Misinformation detection for e-commerce

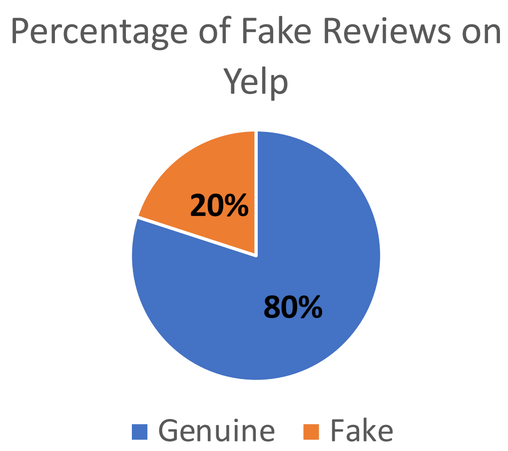
Sihong Xie, Assistant Professor

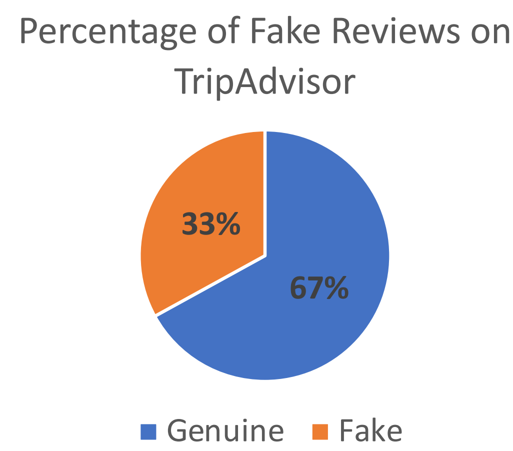
Computer Science and Engineering

Lehigh University

Misinformation are prevalent

Estimated percentage of fake reviews on popular e-commerce websites.

