Discovering Networks of Fraudulent Users in Online Review Platforms

Leman Akoglu

University of Texas Austin
McCOMBS School of Business

October 18, 2017

Carnegie Mellon University
Heinz College

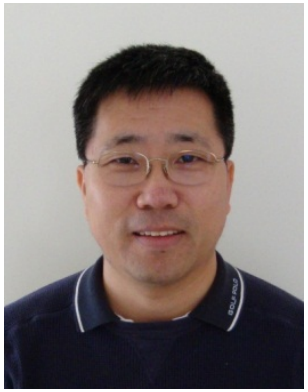
Acknowledgments

Shebuti Rayana
SBU

Junting Ye
SBU

Yejin Choi
UW

Christos Faloutsos
CMU



Bing Liu
UIC

Online reviews

- major source of information

Advertisement

Reviews

Love it
By Preston Rhubee on June 17, 2015
Good for casual wear not workouts
By Jake Klinvex on February 21, 2015
Color
Accuracy not included
By john klett on January 27, 2015
Color: Black | Size: Large (6.2-7.6 in) | Verified Purchase
Couldn't be more disappointed! I followed the directions, better and it was never accurate.

Nielsen Media Research: 68% people rely on online reviews
93% of visitors make a purchase after visiting Yelp

- important to businesses!

+1 star-rating increases revenue by 5-9%

- *Reviews, Reputation, and Revenue: The Case of Yelp.com Michael Luca. Harvard Business School 2011*
- *Do online reviews matter?—An empirical investigation of panel data. W. Duan, B. Gu, A. Whinston. DSS 45 (4), 2008*

Online reviews –which ones do you trust?

A Google User reviewed 7 months ago
Overall 3 / 3
Just got home today and noticed what a great job these guys did. Loved the service! Happy customer! A+++

A Google User reviewed a year ago
Overall 3 / 3
this place is FANTASTIC! they are efficient and organized and the doctor i had (Dr Goldstein) was phenomenal. They knew exactly what to do. BEST urgent care I've ever been to!!!!

Google

yelp

0
2
Evelyn E.
Atherton, CA

10/6/2012

The maple mustard tempeh sandwich is possibly one of the best sandwiches I've ever eaten, vegan or not. It's the perfect balance of savory and sweet, with a good smear of garlic aioli (but not too much!), kale that is still toothsome, and tomato and onion to balance it out. Don't miss out... seek out the truck!

Was this review ...? Useful Funny Cool

BRENDA T
UNITED KINGDOM
1 review

"FANTASTIC HOTEL"
Reviewed October 2, 2012 NEW

tripadvisor

We stayed at the James (4 ladies) in September and from the minute we checked in the staff and service were excellent. Perfect location for everything that Chicago has to offer. The concierge was most helpful with anything that we needed to know. This was out first trip to Chicago but I don't think it will be the last. So...

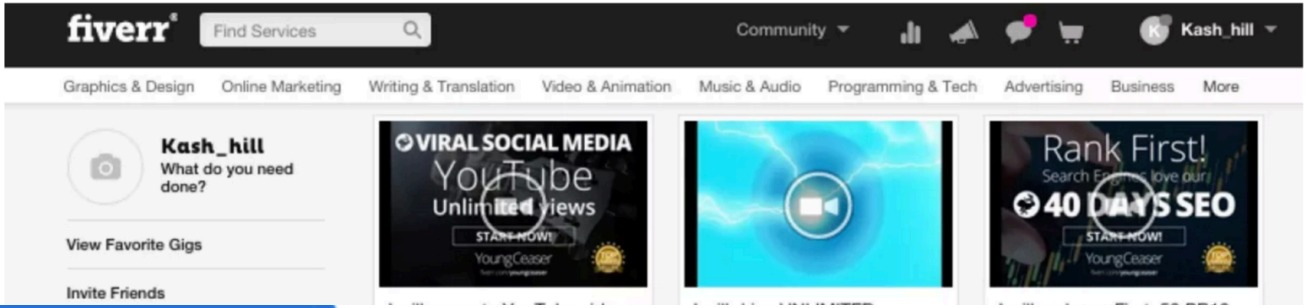
amazon

Sloppy.
The book is a promising reference concept, but the execution is somewhat sloppy. Whatever generator they used was not fully tested. The bulk of each page seems random enough. However at the lower left and lower right of alternate pages, the number is found to increment directly.

How prevalent is the problem?

- Opinion spam is everywhere!
  - 14~20% in Yelp; [Mukherjee et al., ICWSM 2013]
  - 2~6% in Orbitz, Priceline, Expedia, Tripadvisor, etc. [Ott et al., WWW 2012]

How prevalent is the problem?



Employers ask people to...

- Blog about your product
- Post reviews to Websites &
- Add you to Facebook
- Become fan of your group
- Follow you on Twitter
- Digg your website
- and much more...

Tripadvisor banned from claiming its reviews are real

Tripadvisor, the travel review website, has been banned from claiming that all of its hotel and restaurant reviews are real.

How hard is the problem?

Which review is fake?

Date of review: Jun 9, 2006

4 people found this review helpful

My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The location for the shopping great, since I love to shop I could walk for hours!! We will definatly be back to Chicago and we will for sure be back to the James Chicago.

Date of review: Jun 9, 2006

4 people found this review helpful

I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.

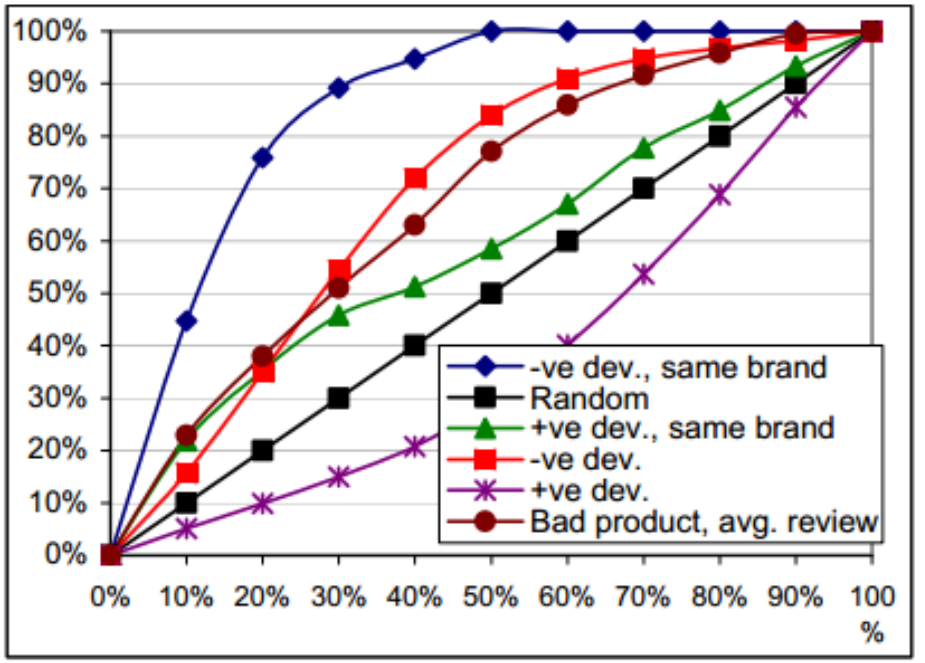
Human accuracy is slightly better than random.*

			TRUTHFUL			DECEPTIVE		
		Accuracy	**P**	**R**	**F**	**P**	**R**	**F**
HUMAN	JUDGE 1	**61.9%**	57.9	87.5	**69.7**	74.4	36.3	48.7
	JUDGE 2	56.9%	53.9	**95.0**	68.8	**78.9**	18.8	30.3
	JUDGE 3	53.1%	52.3	70.0	59.9	54.7	36.3	43.6

* *Finding deceptive opinion spam by any stretch of the imagination.*
M. Ott, Y. Choi, C. Cardie, J. T. Hancock. ACL, 2011.

Behavioral clues [Jindal & Liu WWW'08]

- Extract (36) features
  - #feedbacks, length, time-order, #brand-names, %capitals, avg. reviewer rating, price, sales rank, …
- Train logistic regression classifier
  - + examples: duplicate reviews,
  - **AUC: 78%** (CV)
  - **Lift curves** for non-duplicates:



Linguistic clues

[Ott+ ACL'11]

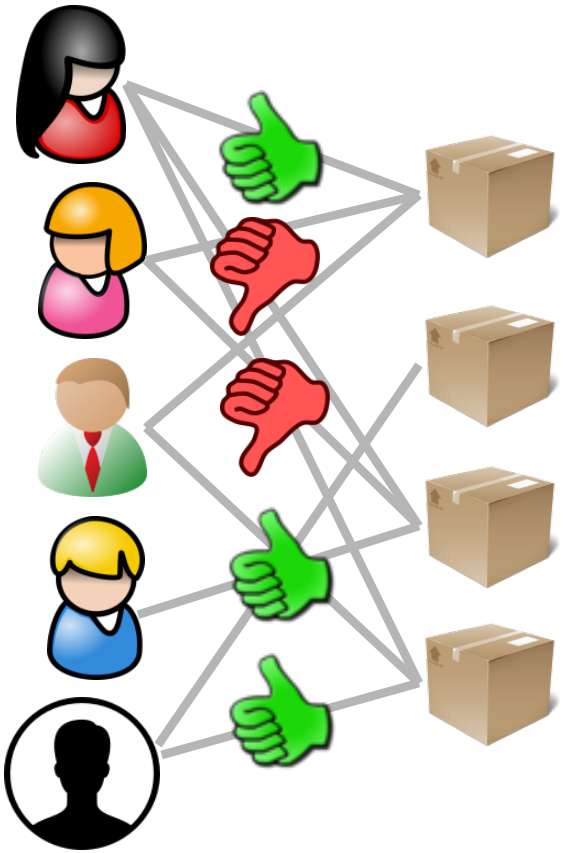
- Extract features based on
  - **Part-of-Speech** (superlatives, past tense, …)
  - **Psycholinguistic** (#words/sentence, rate of misspell, references to money/religion, agreement words, …)
  - **Text** (n-grams)
- Train SVM
  - positive examples from AMT

Approach	Features	Accuracy
GENRE IDENTIFICATION	POS_{SVM}	73.0%
PSYCHOLINGUISTIC DECEPTION DETECTION	$LIWC_{SVM}$	76.8%
TEXT CATEGORIZATION	$UNIGRAMS_{SVM}$	88.4%
	$BIGRAMS^{+}_{SVM}$	89.6%
	$LIWC{+}BIGRAMS^{+}_{SVM}$	**89.8%**
	$TRIGRAMS^{+}_{SVM}$	89.0%

TRUTHFUL	DECEPTIVE
-	chicago
...	★ my
★ on	hotel
★ location	,_and
)	luxury
$allpunct_{LIWC}$	experience
★ floor	hilton
(	business
the_hotel	vacation
★ bathroom	★ i
★ small	spa
helpful	looking
$	while
hotel_.	husband
other	★ my_husband

Review networks

- who-reviews-what



Roadmap

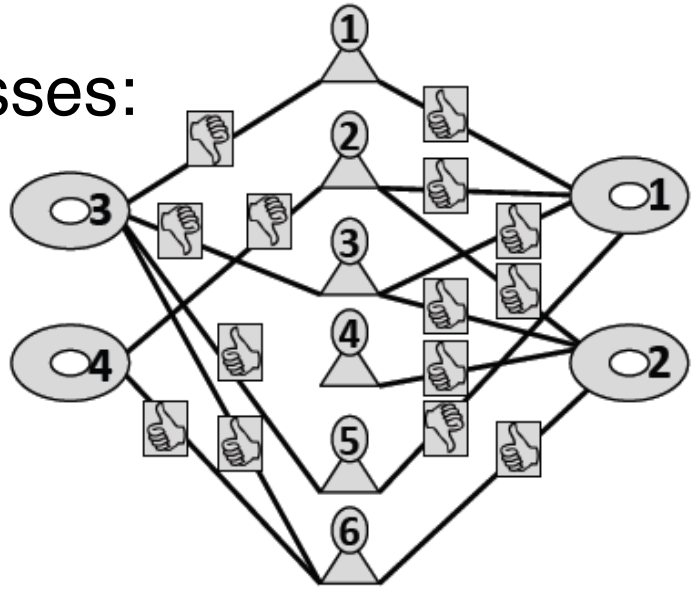
- Intro/Motivation
- Fraud detection
  - Opinion fraud
    - *Network effects* [Akoglu+ ICWSM'13]
    - *Network clues for spammer groups*
    - *Networks & meta-data*
    - *Temporal analysis*
- Open challenges

Network effects

- Given
  - user-product review network (bipartite)
  - review sentiments (+: thumbs-up, -: thumbs-down)
- Classify
  - objects into **type-specific** classes:
    users: `honest' / `fraudster'
    products: `good' / `bad'
    reviews: `genuine' / `fake'

No meta data!

(e.g., timestamp, review text)



Opinion Fraud Detection in Online Reviews using Network Effects
Leman Akoglu, Rishi Chandy, Christos Faloutsos ICWSM 2013

Formulation: Collective Classification

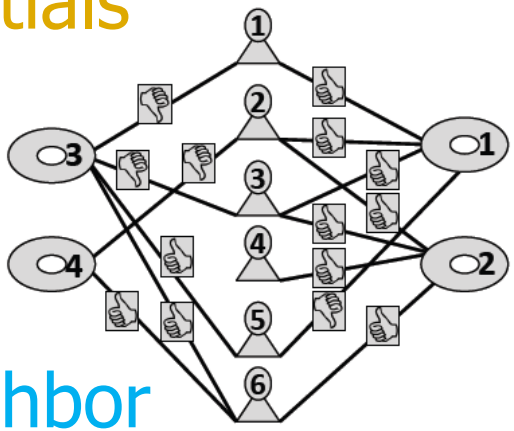
$$\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{Y_i \in \mathcal{Y}^{\mathcal{V}}} \phi_i(y_i) \prod_{e(Y_i^{\mathcal{U}}, Y_j^{\mathcal{P}}, s) \in \mathcal{E}} \psi_{ij}^{s}(y_i, y_j)$$

edge sign

Node labels as random variables

compatibility potentials

$$\phi_i^{\mathcal{U}}(y_i) = \psi_i^{\mathcal{U}}(y_i) \prod_{e(Y_i^{\mathcal{U}}, X_j^{\mathcal{P}}, s) \in \mathcal{E}} \psi_{ij}^{s}(y_i)$$



prior belief

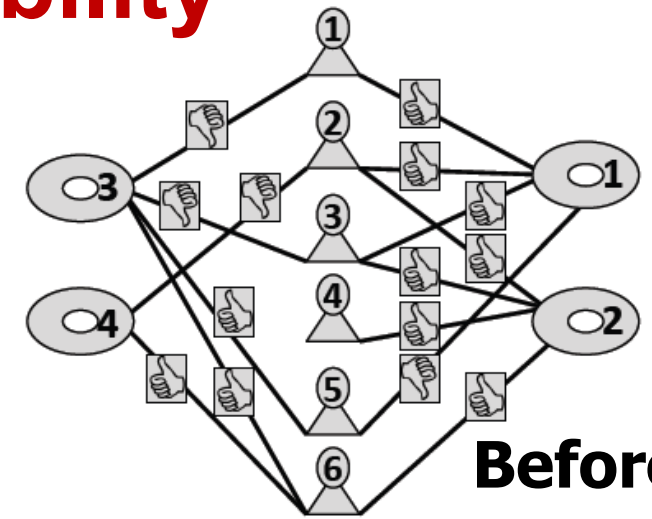
observed neighbor potentials

Formulation: Compatibility

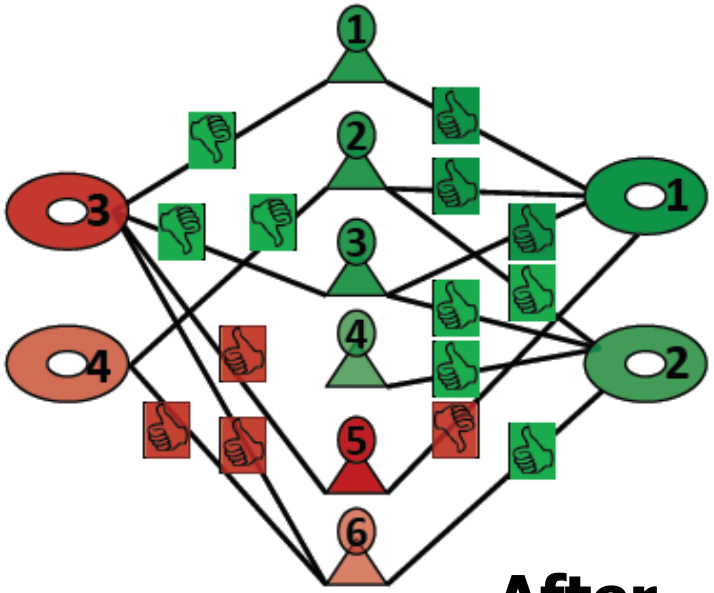
User		Product
honest	$\xrightarrow{-}$	*bad*
honest	$\xrightarrow{+}$	*bad*

s: -	**Products**	
Users	Good	Bad
Honest	ϵ	$1-\epsilon$
Fraud	$1-2\epsilon$	2ϵ

s: +	**Products**	
Users	Good	Bad
Honest	$1-\epsilon$	ϵ
Fraud	2ϵ	$1-2\epsilon$



Before



After

Inference: Loopy belief propagation

- Invented in 1982 [Pearl] to calculate marginals in Bayes nets.
- Also used to estimate marginals (=beliefs), or most likely states (e.g. MAP) in MRFs
- Iterative process in which neighbor variables "talk" to each other, passing **messages**

