

# A GENERAL SUSPICIOUSNESS METRIC FOR DENSE BLOCKS IN MULTIMODAL DATA

---

Meng Jiang, University of Illinois at Urbana-Champaign, USA

Joint work with

Alex Beutel (CMU), Peng Cui (Tsinghua), Bryan Hooi (CMU),  
Shiqiang Yang (Tsinghua), Christos Faloutsos (CMU)



# ROADMAP

**1. Motivation & Problem**

**2. Proposed Method**

**3. Experiments**

# 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. #dayymm
4	254	Nora Apnila J...	Dec 14, 2013	Alhamdulilaahhhh...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
8	258	Social News	Dec 13, 2013	sap hana online training and placenet 2 http://t.co/JlaA41dnV
9	259	Nurianah	Dec 13, 2013	Baca notif fb... ada training dadakaann dari evang kita.... avo wara wiri ca...
10	260	Nora Apnila J...	Dec 13, 2013	Ianjutt di rumah dulu ikutan trainingnyaaaa..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...

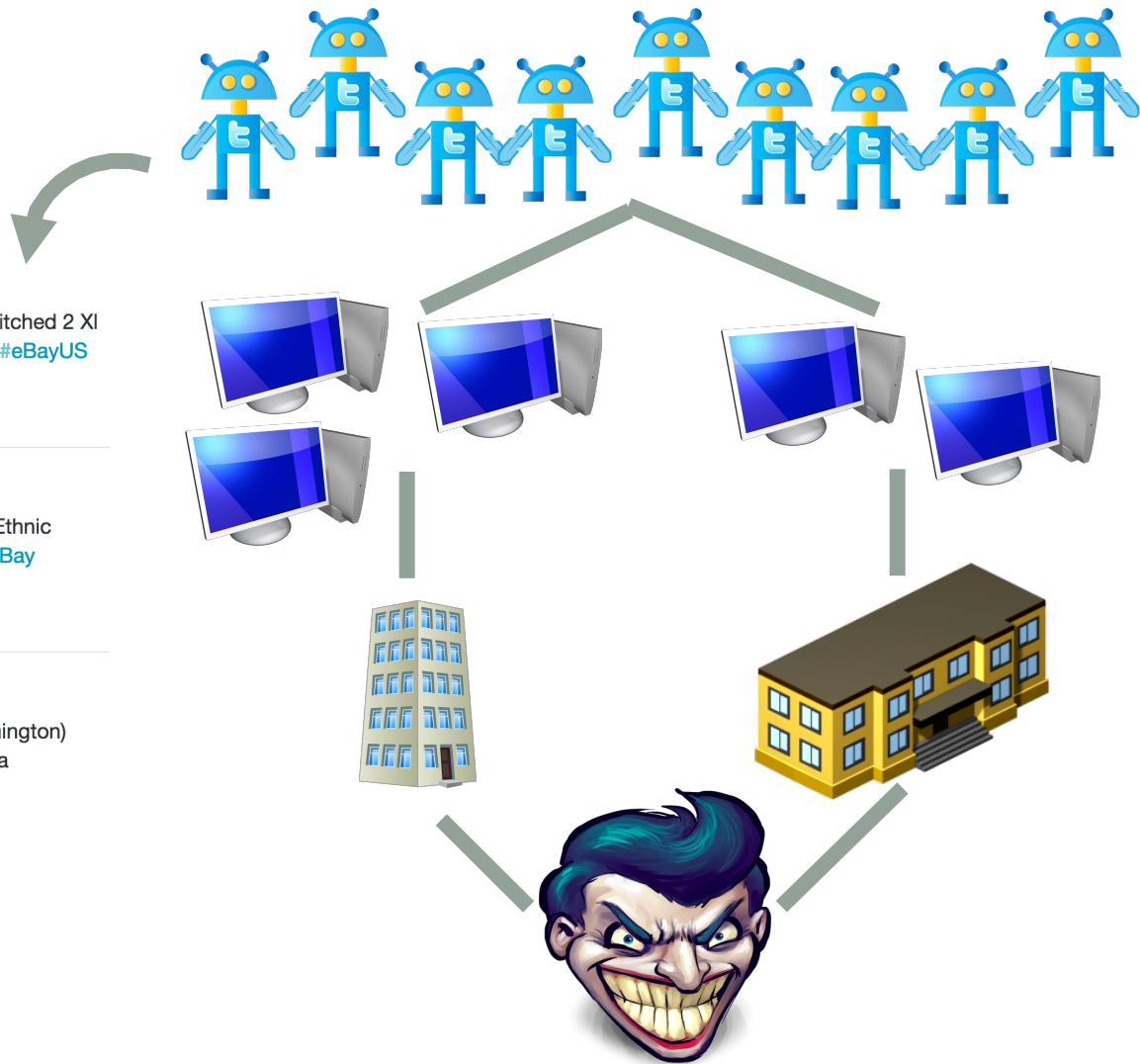
**fraud**

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

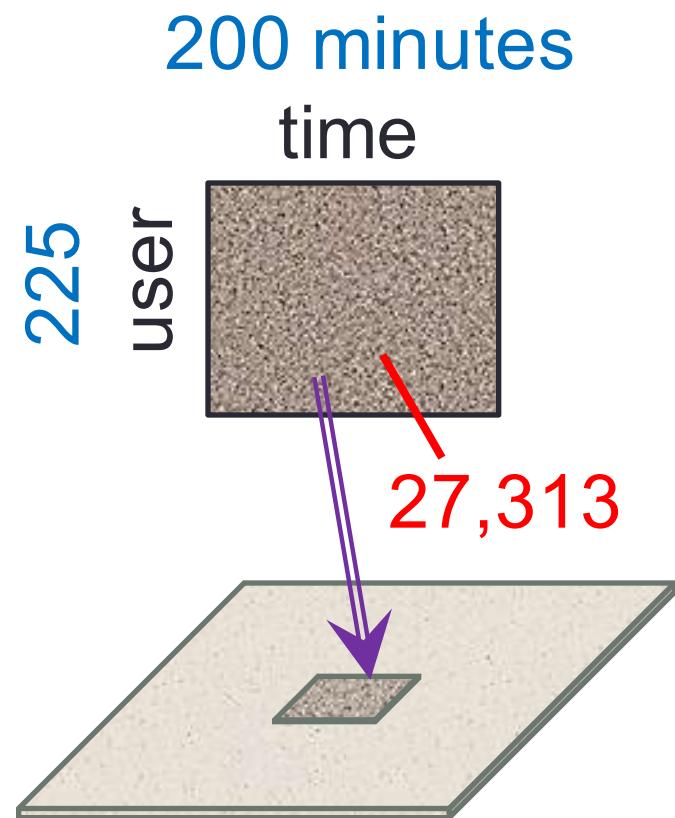
Dataset	Mode				Mass
Retweeting	User	Root ID	IP	Time (min)	#retweet
	29.5M	19.8M	27.8M	56.9K	211.7M
Trending (Hashtag)	User	Hashtag	IP	Time (min)	#tweet
	81.2M	1.6M	47.7M	56.9K	276.9M
Network attacks (LBNL)	Src-IP	Dest-IP	Port	Time (sec)	#packet
	2,345	2,355	6,055	3,610	230,836

