Chapter 11. Outlier Analysis: Methods

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

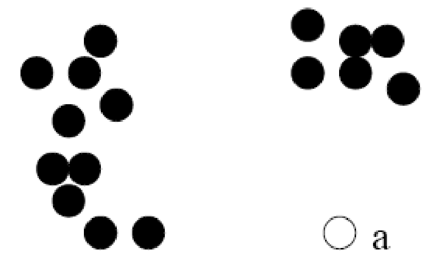
Outlier Analysis

- Basic Concepts
- Outlier Detection Methods
- Statistical Approaches
- **Clustering-Based Approaches**
- Classification-Based Approaches

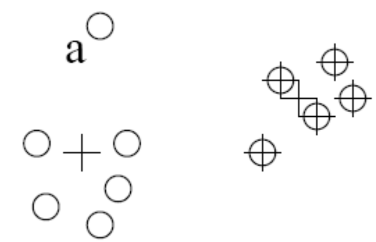
Clustering-Based Outlier Detection (1 & 2):

Not belong to any cluster, or far from the closest one

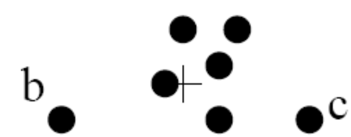
- An object is an outlier if (1) it does not belong to any cluster, (2) there is a large distance between the object and its closest cluster , or (3) it belongs to a small or sparse cluster
- Case I: Not belong to any cluster
  - Identify animals not part of a flock: Using a density-based clustering method such as DBSCAN



- Case 2: Far from its closest cluster
  - Using k-means, partition data points of into clusters
  - For each object o, assign an outlier score based on its distance from its closest center
    - If dist(o, co)/avg_dist(co) is large, likely an outlier

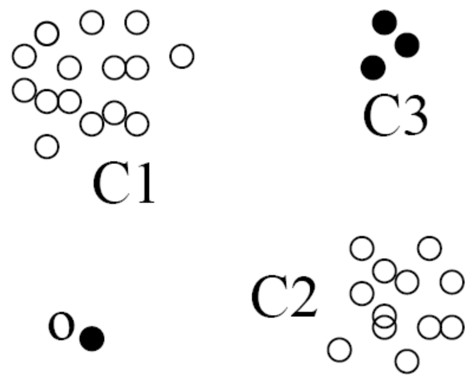


- Ex. Intrusion detection: Consider the similarity between data points and the clusters in a training data set
  - Use a training set to find patterns of “normal” data, e.g., frequent itemsets in each segment, and cluster similar connections into groups
  - Compare new data points with the clusters mined—Outliers are possible attacks



Clustering-Based Outlier Detection (3): Detecting Outliers in Small Clusters

- FindCBLOF: Detect outliers in small clusters
  - Find clusters, and sort them in decreasing size
  - To each data point, assign a cluster-based local outlier factor (CBLOF):
  - If obj p belongs to a large cluster, CBLOF = cluster_size X similarity between p and cluster
  - If p belongs to a small one, CBLOF = cluster size similarity betw. p and the closest large cluster



- Ex. In the figure, o is outlier since its closest large cluster is C1, but the similarity between o and C1 is small. For any point in C3, its closest large cluster is C2 but its similarity from C2 is low, plus |C3| = 3 is small

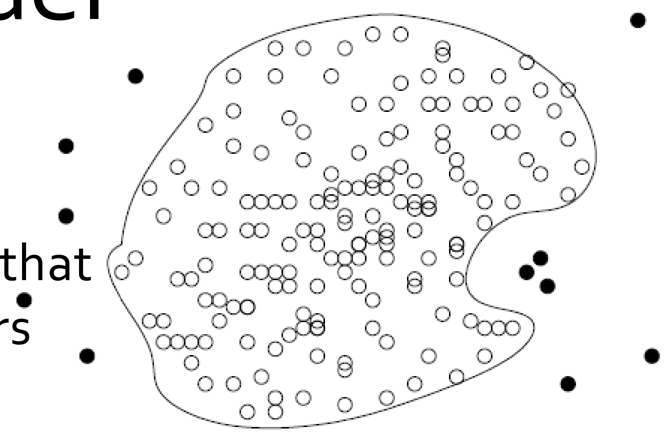
Clustering-Based Method: Strength and Weakness

- Strength
  - Detect outliers without requiring any labeled data
  - Work for many types of data
  - Clusters can be regarded as summaries of the data
  - Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)
- Weakness
  - Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection
  - High computational cost: Need to first find clusters
  - A method to reduce the cost: Fixed-width clustering
    - A point is assigned to a cluster if the center of the cluster is within a pre-defined distance threshold from the point
    - If a point cannot be assigned to any existing cluster, a new cluster is created and the distance threshold may be learned from the training data under certain conditions

Outlier Analysis

- Basic Concepts
- Outlier Detection Methods
- Statistical Approaches
- Clustering-Based Approaches
- **Classification-Based Approaches**

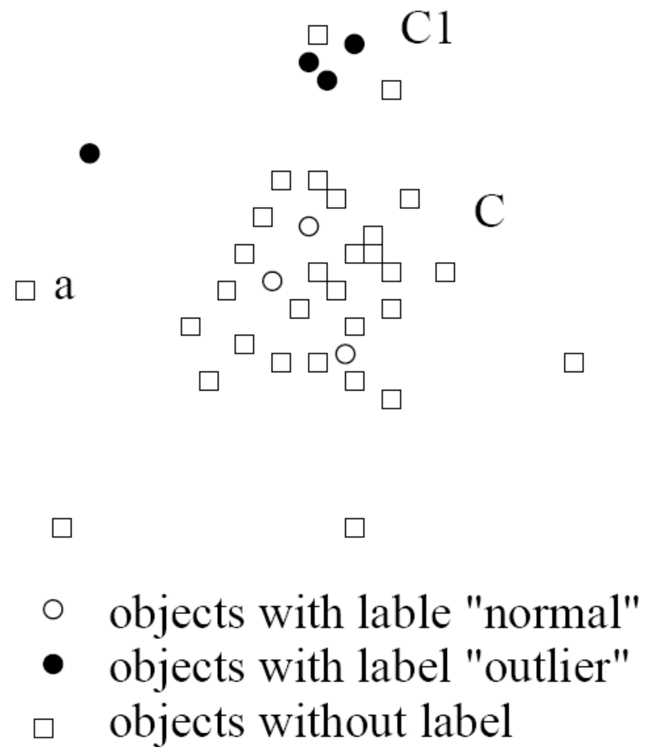
Classification-Based Method I: One-Class Model



- Idea: Train a classification model that can distinguish "normal" data from outliers
- A brute-force approach: Consider a training set that contains samples labeled as "normal" and others labeled as "outlier"
  - But, the training set is typically heavily biased: # of "normal" samples likely far exceeds # of outlier samples
  - Cannot detect unseen anomaly
- One-class model: A classifier is built to describe only the normal class.
  - Learn the decision boundary of the normal class using classification methods such as SVM
  - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
  - Adv: can detect new outliers that may not appear close to any outlier objects in the training set
  - Extension: Normal objects may belong to multiple classes

Classification-Based Method II: Semi-Supervised Learning

- Semi-supervised learning: Combining classification-based and clustering-based methods
- Method
  - Using a clustering-based approach, find a large cluster, C, and a small cluster, C_1
  - Since some objects in C carry the label "normal", treat all objects in C as normal
  - Use the one-class model of this cluster to identify normal objects in outlier detection
  - Since some objects in cluster C_1 carry the label "outlier", declare all objects in C_1 as outliers
  - Any object that does not fall into the model for C (such as *a*) is considered an outlier as well



