Securing Behavior-based Opinion Spam Detection

Shuaijun Ge, Guixiang Ma, Sihong Xie and Philip S. Yu

Dec 13, 2018 @BigData

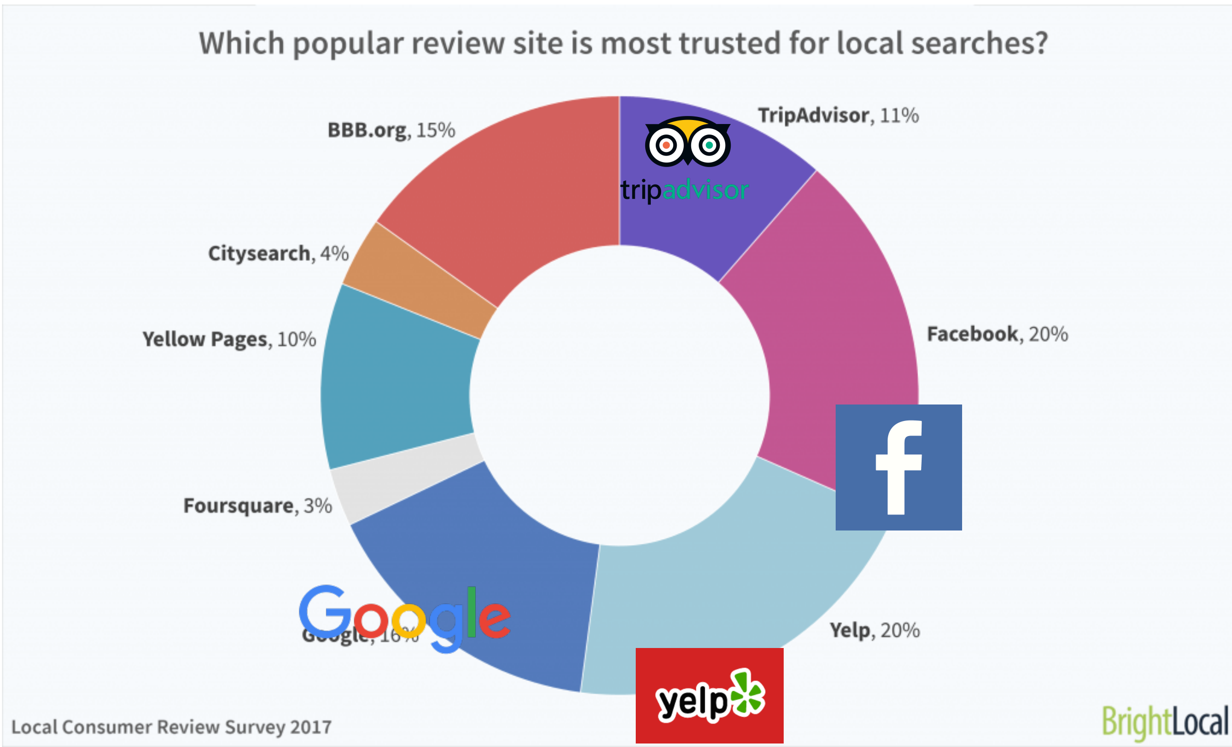
LEHIGH UNIVERSITY

Fake reviews

?

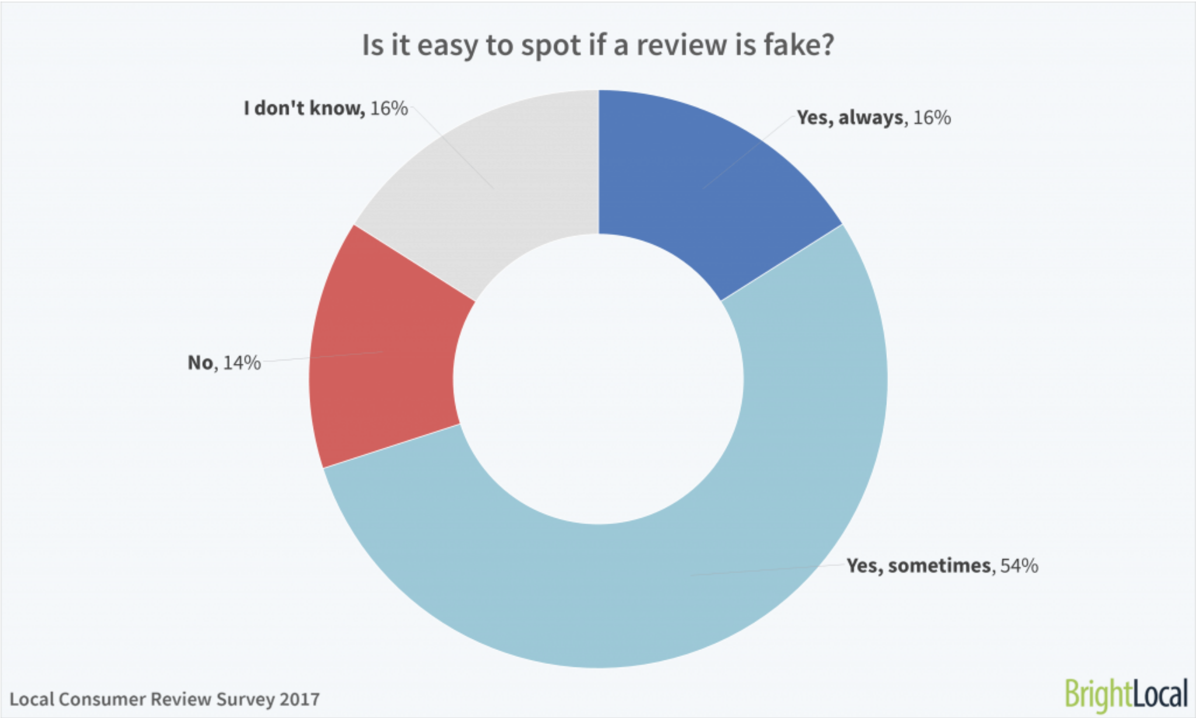
LEHIGH UNIVERSITY

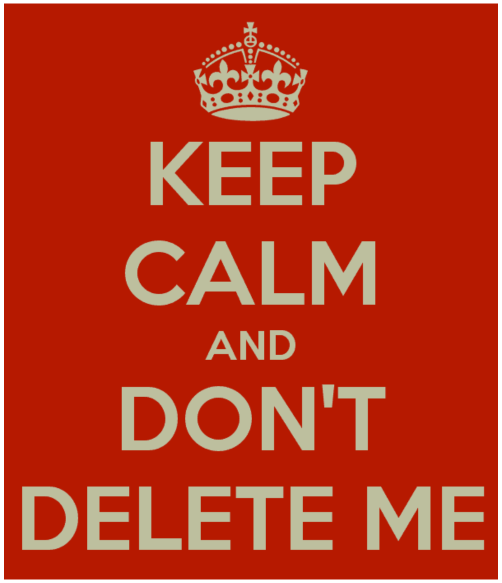
Online reviews



Source: https://www.brightlocal.com/learn/local-consumer-review-survey/ based on a pool of representative sample of 1,031 US-based consumers

The challenges



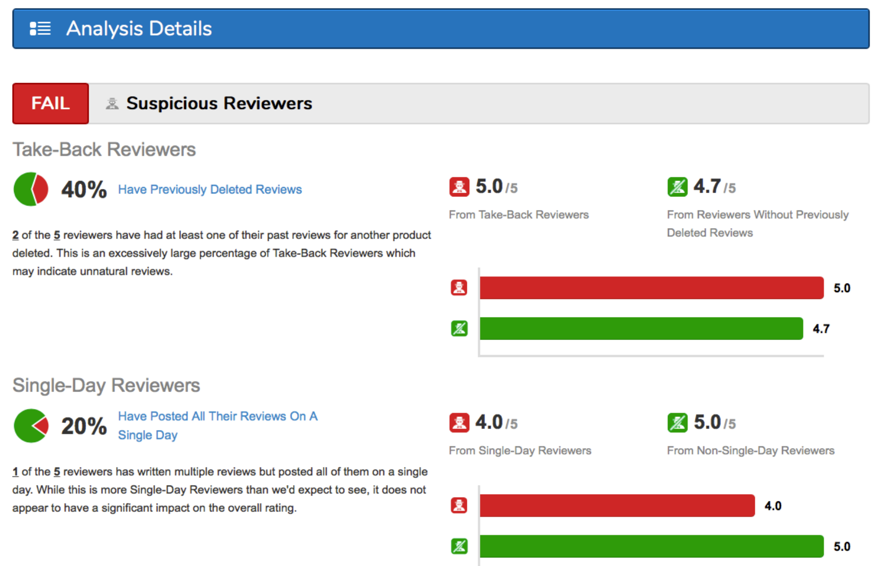


Source: https://www.brightlocal.com/learn/local-consumer-review-survey/ based on a pool of representative sample of 1,031 US-based consumers