# Suspicious Behaviors in Multi-Modal Data

- Wholesalebargain2015 Retweeted  
Real Time Deals @ebayrt · 2h  
 Seattle Mariners Mlb #Majestic Authentic Diamond Blue Stitched 2 XI M... (Sanford) USD 25 ebayrt.co/sports-mem-car... #eBay #eBayUS via @wil30225
- 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
- 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

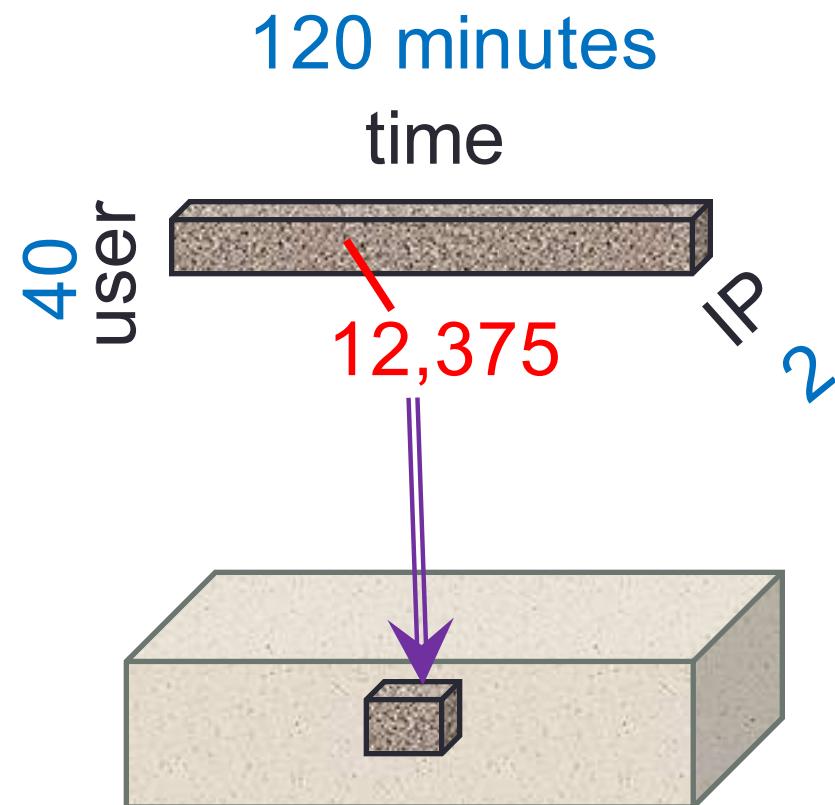


# Dense Blocks Indicates Suspiciousness

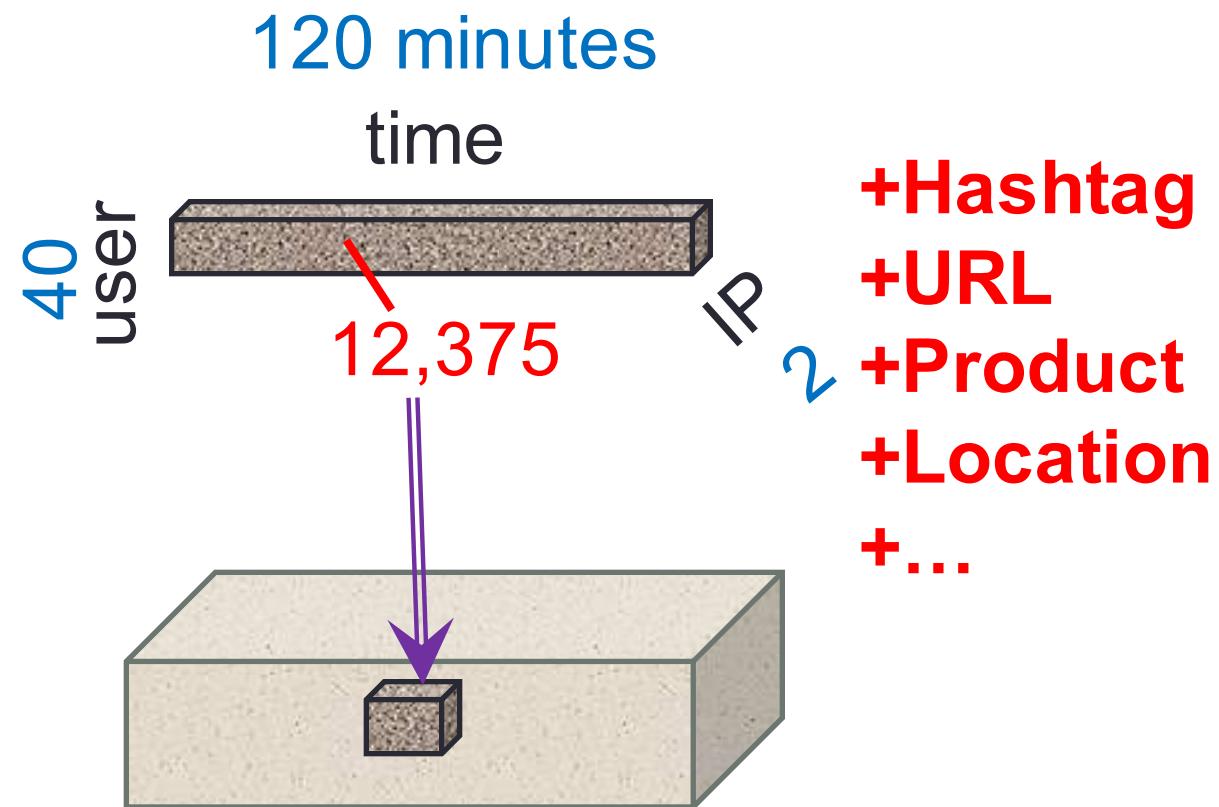


- Wholesalebargain2015 Retweeted  
**Real Time Deals** @ebayrt · 2h  
Seattle Mariners Mlb #Majestic Authentic Diamond Blue Stitched 2 XI  
M... (Sanford) USD 25 ebayrt.co/sports-mem-car... #eBay #eBayUS  
via @wil30225
- Wholesalebargain2015 Retweeted  
**Real Time Deals** @ebayrt · 2h  
Embroidered Navy Blue Aztec Mexican Top/ Long Sleeve Ethnic  
Mod... USD 35 ebayrt.co/clothing-shoes... #Handmade #eBay  
#eBayUS via @smilingbluedog
- 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