Classification-Based Method: Strength and Weakness

- Strength: Outlier detection is fast
- Bottleneck: Quality heavily depends on the availability and quality of the training set, but often difficult to obtain representative and high-quality training data

Summary

- Basic Concepts
- Outlier Detection Methods
- Statistical Approaches
- Clustering-Based Approaches
- Classification-Based Approaches

References (1)

- B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. Biometrika, 1979.
- Malik Agyemang, Ken Barker, and Rada Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. Intell. Data Anal., 2006.
- Deepak Agarwal. Detecting anomalies in cross-classied streams: a bayesian approach. Knowl. Inf. Syst., 2006.
- C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. SIGMOD'01.
- M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Optics-of: Identifying local outliers. PKDD '99
- M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. SIGMOD'00.
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Comput. Surv., 2009.
- D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. Computational Intelligence, 2002.
- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Proc. 2002 Int. Conf. of Data Mining for Security Applications, 2002.
- E. Eskin. Anomaly detection over noisy data using learned probability distributions. ICML'00.
- T. Fawcett and F. Provost. Adaptive fraud detection. Data Mining and Knowledge Discovery, 1997.
- R. Fujimaki, T. Yairi, and K. Machida. An approach to spacecraft anomaly detection problem using kernel feature space. KDD '05
- F. E. Grubbs. Procedures for detecting outlying observations in samples. Technometrics, 1969.

References (2)

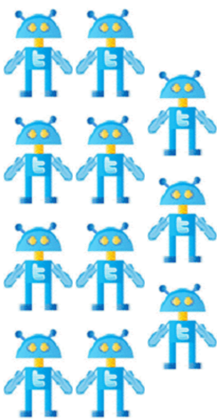
- V. Hodge and J. Austin. A survey of outlier detection methodologies. Artif. Intell. Rev., 2004.
- Douglas M Hawkins. Identification of Outliers. Chapman and Hall, 1980.
- P. S. Horn, L. Feng, Y. Li, and A. J. Pesce. Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation. Clin Chem, 2001.
- W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. PAKDD'06
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98
- M. Markou and S. Singh.. Novelty detection: a review| part 1: statistical approaches. Signal Process., 83(12), 2003.
- M. Markou and S. Singh. Novelty detection: a review| part 2: neural network based approaches. Signal Process., 83(12), 2003.
- S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. ICDE'03.
- A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Netw., 51(12):3448{3470, 2007.
- W. Stefansky. Rejecting outliers in factorial designs. Technometrics, 14(2):469{479, 1972.
- X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. IEEE Trans. on Knowl. and Data Eng., 19(5):631{645, 2007.
- Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. KDD '06:
- N. Ye and Q. Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. Quality and Reliability Engineering International, 2001.

Suspicious Behavior Detection

- Meng Jiang, Peng Cui, and Christos Faloutsos. "Suspicious behavior detection: current trends and future directions." **IEEE Intelligent Systems**, 2016. (Survey paper)

Catching Social Link Farming

Catching Zombie Followers



"Your best followers!"
"Delivery within 1-2 days"

Catching Zombie Followers

"Your best followers!"
"Delivery within 1-2 days"

engineers

product managers

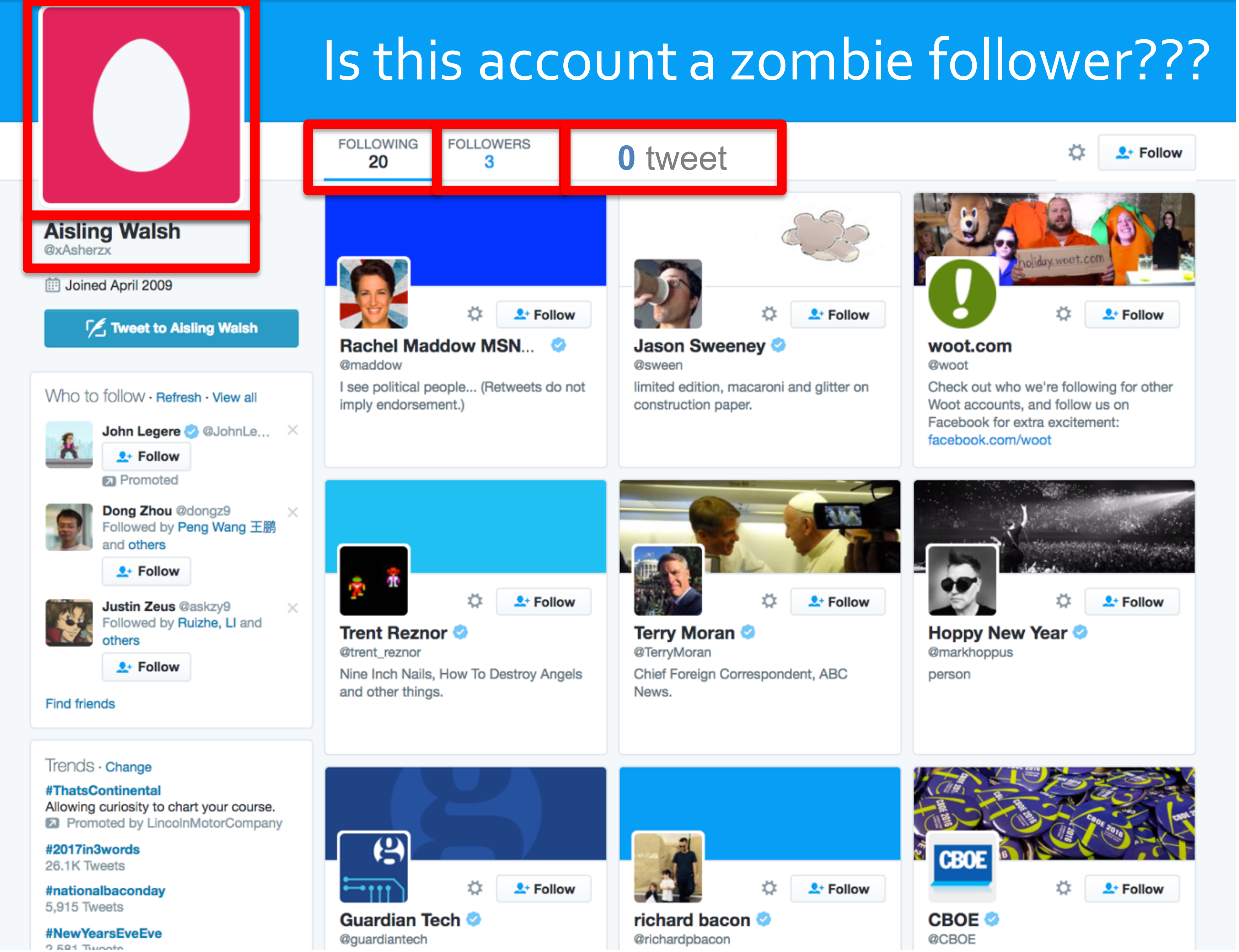
Knowledge from manual inspection: #followees, #followers, #tweets, #hashtags, #urls…

Learning models (classifiers)

