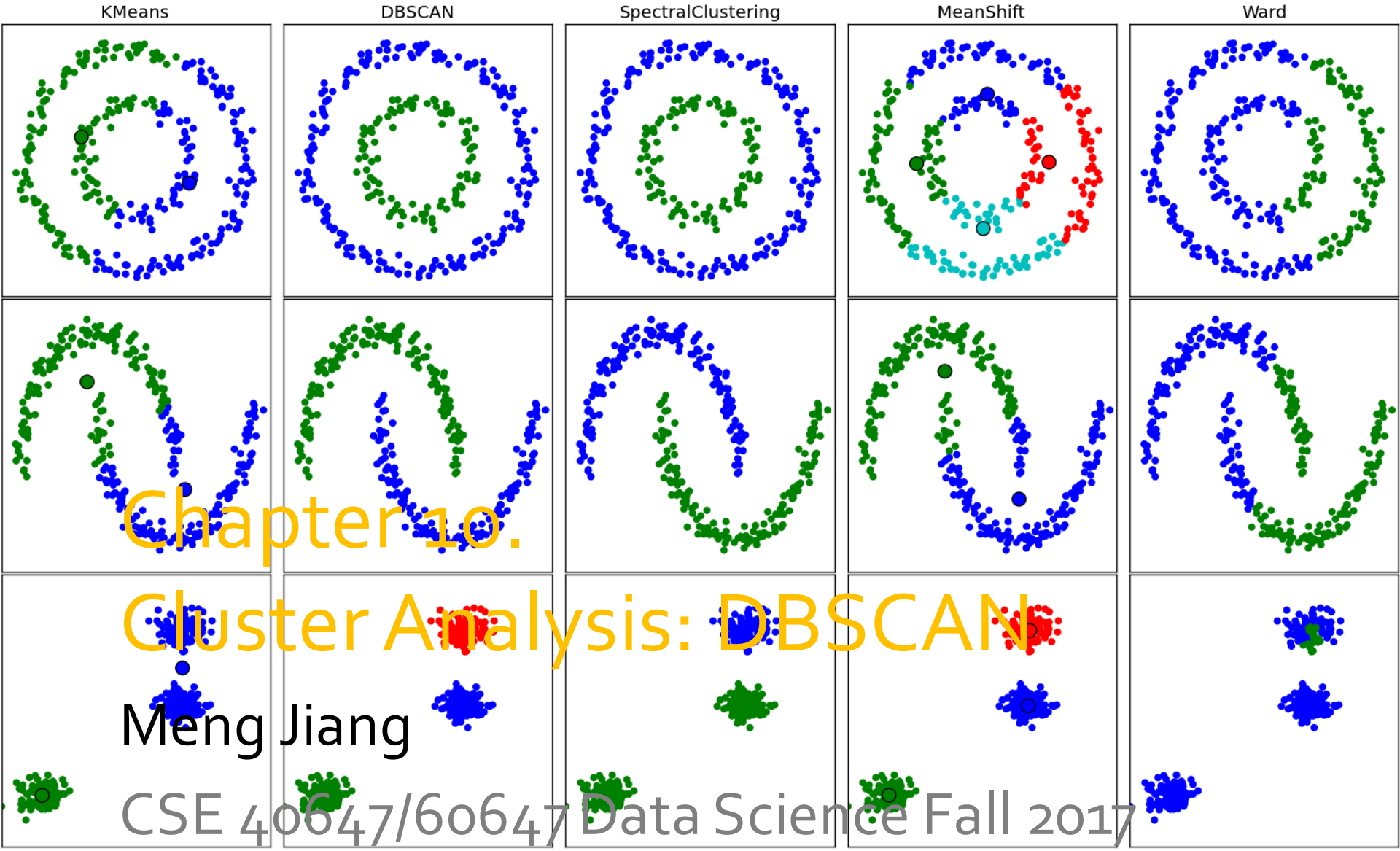


Chapter 10. Cluster Analysis: DBSCAN

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Cluster Analysis

- Cluster Analysis: An Introduction
- Partitioning Methods
- **Density-based Methods**
- Evaluation of Clustering

Density-Based and Grid-Based Clustering Methods

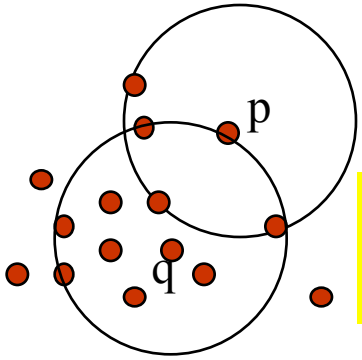
- Density-Based Clustering
  - Basic Concepts
  - **DBSCAN: A Density-Based Clustering Algorithm**
  - OPTICS: Ordering Points To Identify Clustering Structure
- Grid-Based Clustering Methods
  - Basic Concepts
  - STING: A Statistical Information Grid Approach
  - CLIQUE: Grid-Based Subspace Clustering

Density-Based Clustering Methods

- Clustering based on density (a local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan (only examine the local region to justify density)
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99)
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (also, grid-based)

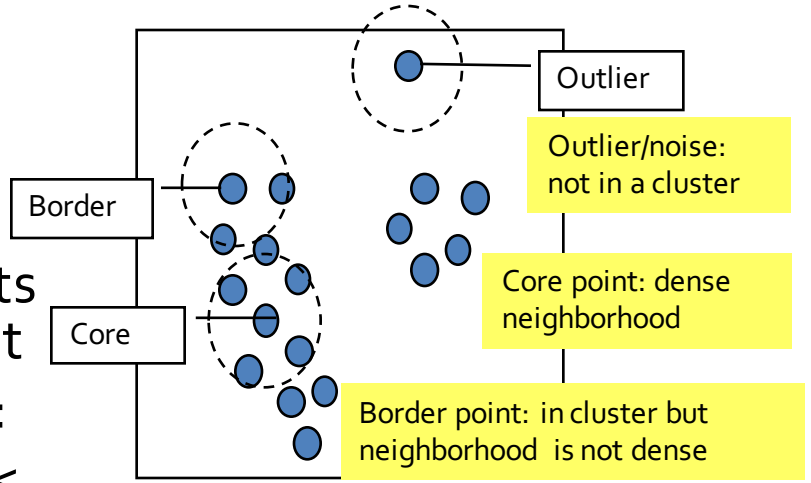
DBSCAN: A Density-Based Spatial Clustering Algorithm

- DBSCAN (M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, KDD'96)
  - Discovers clusters of arbitrary shape: Density-Based Spatial Clustering of Applications with Noise
- A *density-based* notion of cluster
  - A ***cluster*** is defined as a **maximal** set of **density-connected** points
- Two parameters*:*
  - *Eps* (ε): Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in the Eps-neighborhood of a point
- The Eps(ε)-neighborhood of a point *q*:
  - $N_{Eps}(q)$: {p belongs to D | dist(p, q) ≤ Eps}



