Robust Spammer Detection by Nash Reinforcement Learning

Yingtong Dou (UIC) **Guixiang Ma (Intel Labs)**

Philip S. Yu (UIC) **Sihong Xie (Lehigh)**

ydou5@uic.edu

Paper: http://arxiv.org/abs/2006.06069
Slides: http://ytongdou.com/files/kdd20slides.pdf
Code: https://github.com/YingtongDou/Nash-Detect

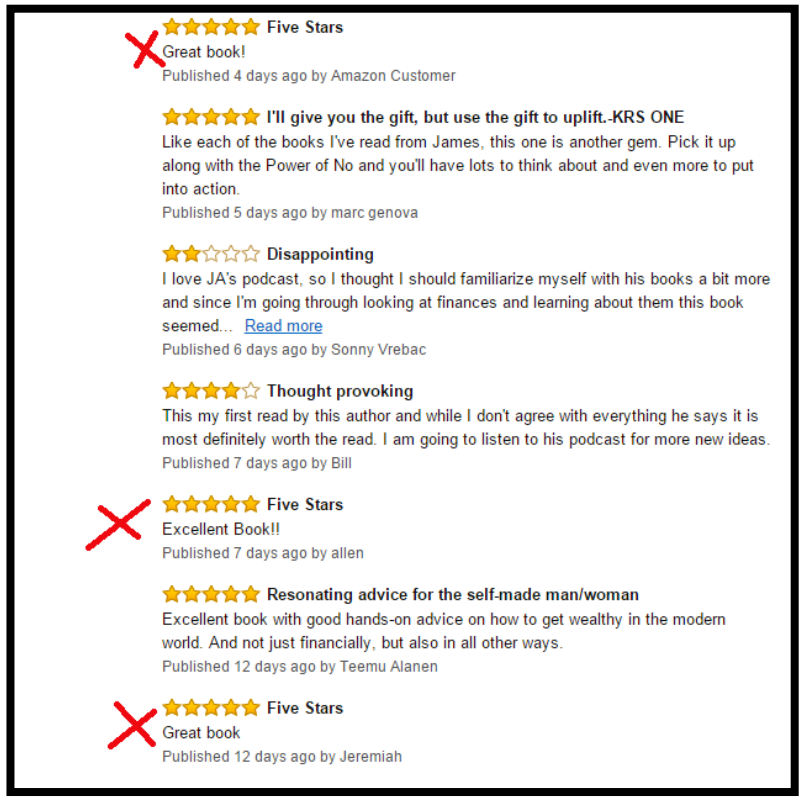
ACM SIGKDD' 20, August 23-27th, Virtual Event, CA, USA

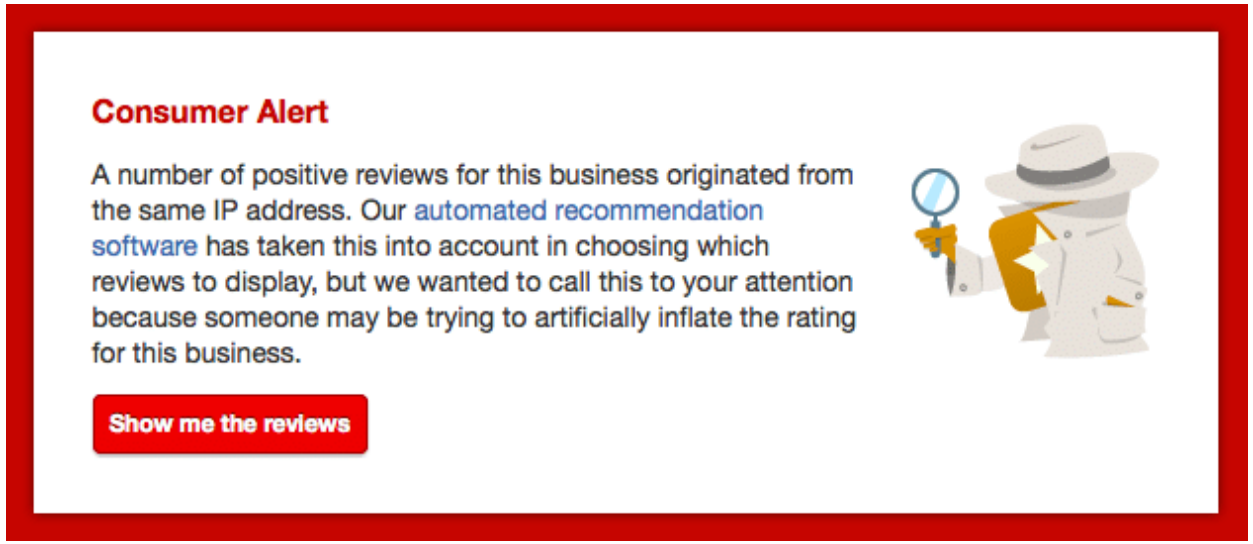
Outline

- **Background:** review spam and spamming campaign
- **Highlight:** previous works vs. our works
- **Methodology I:** practical goals of spammers and defenders
- **Methodology II:** robust training of spam detectors (Nash-Detect)
- **Experiments:** the training and deployment performance of Nash-Detect
- **Conclusion & Future Works**