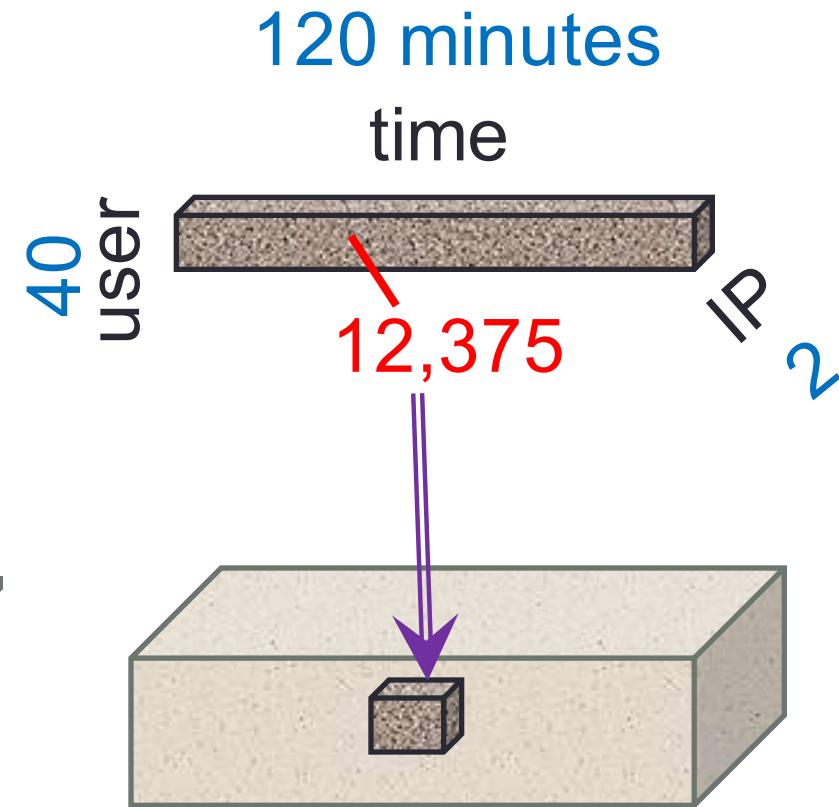
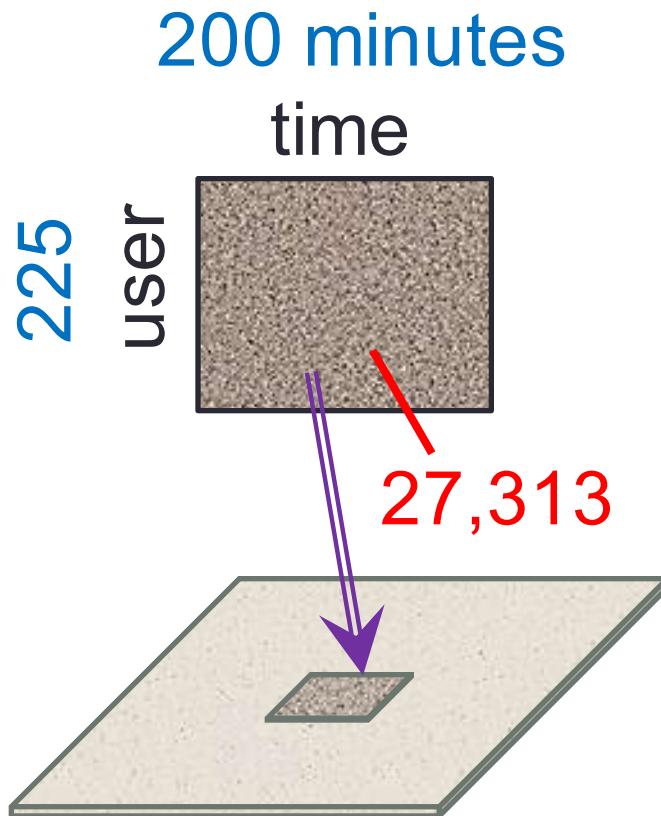
# Dense Blocks Indicates Suspiciousness



# Dense Blocks Indicates Suspiciousness



# Dense Blocks Indicates Suspiciousness



*Question:* Which is more suspicious?  
We need a metric to evaluate the suspiciousness.

# ROADMAP

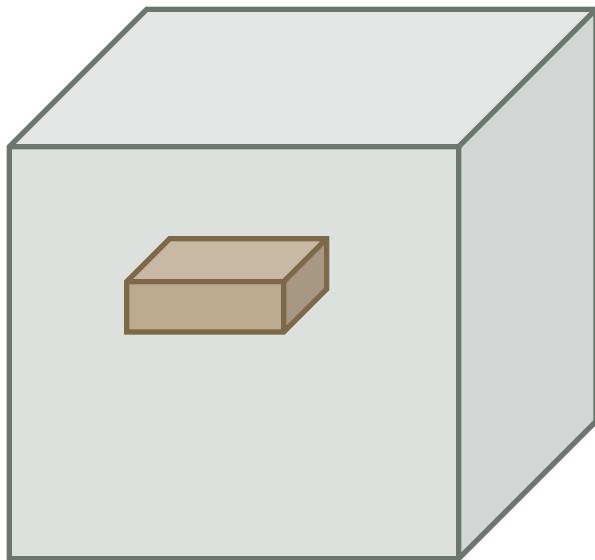
1. Motivation & Problem

2. Proposed Method

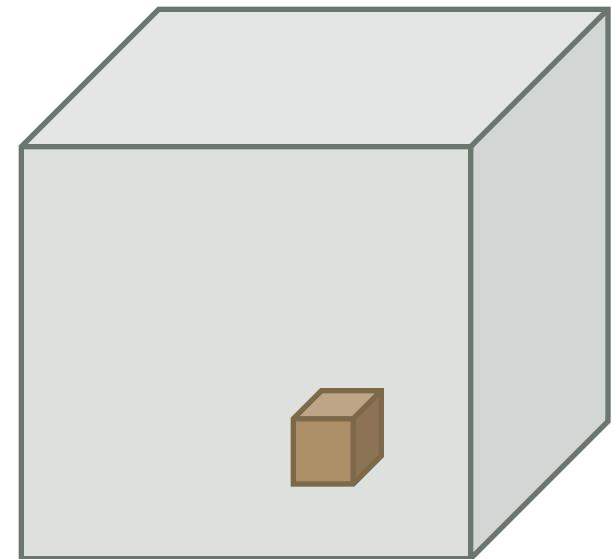
3. Experiments

# Metric Criteria

What properties are required of a good metric?



