Data-Driven Behavioral Analytics:

Observations, Representations and Algorithms for ***Suspicious Behavior Detection***

Meng Jiang

University of Illinois at Urbana-Champaign

Welcome to visit my homepage www.meng-jiang.com !

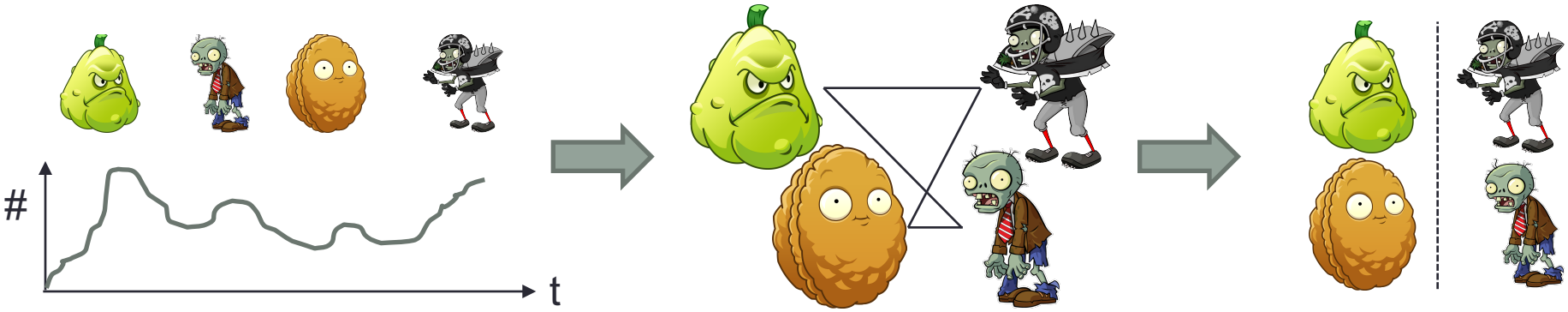
Behavior

- **Human behavior** refers to the ***array of every physical action*** … associated with ***individuals***, as well as the ***human race*** as a whole. (From Wikipedia)

Behavioral Analytics

Methodology

Understanding	Observations	Who, what, where, when, why, how… (scientific view)
Modeling	Representations	Graph, network, matrix, tensor… (mathematical view)
Intervening	Algorithms	Classification, prediction, recommendation, anomaly detection… (application view)



Behavioral Analytics/Modeling

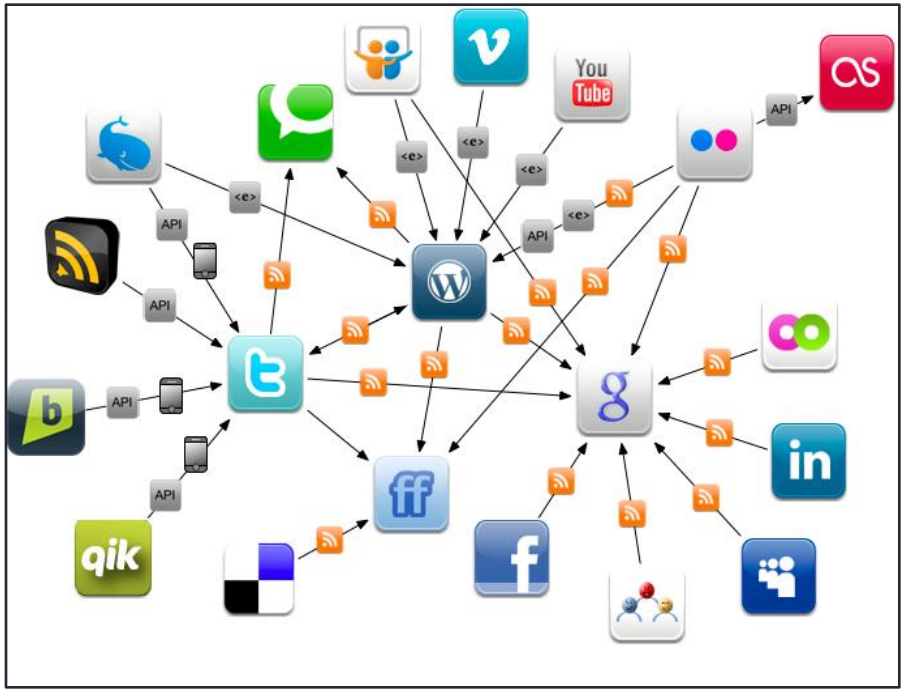
Basic Research Areas

- Six Disruptive Basic Research Areas
  - Engineered Materials (metamaterials and plasmonics)
  - Quantum Information and Control
  - Cognitive Neuroscience
  - Nanoscience and Nanoengineering
  - Synthetic Biology
  - Computational Modeling of Human and Social Behavior

From Experience-Driven to Data-Driven

Physical World

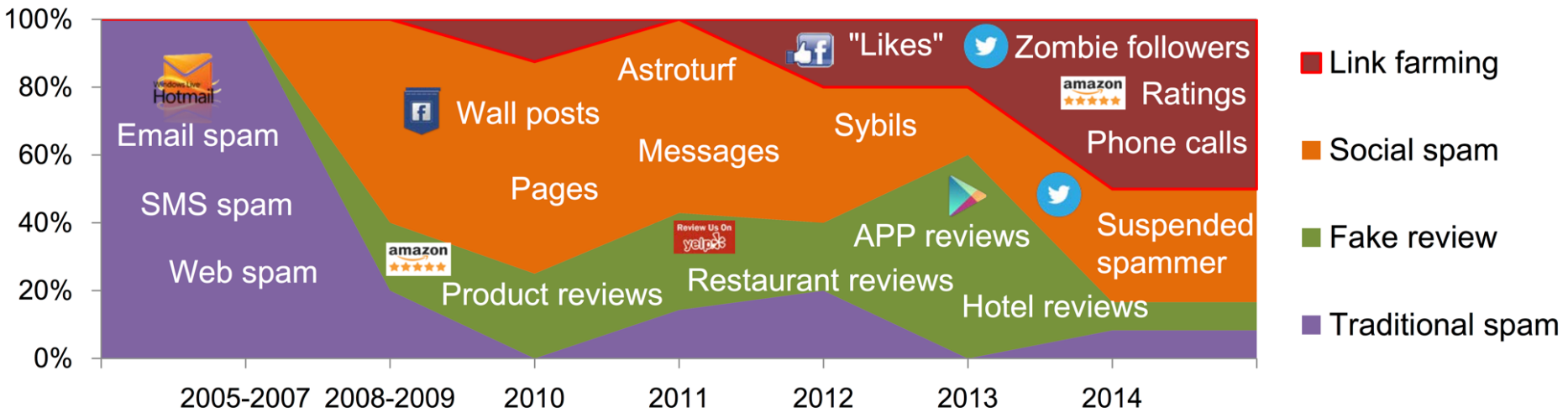
Online Behavioral Data



The human behaviors are broadly and deeply recorded in an unprecedented level.

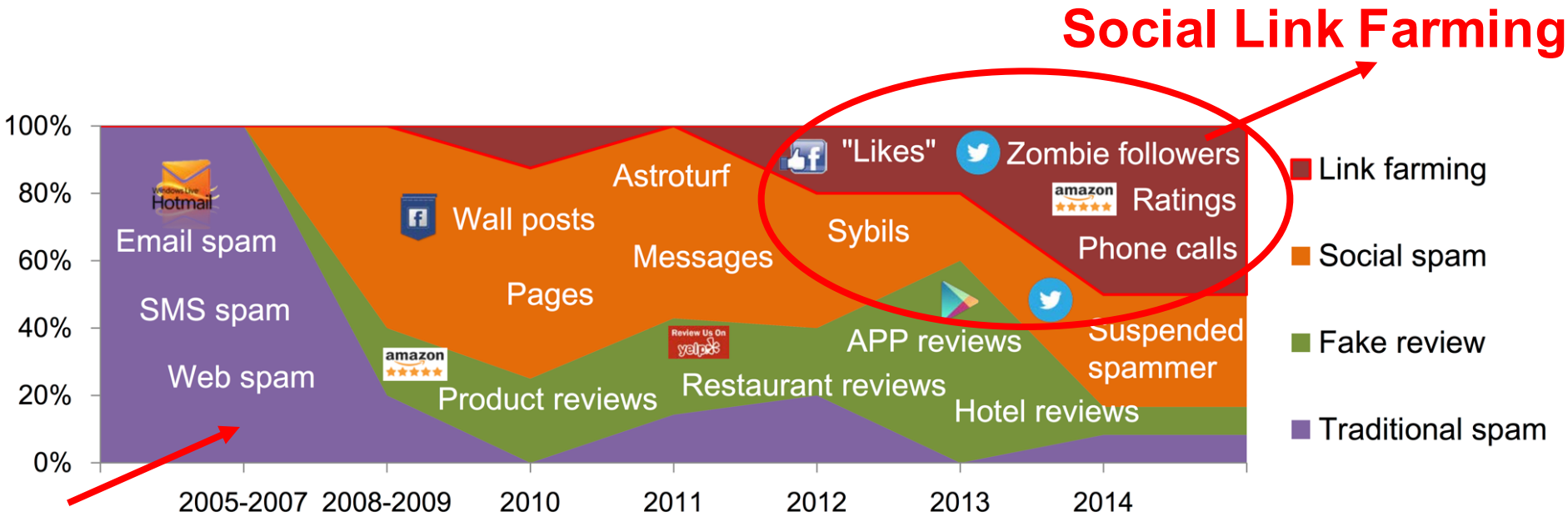
This is the first time that we can get insights of human behaviors and the society from large scale real data.

Data-Driven Approaches for Suspicious Behavior Detection



Meng Jiang, Peng Cui and Christos Faloutsos.
Suspicious Behavior Detection: Current Trends and Future Directions.
IEEE Intelligent Systems (ISSI), 2016.

Data-Driven Approaches for Suspicious Behavior Detection



Meng Jiang, Peng Cui and Christos Faloutsos.
Suspicious Behavior Detection: Current Trends and Future Directions.
IEEE Intelligent Systems (ISSI), 2016.

Roadmap

- Can we spot and catch the suspicious behaviors in a scalable and principled way?
- Behavioral patterns
  - W1. Spotting and catching synchronized behaviors (*KDD'14 Best Paper Finalist*)
  - W2. Evaluating suspiciousness in multiple dimensions (*ICDM'15, TKDE'16*)
  - W3. Representing and summarizing dynamic and multi-contextual behaviors (*KDD'16*)

Roadmap

- Can we spot and catch the suspicious behaviors in a scalable and principled way?
- Behavioral patterns
  - **W1. Spotting and catching synchronized behaviors (*KDD'14 Best Paper Finalist*)**
  - W2. Evaluating suspiciousness in multiple dimensions (*ICDM'15, TKDE'16*)
  - W3. Representing and summarizing dynamic and multi-contextual behaviors (*KDD'16*)

Social Link Farming

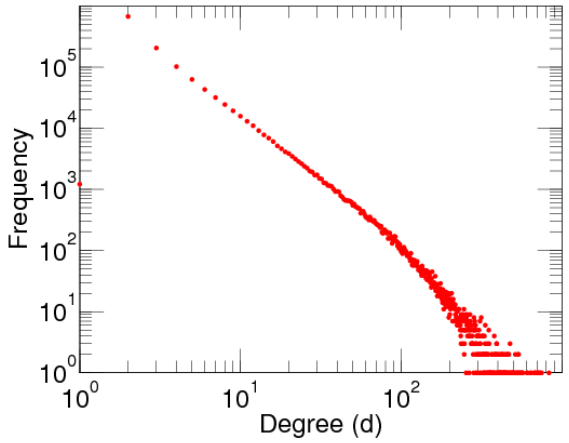
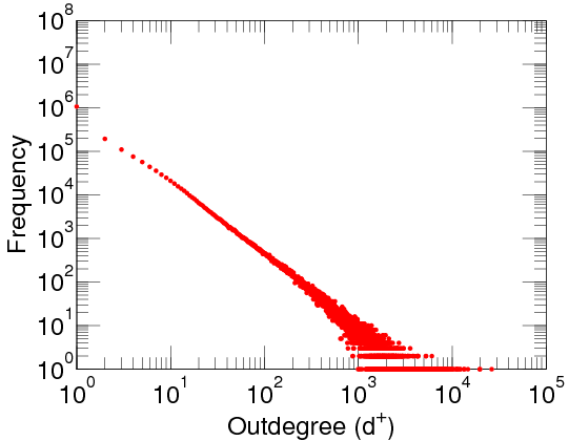
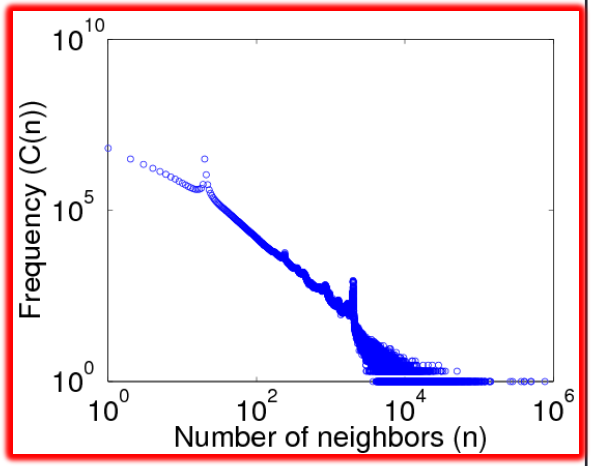
- Selling Twitter followers

