

Misinformation detection for e-commerce

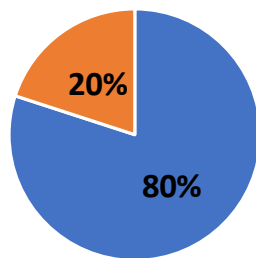
Sihong Xie, Assistant Professor
Computer Science and Engineering
Lehigh University



Misinformation are prevalent

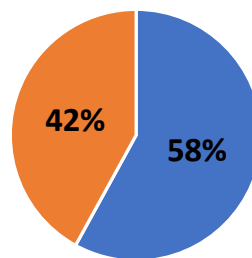
Estimated percentage of fake reviews on popular e-commerce websites.

Percentage of Fake Reviews on
Yelp



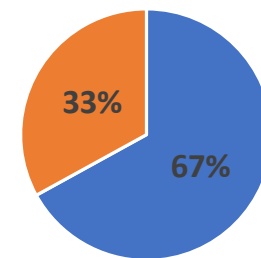
■ Genuine ■ Fake

Percentage of Fake Reviews on
Amazon



■ Genuine ■ Fake

Percentage of Fake Reviews on
TripAdvisor

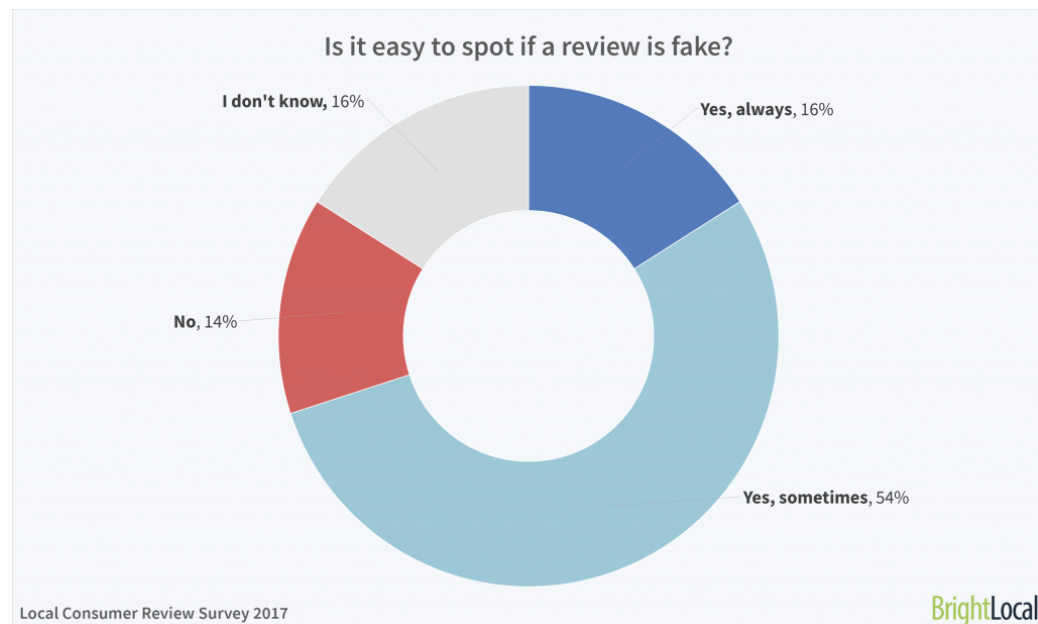


■ Genuine ■ Fake

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

How to spot a fake review

Likely a fake review:

Extreme review

ac11ca
 Published 12 hours ago
 ★★★★★
 These r the greatest headphones ever!! They have high-quality sound (including JBL Pure Bass Sound). You can connect without worry to your Bluetooth. The headphones last an amazing 11-hour battery life. You should definiately buy one.

Relatively short, poorly written, one-sided review full of product details

Few details about the reviewer

Unestablished reviewer with few, one-sided reviews

Few "helpful" votes by other consumers

Unverified purchase

Age:
 Gender:
 Member since July 2019
 1 reviews left
 1.0 average review score

0 people found this helpful

?

Likely a genuine review:

Balanced review

Adrian
 Sydney, AUS
 Age: 34
 Gender: Male
 Member since February 2009
 58 reviews left
 3.8 average review score

Great value headphones with one caveat
 Published 12 hours ago
 ★★★★★
 I purchased these headphones three weeks ago and so far they have been great. I am a big fan of jazz music and the bass really comes through with these headphones, which was a happy surprise given the price. The battery life is also quite good; I usually charge them every few days and it is not a problem. The Bluetooth connected for me without much hassle. However, I do wish the connection was stronger because I often have connection issues when walking around my house. I also wish there was a cable to use to avoid Bluetooth altogether. Overall though, these are great headphones given the price.

Relatively long, well written, two-sided review full of subjective details

Established reviewer with many two-sided reviews

Many "helpful" votes by other consumers

Verified purchase

10 people found this helpful

amazon.in prime

By **Damaya** on 2 March 2018
 Verified Purchase
 Very good product and nice this is awesome product and quality is good
 One person found this helpful

Helpful Not Helpful Report

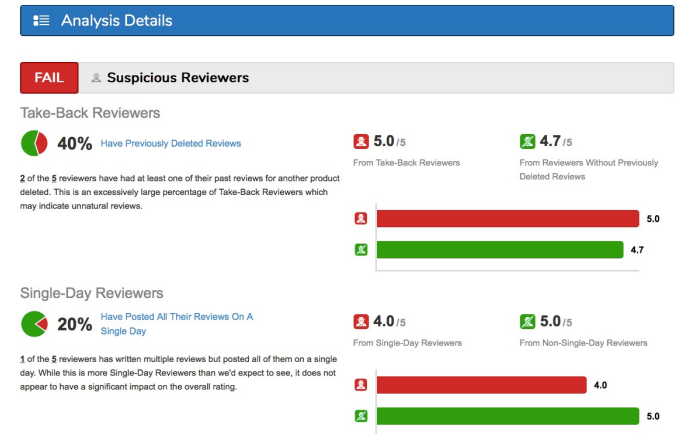
★★★★★ **Five Stars**
 By **Nausad** on 3 March 2018
 Verified Purchase
 Awesome product and quality is good quality and this is nice product and nice this is awesome
 2 people found this helpful

Helpful Not Helpful Report

★★★★★ **Five Stars**
 By **Diviya** on 2 March 2018
 Verified Purchase
 Nice and cool product and quality is good
 One person found this helpful