Fake Reviews are Prevalent

- Near **40%** reviews in Amazon are fake[1]
- Yelp hide suspicious reviews and alert consumers





[1] J. Swearingen. 2017. Amazon Is Filled With Sketchy Reviews. Here's How to Spot Them. https://slct.al/2TBXDpT

Images from https://upserve.com/restaurant-insider/five-key-reasons-shouldnt-buy-yelp-reviews/ http://greyenlightenment.com/detecting-fake-amazon-reviews/

Spamming Campaign

- Dishonest merchants can **easily** buy high-quality fake reviews online
- Machine-generated fake reviews are very **authentic-like**[1]

Generated Reviews (Yelp)
I love this place ! I 've been here several times and I 've never been disappointed . The food is always fresh and delicious . The service is always friendly and attentive . I 've been here several times and have never been disappointed .
I 've been to this location twice now and both times I 've been very impressed . I 've tried their specialty pizzas and they 're all really good . The only problem is that they 're not open on sundays . They 're not open on sundays .
I have been coming to this place for years and have always had great food and service . They have a great lunch buffet . They have a great selection of food for the price . They do have a lot of seating and I would recommend reservations .
I 've eaten here about 8 times . I 've been introduced to this place . Its always busy and their food is consistently great . I LOVE their food , hence the name . It is so clean , the staff is so friendly , and the food is great . I especially like the chicken pad thai , volcano roll , and the yellow curry .
this is strictly to go . Love , love , love the food ! we usually usually get brisket (oh my) , sandwich (pastrami , or pork , just so good) and now these are my two favorites . It 's great . This is gone (according to our waitress) .

[1] P. Kaghazgaran, M. Alfifi, and J. Caverlee. 2019. Wide-Ranging Review Manipulation Attacks: Model, Empirical Study, and Countermeasures. In CIKM.

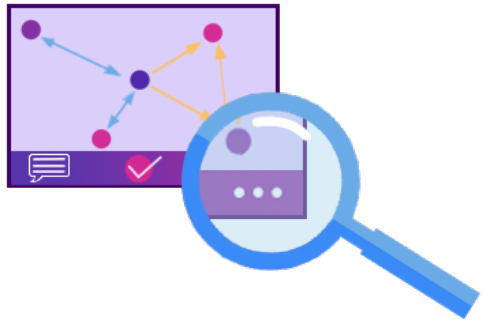
Images from https://mopeak.com/buy-android-reviews/
http://faculty.cs.tamu.edu/caverlee/pubs/kaghazgaran19cikm.pdf

Review Spam Detection

- To detect fake reviews, three major types of spam detectors have been proposed

Text-based Detectors

Behavior-based Detectors

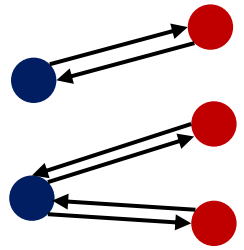


Graph-based Detectors

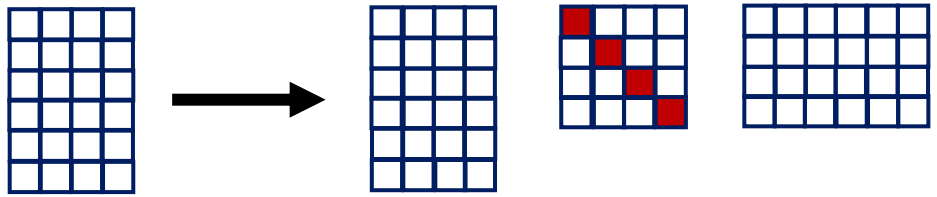
Base Spam Detectors

- **GANG**
- **SpEagle**

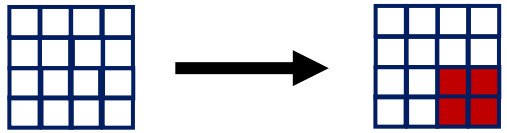
} MRF-based detector



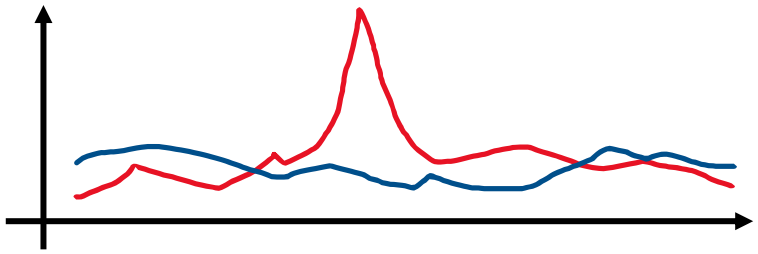
- **fBox** SVD-based detector



- **Fraudar** Dense-block-based detector



- **Prior** Behavior-based detector



Previous Works vs. Our Work

- **Previous works:**
  - Static dataset
  - Accuracy-based evaluation metric
  - Fixed spamming pattern
  - Single detector