Social Link Farming

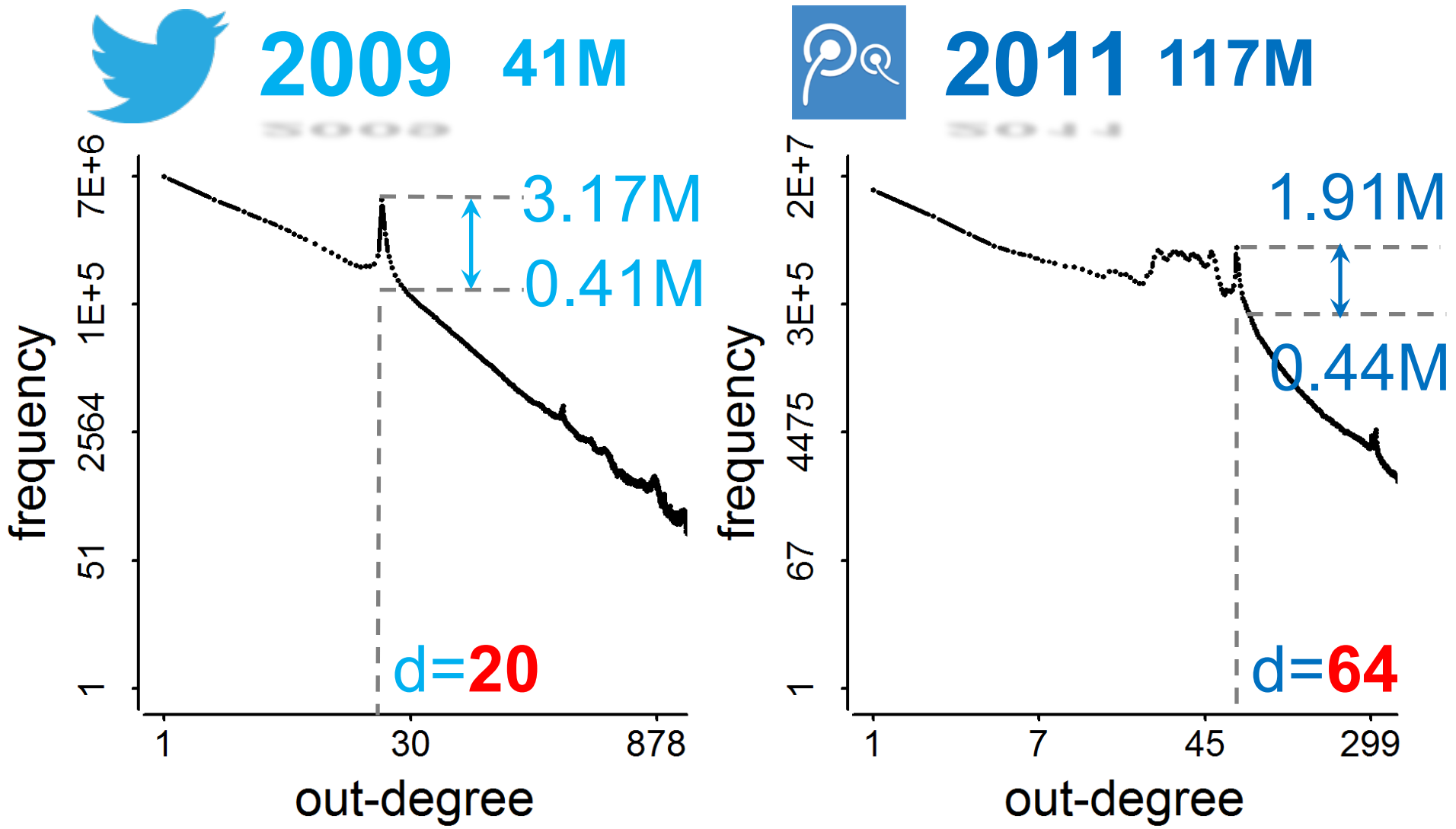
- Selling Facebook Likes

Observation: Power-Law Distribution

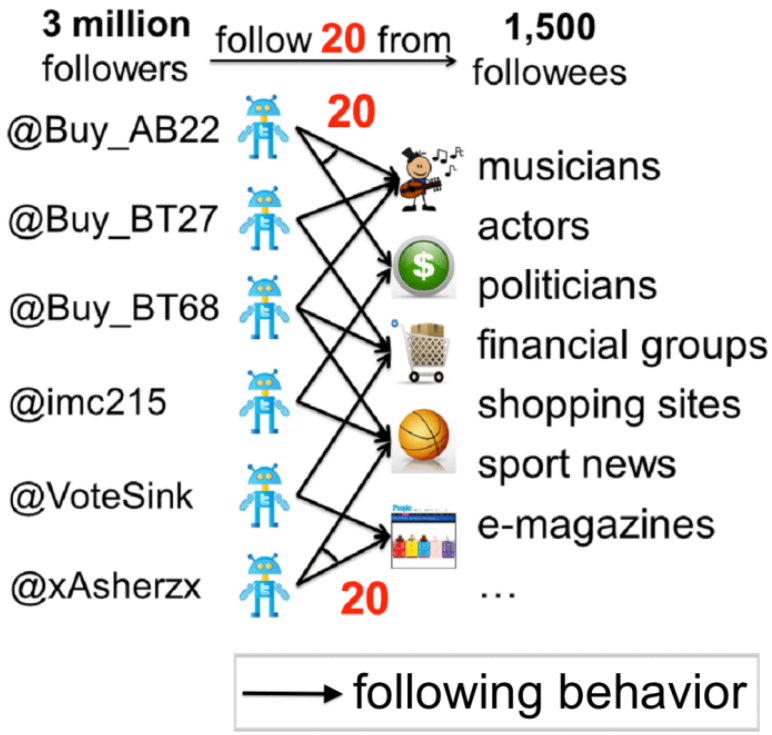
- Out-degree distribution

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

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

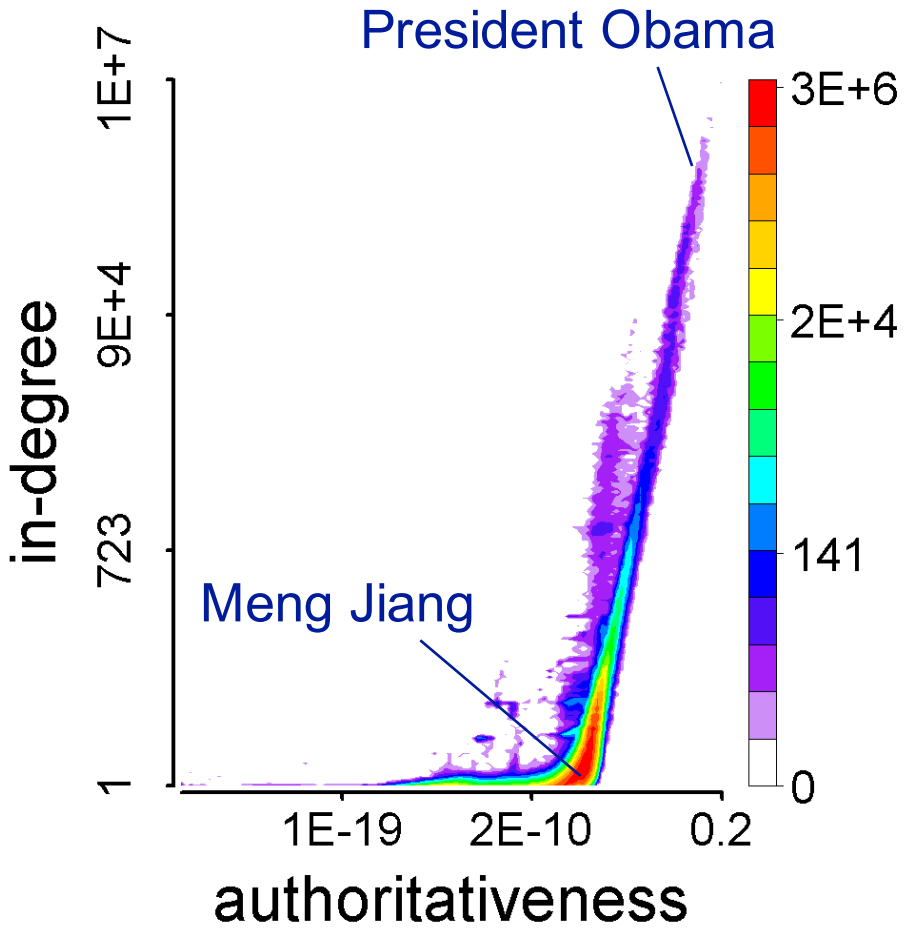

2009
41M
3.17M
0.41M
d=20
frequency
7E+6
1E+5
2564
51
1
1
30
878
out-degree
2011
117M
1.91M
0.44M
d=64
frequency
2E+7
3E+5
4475
67
1
1
7
45
299
out-degree

Observation: Smart Zombie Followers



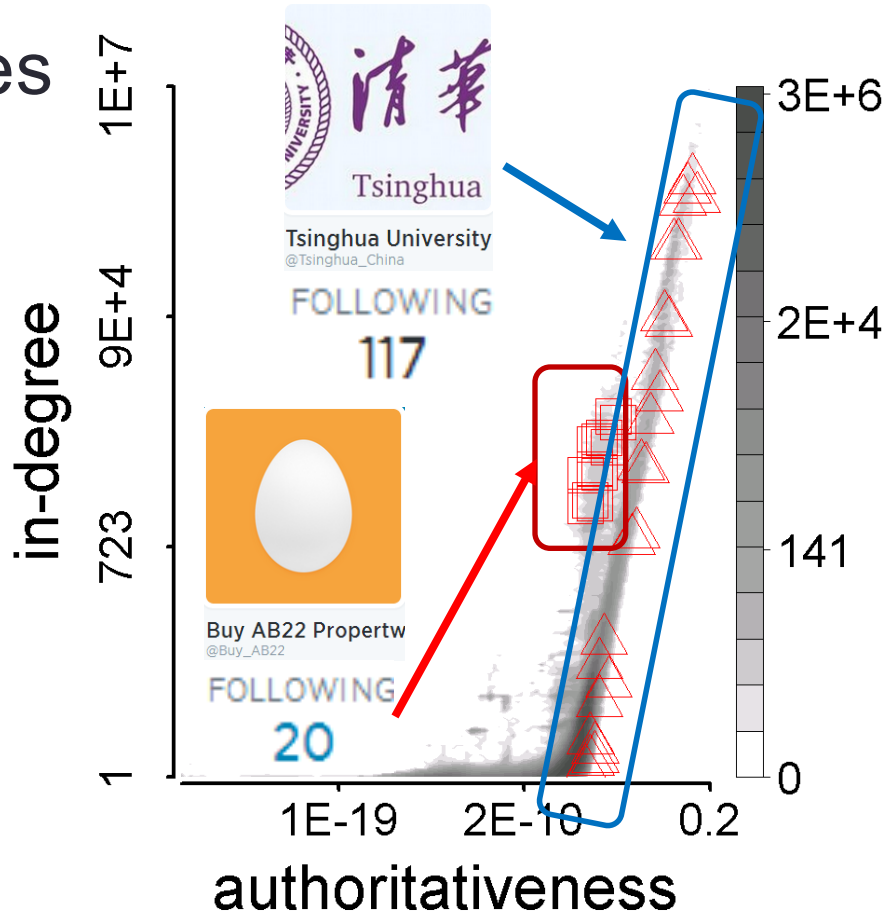
- Taking them individually, it's difficult to tell whether they are fake followers
  - Silent: very few tweets
  - Ordinary: very few followers
  - Unaggressive: only 20/1,500
- Traditional features and classifiers fail
  - #hashtag, #URL…
  - Out-degree, in-degree…

Observation: Feature Space of Followees



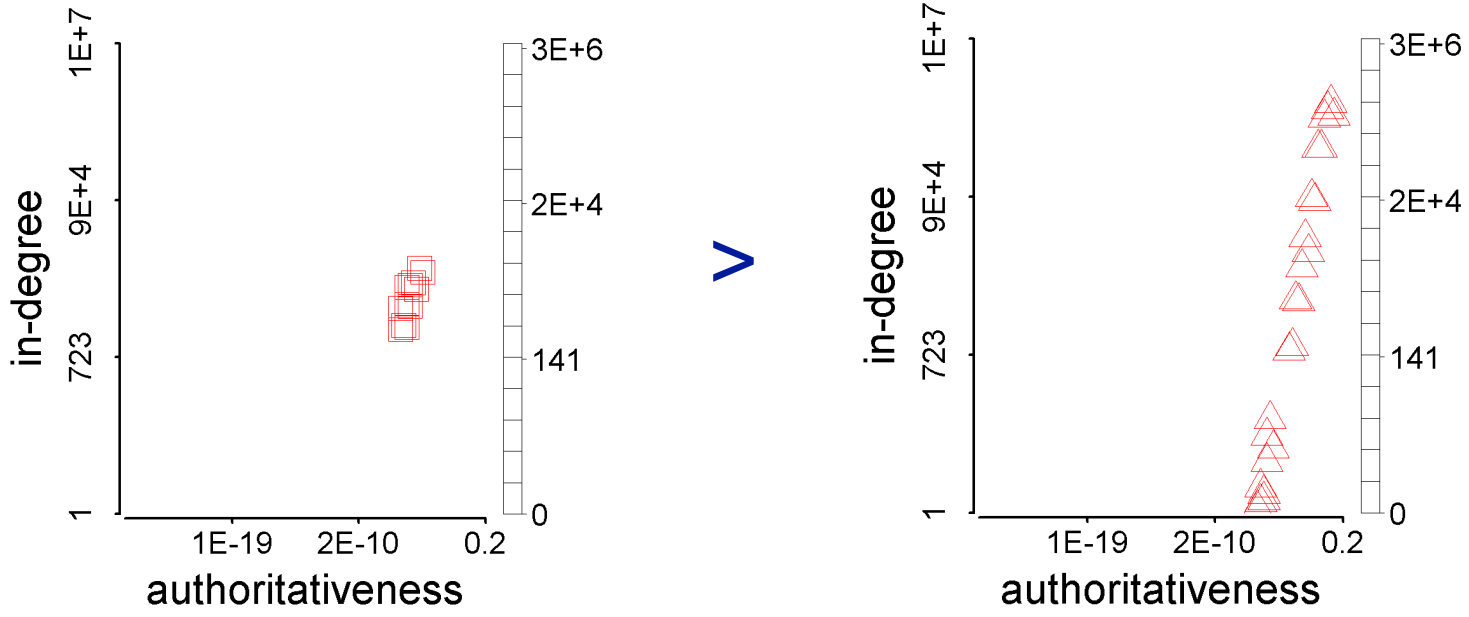
Observation: Who Are Their Followees

- Buy AB22's followees
  - Synchronized
  - Abnormal



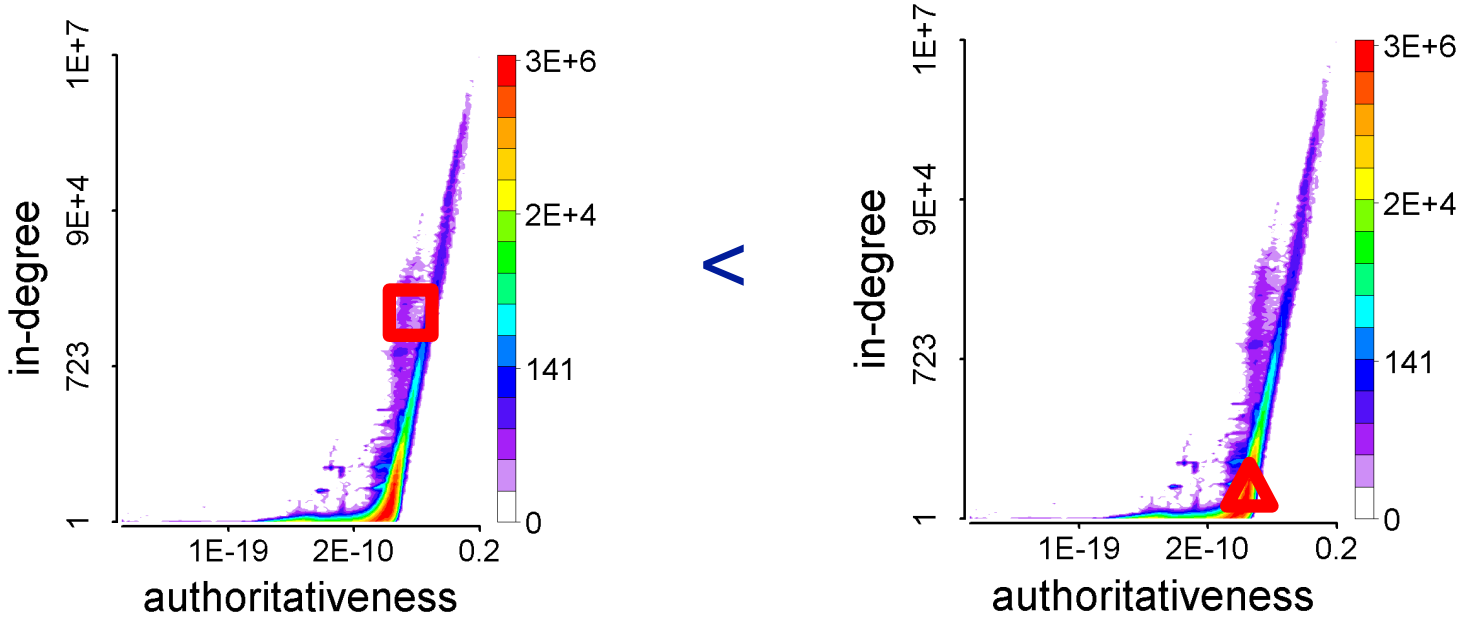
Representation: Synchronicity

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



Representation: Normality

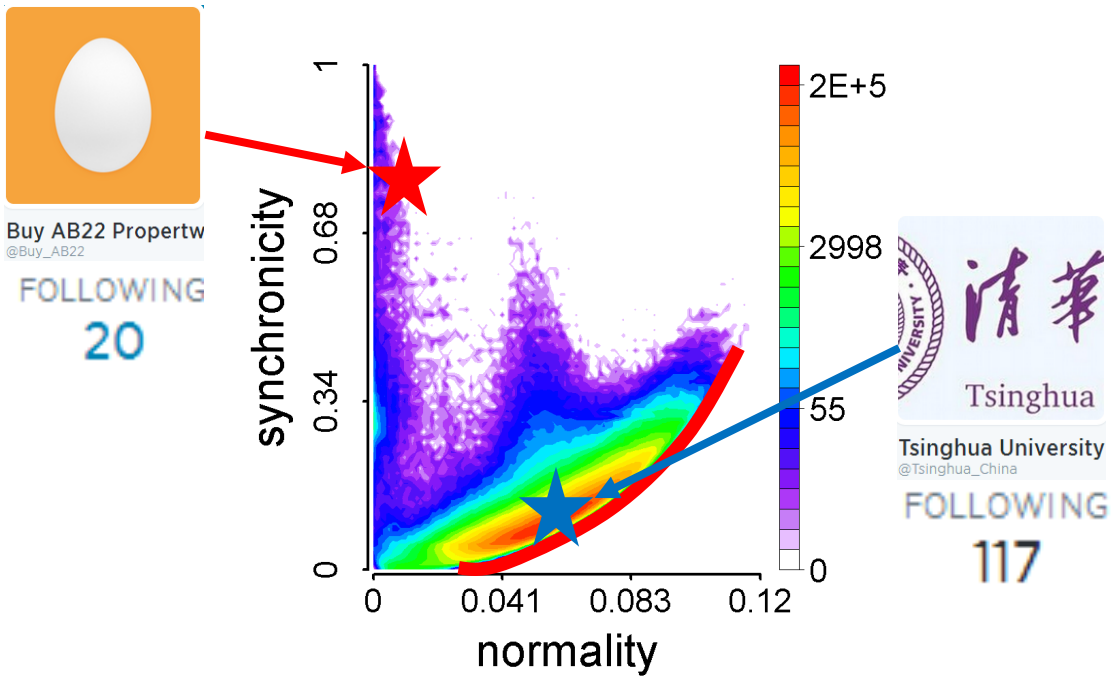
$$norm(u) = \frac{\sum_{(v,v') \in \mathcal{F}(u) \times \mathcal{U}} \mathbf{p}(v) \cdot \mathbf{p}(v')}{d(u) \times N}$$