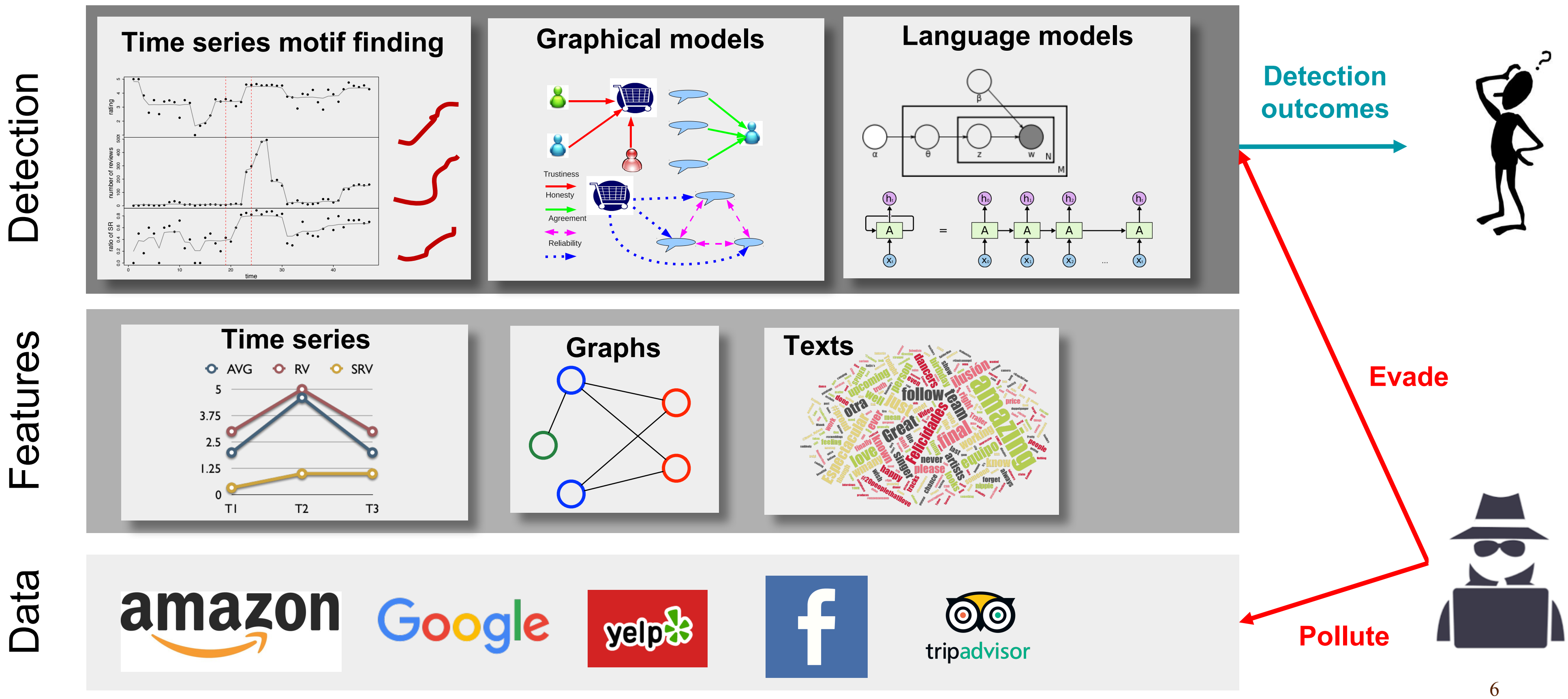
Existing efforts

Outcome + Explanations

ReviewMeta.com
1. Feature engineering
2. Detection models



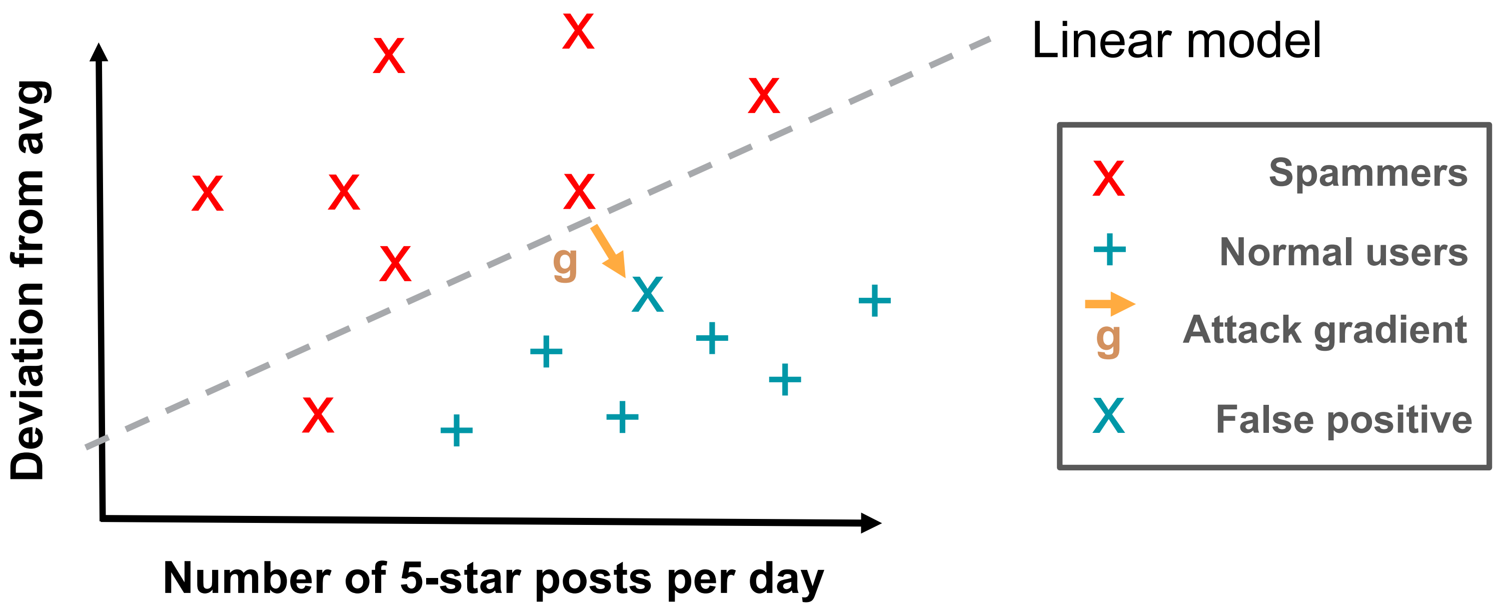
Help make decision



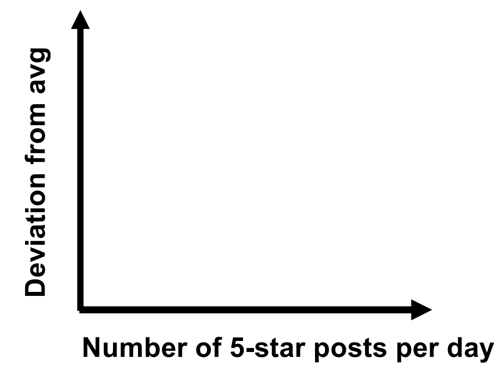
Detection
Features
Data
Time series motif finding
Graphical models
Trustiness
Honesty
Agreement
Reliability
Language models
Time series
AVG
RV
SRV
Graphs
Texts
amazon
Google
yelp
f
tripadvisor
Detection outcomes
Evade
Pollute

Behavior based Attacking

Spamming
Account
Detection

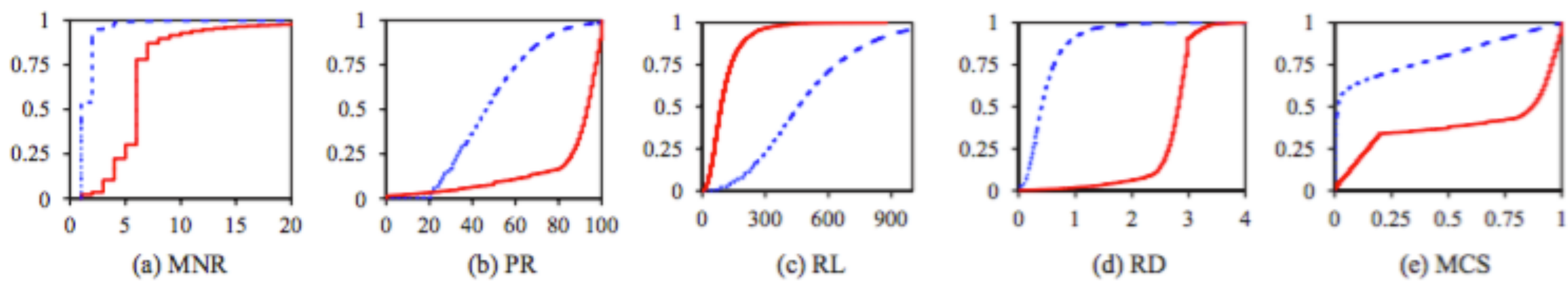


Behavior based Attacking



- **Accessing knowledge of detector** (publications)