$N_1 \times N_2 \times N_3$   
Count data with  
total “mass”  $C$



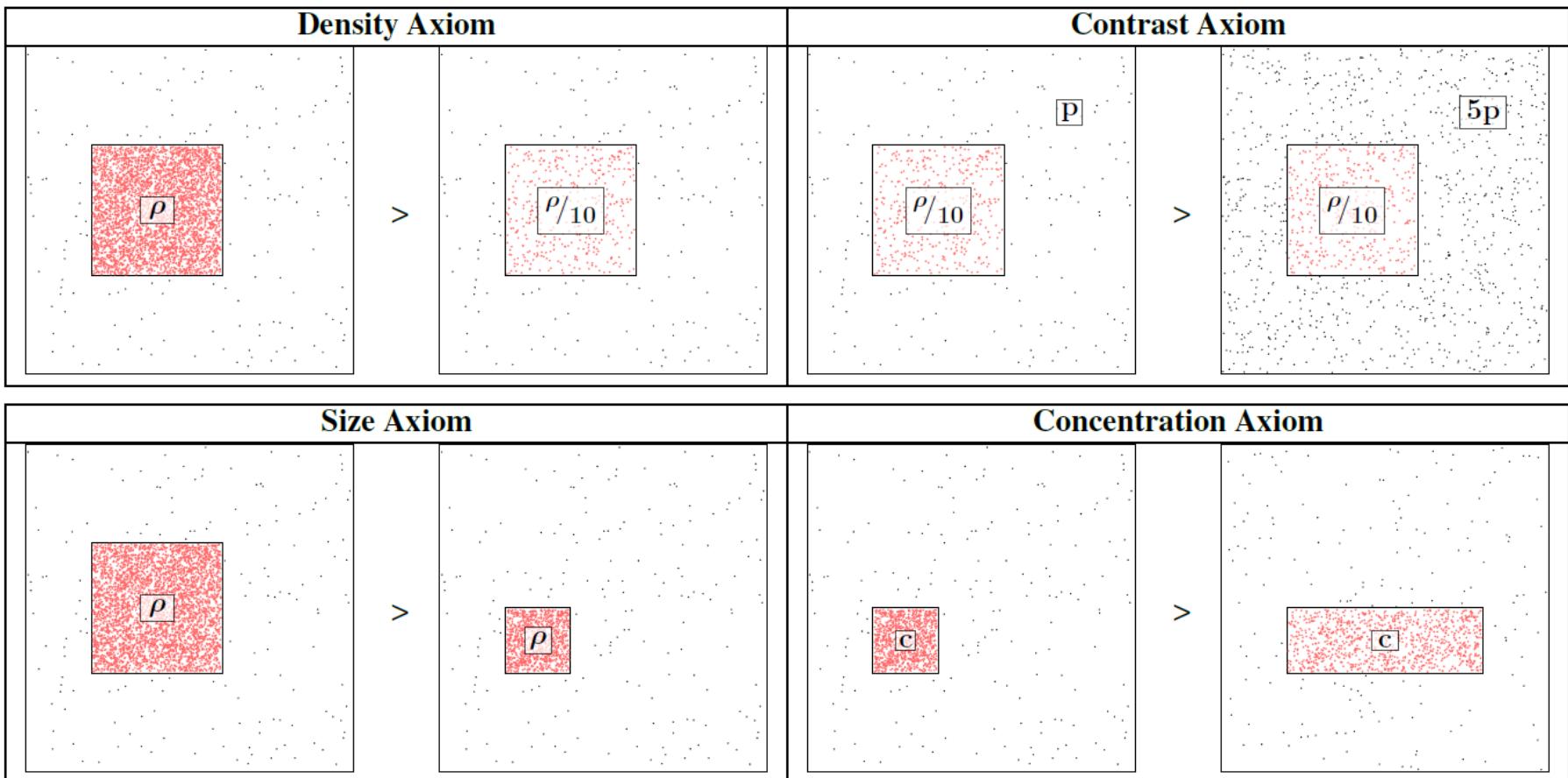
$$f( \begin{matrix} n_1 \times n_2 \times n_3 \\ \text{mass } c \\ \text{density } \rho \end{matrix} )$$

vs

$$f( \begin{matrix} n'_1 \times n'_2 \times n'_3 \\ \text{mass } c' \\ \text{density } \rho' \end{matrix} )$$

# Axioms 1-4

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

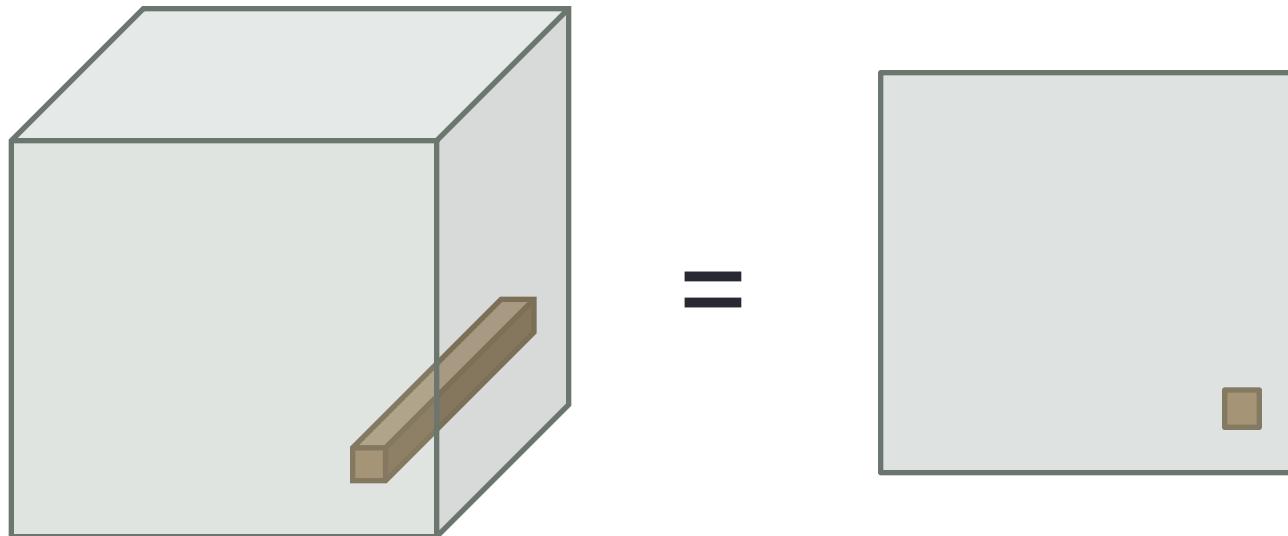


$$p_1 < p_2 \iff \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_1) > \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_2)$$