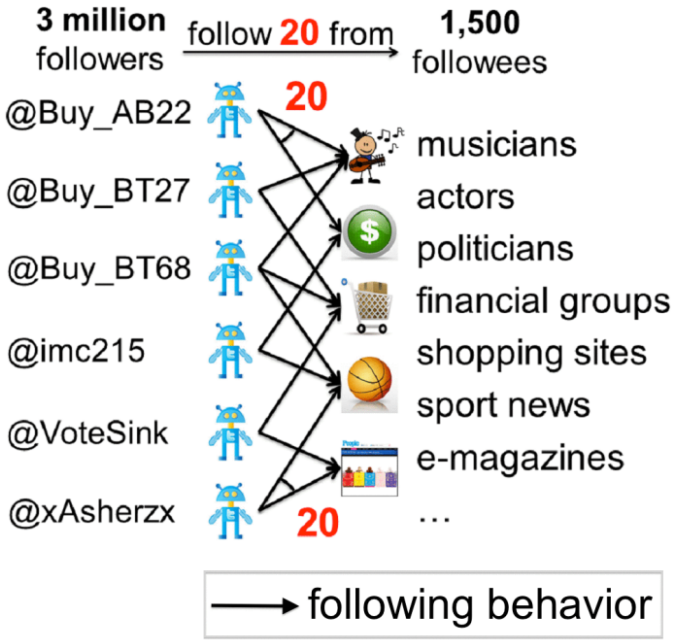
Algorithm: CatchSync

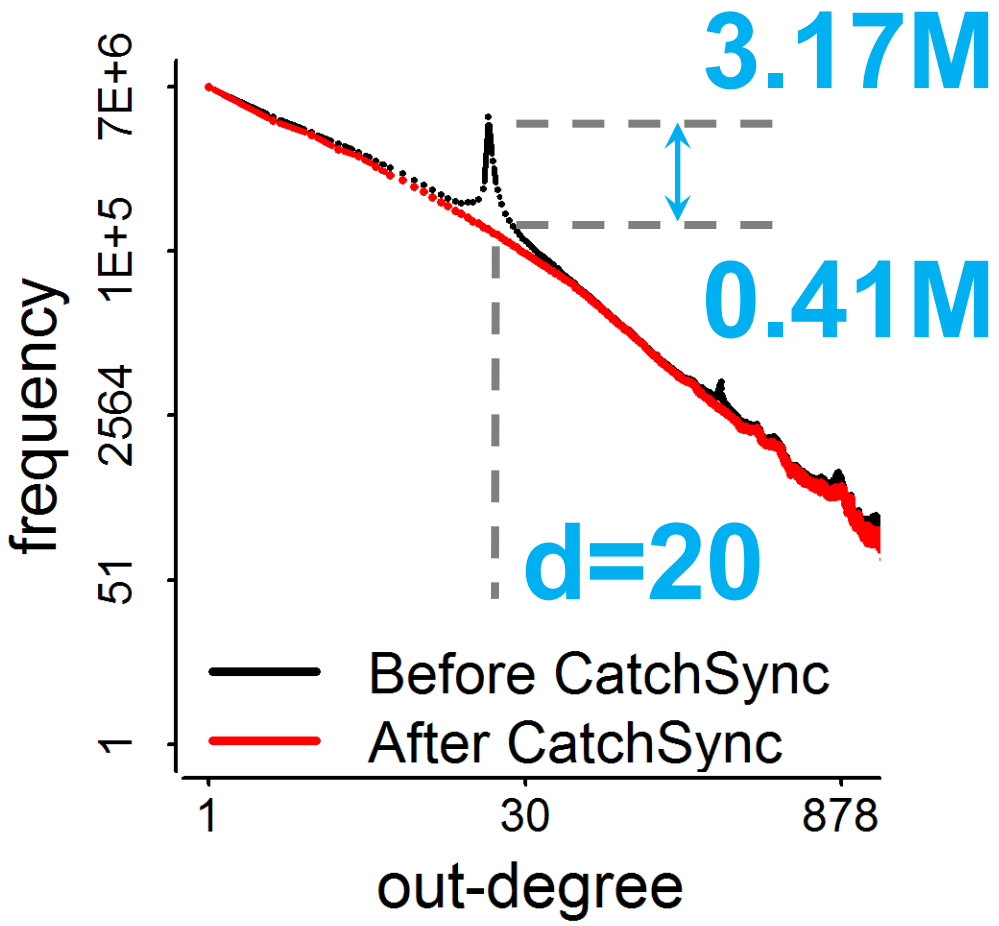
- Synchronicity-Normality: theoretical parabolic lower limit

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



Results

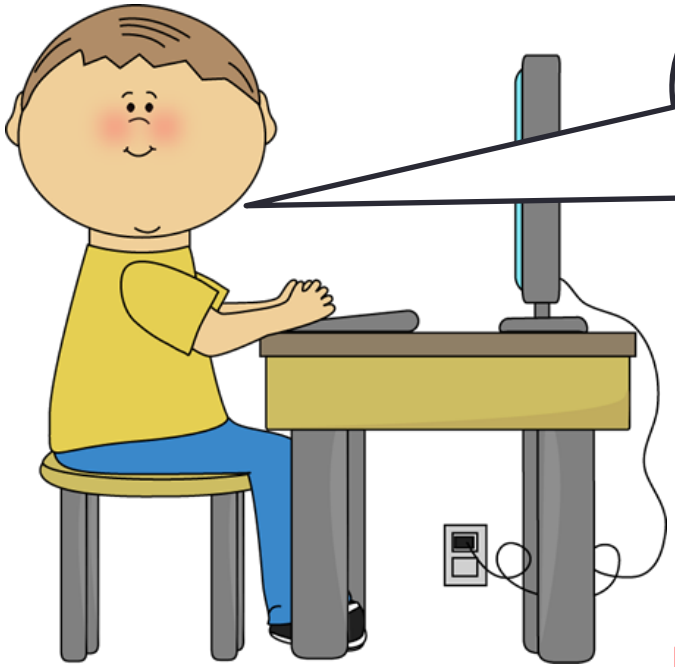

3 million followers
follow 20 from
1,500 followees
@Buy_AB22
@Buy_BT27
@Buy_BT68
@imc215
@VoteSink
@xAsherzx
20
20
musicians
actors
politicians
financial groups
shopping sites
sport news
e-magazines
…
following behavior


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

Roadmap

- Can we spot and catch the suspicious behaviors in a scalable and principled way?
- Behavioral patterns
  - W1. Spotting and catching synchronized behaviors (*KDD'14 Best Paper Finalist*)
  - **W2. Evaluating suspiciousness in multiple dimensions (*ICDM'15, TKDE'16*)**
  - W3. Representing and summarizing dynamic and multi-contextual behaviors (*KDD'16*)

Suppose You Work in Twitter



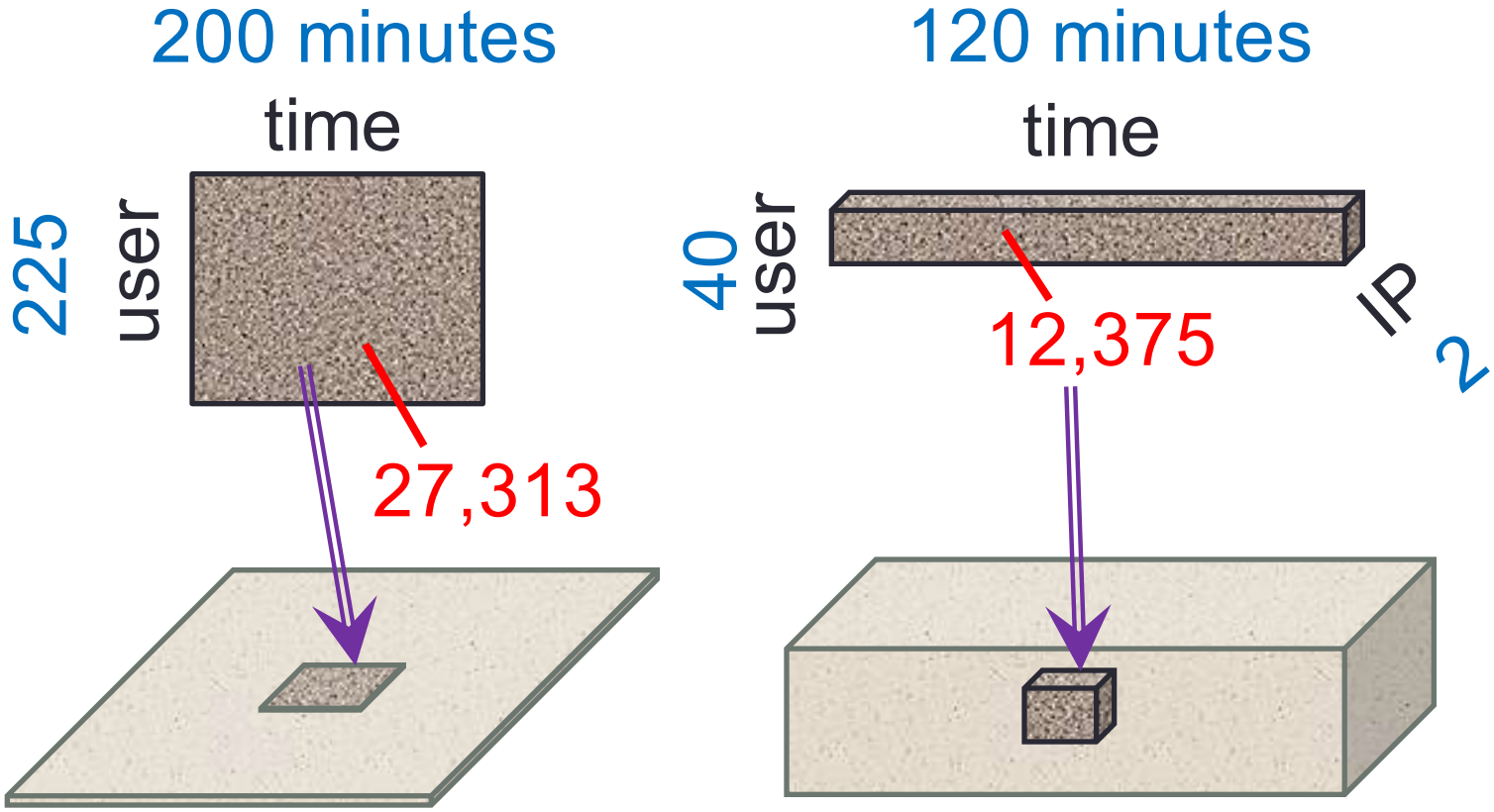
My boss wants me to **catch fraud** in such a big table – **billions of records, tens of columns**!!! How?!

	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	Nurianah	Dec 13, 2013	Baca notif fb .. ada training dadakaann dari evang 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...	

Observation: Multidimensional Data

Dataset	Dimension/Mode (column names)				Mass (#line)
Weibo's Retweeting	User	Root ID	IP	Time (min)	#retweet
	29.5M	19.8M	27.8M	56.9K	211.7M
Weibo's 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

Representation:
Dense Block Indicates Suspiciousness

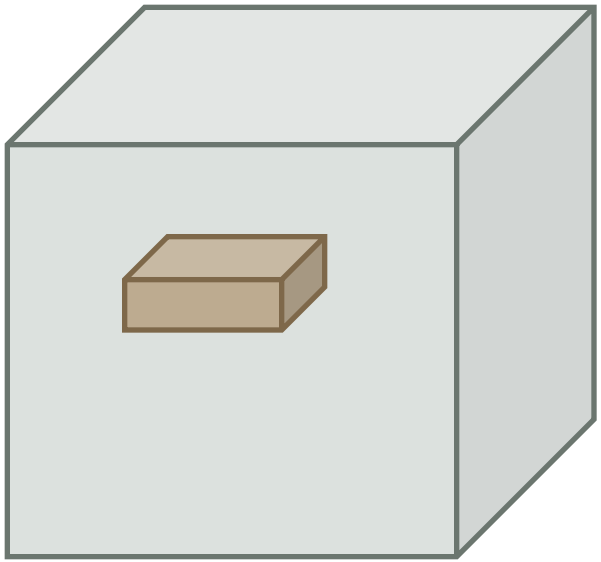


Q: Which is more suspicious?

We need a metric to evaluate the suspiciousness.

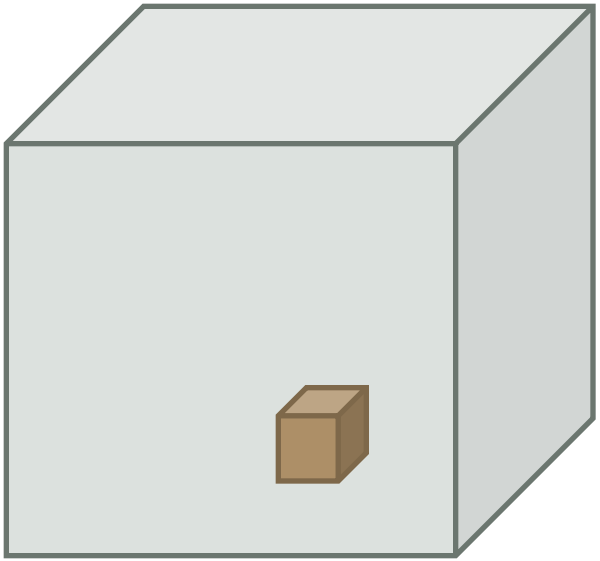
Criteria for Suspiciousness Metric

What properties are required of a good metric?



$N_1 \times N_2 \times N_3$

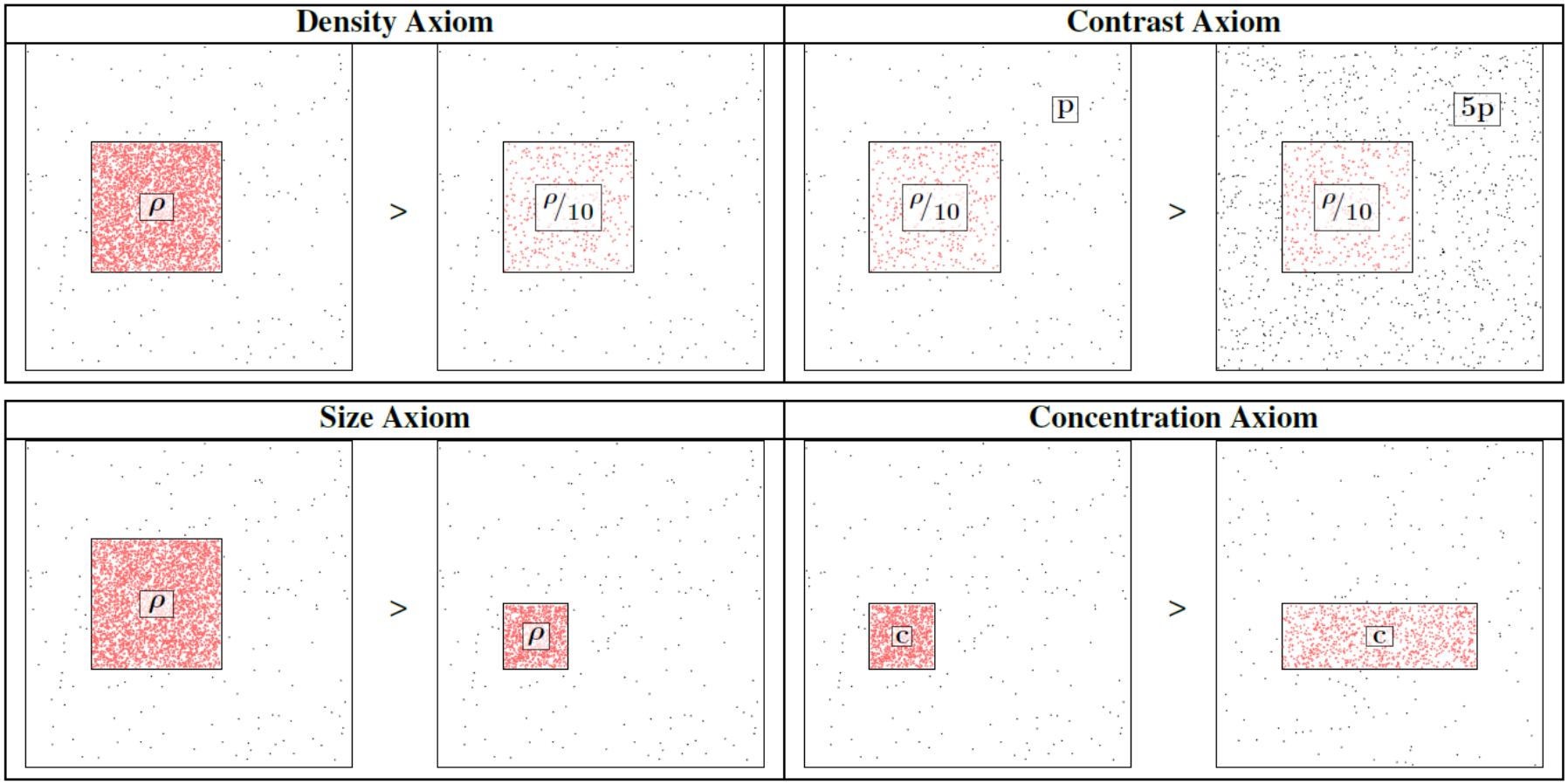
Count data with total "mass" *C*



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

Axiom 1-4

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

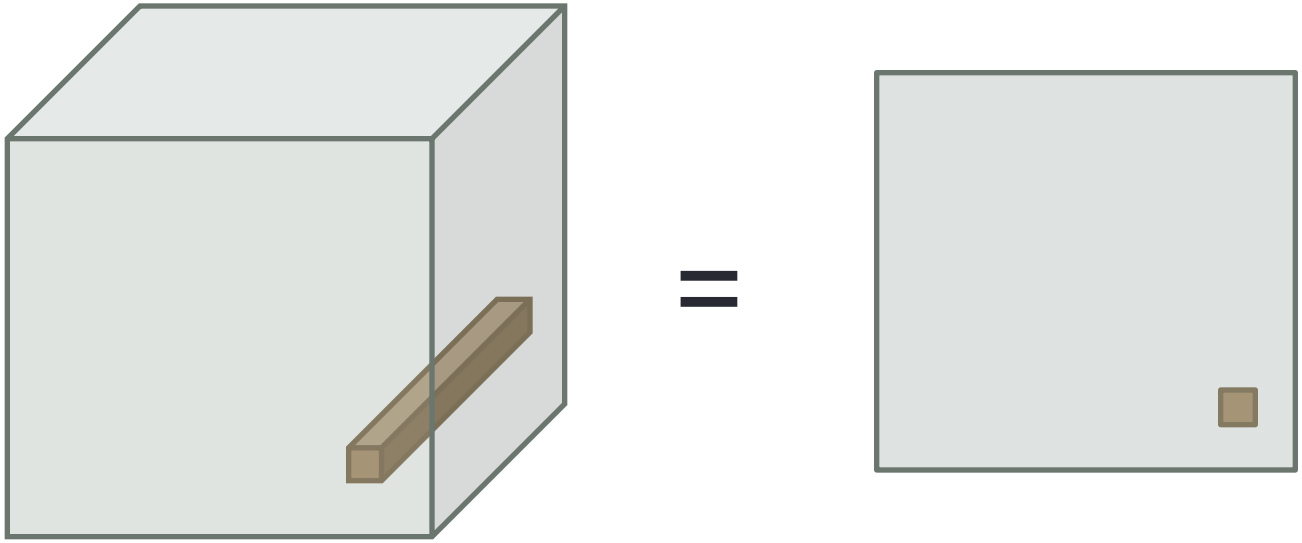


$$p_1 < p_2 \iff \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_1) > \hat{f}(\mathbf{n}, \rho, \mathbf{N}, p_2)$$