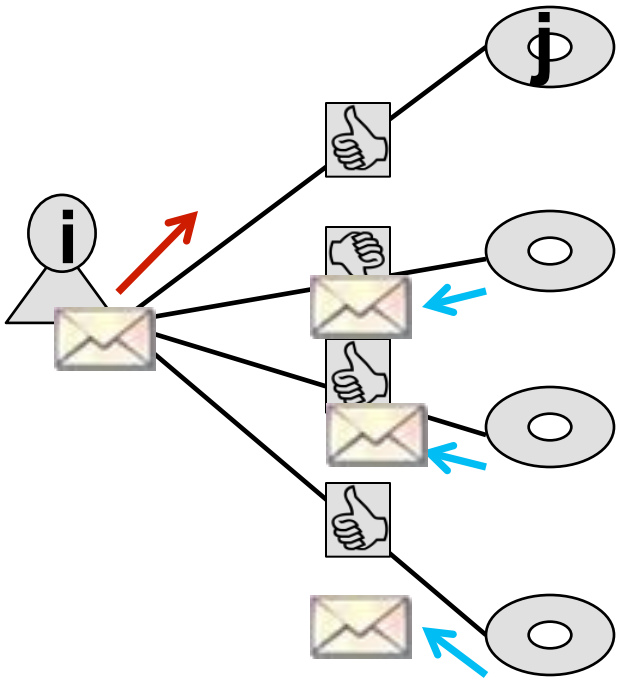
"I (variable x1) believe you (variable x2) belong in these states with various likelihoods…"

- When consensus reached, calculate belief

Inference

Repeat for each node:

$$m_{i\to j}(y_j) = \alpha_1 \sum_{y_i \in \mathcal{L}^{\mathcal{U}}} \psi_{ij}^{s}(y_i, y_j)\, \phi_i^{\mathcal{U}}(y_i) \prod_{Y_k \in \mathcal{N}_i \cap \mathcal{Y}^{\mathcal{P}} \setminus Y_j} m_{k\to i}(y_i), \quad \forall y_j \in \mathcal{L}^{\mathcal{P}}$$



$m_{k\to i}(y_i)$

At convergence:

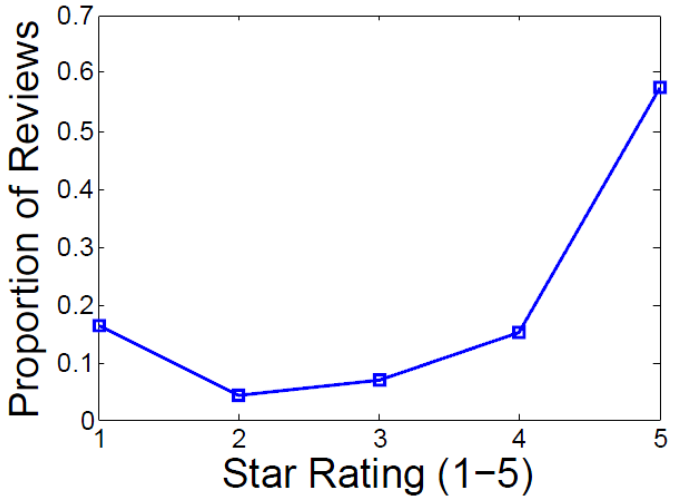
$$b_i(y_i) = \alpha_2 \phi_i^{\mathcal{U}}(y_i) \prod_{Y_j \in \mathcal{N}_i \cap \mathcal{Y}^{\mathcal{P}}} m_{j\to i}(y_i)$$

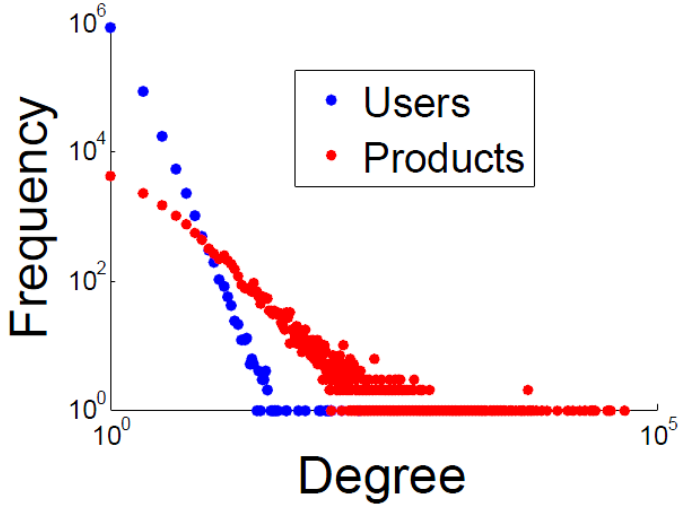
$$\text{score}_i(fraud) = b_i(y_i : fraud)$$

$$\text{score}_{e(i,j)}(fake) = m_{j\to i}(y_i : fraud)$$

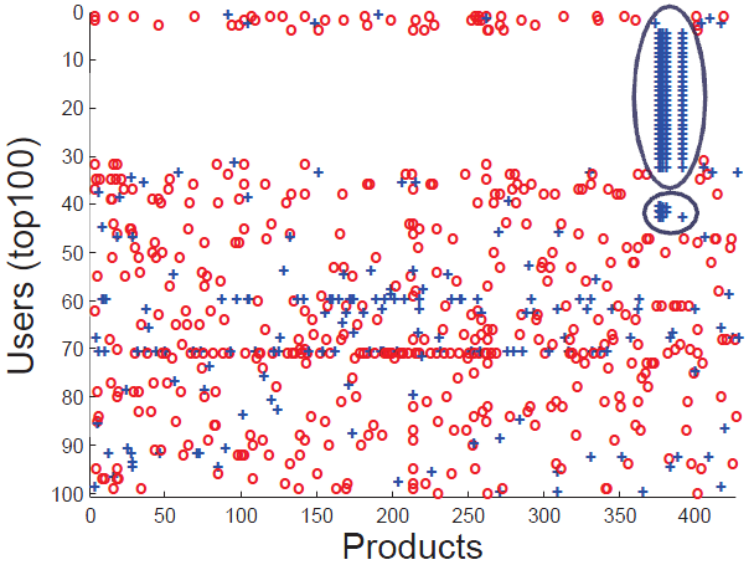
Review data (SWM)

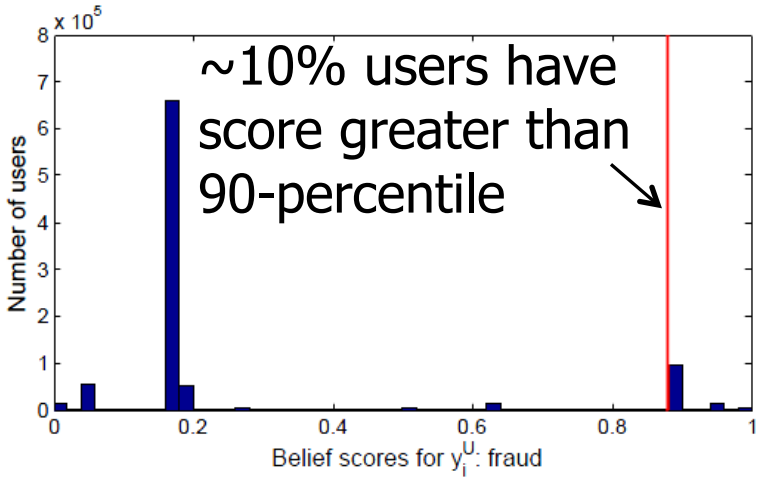
- 966842 users, 15094 software products (apps), 1132373 reviews
- ratings: 1 (worst) to 5 (best)

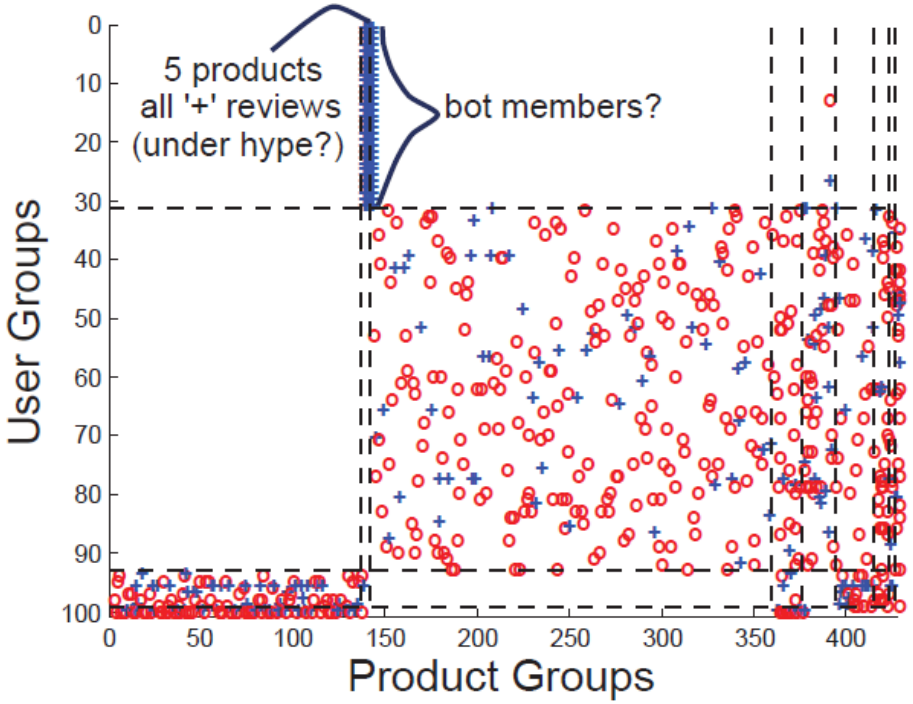




Top scorers





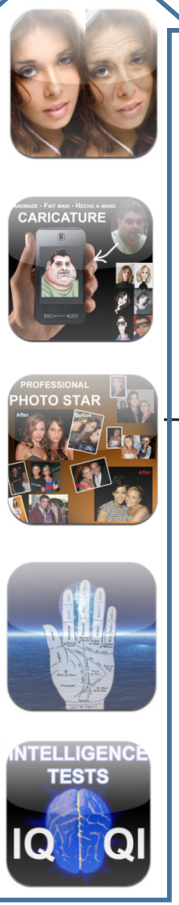


+ positive (4-5) rating

o negative (1-2) rating

‘Fraud-bot’ member reviews

Same developer! Duplicated text! Same day activity!



Suggestion — tebayor — Version 1 - Sep 13, 2011
My face...great, but what I really want to know about are my breasts!

Satisfied — tebayor — Version 1 - Sep 17, 2011
I'm extremely satisfied with my caricature. Well done.

So great! — tebayor — Version 1 - Sep 17, 2011
Seems a real one. It's a lot of fun

Good app — tebayor — Version 1.0 - Sep 17, 2011
Not as stupid as I thought!

Really happy — tebayor — Version 1.0 - Sep 17, 2011
It's actually a personal one on one reading. I'm really happy with the outcome.

Rides — merquezcito — Version 1 - Sep 1, 2011
I just discovered old

A beautiful gift — merquezcito — Version 1 - Sep 1, 2011
I got one made as a gift for my god-daughter

Perfect — merquezcito — Version 1 - Sep 1, 2011
Nothing else to say

Good app — merquezcito — Version 1.0 - Sep 1, 2011
Not as stupid as I thought!

Ooh la la — Muquiwara78 — Version 1 - Aug 25, 2011
Qd I see the head of my old guy, it scares me. At the same time, I will be like!

Satisfied — Muquiwara78 — Version 1 - Aug 25, 2011
I'm extremely satisfied with my caricature. Well done.

ha ha ha — Muquiwara78 — Version 1 - Aug 25, 2011
I want to be that close of J Lo!

Simple IQ test — Muquiwara78 — Version 1.0 - Aug 25, 2011
Good app.

Thanks — Muquiwara78 — Version 1.0 - Aug 25, 2011
I'm extremely satisfied

Deadly — manjuli — Version 1 - Oct 23, 2011
It does not care a shot of being old. Fortunately it is not immediately

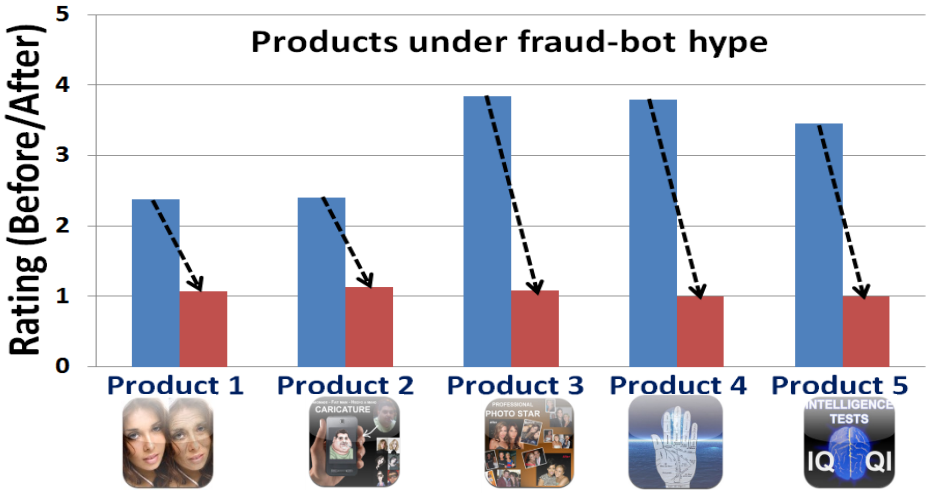
Good job — manjuli — Version 1.1 - Dec 3, 2011
I really like this app. I want another!!!

