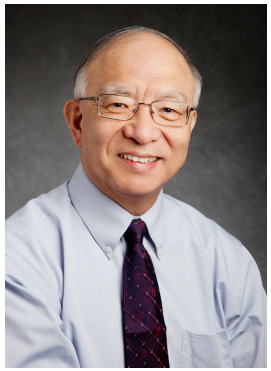
CatchTartan:
Representing and Summarizing Dynamic Multicontextual Behaviors

Meng Jiang (UIUC), Christos Faloutsos (CMU), Jiawei Han (UIUC)





What is Tartan?

Visited CMU in 2012-13

Watched lots of Tartans' games…

What is Behavior? Is it valuable?

Behavior: interactions made by **individuals or organisms** in conjunction with **themselves** or their **environment**. (Wikipedia)

- Tweeting behavior

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

- Publishing-paper behavior

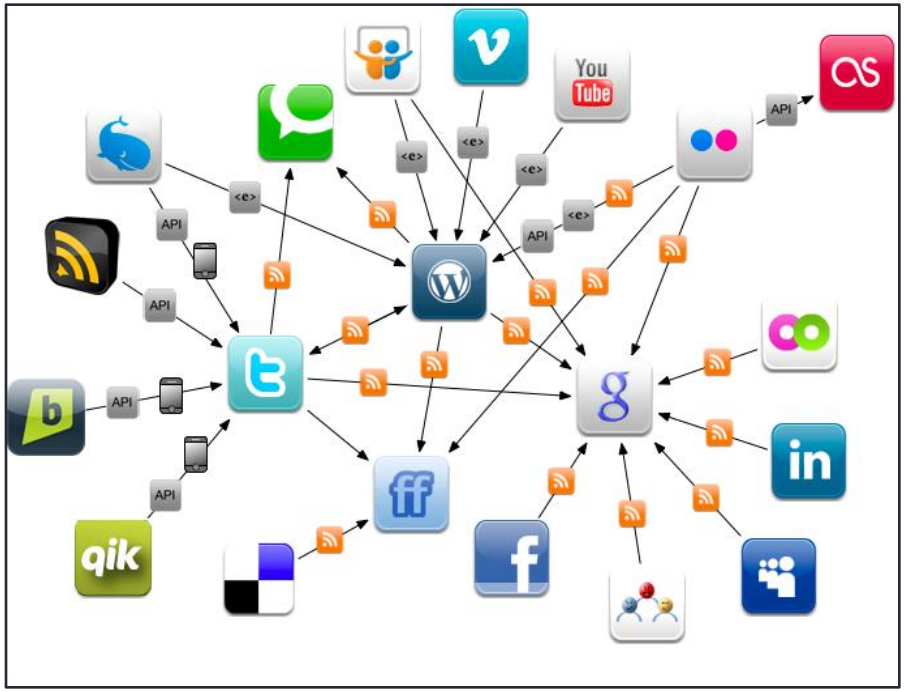
2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Q: *What can we discover from behavioral data?*
Ex. *Given every phone call / message between the military leaders, scientists, businesspersons, Find …*

Why We Talk about Behavior Today?

Physical Environment

Online Environment



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.

Representing and Summarizing Behavior

Representing	Raw data to Math
Summarizing	Patterns: trends, events, campaigns…
Understanding	Factors underlying the patterns: influence, intentions…
Predicting	What will happen in the future?
Intervening	Recommendation, spam/fraud detection…

Given the behavioral data (e.g., DBLP data, tweets)

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

Return behavioral summaries (e.g., research trends, events)

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

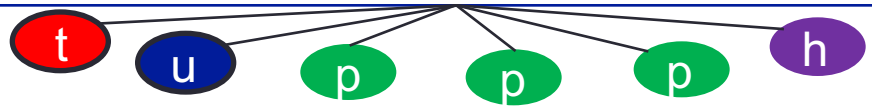
[3] "Optimizing search engines using clickthrough data"

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

Behaviors: Dynamic and Multicontextual

- 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

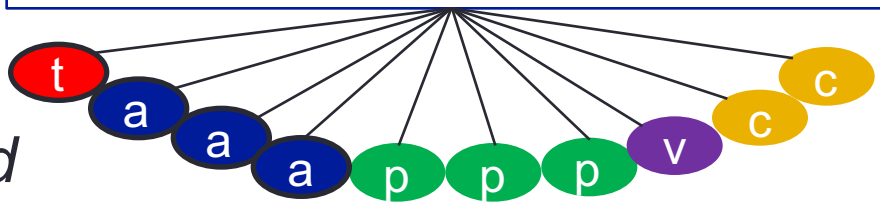
Time slice	User	Location	Phrase	Hashtag	URL
20:00-20:30	@ebekahwsm	∅	{best halftime show, in the history, halftimes shows}	{#SuperBowl}	∅