ReviewMeta.com

1. Feature engineering
2. Detection models



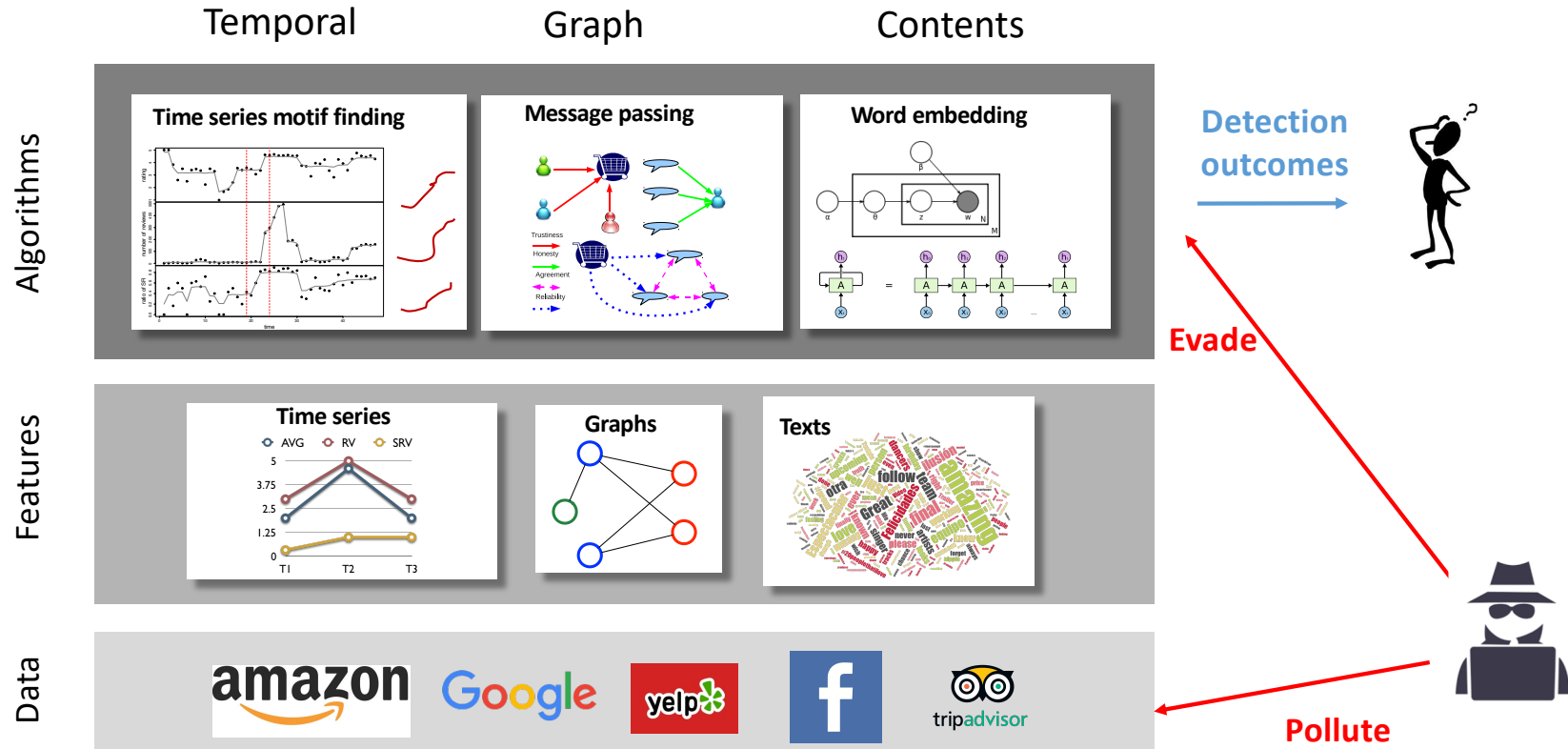
Pros
 Cons

Rely on yourselves
 Not everyone can spot fake reviews

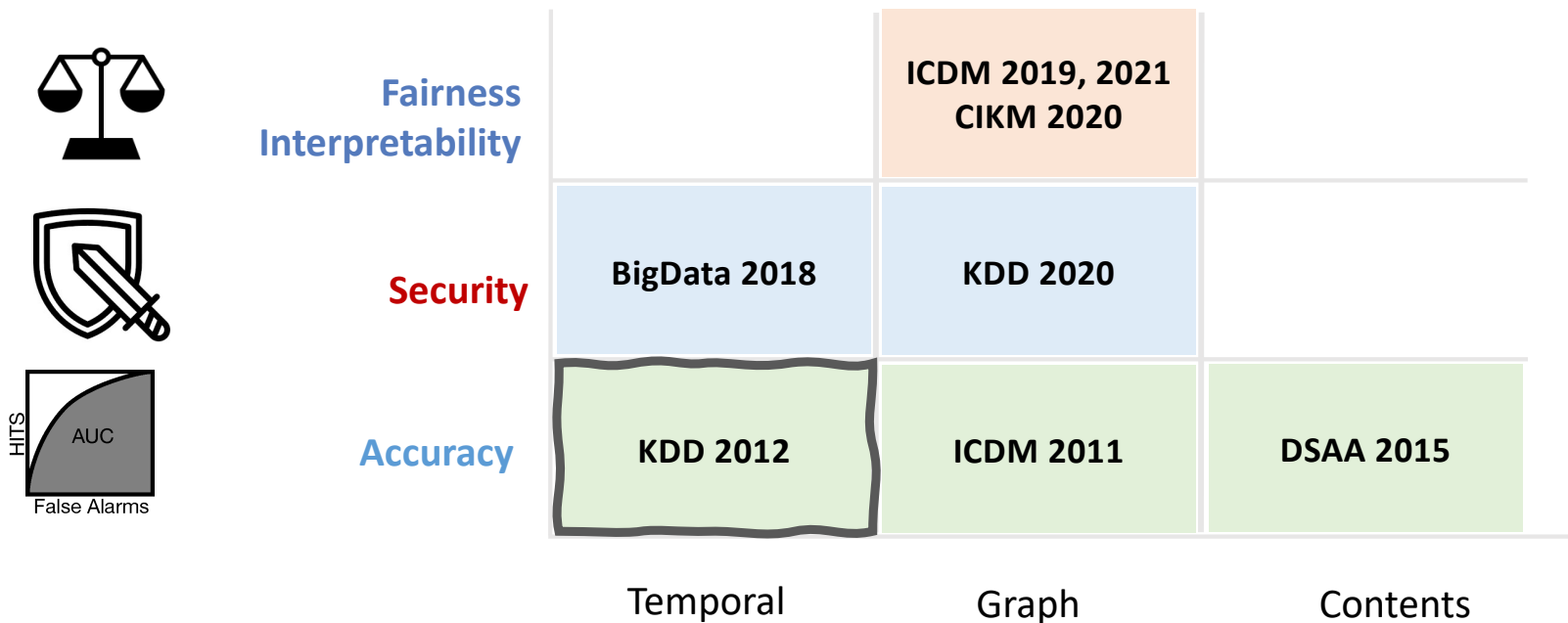
Convenient and easy
 Can be gamed

More in-depth automatic analysis
 Can be gamed

Misinformation detection architecture



Overview



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?



roneconner

★★★★★ 5/5

posted Jul-07-2012

"My experience with B&H Photo-Video-Pro Audio was excellent. Order arrived sooner than expected and in good shape. I am very satisfied with my new Acer Iconia Tab A200 and accessories, and the Acer tablet was 50\$ less than my local Best Buy and paid no tax or shipping to boot."



Steve-29

★★★★★ 5/5

posted Jul-07-2012

"Good prices, easy-to-use website, efficient delivery, they make it tough to consider going elsewhere."



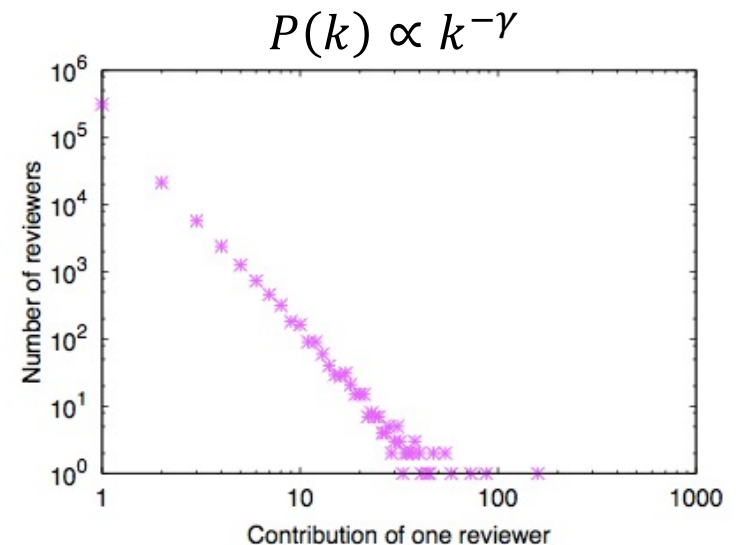
misterfupr

★★★★★ 5/5

posted Jul-07-2012

"Found everything I was looking for in a single stop. Super fast shipping. "