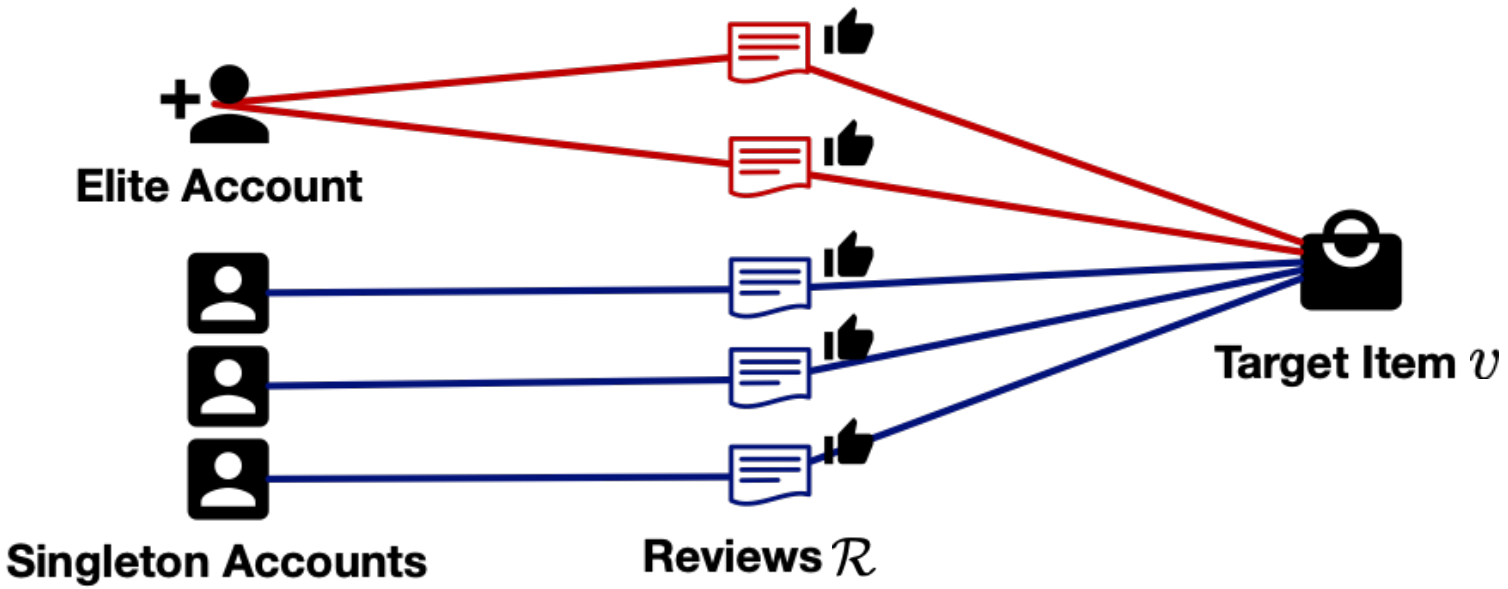
- **Our work:**
  - Dynamic game between spammer and defender
  - Practical evaluation metric
  - Evolving spamming strategies
  - Multiple detectors ensemble

Turning Reviews into Business Revenues

- In Yelp, product's rating is correlated to its revenue[1]

Revenue Estimation & Practical Effect:

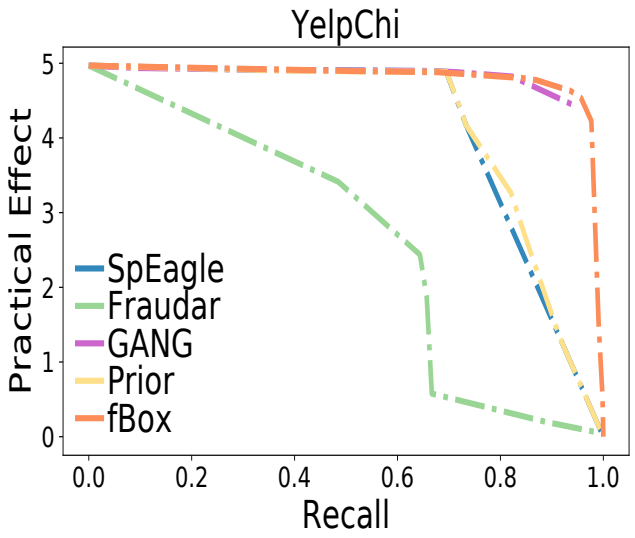
$$f(v;\mathcal{R}) = \beta_0 \times \boxed{\mathrm{RI}(v;\mathcal{R})} + \beta_1 \times \boxed{\mathrm{ERI}(v;\mathcal{R}_E(v))} + \alpha$$