Behaviors: Dynamic and Multicontextual

- 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 …}

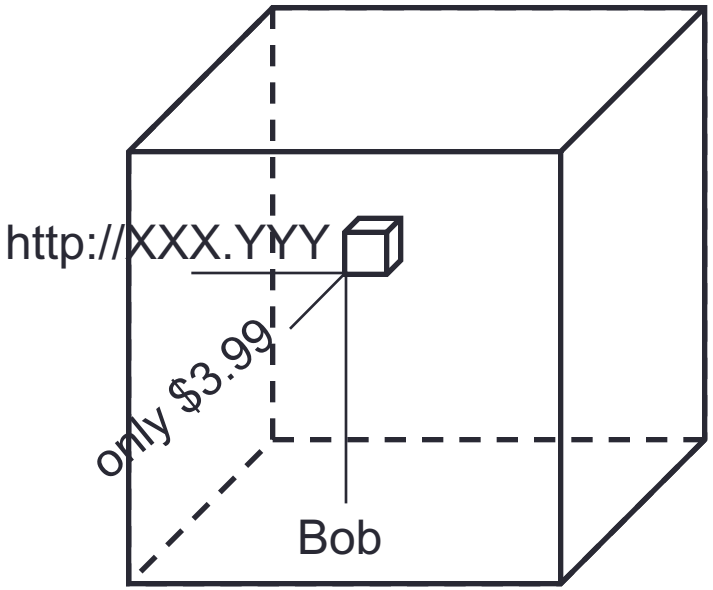
Summarizing Behaviors

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

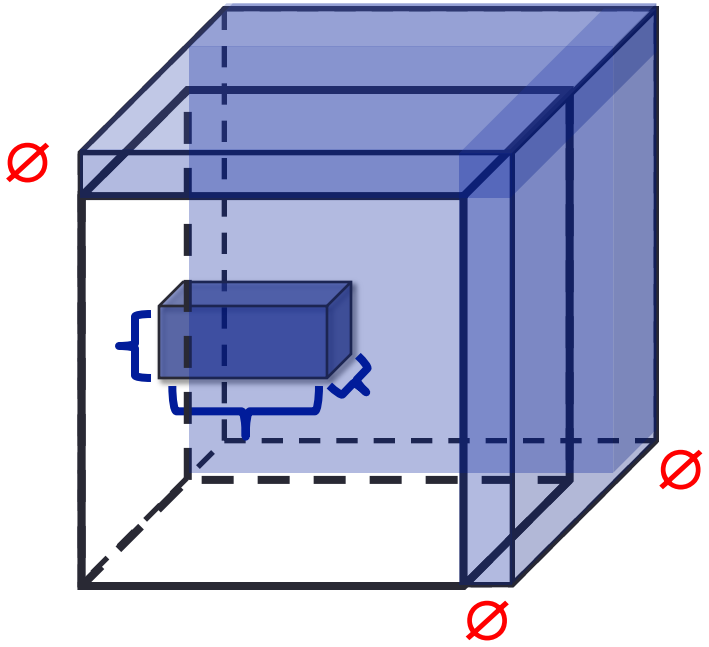
Term	Definition
Dimension	The type of a contextual factor (e.g., location, phrase; author, keyword)
(Dimensional) value	The contextual factor in the dimension
Time slice	The period for consecutive behaviors
Behavior	A set of dimensions, a set of values in each dimension, a time slice for the timestamp

Tensor Fails

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)

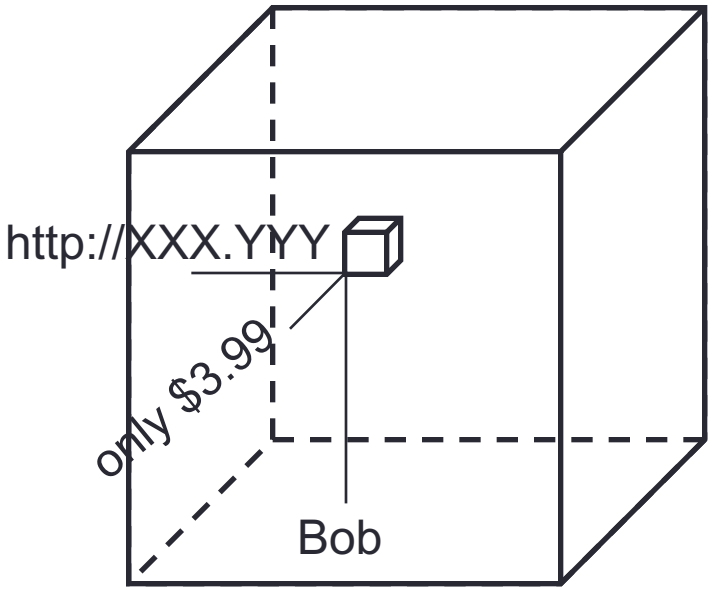


- **Representation: (multicontextual)**
  - **Empty values?**

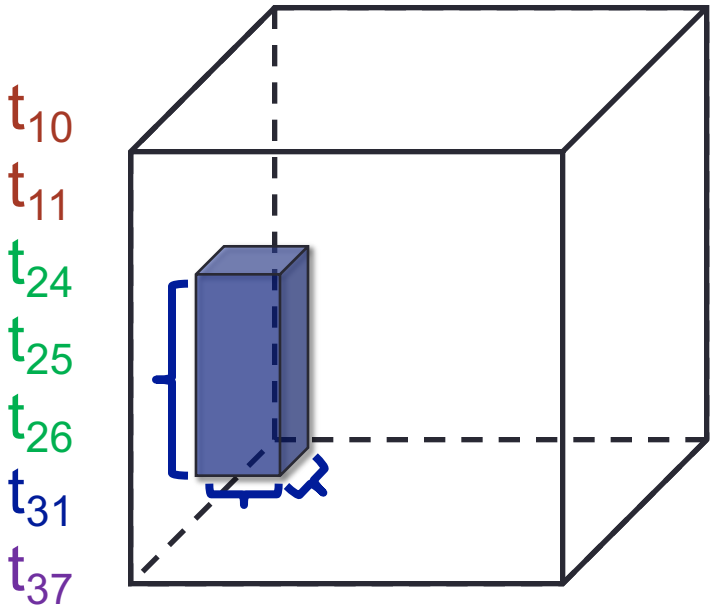


Tensor Fails (cont.)

