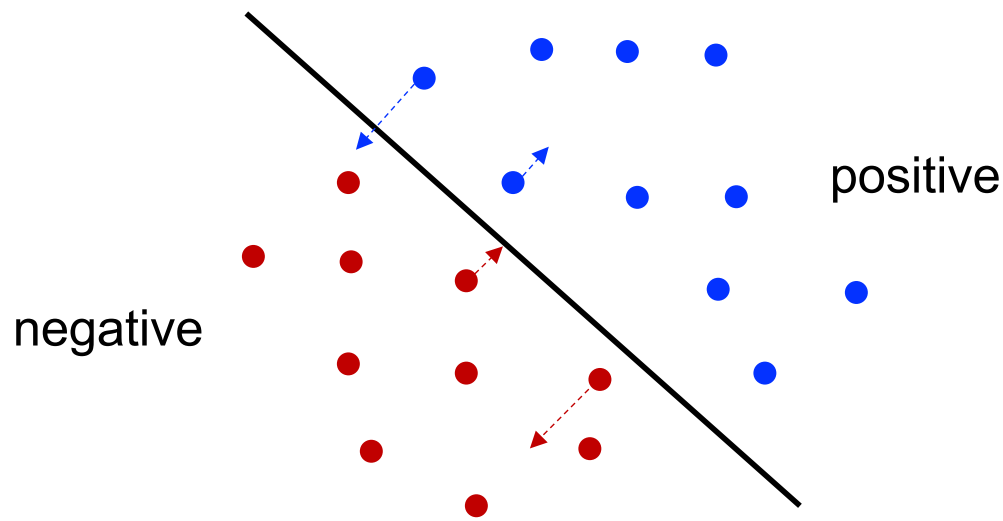
$$\begin{cases} r > 0 \rightarrow \text{prefers positive} \\ r < 0 \rightarrow \text{prefers negative} \end{cases}$$



PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

- Each data point is an economic agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive
- Manipulating feature from $\boldsymbol{x}$ to $\boldsymbol{z}$ incurs cost $c(\boldsymbol{x} - \boldsymbol{z})$
  - c is an arbitrary semi-norm



PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

- Each data point is an economic agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive
- Manipulating feature from $\boldsymbol{x}$ to $\boldsymbol{z}$ incurs cost $c(\boldsymbol{x} - \boldsymbol{z})$
  - c is an arbitrary semi-norm
- Given classifier $f: X \rightarrow \{0, 1\}$, data point $(\boldsymbol{x}, y, r)$ will manipulate its feature to $\boldsymbol{z}$ that maximizes utility

$$\underbrace{r \cdot \mathbb{I}(f(\boldsymbol{z}) = 1)}_{\text{reward from classification outcome}} - \underbrace{c(\boldsymbol{x} - \boldsymbol{z})}_{\text{Manipulation cost}}$$

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

- Each data point is an economic agent, represented by $(\boldsymbol{x}, y, r)$
  - $r \in \mathbb{R}$ capture the point's incentive of being classified as positive
- Manipulating feature from $\boldsymbol{x}$ to $\boldsymbol{z}$ incurs cost $c(\boldsymbol{x} - \boldsymbol{z})$
  - c is an arbitrary semi-norm
- Given classifier $f: X \to \{0, 1\}$, data point $(\boldsymbol{x}, y, r)$ will manipulate its feature to

$$\boldsymbol{z}^*(\boldsymbol{x}, r; f) = \arg\max_{\boldsymbol{z} \in X} \left[r \cdot \mathbb{I}(f(\boldsymbol{z}) = 1) \; - \; c(\boldsymbol{x} - \boldsymbol{z}) \right]$$

This is a game now!