Axiom 5: Cross Dimensions

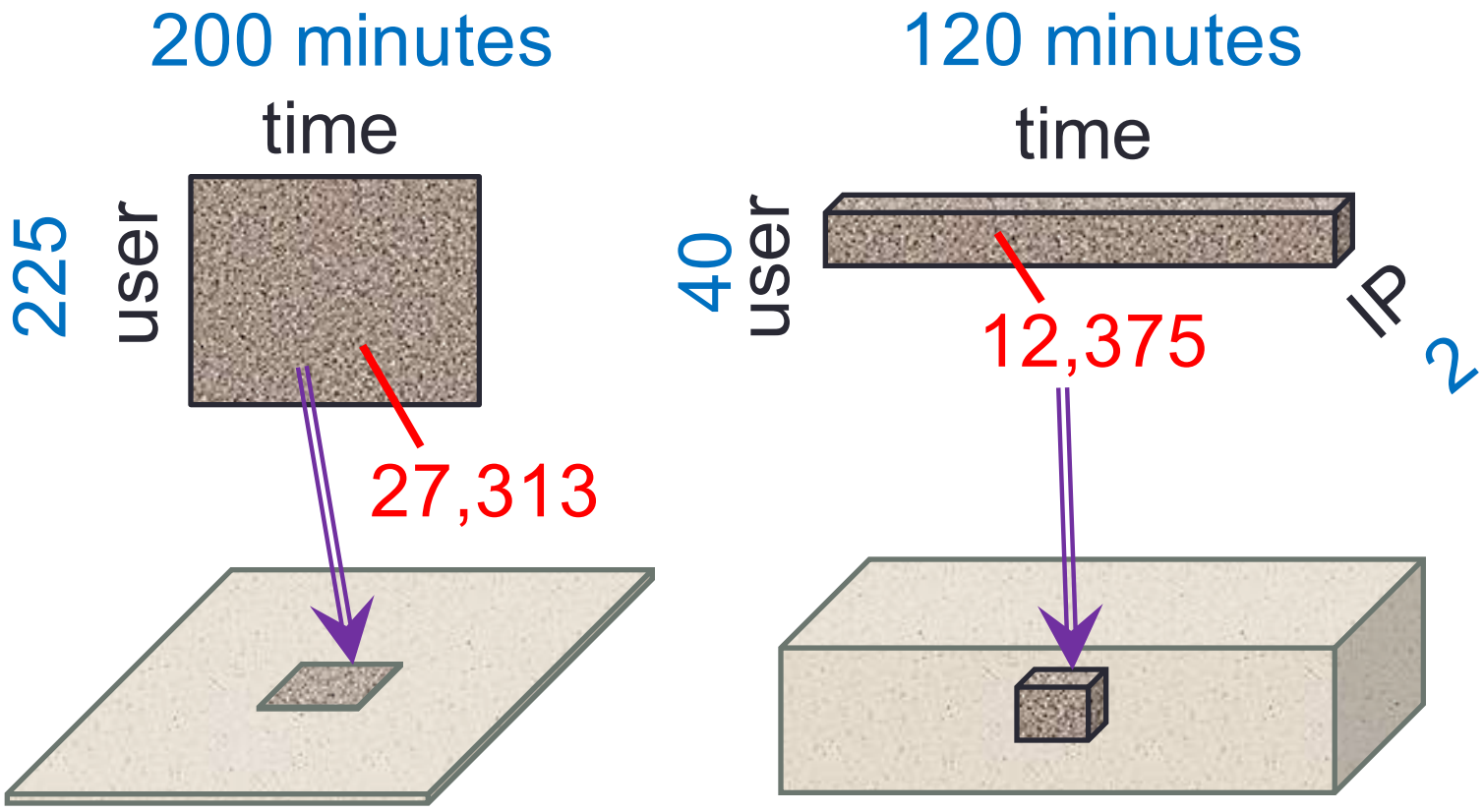
$$f_{K-1}\left([n_k]_{k=1}^{K-1}, c, [N_k]_{k=1}^{K-1}, C\right) = f_K\left(([n_k]_{k=1}^{K-1}, N_K), c, [N_k]_{k=1}^{K}, C\right)$$

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



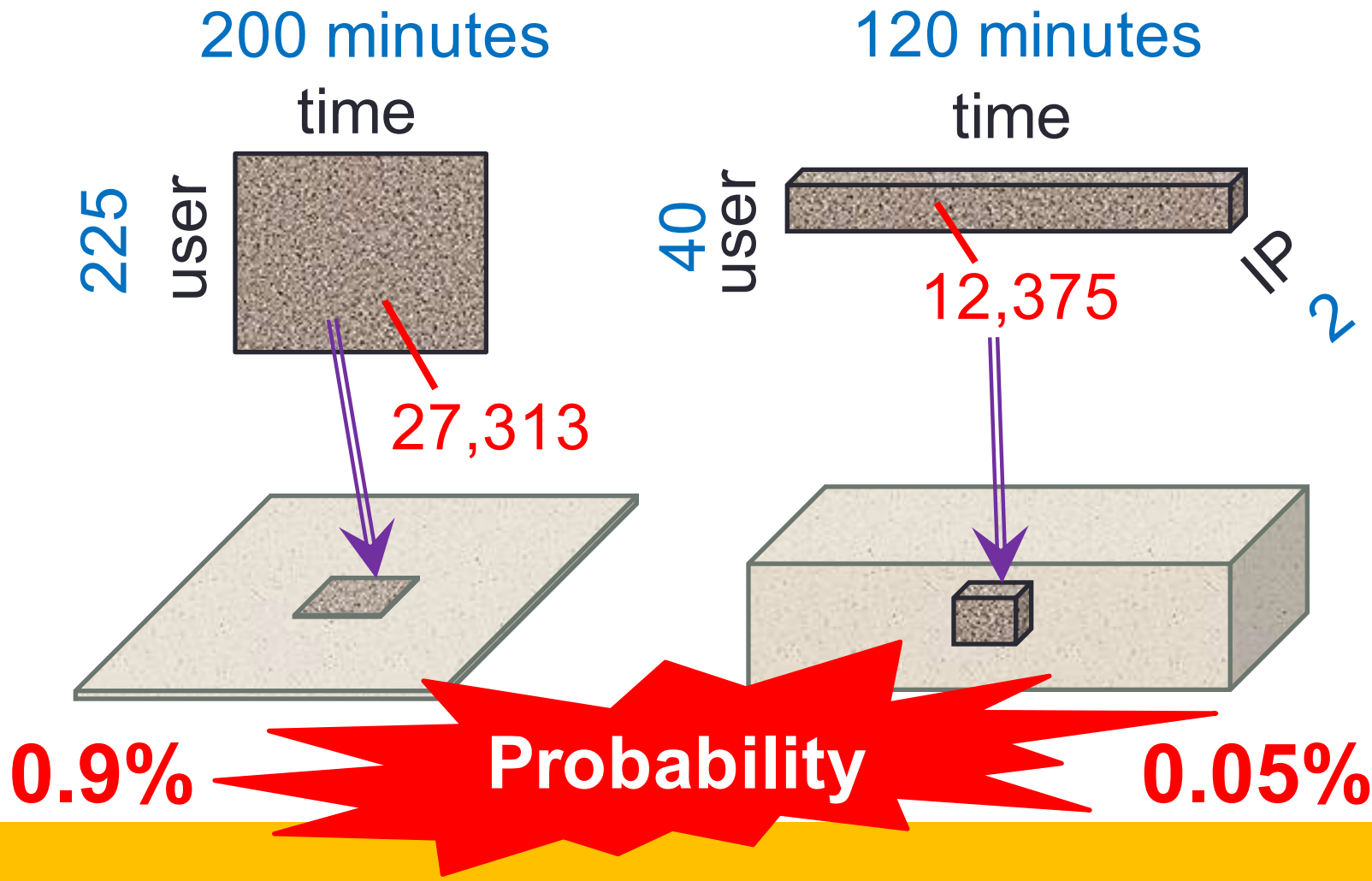
- New information (more modes) can only make our blocks more suspicious

Scoring the Suspiciousness



Q: Which is more suspicious?

Scoring the Suspiciousness



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*

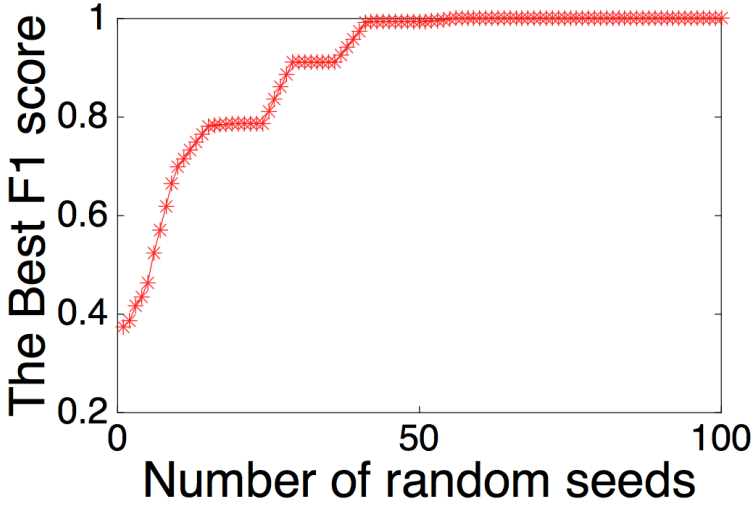
$$f(\mathbf{n}, c, \mathbf{N}, C) = c(\log \frac{c}{C} - 1) + C \prod_{i=1}^{K} \frac{n_i}{N_i} - c \sum_{i=1}^{K} \log \frac{n_i}{N_i}$$

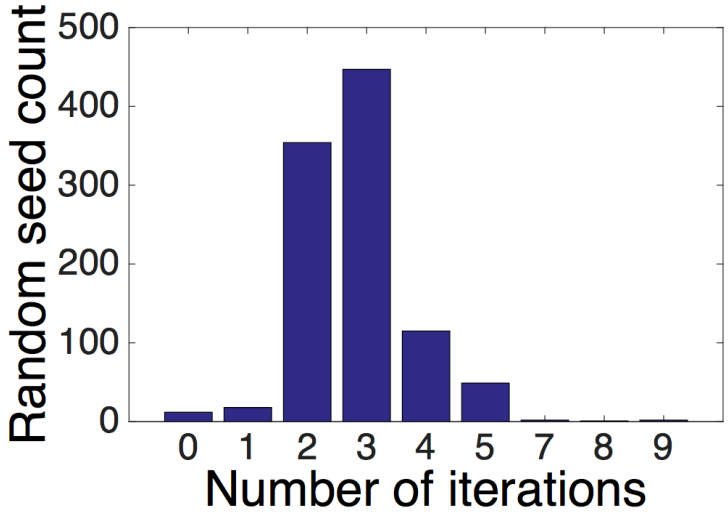
Using ρ as the block's density and p is the data's density, we have the simpler formulation

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

Algorithm: CrossSpot

- Local search to maximize the metric
  - Start with seed blocks
  - Parameter-free: iteratively update the blocks
  - Scalable: parallelize to multiple machines





Advantages

	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				

Results

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!

Results

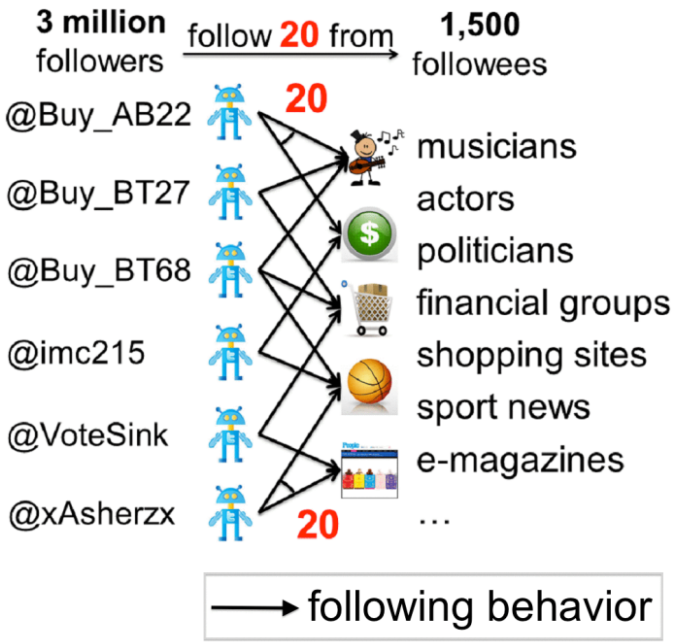
	#	Src-IP × dst-IP × port × second	Mass c	Suspiciousness
CROSSSPOT	1	411 × 9 × 6 × **3,610**	47,449	552,465
	2	533 × 6 × 1 × **3,610**	30,476	400,391
	3	5 × 5 × 2 × **3,610**	18,881	317,529
	4	11 × 7 × 7 × **3,610**	20,382	295,869
HOSVD	1	15 × 1 × 1 × 1,336	4,579	80,585
	2	1 × 2 × 2 × 1,035	1,035	18,308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

Roadmap

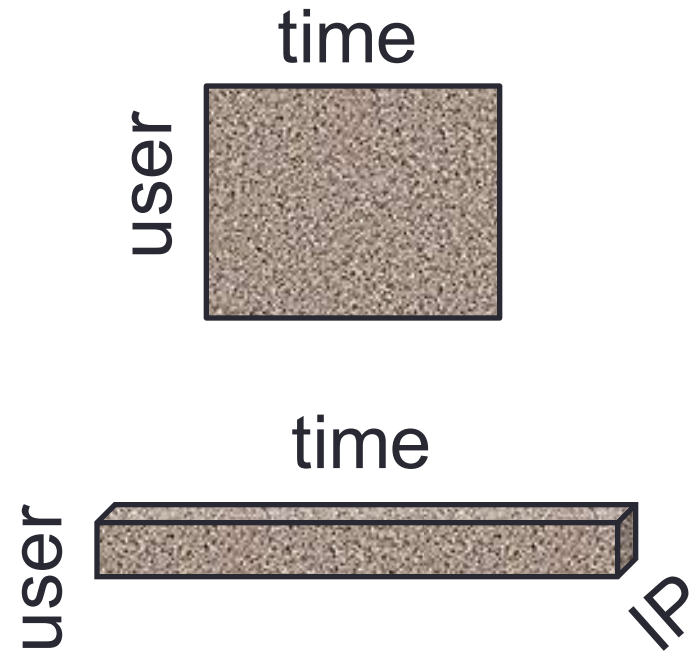
- Can we spot and catch the suspicious behaviors in a scalable and principled way?
- Behavioral patterns
  - W1. Spotting and catching synchronized behaviors (*KDD'14 Best Paper Finalist*)
  - W2. Evaluating suspiciousness in multiple dimensions (*ICDM'15, TKDE'16*)
  - **W3. Representing and summarizing dynamic and multi-contextual behaviors (*KDD'16*)**

Rethink about the Representations

- Graph/Matrix



- Matrix/Tensor



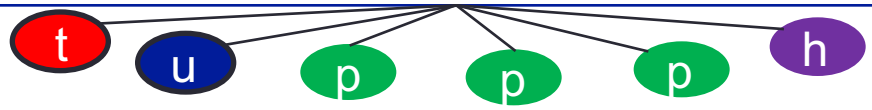
Q: Can they represent every human behavior?