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)

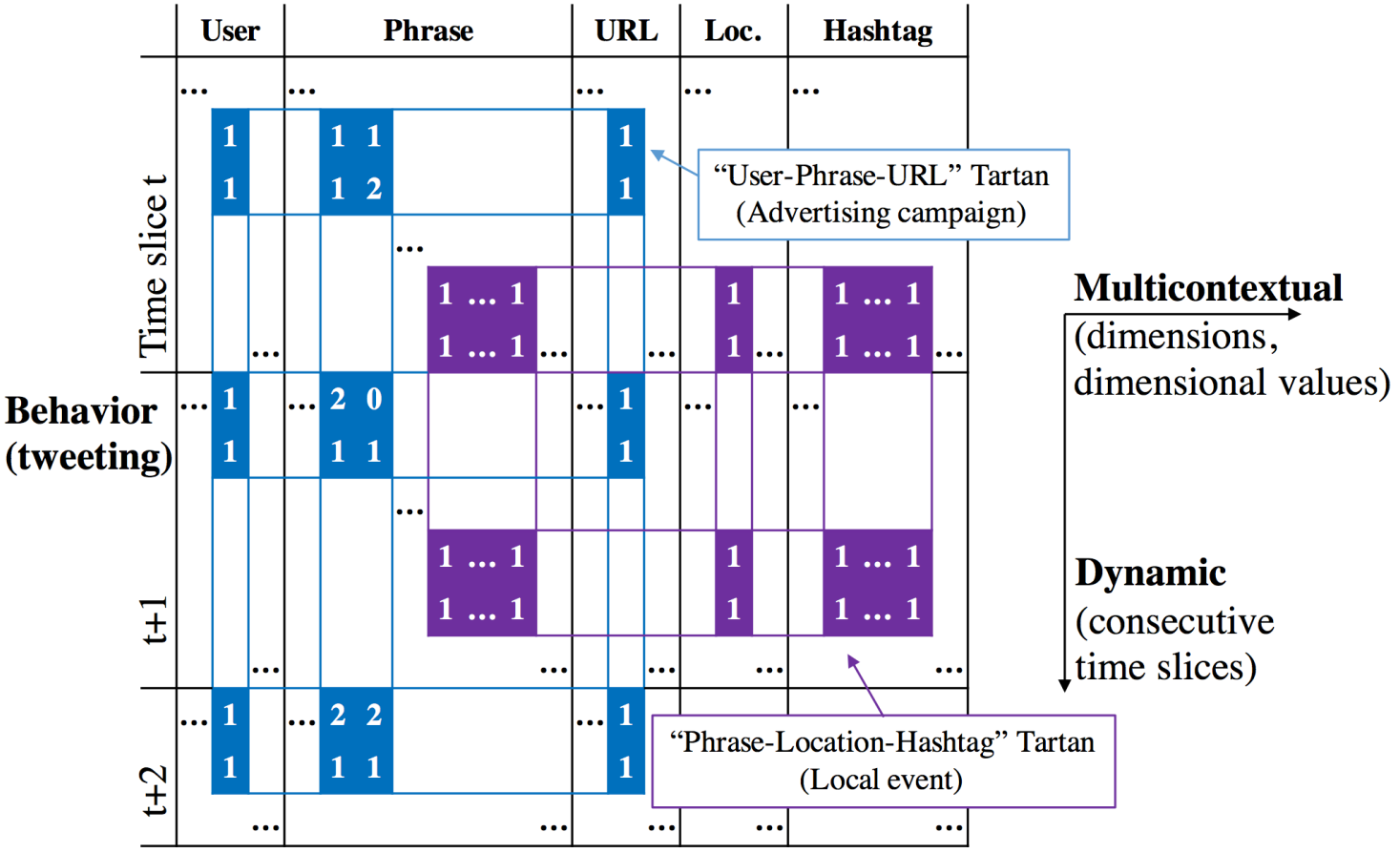


- **Summarization: (dynamic)**
  - **Temporal patterns?**



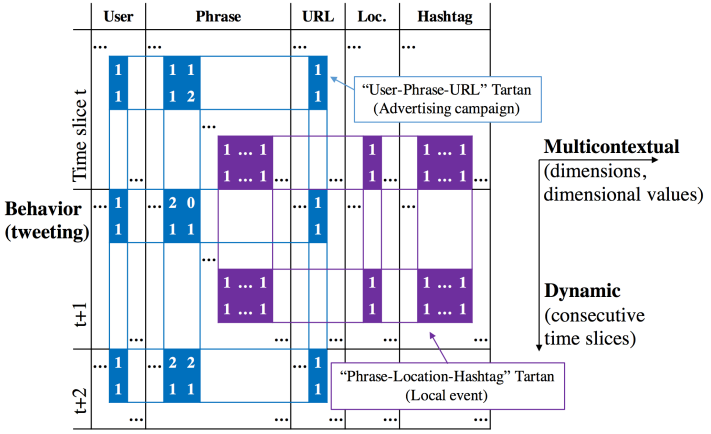
Our Representations for Behavior and Behavioral Summary

- **Behavior: "Two-level matrix"**
- **Behavioral summary: "Tartan"**



The Problem of Behavioral Summarization

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* $\mathcal{A} = \{\mathcal{D}, \mathcal{V}_d|_{d\in\mathcal{D}}, \mathcal{T}, \mathcal{B}^{(t)}|_{t\in\mathcal{T}}\}$.