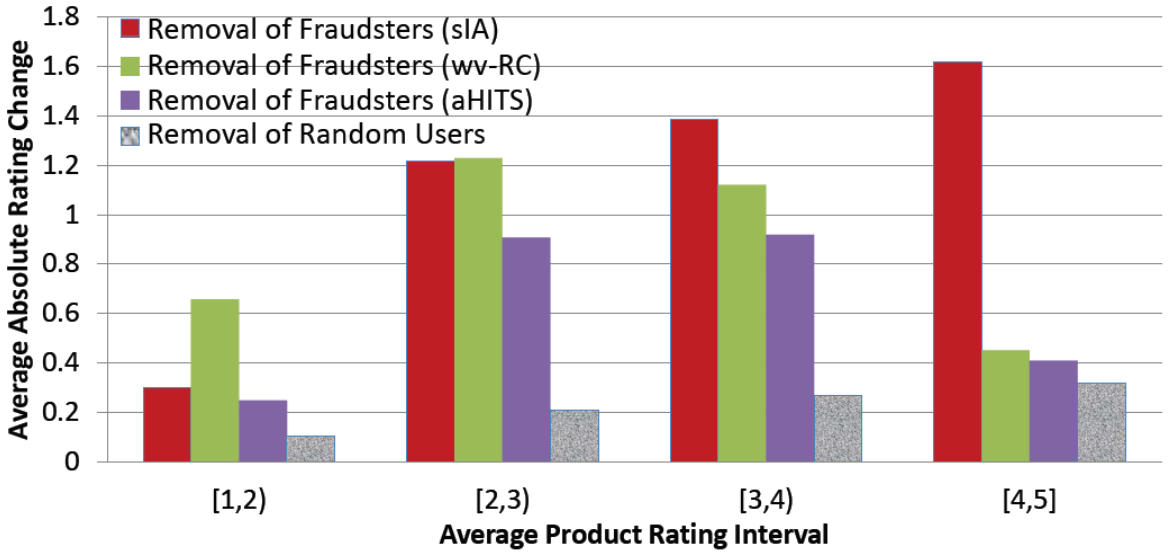
ha ha ha — manjuli — Version 1 - Oct 23, 2011
I want to be that close of J Lo!

Simple IQ test — manjuli — Version 1.0 - Oct 23, 2011
Good app.

Thanks — manjuli — Version 1.0 - Oct 23, 2011
I'm extremely satisfied

Top scorers removed





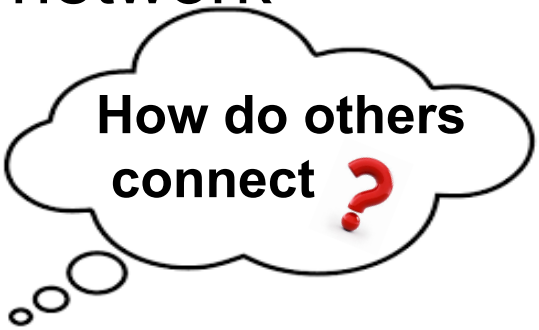
Roadmap

- Intro/Motivation
- Fraud detection
  - Opinion fraud
    - *Network effects* [Akoglu+ ICWSM'13]
    - *Network clues for spammer groups* [Ye & Akoglu PKDD'15]
    - *Networks & meta-data*
    - *Temporal analysis*
- Open challenges

Spammer groups

- Spammers organize into **groups**:
  - Impact maximized – dominate the sentiments
  - Effort / workload shared
  - Easier to hide: nobody stands out

- **Key**: spammers **unaware** of global network



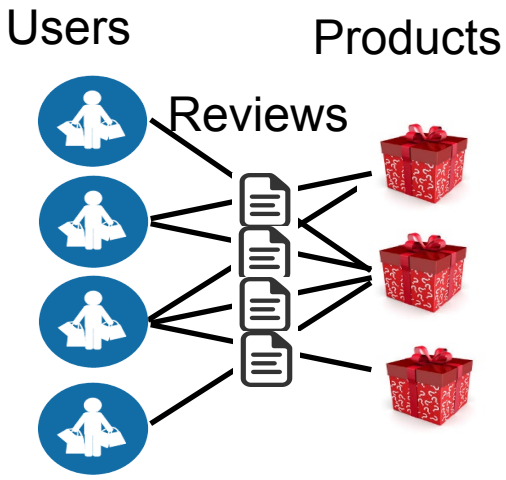
Discovering Opinion Spammer Groups by Network Footprints
Junting Ye and Leman Akoglu ECML/PKDD, 2015.

Spammer groups

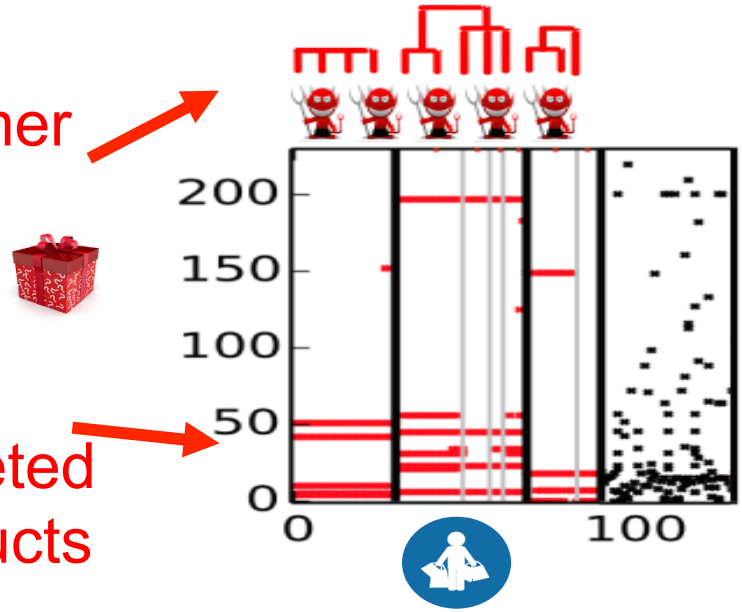
- Input:
  - A user-product bipartite review network

$$G = (U, P, E)$$

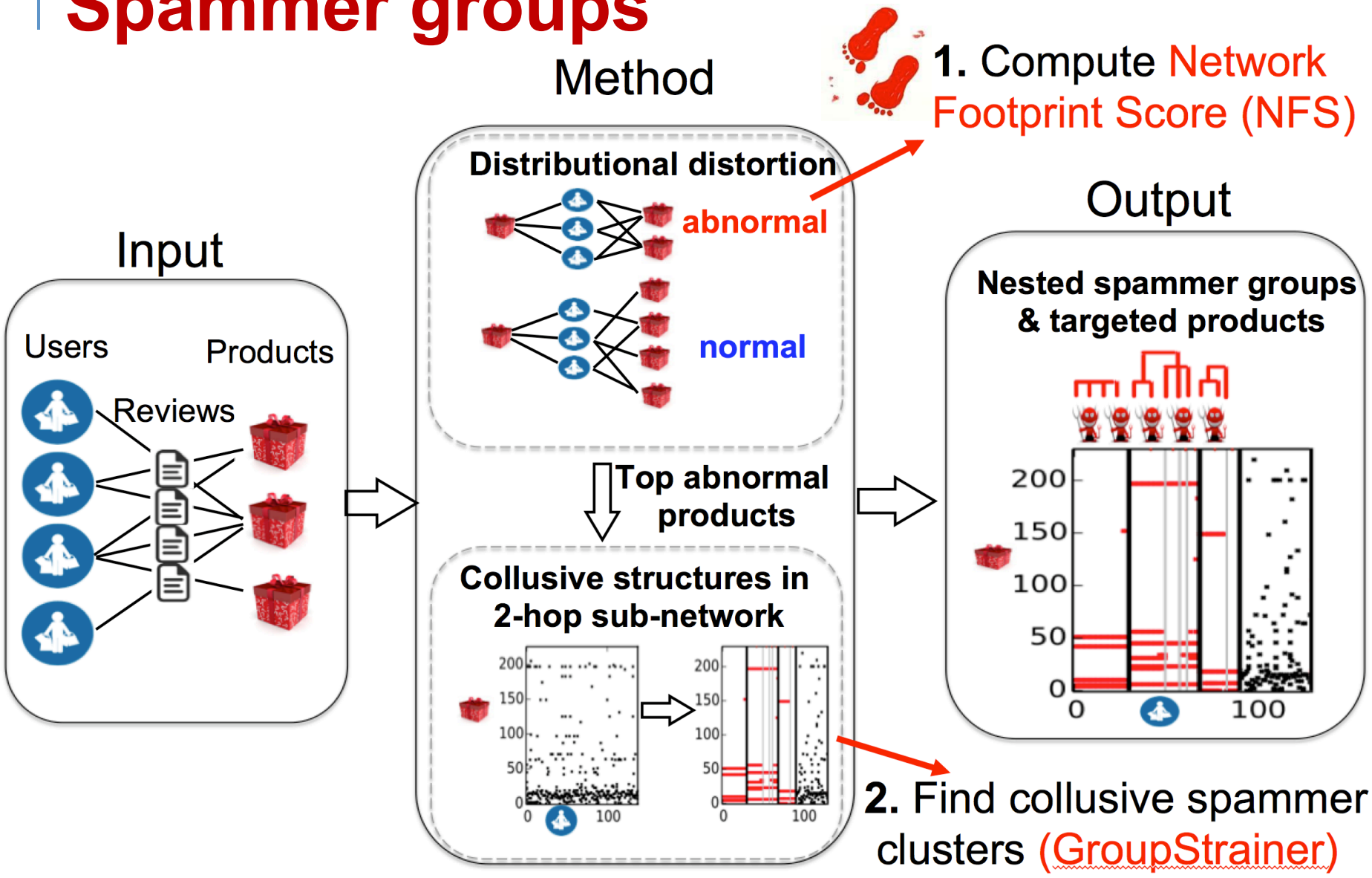
U: the set of users, P: set of products, E: set of review links.



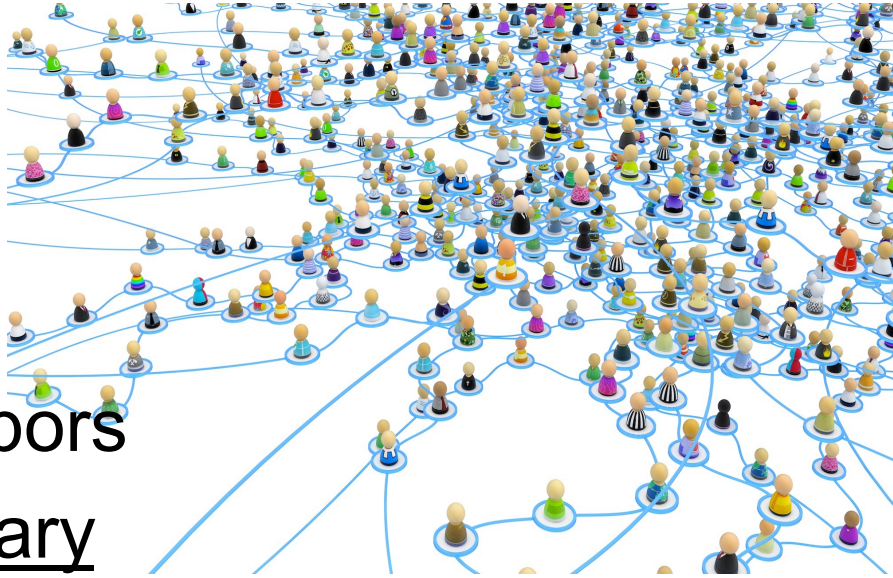
- Output:
  - Nested spammer groups;
  - Targeted products;



Spammer groups

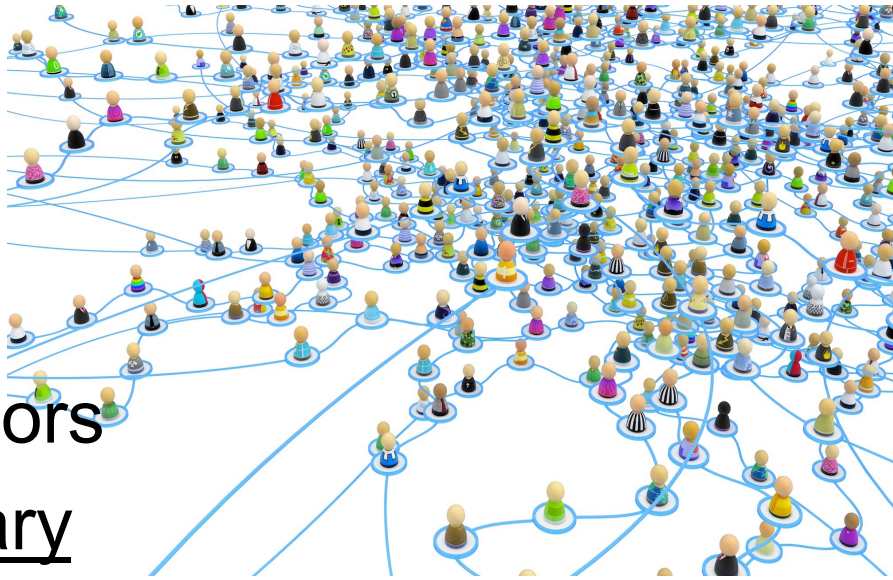


Network clues



- **Observation 1:**
  Neighbor diversity
  - Diverse group of neighbors
  - → Neighbor centralities vary
    (not concentrated)

Network clues



- **Observation 1:**
  Neighbor diversity
  - Diverse group of neighbors
  - → Neighbor centralities vary (not concentrated)

- Quantification:

i: index of products

p: centrality density histogram

$$H_c(i) = -\sum_{k=1}^{K} p_k^{(i)} \log p_k^{(i)}$$

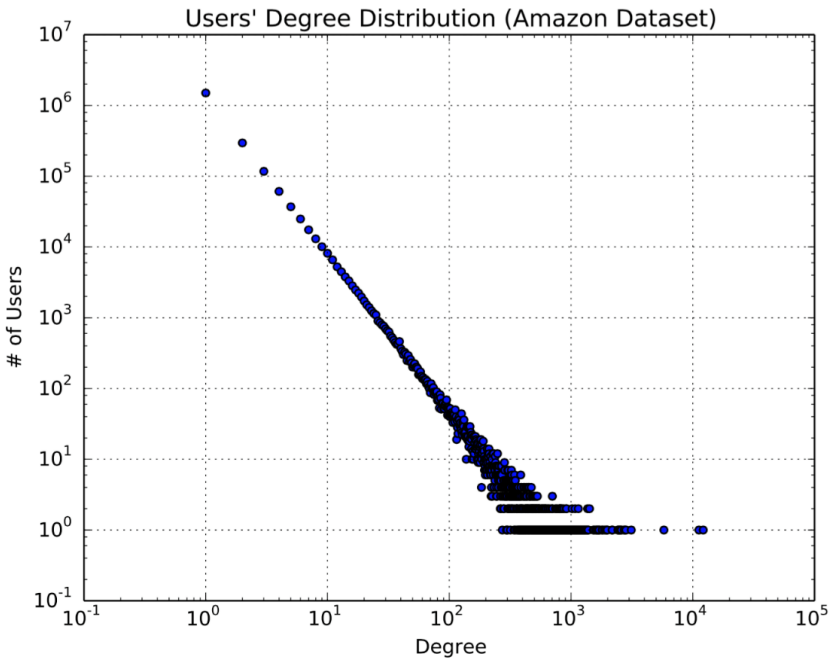
c: type of centrality (Degree or Pagerank)

k: index of bins

Network clues

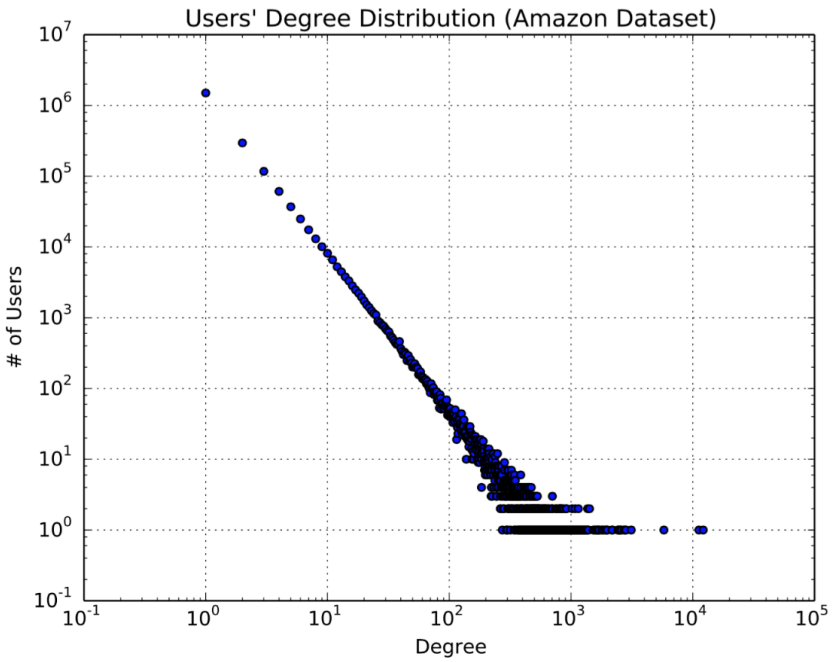
- **Observation 2:** Self-similarity
  - Local distributions similar to global ones