***What yelp fake review filter might be doing*, ICWSM, 2013**



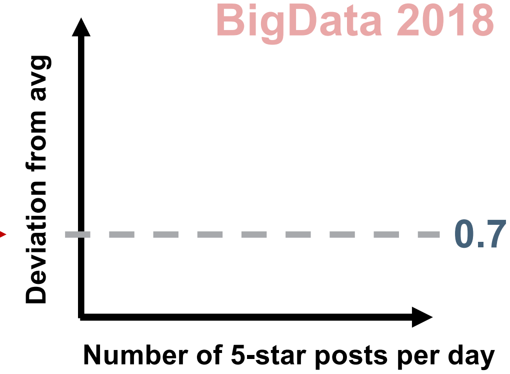
max # of review per day | **Positive ratio** | **Review length** | **Rating deviation** | **Maximum content similarity**

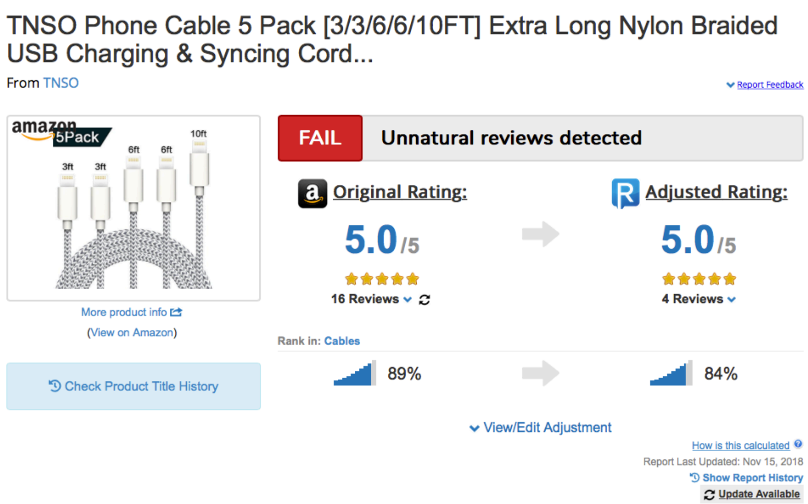
LEHIGH UNIVERSITY

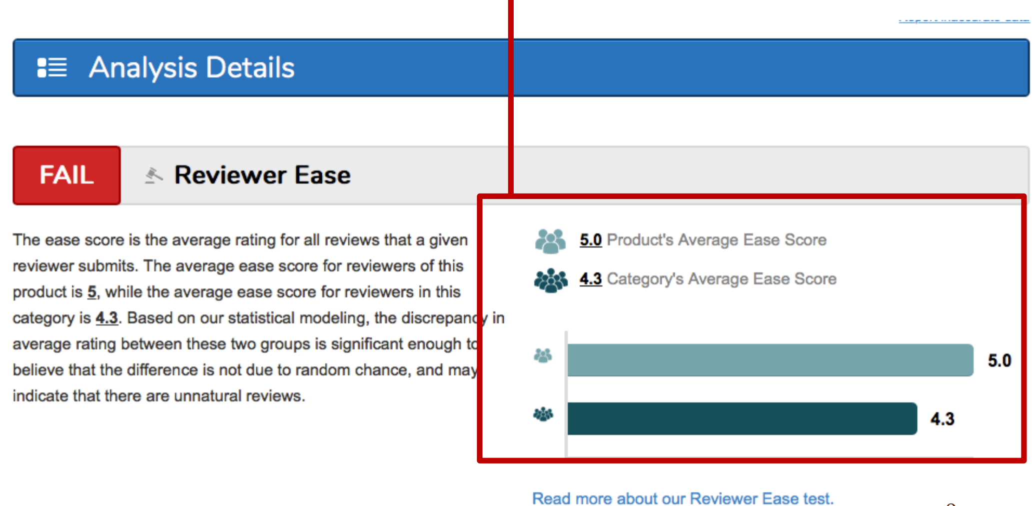
BigData 2018

Behavior based Attacking

- **Accessing knowledge of detector** (Detection websites)

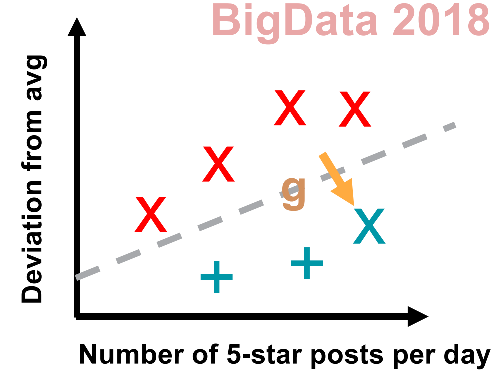






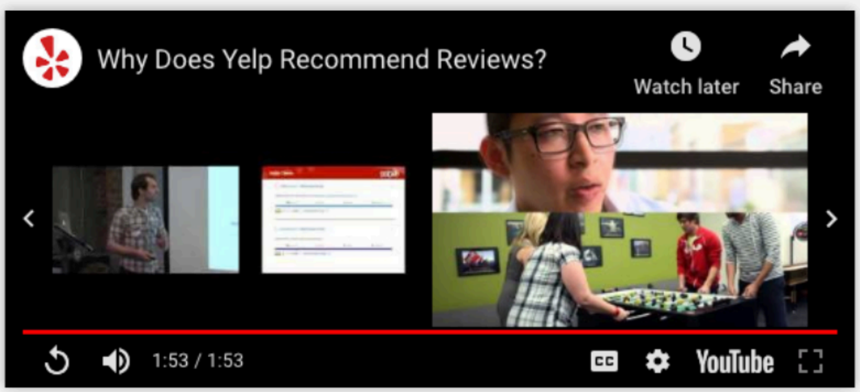
Behavior based Attacking

BigData 2018



- **Accessing knowledge of detector** (Released data)

619 other reviews that are not currently recommended



We use automated software to recommend the reviews we think will be the most helpful to the Yelp community based primarily on quality, reliability and the reviewer's activity on Yelp. Advertisers get no special treatment. The reviews below didn't make the cut and are therefore not factored into this business's overall star rating. Watch the video above or check out our FAQ for more details.

Lindsey L.
Denver, CO
114 friends
17 reviews
9 photos

6/3/2018
Pulled up to the restaurant, cooks were fighting customers on the street, left the restaurant. Mind you... it was 2AM

Jade S.
Concord, CA
3 friends
66 reviews
33 photos

1/29/2018
Its a market that sells food. Food was ok. If your hungry and it's late go for it. If you want good mexican food... keep searching