CATCHTARTAN

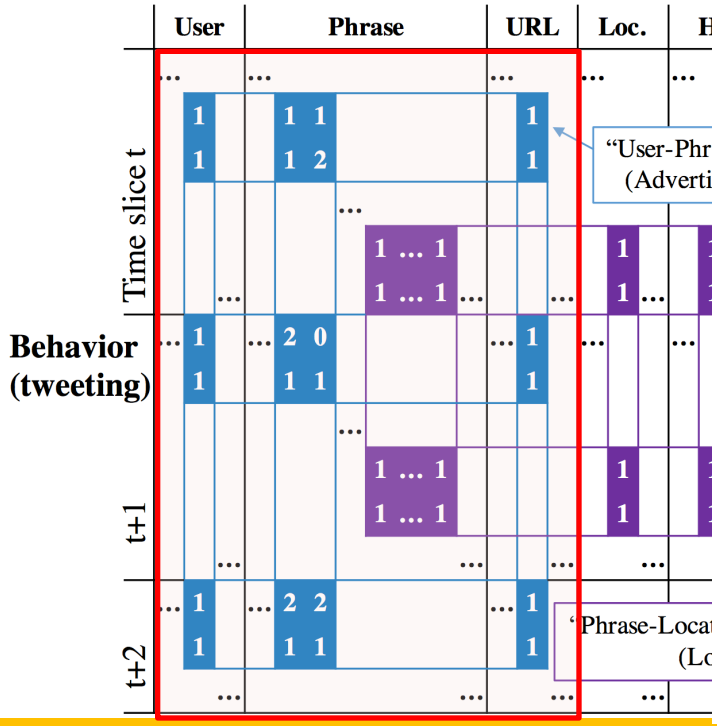
- Employing a lossless encoding scheme
  - The *Minimum Description Length* (MDL) principle
  - Estimating the number of bits that encoding the Tartan can save from merging the meaningful pattern into the encoding of the data

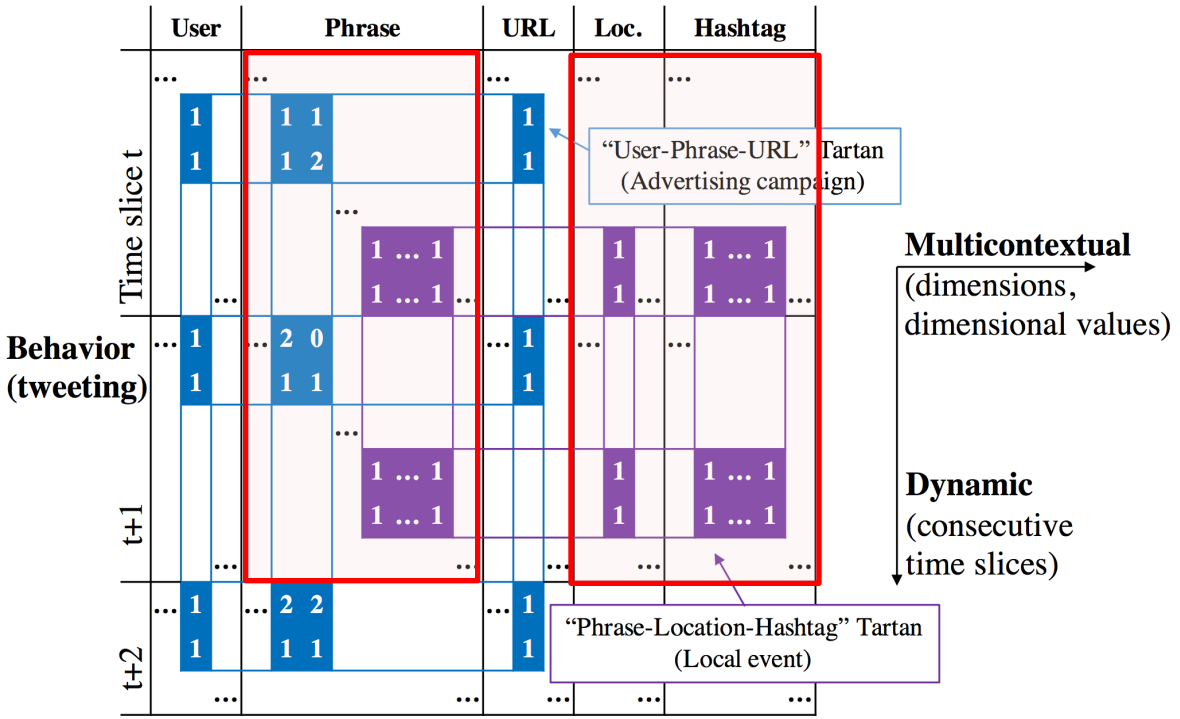
	FSG [18]	GRAPH-CUBE [33]	EVENT-CUBE [29]	MDC [21]	BOW [6]	FEMA [9]	COM2 [2]	CROSS-SPOT [8]	GRAPH-SCOPE [27]	VOG [15]	TIME-CRUNCH [26]	**CATCH-TARTAN**
Principled scoring								√	√	√	√	√
Parameter-free	√				√		√	√	√	√	√	√
Multidimensional		√	√	√	√	√	√	√				√
Multicontextual												√
Timestamp value			√	√		√	√	√	√		√	√
Dynamics							√		√		√	√

Objective Function to Maximize

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

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





$$\mathcal{X}^{\mathcal{A}} = \{\mathcal{X}_d^{(t)}(b,i) | d \in \mathcal{D}, t \in \mathcal{T}, i \in \{1,\ldots,N_d\}, b \in \{1,\ldots,E^{(t)}\}\}.$$

Objective Function to Maximize (cont.)