# Axiom 5: Multimodal

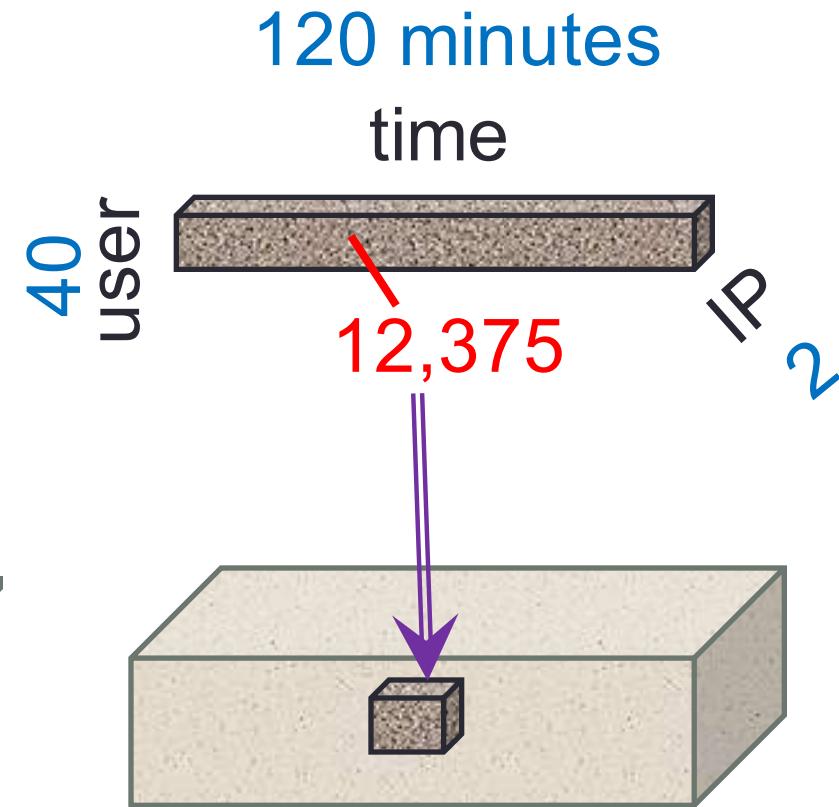
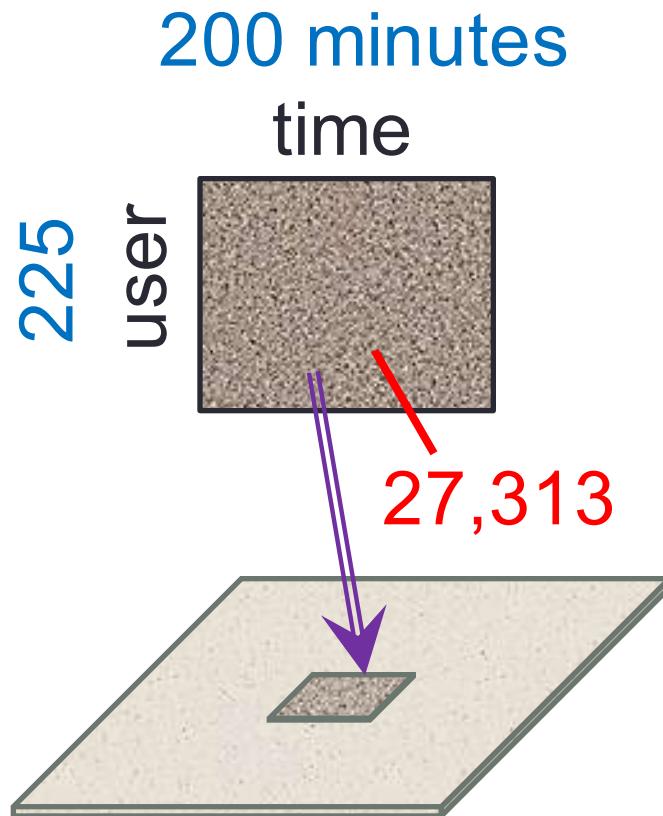
$$f_{K-1} \left( [n_k]_{k=1}^{K-1}, c, [N_k]_{k=1}^{K-1}, C \right) = f_K \left( ([n_k]_{k=1}^{K-1}, N_K), c, [N_k]_{k=1}^K, C \right)$$

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

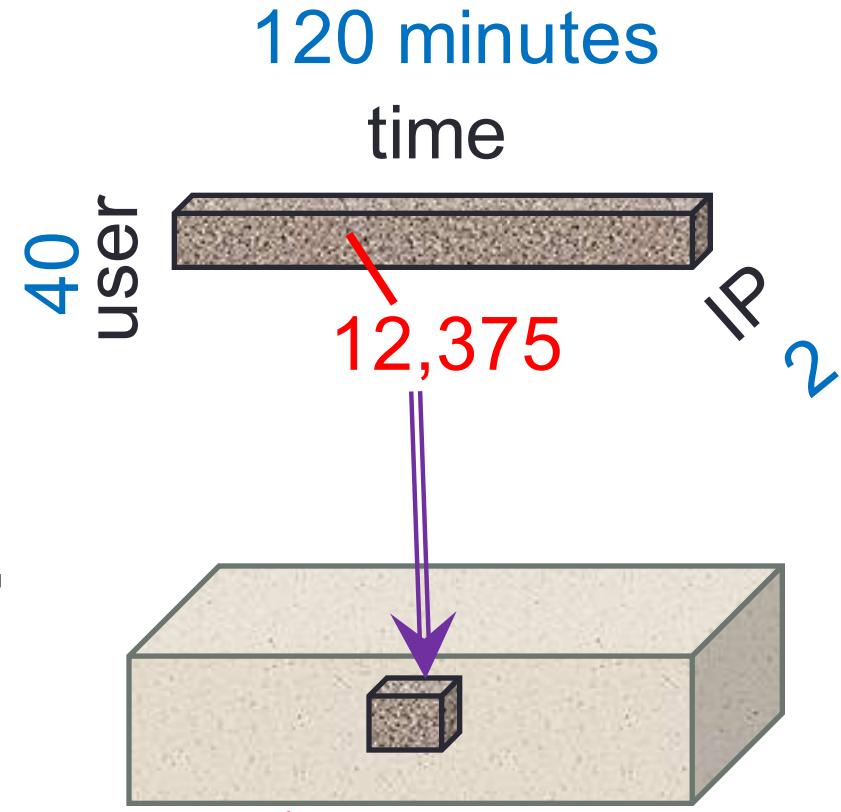
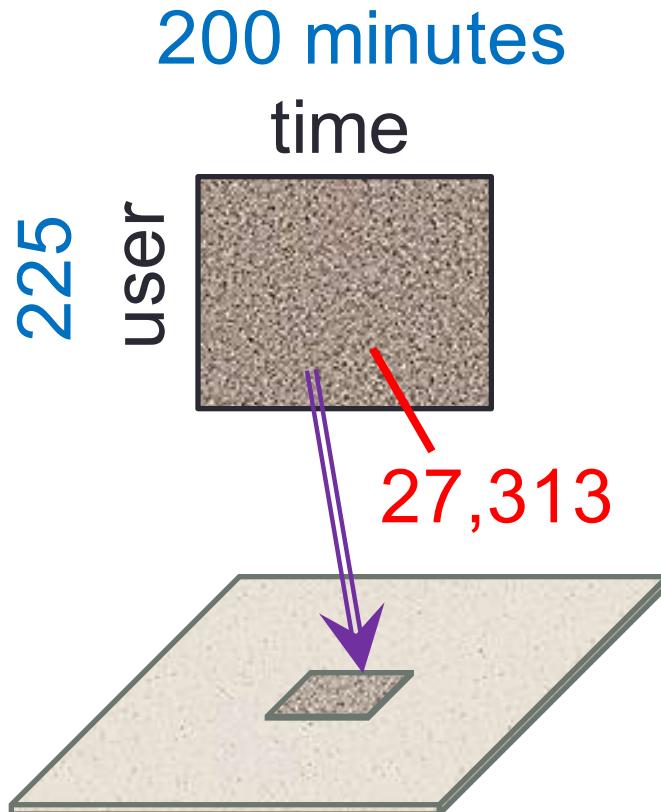