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



Exploiting an invariance of spamming

Invariance:

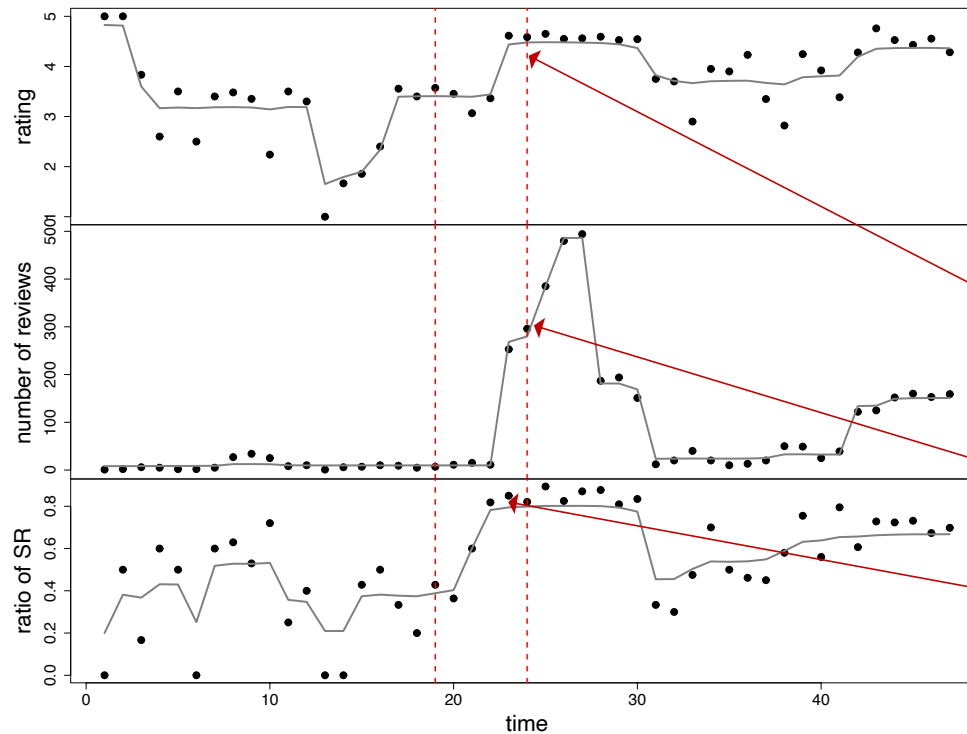
to manipulate ratings, a **large number of consistent ratings** must be posted in a **short time**.

Temporal features
for each window:

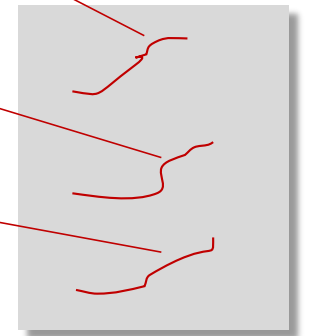
AVG rating

Review Volume

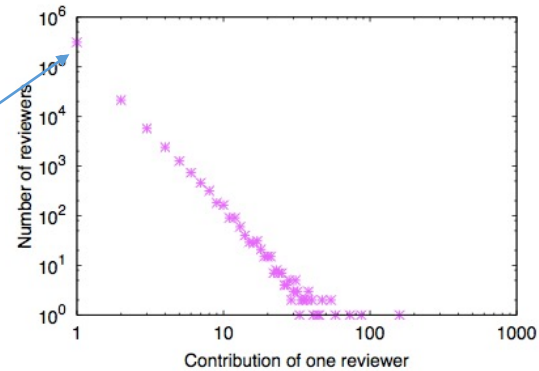
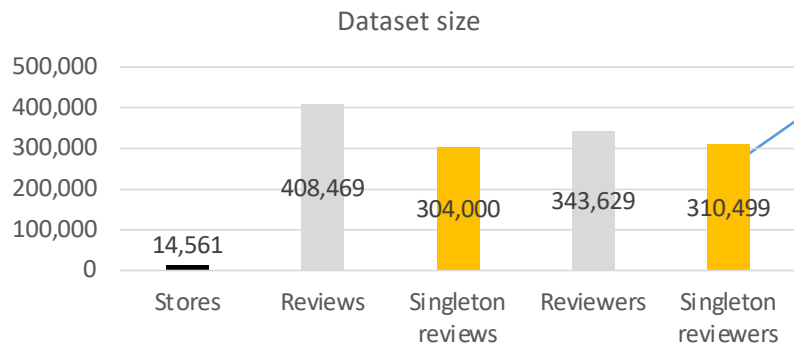
Singleton Review Volume



Burst motif detection
on all 3 series.



Results

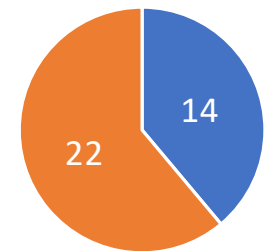
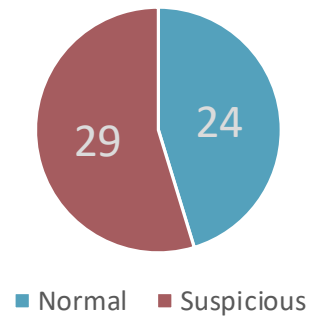


Manual labeling of dishonest businesses

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

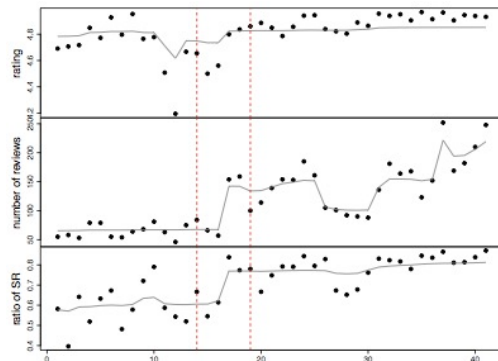
	Evaluator 1	Evaluator 2	Evaluator 3
Evaluator 1	17	14	16
Evaluator 2	-	20	19
Evaluator 3	-	-	24