We need to check the basic characteristics of behavior.

Observation: Dynamic and Multi-contextual

- Tweeting behavior

20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl



Contextual factors:

Dynamic — *One-guaranteed value* — *Empty (set of) value* — *Set value* — *Empty (set of) value*

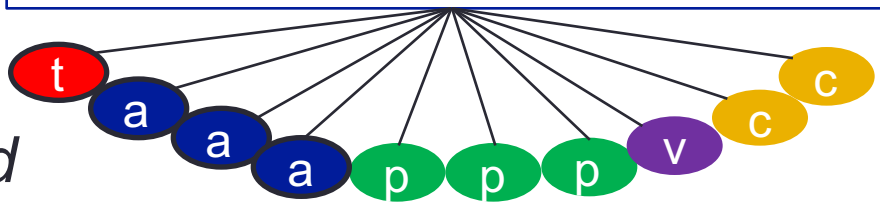
Time slice	User	Location	Phrase	Hashtag	URL
20:00-20:30	@ebekahwsm	∅	{best halftime show, in the history, halftimes shows}	{#SuperBowl}	∅

Observation: Dynamic and Multi-contextual

- Publishing-paper behavior

2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Contextual factors:



Dynamic | *Set value* | *One-guaranteed value* | *Set value* | *Set value*

Time slice	Author	Venue	Keyword	Cited papers
2009	{P. Melville, W. Gryc, R. Lawrence}	SIGKDD	{sentiment analysis, lexical knowledge, text classification}	{p81623, p84395, p95393, p95409, p99073, p116349 …}

Observation: Dynamic and Multi-contextual

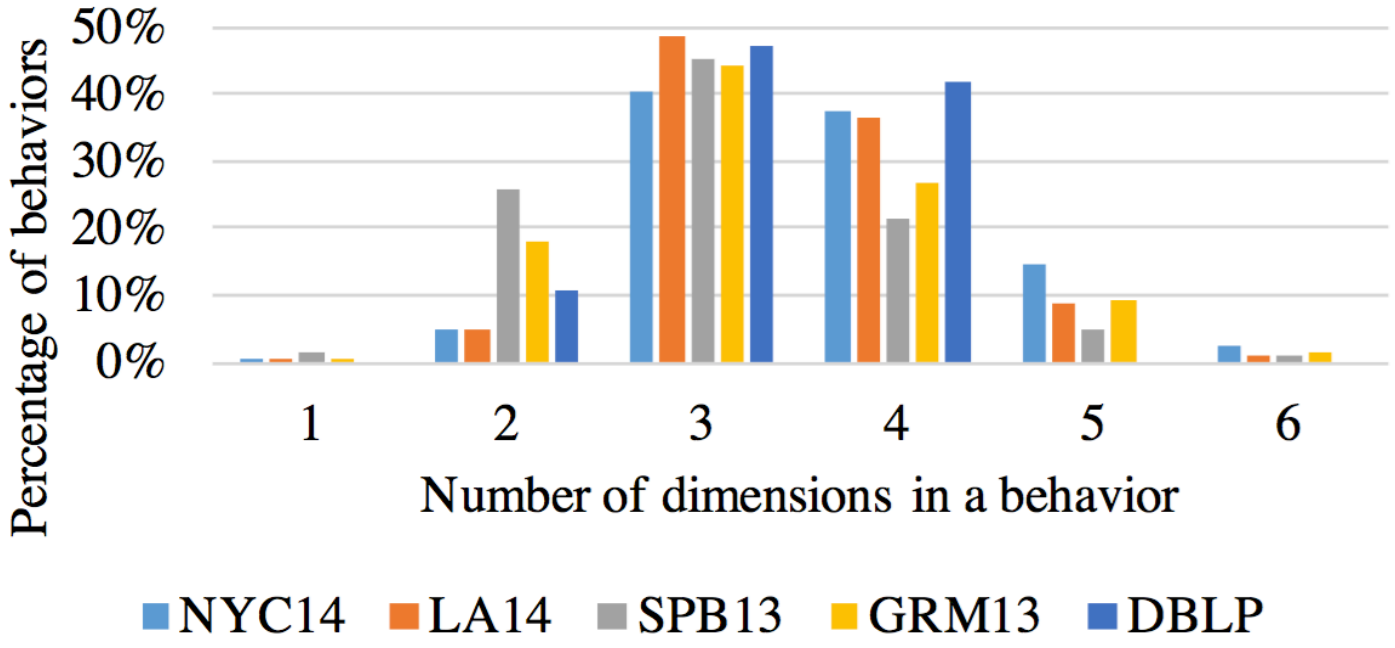
- *Dynamic:* taking a set of consecutive time slices
- *Multi-contextual:* taking a set of dimensions and a set of dimensional values in each dimension

Term	Definition
Dimension	The type of a contextual factor (e.g., location, phrase; author, keyword)
(Dimensional) value	The contextual factor in the dimension
Time slice	The period for consecutive behaviors
Behavior	A set of dimensions, a set of values in each dimension, a time slice for the timestamp

Observation: Dynamic and Multi-contextual

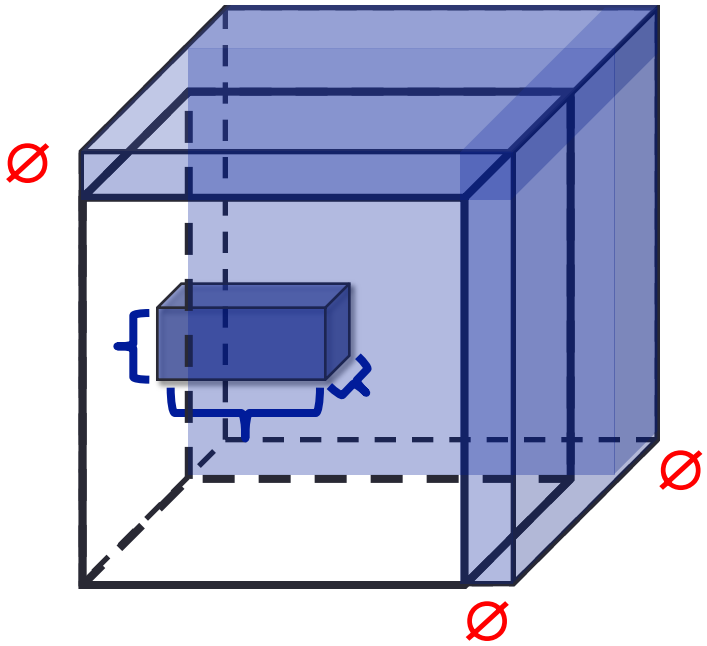
Dataset	#Tweet	#User	#Loc	#Phrase	#Hashtag	#URL	# RT @User	#@User	Time Period
NYC14	10,111,725	329,779	690	1,082,463	587,527	2,766,557	24,439	955,764	113 days
LA14	402,036	14,949	55	257,301	24,711	76,950	795	42,951	113 days
SPB13	2,072,402	1,456,992	9,306	416,461	105,473	140,874	284,647	223,261	25 half-hours
GRM13	2,606,933	1,457,664	5,750	433,548	81,582	334,707	235,097	160,184	52 half-hours

Dataset	#Paper	#Author	#Venue	#Keyword	#Cited paper				Time Period
DBLP	112,157	117,934	55	33,285	62,710				35 years

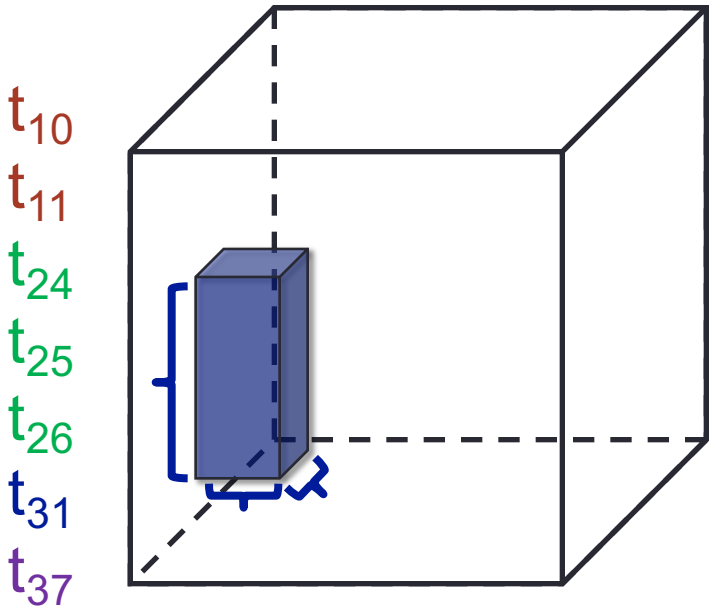


Tensor Fails

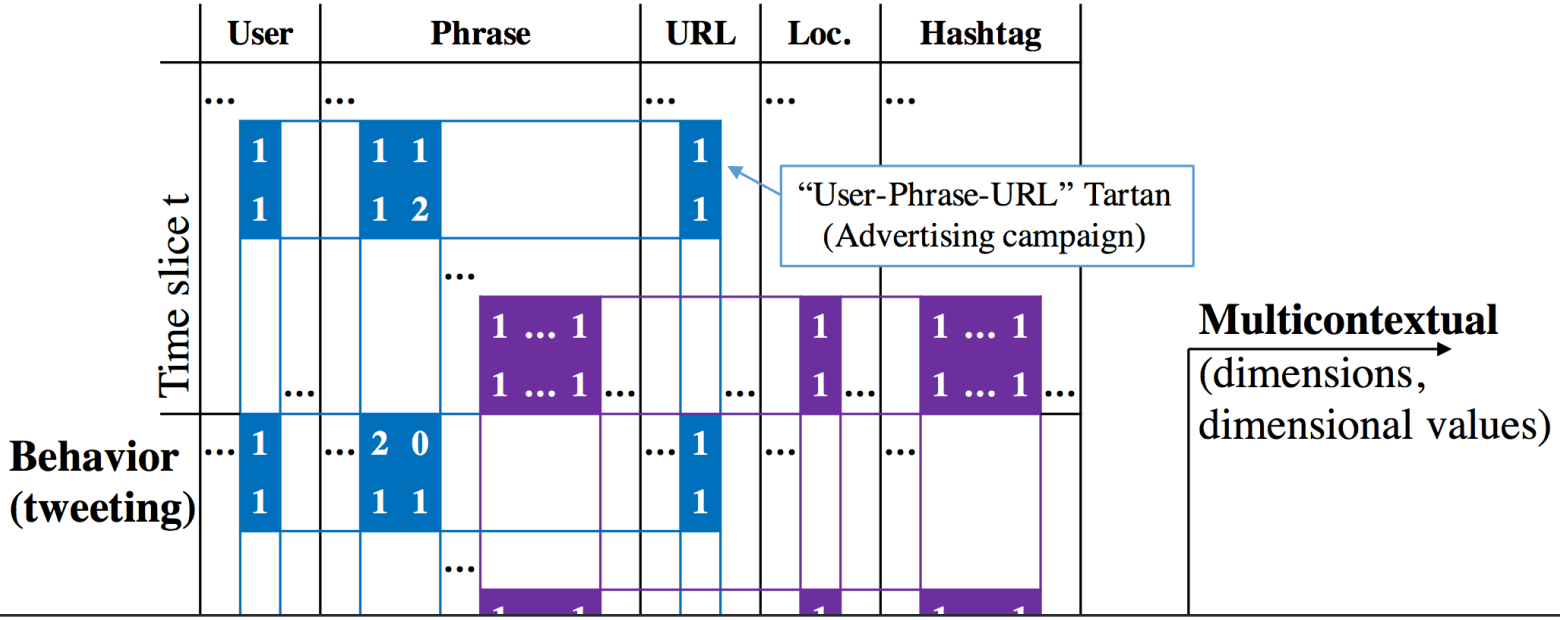
- Multi-contextual
  - Set value
  - Empty value



- Dynamic
  - Temporal values

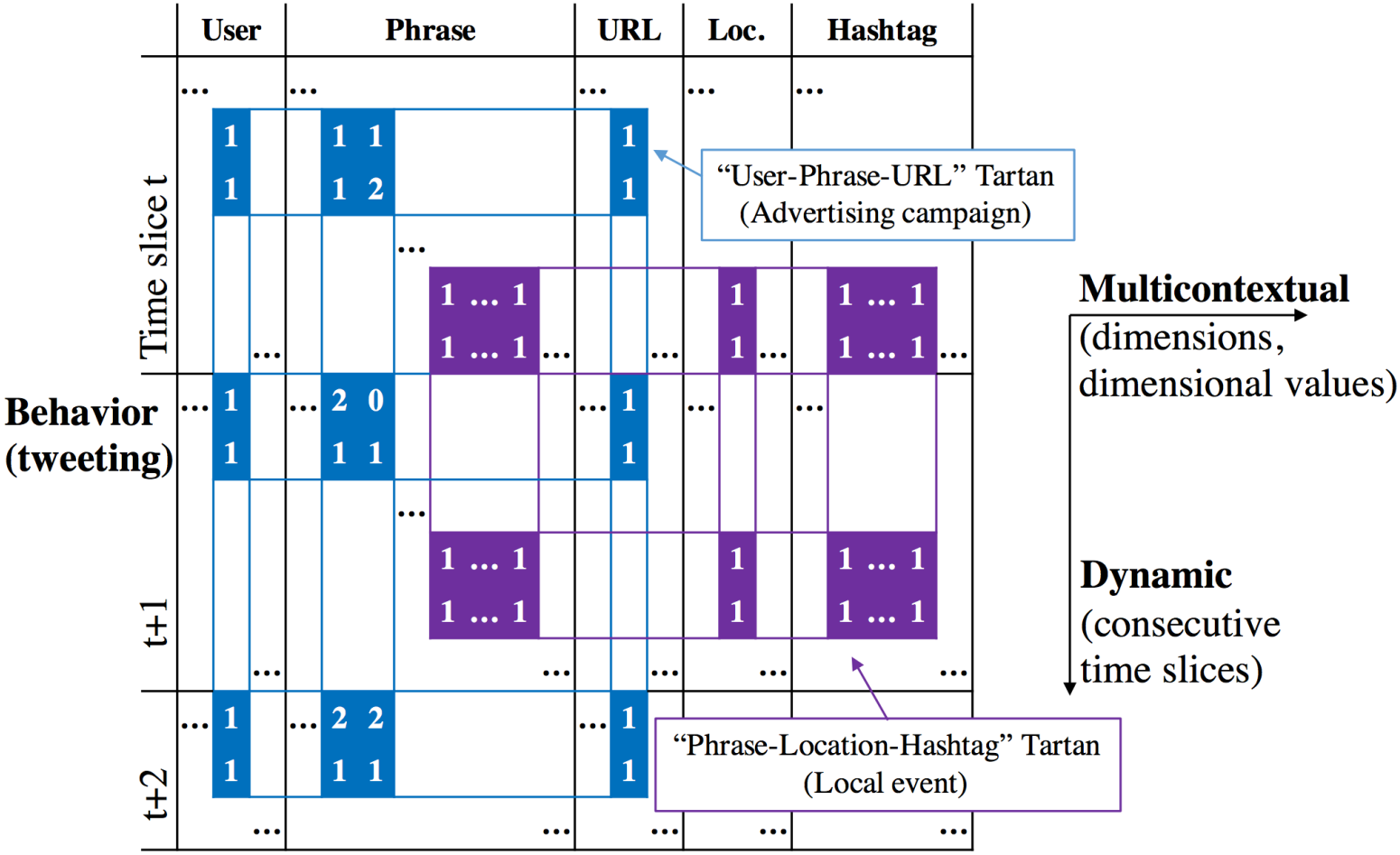


Representation: "Two-Level Matrix"

