CATCHSYNC: CATCHING SYNCHRONIZED BEHAVIOR IN LARGE DIRECTED GRAPHS

Meng Jiang, Tsinghua University, Beijing, China

Joint work with Peng Cui, Alex Beutel,
Christos Faloutsos and Shiqiang Yang
August 26, 2014 – NYC, USA

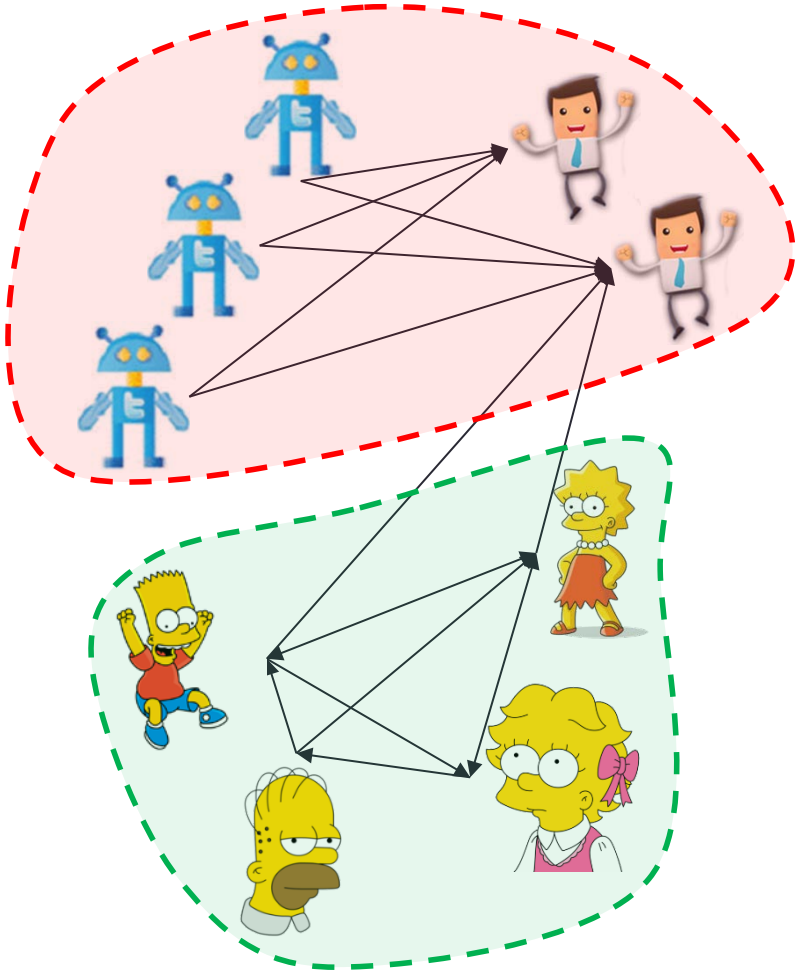
Fraud Detection: Graph Analysis Problem

[www.buyfollowz.org]

5,000 FOLLOWERS	2,000 FOLLOWERS	1,000 FOLLOWERS	10,000 FOLLOWERS	20,000 FOLLOWERS
400 FREE	300 FREE	200 FREE	500 FREE	1000 FREE
$69.99	$29.99	$15.99	$119.99	$229.99
Delivery within 3-4 days	Delivery within 2-3 days	Delivery within 1-2 days	Delivery within 4-5 days	Delivery within 5-8 days
Buy Now	Buy Now	Buy Now	Buy Now	Buy Now
Save + 3%	Save + 2%		Save + 14%	Save + 34%

[buymorelikes.com]

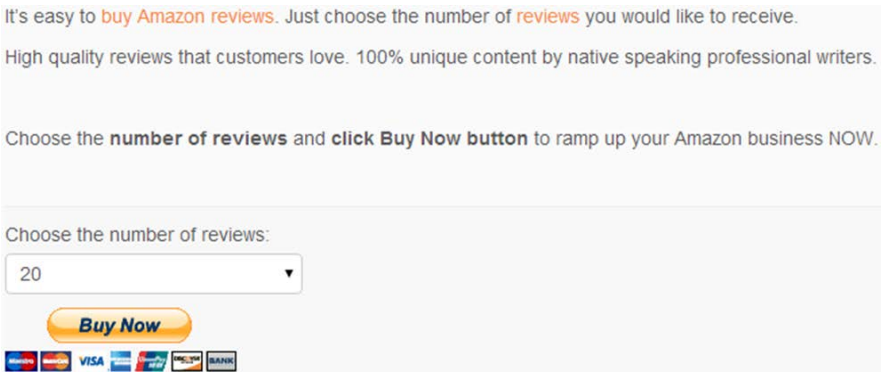
25,000 Facebook Likes	50,000 Facebook Likes	100,000 Facebook Likes	200,000 Facebook Likes
$265	$525	$1,000	$1,750
Lifetime Replacement Warranty	Lifetime Replacement Warranty	Lifetime Replacement Warranty	Lifetime Replacement Warranty
Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service	Dedicated 24/7 Customer Service
100% Risk Free, Try Us Today	100% Risk Free, Try Us Today	100% Risk Free, Try Us Today	100% Risk Free, Try Us Today
Order starts within 24 - 48 hours	Order starts within 24 - 48 hours	Order starts within 24 -48 hours	Order starts within 24 -48 hours
Order completed within 22 days	Order completed within 35 days	Order completed within 35 days	Order completed within 35 days

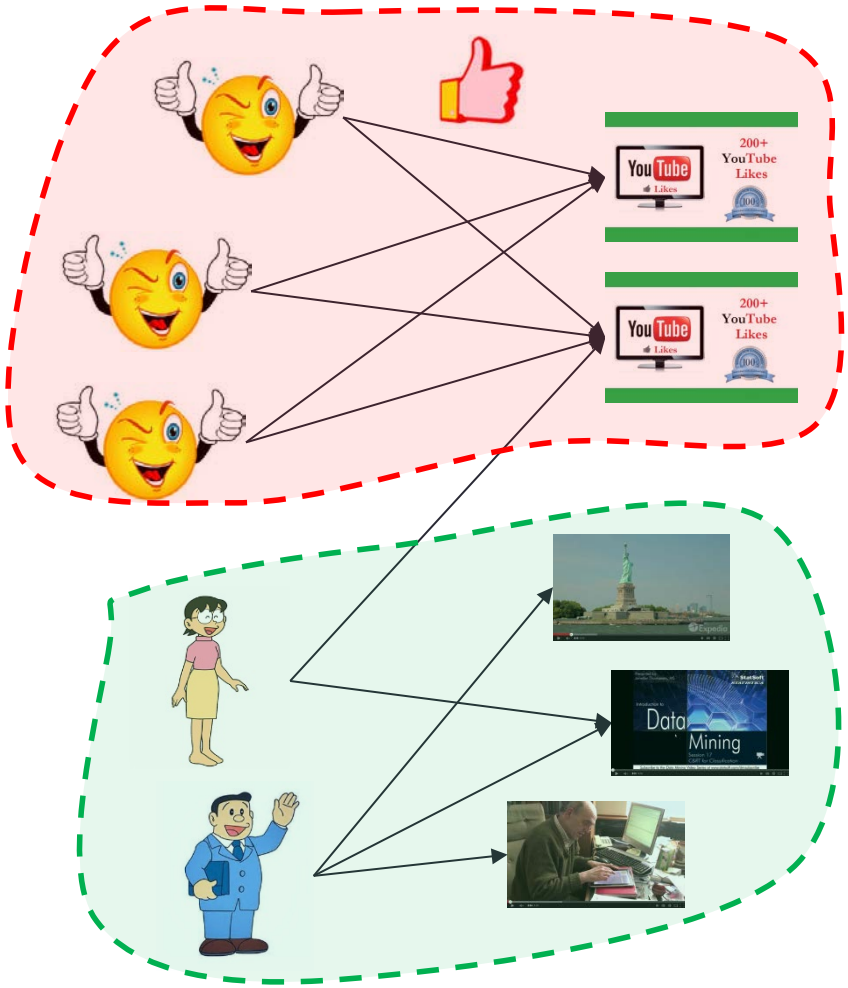


Fraud Detection: Graph Analysis Problem

[buycheaplikes.com]

[reviewsteria.com]





Our Goals

- Given: A graph (large-scale, directed, etc.)
- Find: Frauds = Anomalous edges

- Goals:
  - G1. Find patterns that distinguish fraudsters from normal users
  - G2. Design algorithms that catch fraudsters

OUTLINE

1. Background

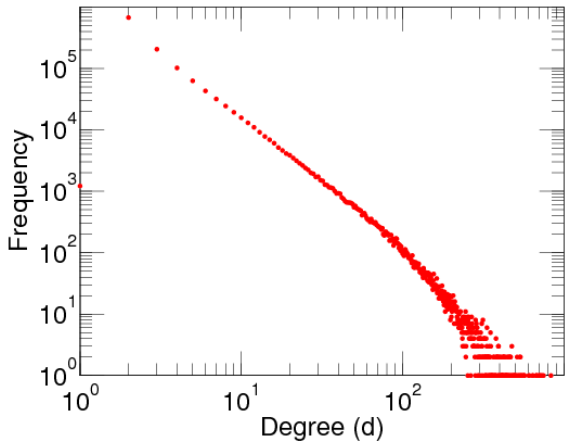
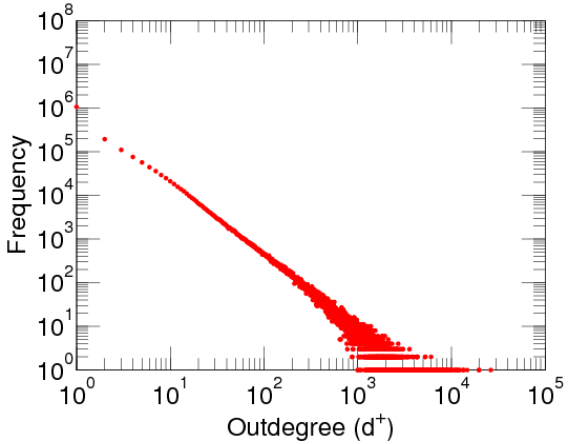
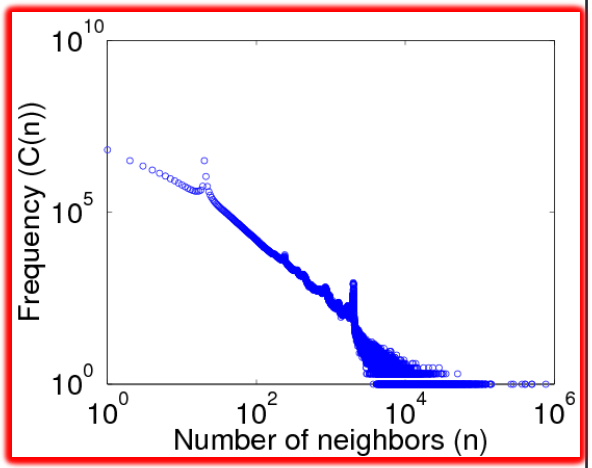
2. Fraudulent Pattern

3. The Algorithm

4. Experiments

Anomalies in Degree Distributions

- Power-law distribution

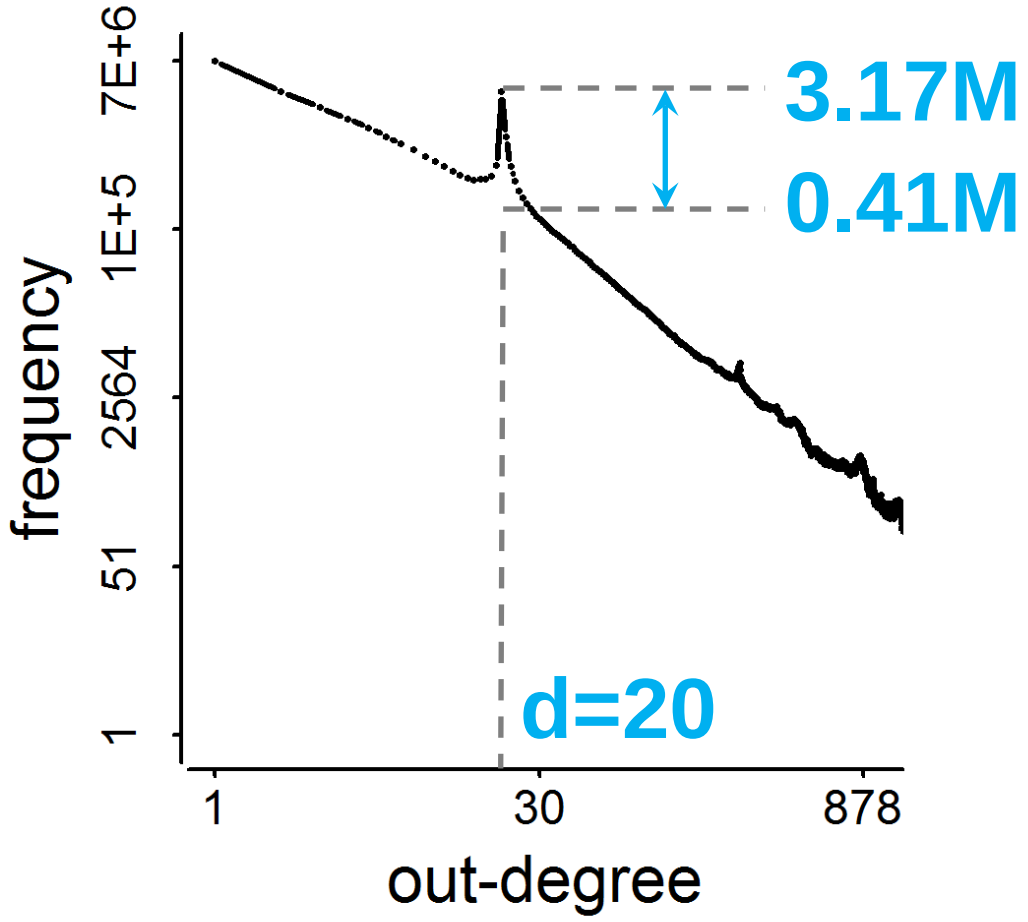
DBLP **Author**-publication	**Flickr** **User**-user	**Twitter** **Who**-follows-whom
		

[konect.uni-koblenz.de/networks/]