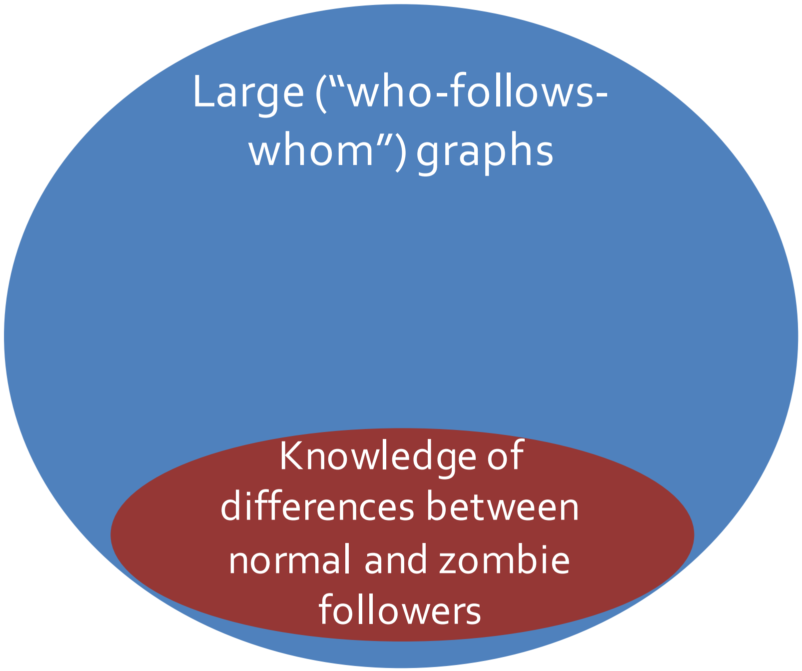
Poor accuracy
(**serious complaints** from users)

Fake account detection [Egele and Stringhini et al. NDSS'13; Yang and Wilson et al. TKDD'14; Viswanath and Bashir et al. USENIX Security Symposium'14]

Is this account a zombie follower???



Methodology



Methodology

Large ("who-follows-whom") graphs

Knowledge of differences between normal and zombie followers

Behavioral theory
(Think what they have to do, not what they can do!)

Consistently Connecting to Customers

Methodology

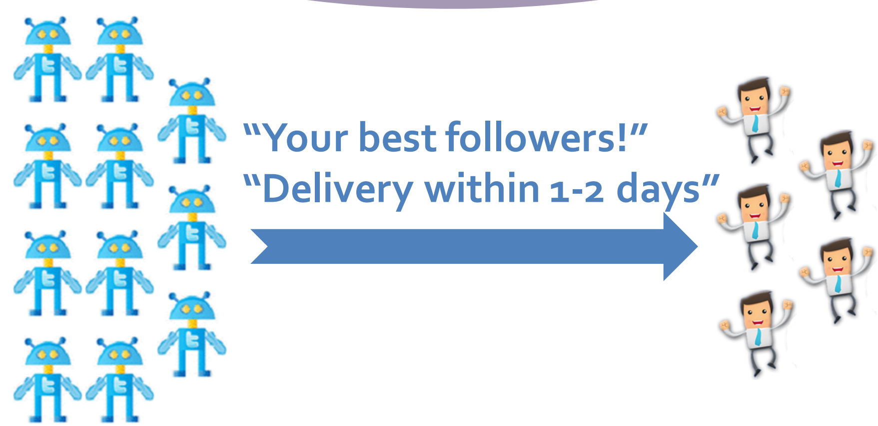
Large ("who-follows-whom") graphs

Knowledge of differences between normal and zombie followers

Behavioral theory (Think what they have to do, not what they can do!)

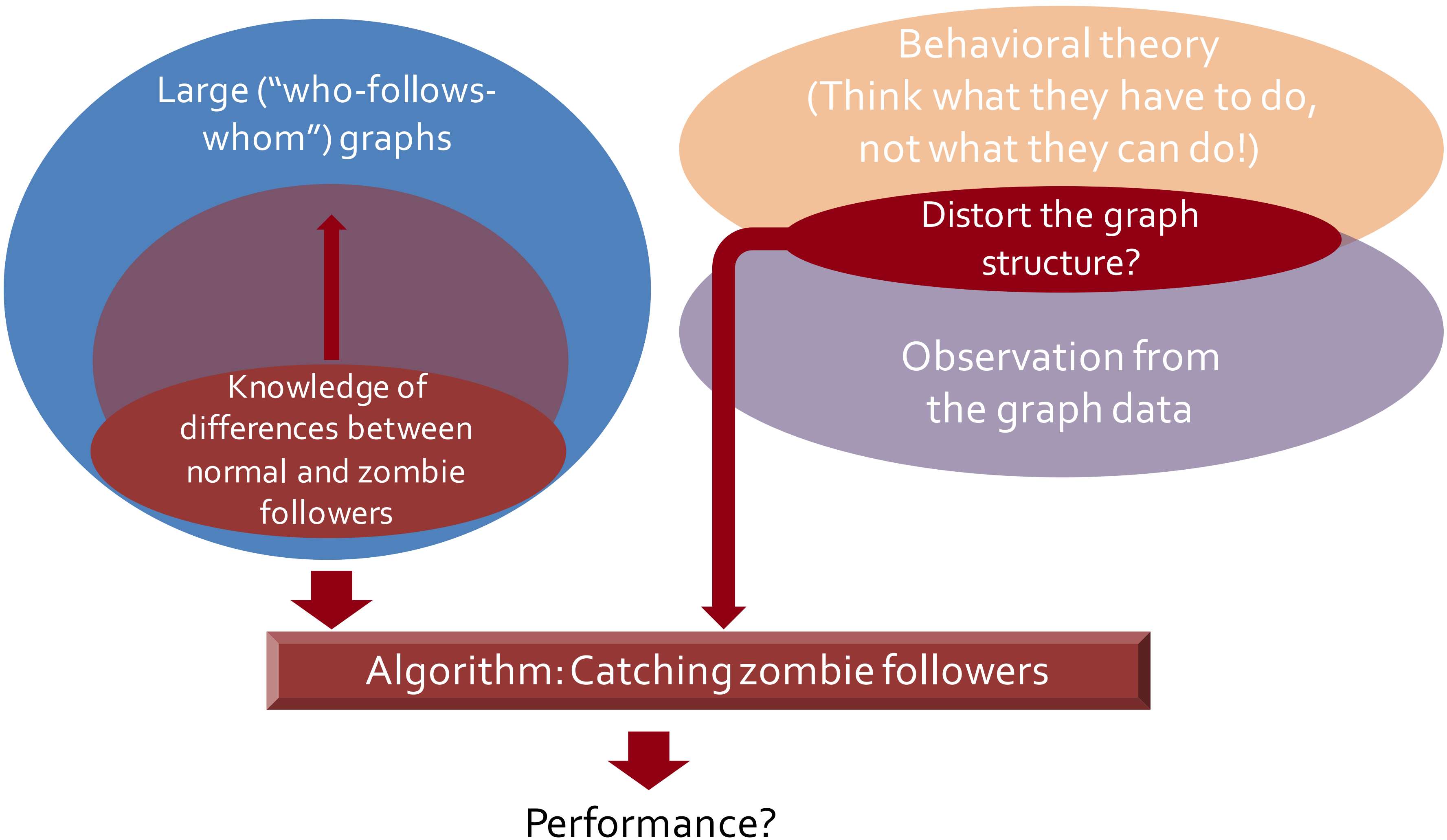
Distort the graph structure?