Burst detector Performance



Case study

A detected 15-day window

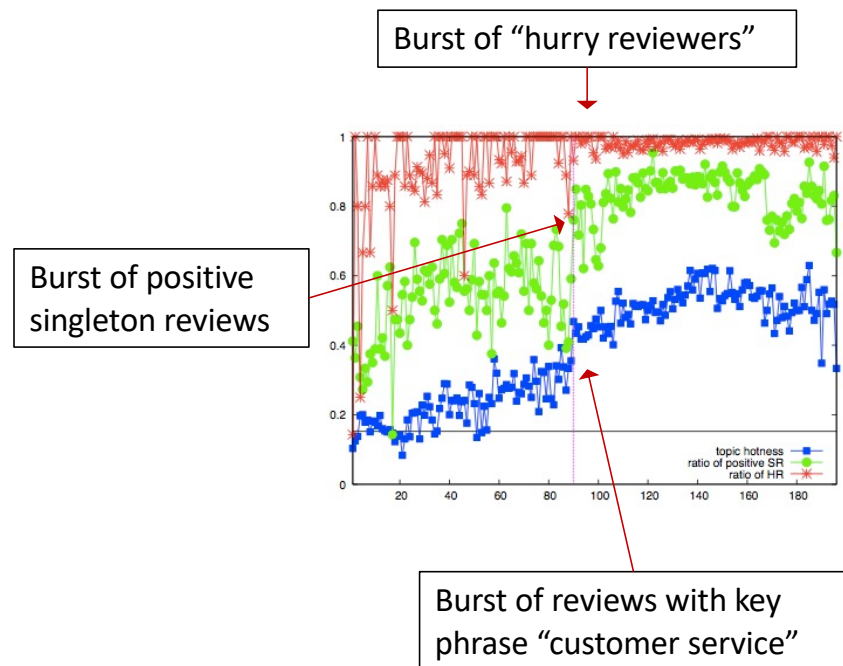


AVG rating: 4.4 → 4.79

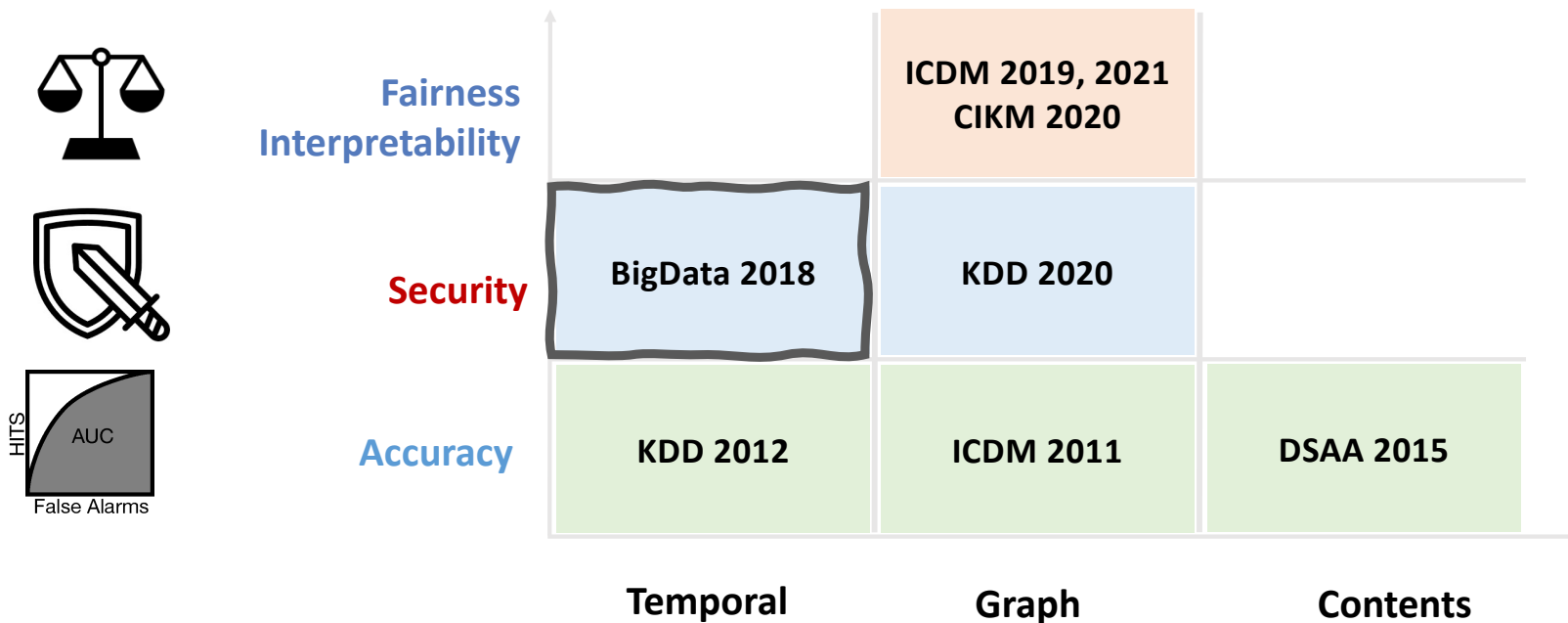
Volume of reviews: 57 → 154

Ratio of singleton reviews: 61% → 83%

More evidences



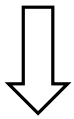
Overview of my research



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

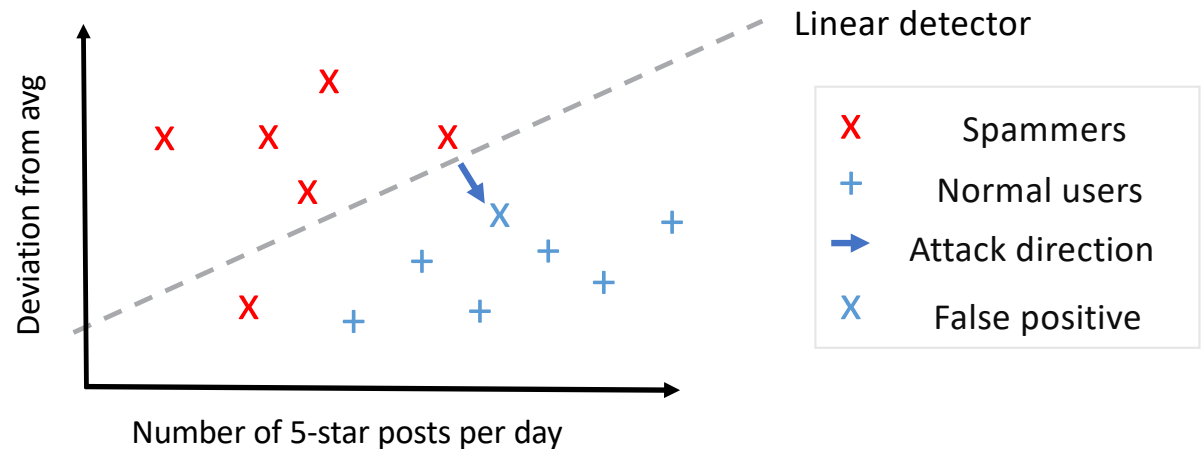
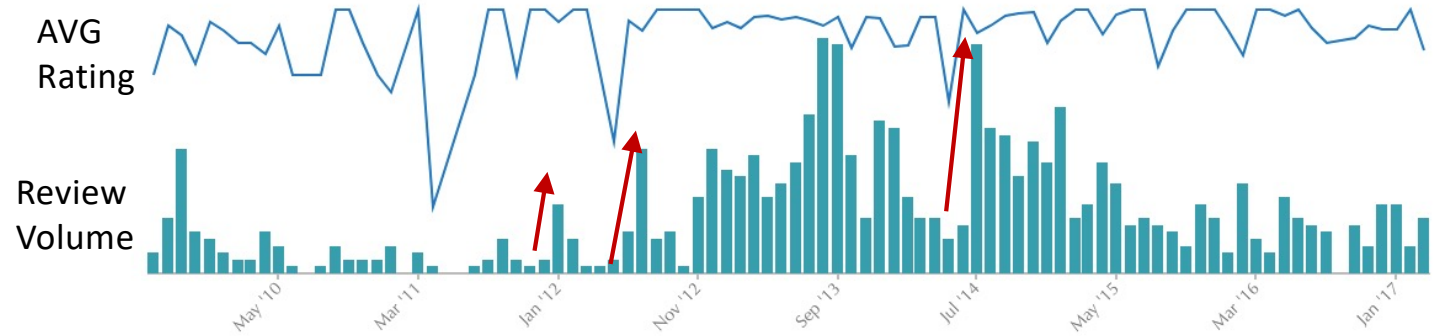
Evading a spam detector