Noreen C.
Los Angeles, CA
15 friends
38 reviews
8 photos

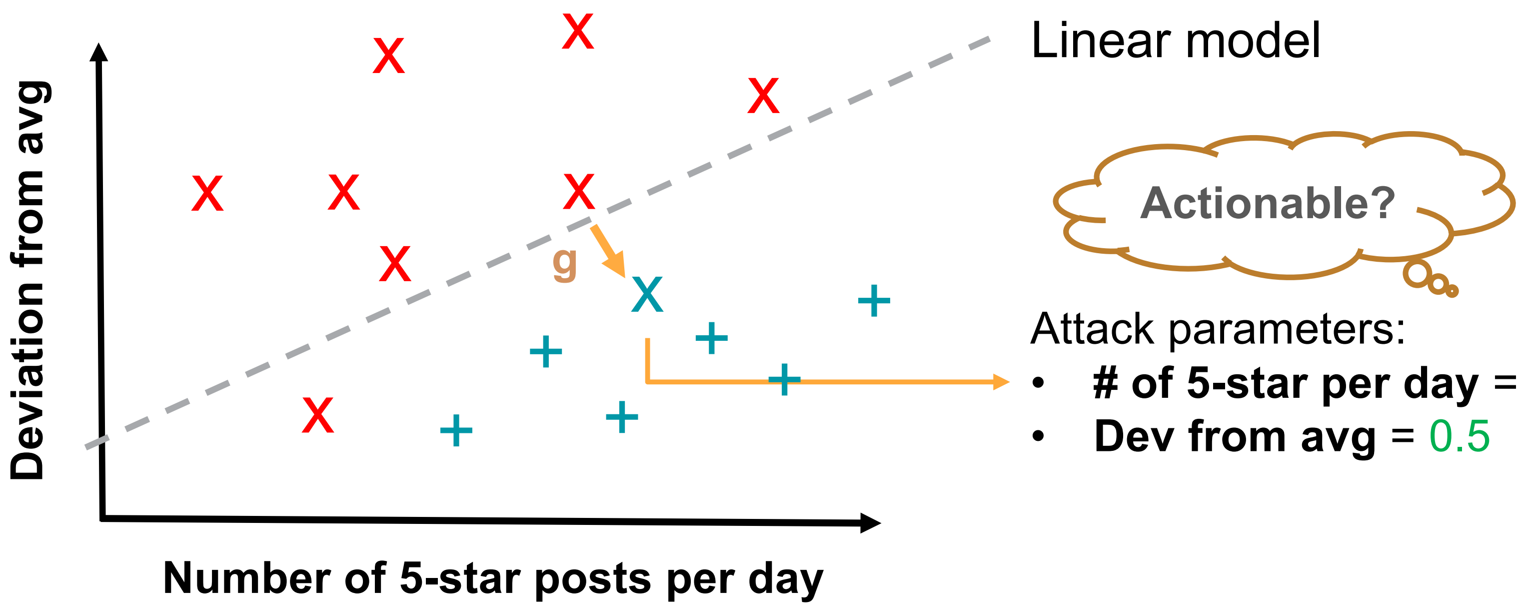
10/23/2018
The best Mexican food in the area. I recommend this place to anyone who wants authentic Mexican food. They have a simple, and well-flavored menu, good service and I would recommend the Quesadillas and Burrito.

Continue reading other reviews that are not currently recommended

BigData 2018

Behavior based Attacking

To defend: need to generate the attacks.



Attack parameters:

- **# of 5-star per day** = 4
- **Dev from avg** = 0.5

Behavior based Attacking

To defend: need to generate the attacks. How?

Actionable attack 1
- post 4 ★★★★★ per day
- post 1 ★★★★☆ per week

Actionable attack 2
- post 3 ★★★★★ per day

Actionable attack 3
- post 4 ★★★★★ per day
- post 1 ★★★★☆ per day
- post 1 ★★★☆☆ per day

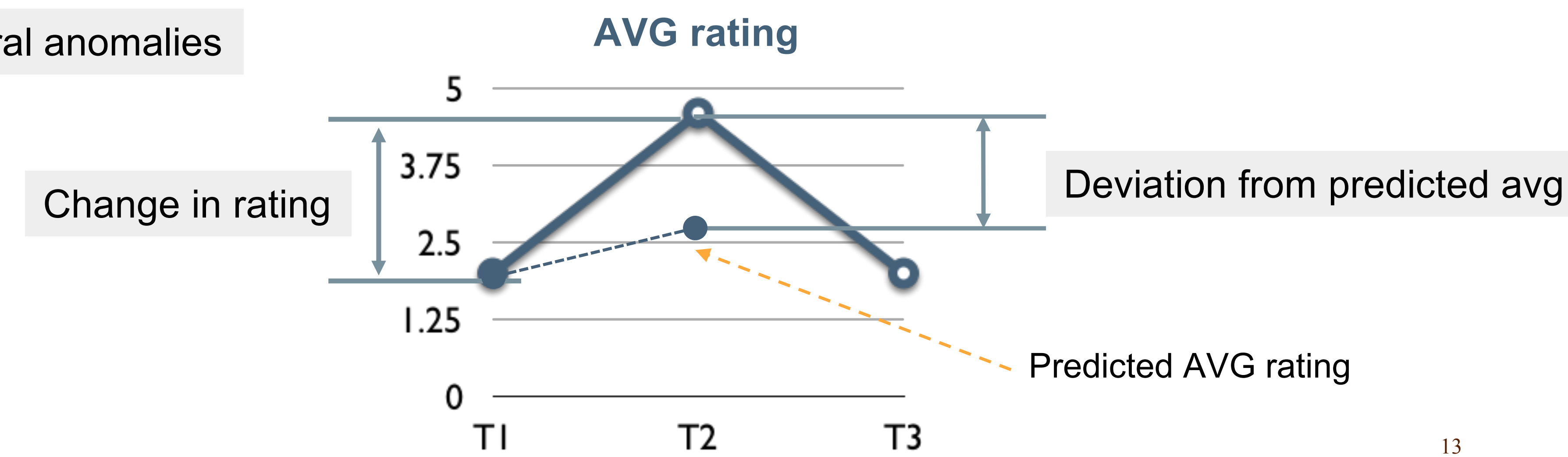
Actionable?

Attack parameters:
- **# of 5-star per day** = 4
- **Dev from avg** = 0.5

Behavior based **Attacking**

Spammer objective function = (risk of being detected) – (profit of spamming)

Temporal anomalies



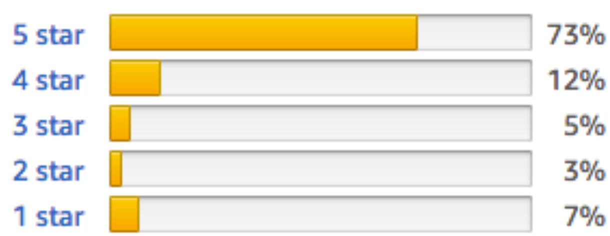
Behavior based Attacking

Spammer objective function = (risk of being detected) – (profit of spamming)

Rating distribution anomaly

$$\mathrm{KL}(\boldsymbol{p}||\bar{\boldsymbol{p}}) = \sum_{i=1}^{5} p_i \log \frac{p_i}{\bar{p}_i}$$

$\bar{\mathbf{p}}$: Background



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

Behavior based Attacking

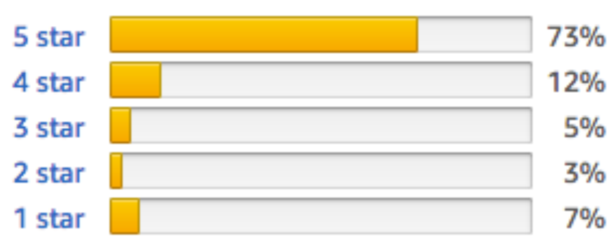
Spammer objective function = (risk of being detected) – (profit of spamming)

Rating distribution anomaly

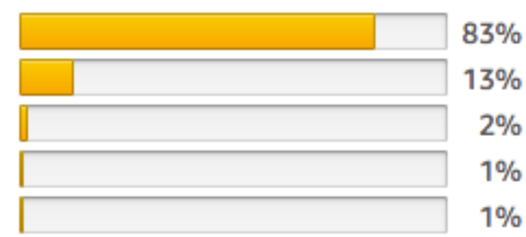
$$\mathrm{EN}(t) = -\sum_{i=1}^{5} p_i(t) \log p_i(t)$$