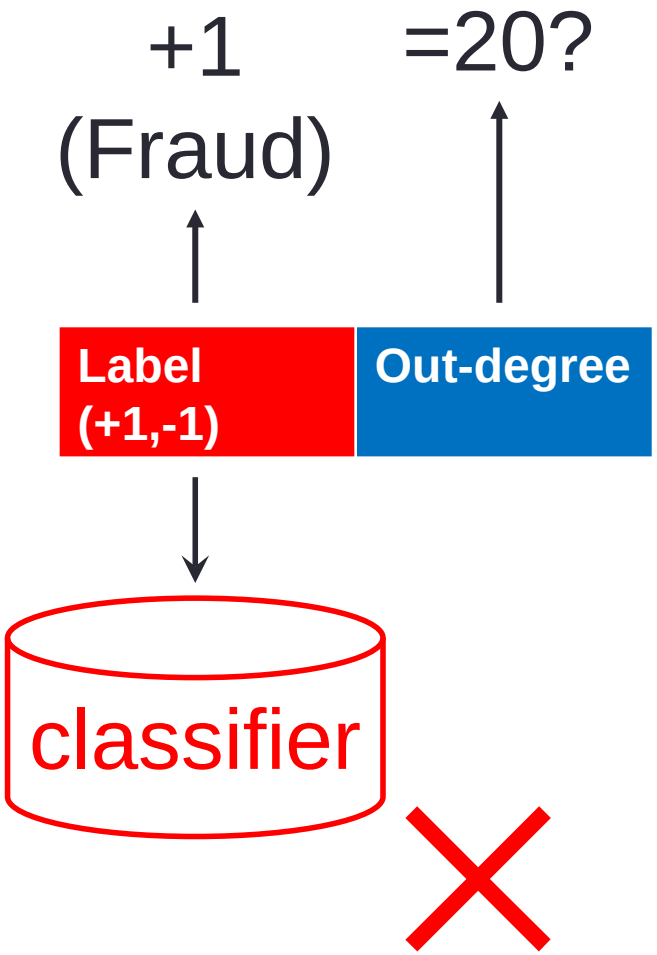
Anomalies in Degree Distributions

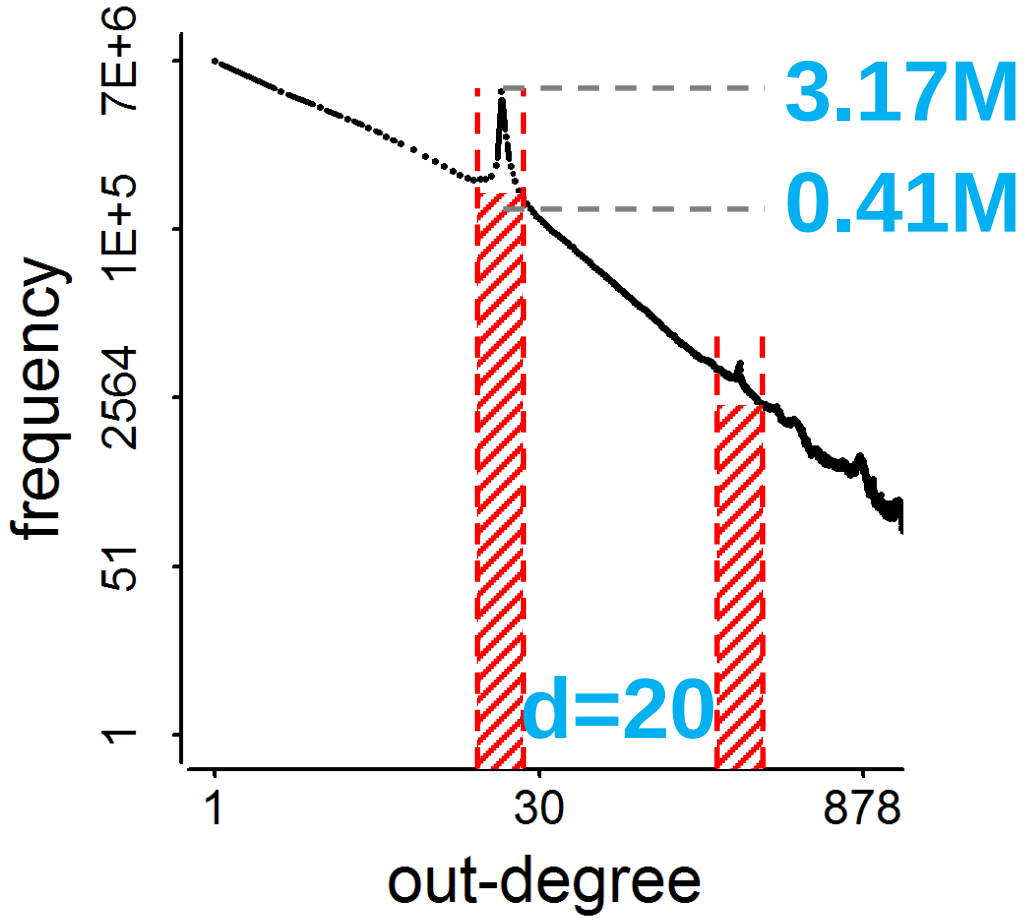
2009

41M



Linear Classifier with “Degree”: Fail

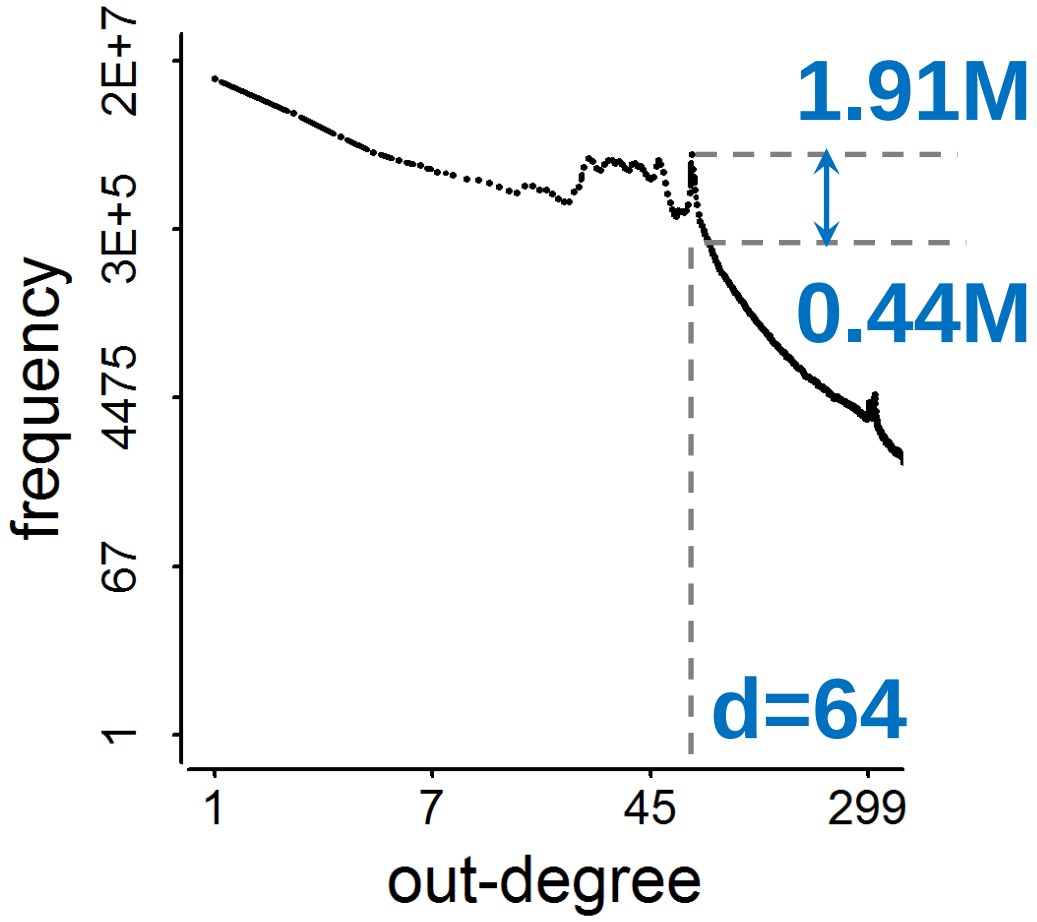




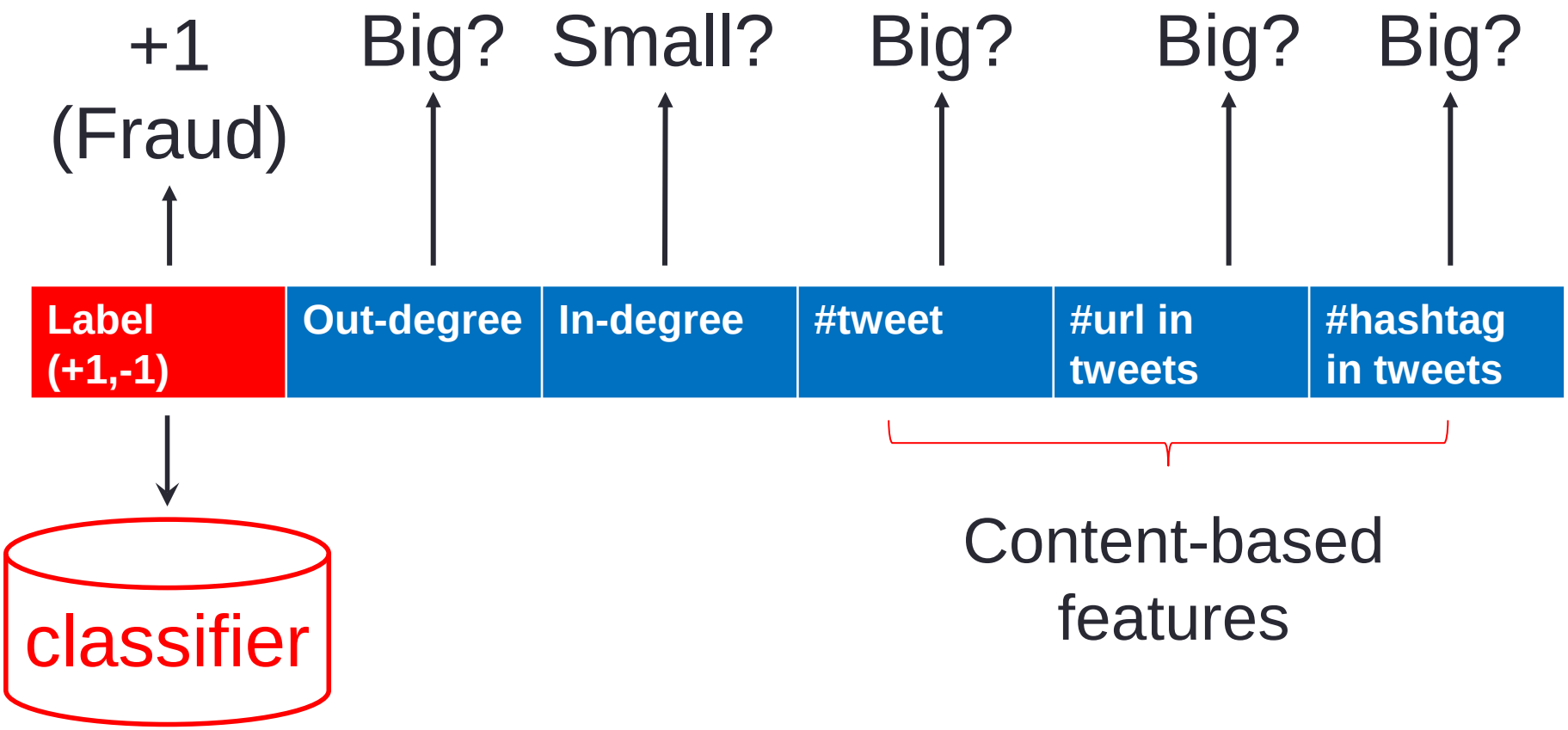
Graph Structure Distorted

2011

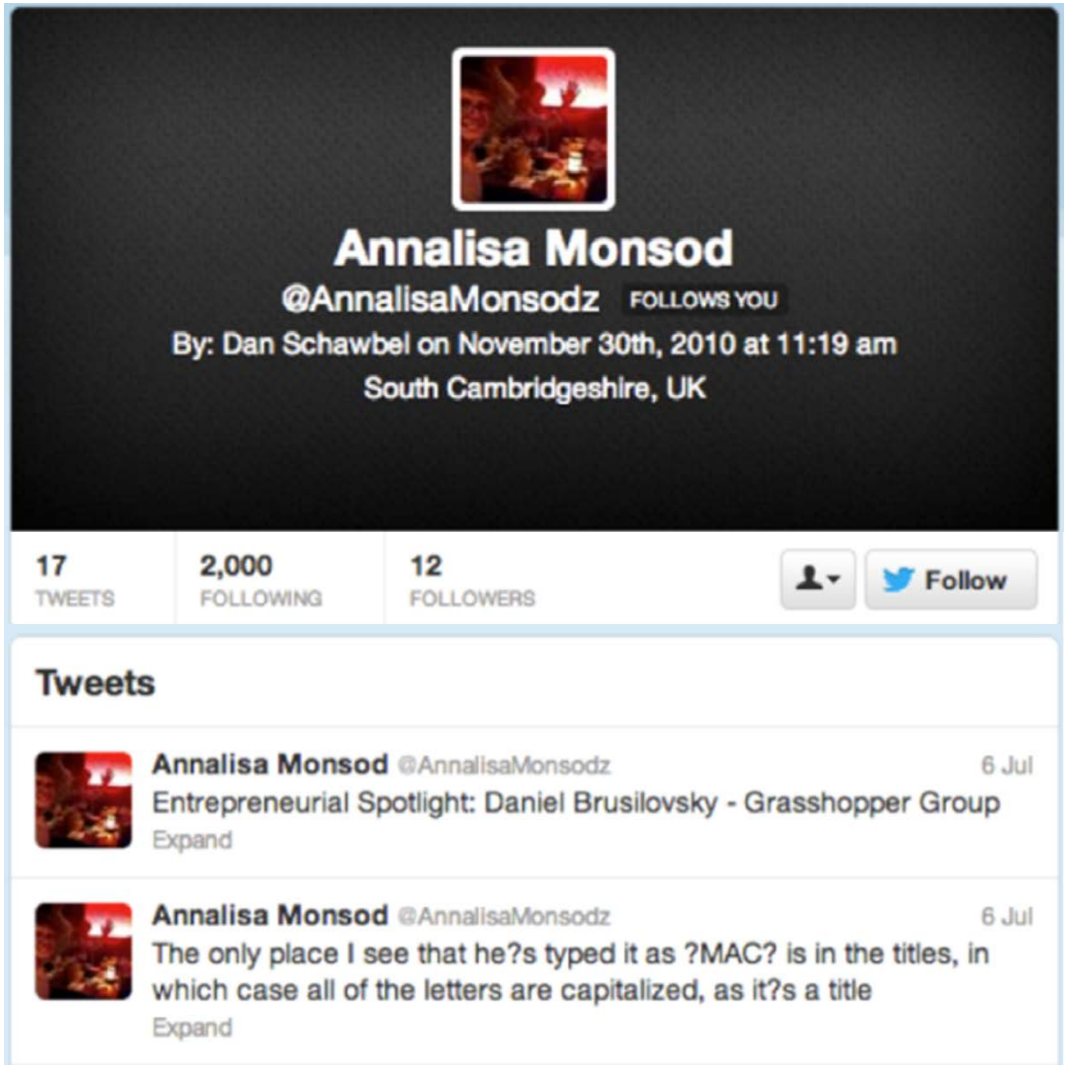
117M



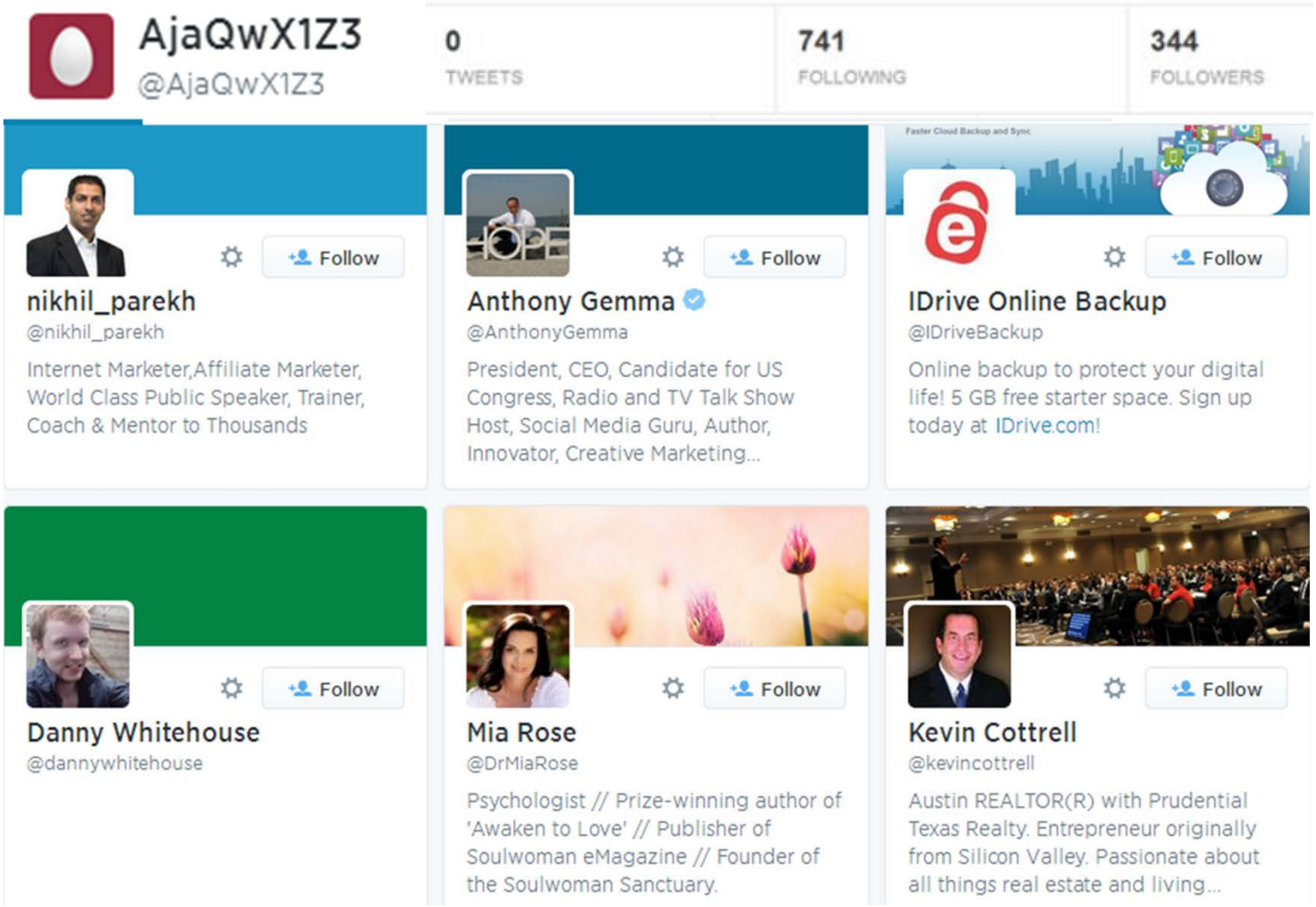
Traditional Fraud Detection



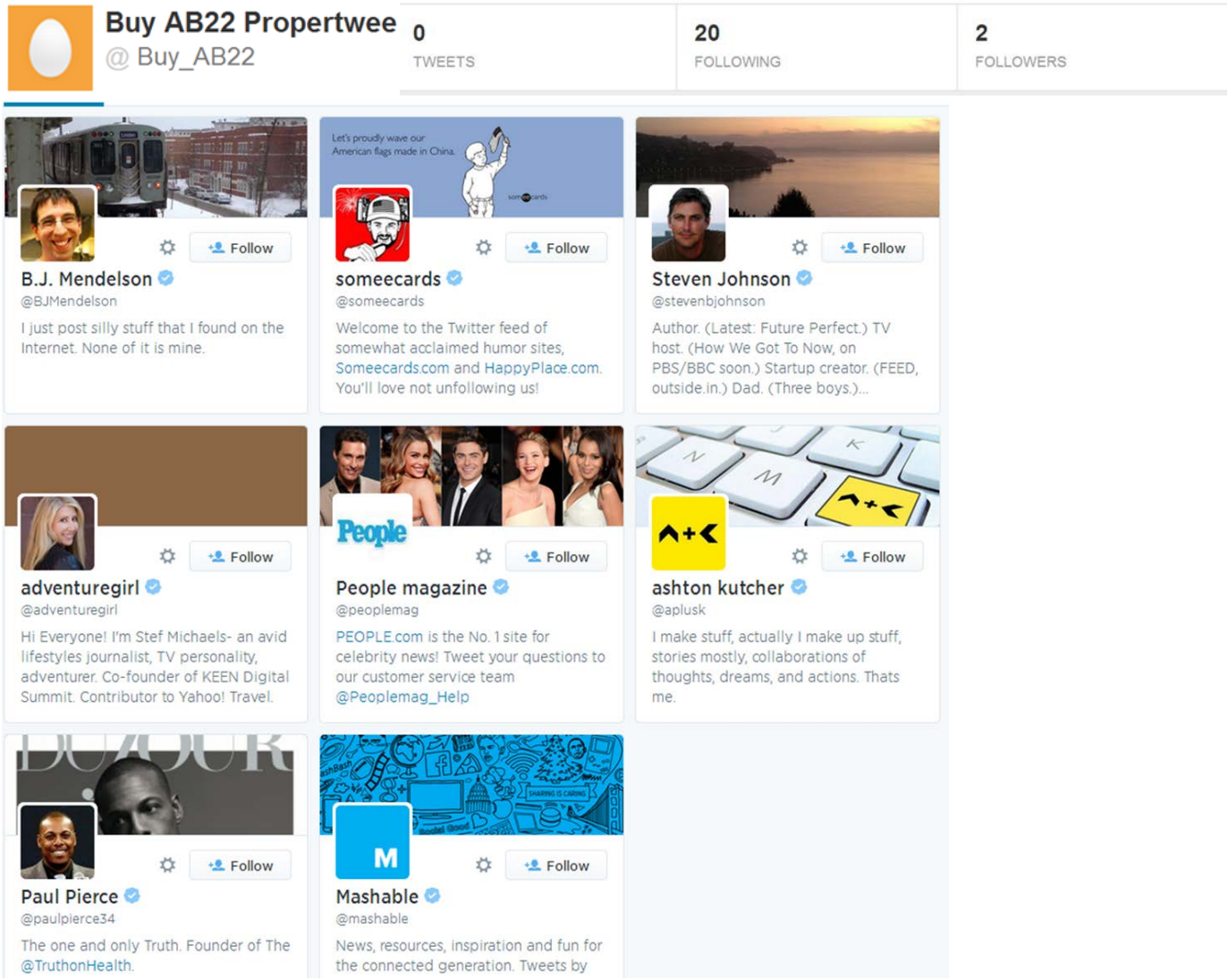
Empty Profile?



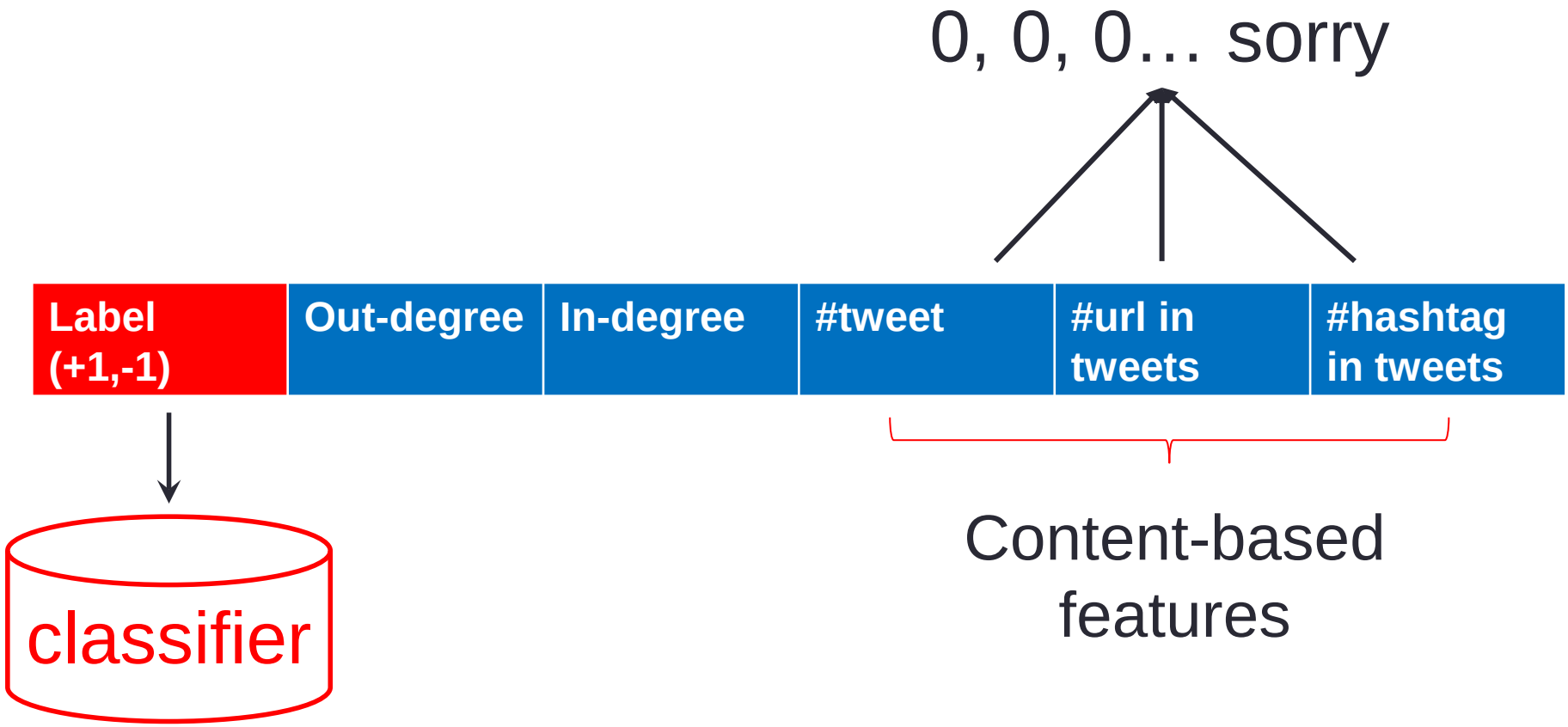
Few Followers?



Many Followings?



Content: Unavailable? Look Normal?



Behavior is the Key

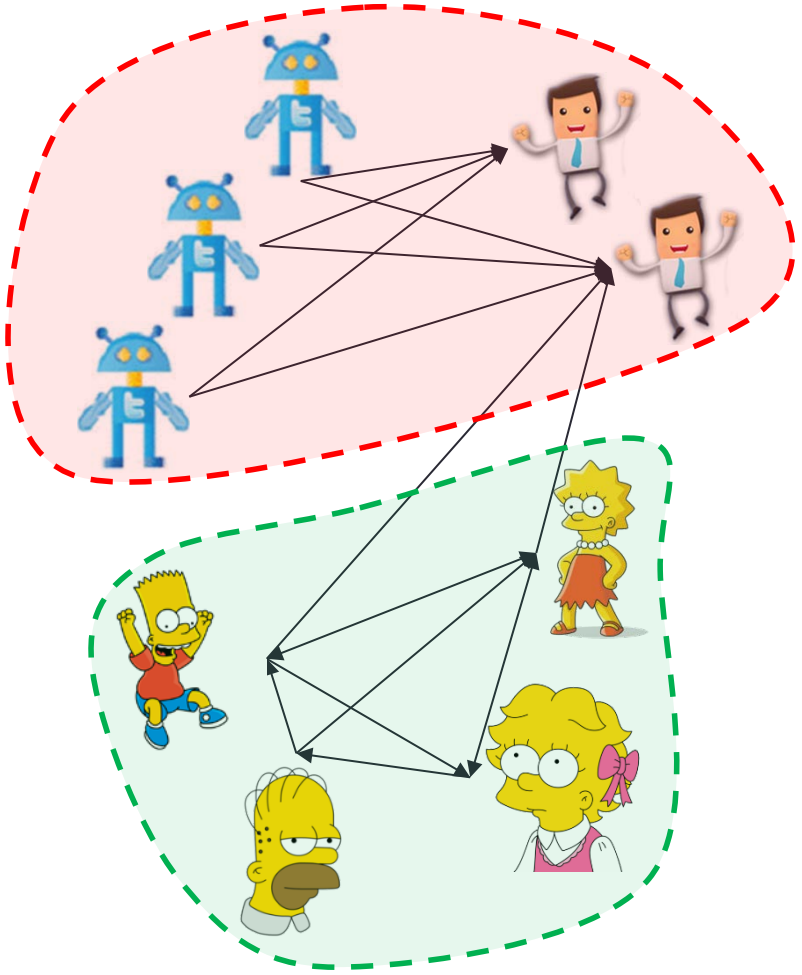
Monetary Incentive