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

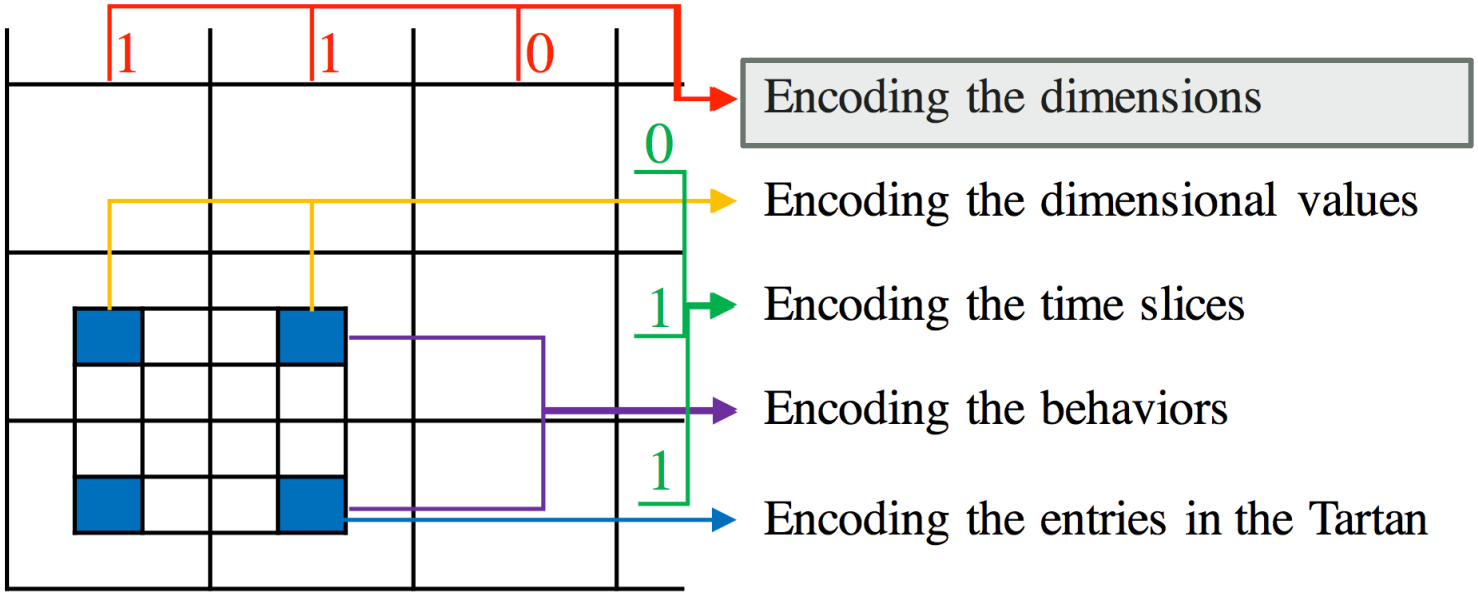
$$\begin{aligned} V &= \left(\sum_{d \in \mathcal{D}} N_d\right)\left(\sum_{t \in \mathcal{T}} E^{(t)}\right). \\ C &= \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \{1,\ldots,E^{(t)}\}, i \in \{1,\ldots,N_d\}} \mathcal{X}_d^{(t)}(b,i). \end{aligned}$$

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

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

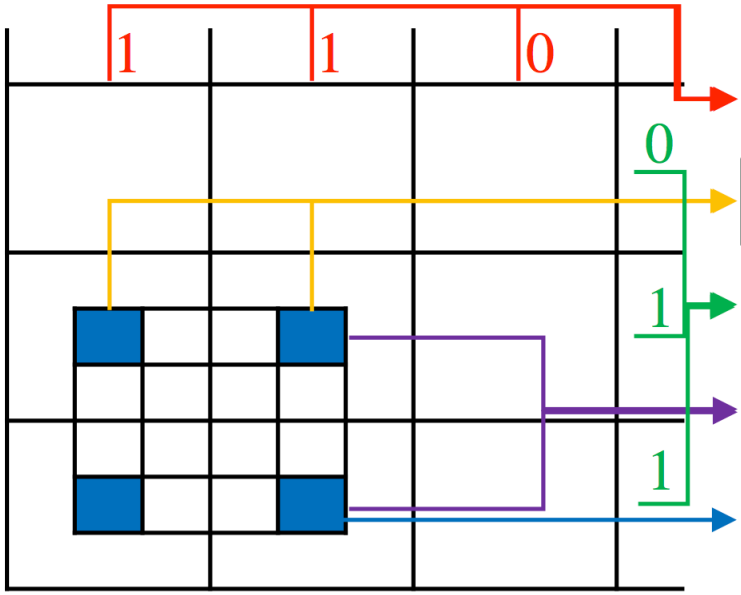
$$L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}) = g(V + C - v - c, C - c);$$