Observation from the graph data



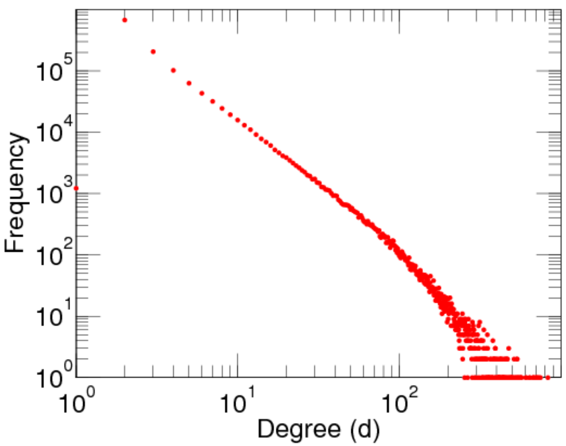
Consistently Connecting to Customers

Methodology

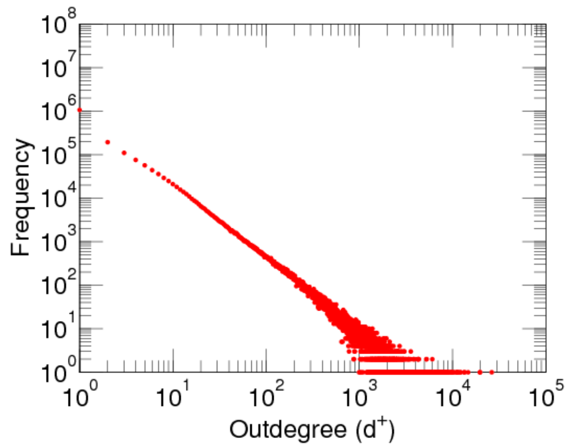


Out-Degree Distributions: Power Law Expected

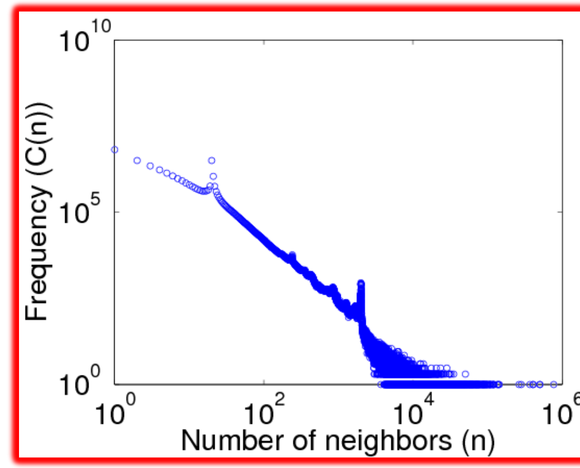
DBLP
Author-publication



Flickr