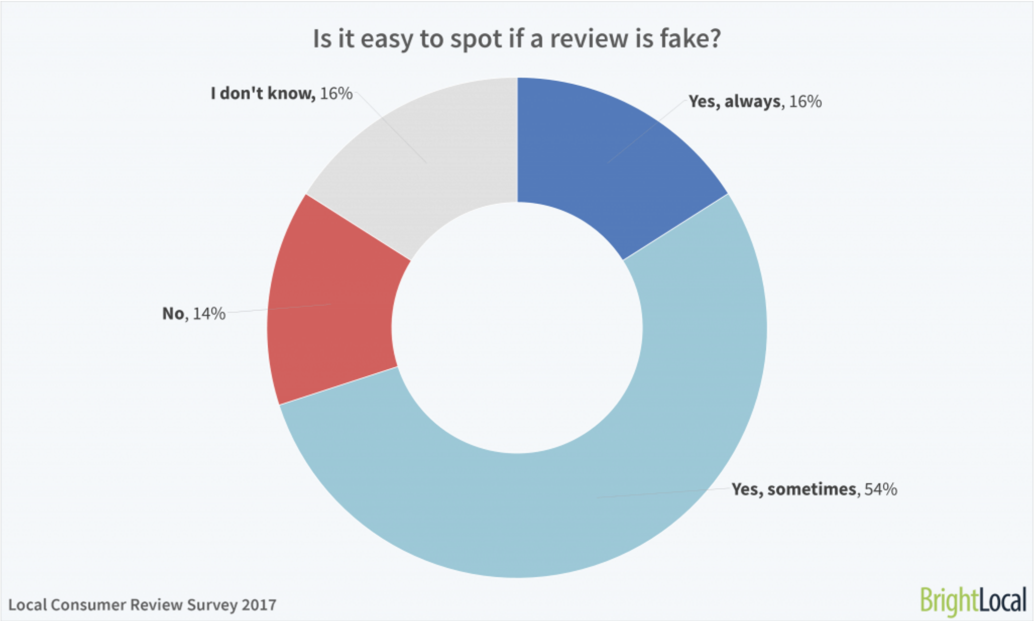




Source: BusinessInsider, ChicagoTribute

Misinformation are hard to spot

Based on a 2017 pool of representative 1,031 US-based consumers

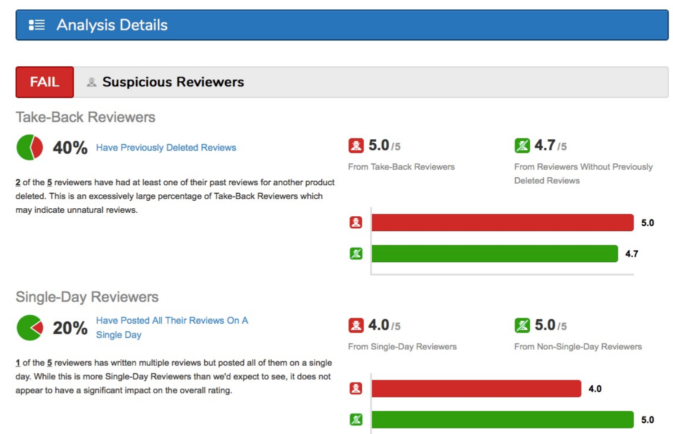


Source: https://www.brightlocal.com/learn/local-consumer-review-survey/

Existing efforts

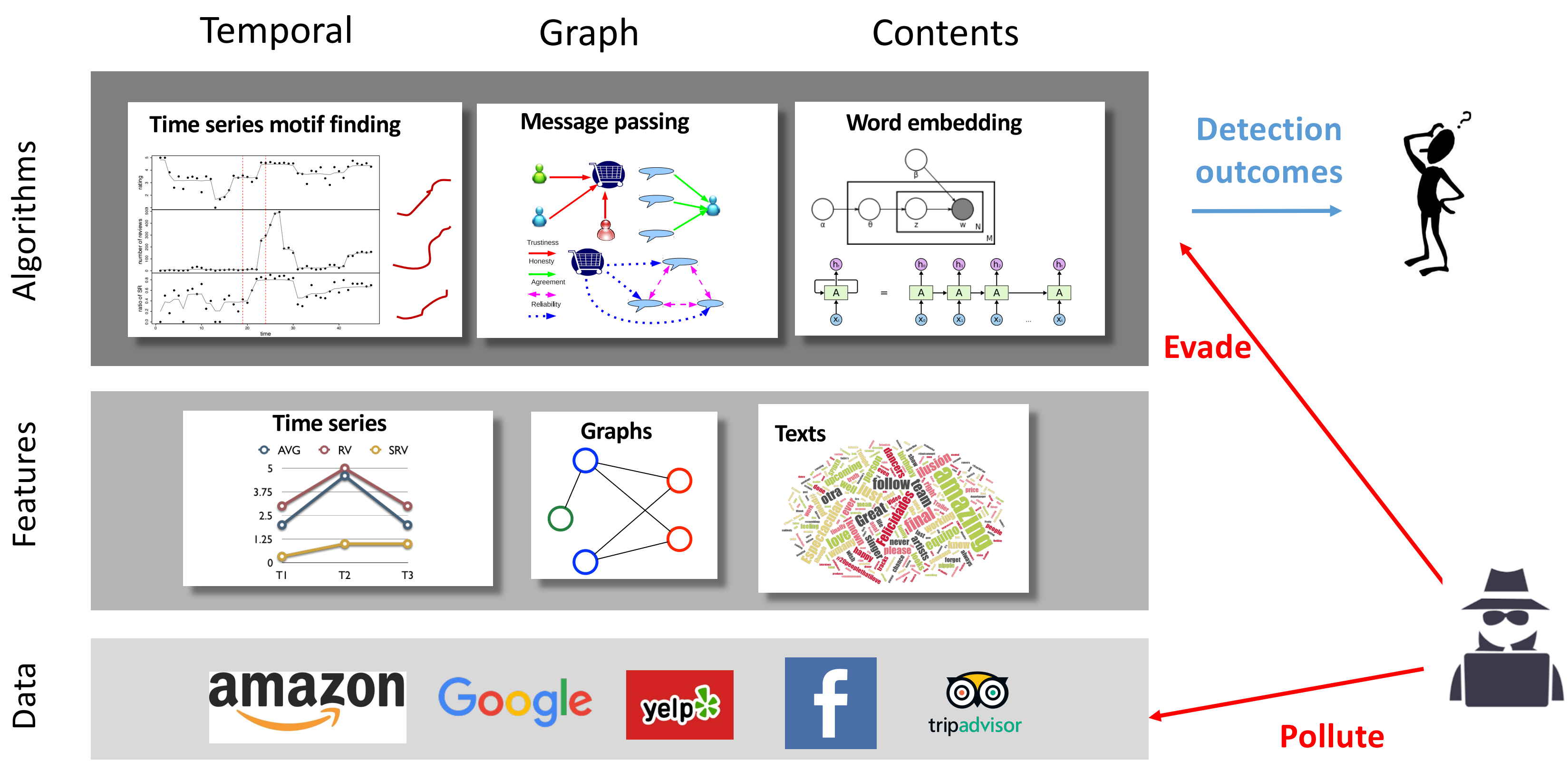
ReviewMeta.com

1. Feature engineering
2. Detection models

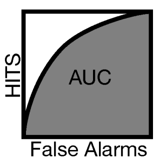


Pros	Rely on yourselves	Convenient and easy	More in-depth automatic analysis
Cons	Not everyone can spot fake reviews	Can be gamed	Can be gamed

Misinformation detection architecture



Overview



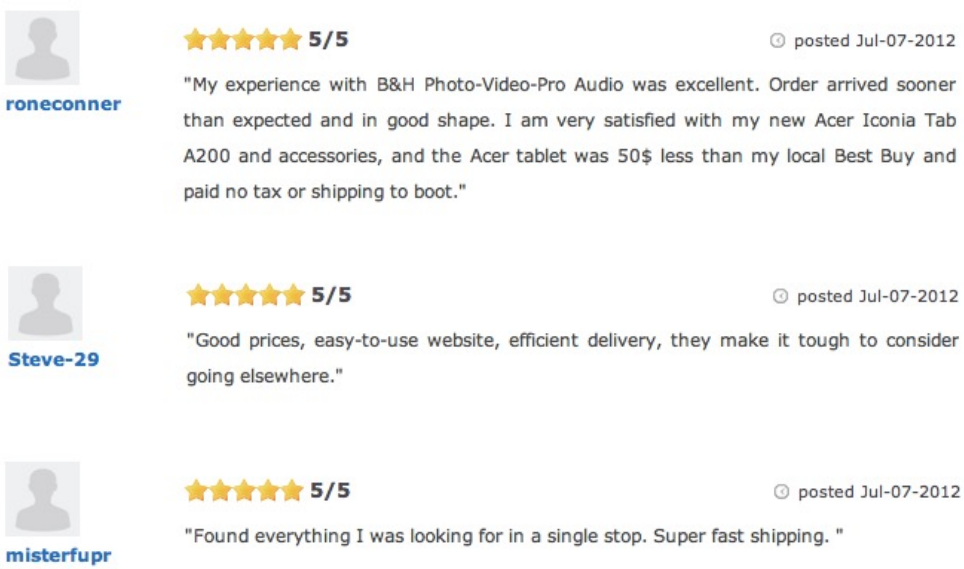
	Temporal	Graph	Contents
Fairness Interpretability		ICDM 2019, 2021 CIKM 2020	
Security	BigData 2018	KDD 2020	
Accuracy	KDD 2012	ICDM 2011	DSAA 2015

Review Spam Detection via Temporal Pattern Discovery

Sihong Xie, Guan Wang, Shuyang Lin, Philip S. Yu

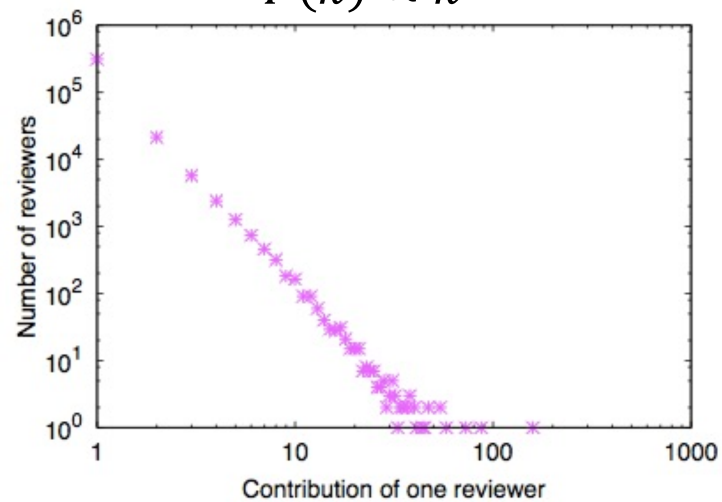
Why detection is so hard

Each account posts just one review.
Can you spot the fake ones?