  → Neighbors have power-law-like centrality distributions



Network clues

- **Observation 2:** Self-similarity
  - Local distributions similar to global ones
  - → Neighbors have power-law-like centrality distributions



- Quantification:

centrality of product *i*'s neighbors

centrality of all users

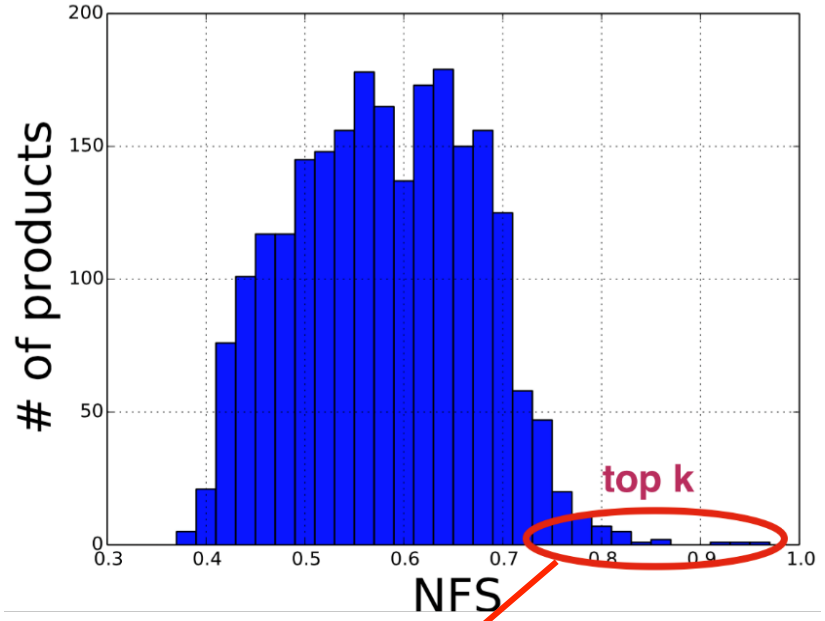
$$KL_c(P^{(i)} \| Q) = \sum_k p_k^{(i)} \log \frac{p_k^{(i)}}{q_k}$$

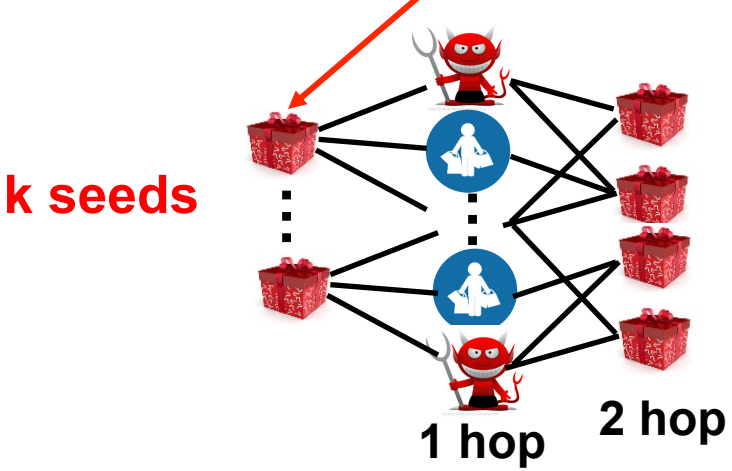
c: type of centrality (Degree or Pagerank)

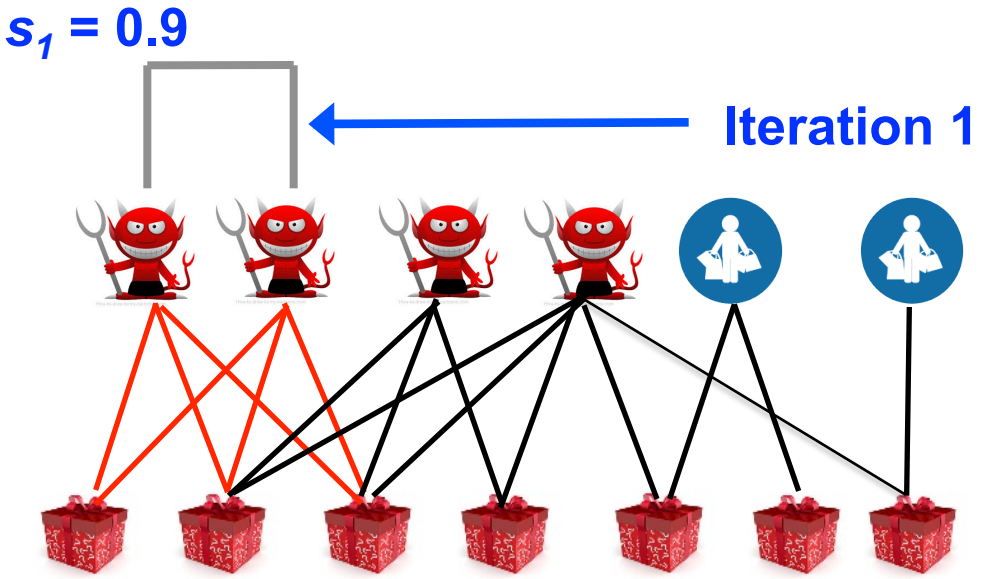
k: index of bins

Group detection

Distribution of NFS

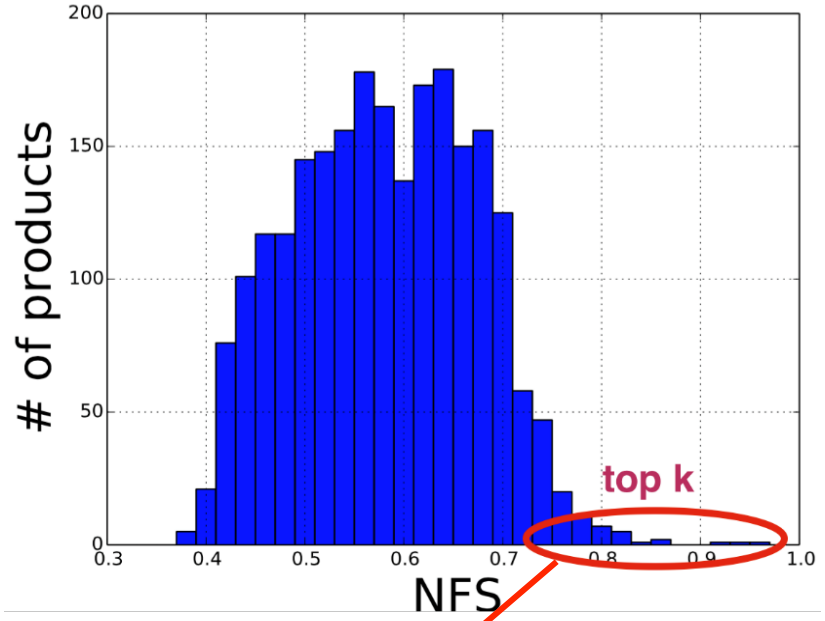


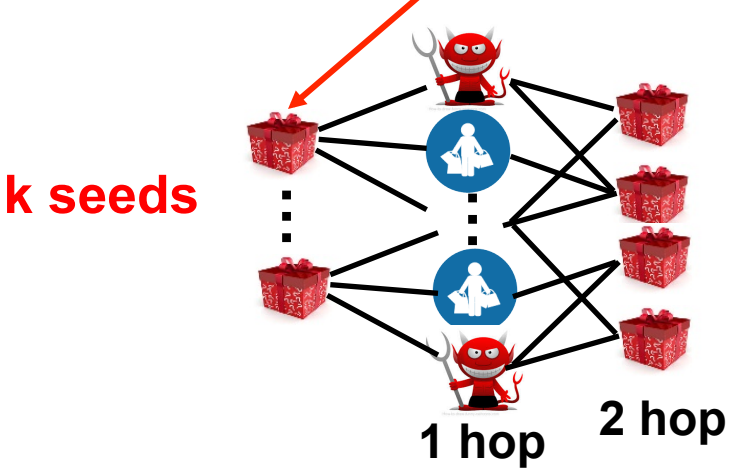


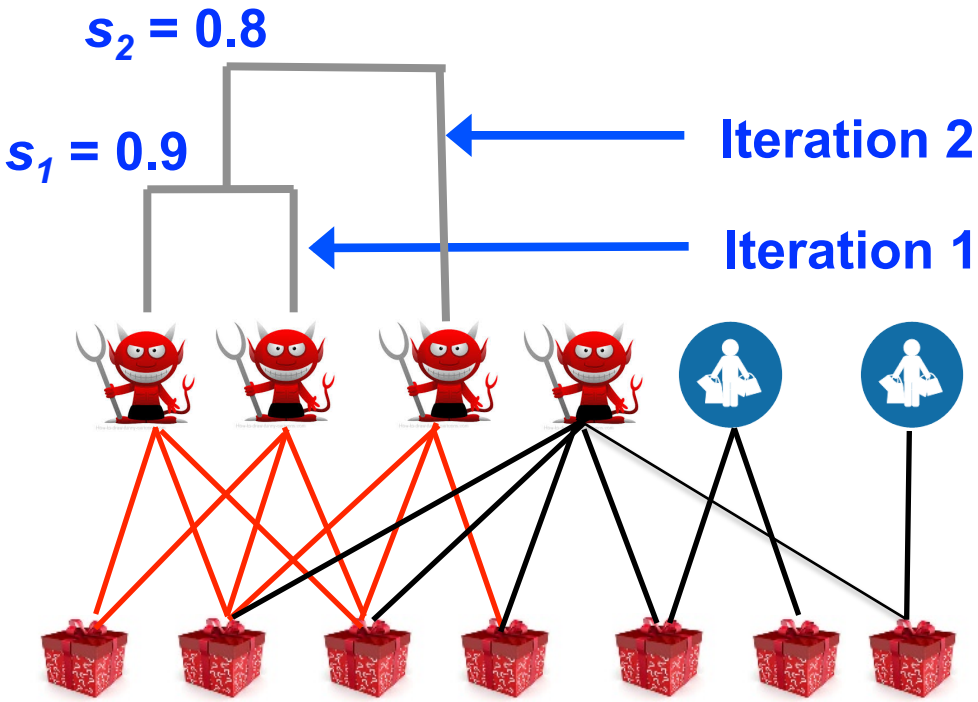


Group detection

Distribution of NFS

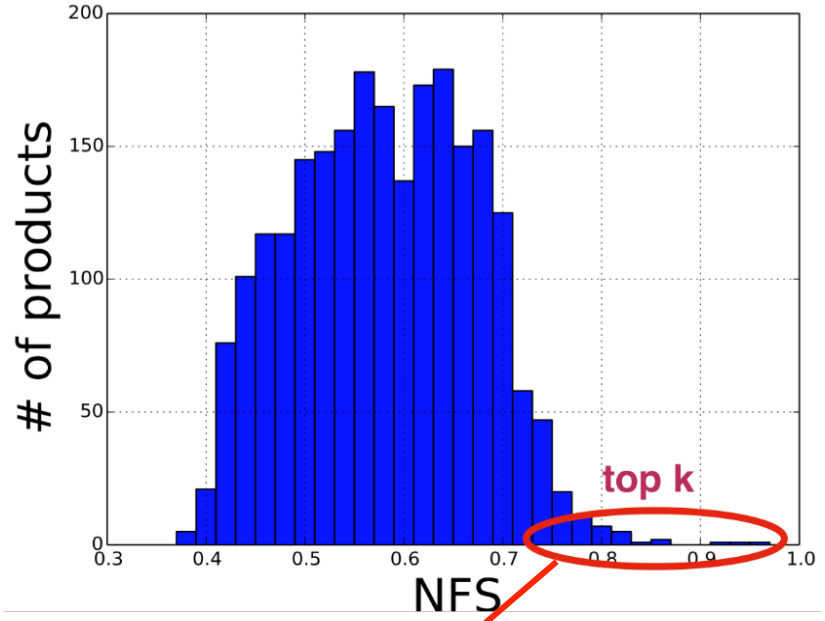


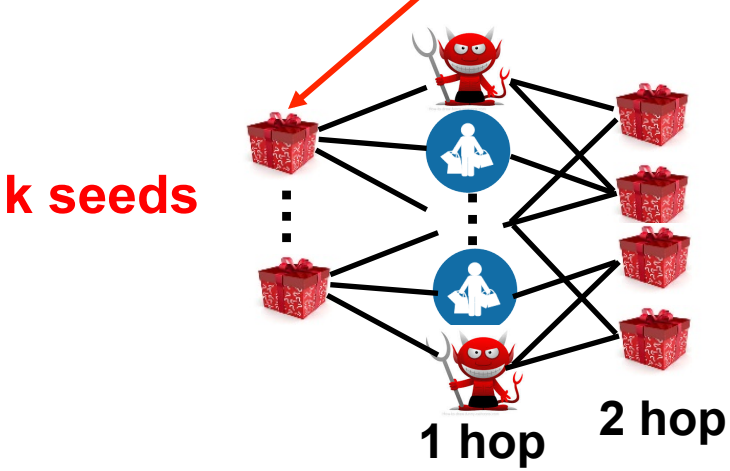


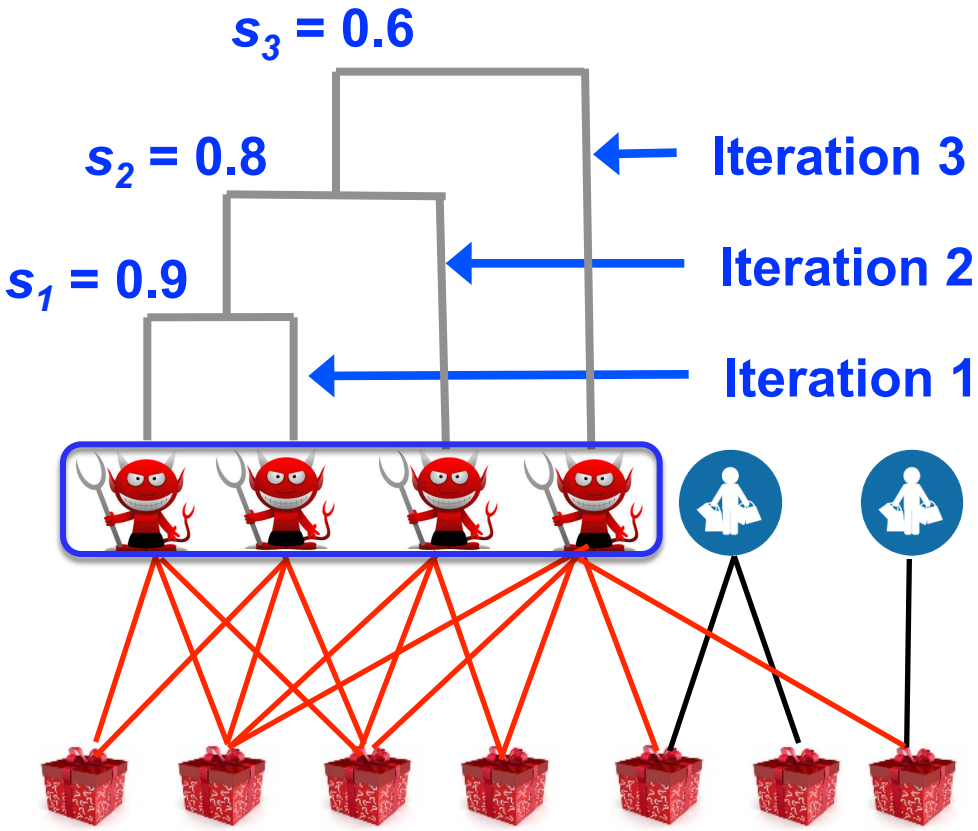


Group detection

Distribution of NFS







Datasets

- **Synthetic** datasets: (4 datasets, various generators and sizes)
  - Chung-Lu Generator [Chung et al., Internet Mathematics, 2003]
    - Degrees from a power-law dist.n with exponent 2.9 and 2.1
    - Camouflage a1) 10%, a2) 30% to b1) popular or b2) random products
  - Random Typing Generator (RTG) [Akoglu et al., PKDD, 2009]