User-user



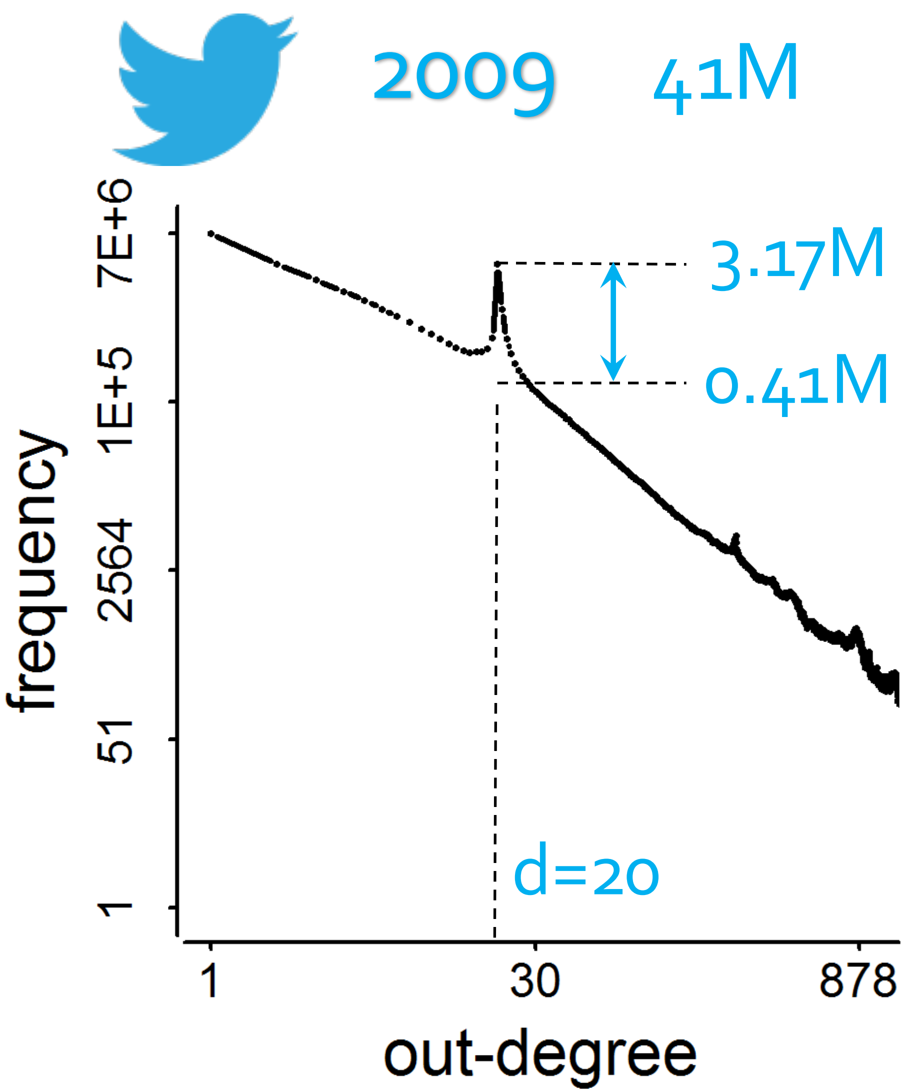
Twitter
Who-follows-whom

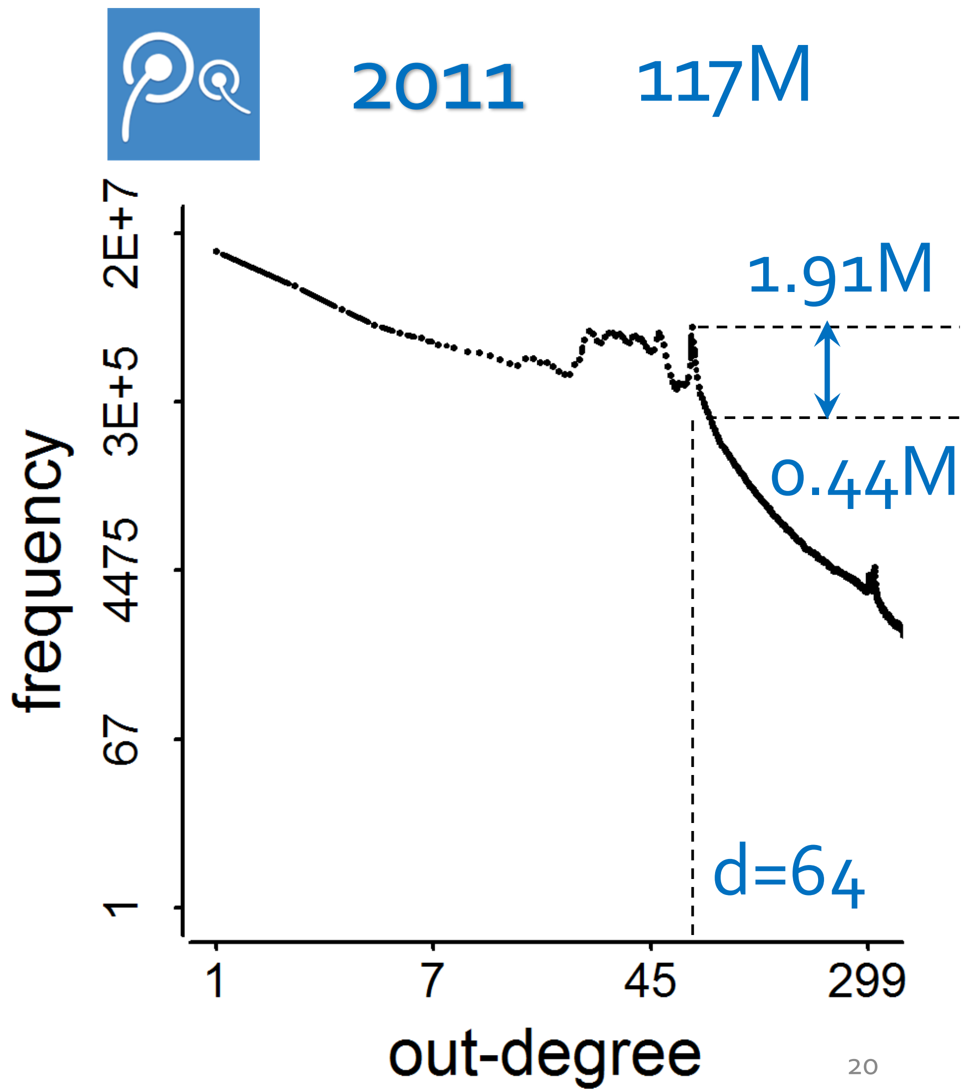


[konect.uni-koblenz.de/networks/]

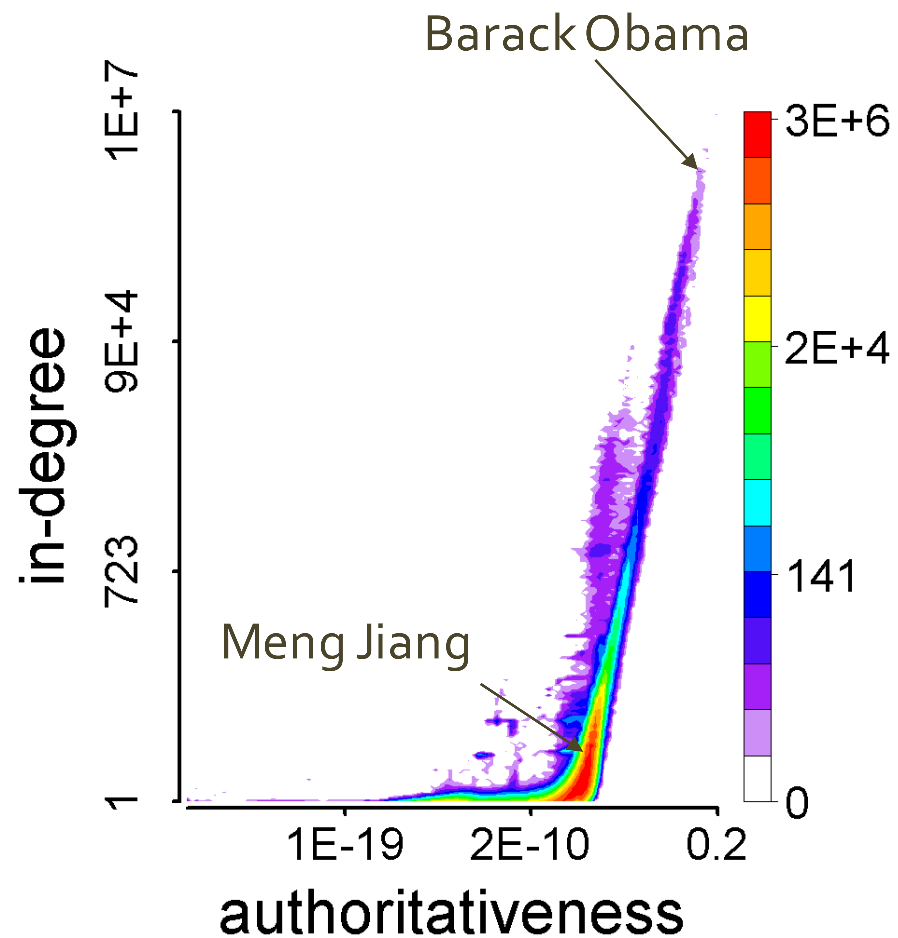
Power-law distributions in networks [Faloutsos et al. SIGCOMM'99; Chung et al. PNAS'02]

Spikes!