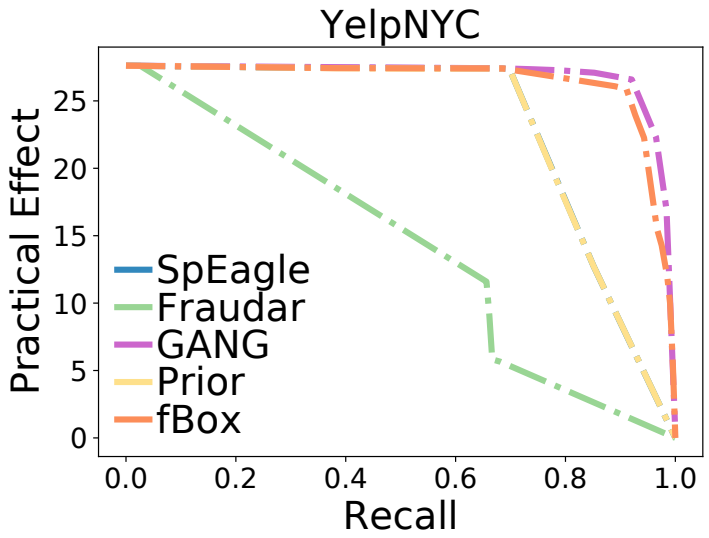


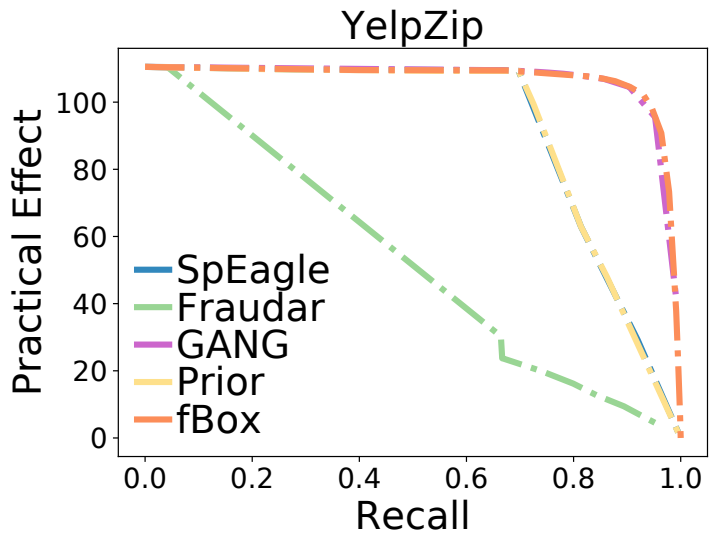
[1] M. Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com. HBS Working Paper (2016).

Practical Effect is Better than Recall

- We run five detectors individually against five attacks
- When detector recalls are **high (>0.7)**, the practical effects are **not reduced**







Spammer's Practical Goal

Spamming Practical Effect: $\mathrm{PE}(v; \mathcal{R}, \mathsf{p}, \mathsf{q}) = f(v; \mathcal{R}(\mathsf{p}, \mathsf{q})) - f(v; \mathcal{R})$

Revenue after attacks

Revenue before attacks

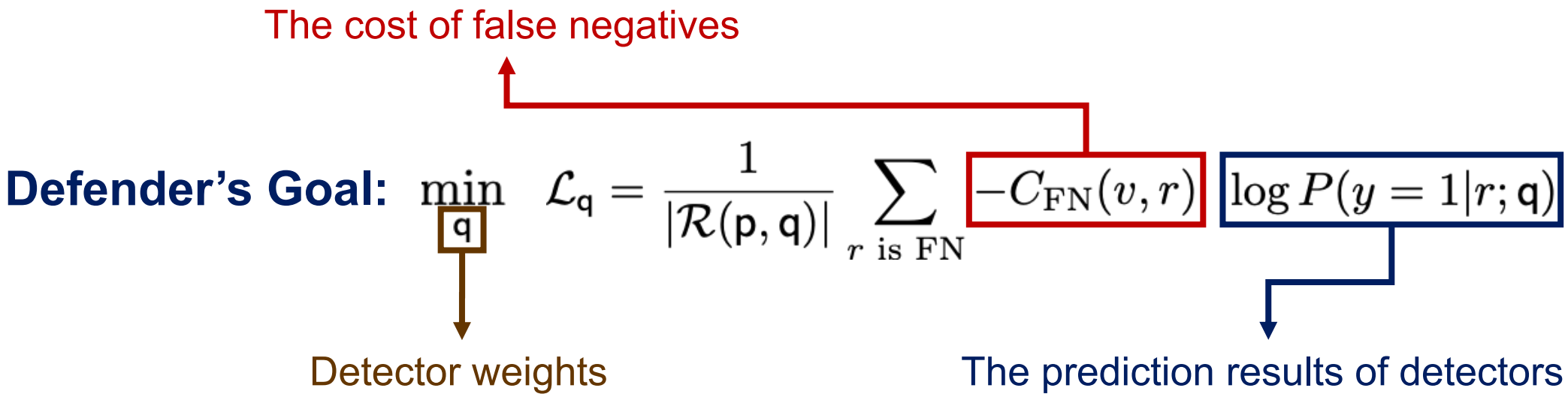
- To promote a product, the practical goal of the spammer is to **maximize** the PE.