Number of accounts posting a number of reviews follows a power law distribution.

$$P(k) \propto k^{-\gamma}$$



Exploiting an invariance of spamming

Invariance:

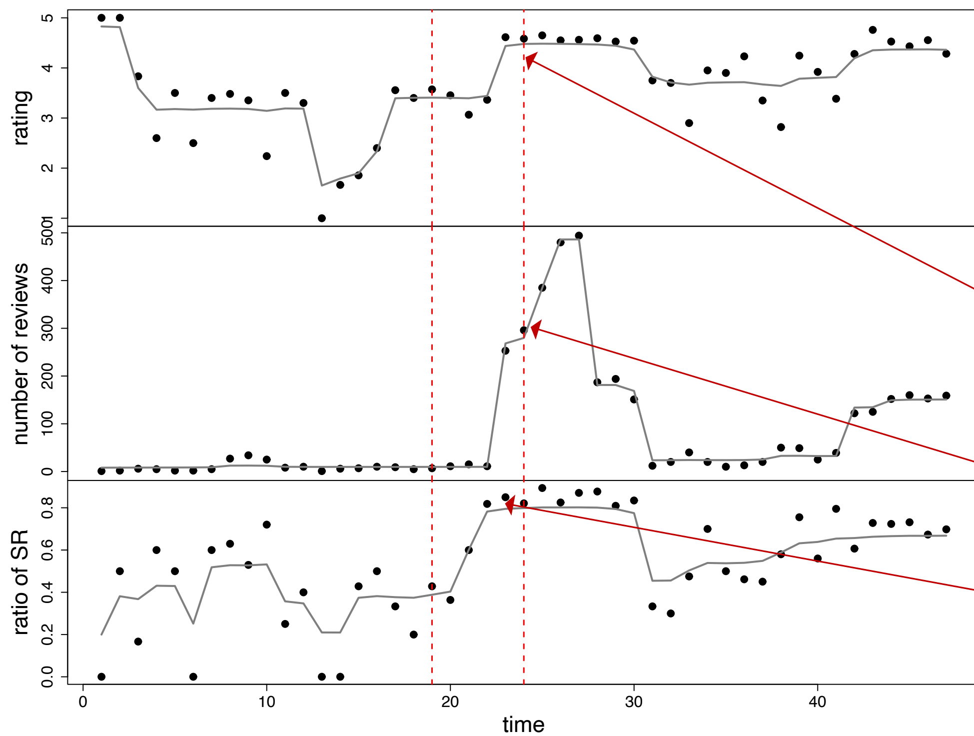
to manipulate ratings, a **large number** of **consistent ratings** must be posted **in a short time**.

Temporal features for each window:

AVG rating

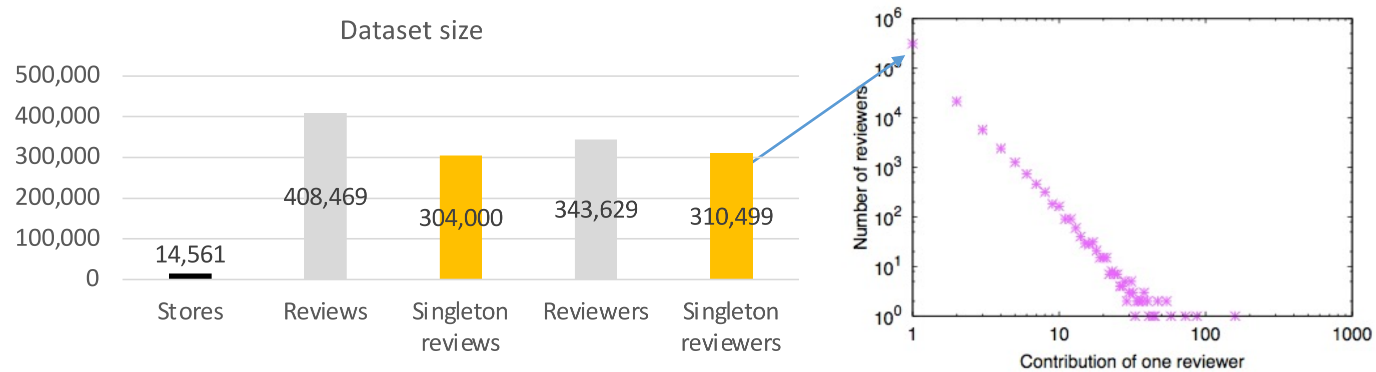
Review Volume

Singleton Review Volume



Burst motif detection on all 3 series.

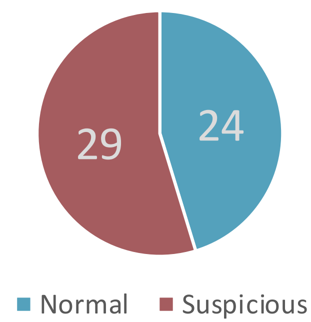
Results



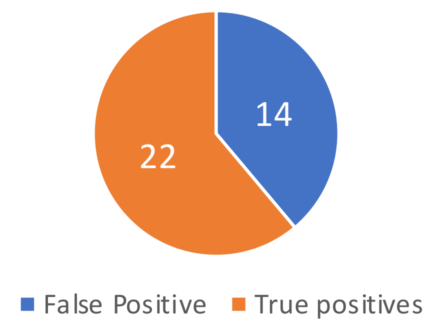
Manual labeling of dishonest businesses

- Hard to evaluate the recall rate.
- Only label the top 53 stores with most reviews.
- Humans background-checked stores on Google and BBB.

	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1	17	14	16
Evaluator 2	-	20	19
Evaluator 3	-	-	24

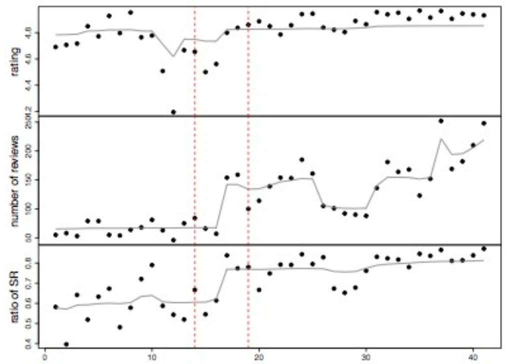


Burst detector Performance



Case study

A detected 15-day window



AVG rating: 4.4 → 4.79

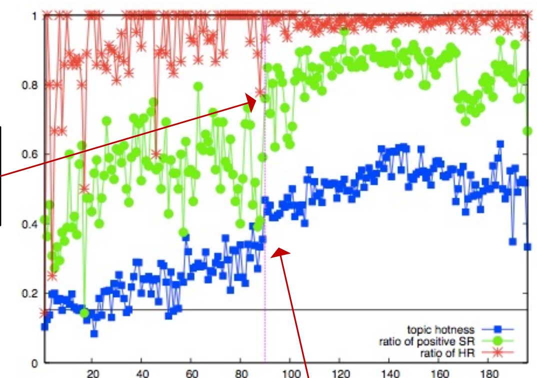
Volume of reviews: 57 → 154

Ratio of singleton reviews: 61% → 83%

More evidences

Burst of "hurry reviewers"