Encoding the Tartan: Dimensions



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

Encoding the Tartan: Dimensional Values



Encoding the dimensions

Encoding the dimensional values

Encoding the time slices

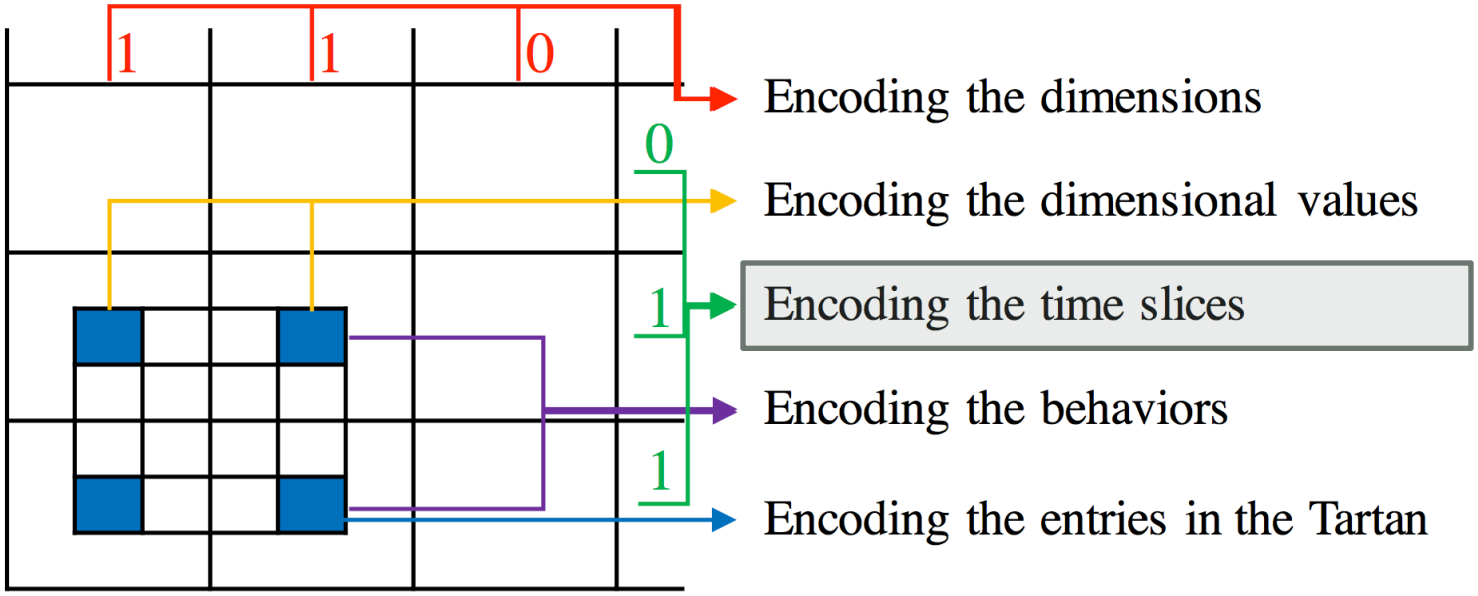
Encoding the behaviors

Encoding the entries in the Tartan

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

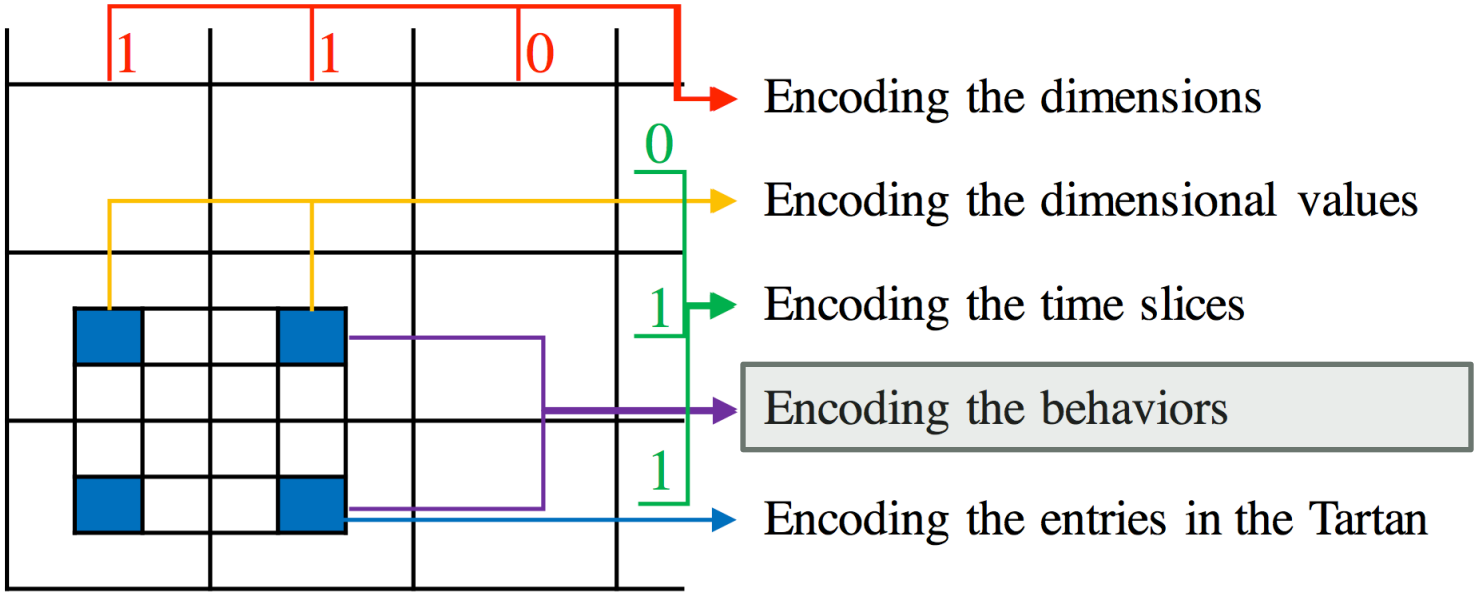
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding the Tartan: Time Slices