Spammer's Goal: $\max_{\mathsf{p}} \max\{0, \mathrm{PE}(v; \mathcal{R}, \mathsf{p}, \mathsf{q}))\}$

Spamming strategy weights

Defender's Practical Goal

- The defender needs to **minimize** the practical effect
- We combine detector prediction results with the practical effect to formulate a **cost-sensitive loss**



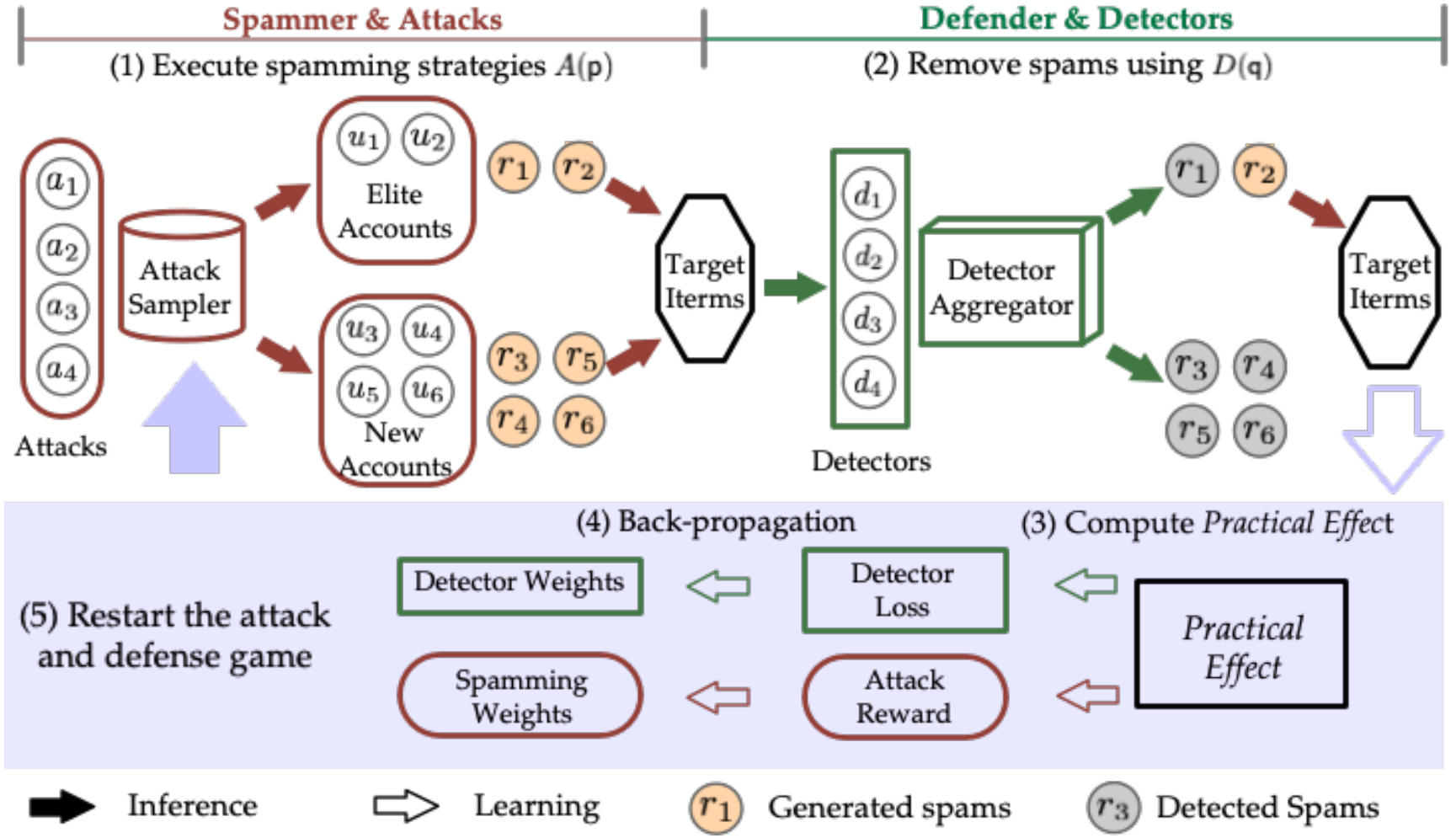
A Minimax-Game Formulation

Minimax Game Objective: $\min_{\mathsf{q}} \max_{\mathsf{p}} \sum_{v \in \mathcal{V}_T} \max\{0, \mathrm{PE}(v; \mathcal{R}, \mathsf{p}, \mathsf{q})\}$

- The objective function is not differentiable

- Our solution: **multi-agent non-cooperative reinforcement learning** and **SGD optimization**

Train a Robust Detector - Nash-Detect



Base Spamming Strategies

- **IncBP:** add reviews with minimum suspiciousness based on belief propagation on MRF
- **IncDS:** add reviews with minimum densities on graph composed of accounts, reviews, and products
- **IncPR:** add reviews with minimum prior suspicious scores computed by behavior features
- **Random:** randomly add reviews
- **Singleton:** add reviews with new accounts