Content

what they **appear** to behave

Behavior/ Links

what they **have** to behave



OUTLINE

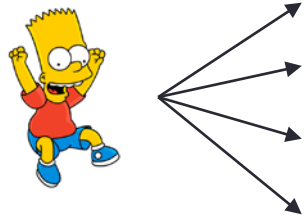
1. Background

2. Fraudulent Pattern

3. The Algorithm

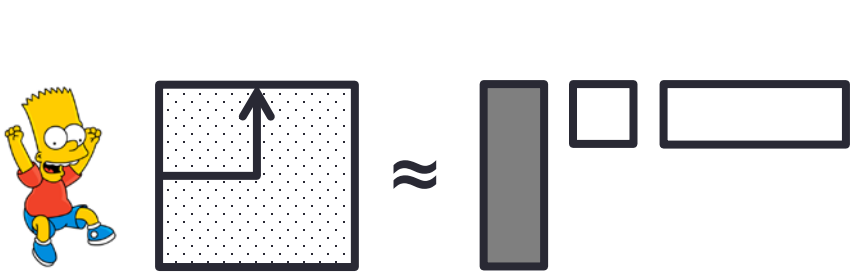
4. Experiments

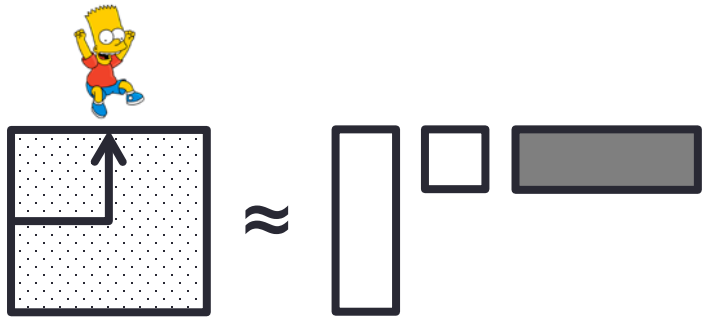
Behavior-based Features



Follower behavior

Followee behavior





Out-degree

1st left singular vector (Hubness)

2nd left singular vector

...

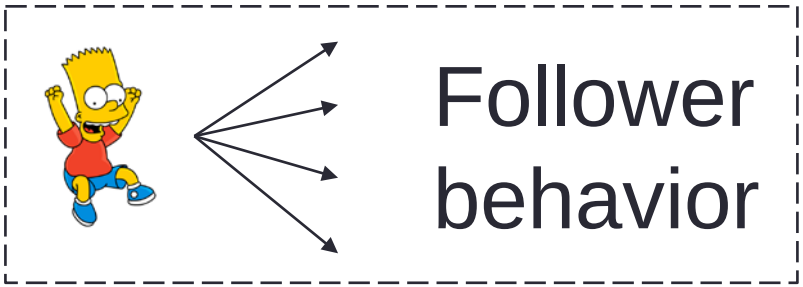
In-degree

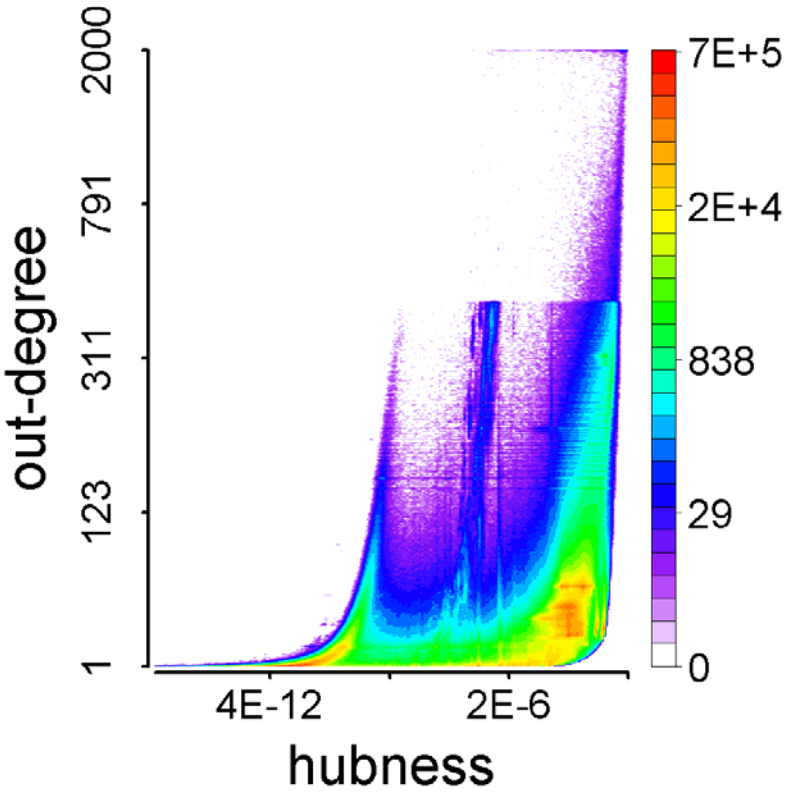
1st right singular vector (Authoritativeness)

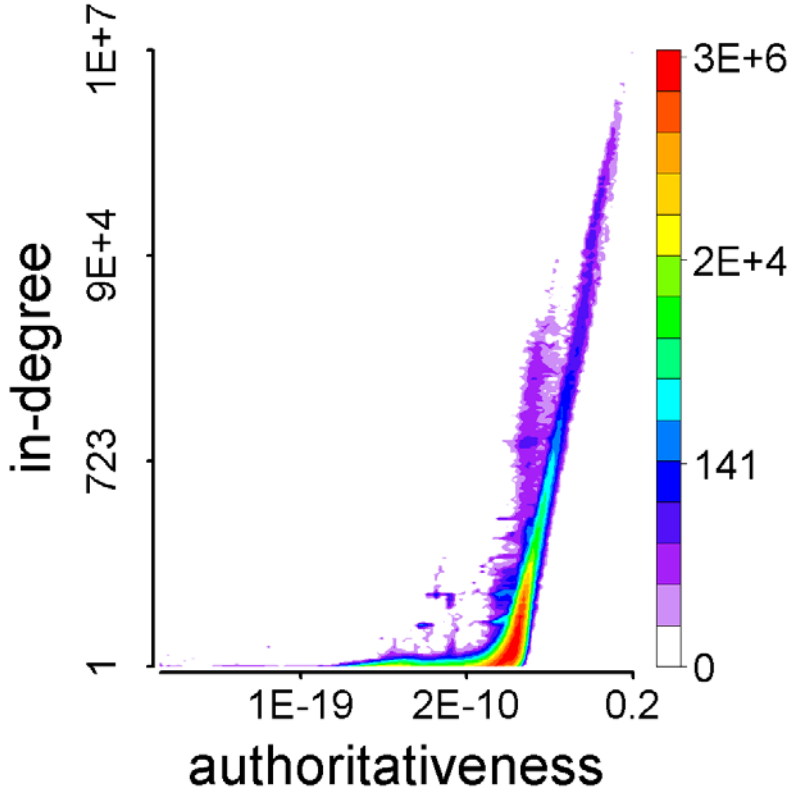
2nd right singular vector

...