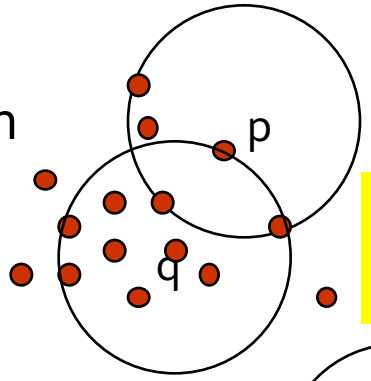
MinPts = 5

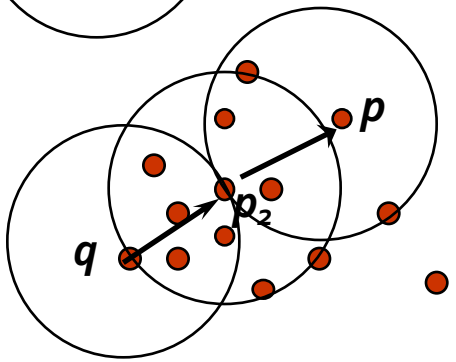
Eps = 1 cm

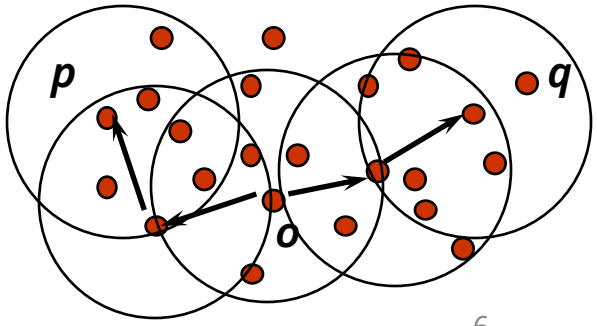


DBSCAN: Density-Reachable and Density-Connected

- **Directly density-reachable**:
  - A point *p* is directly density-reachable from a point *q* w.r.t. *Eps* (ε), *MinPts* if
    - *p* belongs to $N_{Eps}(q)$
    - **core point** condition: $|N_{Eps}(q)| \geq MinPts$
- **Density-reachable**:
  - A point *p* is density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i
- **Density-connected:**
  - A point *p* is density-connected to a point *q* w.r.t. *Eps*, *MinPts* if there is a point *o* such that both *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*



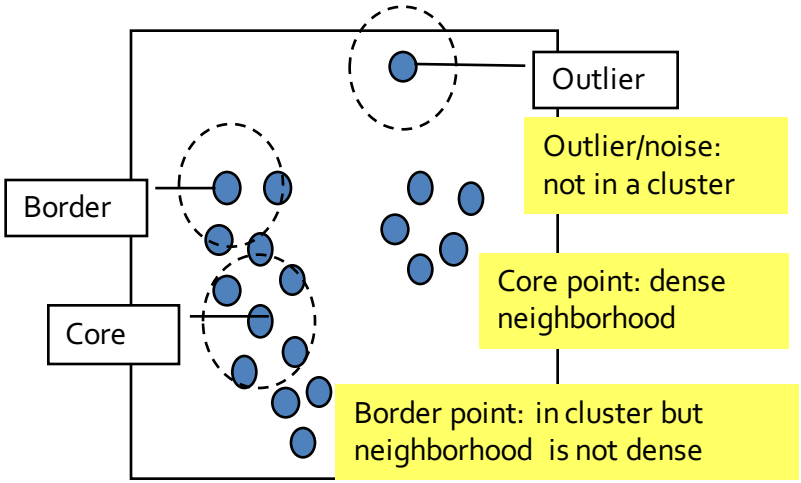




DBSCAN: The Algorithm

- **Algorithm**
  - Arbitrarily select a point *p*
  - Retrieve all points density-reachable

    from *p* w.r.t. *Eps* and *MinPts*
    - If *p* is a core point, a cluster is formed
    - If *p* is a border point, no points are density-reachable from *p*, and DBSCAN visits the next point of the database
  - Continue the process until all of the points have been processed



DBSCAN: The Algorithm

- **Computational complexity**
  - If a spatial index is used, the computational complexity of DBSCAN is O(nlogn), where n is the number of database objects
  - Otherwise, the complexity is $O(n^2)$

https://en.wikipedia.org/wiki/DBSCAN

DBSCAN visits each point of the database, possibly multiple times (e.g., as candidates to different clusters). For practical considerations, however, the time complexity is mostly governed by the number of regionQuery invocations. DBSCAN executes exactly one such query for each point, and if an indexing structure is used that executes a neighborhood query in $O(\log n)$, an overall average runtime complexity of $O(n \log n)$ is obtained (if parameter ε is chosen in a meaningful way, i.e. such that on average only $O(\log n)$ points are returned). Without the use of an accelerating index structure, or on degenerated data (e.g. all points within a distance less than ε), the worst case run time complexity remains $O(n^2)$. The distance matrix of size $(n^2\text{-}n)/2$ can be materialized to avoid distance recomputations, but this needs $O(n^2)$ memory, whereas a non-matrix based implementation of DBSCAN only needs $O(n)$ memory.

DBSCAN Is Sensitive to the Setting of Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

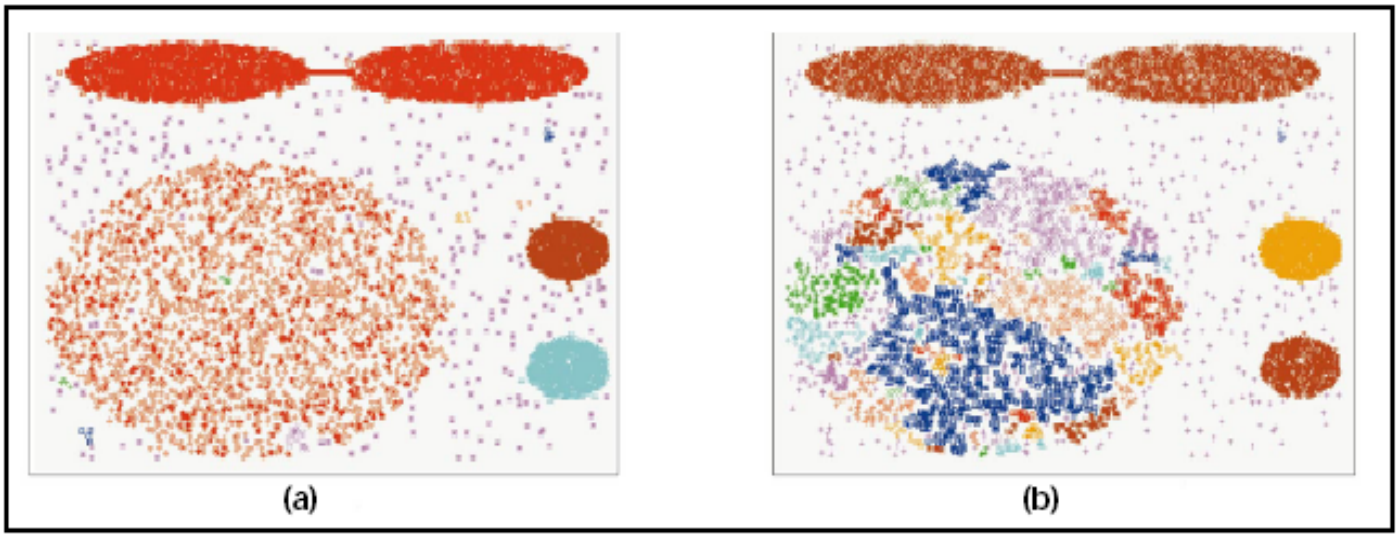
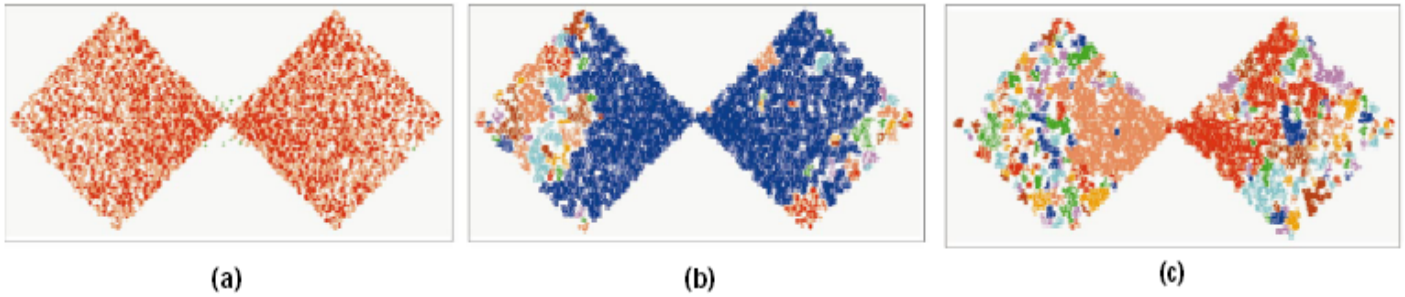