$$\Delta\mathrm{EN} = \mathrm{EN}(t+1) - \mathrm{EN}(t)$$

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

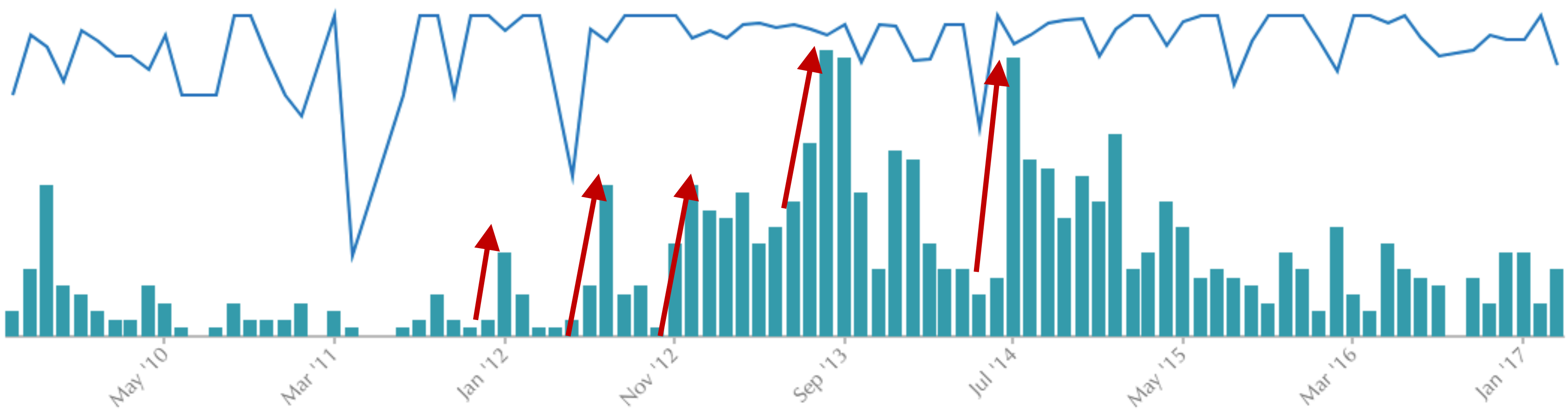


$\mathbf{p}$: Rating dist at time t+1



Behavior based Attacking

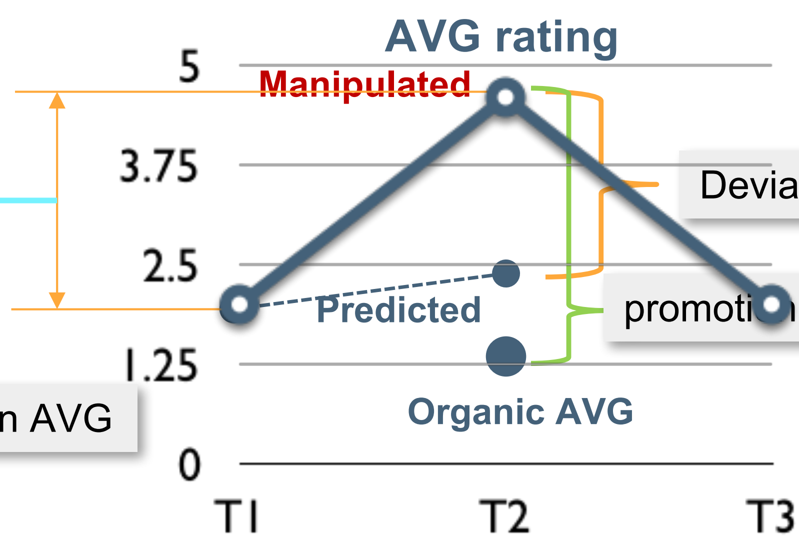
Spammer maximizes [risk of being detected –profit of spamming]



Behavior based Attacking

Find amout of promotion

Cap of all ratings <= 5



Deviation from predicted avg

promotion >= 0

promotion

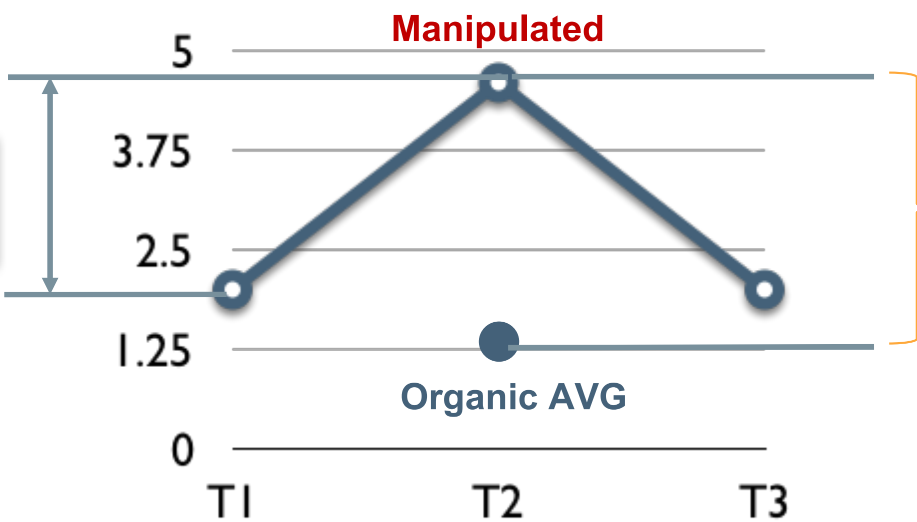
Temporal change in AVG

are set to 80th percentiles of the corresponding changes estimated from the historic data

Behavior based Attacking

Find a proper amount of promotion in AVG rating δ

Large temporal change in AVG?



$\max_{\delta} \quad \delta$

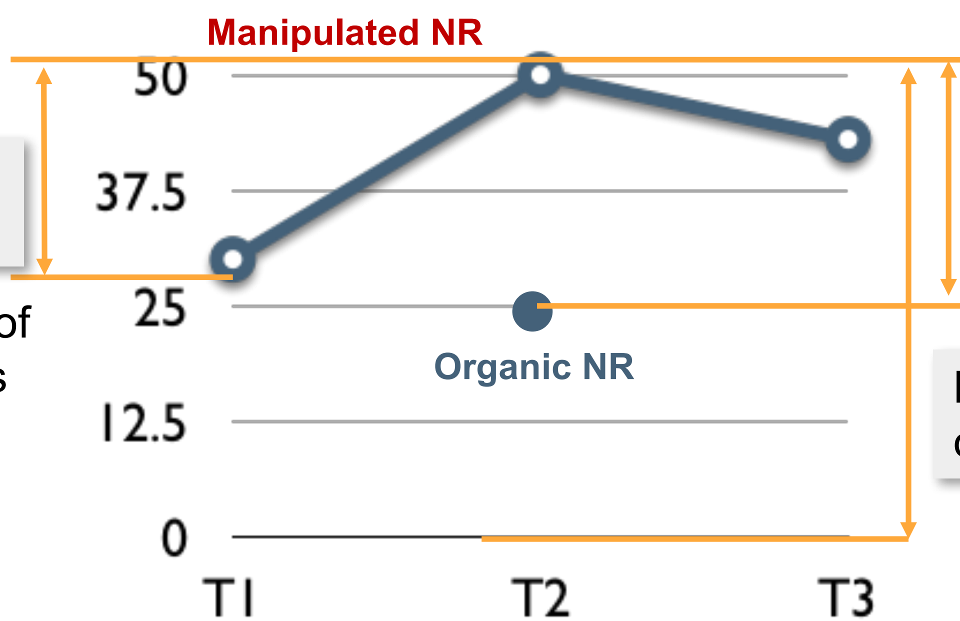
are set to 80th percentiles of the corresponding changes estimated from the historic data

Behavior based Attacking

find a proper number of spamming ratings n_δ

Large incremental in the number of reviews?

<= 80th percentile of historic increments



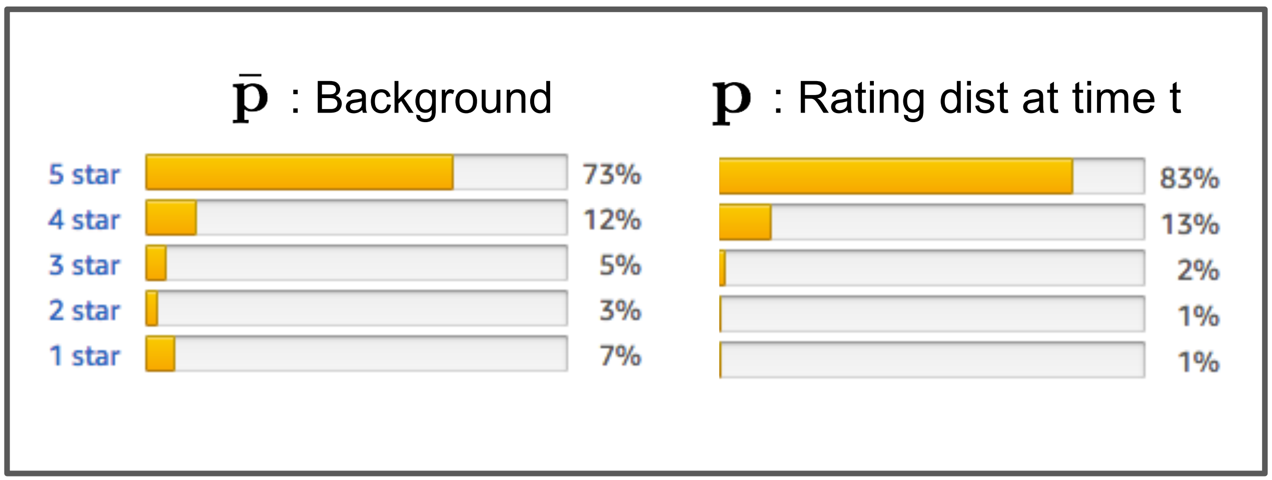
Number of spams n_δ

Large absolute number of reviews?

<= 80th percentile of historic NR

Behavior based Attacking

Compute an evasive rating distribution $\mathbf{p}$



$$\min_{\mathbf{p}} \quad \mathrm{KL}(\mathbf{p}||\bar{\mathbf{p}})$$

Optimal rating distribution found by the dual problem.