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

SIGMA LAB
Strategic IntelliGence in Machine Agents

A Unified Model of Strategic Classification

General Strategic Classification

Input: n training data points $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n) \sim \mathcal{D}$

Learning goal: compute a classifier f that predicts well based only on the manipulated feature $\boldsymbol{z}^*(\boldsymbol{x}, r; f)$ during testing

$$\boldsymbol{z}^*(\boldsymbol{x}, r; f) = \arg\max_{\boldsymbol{z} \in X} \left[r \cdot \mathbb{I}(f(\boldsymbol{z}) = 1) \;-\; c(\boldsymbol{x} - \boldsymbol{z}) \right]$$

Also called testing time attack

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

General Strategic Classification

Input: n training data points $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n) \sim \mathcal{D}$

Learning goal: compute a classifier f that predicts well based only on the manipulated feature $\boldsymbol{z}^*(\boldsymbol{x}, r; f)$ during testing

Some notably special cases

- ✓ $r = 0$ → classic classification
- ✓ $r = 1$ → strategic classification (cf. [Hardt et al.'16])
- ✓ $r = -y$ → adversarial classification (cf. [Cullina et al.'18])

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

A Unified Model of Strategic Classification

General Strategic Classification

Input: n training data points $(\boldsymbol{x}_1, y_1, r_1), \cdots, (\boldsymbol{x}_n, y_n, r_n) \sim \mathcal{D}$

Learning goal: compute a classifier f that predicts well based only on the manipulated feature $\boldsymbol{z}^*(\boldsymbol{x}, r; f)$ during testing

But will this general problem still be learnable? ✓

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

Recall classic ML setup

✓ Learnability (sample complexity) of a hypothesis class is governed by its VC-dimension

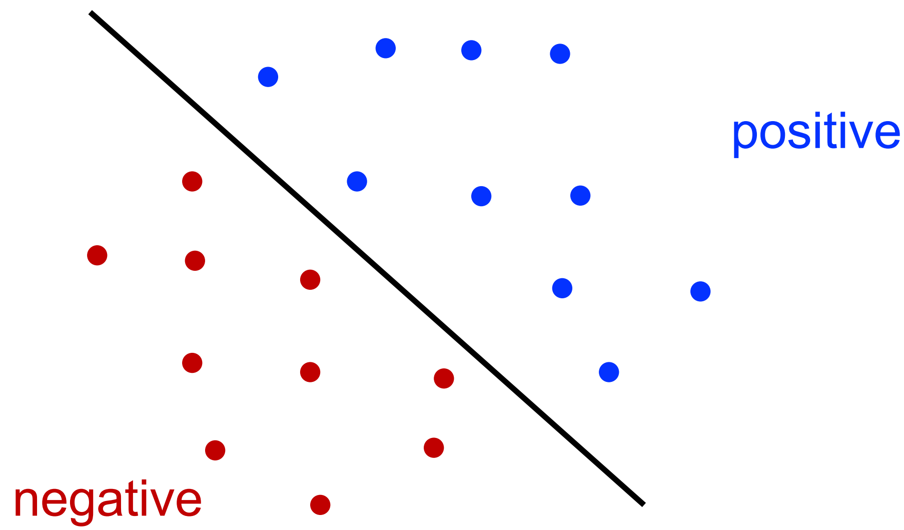
The Learnability of Strategic Classifiers

… is governed by a variant, coined strategic VC-dimension (SVC)

The Learnability of Strategic Classifiers

… is governed by a variant, coined strategic VC-dimension (SVC)

- Defined over the equilibrium of the classification outcome



PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

The Learnability of Strategic Classifiers

… is governed by a variant, coined strategic VC-dimension (SVC)

Theorem. Any strategic classification instance is (PAC) learnable with sample complexity

$$n(\epsilon, \delta) = \Theta(\frac{SVC + \log(1/\delta)}{\epsilon^2})$$

where ϵ is accuracy loss and δ is the failure probability.

1. Unifies learnability of all previous special cases

- Generalizes the fundamental theorem of classic PAC learning ($r = 0$)
- Recovers a few major learnability results in recent literature
  - Sample complexity of [Hardt et al.'16] follows from their $SVC = 3$
  - Learnability of adversarial classifier [Cullina et al.'18] follows by $r = -y$

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

The Learnability of Strategic Classifiers

… is governed by a variant, coined strategic VC-dimension (SVC)

Theorem. Any strategic classification instance is (PAC) learnable with sample complexity

$$n(\epsilon, \delta) = \Theta\left(\frac{SVC + \log(1/\delta)}{\epsilon^2}\right)$$

where ϵ is accuracy loss and δ is the failure probability.