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



Ack. Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999

References: (III) Density-based Methods

- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases. KDD'96
- W. Wang, J. Yang, R. Muntz, STING: A Statistical Information Grid Approach to Spatial Data Mining, VLDB'97
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD'98
- A. Hinneburg and D. A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD'98
- M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering Points to Identify the Clustering Structure. SIGMOD'99
- M. Ester. Density-Based Clustering. In (Chapter 5) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications . CRC Press. 2014
- W. Cheng, W. Wang, and S. Batista. Grid-based Clustering. In (Chapter 6) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press. 2014

ICDM 2015

IEEE International Conference on Data Mining

A General **Suspiciousness** Metric for **Dense Blocks** in **Multimodal** Data

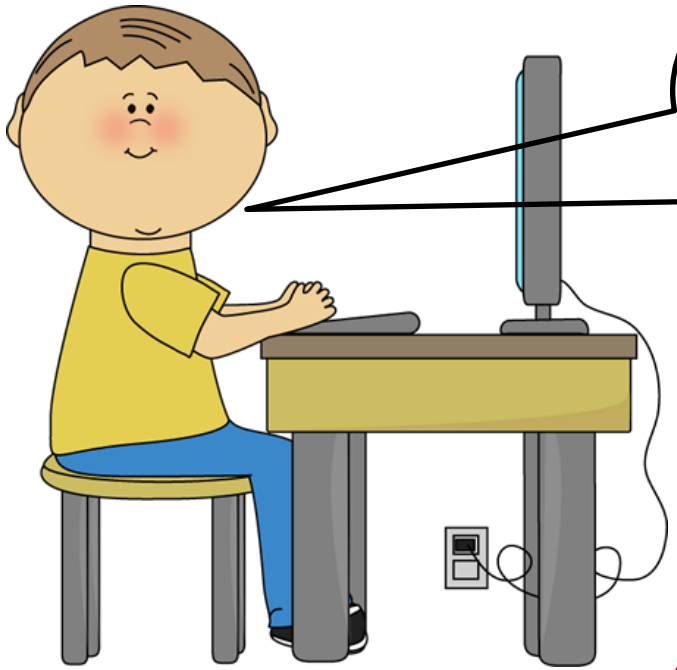
Meng Jiang

Joint work with

Alex Beutel, Peng Cui, Bryan Hooi,

Shiqiang Yang, Christos Faloutsos

Suppose You Work in Twitter



My boss wants me to **catch fraud** in such a big table – **billions of records, tens of columns**!!! How?!

	ID	USER_NAME	CREATED_AT	TEXT	HASH_TAGS
1	251	SpiritSofts	Dec 14, 2013	SAP HANA ONLINE TRAINING COURSE CONTENT http://t.co/2DefOMC0Vi	
2	252	Blue net studiO	Dec 14, 2013	sap hana online training and placenet 2 http://t.co/S1wGh8n5Kk	
3	253	Hana Kingham	Dec 14, 2013	Right film fest today: love actually, elf, gravity, training day. #dayyym	dayyym,
4	254	Nora Apnila J...	Dec 14, 2013	Alhamdulilaaahhhh...selesai ikutin kelanjutan training dadakan mb Hana ...	
5	255	ZaranTech	Dec 14, 2013	I added a video to a @YouTube playlist http://t.co/O3qD9wfi8K SAP BUSI...	
6	256	ZaranTech	Dec 14, 2013	I added a video to a @YouTube playlist http://t.co/XxrfuCUqAS SAP BUSI...	
7	257	Helmich op t...	Dec 14, 2013	Reserveer alvast 15 januari 2014 training HANA Essentials #SAP #HANA	SAP,HANA,
8	258	Social News	Dec 13, 2013	sap hana online training and placenet 2 http://t.co/JlaA41IdnV	
9	259	Nurjanah	Dec 13, 2013	Baca notif fb .. ada training dadakaann dari eyang kita.... ayo wara wiri ca...	
10	260	Nora Apnila J...	Dec 13, 2013	lanjutt di rumaah dulu ikutan trainingnyaaa..mau buru buru pulang see u...	
11	261	madhu	Dec 13, 2013	SAP HANA TRAINING	SAP HANA PLACEMENT	SAP HANA INSTITUTE I...	
12	262	Hana O'Neill	Dec 13, 2013	@sarahsilvanator no I have life guard training Saturday and my final test t...	
13	263	arjun	Dec 13, 2013	sap grc online training |sap hana sap security online training@YEKTEK - A...	

fraud

Massive **Multi-Modal** Data: Lines (Mass) & Columns (Mode)

Dataset	Mode				Mass
Retweeting	**User**	**Root ID**	**IP**	**Time (min)**	**#retweet**
	29.5M	19.8M	27.8M	56.9K	211.7M
Trending (Hashtag)	**User**	**Hashtag**	**IP**	**Time (min)**	**#tweet**
	81.2M	1.6M	47.7M	56.9K	276.9M
Network attacks (LBNL)	**Src-IP**	**Dest-IP**	**Port**	**Time (sec)**	**#packet**
	2,345	2,355	6,055	3,610	230,836

Suspicious Behaviors in **Multi-Modal** Data

Wholesalebargain2015 Retweeted

Real Time Deals @ebayrt · 2h

Seattle Mariners Mlb #Majestic Authentic Diamond Blue Stitched 2 Xl M... (Sanford) USD 25 ebayrt.co/sports-mem-car… #eBay #eBayUS via @wil30225

2

Wholesalebargain2015 Retweeted

Real Time Deals @ebayrt · 2h

Embroidered Navy Blue Aztec Mexican Top/ Long Sleeve Ethnic Mod... USD 35 ebayrt.co/clothing-shoes… #Handmade #eBay #eBayUS via @smilingbluedog