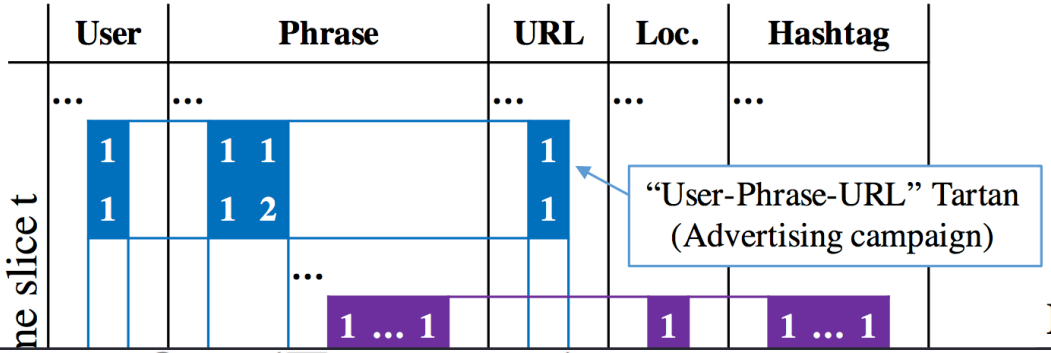


DEFINITION 1 (TWO-LEVEL MATRIX (BEHAVIORAL DATA)). *A two-level matrix $\mathcal{X}$ consists of $\sum_{d=1}^{D} N_d$ columns (dimensional values) and $\sum_{t=1}^{T} E^{(t)}$ rows (behaviors), in which $\mathcal{X}_d^{(t)}(b, i)$ denotes how many times the i-th value in the d-th dimension appears in the b-th behavior at the t-th time slice. The top level consists of D dimensions and T time slices.*

Representation: "Two-Level Matrix"



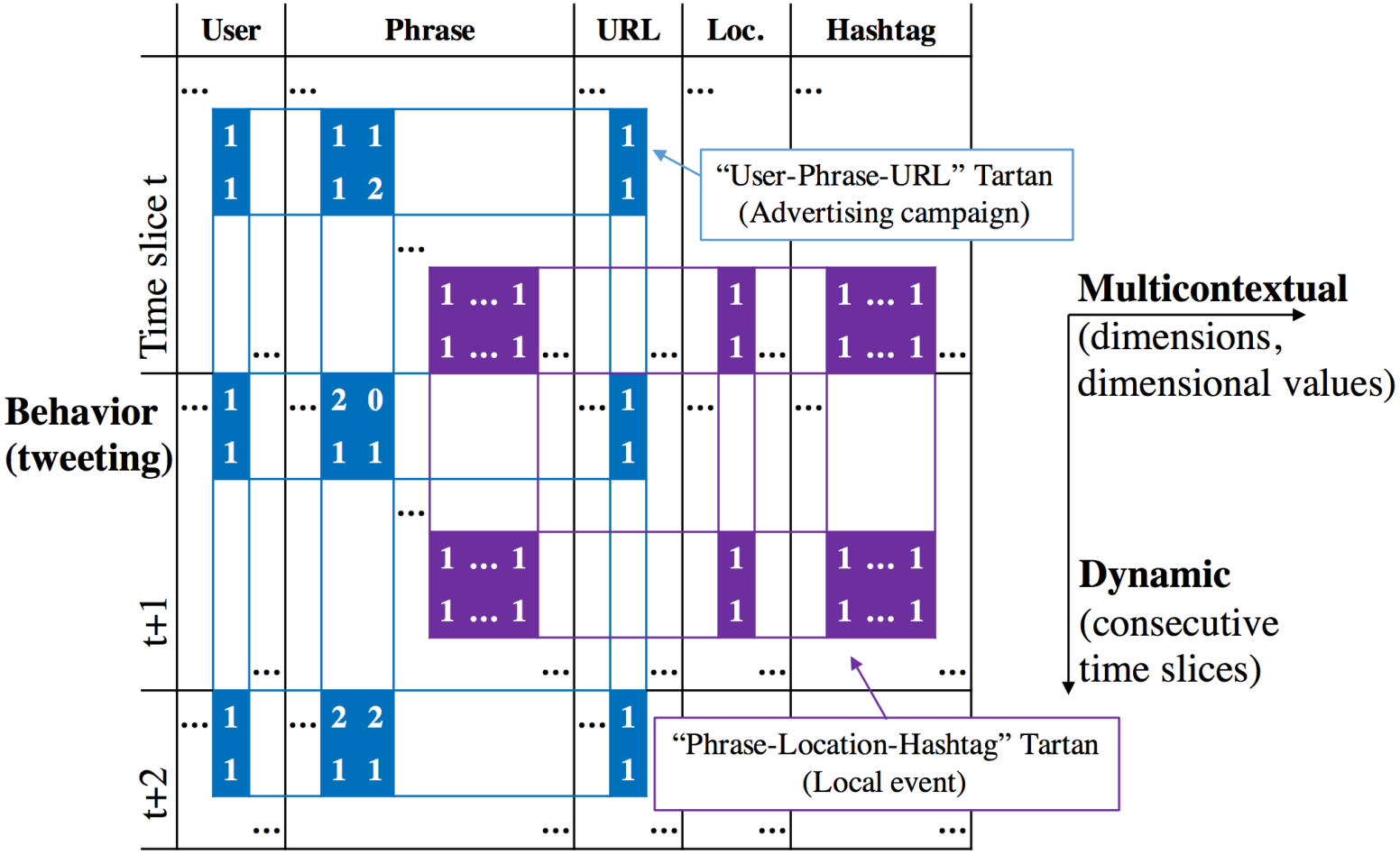
Representation: "Tartan"



DEFINITION 2 (TARTAN (BEHAVIORAL SUMMARY)). *A behavioral summary $\mathcal{A}$ has five components:*

- *a set of dimensions* $\mathcal{D} \subseteq \{1, \ldots, D\}$;
- *a set of values* $\mathcal{V}_d \subseteq \{1, \ldots, N_d\}$ *in the dimension* $d \in \mathcal{D}$;
- *a list of consecutive time slices* $\mathcal{T} = [t_{start}, t_{end}] \subseteq [1, T]$;
- *a set of behavior entries* $\mathcal{B}^{(t)} \subseteq \{1, \ldots, E^{(t)}\}$ *in the time slice* $t \in \mathcal{T}$;
- *the behavior-value entries* $\{\mathcal{X}_d^{(t)}(b, i) | d \in \mathcal{D}, t \in \mathcal{T}, b \in \mathcal{B}^{(t)}, i \in \mathcal{B}^{(t)}\}$

Representation: “Tartan”



The Summarization Problem

- Given the "Two-Level" Matrix, find the Tartans
  - Tartan: sets of meaningful dimensions, values, time slices and behaviors
  - Principle matric function f(Tartan, Data)

PROBLEM 1 (BEHAVIORAL SUMMARIZATION). ***Given*** *the behavioral data (a two-level matrix)* $\mathcal{X} = \{D, N_d|_{d=1}^{D}, T, E^{(t)}|_{t=1}^{T}\}$, ***find*** *a list of behavioral summaries (Tartans)* $\tilde{\mathcal{A}} = \{\ldots, \mathcal{A}, \ldots\}$ *ordered by a principled metric function* $f(\mathcal{A}, \mathcal{X})$ *which defines how well the sets of meaningful dimensions, values, time slices and behaviors are partitioned and how well the meaningful subset of data is summarized, where* $\mathcal{A} = \{\mathcal{D}, \mathcal{V}_d|_{d\in\mathcal{D}}, \mathcal{T}, \mathcal{B}^{(t)}|_{t\in\mathcal{T}}\}$.

Algorithm: CatchTartan

- Employing a lossless encoding scheme based on the *Minimum Description Length* (MDL) principle
- Estimating the number of bits that encoding the Tartan can save from merging the meaningful pattern into the encoding of the data

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}).$$

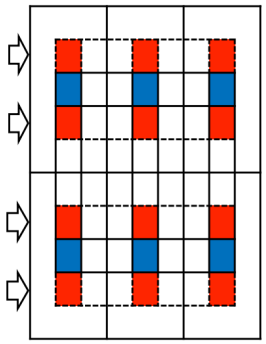
Tartan **Data** **First-level matrix** **Individual entries**

Algorithm: CatchTartan

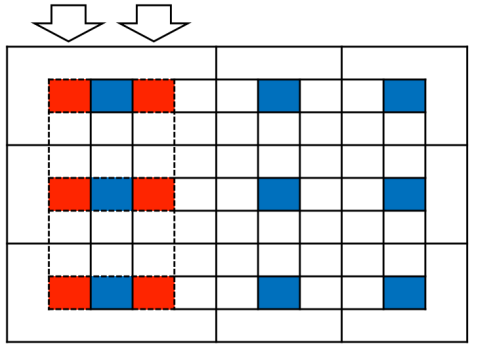
Algorithm 1 CATCHTARTAN : Catching the dynamic multicontextual Tartans for behavioral summaries

Require: the behavioral data $\mathcal{X} = \{D, N_d|_{d=1}^{D}, T, E^{(t)}|_{t=1}^{T}\}$

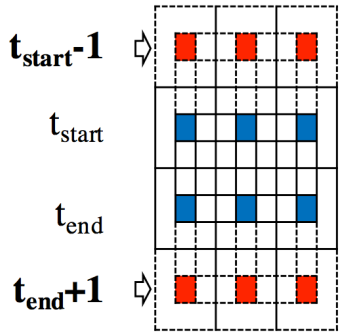
$\tilde{\mathcal{A}} = \{\}$
while the threads run **do**
  generate a seed Tartan $\mathcal{A} = \{\mathcal{D}, \mathcal{V}_d|_{d\in\mathcal{D}}, \mathcal{T}, \mathcal{B}^{(t)}|_{t\in\mathcal{T}}\}$
  while not converged **do**
    for each time slice $t \in \mathcal{T} = [t_{start}, t_{end}]$ **do**
      Update the set of behaviors $\mathcal{B}^{(t)}$ (see Figure 5a) by maximizing the scoring function $f(\mathcal{A}, \mathcal{X})$
    end for
    for each dimension $d \in \mathcal{D}$ **do**
      Update the set of values $\mathcal{V}_d$ (see Figure 5b)
    end for
    Update the consecutive time slices: check if includes the (t_{start}-1)-th and (t_{end}+1)-th slices (see Figure 5c)
    for each dimension $d \notin \mathcal{D}$ **do**
      Check if includes the dimension (see Figure 5d)
    end for
  end while
  $\tilde{\mathcal{A}} \leftarrow \tilde{\mathcal{A}} \cup \mathcal{A}$ sorted in descending order by $f(\mathcal{A}, \mathcal{X})$
end while
return $\tilde{\mathcal{A}}$: the list of Tartans in $\mathcal{X}$



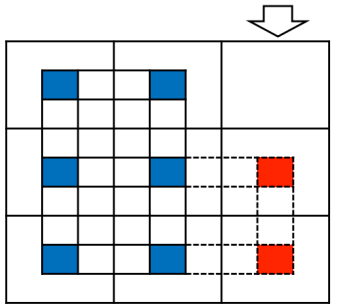
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.

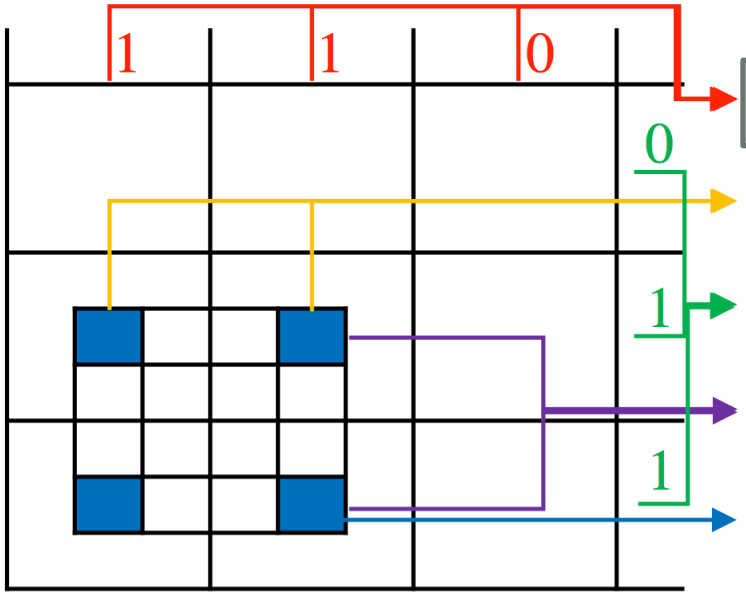


(d) Update the set of dimensions.

Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Encoding the Tartan: Dimensions



Encoding the dimensions

Encoding the dimensional values

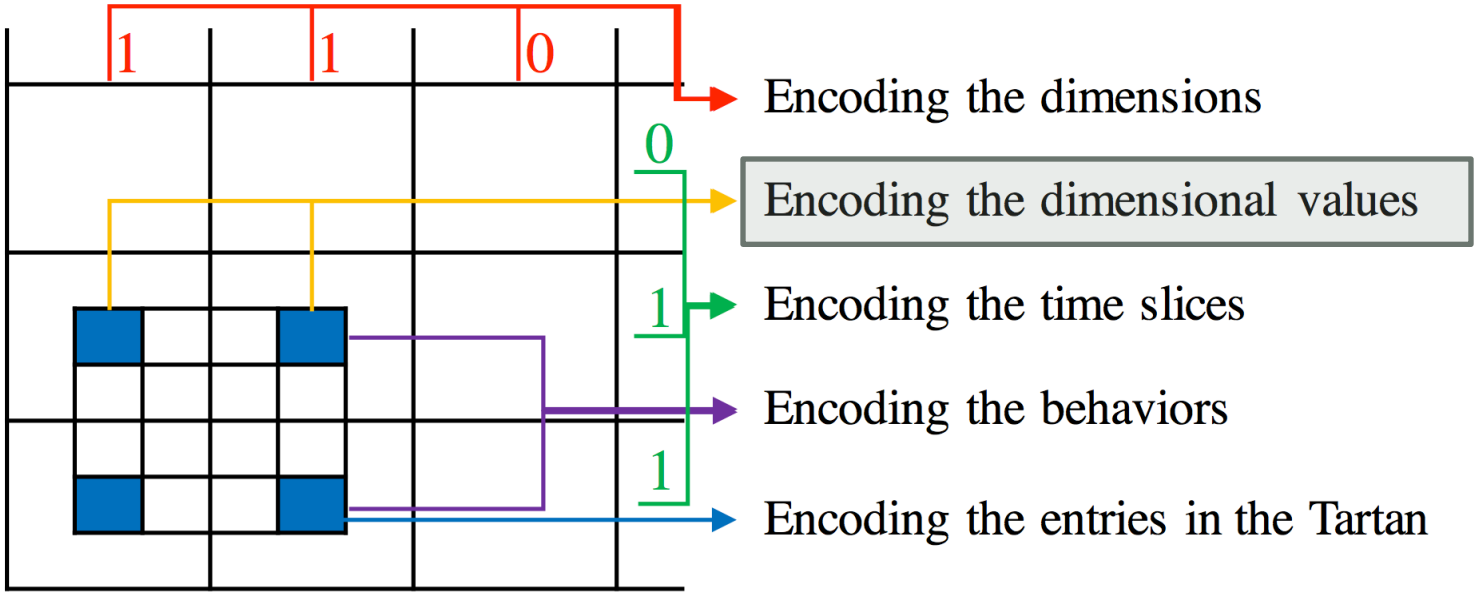
Encoding the time slices

Encoding the behaviors

Encoding the entries in the Tartan

$$
\begin{aligned}
H_{\mathcal{D}}(X) &= -\sum_{x\in\{0,1\}} P(X=x)\log P(X=x) \\
&= -\left(\frac{D^{\mathcal{A}}}{D}\log\frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D}\log\frac{D-D^{\mathcal{A}}}{D}\right). \\
L_{\mathcal{D}}(\mathcal{A}) &= \log^* D + \log^* D^{\mathcal{A}} + D\cdot H_D(X) \\
&= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),
\end{aligned}
$$

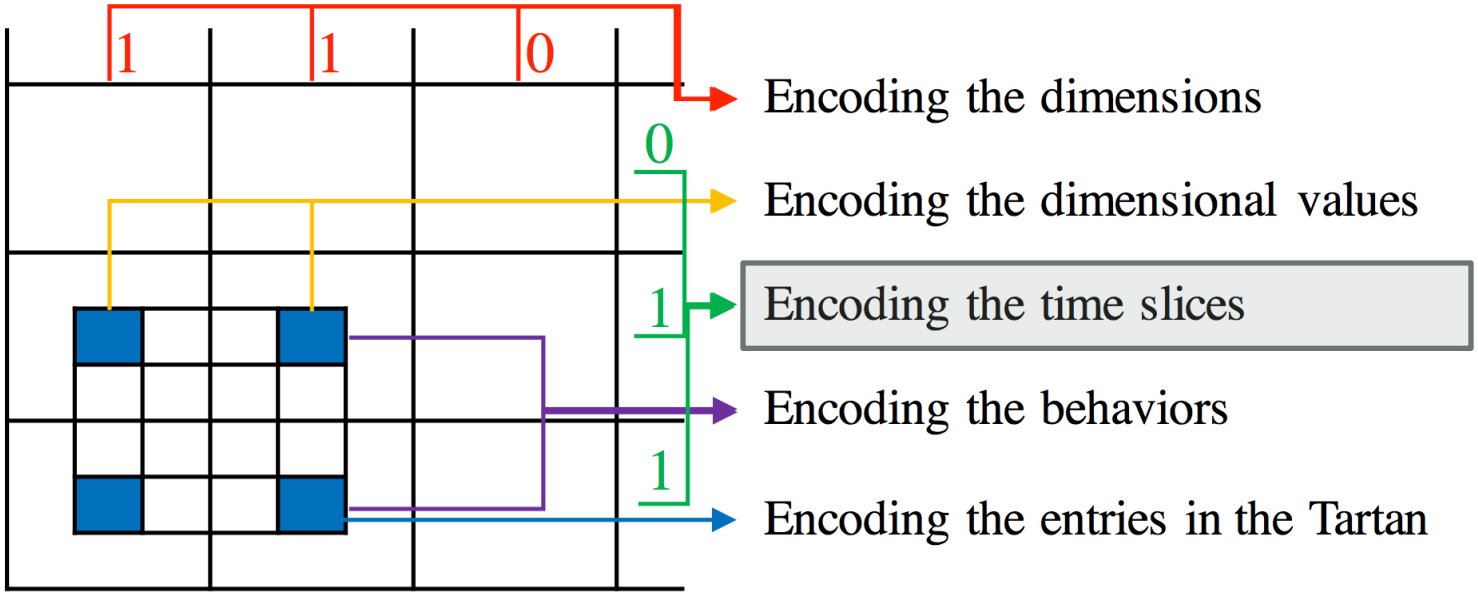
Encoding the Tartan: Dimensional Values



$$H_{\mathcal{V}_d}(X) = -\left(\frac{n_d}{N_d}\log\frac{n_d}{N_d} + \frac{N_d - n_d}{N_d}\log\frac{N_d - n_d}{N_d}\right).$$