- ▶ New information (more modes) can only make our blocks more suspicious

# 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

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

**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

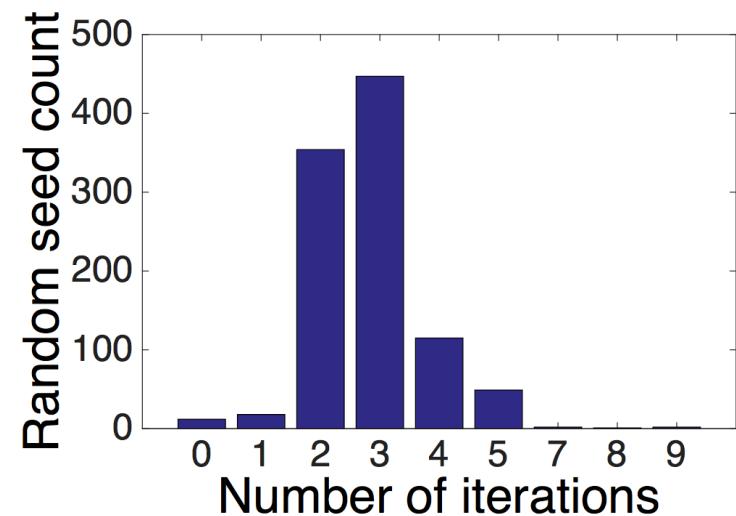
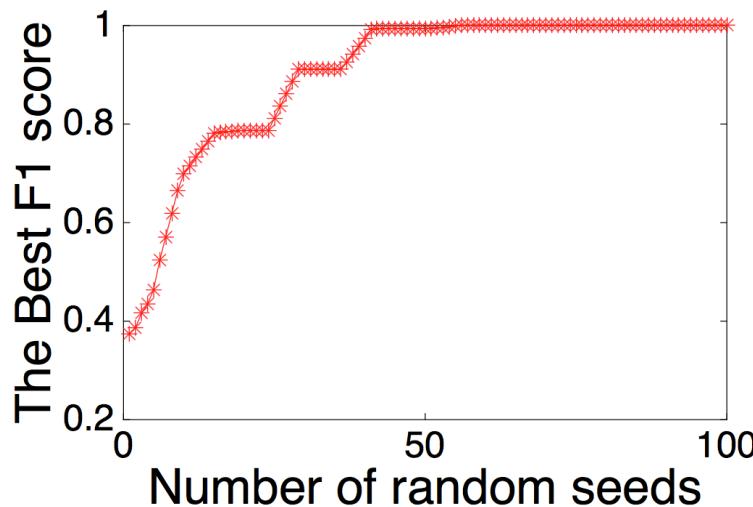
$$f(\mathbf{n}, c, \mathbf{N}, C) = c(\log \frac{c}{C} - 1) + C \prod_{i=1}^K \frac{n_i}{N_i} - c \sum_{i=1}^K \log \frac{n_i}{N_i}$$

Using  $\rho$  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”+CrossSpot

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

Metrics	Method	Scores Blocks	Axioms					Multi-modal
			Density	Size	Concentration	Contrast		
		1	2	3	4	5		
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]		✗					
	TrustRank [14, 8]		✗					
	BP [28, 1]		✗					

# ROADMAP

1. Motivation & Problem

2. Proposed Method

3. Experiments

# Performance: Synthetic Data

## ■ Experiments: Synthetic data



	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%	<b>0.983</b>	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	<b>100%</b>	<b>99.9%</b>	<b>94.9%</b>	<b>95.4%</b>	0.978	<b>0.967</b>	<b>0.972</b>

# Performance: Manipulating Trends

User × hashtag × IP × minute	Mass $c$	Suspiciousness
$582 \times 3 \times 294 \times \mathbf{56,940}$	5,941,821	111,799,948
$188 \times 1 \times 313 \times \mathbf{56,943}$	2,344,614	47,013,868
$75 \times 1 \times 2 \times 2,061$	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	$411 \times 9 \times 6 \times \mathbf{3,610}$	47,449	552,465
	2	$533 \times 6 \times 1 \times \mathbf{3,610}$	30,476	400,391
	3	$5 \times 5 \times 2 \times \mathbf{3,610}$	18,881	317,529
	4	$11 \times 7 \times 7 \times \mathbf{3,610}$	20,382	295,869
HOSVD	1	$15 \times 1 \times 1 \times 1,336$	4,579	80,585
	2	$1 \times 2 \times 2 \times 1,035$	1,035	18,308
	3	$1 \times 1 \times 1 \times 1,825$	1,825	34,812
	4	$1 \times 13 \times 6 \times 181$	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

Thank you!

- Meng Jiang, UIUC
- mjiang89@gmail.com
- [www.meng-jiang.com](http://www.meng-jiang.com)



# (Erdös-Rényi-)Poisson Model

$$X_i \sim \text{Poisson}(p)$$

$$f(Y) = -\log \left( \prod_{i \in Y} \text{Poisson}(Y_i | p) \right)$$

Suspiciousness metric is the negative log-likelihood of the sub-block's mass

# Suspiciousness Metric

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

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

Suspiciousness metric is the negative log-likelihood of the sub-block's mass

# Suspiciousness Metric

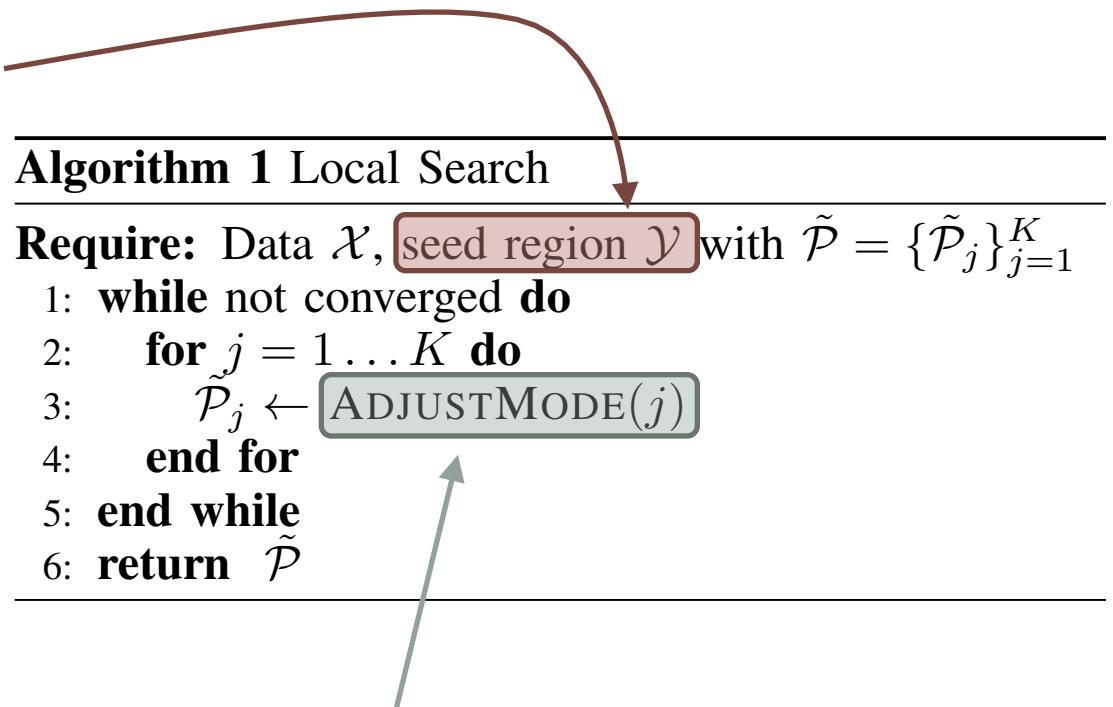
$$f(\mathbf{n}, c, \mathbf{N}, C) = c \left( \log \frac{c}{C} - 1 \right) + C \prod_{i=1}^K \frac{n_i}{N_i} - c \sum_{i=1}^K \log \frac{n_i}{N_i}$$

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

Satisfies all axioms!