- **Real-world** datasets:
  - SWM [Akoglu et al., ICWSM 2013]
  - Amazon [Jindal and Liu, WSDM 2008]

Table 1. Summary of synthetic and real-world datasets used in this work.

	Synthetic Data				Real-world Data	
	Chung-Lu1	Chung-Lu2	RTG1	RTG2	iTunes	Amazon
# of users	532,742	2,133,399	604,520	876,627	966,808	2,146,074
# of products	157,768	665,381	604,805	876,950	15,093	1,230,916
# of edges	1,299,059	5,191,053	3,097,342	4,644,572	1,132,329	5,838,061

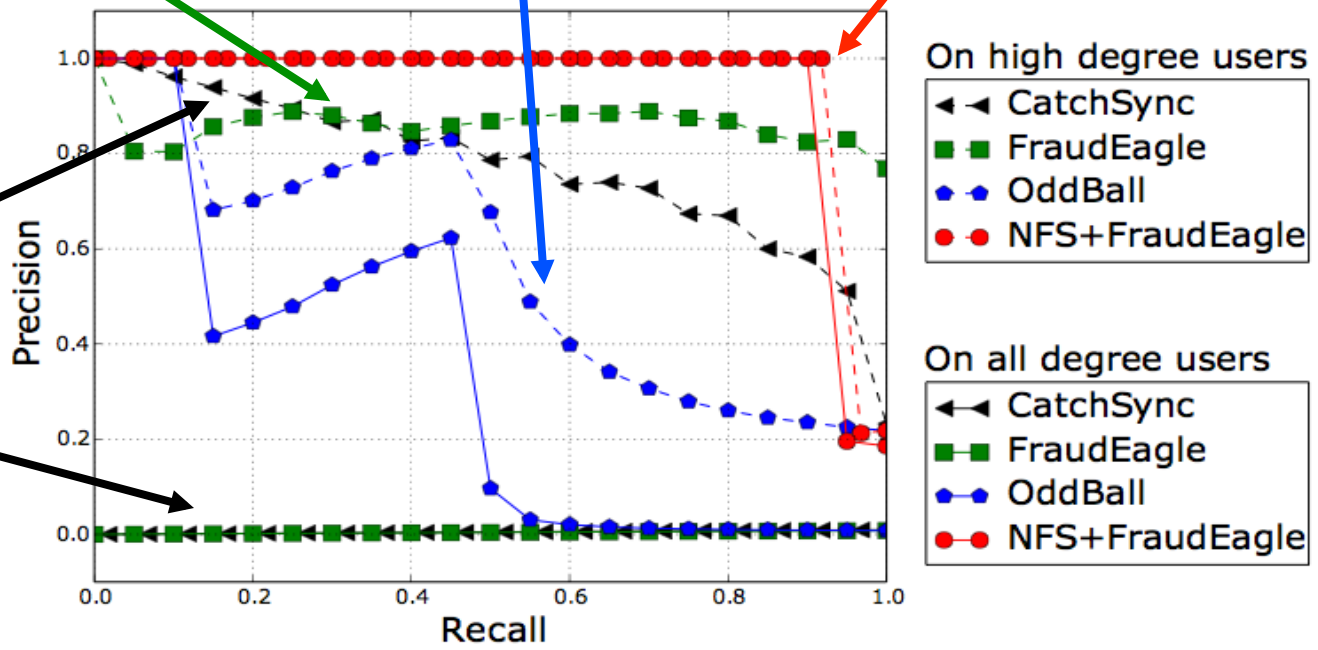
NFS on Synthetic Graphs

FraudEagle:
- Label propagation

OddBall:
- Detects near-cliques and star shapes

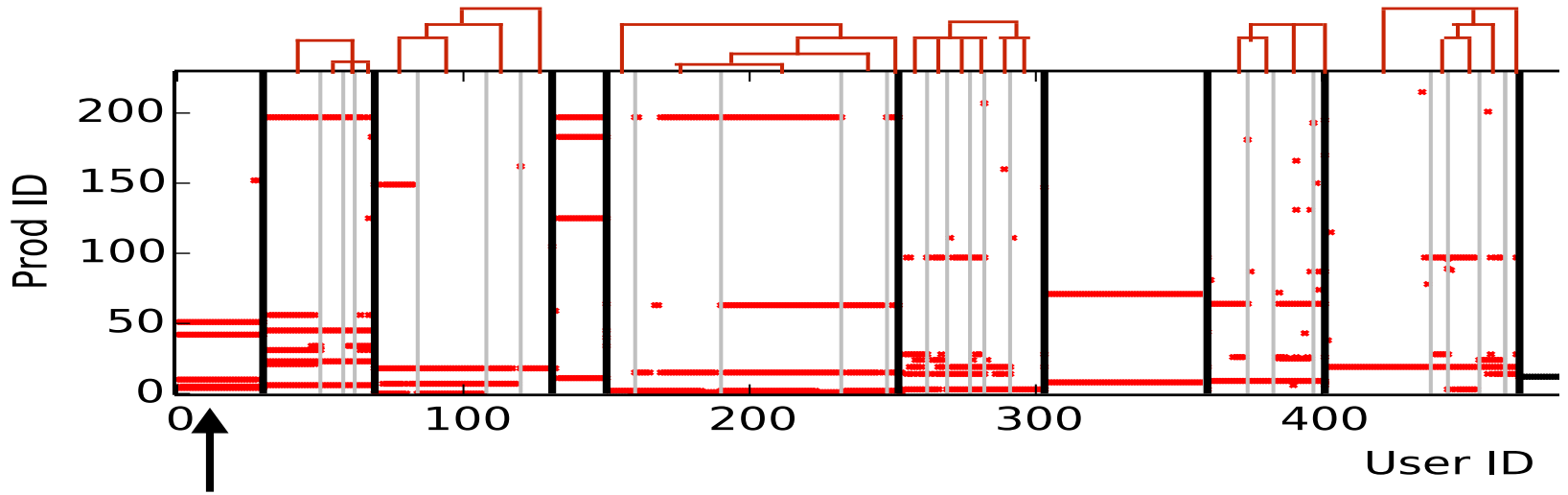
NFS + FraudEagle
- Solid line: all users
- Dash line: high-deg users

CatchSync
- Good on high-degree users
- Poor on all users



AUC of **Pre-Rec curve** on RTG2 (30% random camouflage)

Real data (SWM)



Group: (1) (2) (3) (4) (5) (6) (7) (8) (9)

ID	#P	#U	t, ⋆	*Dup*	Developer
1	5	31	s, c	51/154	all same
2	8	38	c, s	29/202	2 same
3	4	61	s, c	34/144	all inaccessible
4	4	17	c, s	0/68	1 inaccessible
5	5	102	c, s	8/326	different
6	6	50	s, c	4/173	all same
7	2	56	c, c	12/112	different
8	4	42	c, c	8/112	2 same
9	6	67	s, c	0/137	all same

t: time stamps
⋆: ratings
s: scattered
c: concentrated

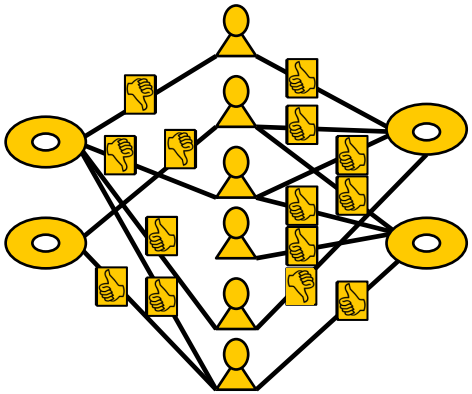
Real data (Amazon)

Amazon				
#P	#U	t, ⋆	*Dup*	Category, Author
10	20	c, c	90/138	Books, all same
4	12	s, c	32/47	Books, all same
7	9	c, c	44/60	Books, all same
7	19	s, c	5/88	Books, all same
23	42	c, c	2/468	Music, all same
8	17	s, c	9/73	Books, 4/8 same
6	18	s, c	4/94	Movies&TV, all same

Roadmap

- Intro/Motivation
- Fraud detection
  - Opinion fraud
    - *Network effects* [Akoglu+ ICWSM'13]
    - *Network clues for spammer groups* [Ye & Akoglu PKDD'15]
    - *Networks & meta-data* [Rayana & Akoglu KDD'15]
    - *Temporal analysis*
- Open challenges

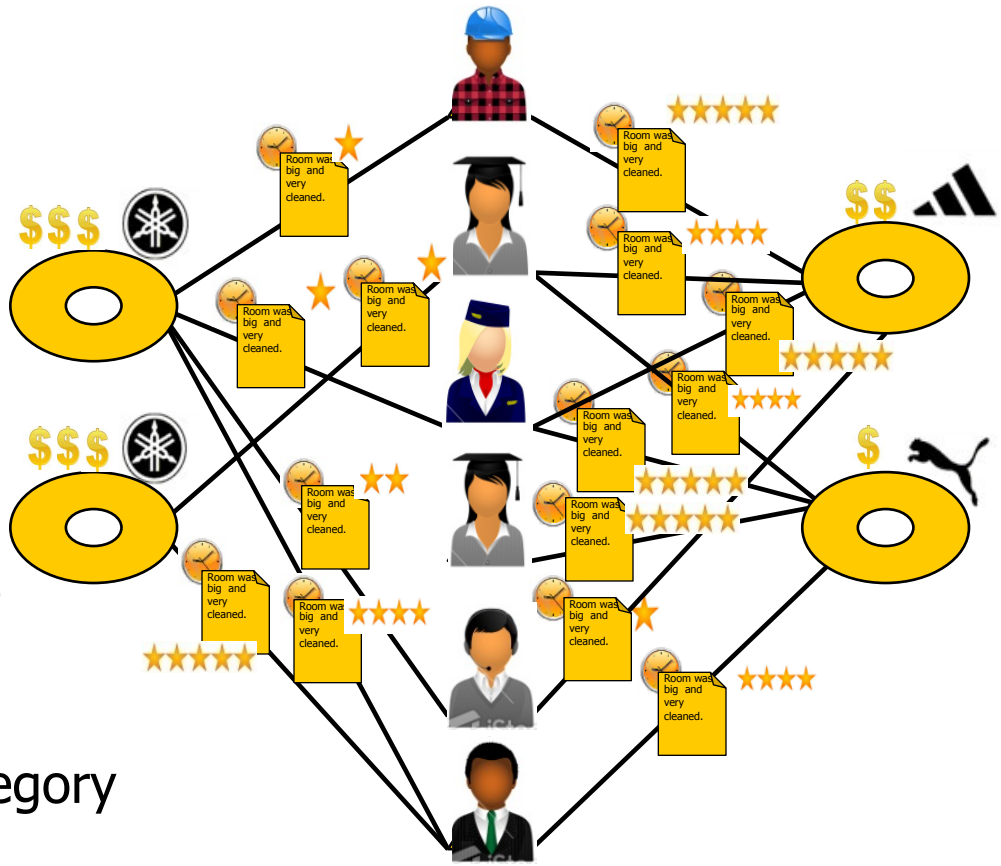
The ‘big’ picture



Spent the weekend in NYC to celebrate my wife's birthday. Walking distance to the theater district and shopping. Had a fantastic view of

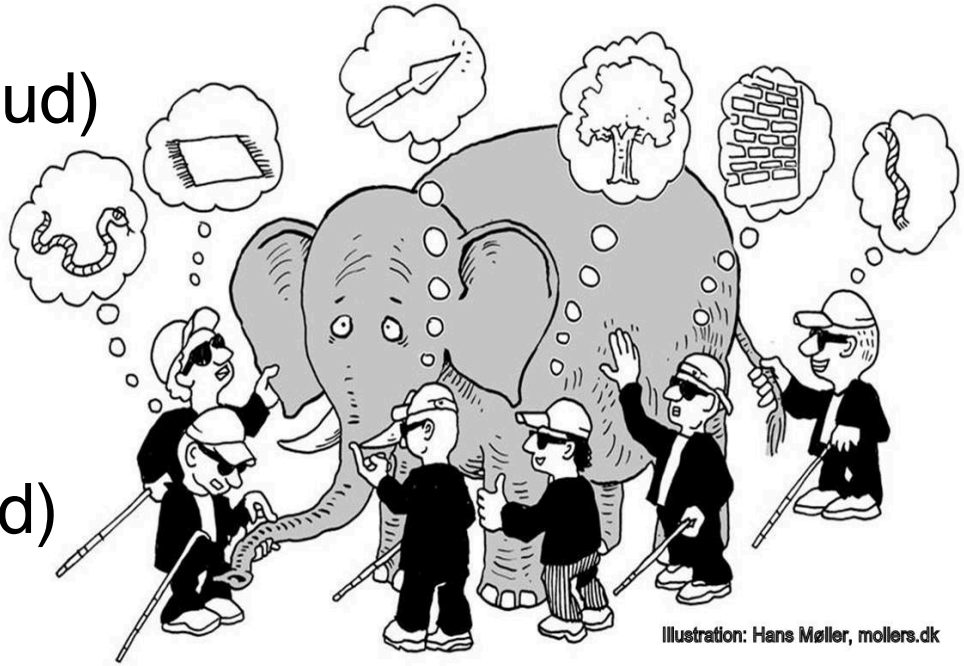
meta-data:

- text
- writing style
- timestamp
- IP address
- product-category
- price
- brand
- …