2009
41M
7E+6
1E+5
2564
51
1
frequency
3.17M
0.41M
d=20
1
30
878
out-degree

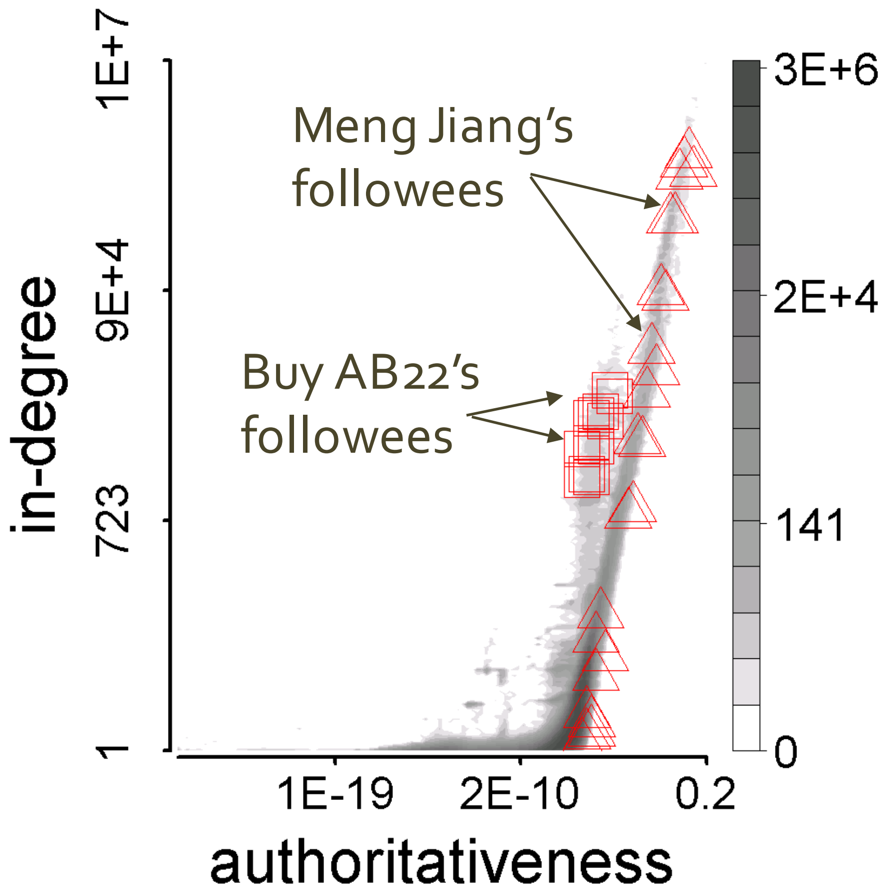

2011
117M
2E+7
3E+5
4475
67
1
frequency
1.91M
0.44M
d=64
1
7
45
299
out-degree

How We/They Connect to Our/Their Followees

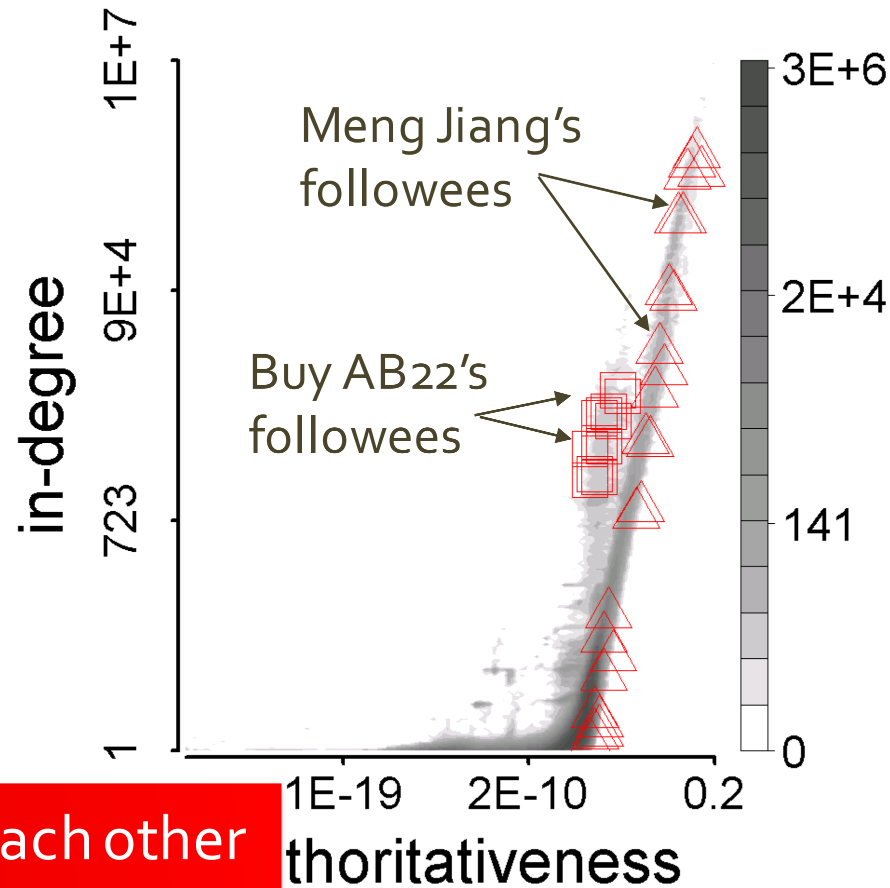


The HITS algorithm. Kleinberg. "Authoritative sources in a hyperlinked environment." JACM'99.

How We/They Connect to Our/Their Followees



How We/They Connect to Our/Their Followees

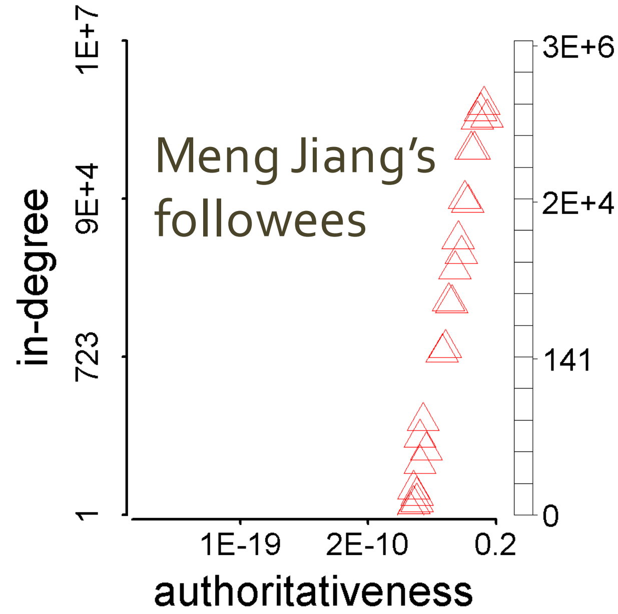


Synchronized: too similar with each other
Abnormal: too different from the majority

Definition: Synchronicity

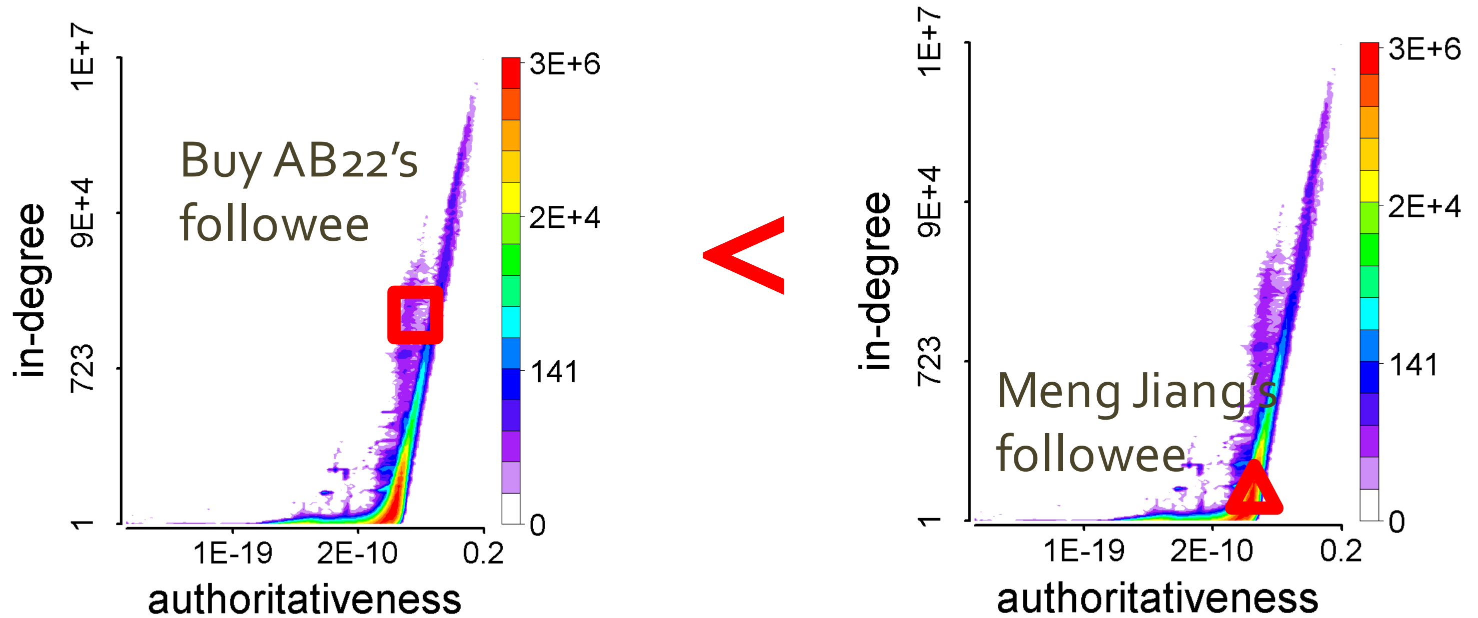
$$sync(u) = \frac{\sum_{(v,v')\in\mathcal{F}(u)\times\mathcal{F}(u)} \mathbf{p}(v)\cdot\mathbf{p}(v')}{d(u)\times d(u)}$$