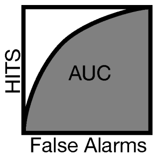
Burst of positive singleton reviews



Burst of reviews with key phrase "customer service"

Overview of my research

	Temporal	Graph	Contents
Fairness Interpretability		ICDM 2019, 2021 CIKM 2020	
Security	BigData 2018	KDD 2020	
Accuracy	KDD 2012	ICDM 2011	DSAA 2015



Securing Behavior-based Opinion Spam Detection
Shuaijun Ge, Guixiang Ma, **Sihong Xie**, and Philip S. Yu

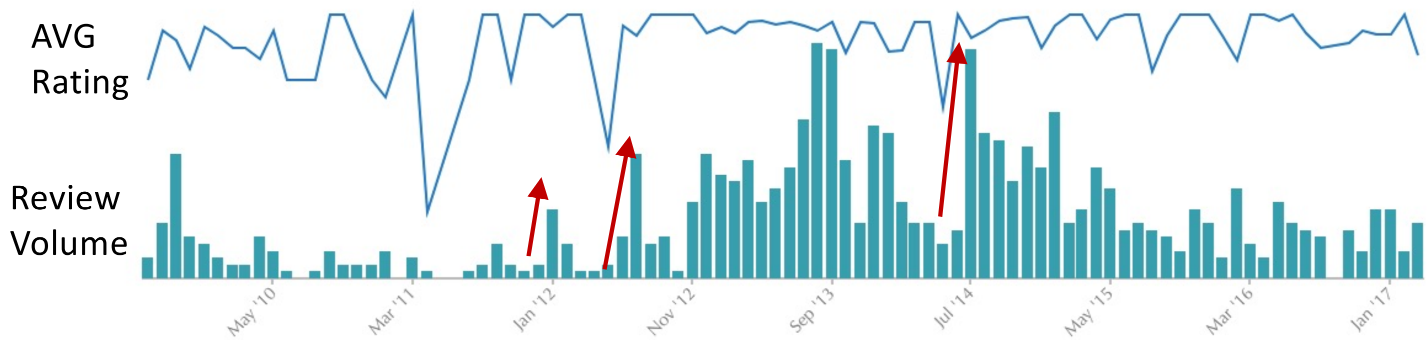
Evading a spam detector

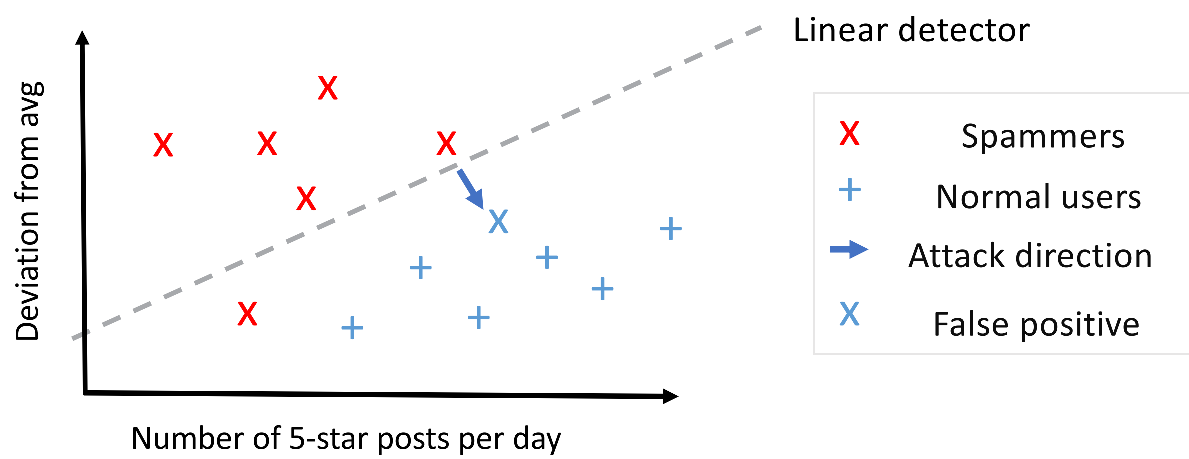
A strategic spammer will be more careful in posting fake reviews.

⇩

The strategic spammer will try to avoid the detection while manipulate the rating.

Risk of being detected vs. Profit of spamming

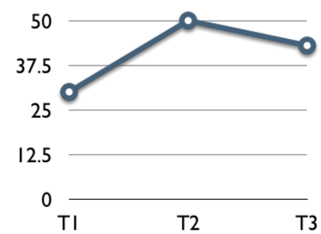




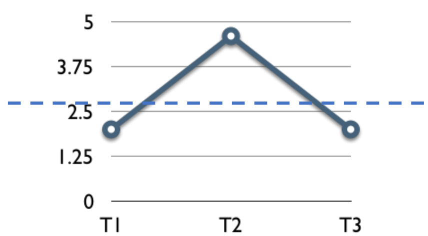
Evading a spam detector

Multiple detection signals need to be evaded:

- Number of reviews
- Change in the number of reviews



- Deviation from baseline average rating
- Change in rating



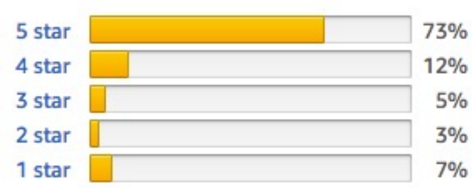
Rating optimization

Max: rating
Min: risk

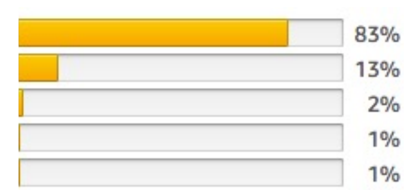
$n_\delta = 60$

- Rating distribution
- Change in the rating distribution

$\bar{\mathbf{p}}$: Background/last distribution



$\mathbf{p}$: Rating dist at time t



Data augmentation for robust detection

Probe parameters | Attack simulation | Attack in the wild!