2. Implies learnability of new setups with heterogeneous data preferences

30-Year Fixed	*V.S*	15-Year Fixed

Classify the approval to different loan types

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

Instantiation to Linear Classification

Theorem. The SVC of d-dimensional linear classifiers is at most $d + 1$.

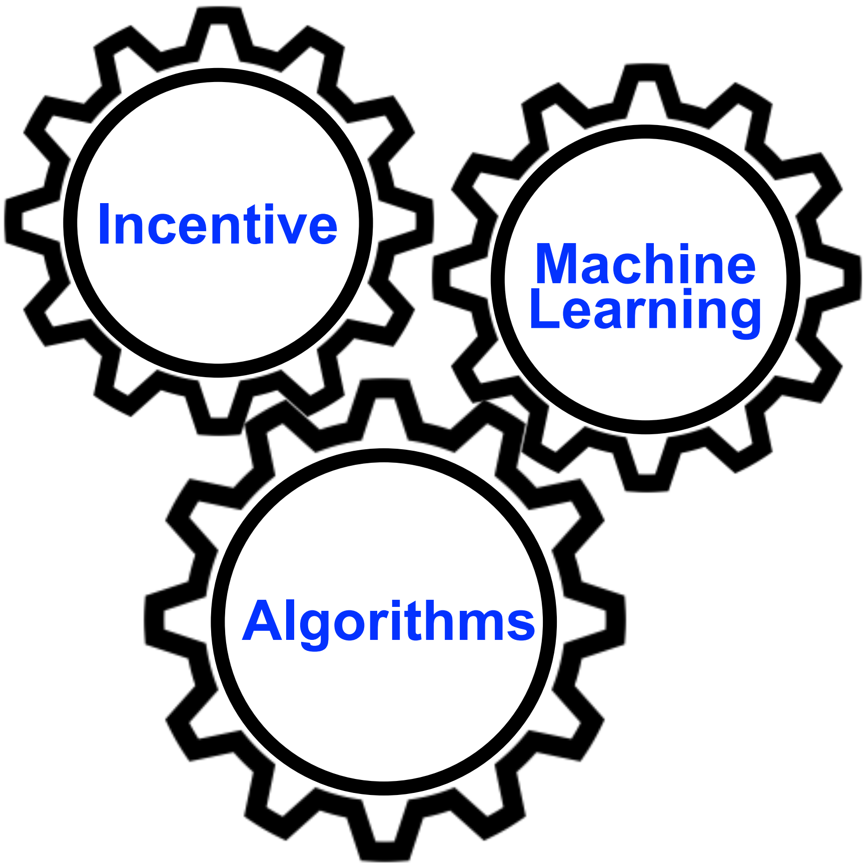
- $d + 1$ is the VC of linear classifiers in classic setup
- Learning strategic linear classifiers is no harder statistically

However, it is computationally harder

Theorem. Empirical risk minimization for strategic linear classification is NP-hard.

PAC-Learning for Strategic Classification. Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao. Journal of Machine Learning Research, 2023

Summary



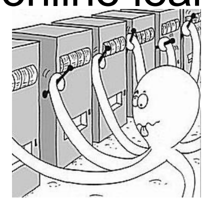
in both foundational models and pressing real-world problems

Summary

Incentives in recommendation policy design

Learning to play against adversaries

Strategic behaviors in online learning



Vignette 1

Elicit truthful information to improve statistical estimation

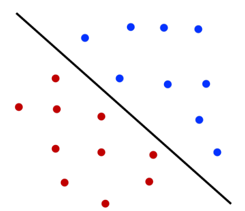
ICML
International Conference On Machine Learning

NEURAL INFORMATION PROCESSING SYSTEMS

REJECTED

Vignette 2

PAC-Learning in strategic environments



SIGMA LAB
Strategic IntelliGence in Machine Agents

Thank You

Questions?

haifengxu@uchicago.edu

National Science Foundation

Army Research Lab

Google Research

SIGMA LAB
Strategic IntelliGence in Machine Agents