Behavior based Attacking

The found evasive rating distribution $\mathbf{p}$

$$n_\delta = 60$$

x 50

x 0

x 0

x 3

x 7

Behavior based Attacking

Flexible attacks generation.

Evade time series based detectors n_δ δ

E1: NR

E2: NR + ΔNR

E3: NR
+ deviation in AVG rating

E4: NR
+ deviation in AVG rating
+ change in AVG rating

Evade both time series and rating distribution based detectors $\mathbf{p}$

E5: KL-DIV

E6: KL-DIV
+ change in entropy

⋮

E9: Max Entropy
+ change in entropy

For short history targets

E-A: NR + ΔNR
+ change in AVG rating
+ Max Entropy

E-B: NR + ΔNR
+ change in AVG rating

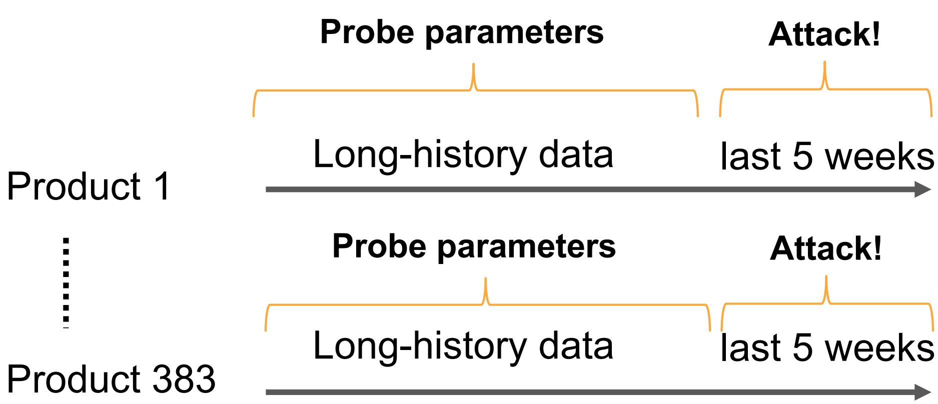
Behavior based Attacking

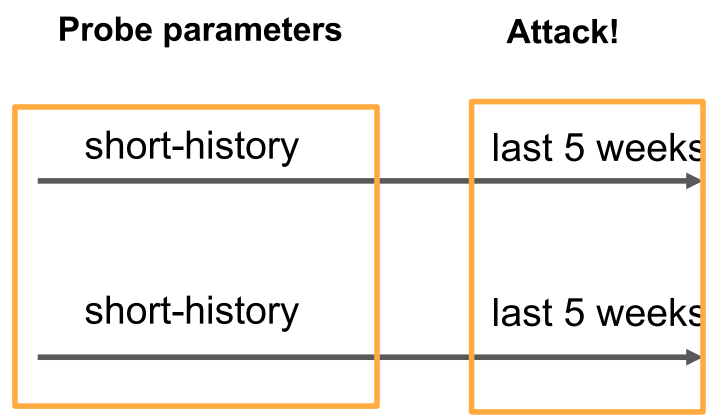
Targets with long review histories

- Products with >= 1,000 reviews
- Reviews span more than 37 months (Yelp) / weeks (Amazon)
- 1,175,088 reviews / 383 products
- 247,117 reviews / 327 restaurants.

Targets with short review histories

- The remaining products / restaurants are used.
- Longitudinal data are too sparse for each target.