First 30 weeks ... last 5 weeks

Training data generated from Evasion 1

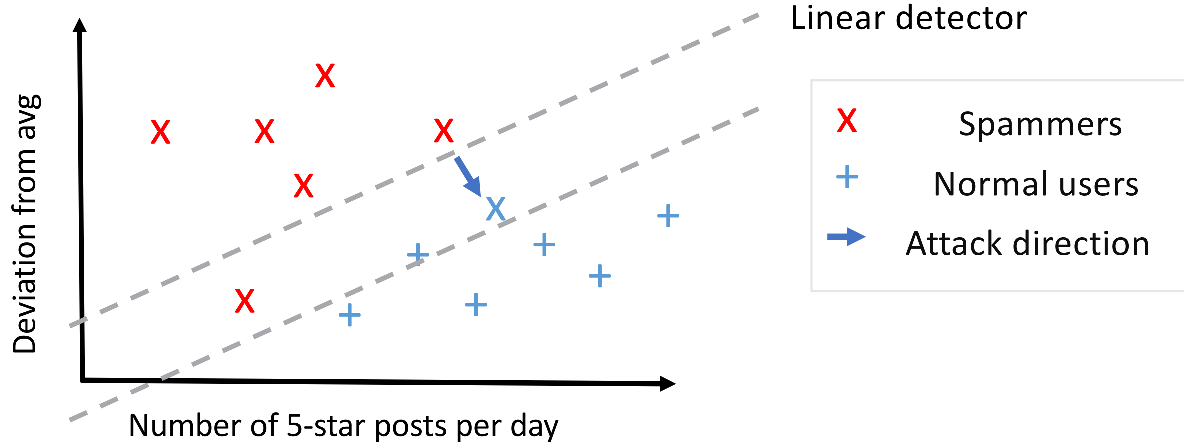
Training data generated from Evasion 2

Training data generated from Evasion 9

Model re-training and ensemble

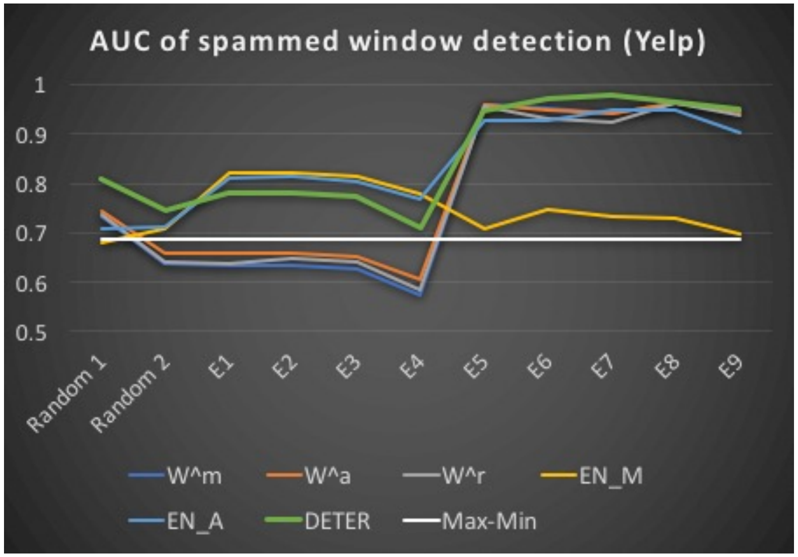
Pooling and Re-training (DETER)

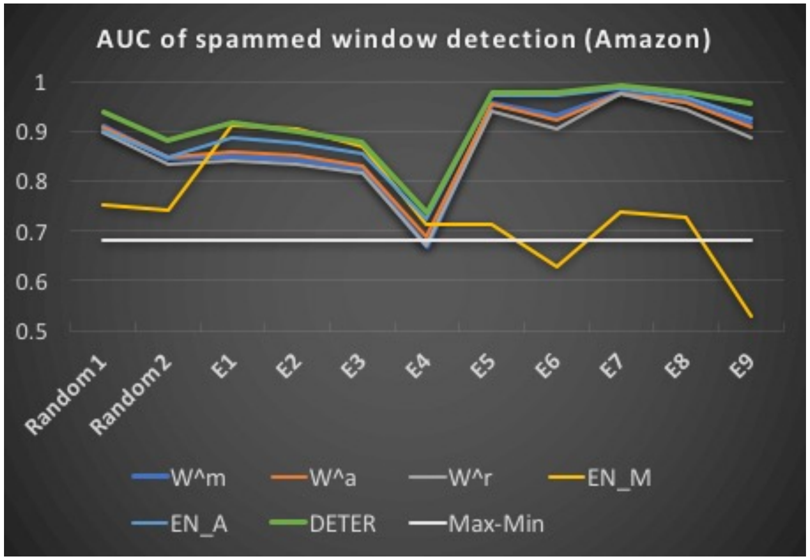
This technique is applicable to stores with sparse review data.



Robustness of the re-trained detector

Base detectors using statistics of time windows:
number of reviews, positive review ratio, change in rating distribution, …

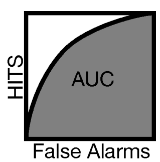




W^m: Max of signals W^a: Avg of signals W^r: Random selection EN_M: Re-train Max

EN_A: Re-train Avg DETER: Re-train Pool Max-min: Game equilibrium

Overview

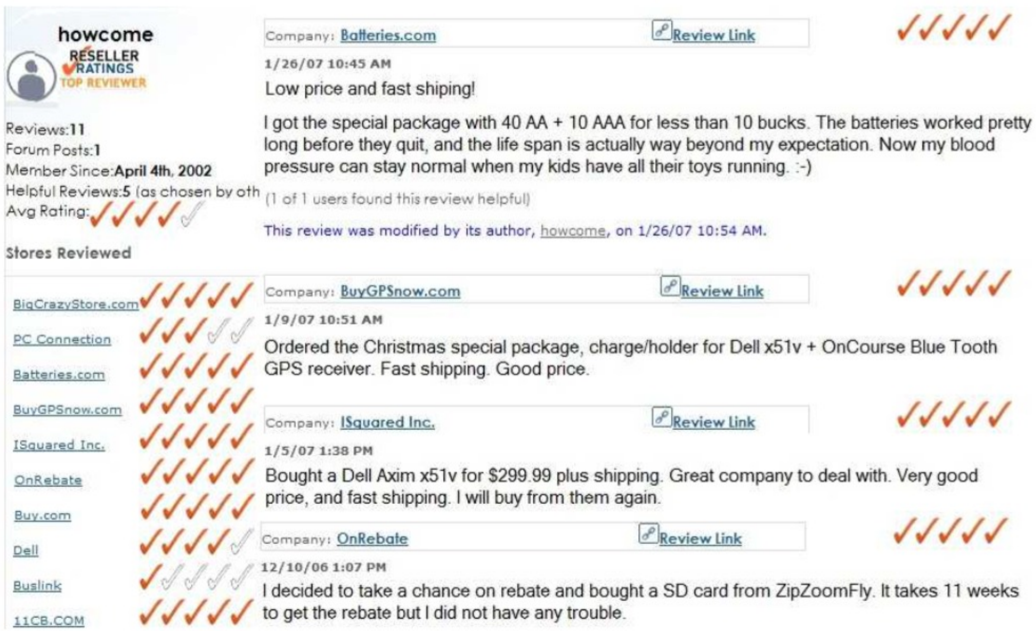


	Temporal	Graph	Contents
Fairness Interpretability		ICDM 2019, 2021 CIKM 2020	
Security	BigData 2018	KDD 2020	
Accuracy	KDD 2012	ICDM 2011	DSAA 2015

Review Graph Based Online Store Review Spammer Detection
Guan Wang, **Sihong Xie**, Bing Liu, Philip S. Yu

Detecting productive spammers

Can a reviewer with a long history and many reviews be a spammer?