Experimental Settings

- Dataset statistics and spamming attack settings

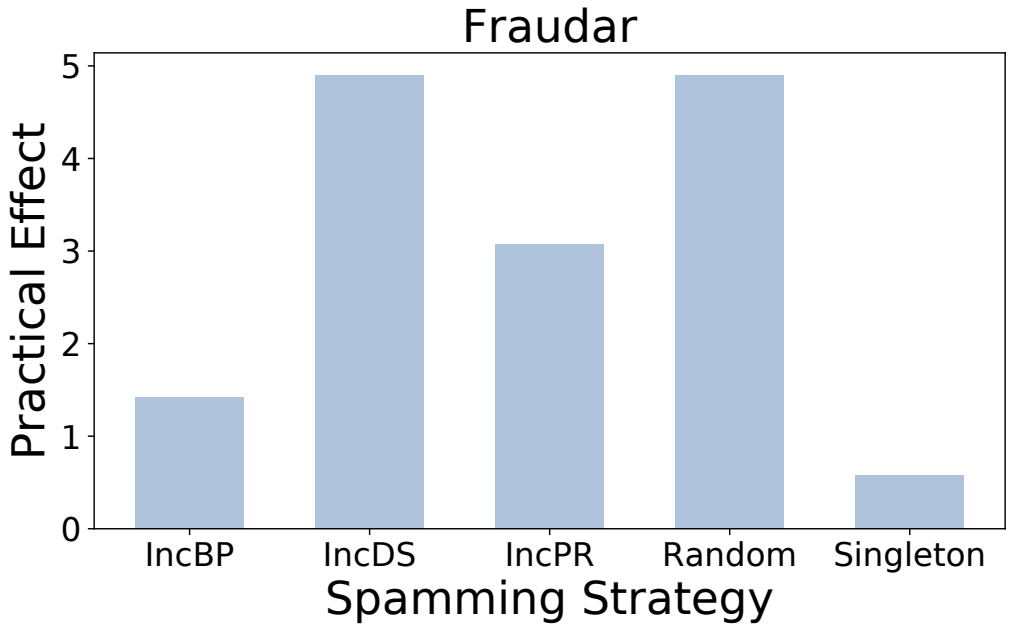
Dataset	# Accounts	# Products	# Reviews	# Controlled elite accounts	# Target products	# Posted fake reviews
YelpChi	38063	201	67395	100	30	450
YelpNYC	160225	923	359052	400	120	1800
YelpZip	260277	5044	608598	700	600	9000

- The spammer controls **elite and new accounts**

- The defender removes **top k** suspicious reviews

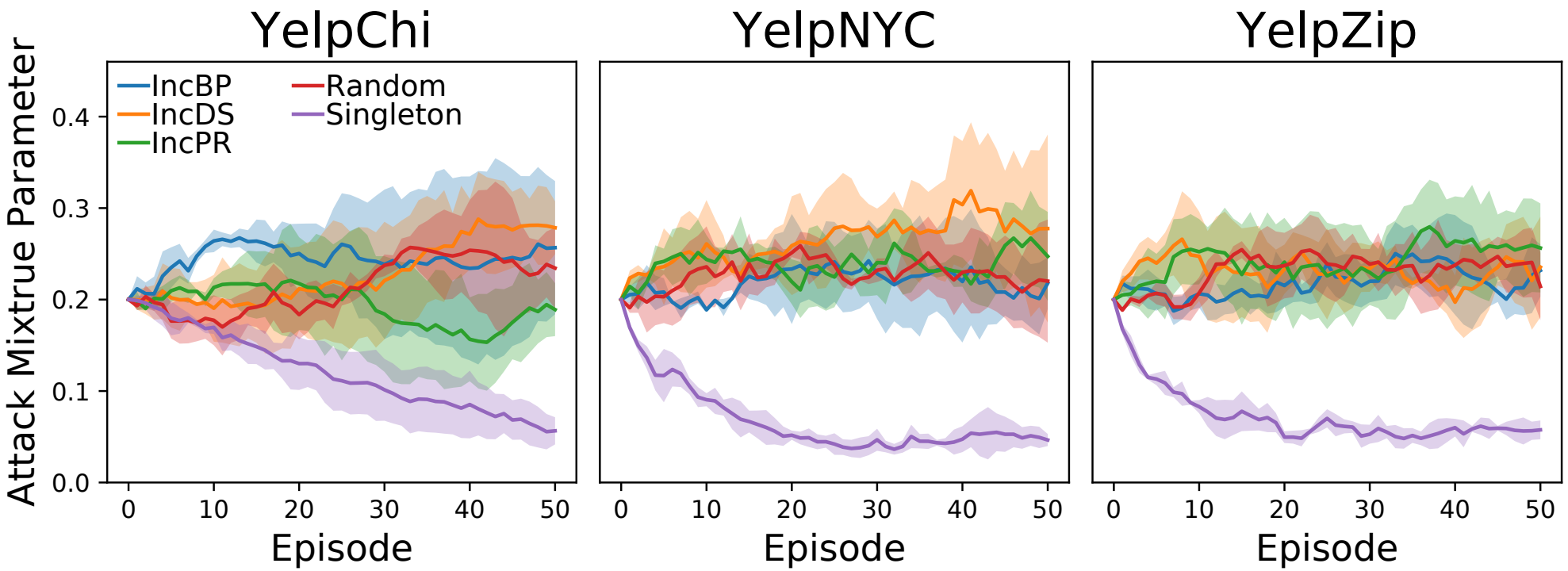
Fixed Detector's Vulnerability

- For a fixed detector (**Fraudar**), the spammer can switch to the spamming strategy with the max practical effect (**IncDS**)