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

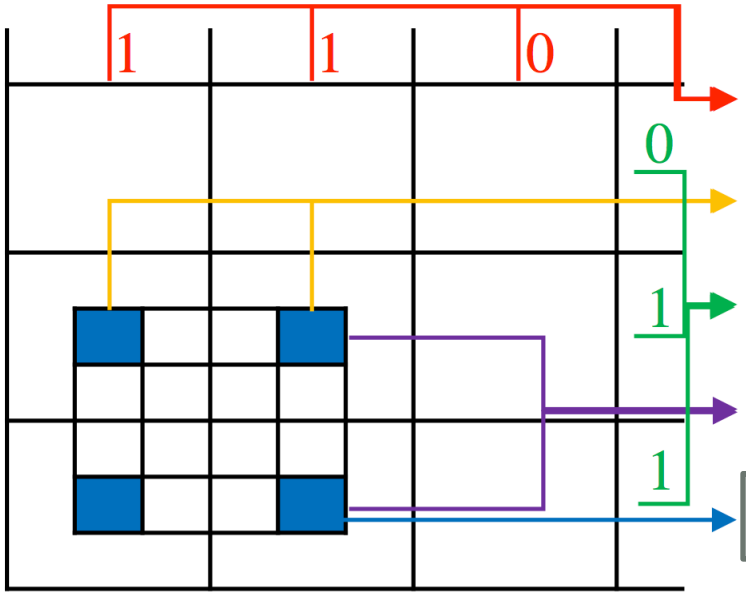
Encoding the Tartan: Behaviors



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

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

Encoding the Tartan: Entries



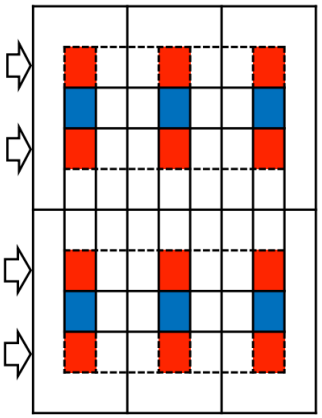
$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

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

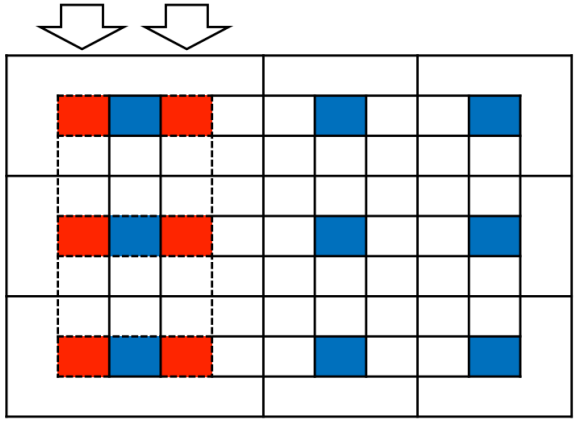
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

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

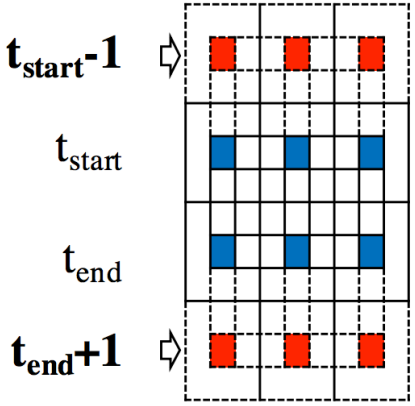
Greedy Search for the Local Minimum



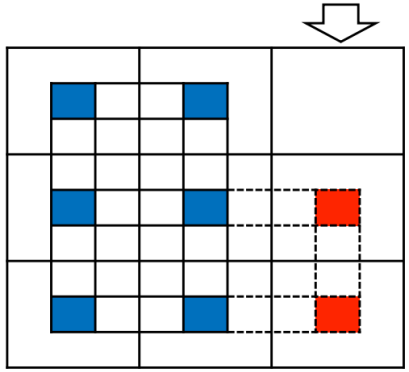
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Time complexity:

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

Qualitative Analysis: DBLP data

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

Qualitative Analysis: Super Bowl 2013

16:30

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)	

Tartan #1: (1 dim) 16:30-17:30

17:00

17:30

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	

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

18:00

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)	

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

18:30

19:00

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	

Tartan #4: (3 dims) 20:00-21:00

19:30

20:00

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)	