- Diverse review texts
- Diverse ratings
- Spreaded out temporally

A strategic spammer can build credibility over time to hijack ratings.

Dependent trustworthiness on a graph

Algorithm:
iteratively and alternatively calculate Trustworthiness, Honesty, and Reliability, relying on previously computed quantities.

Similar trustworthiness with similar-minded reviews

$$A(v, \Delta t) = \sum_{i \in S_{v,a}} T(\kappa_i) - \sum_{j \in S_{v,d}} T(\kappa_j)$$

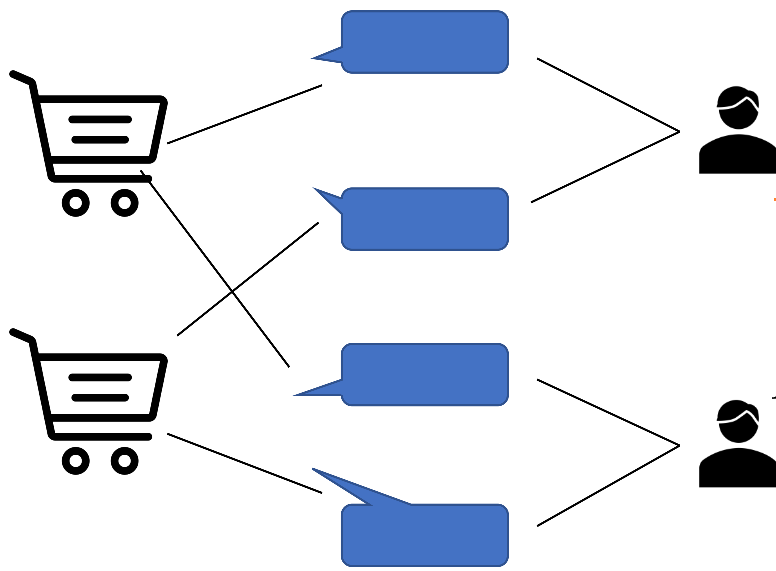
$$A_n(v, \Delta t) = \frac{2}{1 + e^{-A(v, \Delta t)}} - 1$$

Honesty of the review *v* $H(v) = |R(\Gamma_v)| A_n(v, \Delta t)$

Reliability of the business *s*

$$R(s) = \frac{2}{1 + e^{-\zeta}} - 1$$

$$\zeta = \sum_{v \in U_s, T(\kappa_v) > 0} T(\kappa_v)(\Psi_v - \mu)$$

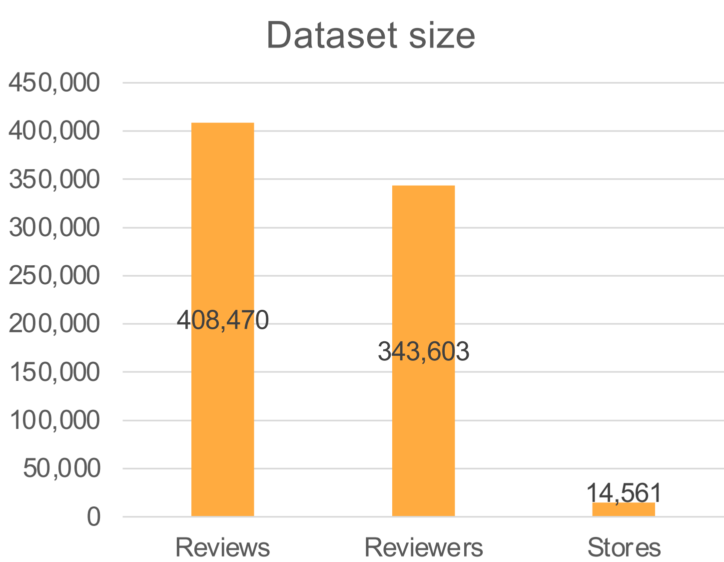


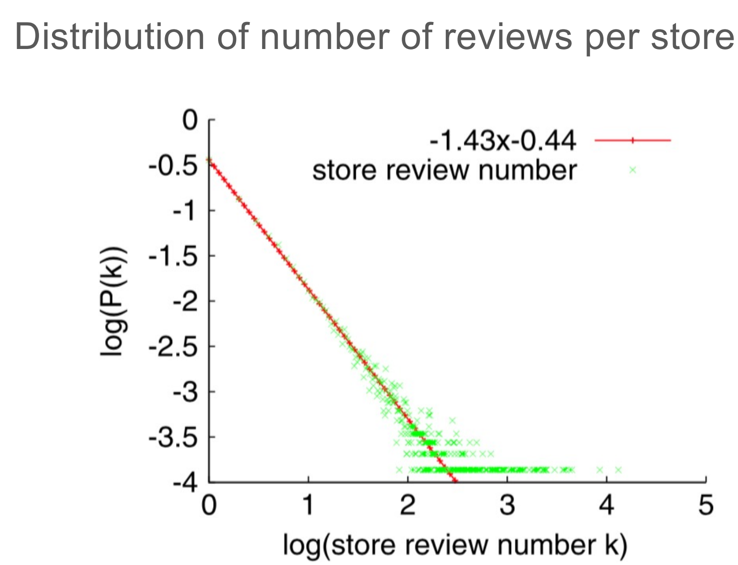
Trustworthiness of the reviewer *r*

$$T(r) = \frac{2}{1 + e^{-H_r}} - 1$$

$$H(r) = \sum_i \text{honest scores of } i\text{-th review by } r$$

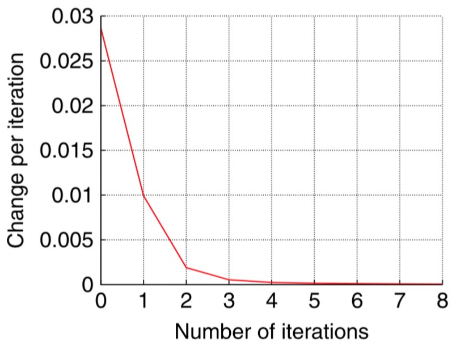
Experiments





Experiments

Convergence of rustworthiness



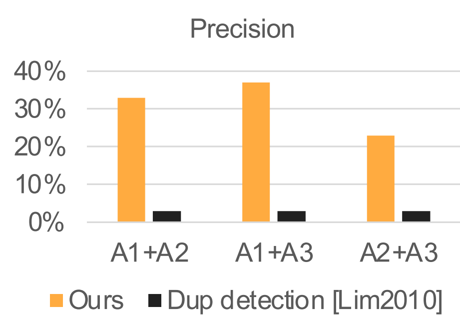
Quantitative evaluation

- Focused on precision@100

Evaluator inspection outcome and agreement

	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1	**49**	33	37
Evaluator 2	-	**35**	23
Evaluator 3	-	-	**40**