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



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

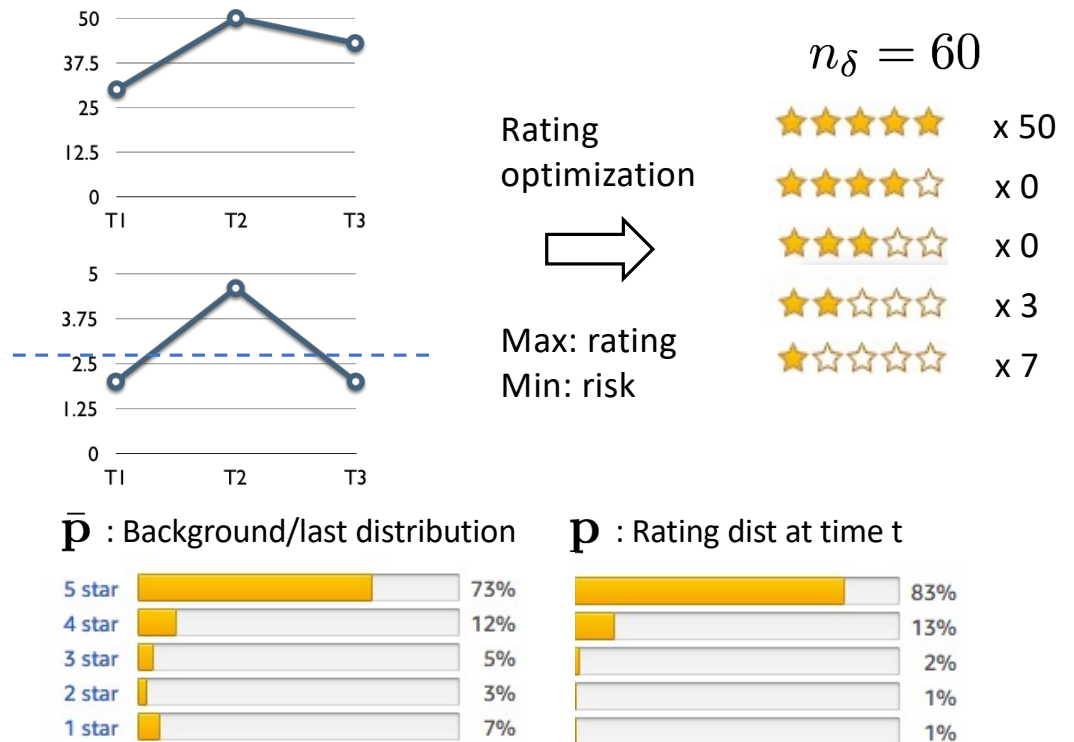
Risk of being detected vs. Profit of spamming



Evading a spam detector

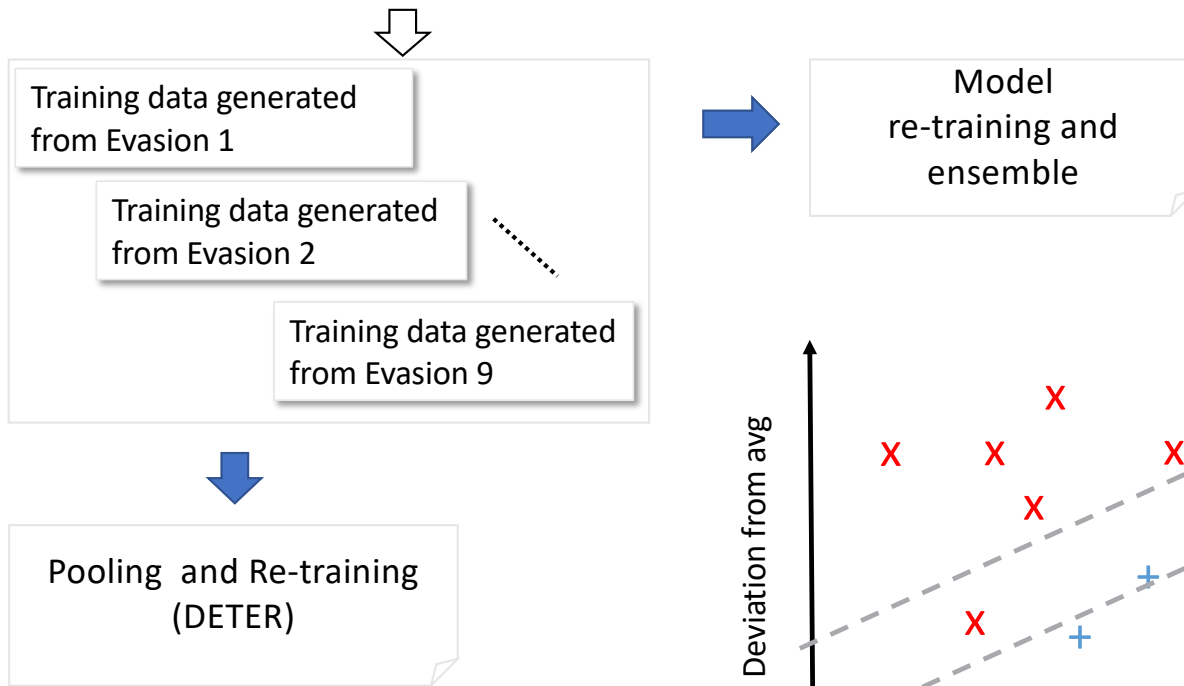
Multiple detection signals need to be evaded:

- Number of reviews
- Change in the number of reviews
- Deviation from baseline average rating
- Change in rating
- Rating distribution
- Change in the rating distribution

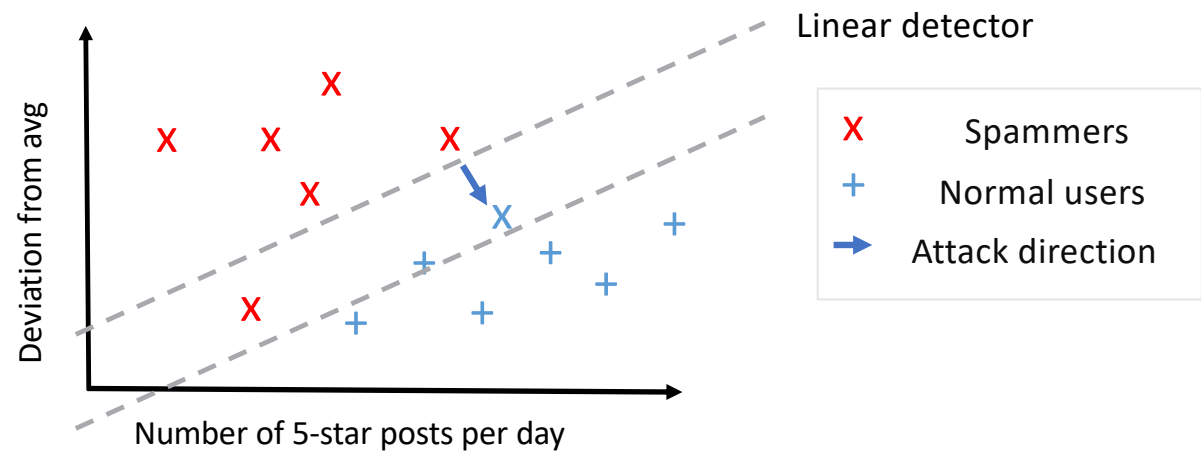


Data augmentation for robust detection

Probe parameters Attack simulation Attack in the wild!
First 30 weeks last 5 weeks