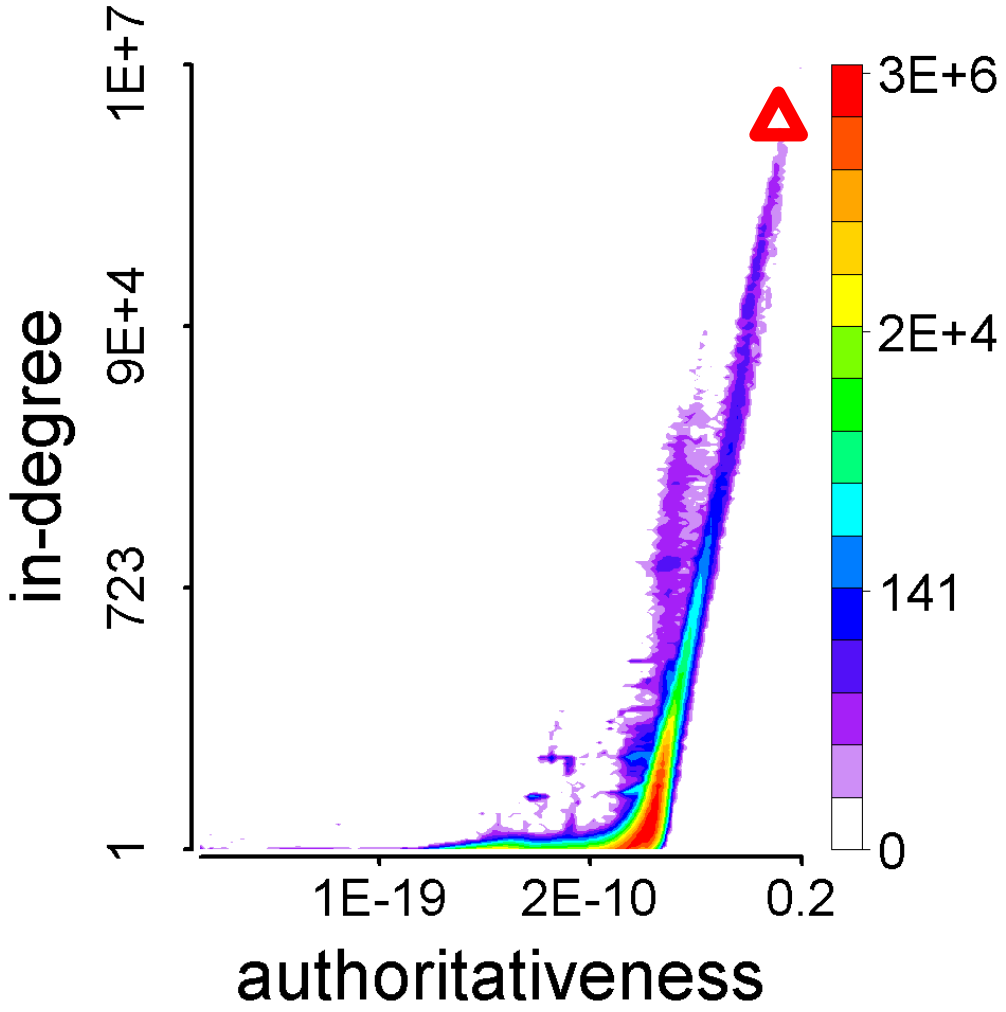
Behavior-based Feature Space







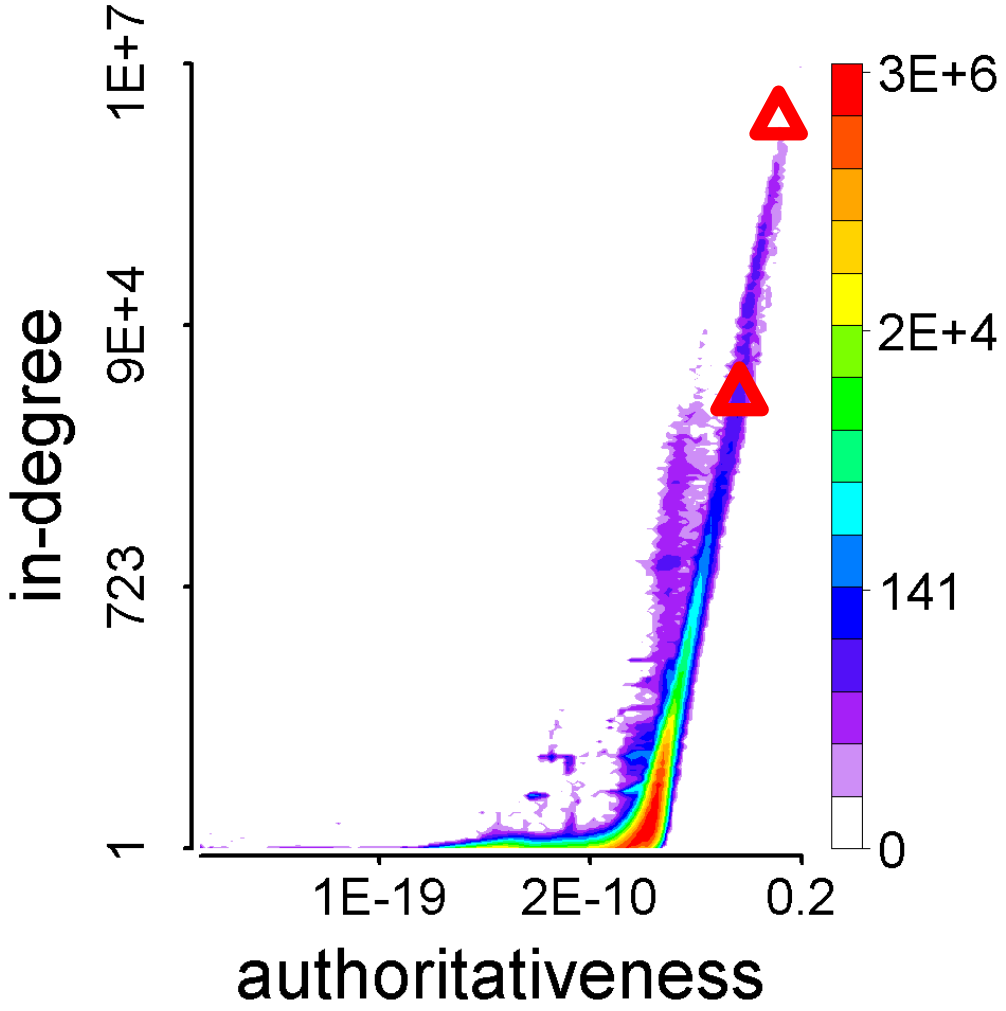
Fraudulent Behavior Patterns



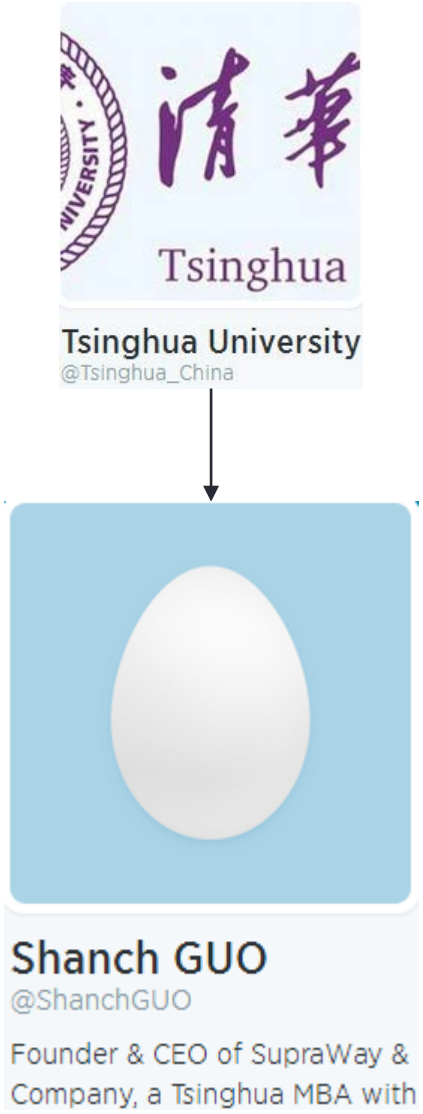
Fraudulent Behavior Patterns

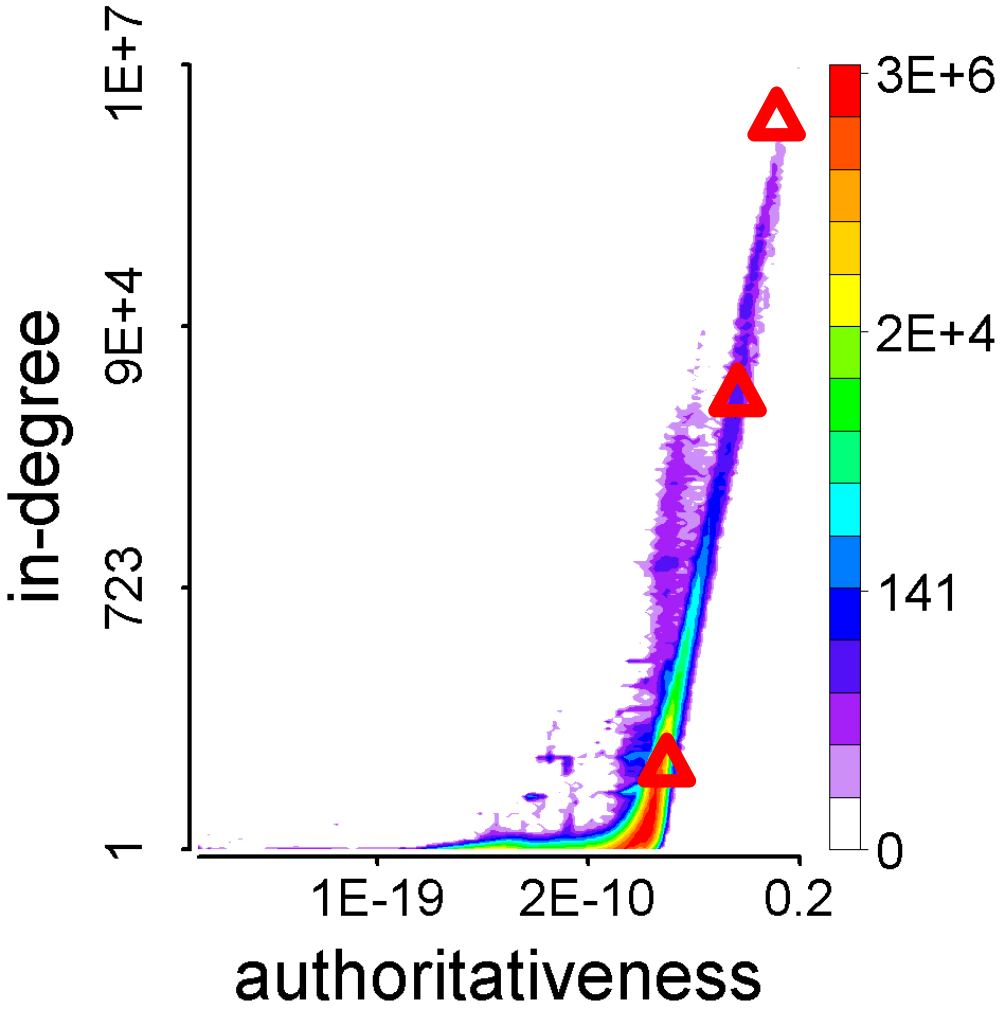
FOLLOWING
117

FOLLOWERS
25.3K

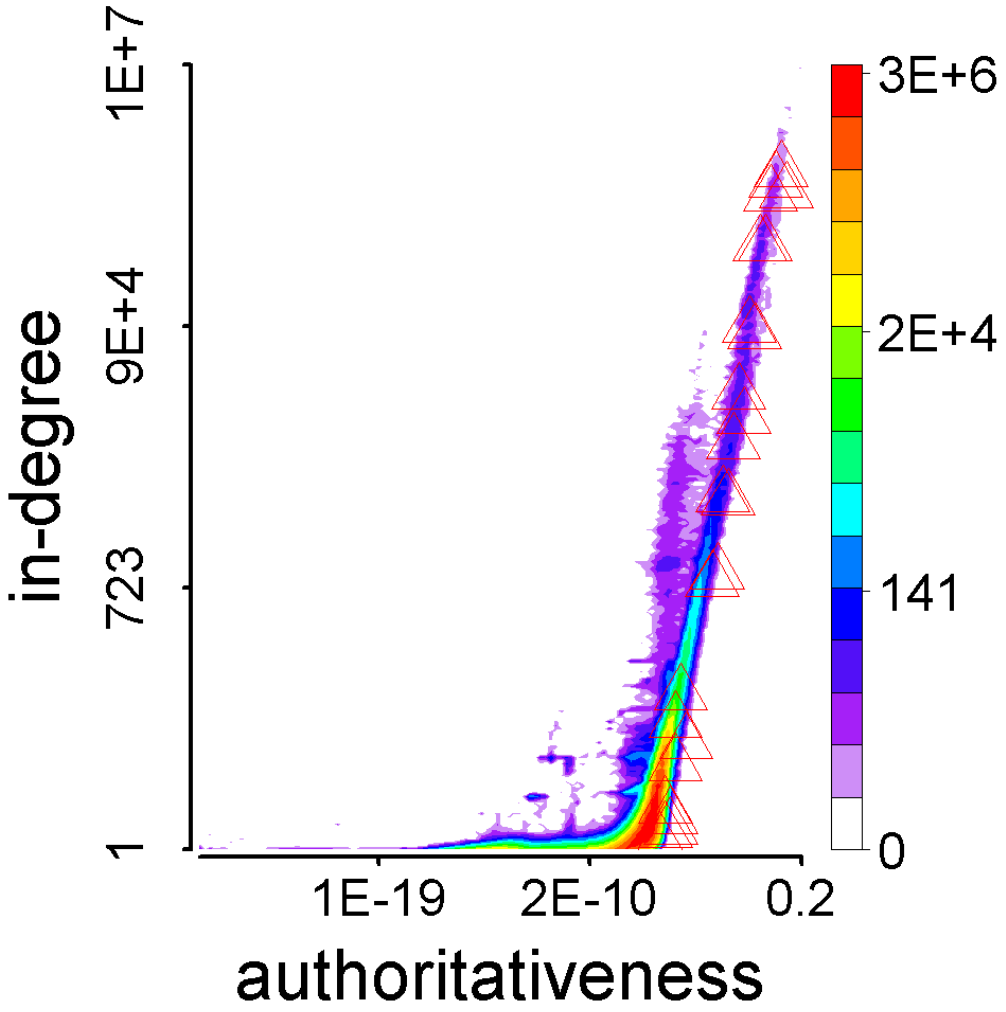


Fraudulent Behavior Patterns

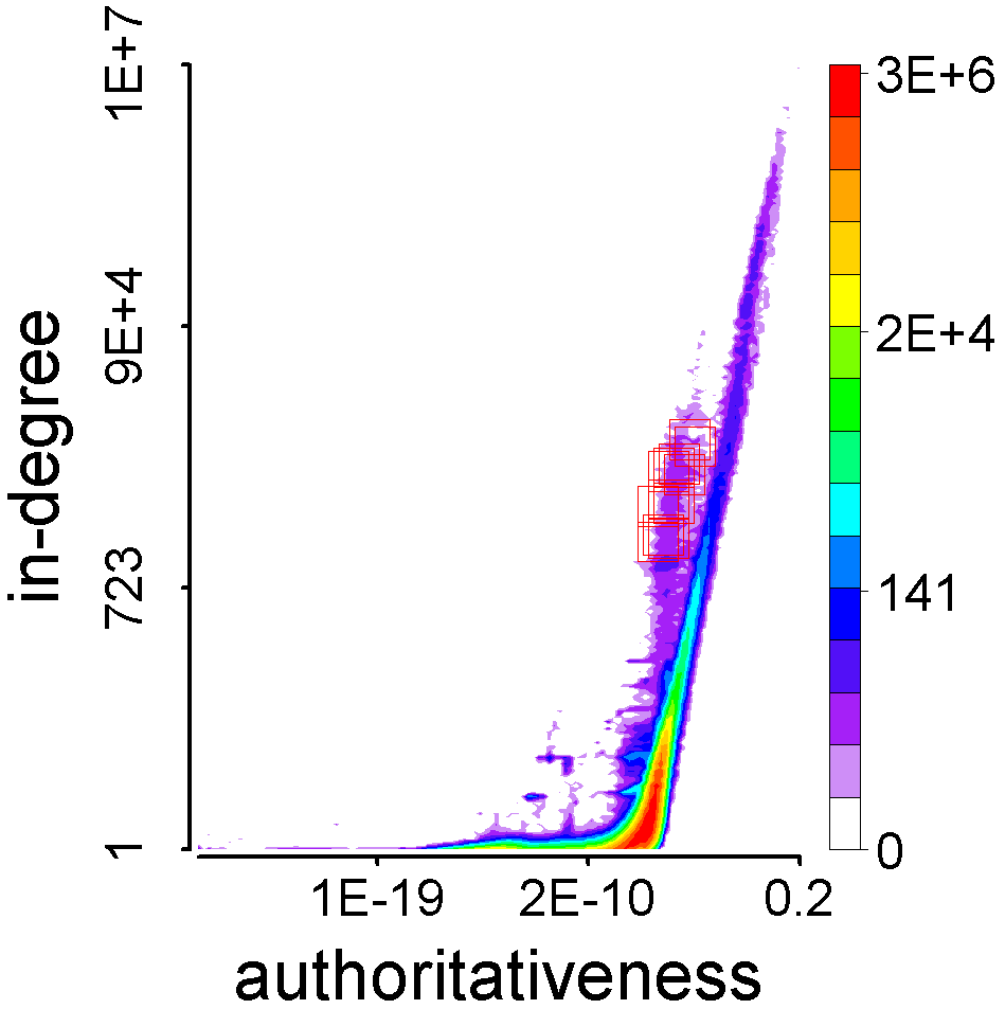




Fraudulent Behavior Patterns

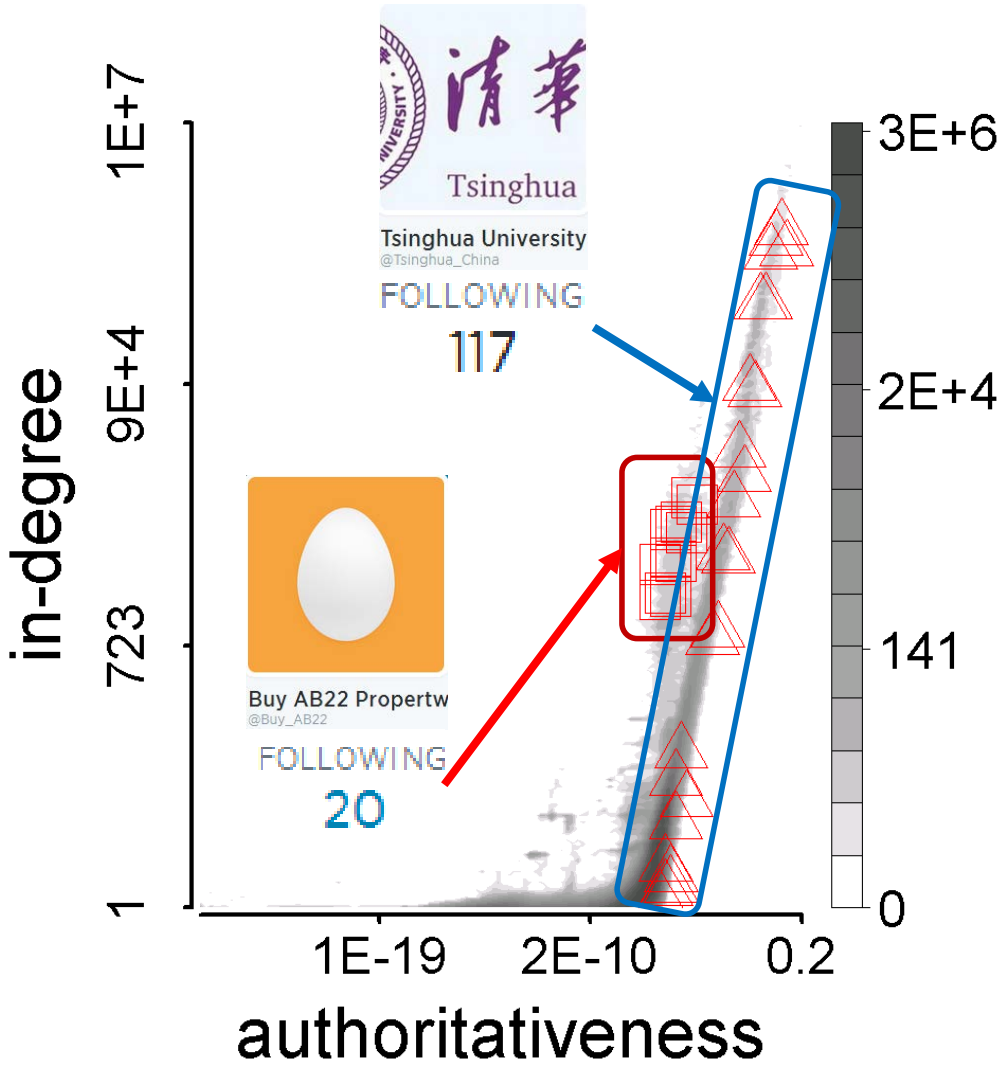


Fraudulent Behavior Patterns



Fraudulent Behavior Patterns

- Synchronized
- Abnormal