Evaluator agreement are statistically significant (kappa=60.3%)



Qualitative evaluation

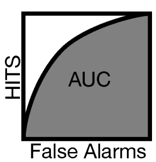
Top reliable stores

Store Name	Resellerratings Rating	BBB Rating
TigerDirect	7.44	A
SuperMediaStore	9.27	A^+
OneCall	9.33	A^+
Newegg	9.77	A^+
Mwave	9.18	B^-
LA Police Gear	9.11	A^-
iBuyPower	8.33	B^-
FrozenCPU	9.44	A^+
eWiz	9.08	C
eForcity	8.55	A^-

Bottom reliable stores

Store Name	Resellerratings Rating	BBB Rating
86^{th} Street Photo	0.30	F
Best Price Cameras	1.43	F
Dealer Cost Car Audio	1.23	F
USA Photo Nation	0.20	F
Camera Addict	0.59	F
CCI Camera City	0.44	A^+
OC System	3.00	F
Shop Digital Direct	0.35	F
Camera Giant	0.21	F
Infiniti Photo	0.28	F

Reinforcement learning for robust grap-based detection



	Temporal	Graph	Contents
Fairness Interpretability		ICDM 2019, 2021 CIKM 2020	
Security	BigData 2018	$\boxed{\text{KDD 2020}}$	
Accuracy	KDD 2012	ICDM 2011	DSAA 2015

Robust Spammer Detection by Nash Reinforcement Learning
Yingtong Dou, Guixiang Ma, Philip S. Yu, and Sihong Xie

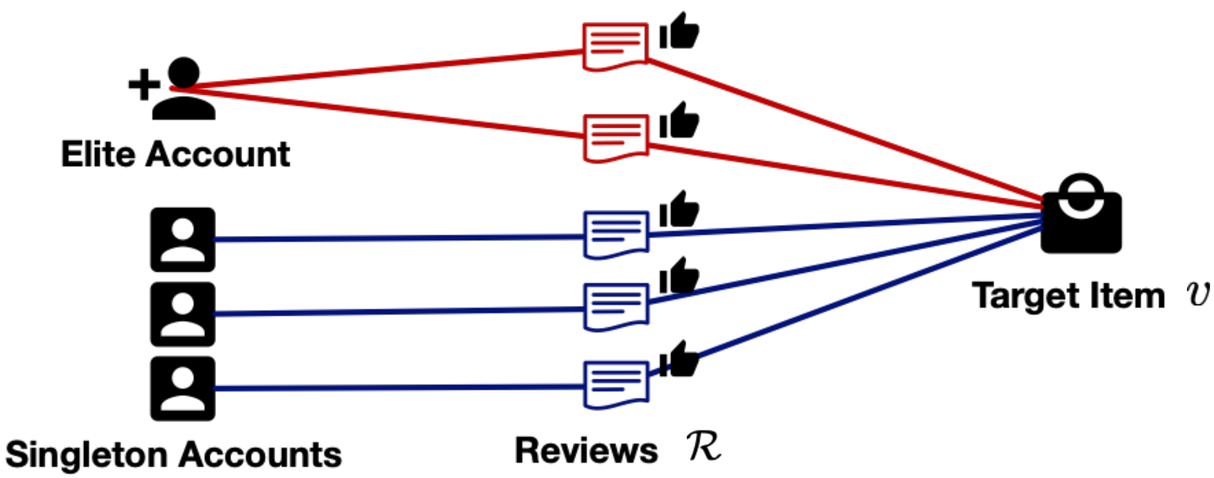
Reinforcement learning for robust grap-based detection

- **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

Rating and revenues

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

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



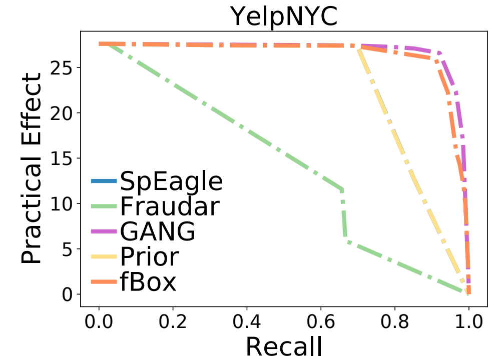
[1] M. Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com. HBS Working Paper (2016).

Spammer and detector goals

p: Spamming strategy

q: Detector strategy

Practical effect and detection recall are not in the same battlefield.



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

Revenue after attacks and detection.