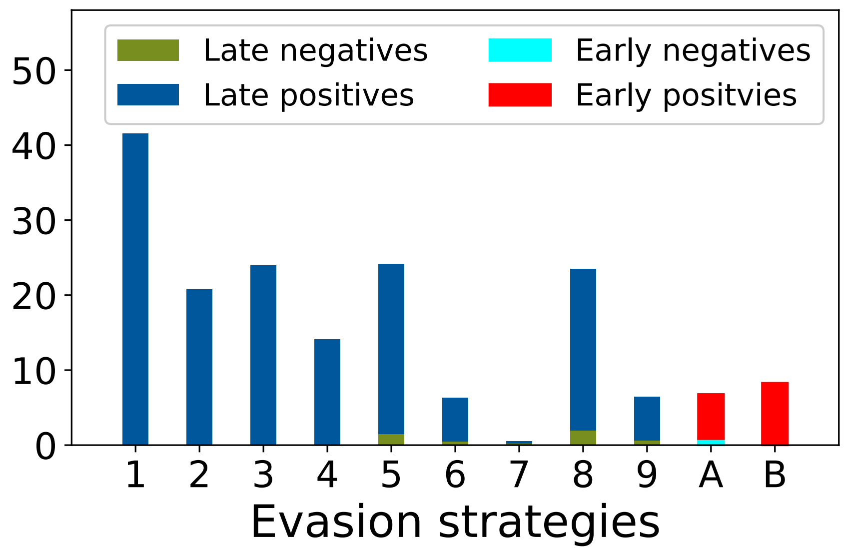
LEHIGH UNIVERSITY

BigData 2018

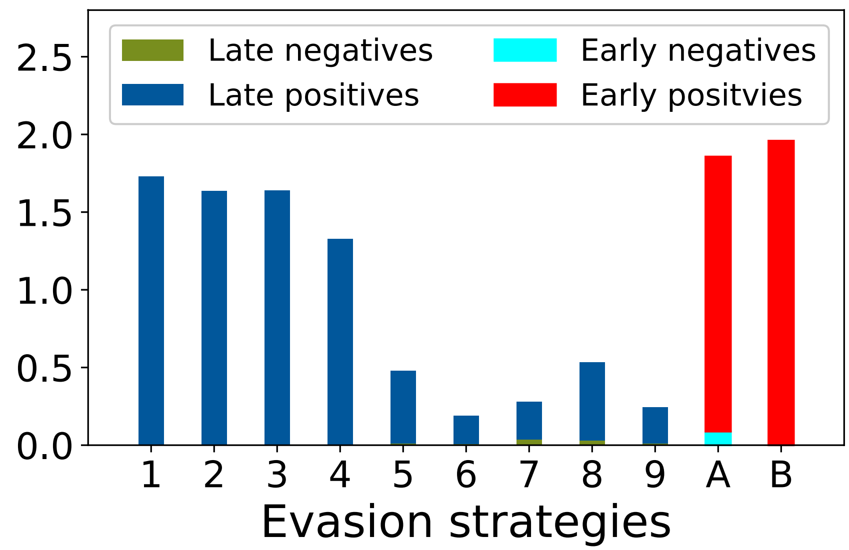
Behavior based Attacking

Average spams posted by each attack

amazon

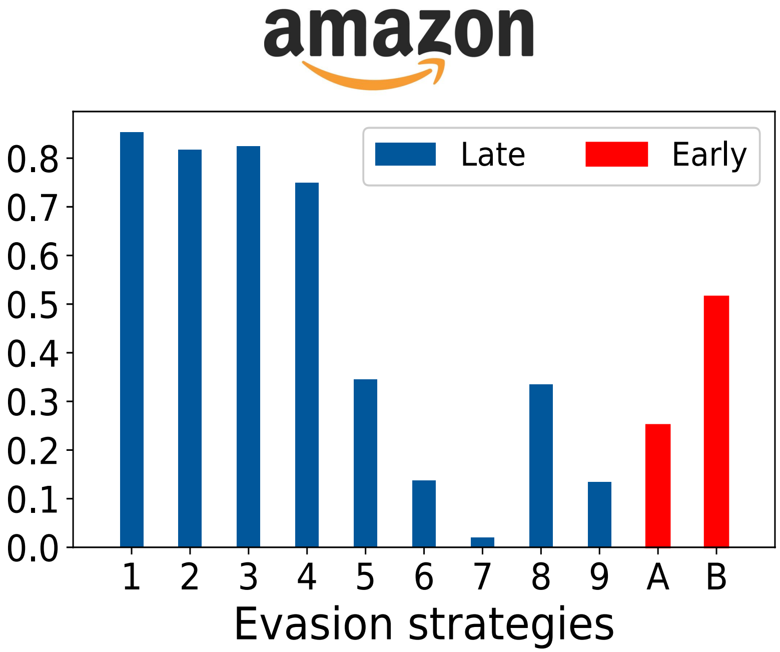


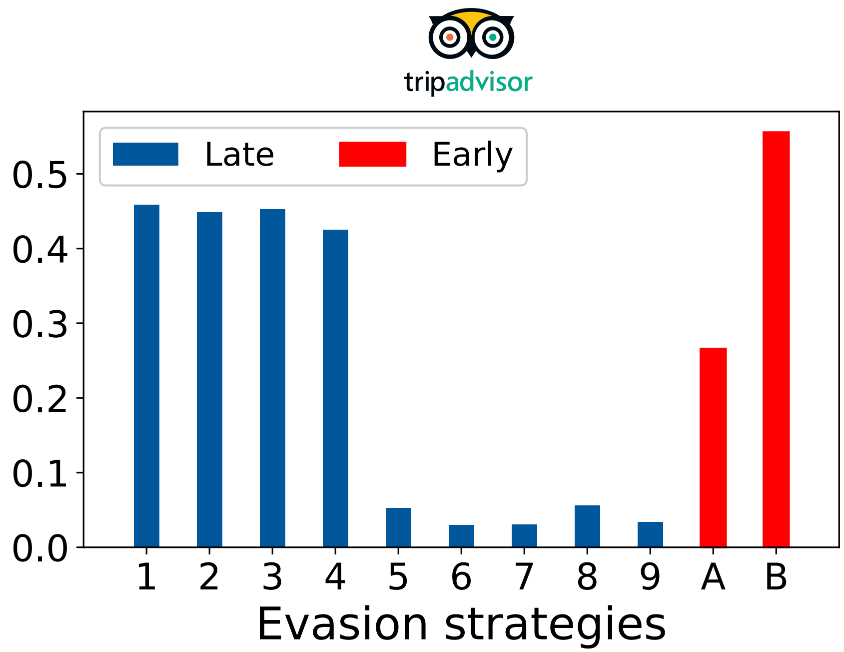
tripadvisor



Behavior based Attacking

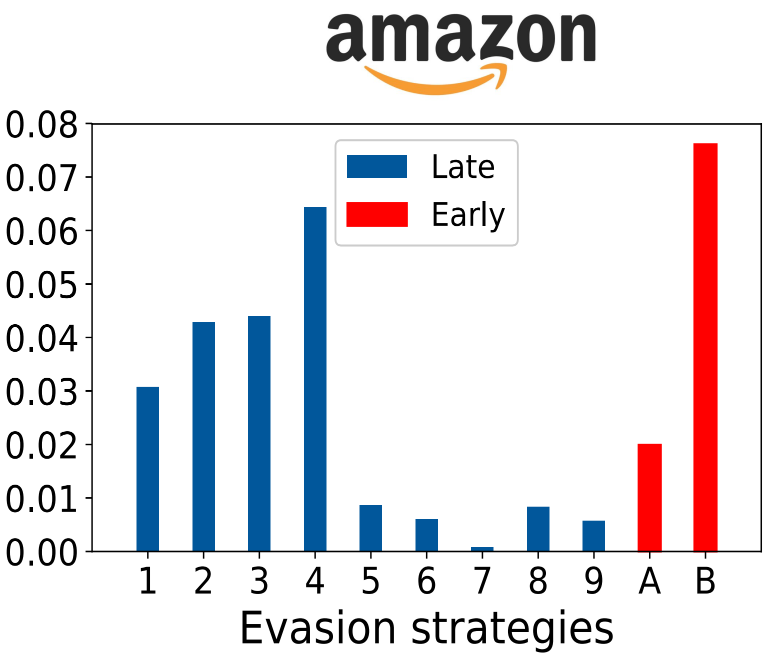
Attacking rate (% of windows can be spammed)