# Search Algorithm

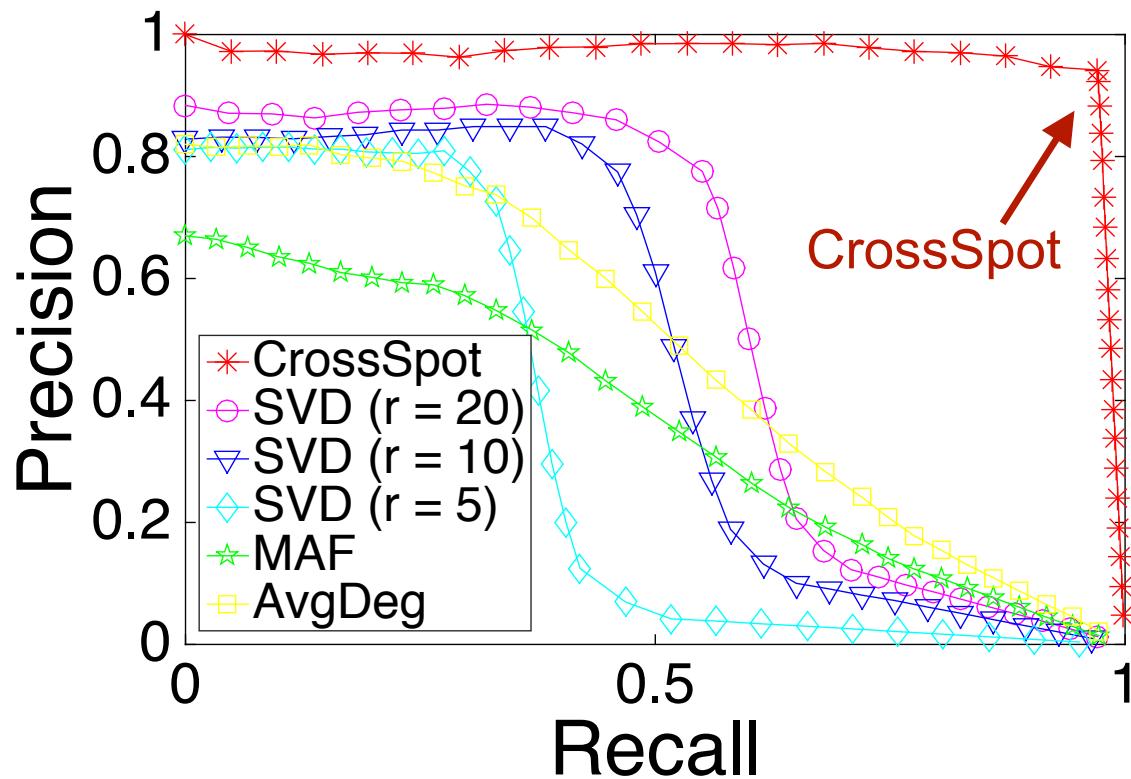
Can use previous methods  
to seed algorithm



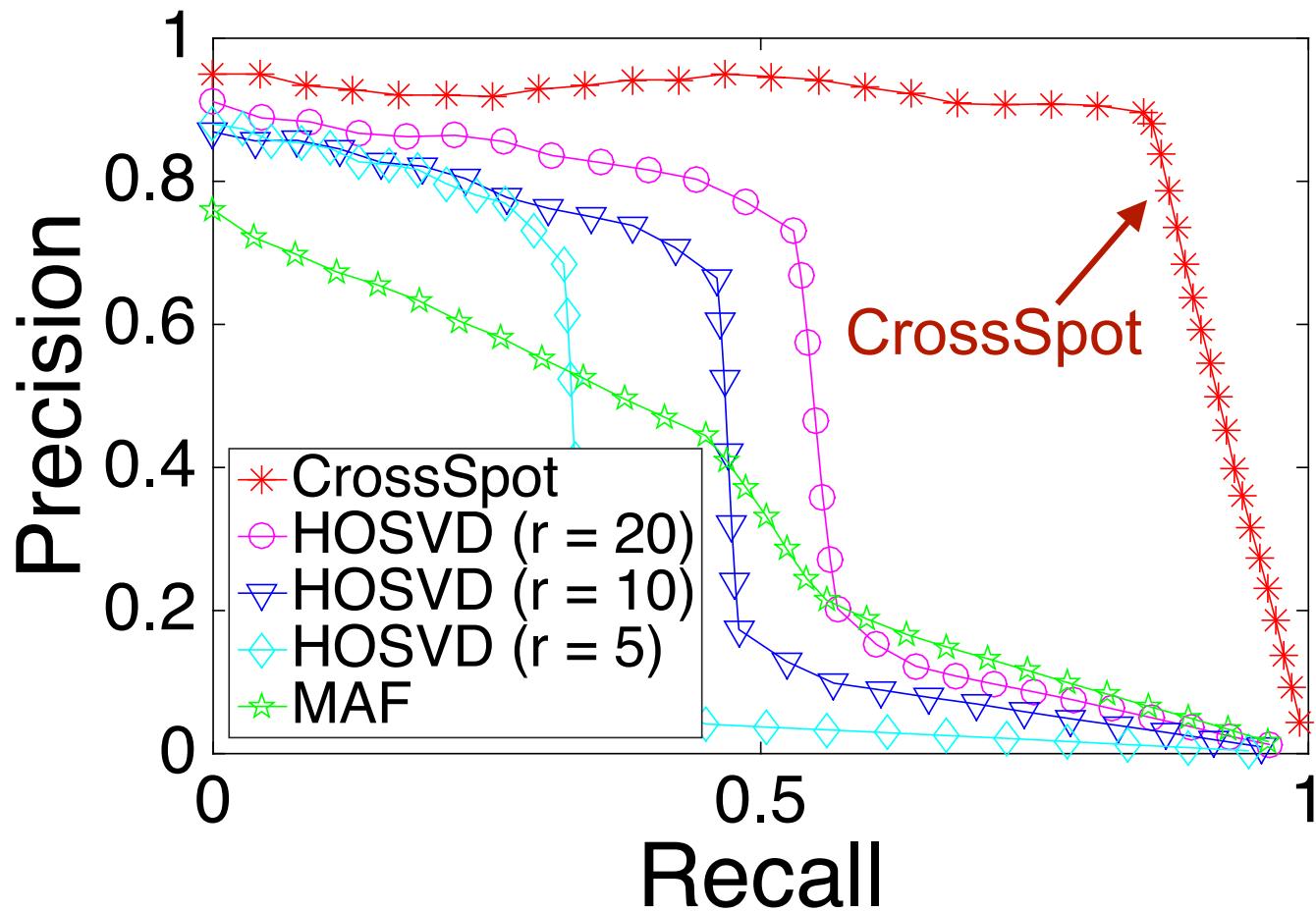
Find optimal\* subset of indices in mode  $j$  in  $O(N_j \log N_j)$  time.