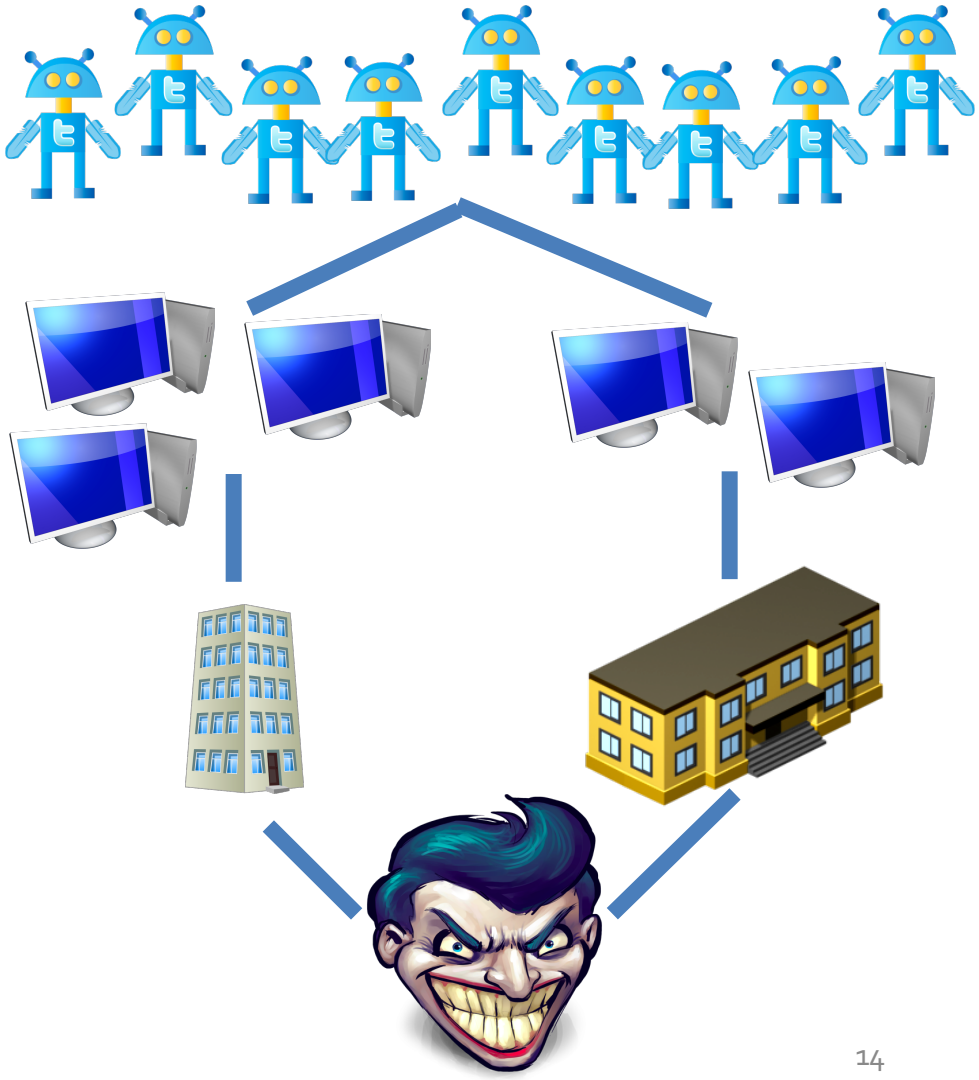
4

Wholesalebargain2015 Retweeted

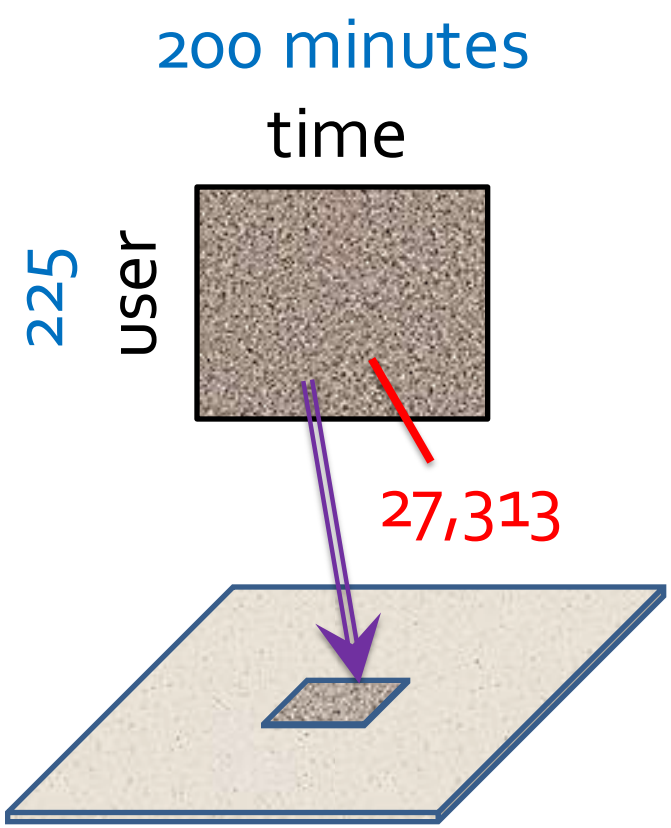
Real Time Deals @ebayrt · 1h

Contractubex Children Cartoon Boxing Gloves Red (Bloomington) USD 21.78 ebayrt.co/sporting-goods… #eBay #eBayUS via @GaroldFrenz

4



Dense Blocks Indicates **Suspiciousness**



Wholesalebargain2015 Retweeted

Real Time Deals @ebayrt · 2h

Seattle Mariners Mlb #Majestic Authentic Diamond Blue Stitched 2 Xl M... (Sanford) USD 25 ebayrt.co/sports-mem-car… #eBay #eBayUS via @wil30225

2

Wholesalebargain2015 Retweeted

Real Time Deals @ebayrt · 2h

Embroidered Navy Blue Aztec Mexican Top/ Long Sleeve Ethnic Mod... USD 35 ebayrt.co/clothing-shoes… #Handmade #eBay #eBayUS via @smilingbluedog