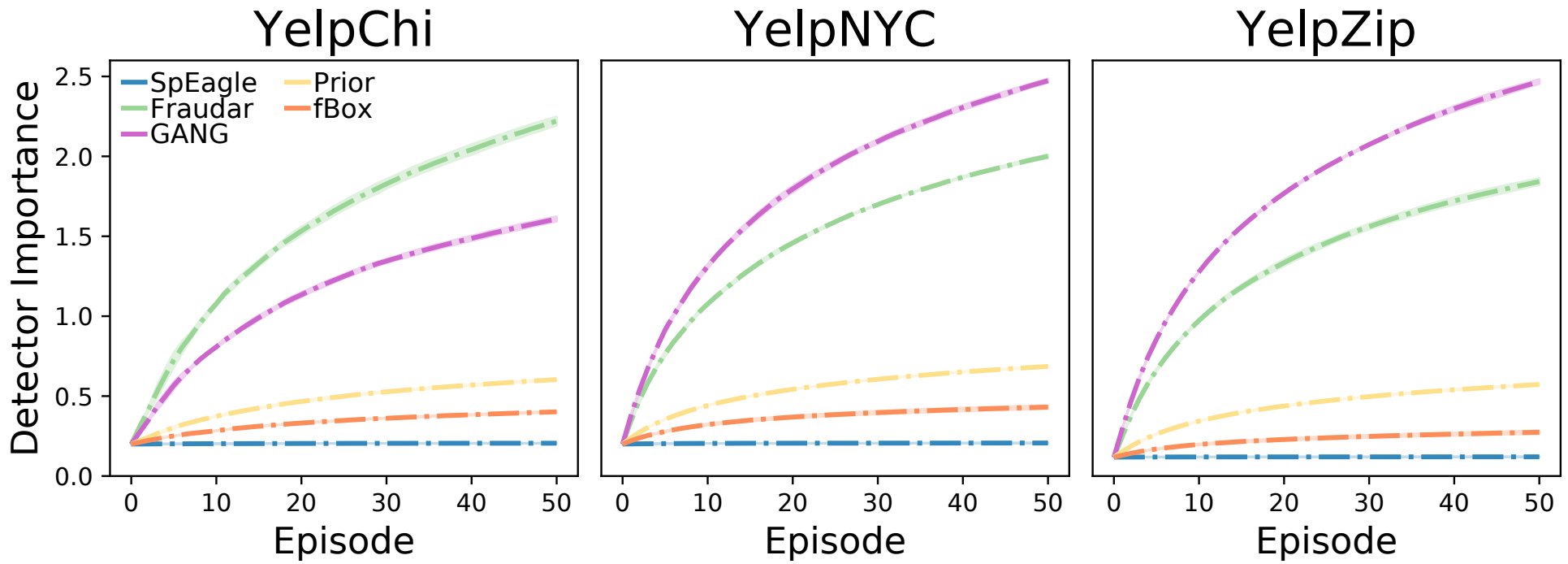
Nash-Detect Training Process

- **Singleton** attack is less effective than other four attacks.