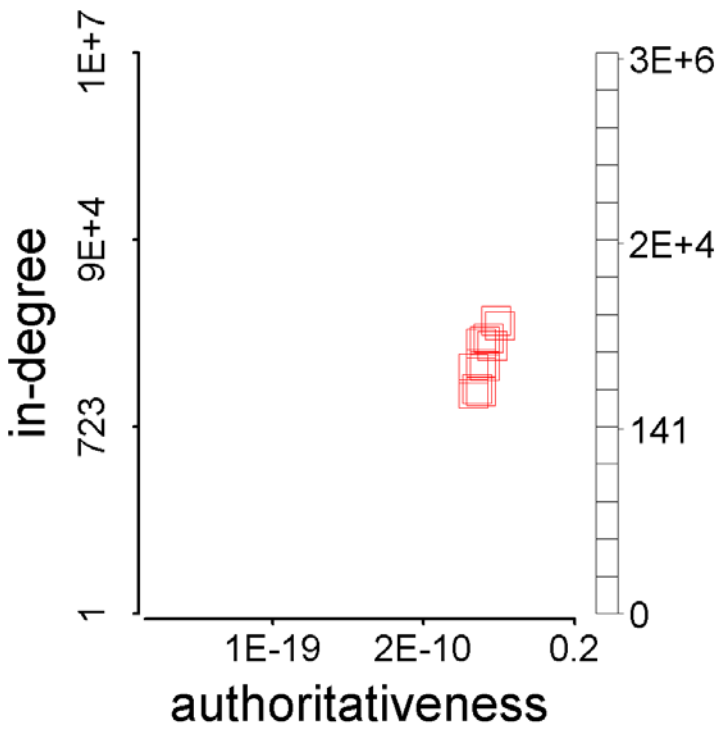
OUTLINE

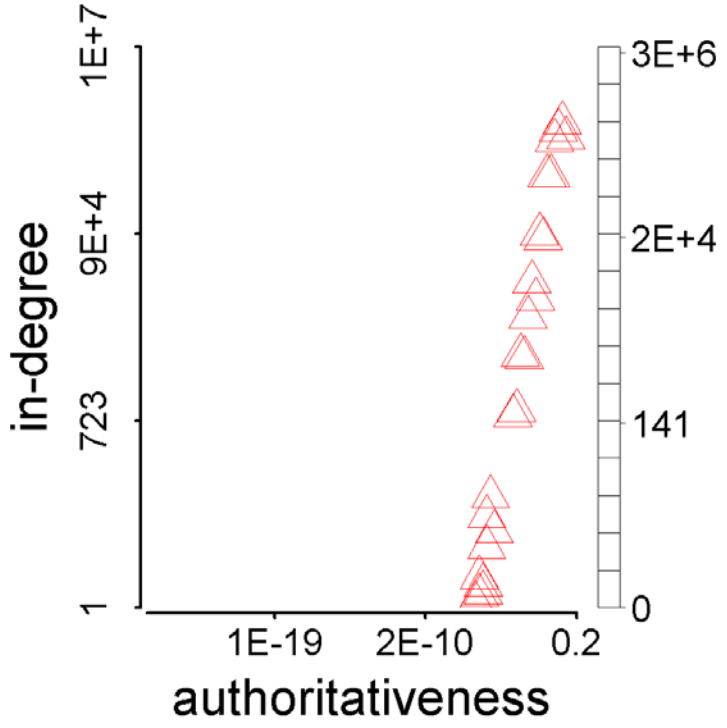
1. Background
2. Fraudulent Pattern
3. **The Algorithm**
4. Experiments

Synchronicity and Normality

- Synchronicity

$$sync(u) = \frac{\sum_{(v,v') \in \mathcal{F}(u) \times \mathcal{F}(u)} \mathbf{p}(v) \cdot \mathbf{p}(v')}{d(u) \times d(u)}$$

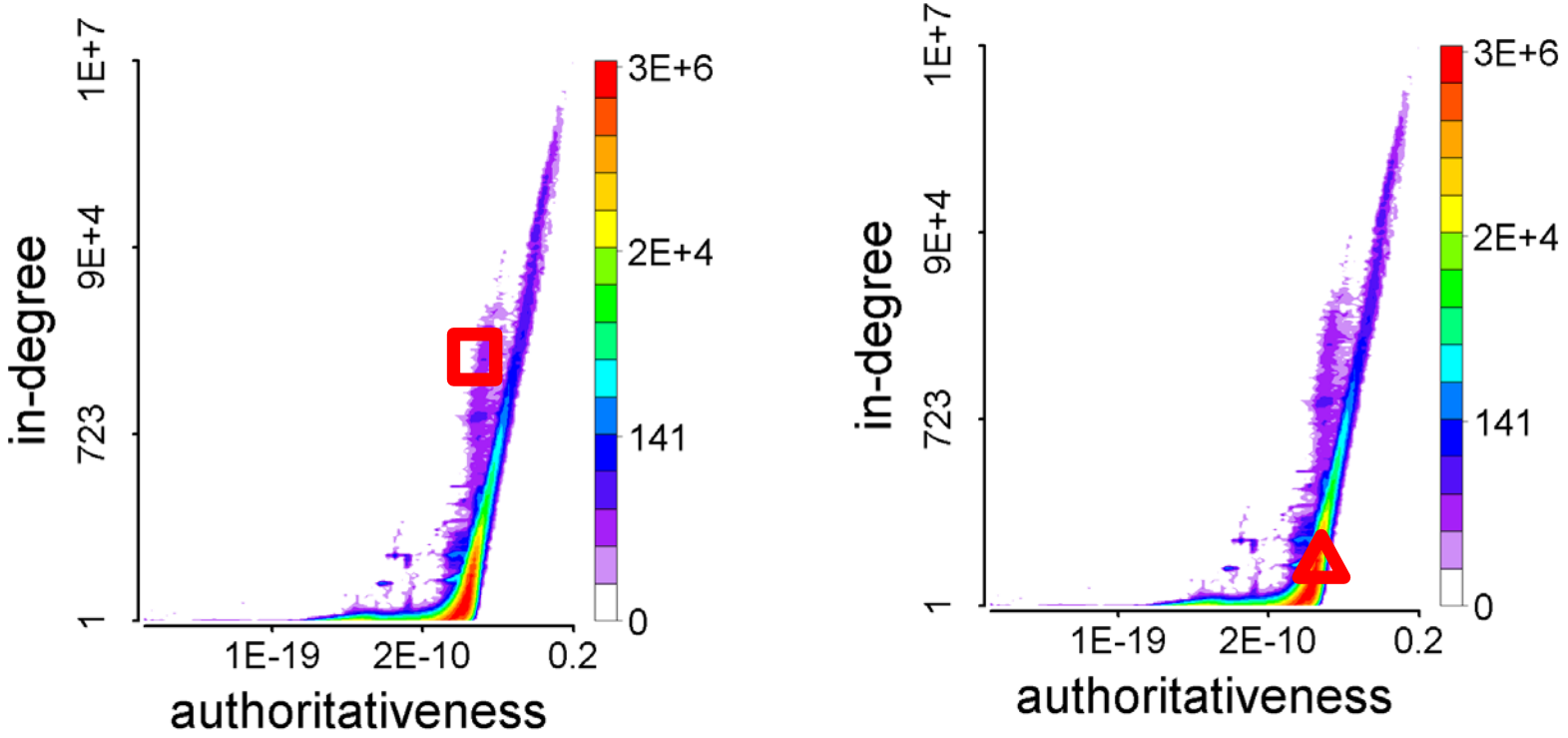




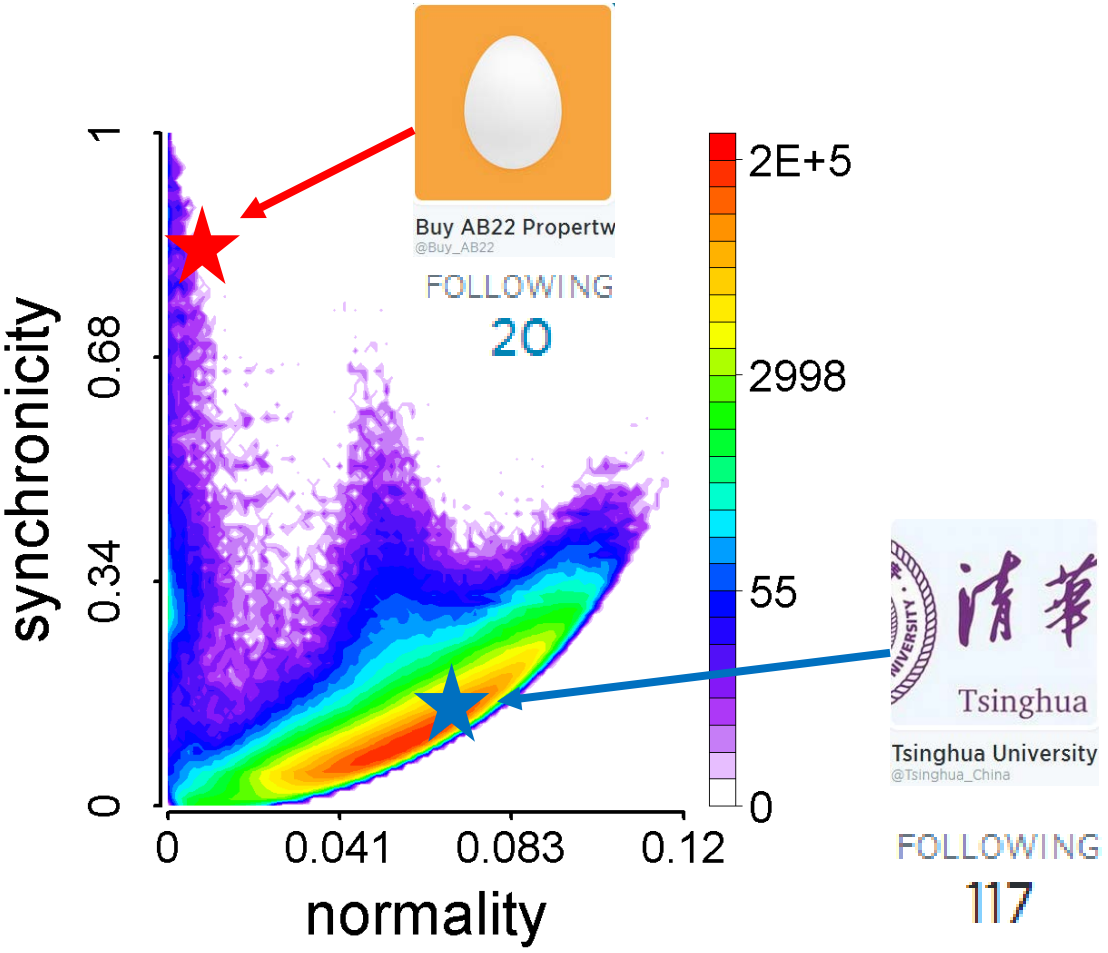
Synchronicity and Normality

- Normality

$$norm(u) = \frac{\sum_{(v,v') \in \mathcal{F}(u) \times \mathcal{U}} \mathbf{p}(v) \cdot \mathbf{p}(v')}{d(u) \times N}$$