>



Definition: Normality

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



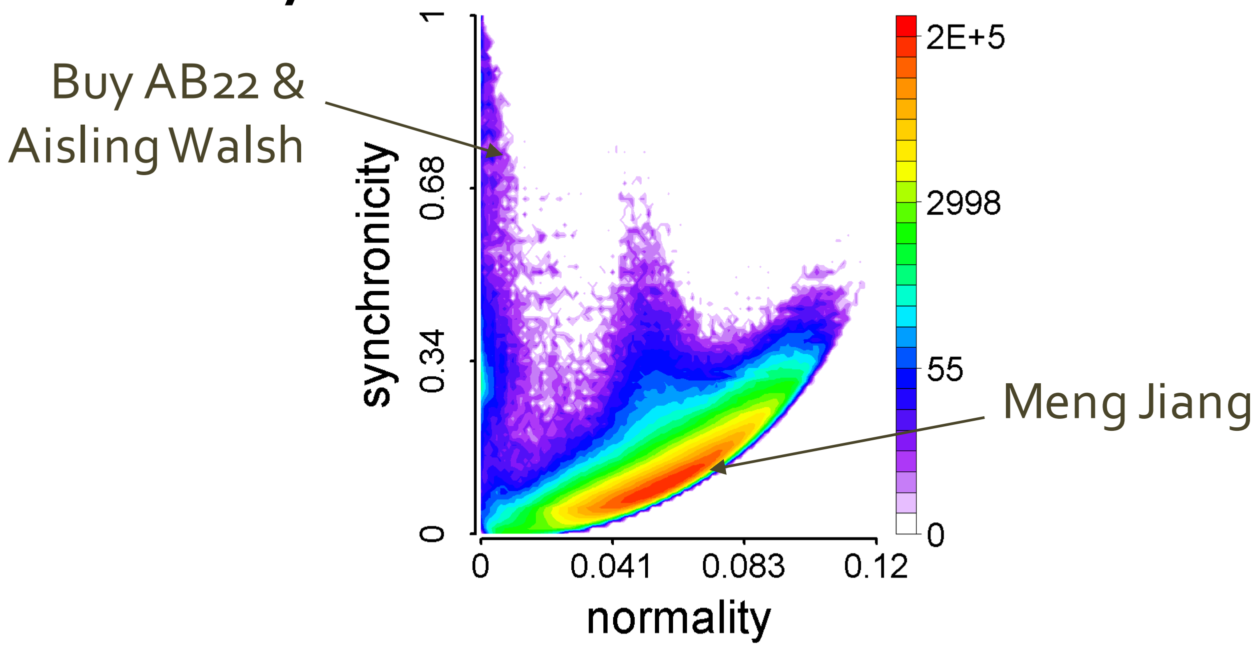
When is the Synchronicity Too High?

Problem: Given a normality value (n) of a follower, find the minimal synchronicity value (s_{min}).

Theorem:

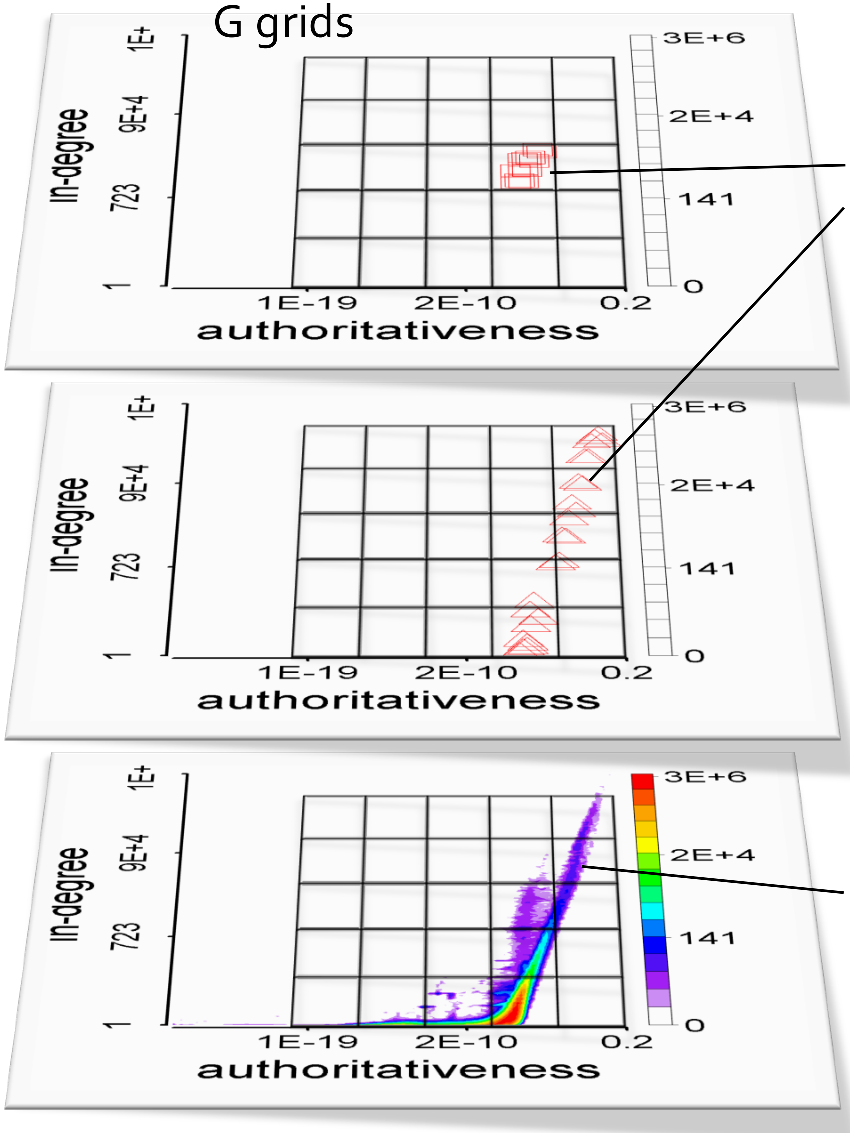
$$s_{min} = \frac{-G\,\mathbf{n^2} + 2\,n - s_b}{1 - G\,s_b} \qquad \textit{(parabolic lower limit)}$$

Our CatchSync:



Proof

G grids



fp_g: #foreground points in grid g

$\sum fp_g = F = d(u)$ (#followees of u)