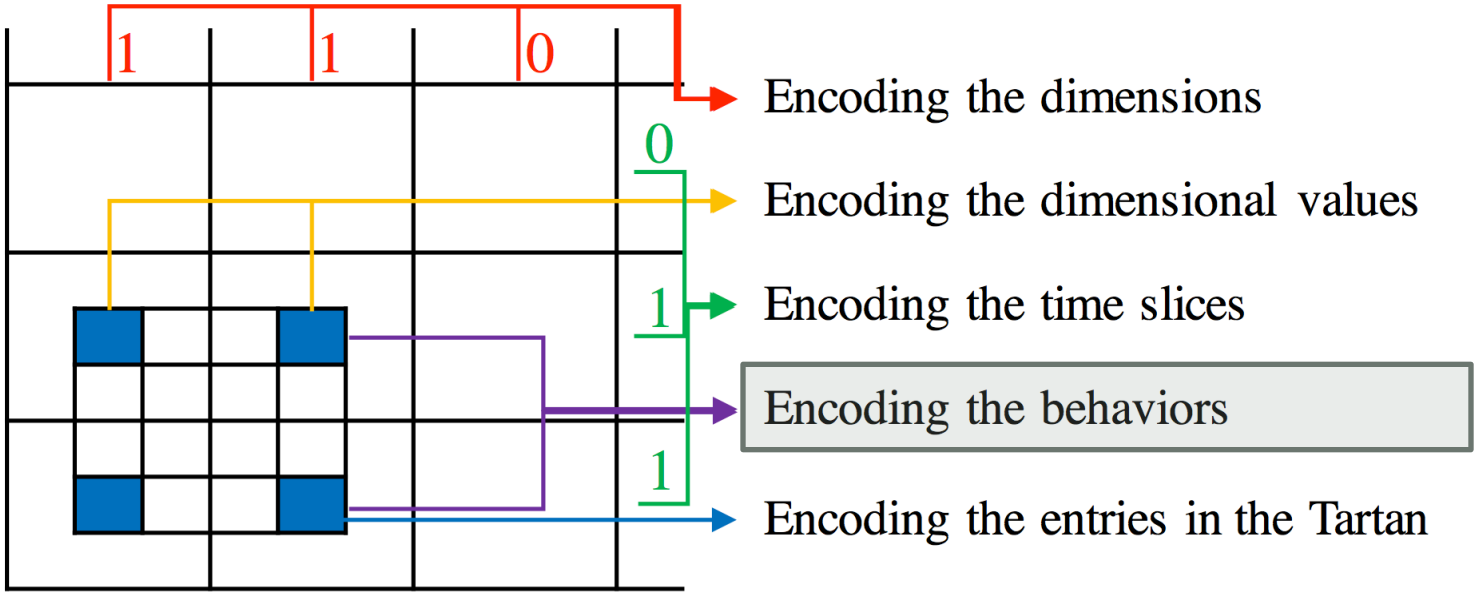
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding the Tartan: Time Slices



$$L_{\mathcal{T}}(\mathcal{A}) \quad = \quad \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

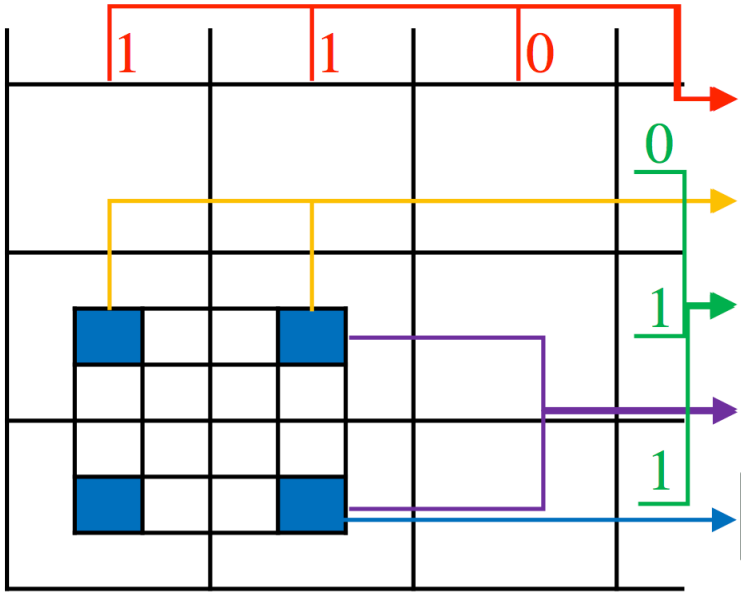
Encoding the Tartan: Behaviors



$$H_{\mathcal{B}^{(t)}}(X) = -\left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)}-e^{(t)}}{E^{(t)}} \log \frac{E^{(t)}-e^{(t)}}{E^{(t)}}\right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t\in\mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)})\right).$$

Encoding the Tartan: Entries



$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

$$c = \sum_{d\in\mathcal{D},t\in\mathcal{T}} \sum_{b\in\mathcal{B}^{(t)},i\in\mathcal{V}_d} \mathcal{X}_d^{(t)}(b,i).$$

$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v+c)H_{\mathcal{A}}(X) = g(v+c,c).$$

The Objective Function

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}).$$

$$\begin{aligned} V &= \left(\sum_{d \in \mathcal{D}} N_d\right)\left(\sum_{t \in \mathcal{T}} E^{(t)}\right). \\ C &= \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \{1,\ldots,E^{(t)}\}, i \in \{1,\ldots,N_d\}} \mathcal{X}_d^{(t)}(b,i). \end{aligned}$$

$$\begin{aligned} L(\mathcal{X}^{\mathcal{A}}) &= g(V + C, C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) \\ &\quad + \sum_{d \in \mathcal{D}} \log^* N_d + \sum_{t \in \mathcal{T}} \log^* E^{(t)}. \end{aligned}$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{\mathcal{A}}(\mathcal{A}).$$

$$L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}) = g(V + C - v - c, C - c);$$

Results: DBLP

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Results: Super Bowl 2013

16:30 – 17:00 — Tartan #1: (1 dim) 16:30-17:30

16:30:31 My prediction Ravens 34 Niners 31
16:30:57 Ready for the big game :D, my prediction 24-20 SF #SuperBowl
16:31:14 My prediction for superbowl.. 48.. Jets over Bears 17-13 Mark Sanchez MVP
16:32:24 I predict Baltimore Ravens will win 27 to 24 or 25 or 26. Basically it will be a close game.

"my prediction"

user	phrase	hashtag	URL	3,397 tweets
(3,325)	226	(0)	(0)	

17:30 – 18:00 — Tartan #2: (3 dims) 17:00-18:00

17:30:51 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:01 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:16 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:19 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist

"make your prediction"

user	phrase	RT @user	URL	196 tweets
(196)	4	1	1	

18:00 – 18:30 — Tartan #3: (2 dims) 18:30-19:30

18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.

"7-3", "1st Qtr"

user	phrase	RT @user	URL	215 tweets
(213)	21	3	(0)	

19:00 – 19:30 — Tartan #4: (3 dims) 20:00-21:00

20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs
20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs

"halftime show"

user	phrase	RT @user	URL	617 tweets
(617)	11	4	4	

20:00 – 20:30 — Tartan #5: (3 dims) 20:00-21:00

20:20:47 (Manhattan, NY)...and every one of those girls took #ballet #Beyonce #superbowl
20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl
20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. EVER. #Beyonce #superbowl #halftimeshow

"beyonce", #beyonce, #superbowl, #DestinysChild

location	phrase	hashtag	URL	166 tweets
2	55	17	(0)	

21:00 – 21:30 — Tartan #6: (2 dims) 21:00-22:00

21:44:42 Ahora si pff #49ers 23-28 #Ravens
21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers
21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBolwXLVII #SuperBowl #NFL

"28-23", #49ers, #Ravens

user	phrase	hashtag	URL	653 tweets
(650)	69	11	(0)	

22:00 — Tartan #7: (1 dim) 22:00-23:30

22:42:27 Congratulations Ravens!!!!
22:42:43 Congratulations Ray Lewis and the Ravens.
22:42:43 Game over.! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep .!
22:42:52 "@LetThatBoyTweet: Game over. Ravens win the Super Bowl."

"congratulations", "game over"

user	phrase	hashtag	URL	1,950 tweets
(1942)	248	(0)	(0)	

Results: NYC 2014

Aug 15 Sep 1 Sep 15 Oct 1 Oct 15 Nov 1 Nov 15

User	Location	Phrase	Hashtag	#Tweet
18 @queen_toni_ @nachoiall	**9** Bronx, NY Staten Island, NY	**42** “have pizza”, “eat candy”…	**19** #vote2sos, #bsmg, #coast2coast…	**1,734** Aug 16- Aug 19

2014-08-16 04:35:17 queen_toni_: AND WE CAN HAVE PIZZA BLAST MUSIC EAT CANDY AND WATCH MOVIES @Michael5SOS #vote5sos http://t.co/qZGv93B82P
2014-08-16 04:35:33 queen_toni_: AND WE CAN HAVE PIZZA BLAST MUSIC EAT CANDY AND WATCH MOVIES @Michael5SOS #vote5sos http://t.co/mrhy1dckG8
2014-08-16 04:36:10 queen_toni_: AND WE CAN HAVE PIZZA BLAST MUSIC EAT CANDY AND WATCH MOVIES @Michael5SOS #vote5sos http://t.co/VDIEGzPZhD

User	Location	Phrase	#Tweet
22 @DateMeJackDail @picomarlon_pico	**8** Rahway, NJ Manhattan, NY	**26** “I love you”, “new york”, “thank you”, “dream come true”…	**1,632** Oct 1- Oct 4

2014-10-02 03:56:06 JossethHenry:
Thank You, Lord, for every blessing and favor in my life - God never forgets - God is not a human being that He should change His mind
2014-10-02 09:26:35 picomarlon_pico: @ShopOnThePorch thank you
2014-10-02 12:35:39 omgimsoawesome: I love you baby @primetime_joe

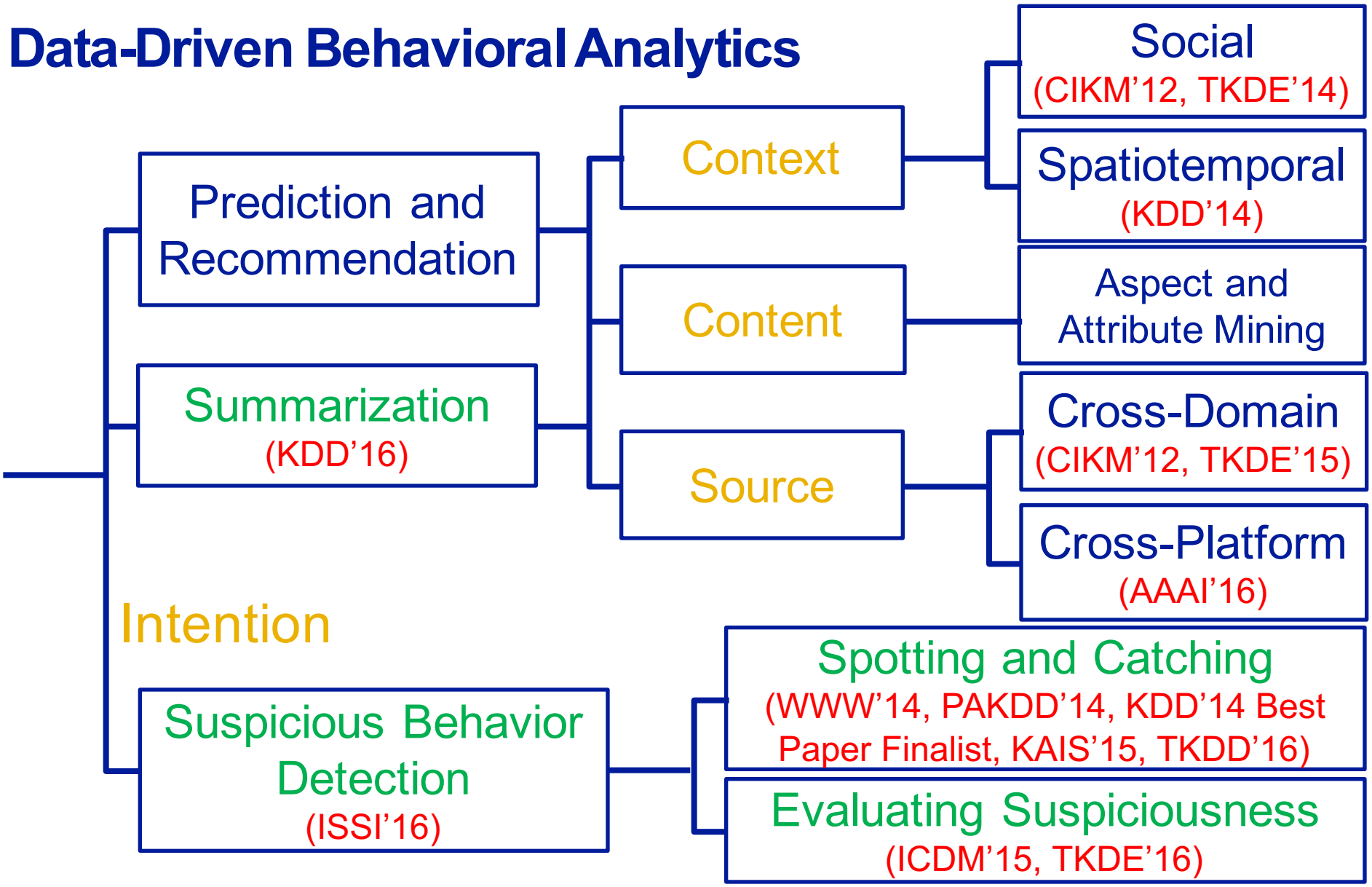
User	Location	Phrase	@User	#Tweet
2 @bryanallantkl @DaniellaGates	**3** Elmwood Park, NJ Lyndhurst, NJ Belleville, NJ	**11** “play bass in tklband new single” “what’s up” “preorder our ep for only”	**3** @tklband @xbrooke_alexis @ashton5sos	**1,585** Nov 14- Nov 18

2014-11-14 17:04:18 bryanallantkl:
@jjnnni What‘s up!???? I play bass in @TKLband NEW SINGLE! http://t.co/Rrn6bTqHVJ preorder our EP for only $3.99! https://t.co/2YiGs7XILe
2014-11-14 17:04:36 bryanallantkl:
@officiallizely What's up!????I play bass in @TKLband NEW SINGLE! http://t.co/PIITd39LM9 preorder our EP for only $3.99! https://t.co/M9EeHdR6P2

User	Location	Phrase	Hashtag	URL	#Tweet
2 @KTPJobs @TFATechRecruit	**1** Manhattan, NY	**4** “Kaplan Test Prep” “Teach for America”…	**16** #Job, #NewYork, #TweetMyJob …	**10** http://t.co/Wt1 xUbqheR…	**80** Nov 16- Nov 23

2014-11-16 06:41:22 KTPJobs:
Kaplan Test Prep: Director of Employee Relations (#NewYork, NY) http://t.co/Wt1xUbqheR #HR #Job #Jobs #TweetMyJobs
2014-11-16 07:06:52 TFATechRecruit:
#OpenSource #Job alert: Director, JavaScript Front End Developer Teach For America #NewYork, NY http://t.co/5UDjsb0Bdk #Jobs
2014-11-16 08:17:09 TFATechRecruit:
Teach For America #IT #Job: Assistant, Information Technology (#NewYork, NY) http://t.co/l2MCuCpeU9 #Jobs #TweetMyJobs

Data-Driven Behavioral Analytics



References

[1] Meng Jiang, Peng Cui, Christos Faloutsos. “Suspicious Behavior Detection: Current Trends and Future Directions”, IEEE Intelligent Systems (ISSI) 2016. (Survey paper, IF=2.34)

[2] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. “CatchSync: Catching Synchronized Behavior in Large Directed Graphs”, ACM SIGKDD 2014 Best Paper Finalist.