4

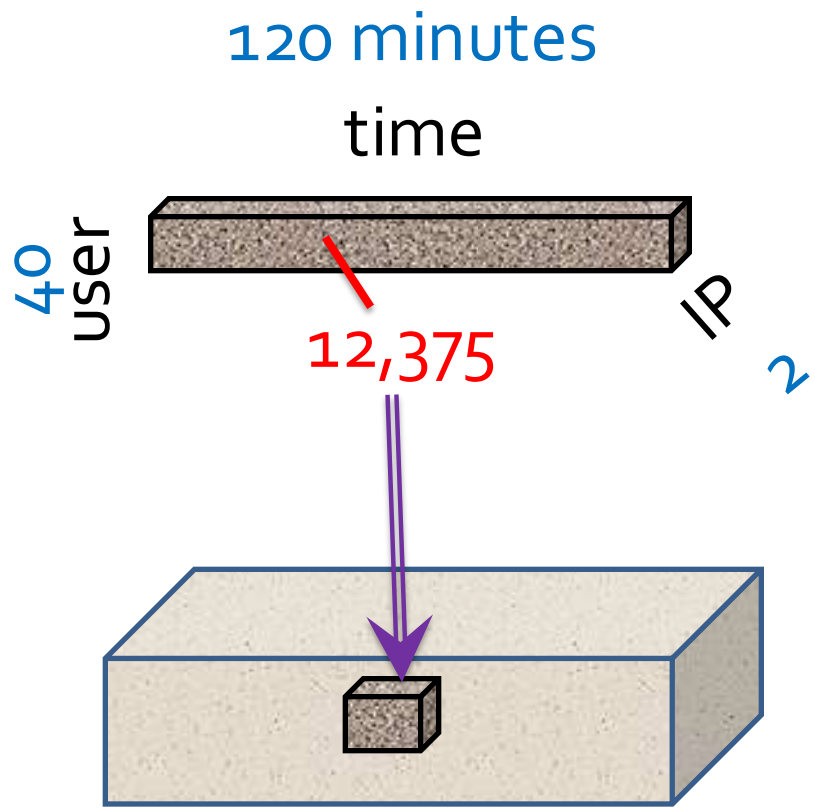
Wholesalebargain2015 Retweeted

Real Time Deals @ebayrt · 1h

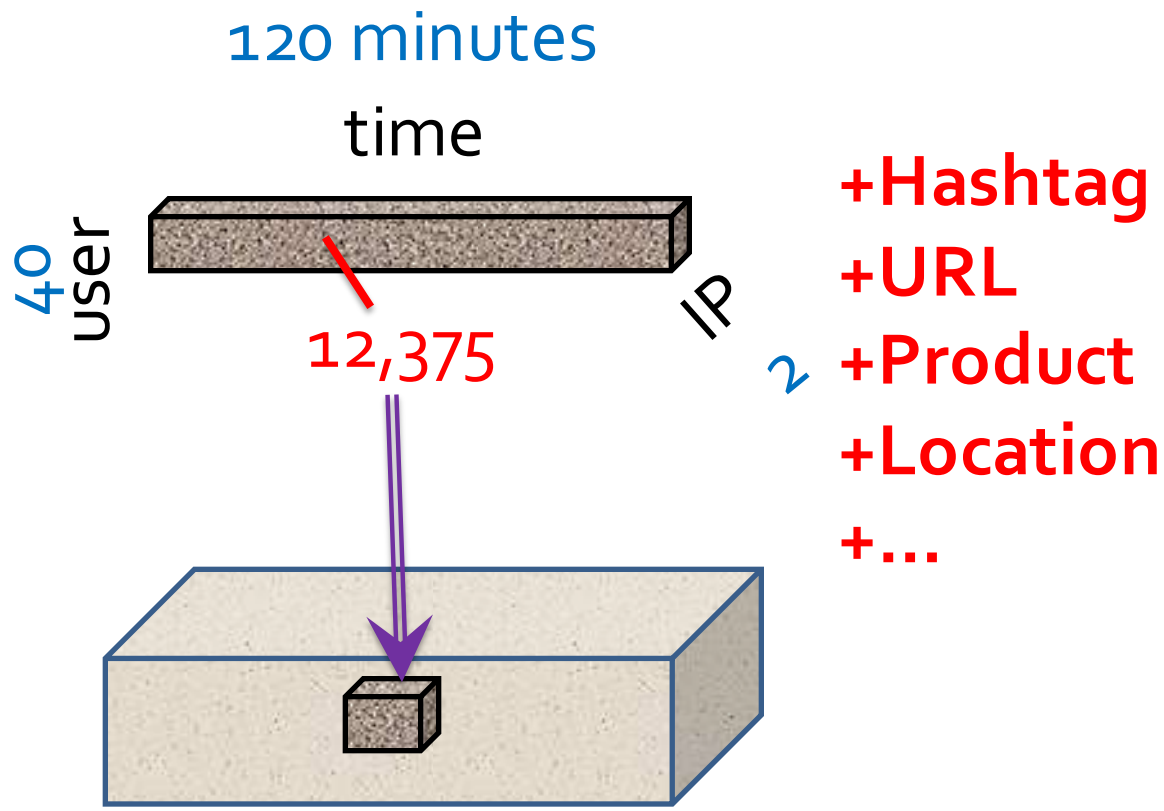
Contractubex Children Cartoon Boxing Gloves Red (Bloomington) USD 21.78 ebayrt.co/sporting-goods… #eBay #eBayUS via @GaroldFrenz

4

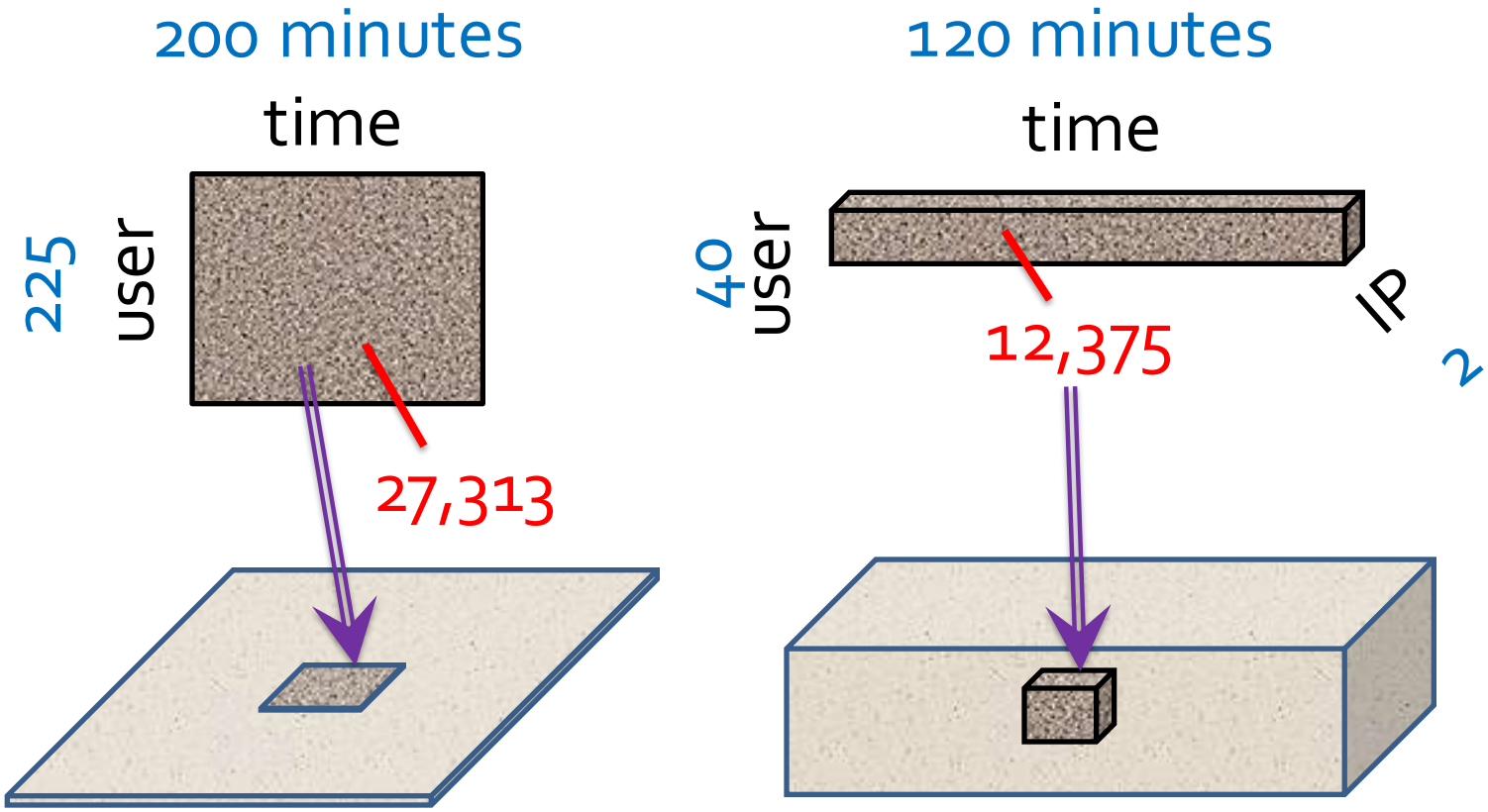
Dense Blocks Indicates **Suspiciousness**



Dense Blocks Indicates **Suspiciousness**



Dense Blocks Indicates **Suspiciousness**

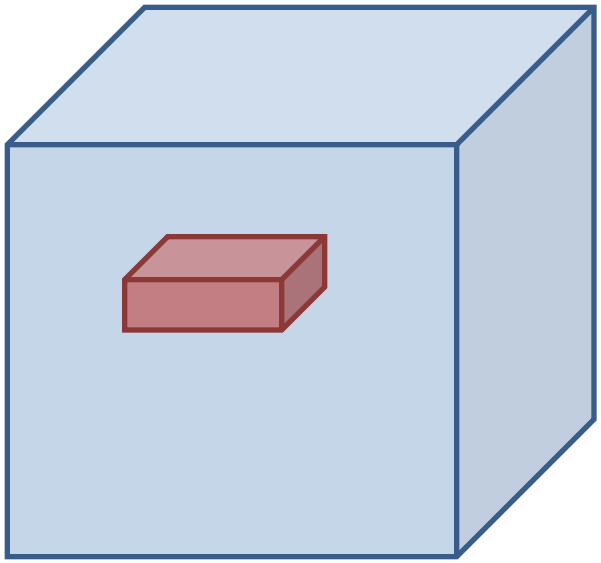


Question: Which is more suspicious?

We need a metric to evaluate the suspiciousness.