Given normality

$$n = \sum (fp_g/F)\,(bp_g/B) = \sum f_g b_g,$$

find minimal synchronicity

$$s_{min} = \sum (fp_g/F)\,(fp_g/F) = \sum f_g^2$$

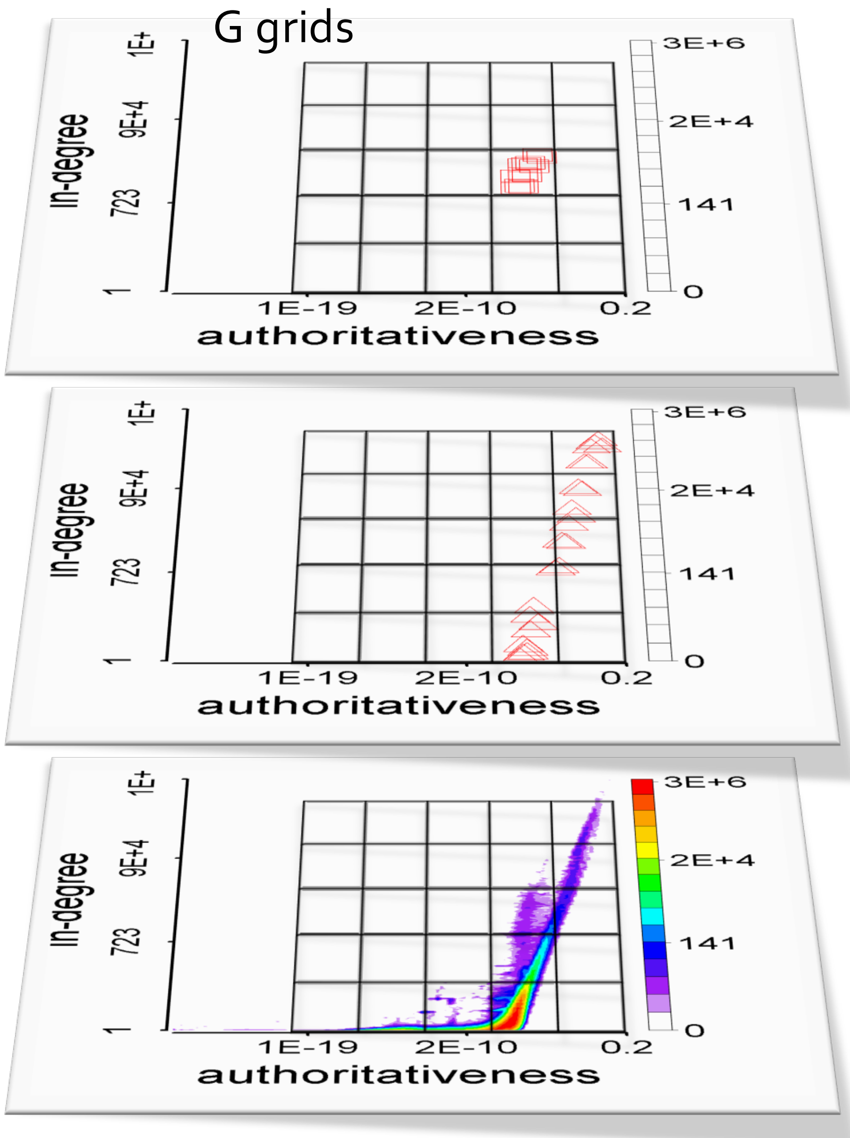
where

$$\sum f_g = 1,\ \sum b_g = 1$$

bp_g: #background points in grid g

$\sum bp_g = B = N$ (#all users)

Proof



Lagrange multiplier:

minimize $s(f_g) = \sum f_g^2$

subject to $\sum f_g = 1$, $\sum f_g b_g = n$

Lagrange function:

$F(f_g, \lambda, \mu) = (\sum f_g^2) + \lambda (\sum f_g - 1) + \mu (\sum f_g b_g - n)$

Gradients:

$$\begin{cases} \nabla_{f_g} F = 2 f_g + \lambda + \mu b_g = 0 \\ \nabla_{\lambda} F = \sum f_g - 1 = 0 \\ \nabla_{\mu} F = \sum f_g b_g - n = 0 \end{cases}$$

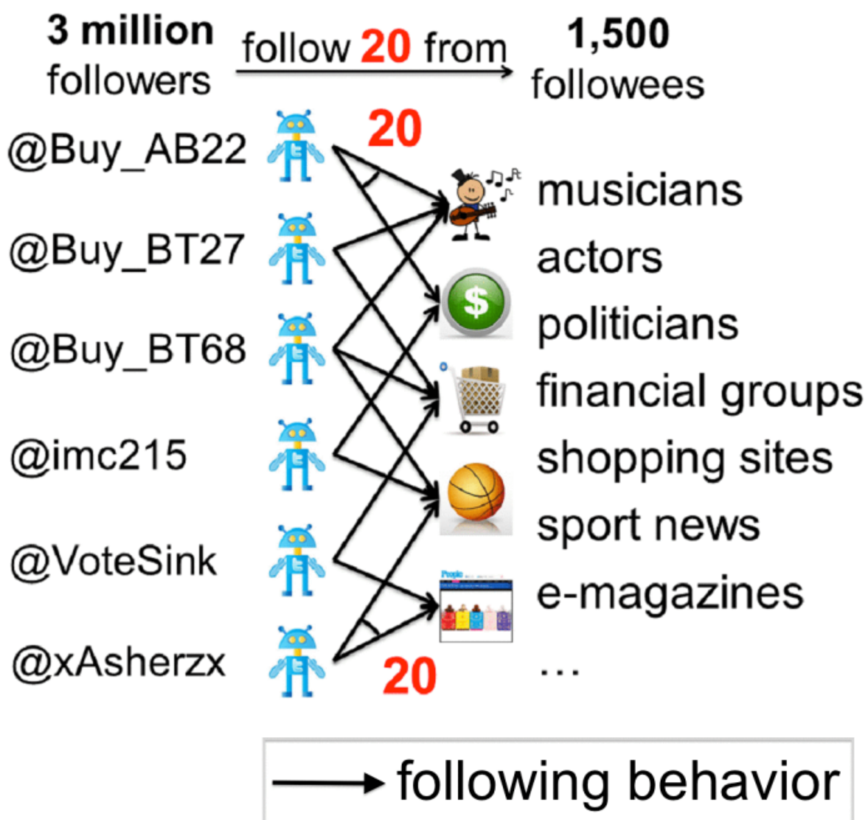
$$\begin{cases} 2 + \lambda G + \mu = 0 & \sum \\ 2n + \lambda + \mu s_b = 0 & \times b_g \sum \\ 2 s_{min} + \lambda + \mu n = 0 & \times f_g \sum \end{cases}$$

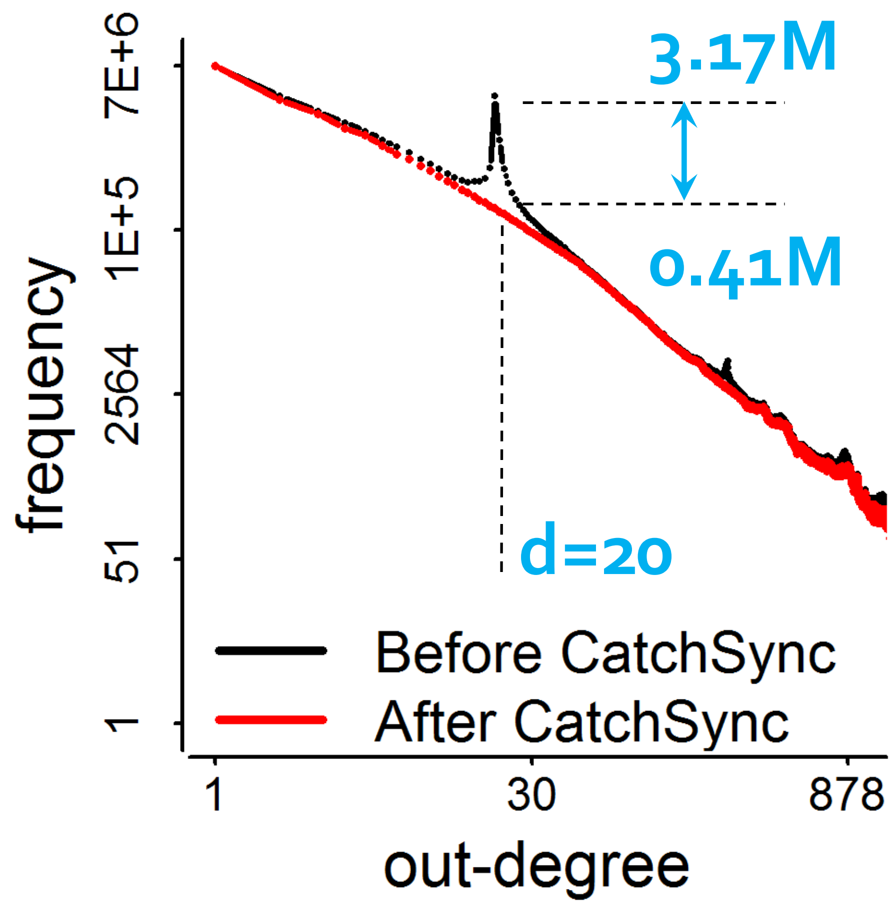
where $s_b = \sum b_g^2$.

Therefore,

$$s_{min} = \frac{-G n^2 + 2n - s_b}{1 - G s_b}$$

The Distribution was Recovered!





Discussion

- What kind of outliers?
  - Unsupervised learning for collective outliers
- Camouflage?