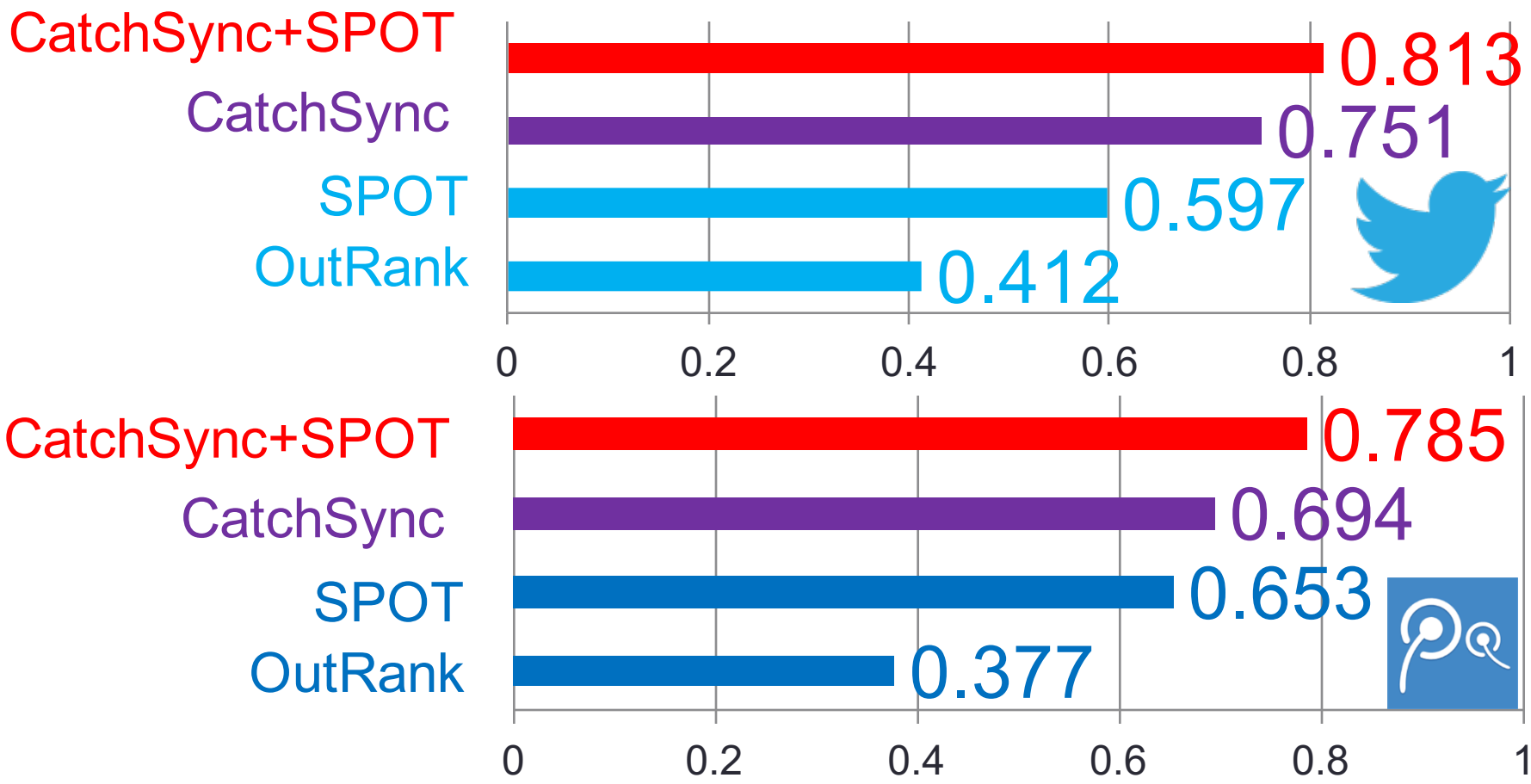
[3] Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, Christos Faloutsos. “A General Suspiciousness Metric for Dense Blocks in Multimodal Data”, IEEE ICDM 2015.

[4] Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, Christos Faloutsos. “Spotting Suspicious Behaviors in Multimodal Data: A General Metric and Algorithms”, IEEE TKDE 2016. (IF=2.07)

[5] Meng Jiang, Christos Faloutsos, Jiawei Han. “CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors”, ACM SIGKDD 2016. (Oral presentation, ACC=8.9%)

THANK YOU!

Results (CatchSync)

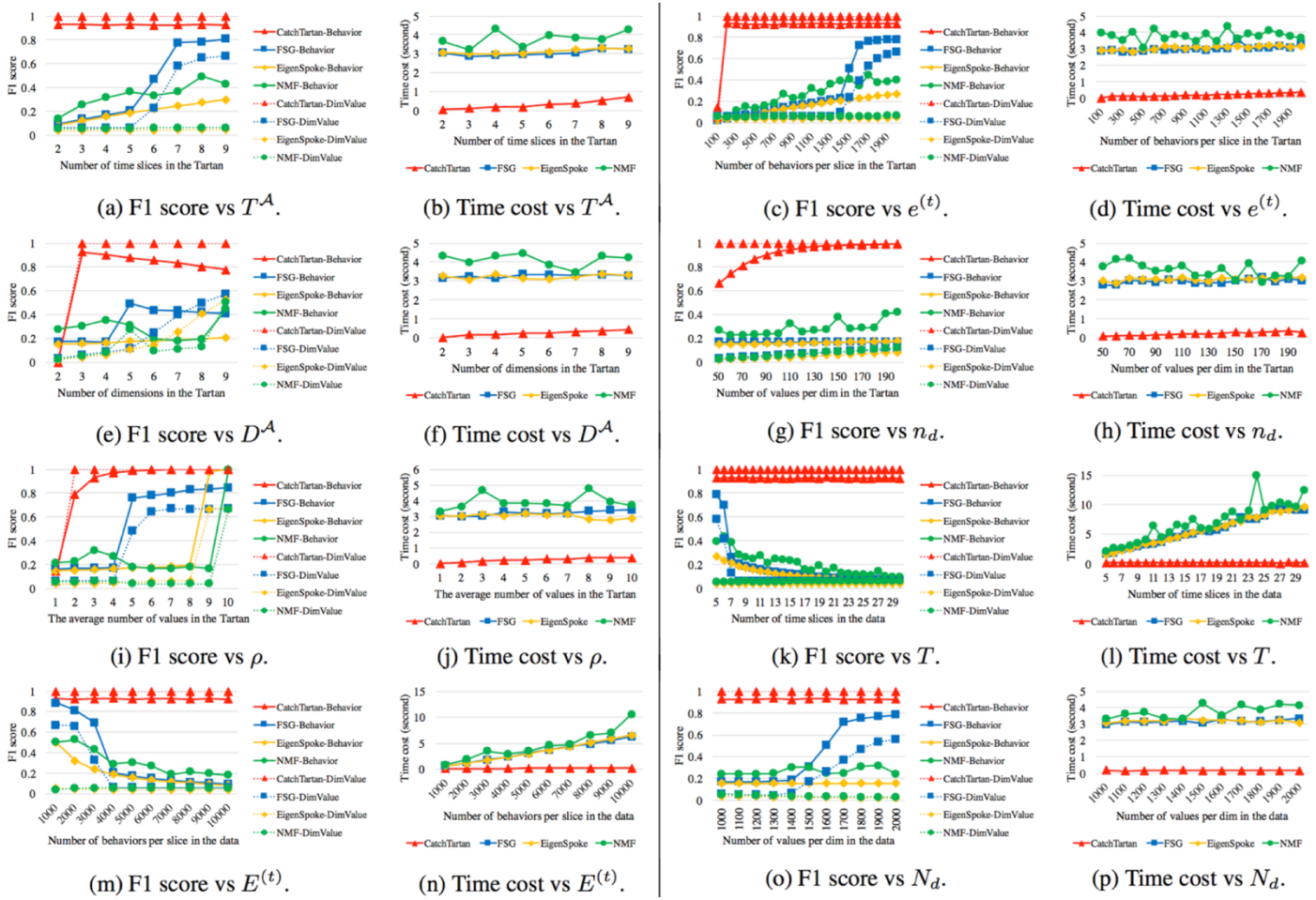


Results (CrossSpot)

- 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**

Results (CatchTartan)



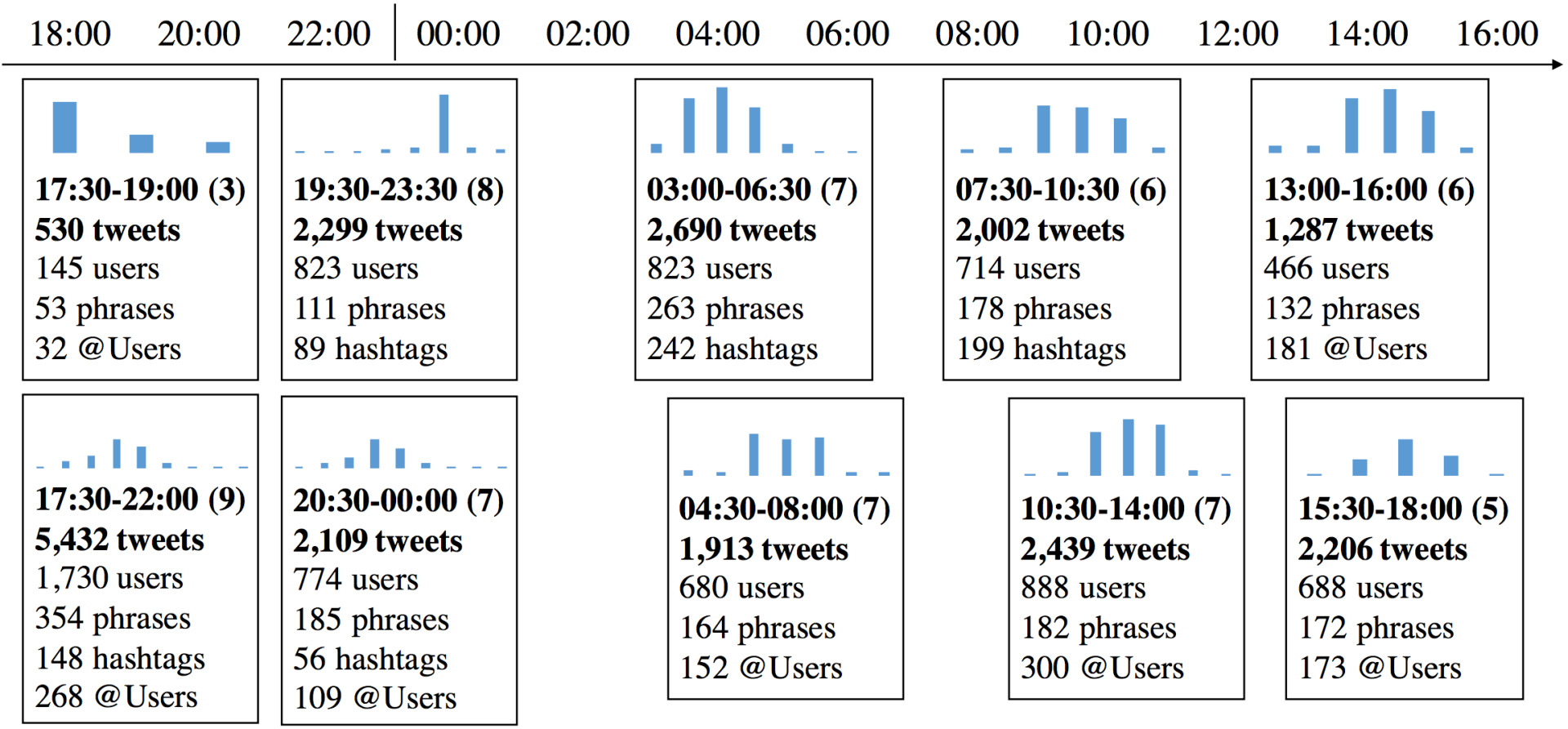
(a) F1 score vs $T^{\mathcal{A}}$. (b) Time cost vs $T^{\mathcal{A}}$. (c) F1 score vs $e^{(t)}$. (d) Time cost vs $e^{(t)}$.

(e) F1 score vs $D^{\mathcal{A}}$. (f) Time cost vs $D^{\mathcal{A}}$. (g) F1 score vs n_d. (h) Time cost vs n_d.

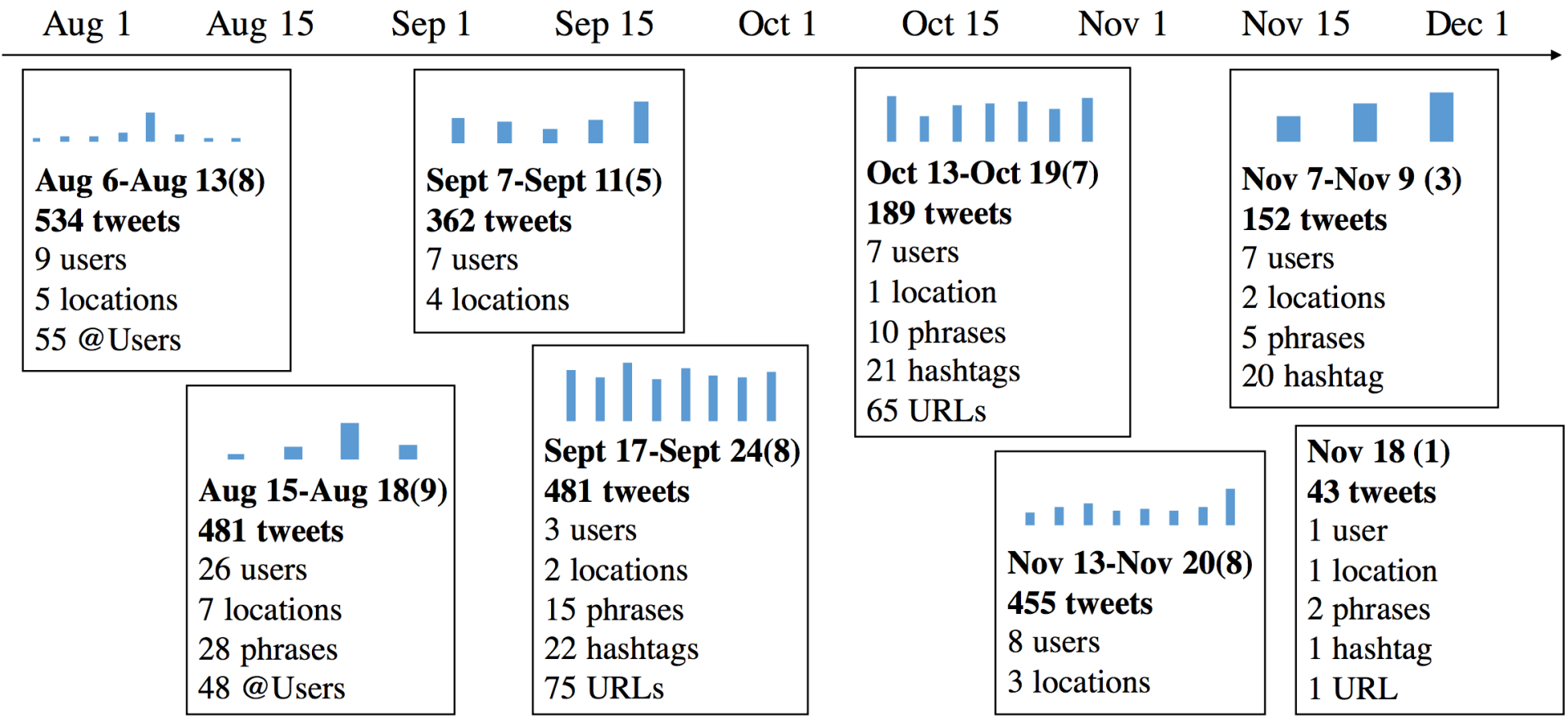
(i) F1 score vs ρ. (j) Time cost vs ρ. (k) F1 score vs T. (l) Time cost vs T.

(m) F1 score vs $E^{(t)}$. (n) Time cost vs $E^{(t)}$. (o) F1 score vs N_d. (p) Time cost vs N_d.

Results (CatchTartan): Grammy 2013



Results (CatchTartan): LA 2014



Results (CatchTartan)

Dataset	Dimensions in a Tartan (the 1st)	#Dims	Pct.	Dimensions of a Tartan (the 2nd)	#Dims	Pct.
NYC14	(User, Location, Phrase)	3	28%	(User, Location, Phrase, Hashtag)	4	13%
LA14	(User, Location, Phrase, @User)	4	40%	(User, Location, Phrase, Hashtag, URL)	5	17%
SPB13	(User, Phrase, URL, @User)	3	67%	(Phrase, Hashtag)	2	21%
GRM13	(User, Phrase, Hashtag, @User)	4	89%	(User, Phrase, @User)	3	6%
DBLP	(Venue, Keyword)	2	38%	(Author, Venue, Keyword)	3	15%