Synchronicity-Normality Plot



Theorem

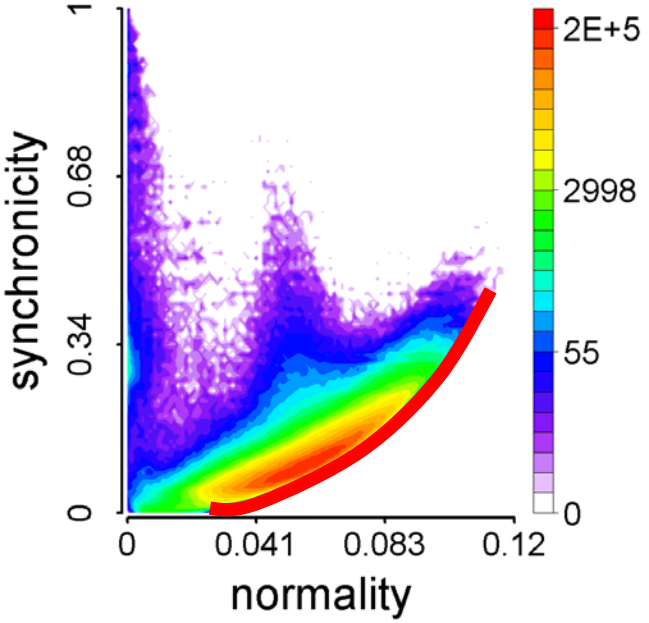
- *For any distribution, there is a **parabolic lower limit** in the synchronicity-normality plot.*

$$s_{min} = (-Mn^2 + 2n - s_b)/(1 - Ms_b)$$

synchronicity

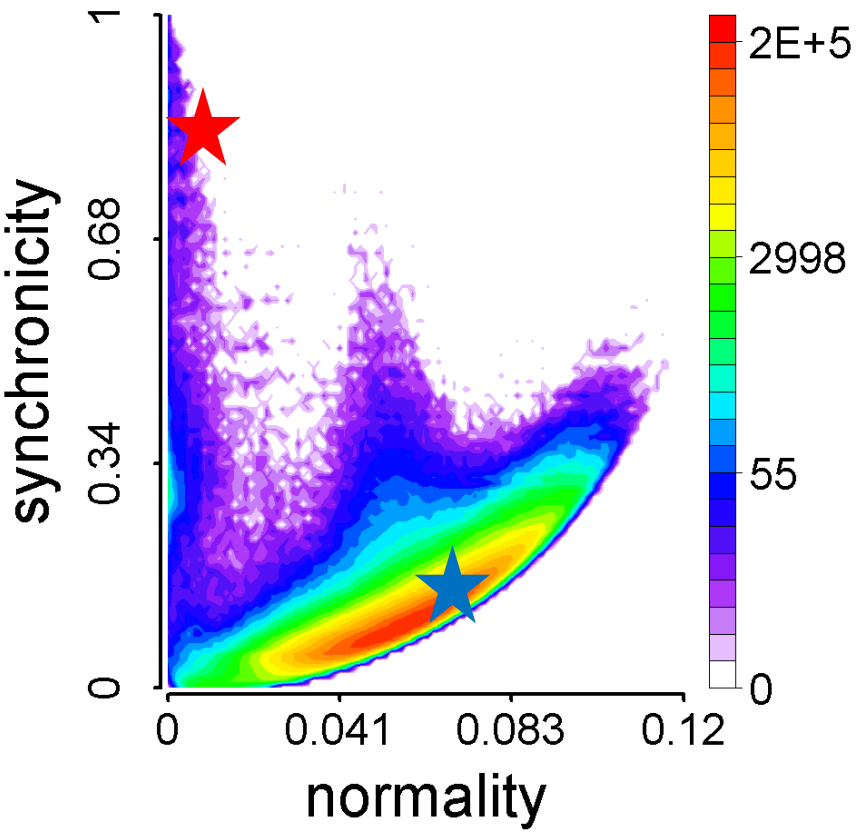
normality

- *Proof.* See our paper ☺



CatchSync Algorithm

- Distance-based anomaly detection
- Fraudsters
  - Big synchronicity
  - Small normality
  - Away from the densest



OUTLINE

1. Background
2. Fraudulent Pattern Mining
3. The Algorithm
4. **Experiments**

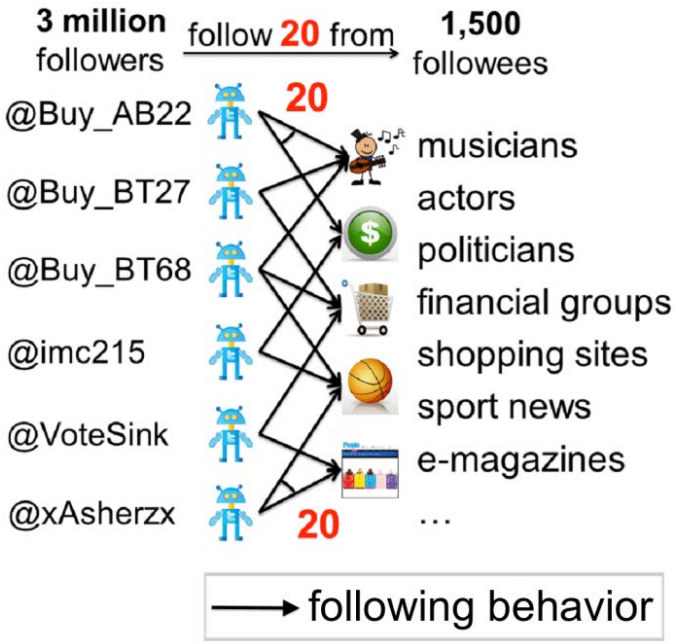
Experiments

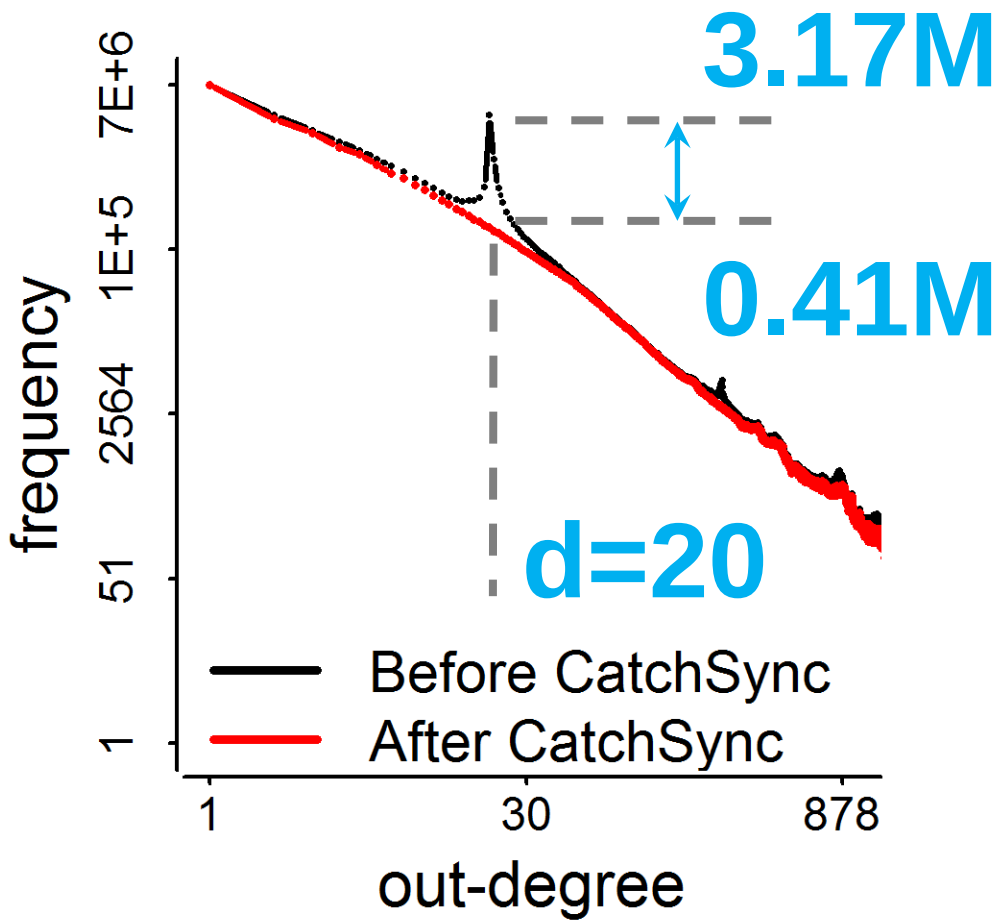
- Q1: Does CatchSync remove anomalies?
  - Degree distribution
  - Feature space
- Q2: Is CatchSync catching actually fraudulent users?
- Q3: Is CatchSync robust?

Q1: Does CatchSync Remove Anomalies?

2009

41M

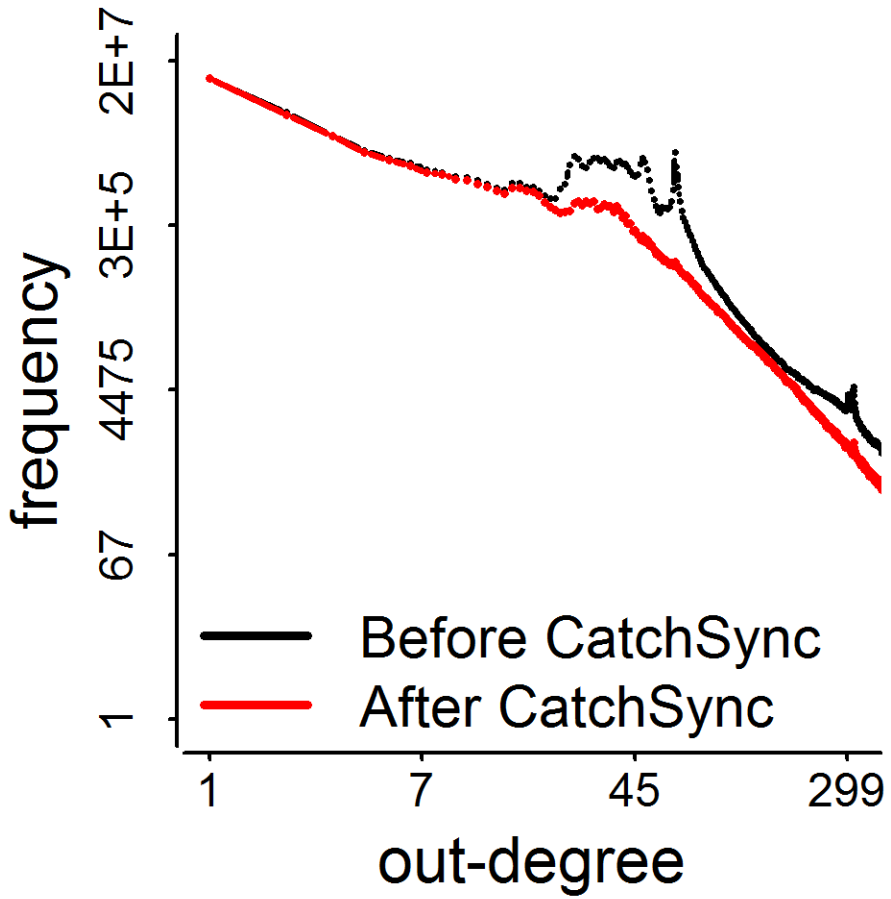




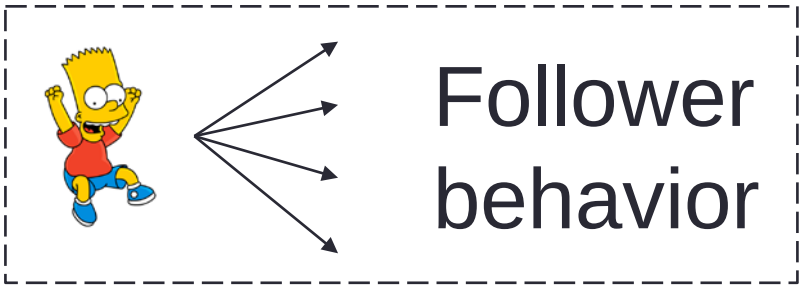
Q1: Does CatchSync Remove Anomalies?