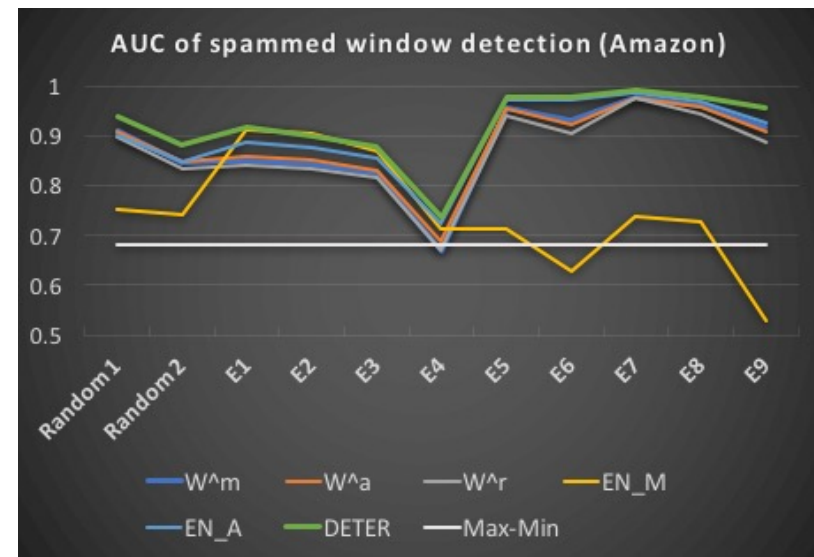
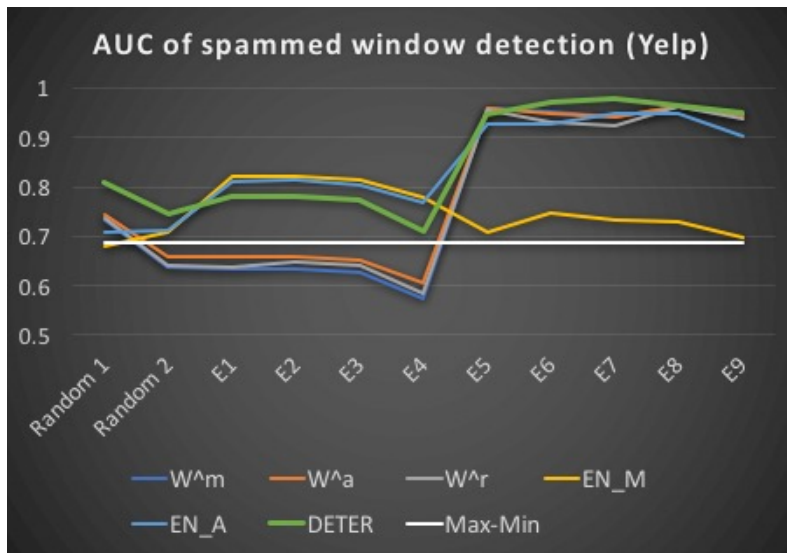


This technique is applicable to stores with sparse review data.



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

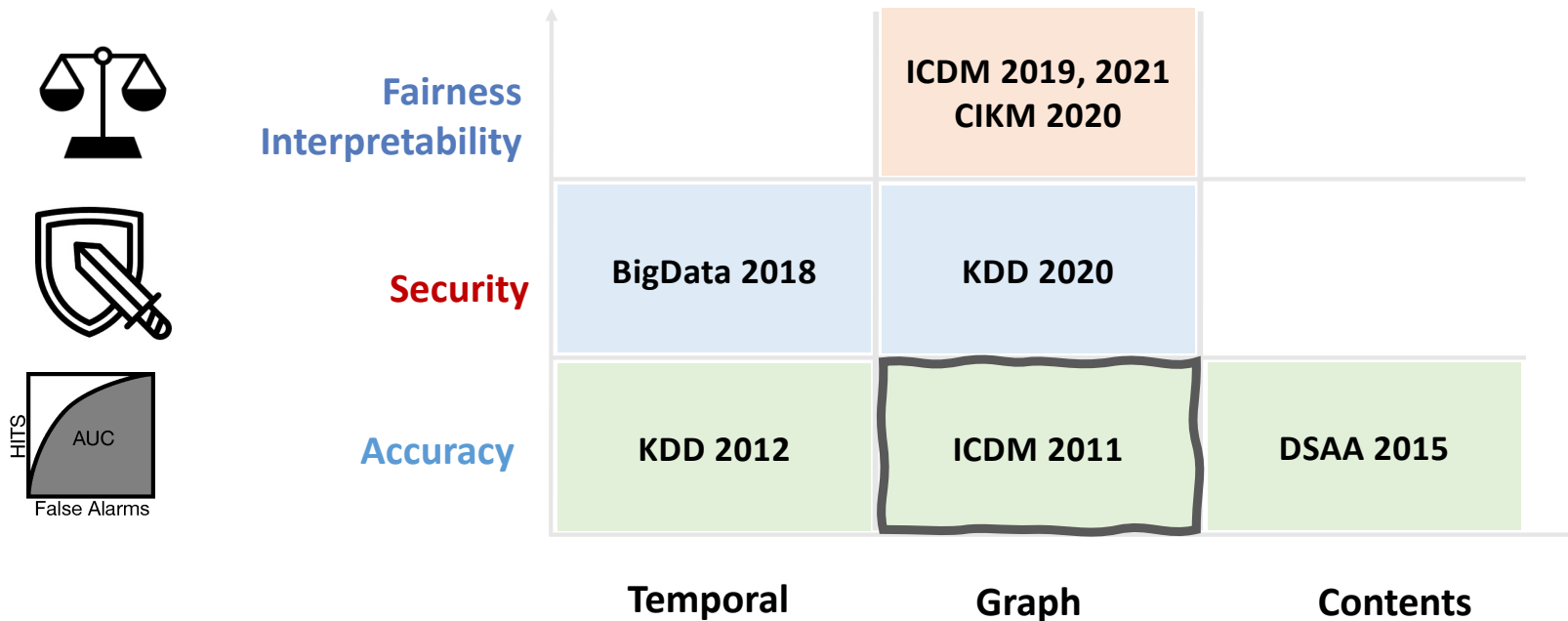
EN_M: Re-train Max

EN_A: Re-train Avg

DETER: Re-train Pool

Max-min: Game equilibrium

Overview



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?

The screenshot shows a user profile for 'howcome' with the following details:

- Profile: howcome, RESELLER, RATINGS, TOP REVIEWER
- Reviews: 11, Forum Posts: 1, Member Since: April 4th, 2002, Helpful Reviews: 5 (as chosen by others), Avg Rating: 5 stars
- Stores Reviewed: A list of 10 companies, each with a 5-star rating and a 'Review Link' button.

The review history shows the following entries:

Company	Date	Review Text	Rating
Batteries.com	1/26/07 10:45 AM	Low price and fast shipping! I got the special package with 40 AA + 10 AAA for less than 10 bucks. The batteries worked pretty long before they quit, and the life span is actually way beyond my expectation. Now my blood pressure can stay normal when my kids have all their toys running. :-)	5 stars
BuyGPSnow.com	1/9/07 10:51 AM	Ordered the Christmas special package, charge/holder for Dell x51v + OnCourse Blue Tooth GPS receiver. Fast shipping. Good price.	5 stars
ISquared Inc.	1/5/07 1:38 PM	Bought a Dell Axim x51v for \$299.99 plus shipping. Great company to deal with. Very good price, and fast shipping. I will buy from them again.	5 stars
OnRebate	12/10/06 1:07 PM	I decided to take a chance on rebate and bought a SD card from ZipZoomFly. It takes 11 weeks to get the rebate but I did not have any trouble.	5 stars

- 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

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

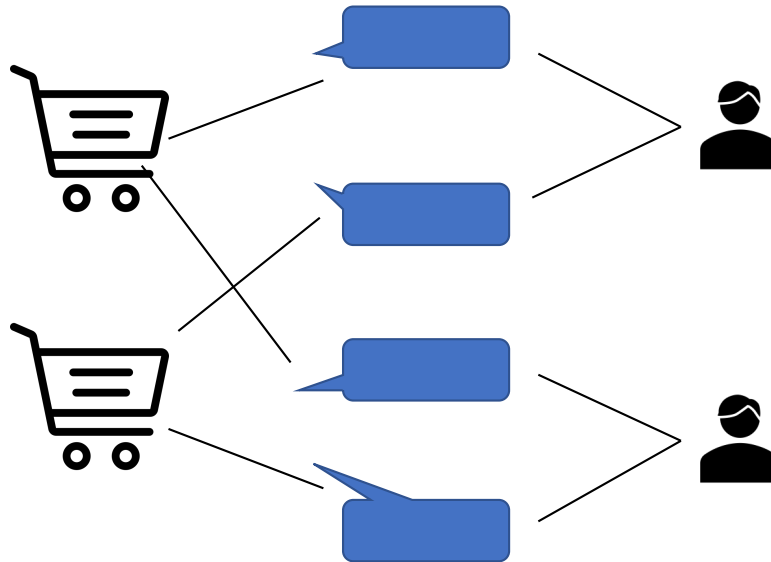
$$A_n(v, \Delta t) = \frac{2}{1 + e^{-A(v, \Delta t)}} - 1$$