\*Optimal given other modes are held constant.

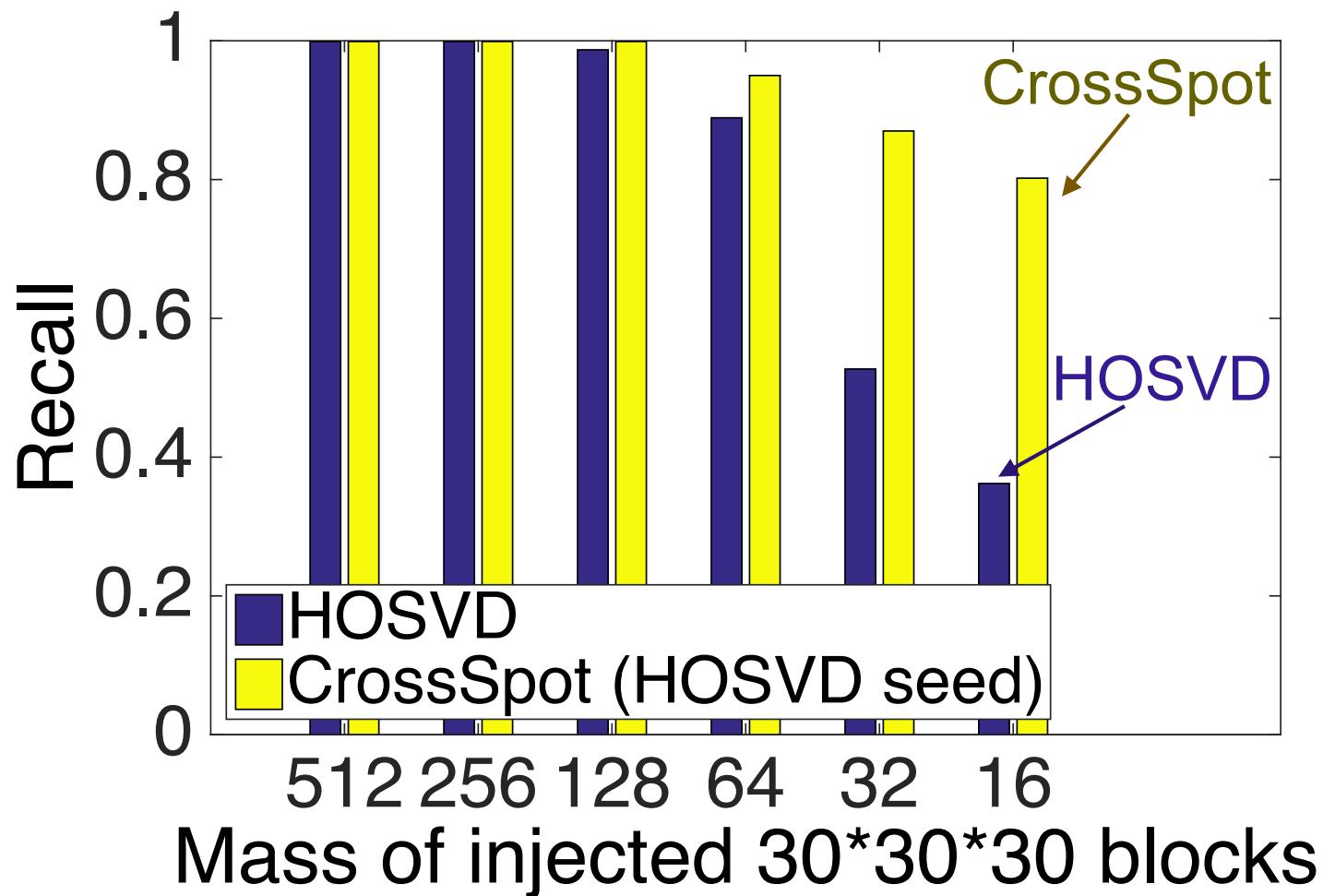
# Synthetic Tests (Matrix)



# Synthetic Tests (3-mode Tensor)



# Synthetic Tests (3-mode Tensor)



# Suspicious Retweet Blocks

	#	User × tweet × IP × minute	Mass $c$	Suspiciousness
CROSSSPOT	1	$14 \times 1 \times 2 \times 1,114$	41,396	1,239,865
	2	$225 \times 1 \times 2 \times 200$	27,313	777,781
	3	$8 \times 2 \times 4 \times 1,872$	17,701	491,323
HOSVD	1	$24 \times 6 \times 11 \times 439$	3,582	131,113
	2	$18 \times 4 \times 5 \times 223$	1,942	74,087
	3	$14 \times 2 \times 1 \times 265$	9,061	381,211

TABLE VII. RETWEETING BOOSTING: WE SPOT A GROUP OF USERS RETWEET “GALAXY NOTE DREAM PROJECT: HAPPY HAPPY LIFE TRAVELLING THE WORLD” IN LOCKSTEP (EVERY 5 MINUTES) ON THE SAME GROUP OF IP ADDRESSES. (RETWEETING LOG IN BLOCK  $225 \times 1 \times 2 \times 200$  IN TABLE VI)

User ID	Time	IP address (city, province)	Retweet comment (Google translator: from Simplified Chinese to English)
USER-A	11-26 10:08:54	IP-1 (Liaocheng Shandong)	Qi Xiao Qi: "unspoken rules count ass ah, the day listening..."
USER-B	11-26 10:08:54	IP-1 (Liaocheng Shandong)	You gave me a promise, I will give you a result...
USER-C	11-26 10:09:07	IP-2 (Liaocheng Shandong)	Clouds have dispersed, the horse is already back to God...
USER-A	11-26 10:13:55	IP-1 (Liaocheng Shandong)	People always disgust smelly socks, it remains to his bed...
USER-B	11-26 10:13:57	IP-2 (Liaocheng Shandong)	Next life do koalas sleep 20 hours a day, eat two hours...
USER-C	11-26 10:14:03	IP-1 (Liaocheng Shandong)	all we really need to survive is one person who truly...
USER-A	11-26 10:18:57	IP-1 (Liaocheng Shandong)	Coins and flowers after the same amount of time...
USER-C	11-26 10:19:18	IP-2 (Liaocheng Shandong)	My computer is blue screen
USER-B	11-26 10:19:31	IP-1 (Liaocheng Shandong)	Finally believe that in real life there is no so-called...
USER-A	11-26 10:23:50	IP-1 (Liaocheng Shandong)	Do not be obsessed brother, only a prop.
USER-B	11-26 10:24:04	IP-2 (Liaocheng Shandong)	Life is like stationery, every day we loaded pen
USER-C	11-26 10:24:19	IP-1 (Liaocheng Shandong)	"The sentence: the annual party 1.25 Hidetoshi premature..."