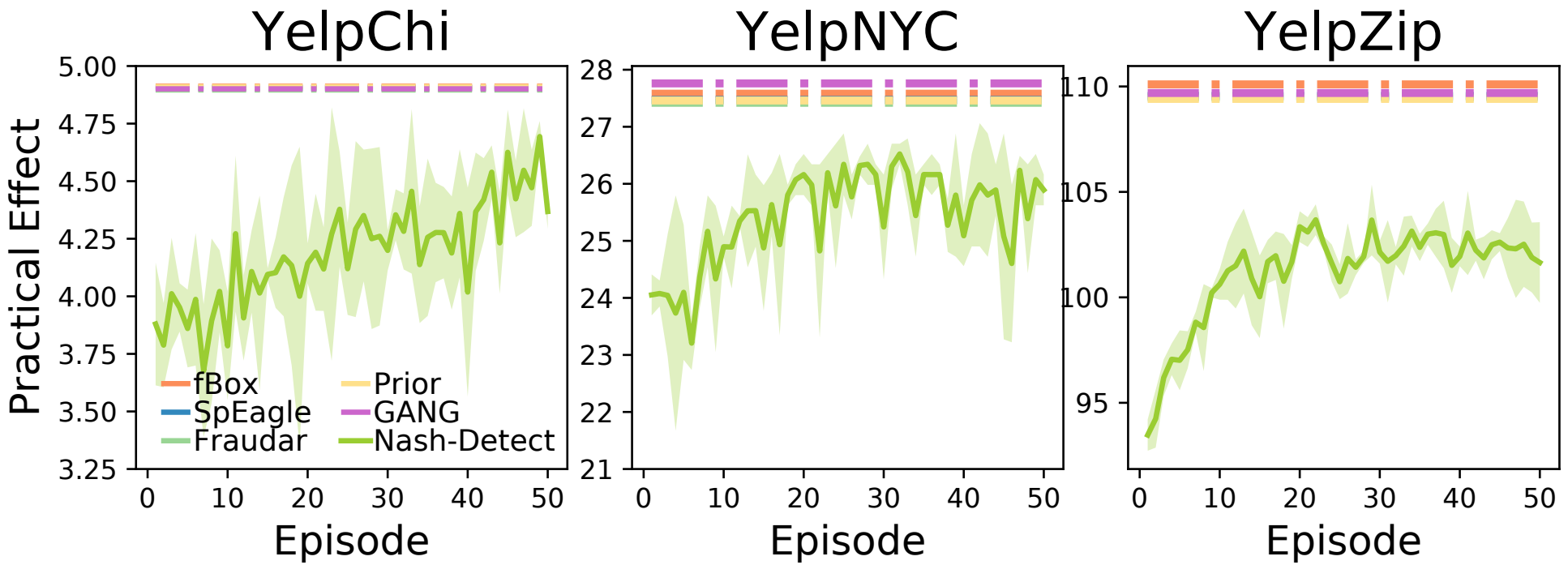
Nash-Detect Training Process

- Nash-Detect can find the optimal detector importance smoothly

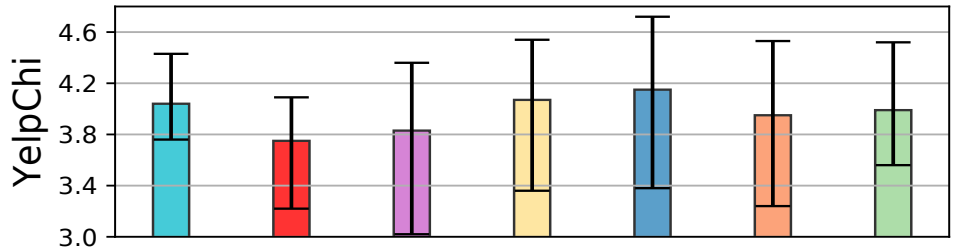


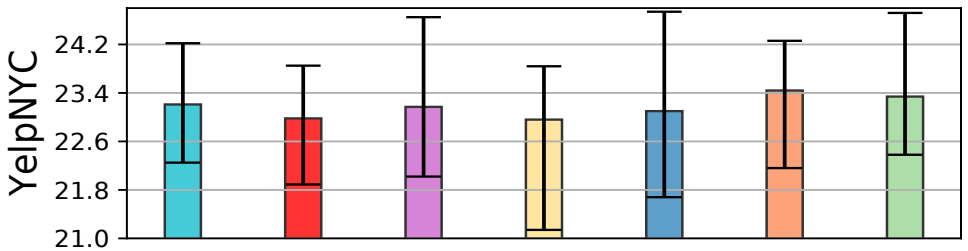
Nash-Detect Training Process

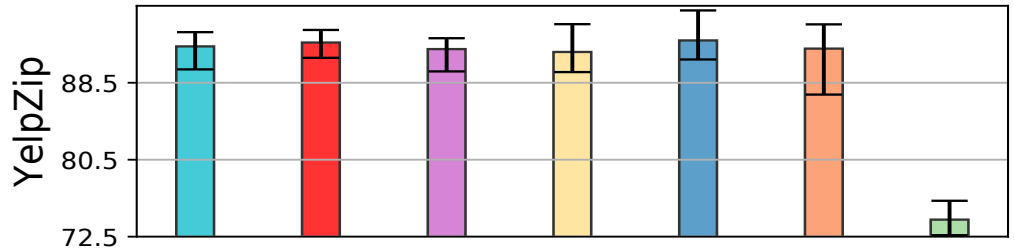
- The practical effect of detectors configured by Nash-Detect are always **less than** the worst-case performances



Nash-Detect Performance in Deployment







Key Takeaways

- New metric
- New spamming strategies
- New adversarial training algorithm

Future Works

- Investigate the attack and defenses of deep learning spam detection methods
- Apply the Nash-Detect framework on other review systems and applications
- Develop advanced attack generation techniques aware of the states of review system

Robust Spammer Detection by Nash Reinforcement Learning

Yingtong Dou (UIC) **Guixiang Ma (Intel Labs)**

Philip S. Yu (UIC) **Sihong Xie (Lehigh)**

ydou5@uic.edu

Paper: http://arxiv.org/abs/2006.06069
Slides: http://ytongdou.com/files/kdd20slides.pdf
Code: https://github.com/YingtongDou/Nash-Detect

ACM SIGKDD' 20, August 23-27th, Virtual Event, CA, USA