Networks and beyond

- Often relational data directly available
  - transactions (auction fraud)
  - phone-call (telecom fraud)
  - referrals (medical fraud)
  - co-worker (securities fraud)
  - …
- Often there is also meta data available
  - firm-size (securities fraud)
  - file name (malware)
  - company type, location (corporate fraud)
  - …



Previously

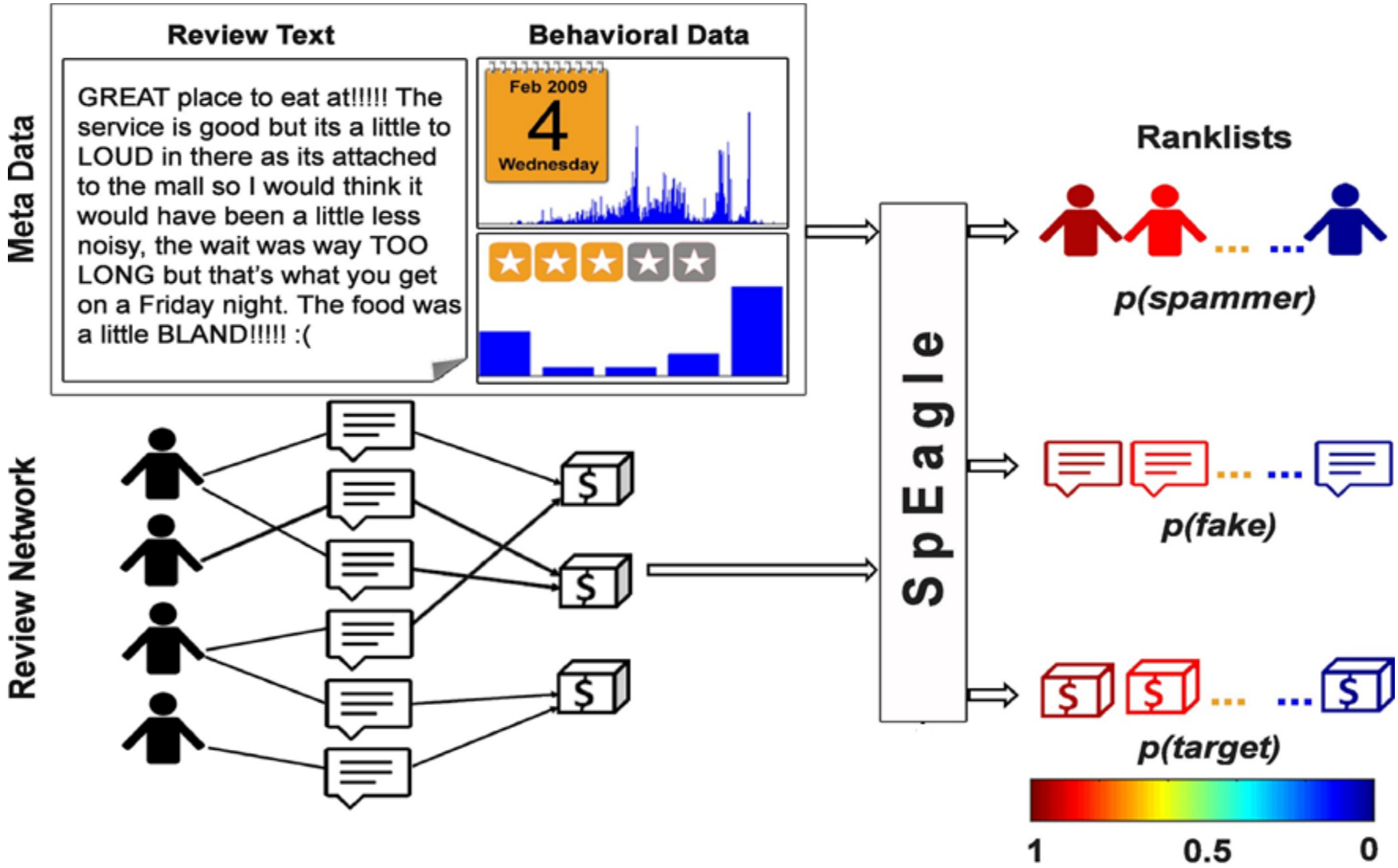
	Review Network	Review Text	Review Behavior	Supervision
Ott'2011		✓		supervised
Mukherjee' 2013		✓	✓	supervised
Jindal'2008			✓	supervised
Co-training [Li'2011]			✓	semi-supervised
Wang'2011	✓		✓	unsupervised
FraudEagle	✓			unsupervised
SpEagle	✓	✓	✓	unsupervised
SpEagle$^+$	✓	✓	✓	semi-supervised

Previously

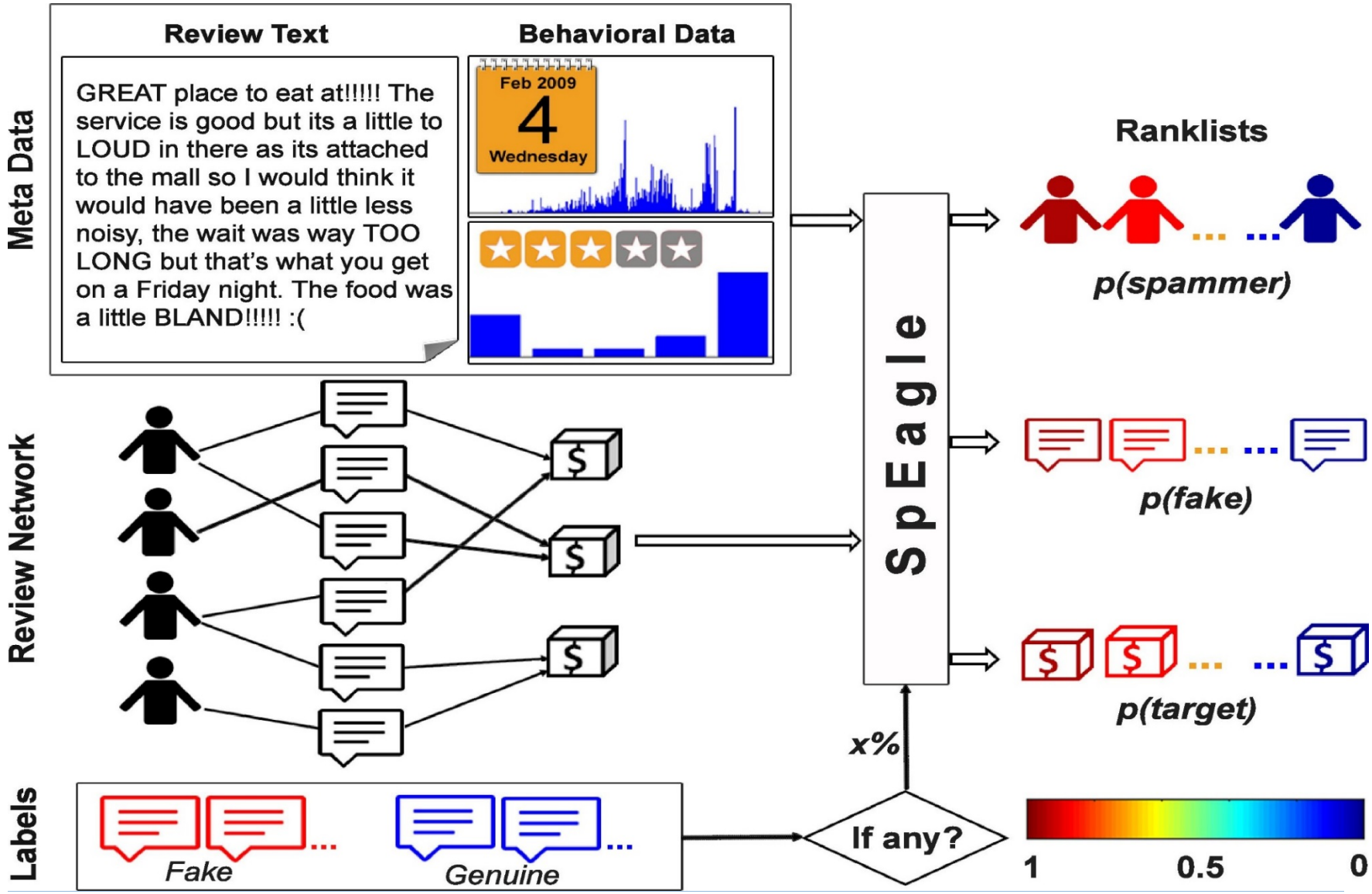
	Review Network	Review Text	Review Behavior	Supervision
Ott'2011		✓		supervised
Mukherjee' 2013		✓	✓	supervised
Jindal'2008			✓	supervised
Co-training [Li'2011]			✓	semi-supervised
Wang'2011	✓		✓	unsupervised
FraudEagle	✓			unsupervised
SpEagle	✓	✓	✓	unsupervised
SpEagle$^+$	✓	✓	✓	semi-supervised

Collective Opinion Spam Detection: Bridging Review Networks and Metadata
Shebuti Rayana and Leman Akoglu ACM SIGKDD, 2015.

Bridging review networks & metadata



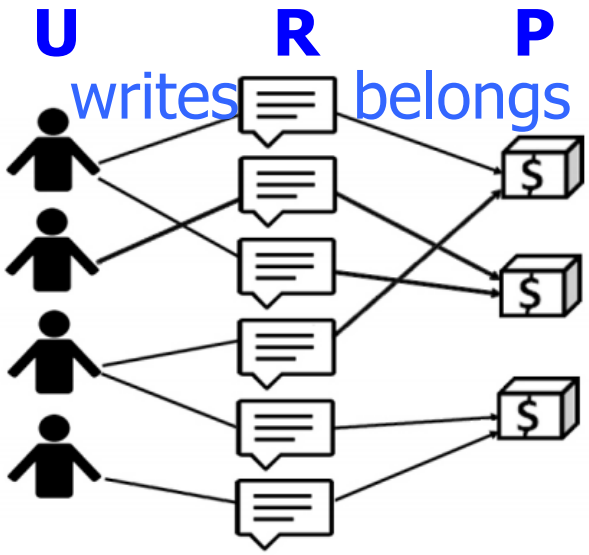
Bridging review networks & metadata



SpEagle

A network classification problem

- **Given**
  - **U**ser-**R**eview-**P**roduct network (tri-partite)
  - Features extracted from metadata (i.e. text, behavior)
    - for users, reviews, and products
- **Classify** network objects into type-specific classes
  - Users ('benign' vs. 'spammer')
  - Products ('non-target' vs. 'target')
  - Reviews ('genuine' vs. 'fake')



SpEagle

- Network classification approach (**unsupervised**)
  - Objective function :

$$\max_{\mathbf{y}} \; P(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}(\mathbf{x})} \prod_{Y_i \in \mathcal{Y}} \phi_i(y_i) \prod_{(Y_i, Y_j) \in E} \psi_{ij}^{t}(y_i, y_j)$$

edge type

Node labels as random variables

edge potential (label-label)

$$\phi_i(y_i) = \psi_i(y_i) \prod_{(Y_i, X_j) \in E} \psi_{ij}^{t}(y_i)$$

prior belief

edge potential (label-observed label)

Priors

Metadata

Features

Spam Scores

Priors

Users: 'benign' 'spammer'
Products: 'non-target' 'target'
Reviews: 'genuine' 'fake'

Feature extraction from metadata

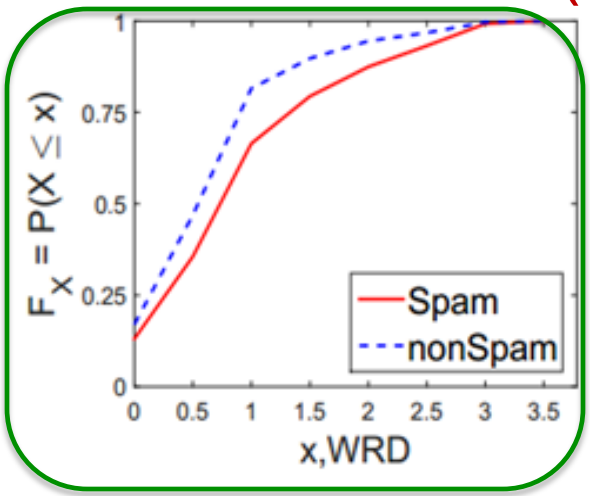
	User Features	Product Features	Review Features
Behavioral	• maximum #reviews/ day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy • burstiness of reviews	• maximum #reviews/day • ratio of +ve/-ve reviews • avg/weighted rating deviation • rating deviation entropy • temporal gaps entropy	• rank order of reviews • absolute/thresholded rating deviation • extremity of rating • early time frame • singleton review
Text	• review length (#words) • avg content similarity • max content similarity	• review length • avg content similarity • max content similarity	• ratio subjective/ objective • description length • ratio of exclamation sent. • freq. of similar reviews • % capital letters • review length • ratio 1st person pronoun

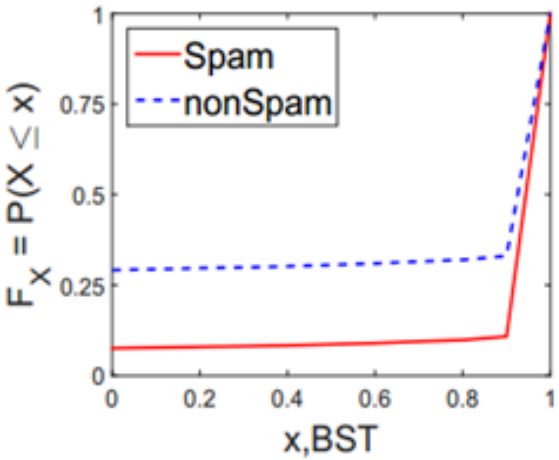
Feature extraction from metadata

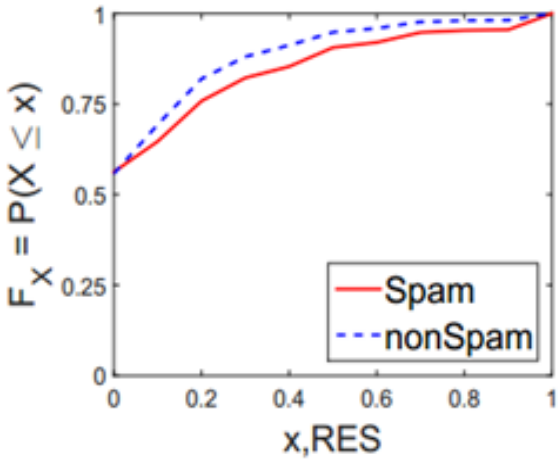
	User Features	Product Features	Review Features
Behavioral	• **Weighted Rating Deviation** $\frac{\sum_{e_{ij} \in E_{i*}} \lvert d_{ij} \rvert w_{ij}}{\sum_{e_{ij} \in E_{i*}} w_{ij}}$ $d_{ij} = r_{ij} - avg_{e \in E_{*j}} r(e)$ $w_{ij} = \frac{1}{(t_{ij})^{\alpha}}$ ↖ temporal order	• **Weighted Rating Deviation**	$\sum_{w} -log(freq(w))$ ↖ words in review • **Description Length**
Text			• ratio subjective/ objective • ratio of exclamation sent. • freq. of similar reviews • % capital letters • review length • ratio 1st person pronoun