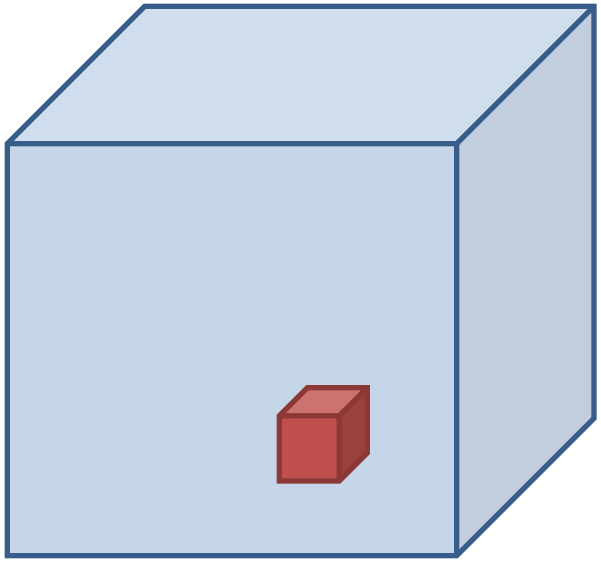
Metric Criteria

What properties are required of a good metric?



$N_1 \times N_2 \times N_3$

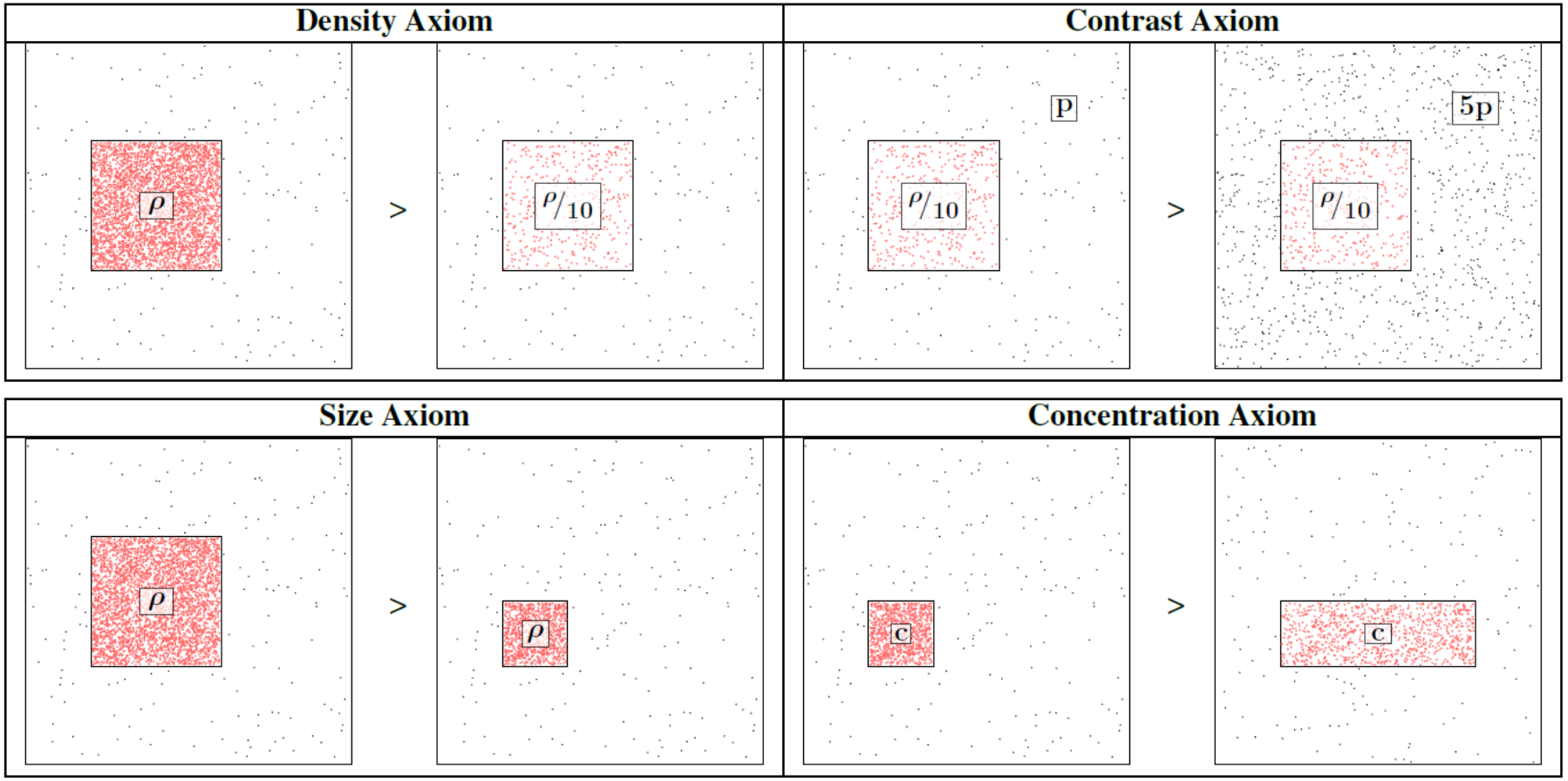
Count data with total "mass" *C*



$f(\ n_1 \times n_2 \times n_3$ mass c density ρ $)$ vs $f(\ n'_1 \times n'_2 \times n'_3$ mass c′ density ρ′ $)$

Axioms 1-4

$$c_1 > c_2 \iff f(\mathbf{n}, c_1, \mathbf{N}, C) > f(\mathbf{n}, c_2, \mathbf{N}, C)$$

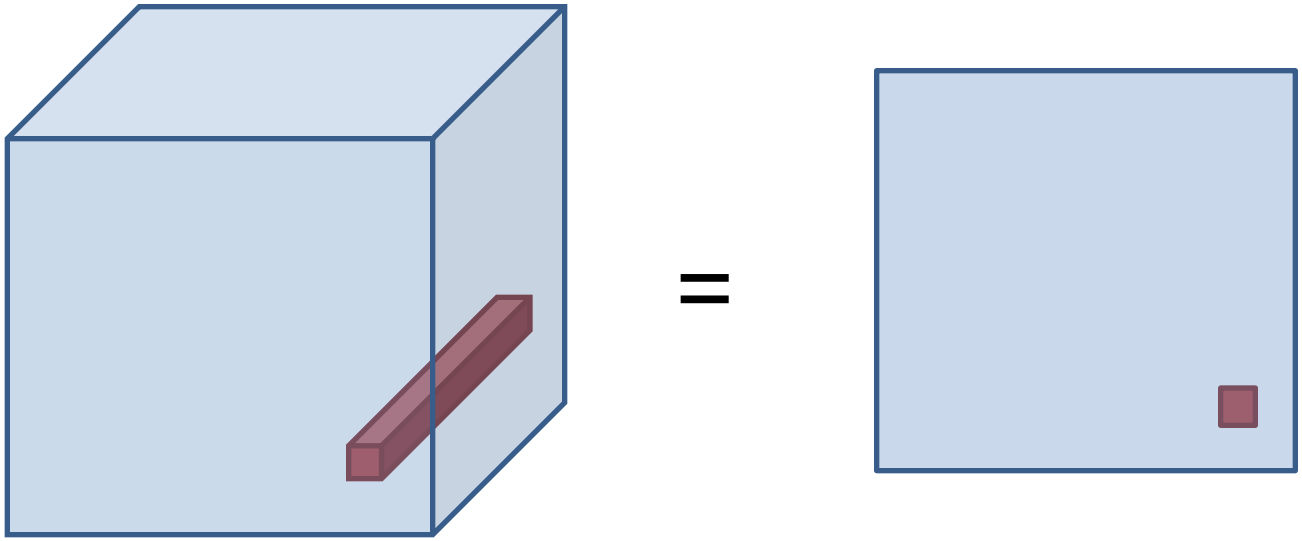


$$p_1 < p_2 \iff \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_1) > \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_2)$$