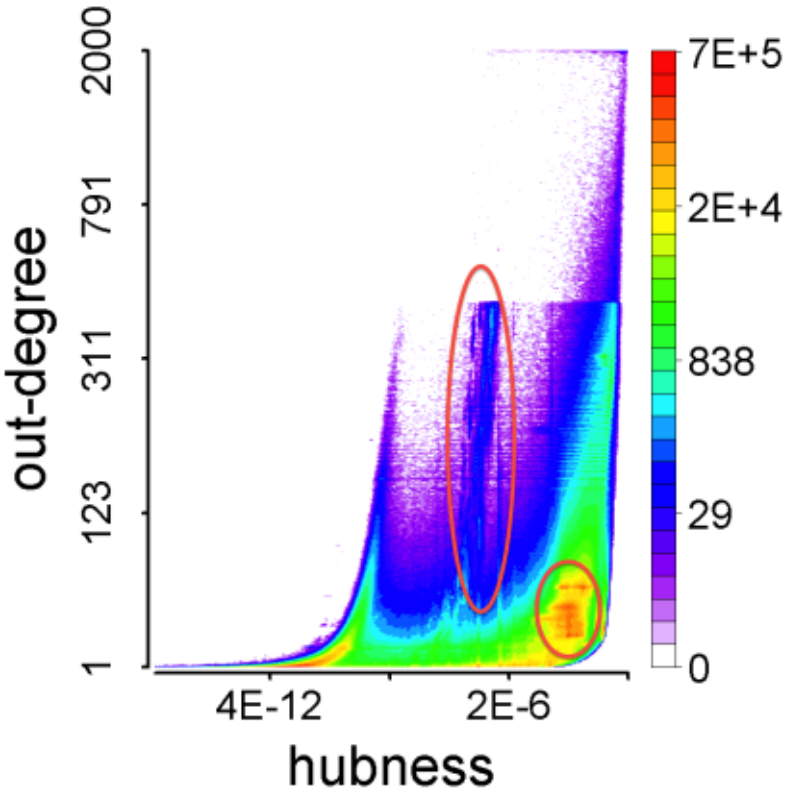
2011

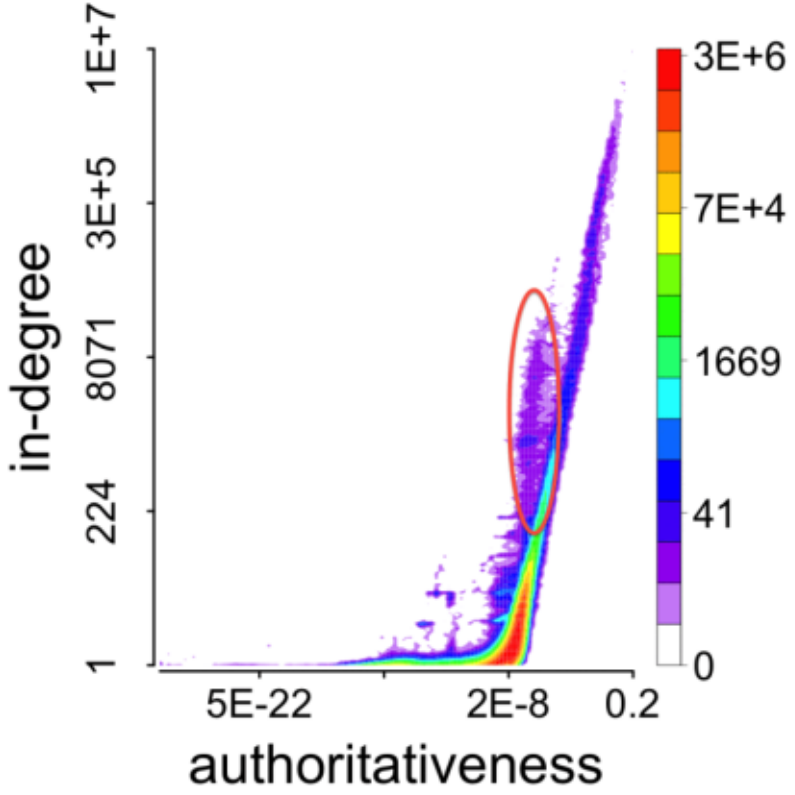
117M



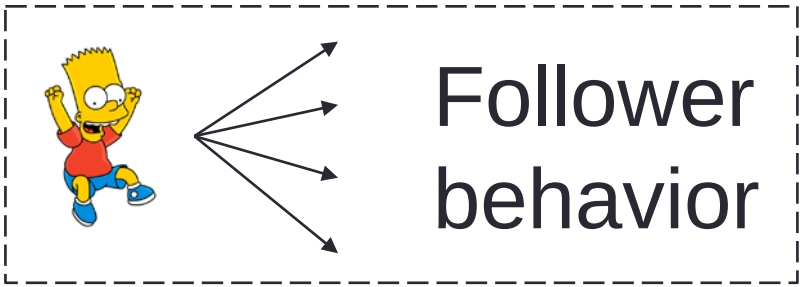
Before CatchSync

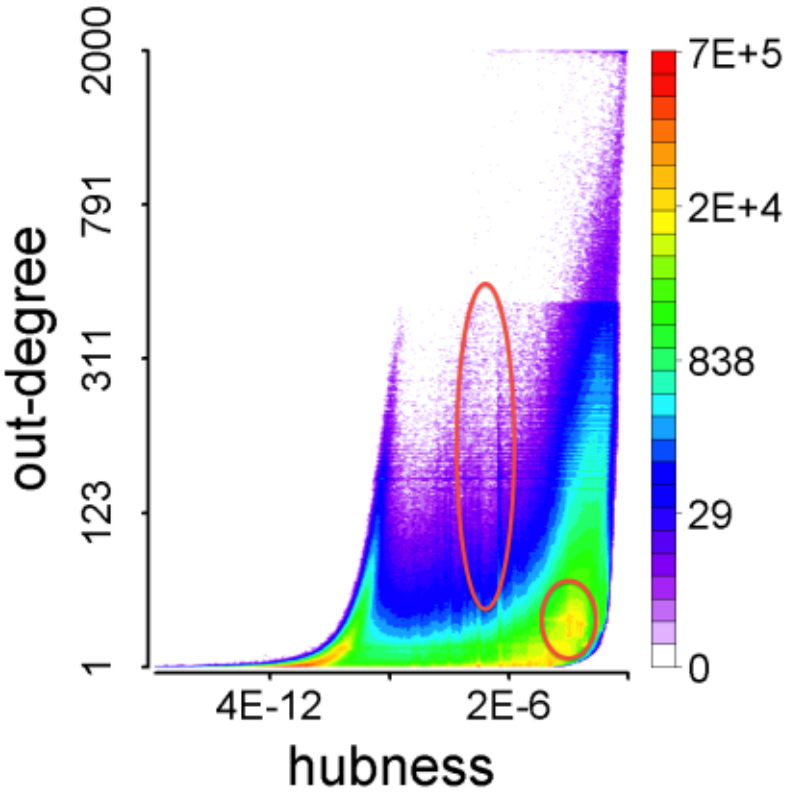


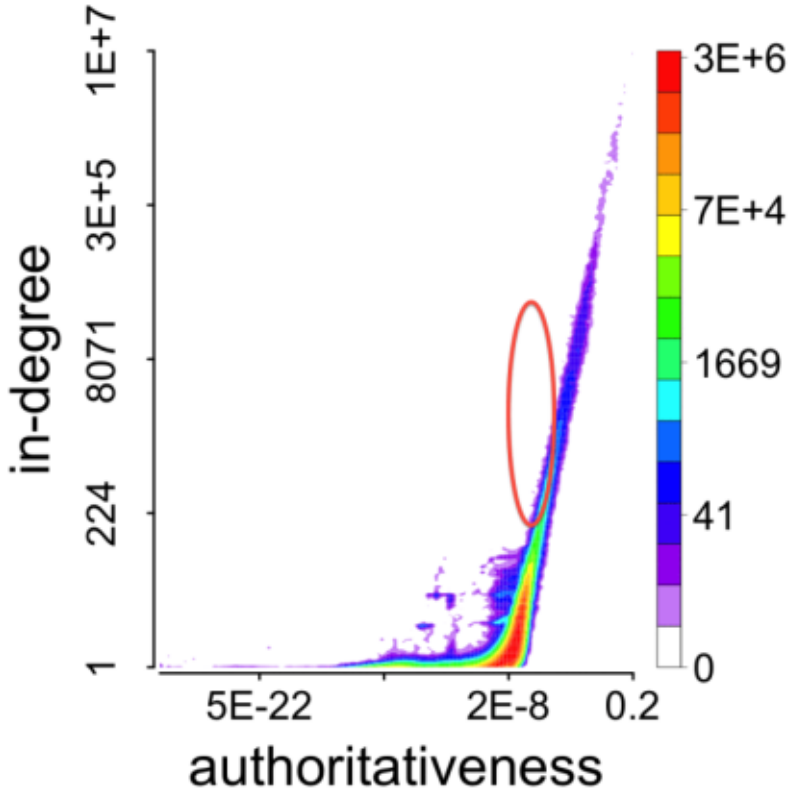




After CatchSync

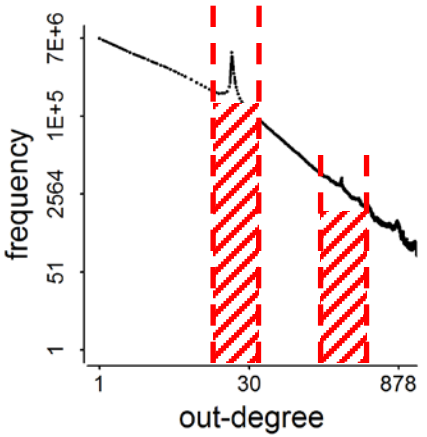




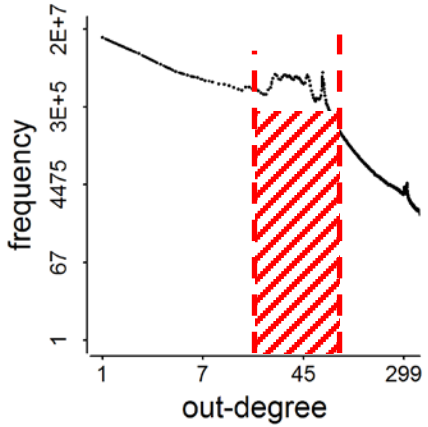


Q2: Is CatchSync Catching Actually Fraudulent Users?

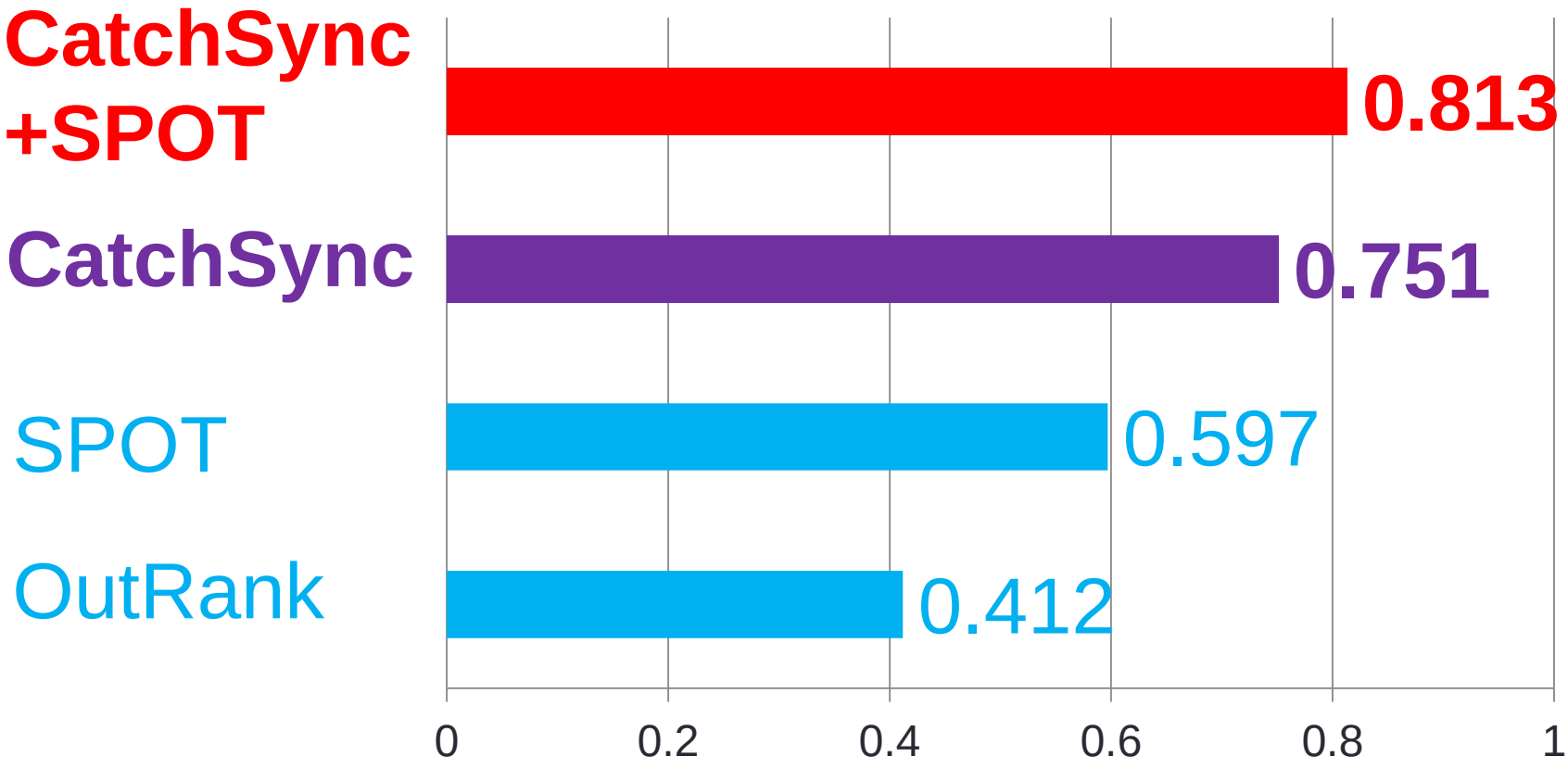
173/1,000



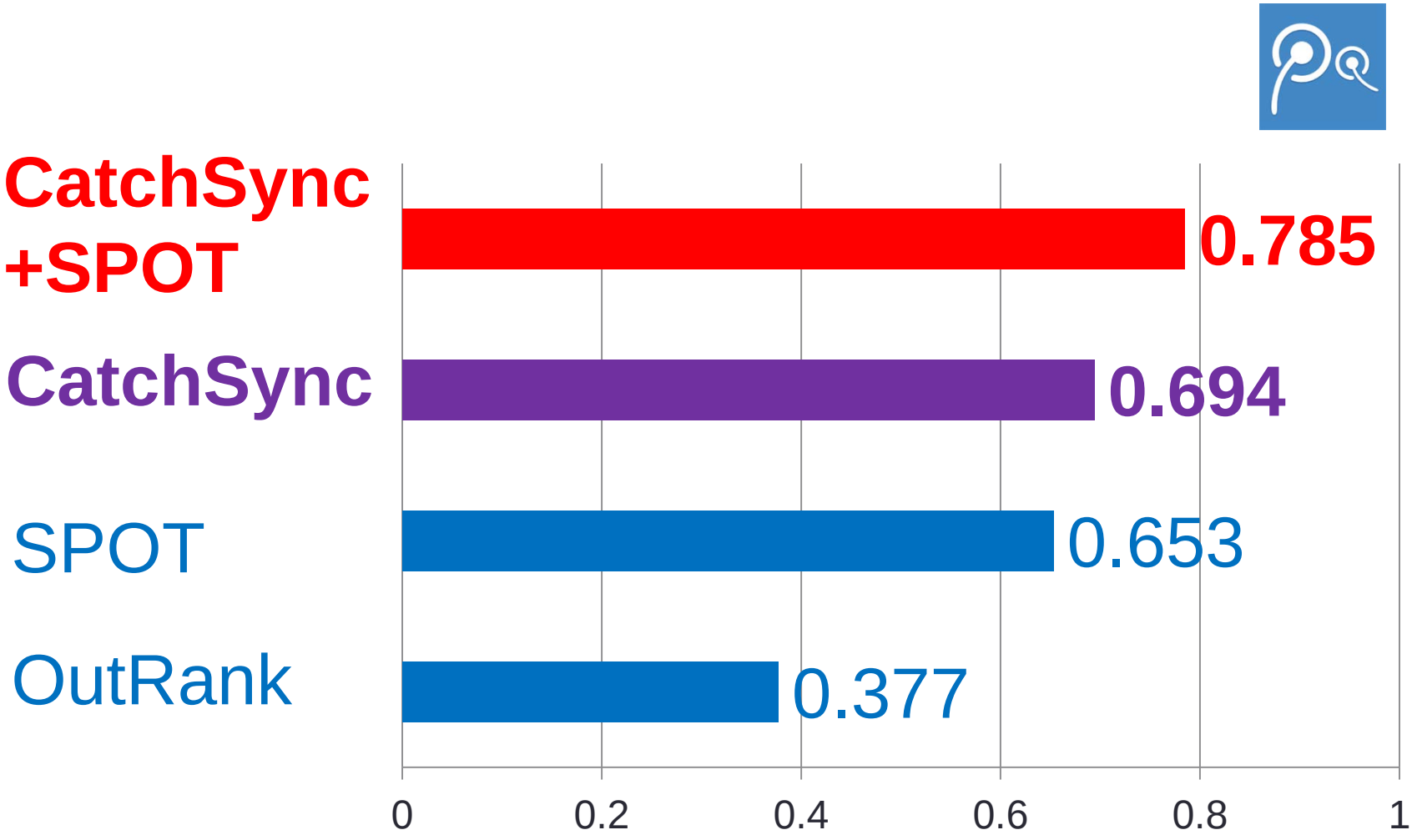
237/1,000



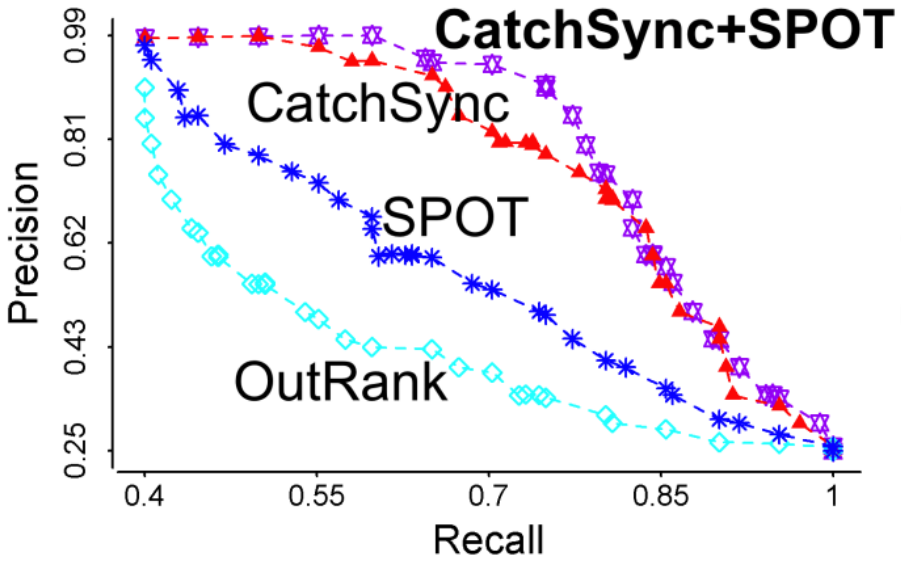
Q2: Is CatchSync Catching Actually Fraudulent Users?