Feature analysis

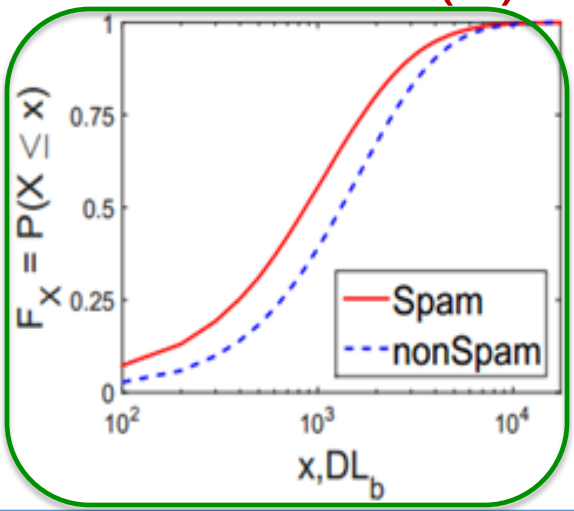
(H)igher more suspicious

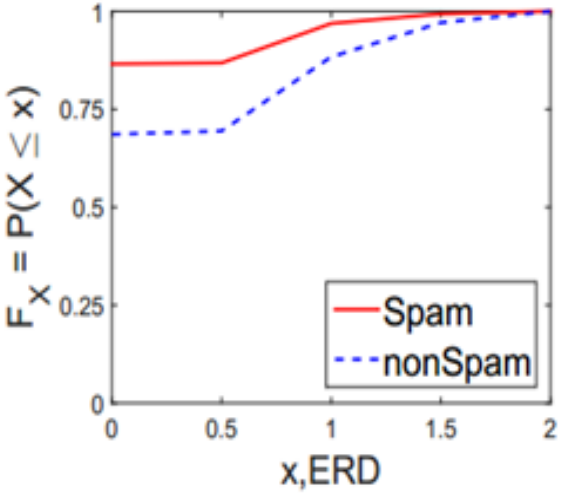


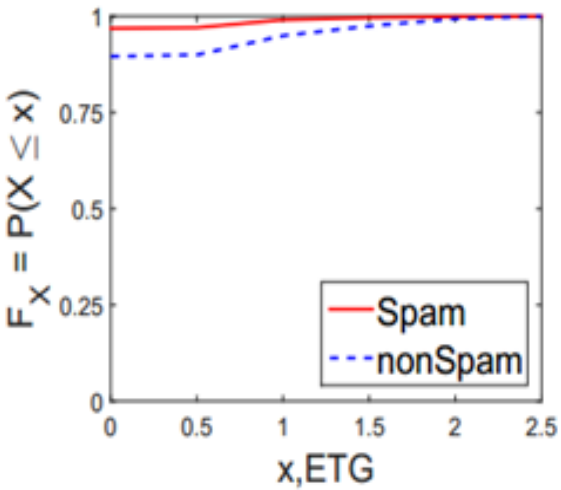




(L)ower more suspicious







Spam Score & Prior Computation

Q: How to handle features with different scales? **A:** Cumulative distribution:

- For each feature l , $1 \le l \le F$ and its corresponding value x_{li} for node i

$$f(x_{li}) = \begin{cases} 1 - P(X_l \le x_{li}), & if\ high\ is\ suspicious\ (H) \\ P(X_l \le x_{li}), & otherwise\ (L) \end{cases}$$

- Combine F values for each node i:

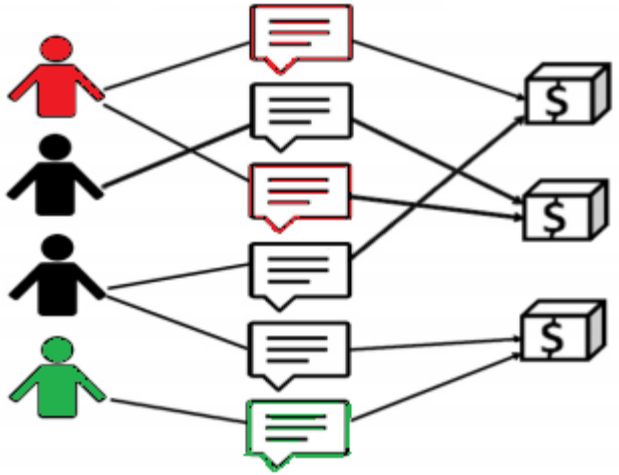
$$S_i = 1 - \sqrt{\frac{\sum_{l=1}^{F} f(x_{li})^2}{F}}$$

- Priors : $\phi_i \leftarrow \{1 - S_i, S_i\}$

SpEagle+: Using available labels

- SpEagle can work **semi-supervised**
  - can incorporate labels seamlessly
  - can use user, review, and/or product labels

- For labeled nodes, priors are set to:
  - $\phi \leftarrow \{\epsilon, 1 - \epsilon\}$ for spam
    (i.e., *fake*, *spammer*, or *target*)
  - $\phi \leftarrow \{1 - \epsilon, \epsilon\}$ for non-spam



Datasets

- 3 **Yelp** datasets[1]: recommended vs. non-recommended
  - **YelpChi** –hotel & restaurant reviews (Chicago)
  - **YelpNYC** –restaurant reviews (New York City)
  - **YelpZip** –restaurants reviews (NJ, VT, CT, PA)

Dataset	#Reviews (filtered %)	#Users (spammer %)[2]	#Products (rest.&hotel)
YelpChi	67,395 (13.23%)	38,063 (20.33%)	201
YelpNYC	359,052 (10.27%)	160,225 (17.79%)	923
YelpZip	608,598 (13.22%)	260,277 (23.91%)	5,044

[1] Datasets are available to the community

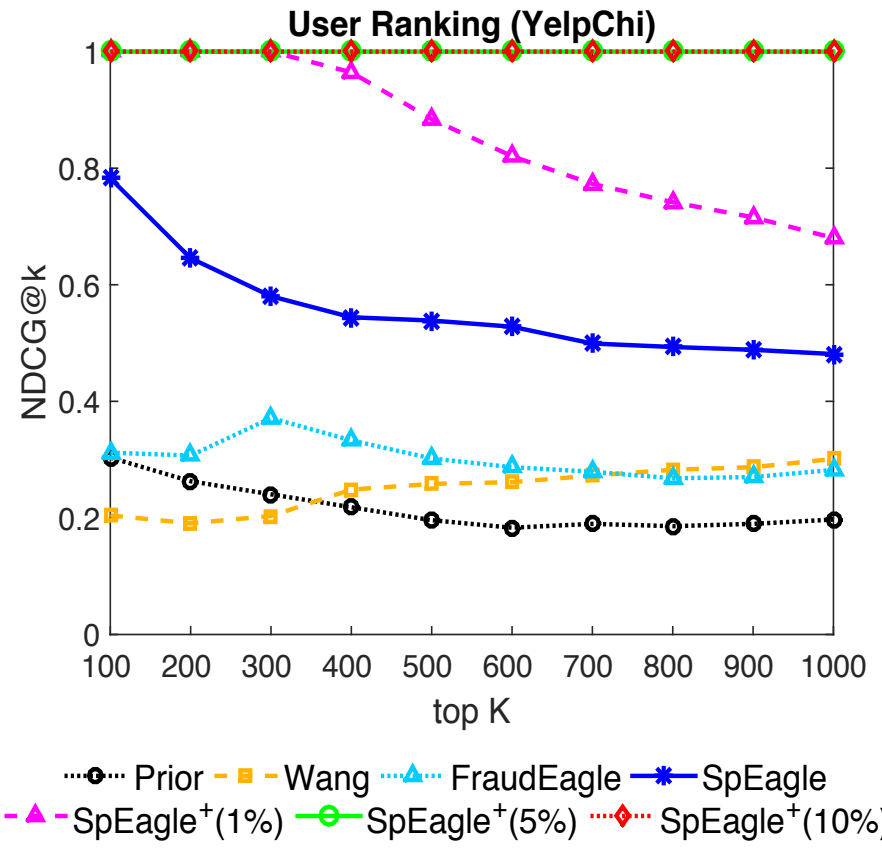
[2] A spammer has at least one filtered review

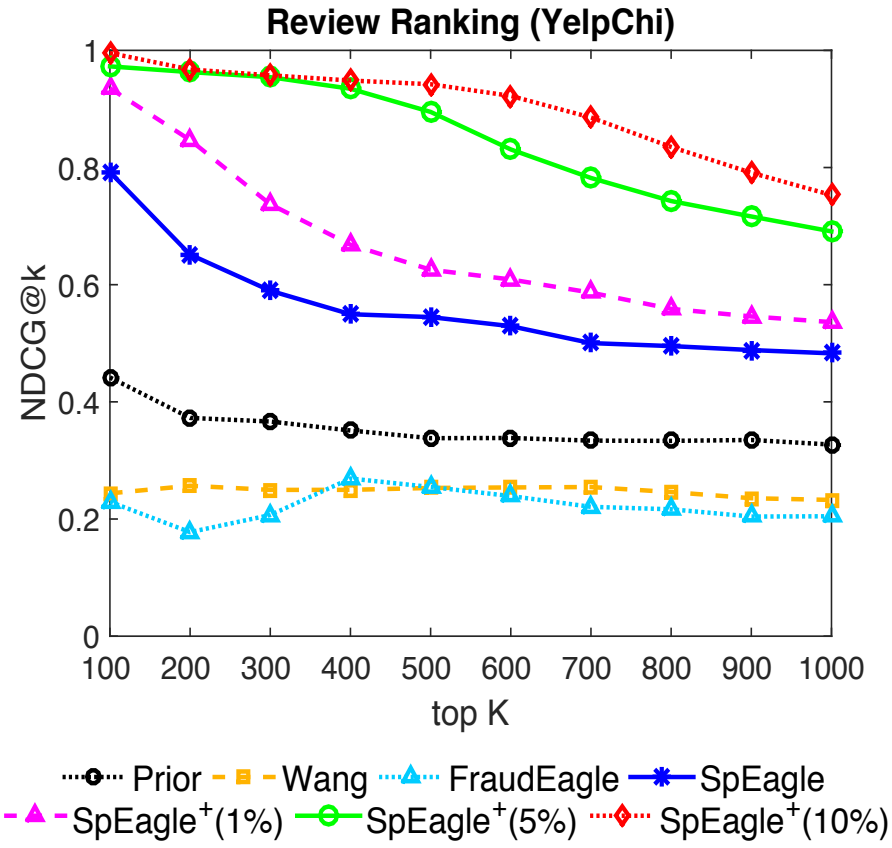
User & Review ranking

	User Ranking					
	AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
RANDOM	0.2024	0.1782	0.2392	0.5000	0.5000	0.5000
FRAUDEAGLE	0.2537	0.2233	0.3091	0.6124	0.6062	0.6175
WANG ET AL.	0.2659	0.2381	0.3306	0.6167	0.6207	0.6554
PRIOR	0.2157	0.1826	0.2550	0.5294	0.5081	0.5269
SPEAGLE	**0.3393**	**0.2680**	**0.3616**	**0.6905**	**0.6575**	**0.6710**

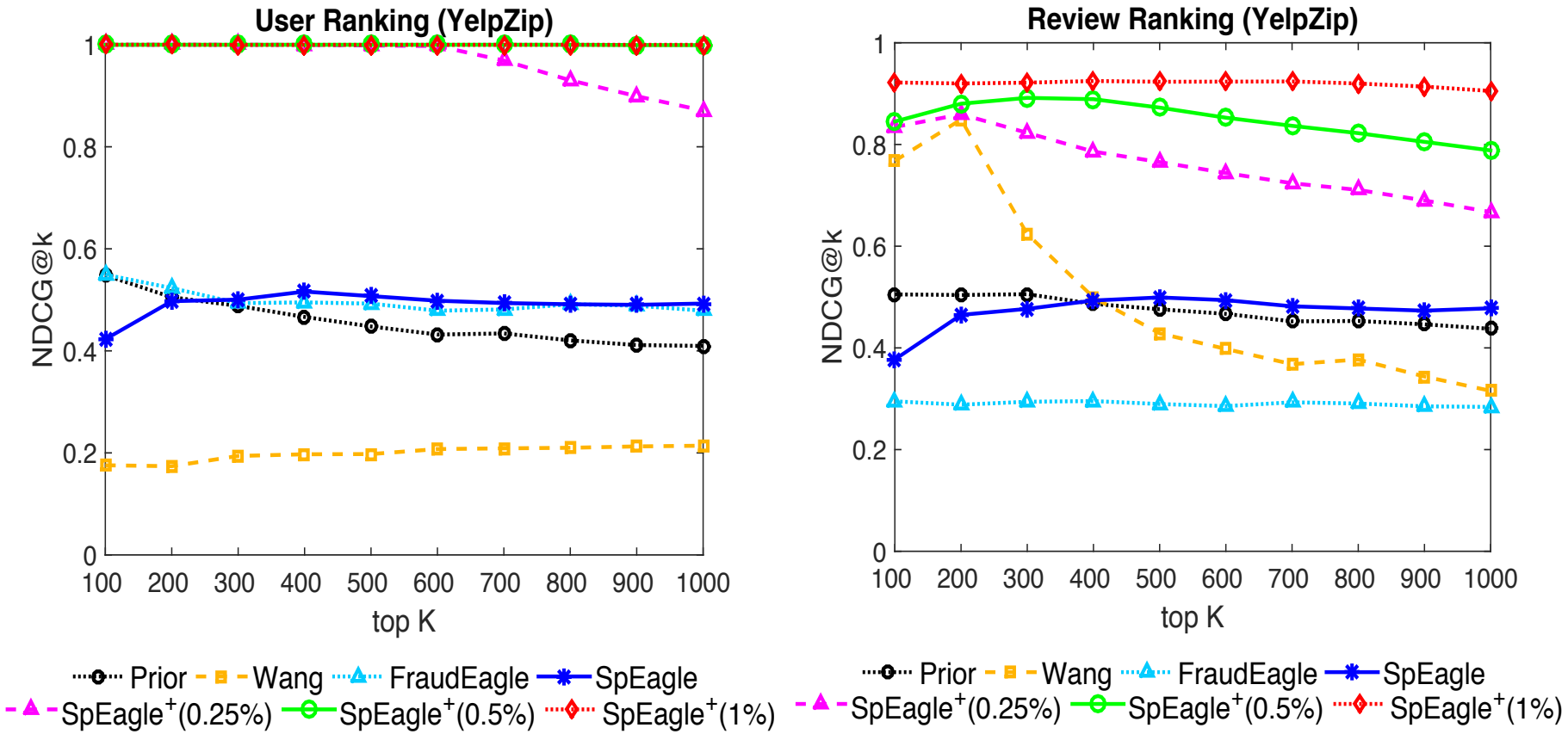
	Review Ranking					
	AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
	0.1327	0.1028	0.1321	0.5000	0.5000	0.5000
	0.1067	0.1122	0.1524	0.3735	0.5063	0.5326
	0.1518	0.1255	0.1803	0.5062	0.5415	0.5982
	0.2241	0.1789	0.2352	0.6707	0.6705	0.6838
	0.3236	**0.2460**	**0.3319**	**0.7887**	**0.7695**	**0.7942**

User & Review ranking





User & Review ranking

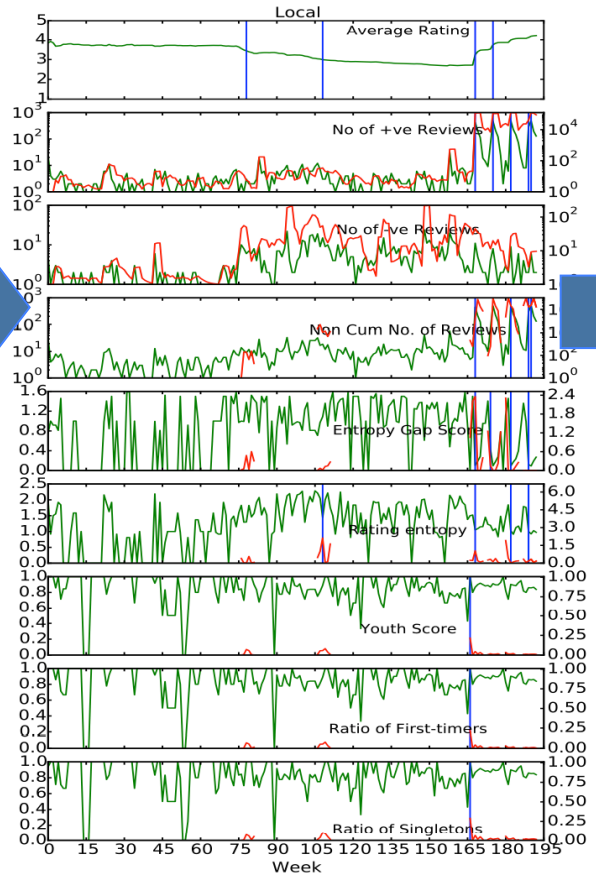


Roadmap

- Intro/Motivation
- Fraud detection
  - Opinion fraud
    - *Network effects* [Akoglu+ ICWSM'13]
    - *Network clues for spammer groups* [Ye & Akoglu PKDD'15]
    - *Networks & meta-data* [Rayana & Akoglu KDD'15]
    - *Temporal analysis* [Ye, Kumar, Akoglu ICWSM'16]
- Open challenges

Problem Statement

Input: products' review streams



Output: targeted products at time *t*.