Axiom 5: Multimodal

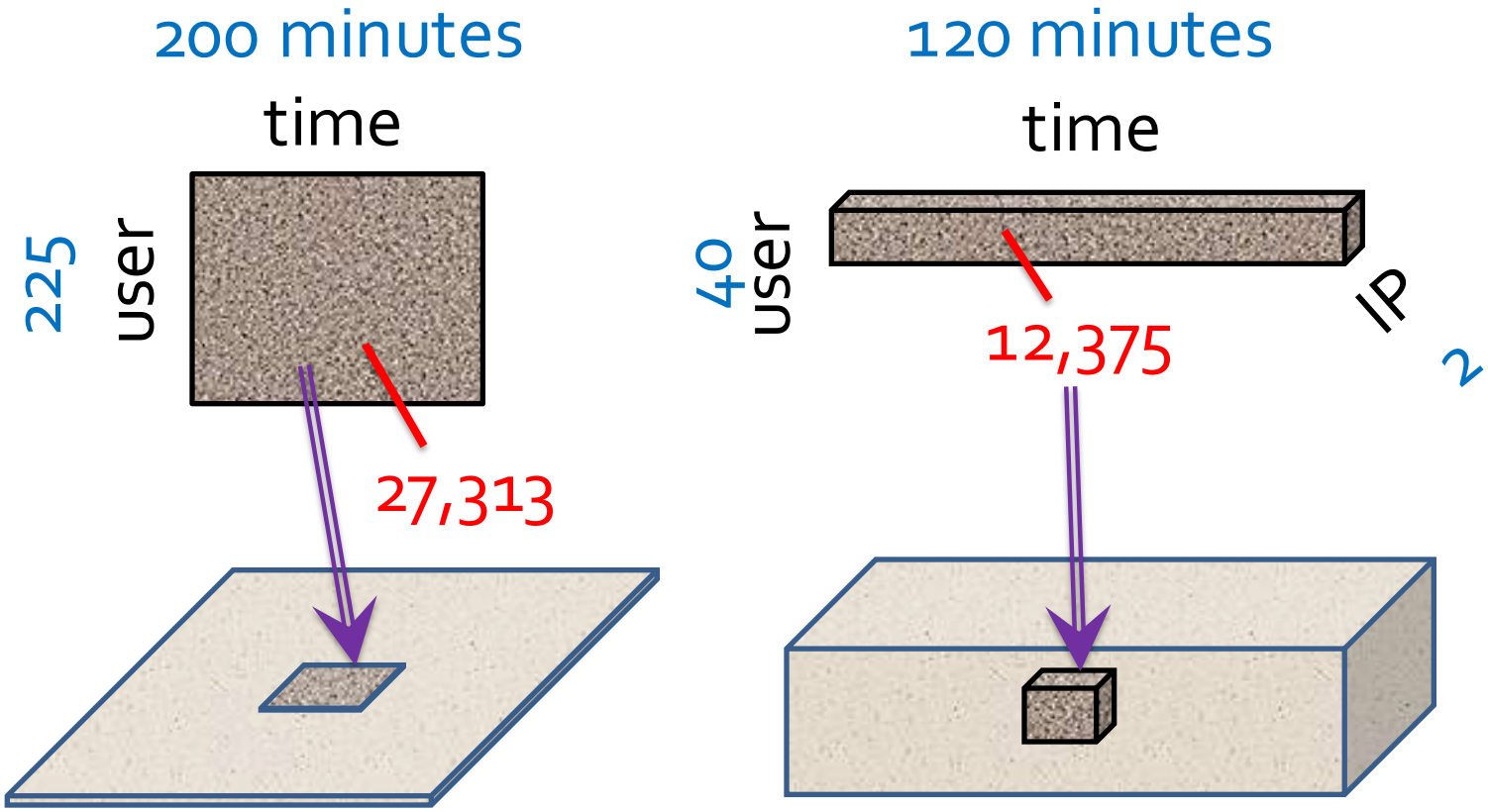
$$f_{K-1}\left([n_k]_{k=1}^{K-1}, c, [N_k]_{k=1}^{K-1}, C\right) = f_K\left(([n_k]_{k=1}^{K-1}, N_K), c, [N_k]_{k=1}^{K}, C\right)$$

Not including a mode is the same as including all values for that mode.

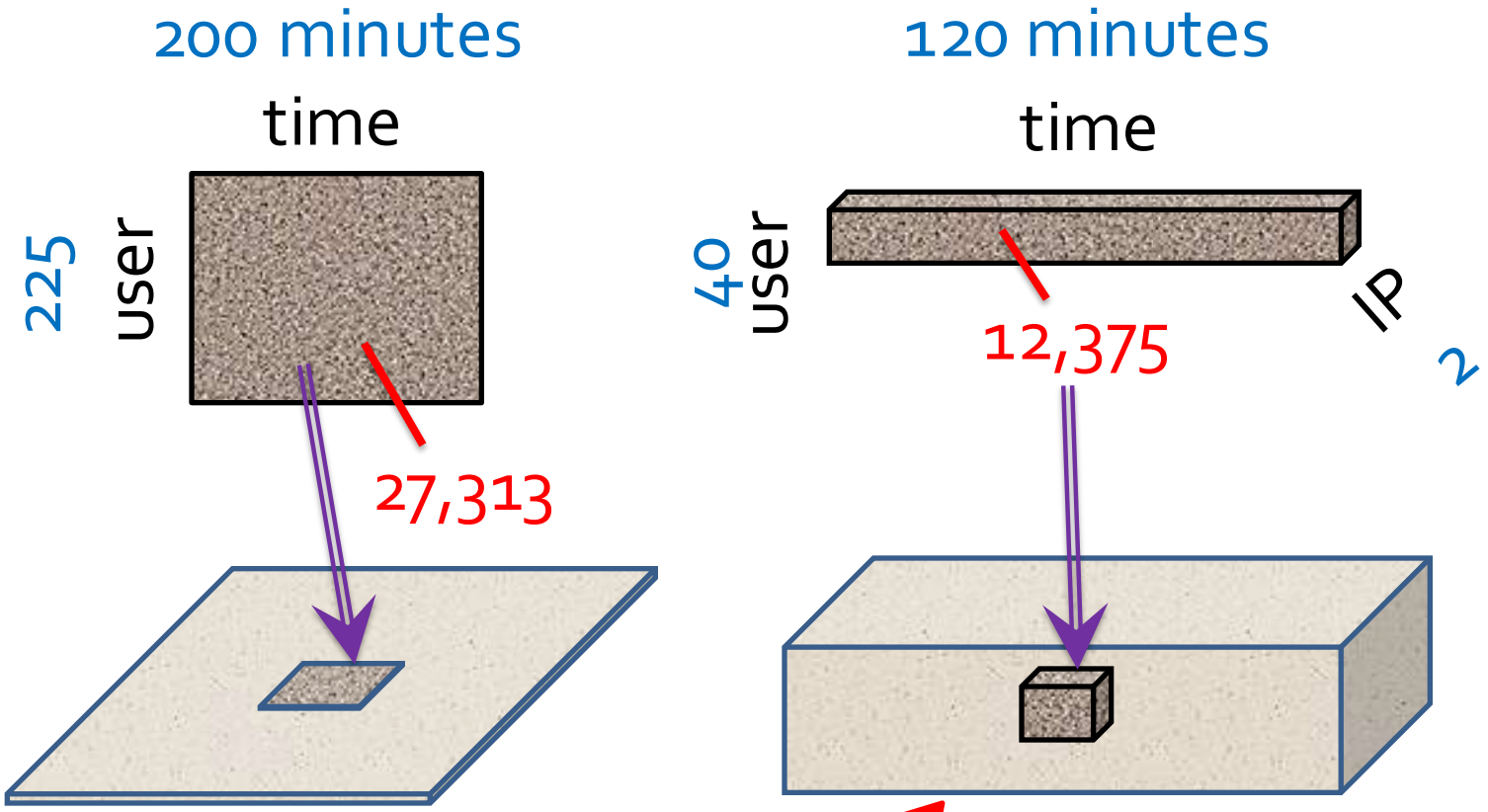


- New information (more modes) can only make our blocks more suspicious

Our Principled Idea: **Scoring Suspiciousness**



Our Principled Idea: **Scoring Suspiciousness**



0.9%

Probability

0.05%

A General Suspiciousness Metric

- Negative log likelihood of block's probability

$$f(n, c, N, C) = -\log\left[Pr(Y_n = c)\right]$$

Lemma *Given an $n_1 \times \cdots \times n_K$ block of mass c in $N_1 \times \cdots \times N_K$ data of total mass C, the suspiciousness function is*