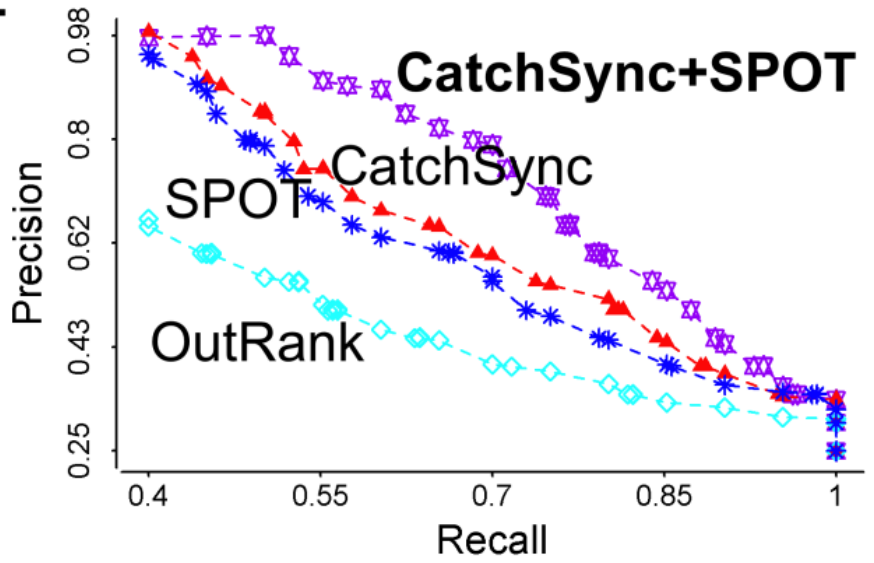


Q2: Is CatchSync Catching Actually Fraudulent Users?



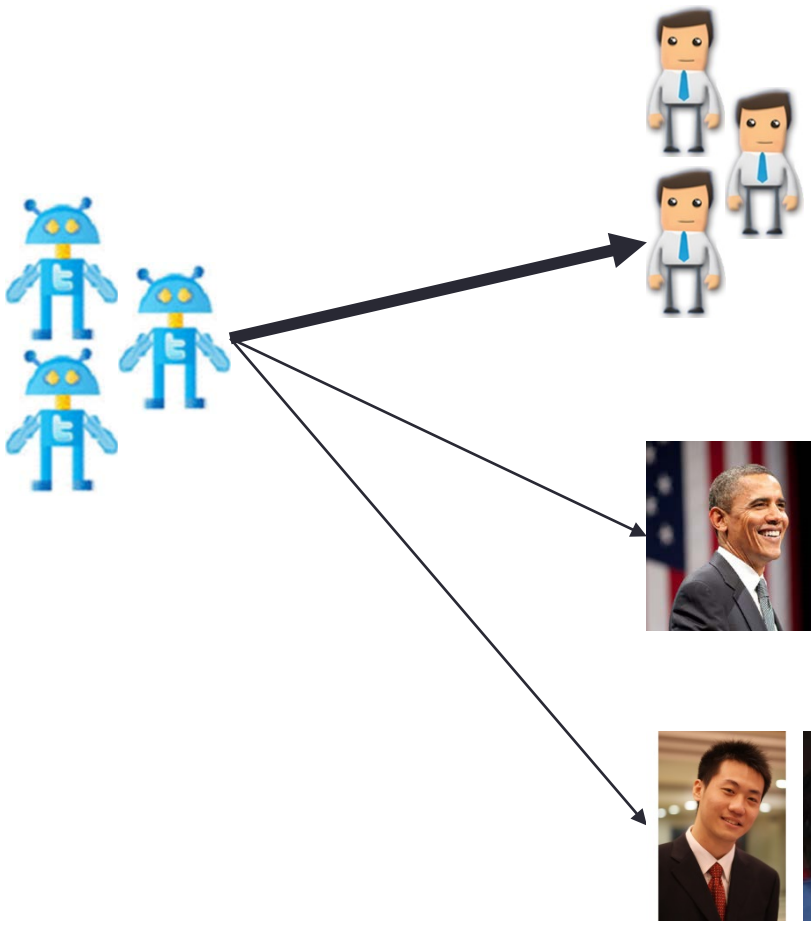
Q2: Is CatchSync Catching Actually Fraudulent Users?





Recall = 80%	Precision in Twitter	Precision in Tencent Weibo
	83.5%	79.4%

Q3: Is CatchSync Robust to Camouflage?

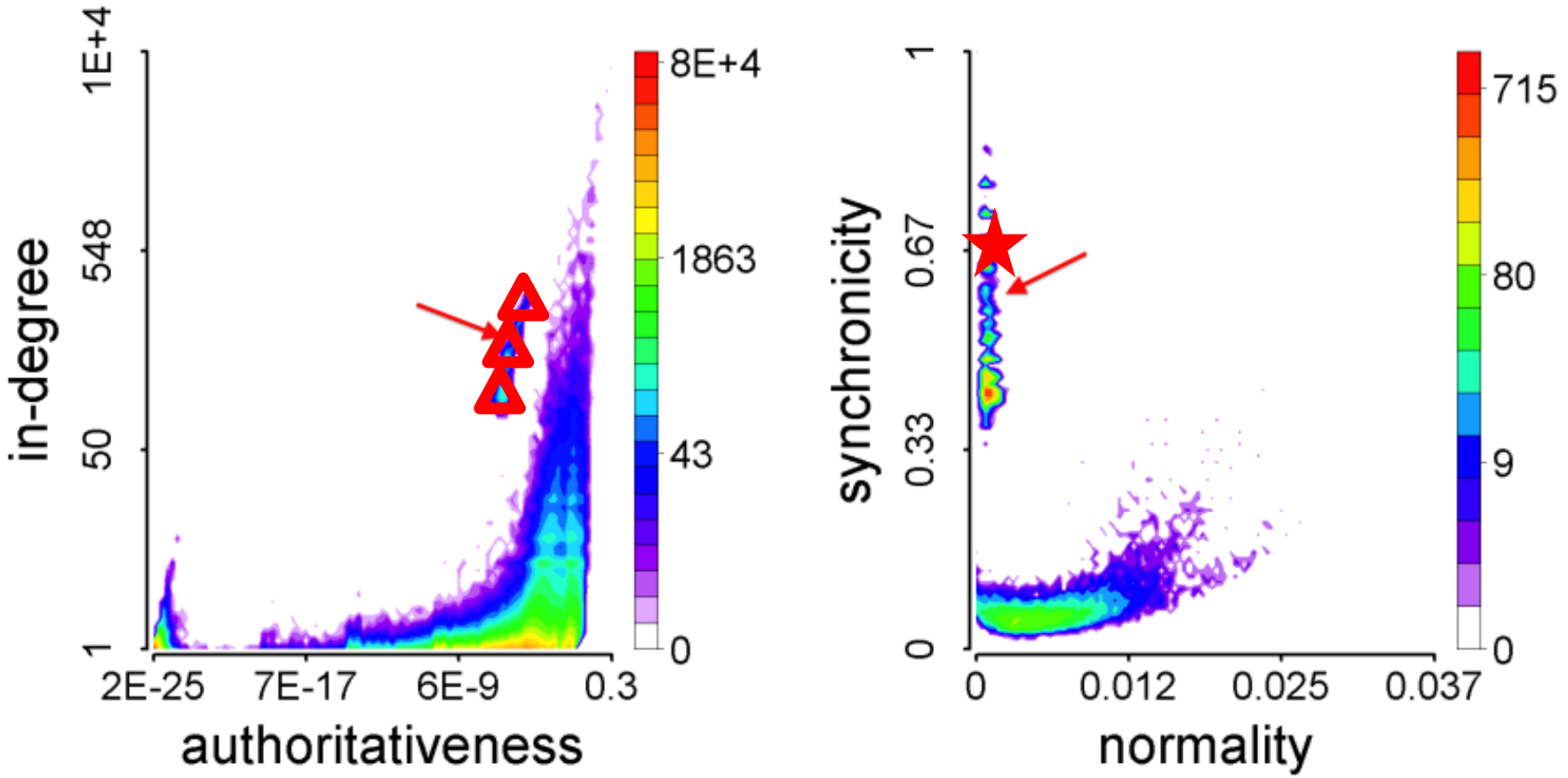


Target

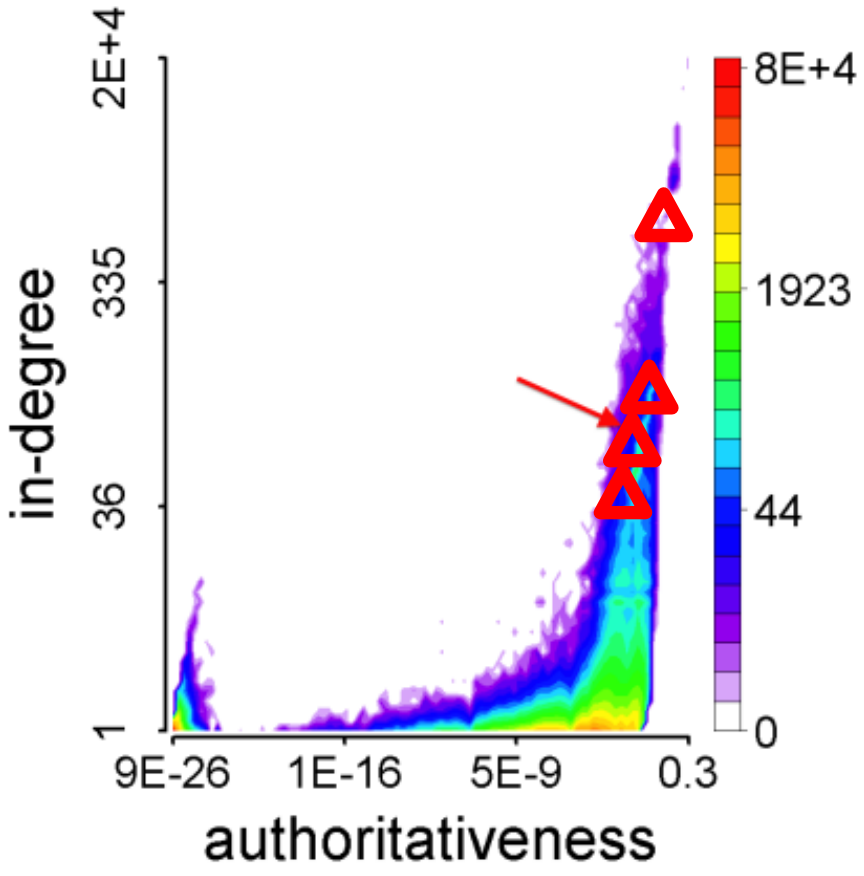
Popular camouflage

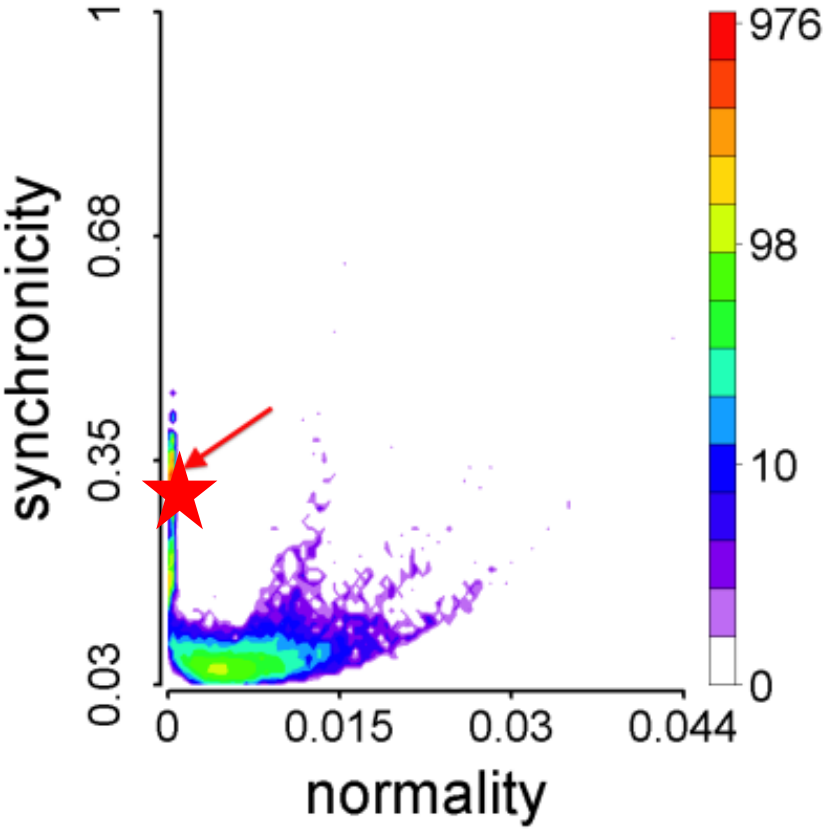
Random camouflage

Q3: Is CatchSync Robust to Camouflage?

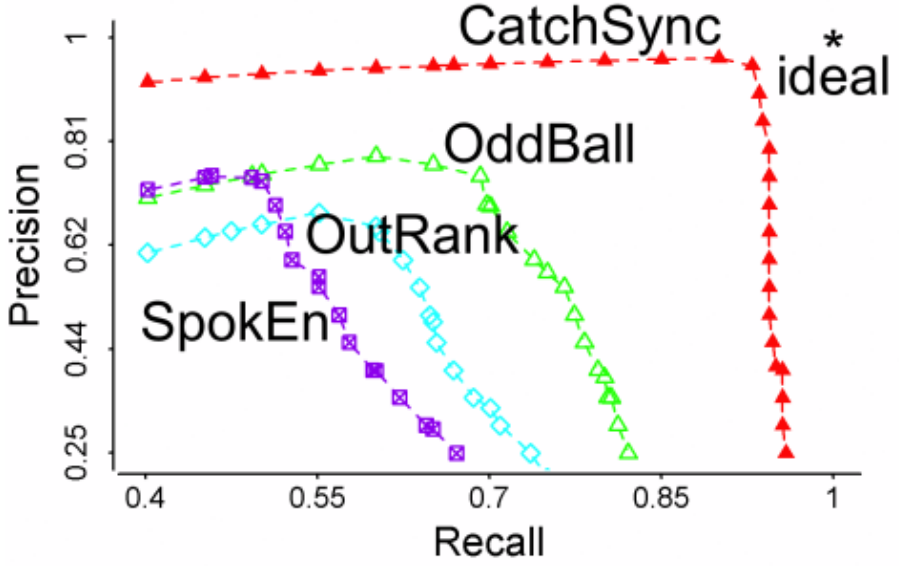


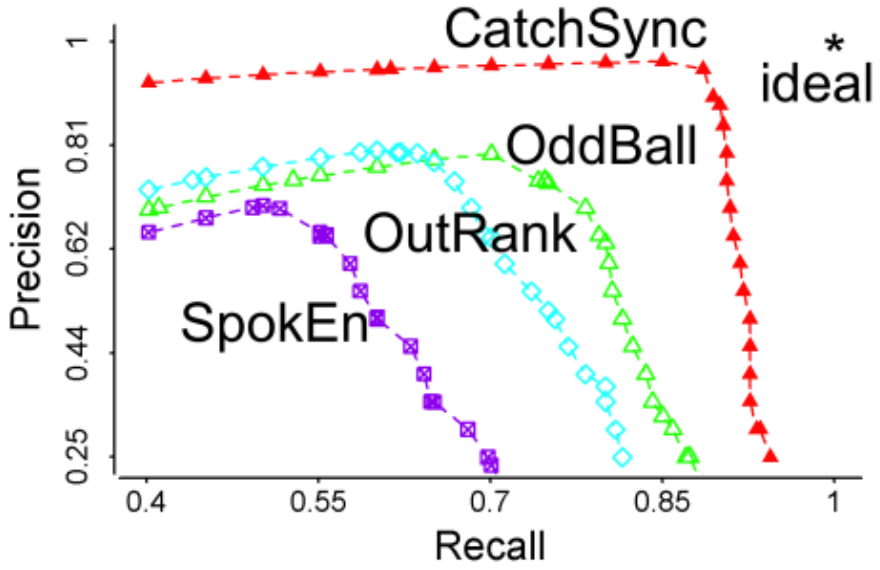
Q3: Is CatchSync Robust to Camouflage?





Q3: Is CatchSync Robust to Camouflage?





Conclusion

- Goals
  - G1. Find patterns that distinguish fraudulent user behavior from normal behavior
  - **A1: Synchronized & Abnormal!**
  - G2. Design algorithms that catch fraudsters
  - **A2: CatchSync!**
    - **Remove spikes**
    - **Content free**
    - **Robust to camouflage**

Questions?




Meng Jiang
mjiang89@gmail.com
http://www.meng-jiang.com