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

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

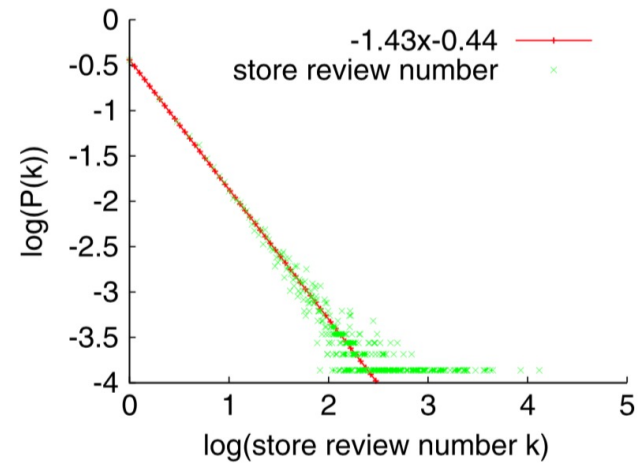
$$H(r) = \sum_i \text{honest scores of } i\text{-th review by } r$$

Experiments

Dataset size

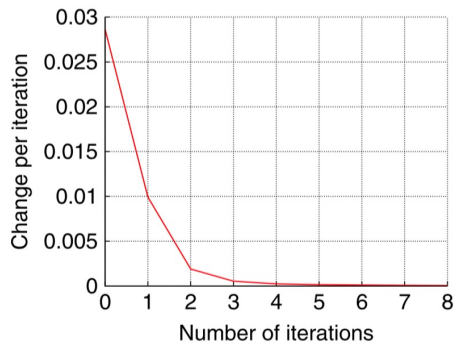


Distribution of number of reviews per store



Experiments

Convergence of rustworthiness



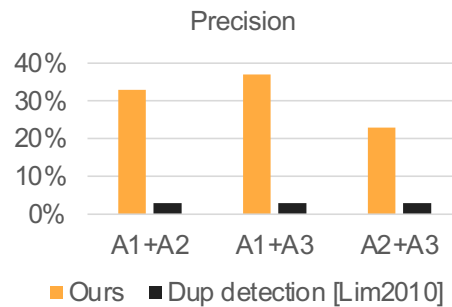
Quantitative evaluation

- 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

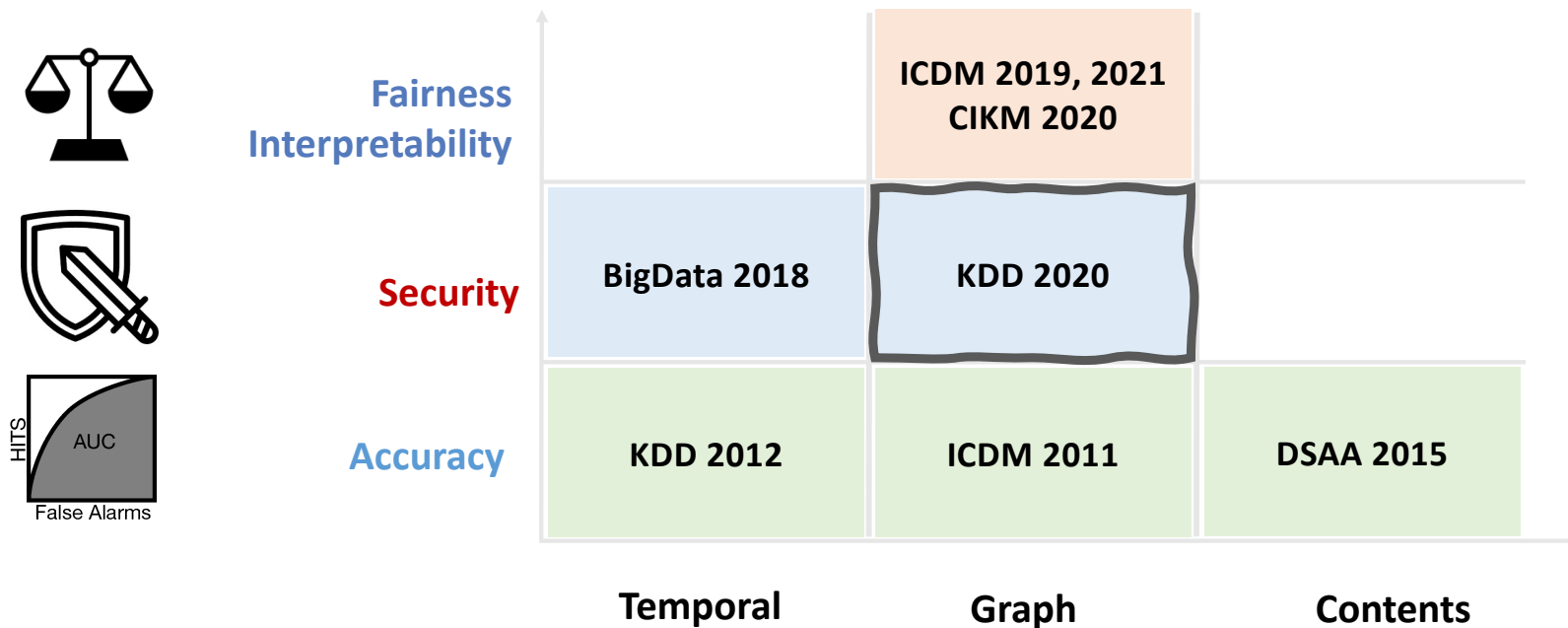
Top reliable stores

Store Name	Reselleratings 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	Reselleratings 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



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

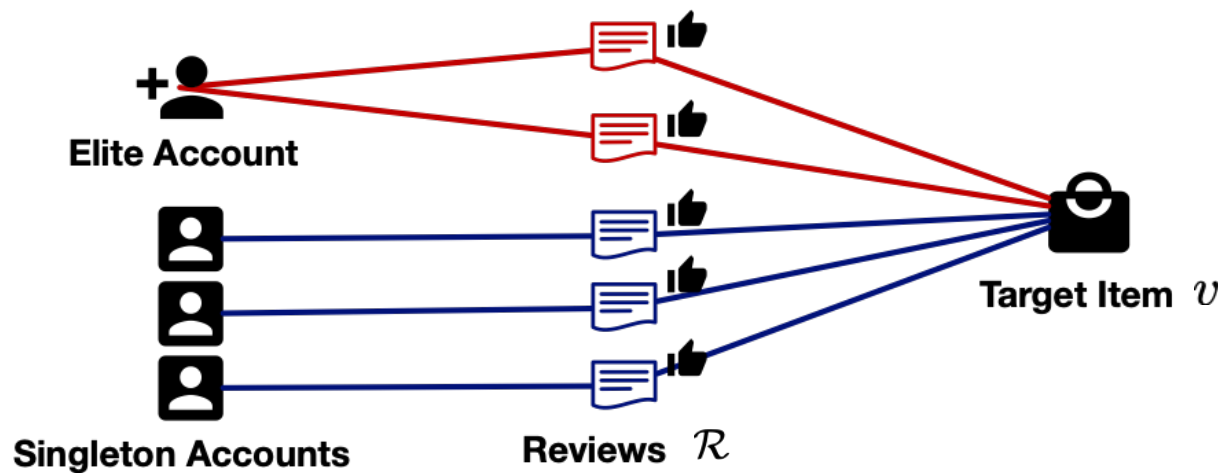
Reinforcement learning for robust grap-based detection

- **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: $f(v; \mathcal{R}) = \beta_0 \times \text{RI}(v; \mathcal{R}) + \beta_1 \times \text{ERI}(v; \mathcal{R}_E(v)) + \alpha$



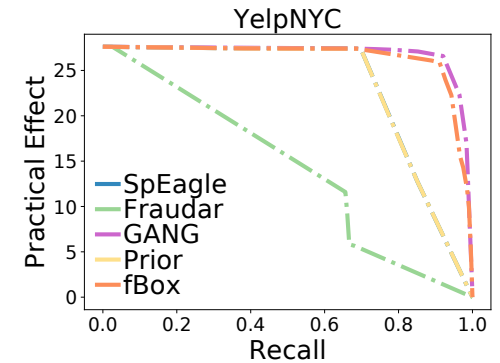
[1] M. Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com. HBS Working Paper (2016).