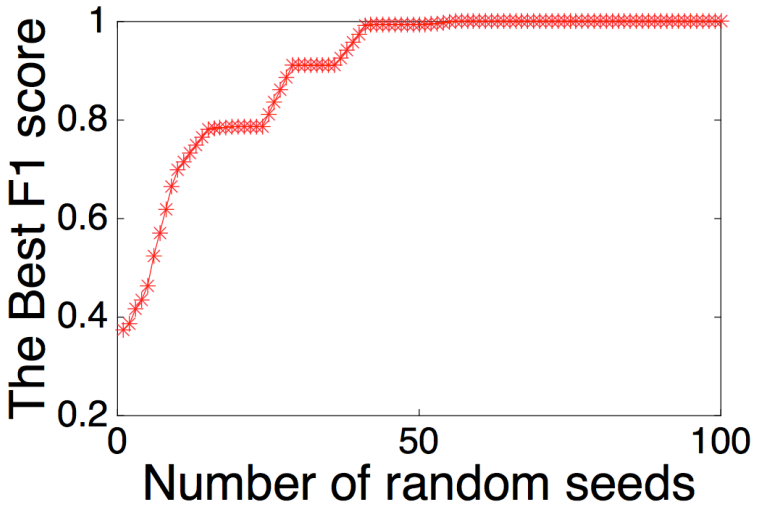
$$f(\mathbf{n}, c, \mathbf{N}, C) = c(\log \frac{c}{C} - 1) + C \prod_{i=1}^{K} \frac{n_i}{N_i} - c \sum_{i=1}^{K} \log \frac{n_i}{N_i}$$

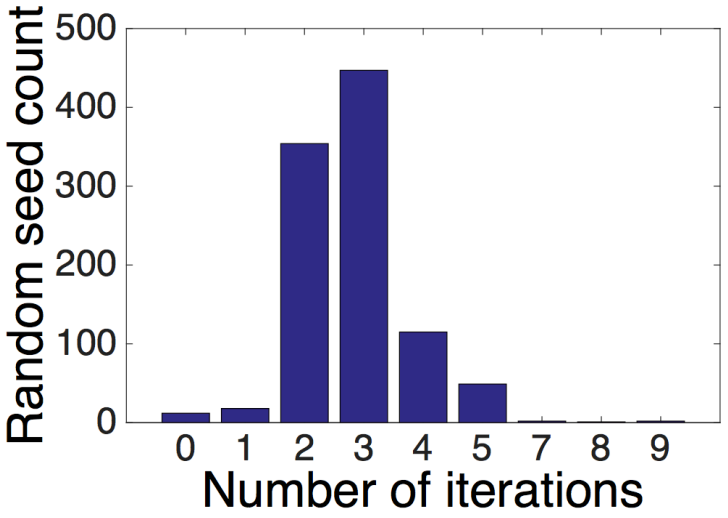
Using ρ as the block's density and p is the data's density, we have the simpler formulation

$$\hat{f}(\mathbf{n}, \rho, \mathbf{N}, p) = \left(\prod_{i=1}^{K} n_i\right) D_{KL}(\rho || p)$$

CrossSpot: Local Search with the Metric

- Seed block, adjust modes, select a mode, adjust values in mode, until convergence.
- Seed selection: HOSVD, or with LockInfer [PAKDD'14]
- Fast convergence





- Parallelize to multiple machines: Scalable!

Advantage: "Suspiciousness" and CrossSpot

- Score dense blocks
- Target multi-modal data
- Satisfy all the axioms

	Method	Scores Blocks	Axioms: Density 1	Size 2	Concentration 3	Contrast 4	Multi-modal 5
Metrics	**SUSPICIOUSNESS**	✔	✔	✔	✔	✔	✔
	Mass	✓	✓	✗	✗	✗	✓
	Density	✓	✓	✗	✓	✗	✗
	Average Degree [9]	✓	✓	✗	✗	✗	N/A
	Singular Value [10]	✓	✓	✓	✓	✗	✗
Methods	**CROSSSPOT**	✔	✔	✔	✔	✔	✔
	Subgraph [30, 10, 36]	✓	✓	✓	✓	✗	N/A
	CopyCatch [6]	✓	✓	✓	✓	✗	N/A
	EigenSpokes [31]	✗	N/A				
	TrustRank [14, 8]	✗	N/A				
	BP [28, 1]	✗	N/A				

Performance: Synthetic Data

- Experiments: Synthetic data
  - 1,000×1,000×1,000 of 10,000 random data
  - Block#1: 30×30×30 of 512 3 modes
  - Block#2: 30×30×1,000 of 512 2 modes
  - Block#3: 30×1,000×30 of 512 2 modes
  - Block#4: 1,000×30×30 of 512 2 modes

	Recall				Overall Evaluation		
	Block #1	Block #2	Block #3	Block #4	Precision	Recall	F1 score
HOSVD (r=20)	93.7%	29.5%	23.7%	21.3%	**0.983**	0.407	0.576
HOSVD (r=10)	91.3%	24.4%	18.5%	19.2%	0.972	0.317	0.478
HOSVD (r=5)	85.7%	10.0%	9.5%	11.4%	0.952	0.195	0.324
CrossSpot	**100%**	**99.9%**	**94.9%**	**95.4%**	0.978	**0.967**	**0.972**

Performance: Manipulating Trends

User × hashtag × IP × minute	Mass c	Suspiciousness
582 × 3 × 294 × **56,940**	5,941,821	111,799,948
188 × 1 × 313 × **56,943**	2,344,614	47,013,868
75 × 1 × 2 × 2,061	689,179	19,378,403

User ID	Time	IP address (city, province)	Tweet text with hashtag
USER-D	11-18 12:12:51	IP-1 (Deyang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-E	11-18 12:12:53	IP-1 (Deyang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-F	11-18 12:12:54	IP-2 (Zaozhuang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-E	11-18 12:17:55	IP-1 (Deyang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!
USER-F	11-18 12:17:56	IP-2 (Zaozhuang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!
USER-D	11-18 12:18:40	IP-1 (Deyang, Shandong)	**#Toshiba Bright Daren#** color personality test to find out your sense...
USER-E	11-18 17:00:31	IP-2 (Zaozhuang, Shandong)	**#Snow#** the Samsung GALAXY SII QQ Service customized version...
USER-D	11-18 17:00:49	IP-2 (Zaozhuang, Shandong)	**#Toshiba Bright Daren#** color personality test to find out your sense...
USER-F	11-18 17:00:56	IP-2 (Zaozhuang, Shandong)	**#Li Ning - a weapon with a hero#** good support activities!

Performance: Network Blocks

	#	Src-IP × dst-IP × port × second	Mass c	Suspiciousness
CrossSpot	1	411 × 9 × 6 × **3,610**	47,449	552,465
	2	533 × 6 × 1 × **3,610**	30,476	400,391
	3	5 × 5 × 2 × **3,610**	18,881	317,529
	4	11 × 7 × 7 × **3,610**	20,382	295,869
HOSVD	1	15 × 1 × 1 × 1,336	4,579	80,585
	2	1 × 2 × 2 × 1,035	1,035	18,308
	3	1 × 1 × 1 × 1,825	1,825	34,812
	4	1 × 13 × 6 × 181	1,722	29,224

Conclusion

- Proposed a general "suspiciousness" metric based on probability for multi-modal behaviors
- CrossSpot: Proposed a local search algorithm for catching suspicious behaviors