Revenue before attacks

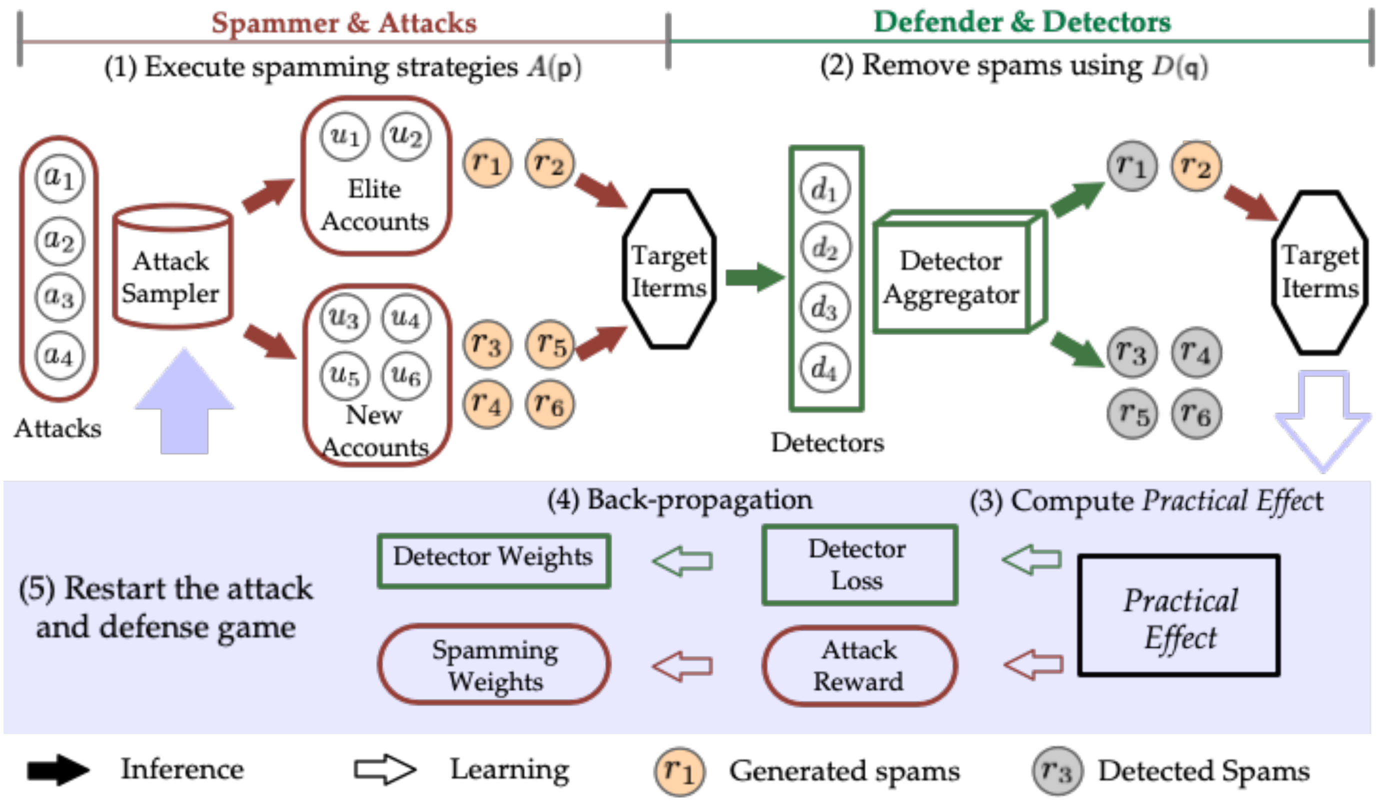
Spammer's Goal: $\max_{\mathsf{p}} \quad \max\{0, \mathrm{PE}(v;\mathcal{R},\mathsf{p},\mathsf{q})\}$

Defender's Goal: $\min_{\mathsf{q}} \quad \mathcal{L}_{\mathsf{q}} = \frac{1}{|\mathcal{R}(\mathsf{p},\mathsf{q})|}\sum_{r \text{ is FN}} -\boxed{C_{\mathrm{FN}}(v,r)}\,\boxed{\log P(y=1|r;\mathsf{q})}$

The cost of false negatives

The prediction results of detectors

Robust detector: Nash-Detect



Experimental settings

Base attack algorithms

1. **IncBP:** add reviews using the least suspicious accounts based on MRF.
2. **IncDS:** add reviews using accounts in the least dense block on review graph.
3. **IncPR:** add reviews using the least suspicious accounts based on behavior features.
4. **Random:** randomly select existing accounts to add reviews.
5. **Singleton:** add reviews with new accounts.

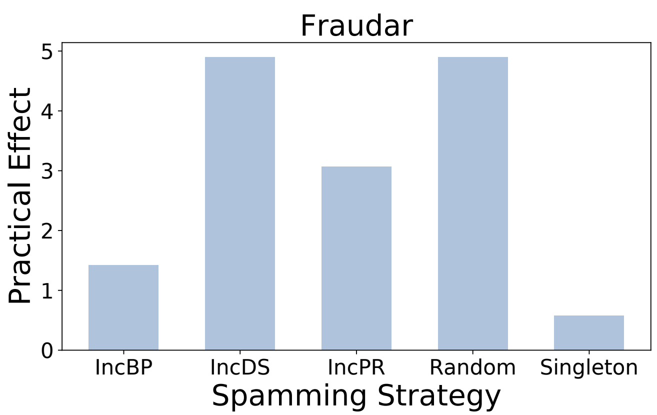
Base detection algorithms

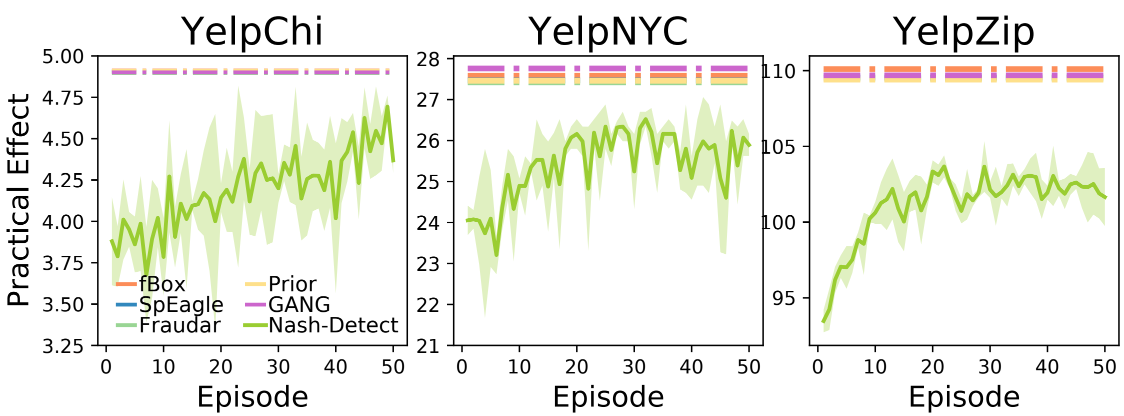
1. **GANG:** MRF-based detector
2. **SpEagle**: MRF-based detector
3. **fBox**: SVD-based detector for finding subtle changes in a large graph.
4. **Fraudar**: Dense-block detector
5. **Prior**: Behavior-based detector (rating changes, deviations, posting volume, etc.)

Overview

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

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





Transparency

- Model debugging:
  - why my algorithm is not detecting these fake reviews?
  - why the false positive rate is so high?

- Users' right to know:
  - why these reviews are removed?

- Auditing:
  - privacy
  - fairness

Fairness

- Auditing: company reputation and legal concerns.

- Are businesses treated equally:
  - some businesses may have advantange
    over others, based on regions, types, size, etc.

- Are customers have equal right to review products/businesses?:
  - It is not right to delete more of the new-comers' reviews,
    though they have a high chance to be spammed.