Spammer and detector goals

p: Spamming strategy

q: Detector strategy

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



Spamming Practical Effect:

$$PE(v; \mathcal{R}, p, q) = \boxed{f(v; \mathcal{R}(p, q))} - \boxed{f(v; \mathcal{R})}$$

Revenue after attacks and detection.

Revenue before attacks

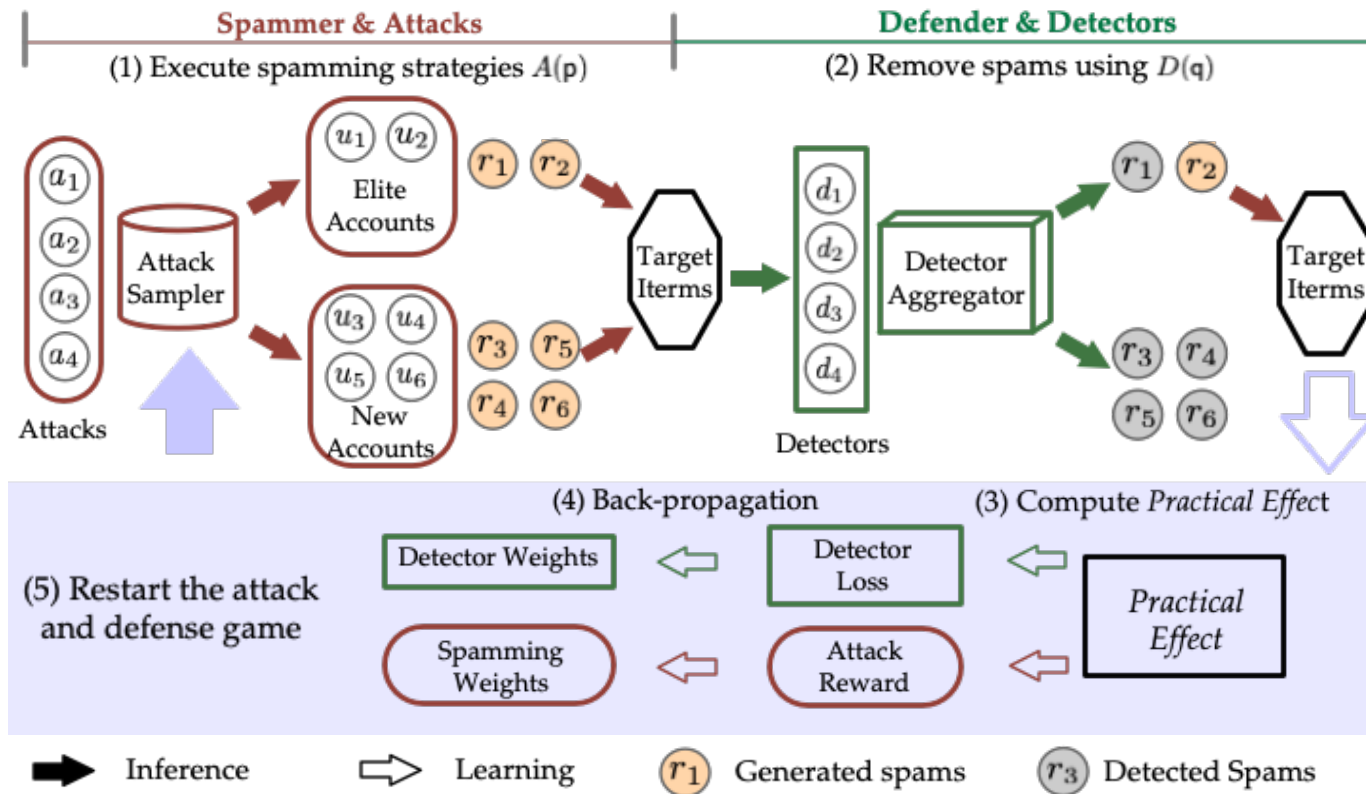
Spammer's Goal: $\max_p \max\{0, PE(v; \mathcal{R}, p, q)\}$

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

The cost of false negatives

The prediction results of detectors

Robust detector: Nash-Detect



Experimental settings

Base attack algorithms

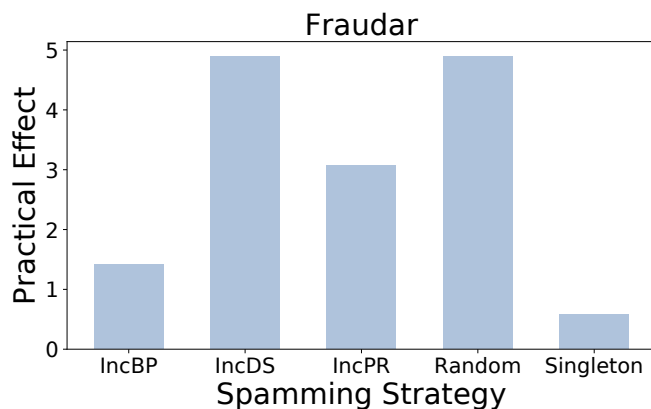
1. **IncBP**: add reviews using the least suspicious accounts based on MRF.
2. **IncDS**: add reviews using accounts in the least dense block on review graph.
3. **IncPR**: add reviews using the least suspicious accounts based on behavior features.
4. **Random**: randomly select existing accounts to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

Base detection algorithms

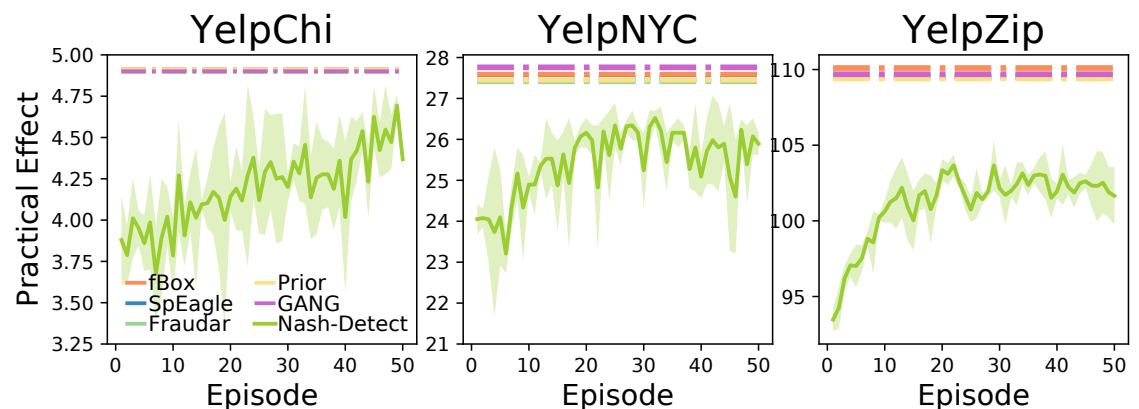
1. **GANG**: MRF-based detector
2. **SpEagle**: MRF-based detector
3. **fBox**: SVD-based detector for finding subtle changes in a large graph.
4. **Fraudar**: Dense-block detector
5. **Prior**: Behavior-based detector (rating changes, deviations, posting volume, etc.)

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 advantage 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.