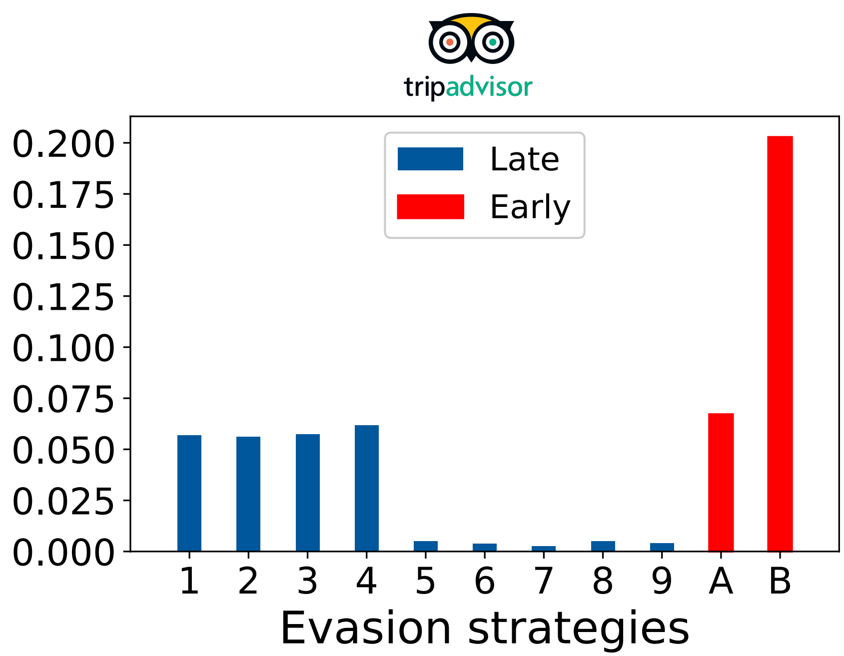




Behavior based Attacking

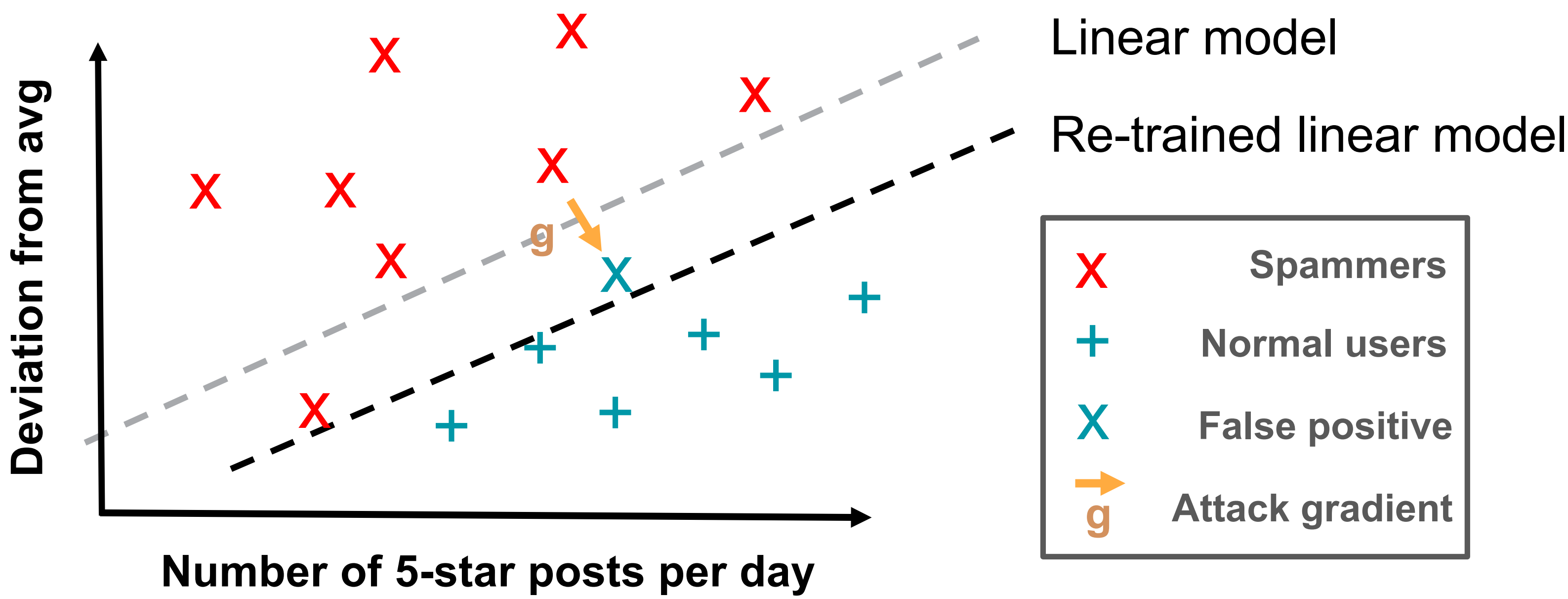
Promotion in ranking per spam



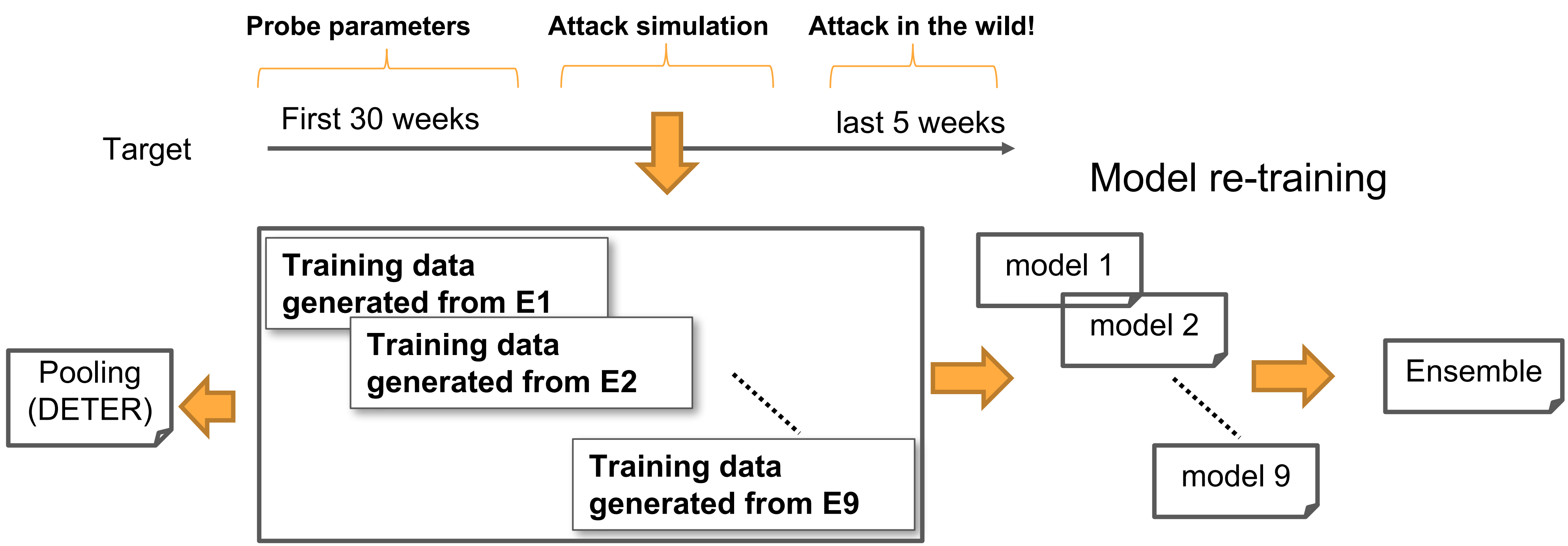


Behavior based Attacking

Secure the detector again



Behavior based Attacking



Behavior based Attacking

Full information detection / evasion game: single spammer

Game 1

Detector

Spammer	KL-DIV	# of reviews
E1	0.9	0.8
E2	0.8	0.9

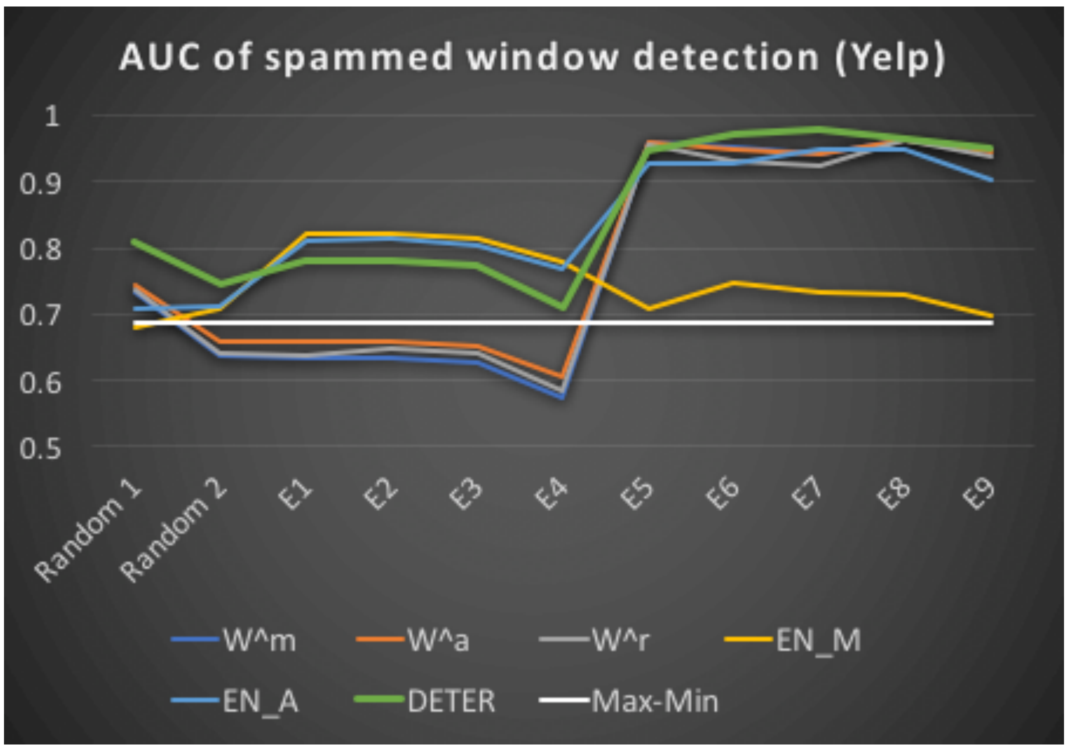
Game 2

Detector

Spammer	KL-DIV	# of reviews
E1	0.9	0.8
E2	equilibrium 0.8	0.7

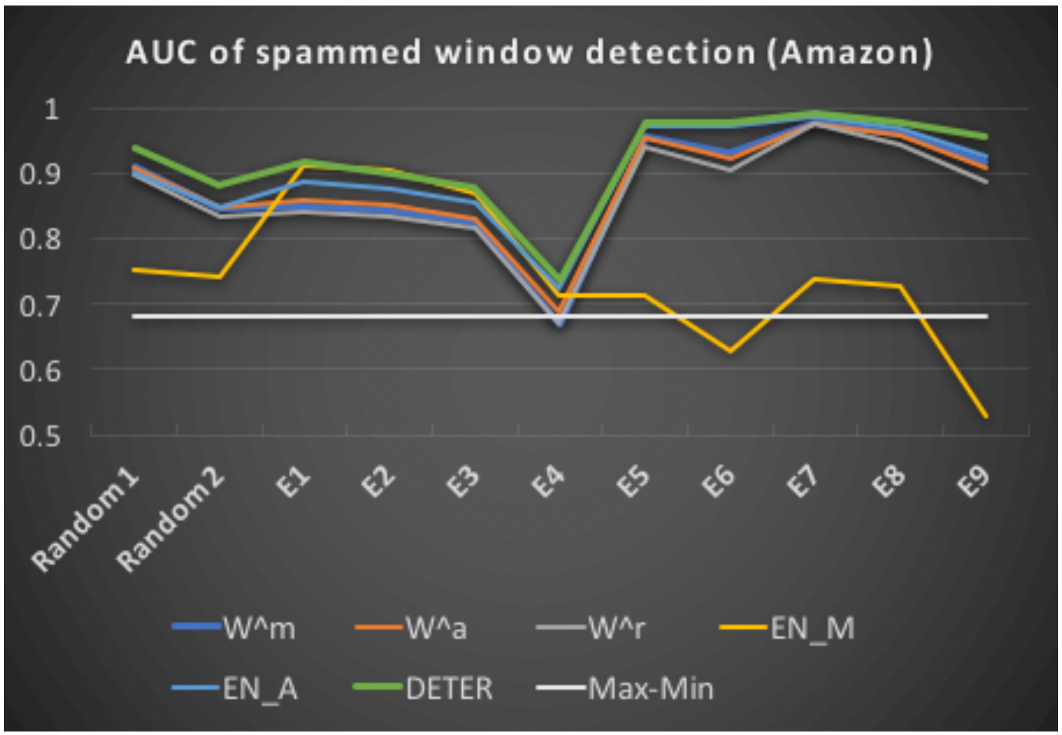
BigData 2018

Behavior based **Attacking**



W^m	Max of signals
W^a	Avg of signals
W^r	Randomly selection
EN_A	Re-train avg
EN_M	Re-train Max
DETER	Re-train Pool
Max-min	Game equilibrium

Behavior based Attacking



W^m	Max of signals
W^a	Avg of signals
W^r	Randomly selection
EN_A	Re-train avg
EN_M	Re-train Max
DETER	Re-train Pool
Max-min	Game equilibrium

Behavior based Attacking

- Unsupervised
- Attack agnostic
- Simple and good performance
- Good for long and short review histories
- Can secure the detector!
- Source codes and data avaiable at:

https://bitbucket.org/Doris_Ge/bigdata18_spam_detection

http://www.cse.lehigh.edu/~sxie/codes.html

Thank you

LEHIGH
UNIVERSITY