Temporal Opinion Spam Detection by Multivariate Indicative Signals
Junting Ye, Santhosh Kumar, Leman Akoglu ICWSM, 2016.

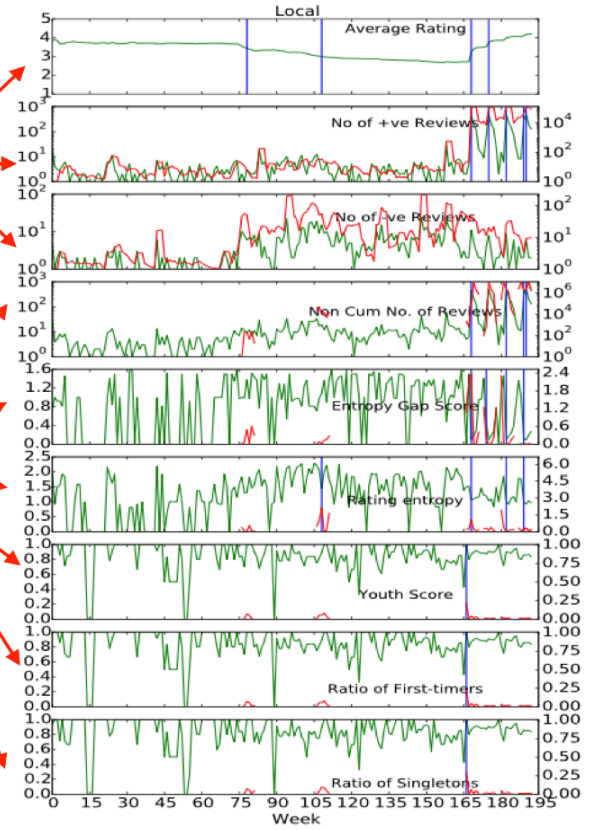
Approach Overview

1. Temporal Signal Extraction;

2. Anomaly Detection in Lead Signal;
   - i. CUSUM for average rating;
   - ii. Autoregressive model (AR) for others;

3. Anomaly Detection in Supporting Signals;
   - i. Analyze local values only when "alarms" triggered by lead signal;
   - ii. Use AR to detect anomalies;

4. Suspiciousness Quantification;
   - i. 4 features to characterize anomalies;
   - ii. Integrate features into single value;



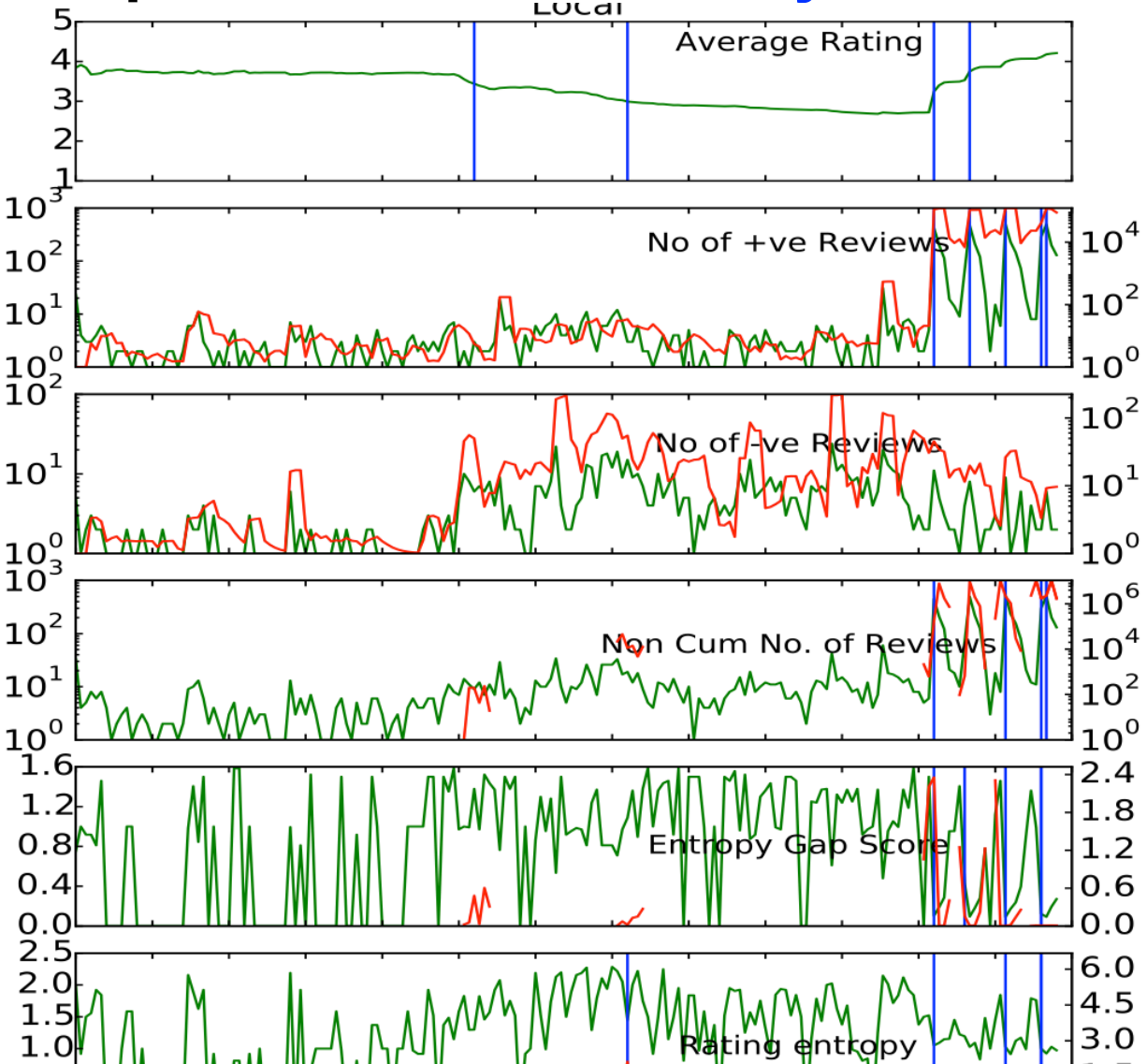
Indicative Signals

Table 1: *Indicative signals of opinion spam.*

Name	Range	Suspicious if
Average Rating	$[1, 5]$	Change
Number of (+/−) Reviews	$[0, \infty]$	Increase
Rating Entropy	$[0, \log_2 5]$	Decrease
Ratio of Singletons	$[0, 1]$	Increase
Ratio of First-timers	$[0, 1]$	Increase
Youth Score	$[0, 1]$	Increase
Temporal Gap Entropy	$[0, maxe^{\dagger}]$	Decrease

Case (SWM)

- **Burst in # of positive reviews: every 7 weeks;**

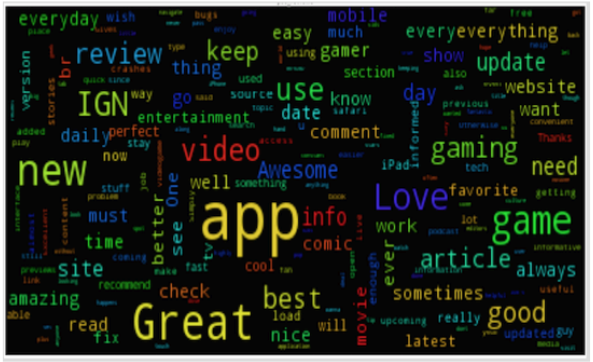


Case (SWM)

- **Duplicate review texts**

(a) 1-2⋆ reviews (weeks 75 to 165)

- "Great app for gamers"
- "Great App For Gaming news"
- "Must have app for gamers"
- "One of my favorite apps"



(b) 4-5⋆ reviews (week 168)

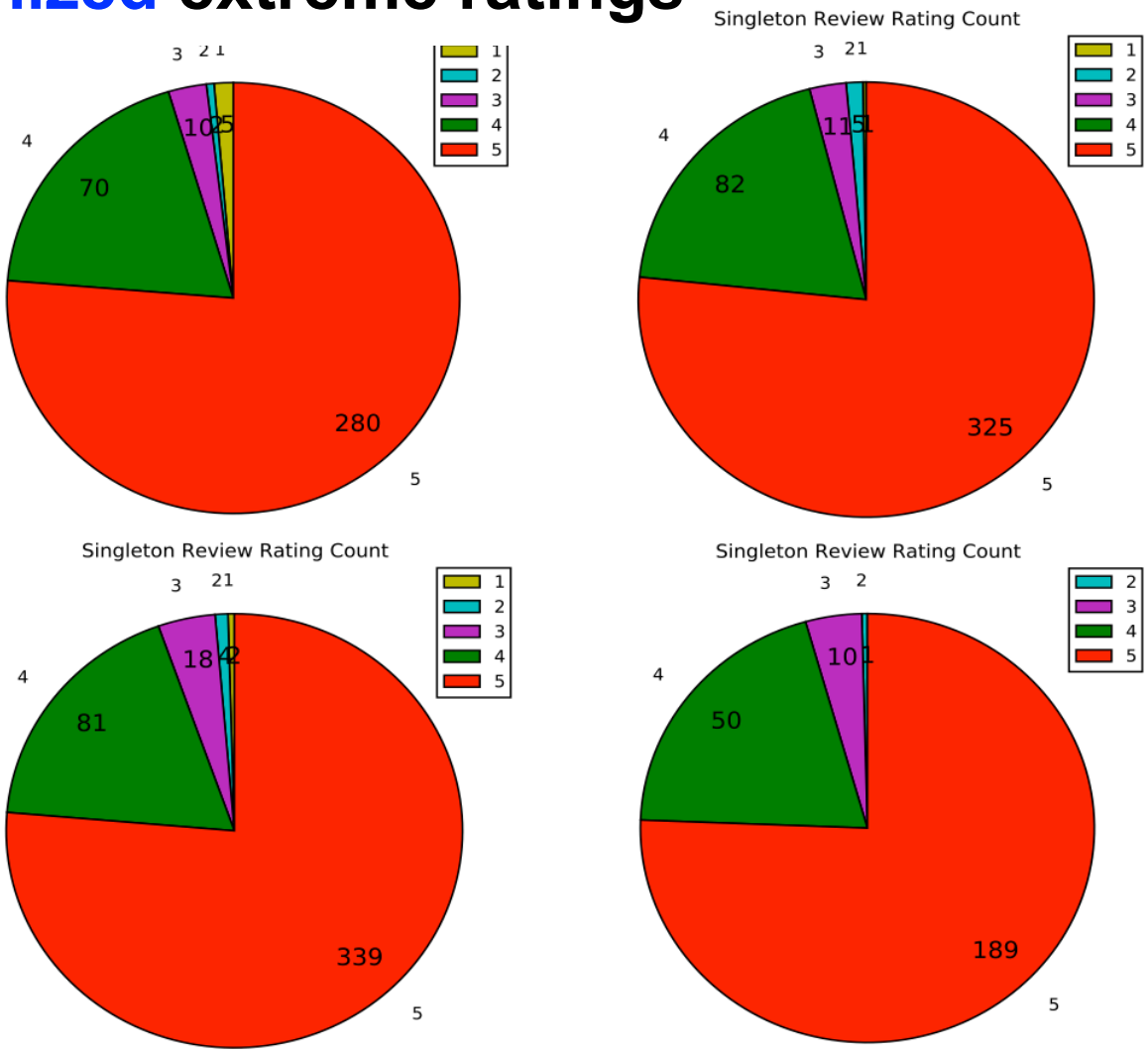
(c) 4-5⋆ reviews (week 175)

(d) 4-5⋆ reviews (week 182)

(e) 4-5⋆ reviews (week 189)

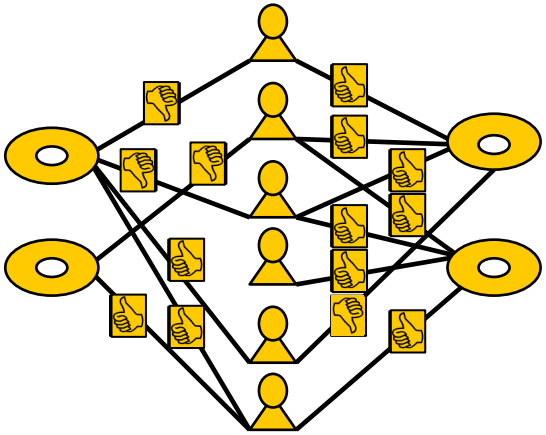
Case (SWM)

- **Synchronized** extreme ratings



Summary – opinion spam

- Fraud is often organized
- Users collude in ratings, time, text, targets
- How to bring together all pieces of the puzzle?
  - network footprints (how fraudsters are "embedded" in the network), behavior, language, time



Roadmap

- Intro/Motivation
- Fraud detection
  - Opinion fraud
- Open challenges

Open Q's for fraud detection

1. In what new ways can we bring together various data sources?
2. How to characterize the various strategies/ mechanisms behind fraud?
3. How to think of prevention (vs. detection) through platform or policy changes?
5. How to measure impact of fraud on consumers? (e.g., counterfeit sale @Amazon)
6. How to quantify adversarial robustness? (fraudsters have limited resources: $$$, time)

References

- Temporal Opinion Spam Detection by Multivariate Indicative Signals
  Junting Ye, Santhosh Kumar, Leman Akoglu. ICWSM, 2016

- BIRDNEST: Bayesian Inference for Ratings-Fraud Detection
  Bryan Hooi, Neil Shah, Alex Beutel, Stephan Gunnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, Christos Faloutsos. SIAM SDM, 2016

- Collective Opinion Spam Detection using Active Inference
  Shebuti Rayana and Leman Akoglu. SIAM SDM, 2016

- Collective Opinion Spam Detection: Bridging Review Networks and Metadata
  Shebuti Rayana and Leman Akoglu. ACM SIGKDD, 2015

- Discovering Opinion Spammer Groups by Network Footprints
  Junting Ye and Leman Akoglu. ECML/PKDD, 2015

- Opinion Fraud Detection in Online Reviews using Network Effects
  Leman Akoglu, Rishi Chandy, Christos Faloutsos. ICWSM, 2013

for Code and Data:

http://www.andrew.cmu.edu/user/lakoglu/pubs.html#code
http://shebuti.com/collective-opinion-spam-detection/
http://odds.cs.stonybrook.edu/

Thanks!