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

20:30

21:00

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)	

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

21:30

22:00

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)	

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

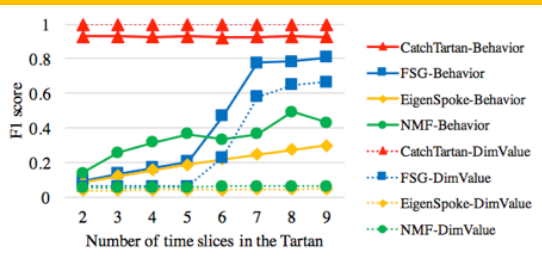
Quantitative Analysis: Accuracy and Efficiency in Synthetic Experiments

❖Tartan distribution

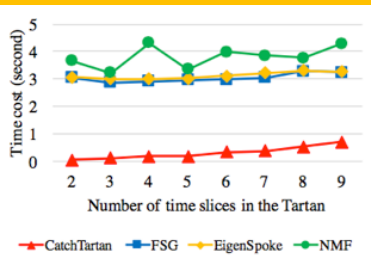
1. $T^{\mathcal{A}} \in [2,\ 9]$, the number of consecutive time slices in the Tartan $\mathcal{A}$, 4 as default;
2. $e^{(t)} \in [100,\ 2,000]$, the number of behaviors in the time slice, $1,000$ as default;
3. $D^{\mathcal{A}} \in [2,\ 9]$, the number of dimensions in $\mathcal{A}$, 3 as default;
4. $n_d \in [50,\ 200]$, the number of values per dimension in $\mathcal{A}$, 100 as default;
5. $\rho \in [1,\ 10]$, the average number of values per dimension in the behaviors, 3 as default;

❖Data distribution

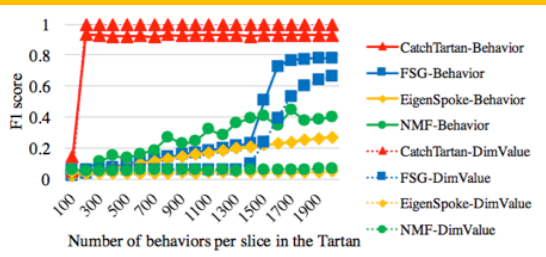
6. $T \in [5, 30]$, the total number of time slices in the dataset, 10 as default;
7. $E^{(t)} \in [1,000,\ 10,000]$, the number of behaviors per time slice in the dataset, $5,000$ as default;
8. $N_d \in [1,000,\ 2,000]$, the number of values per dimension in the data, $1,000$ as default.



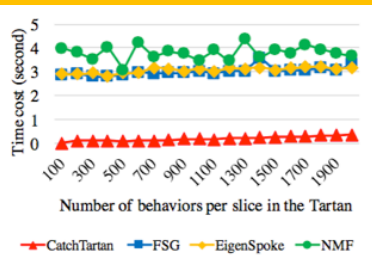
(a) F1 score vs $T^{\mathcal{A}}$.



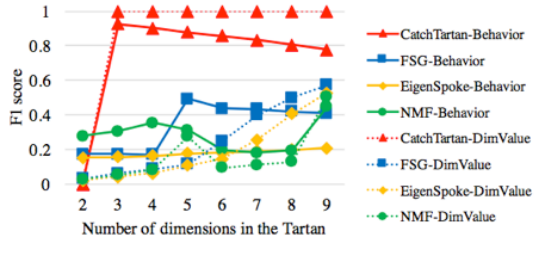
(b) Time cost vs $T^{\mathcal{A}}$.



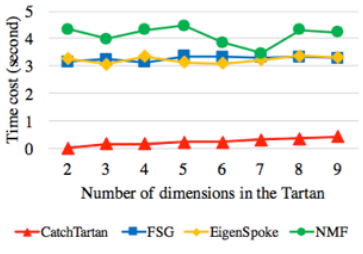
(c) F1 score vs $e^{(t)}$.



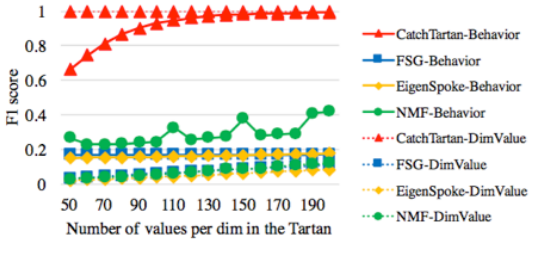
(d) Time cost vs $e^{(t)}$.



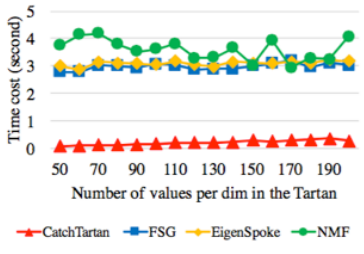
(e) F1 score vs $D^{\mathcal{A}}$.



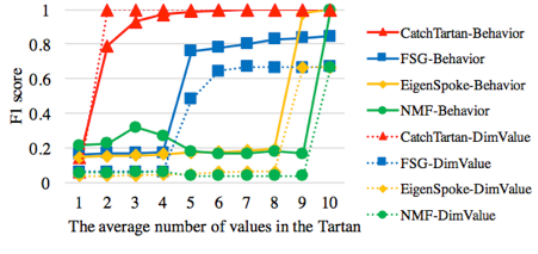
(f) Time cost vs $D^{\mathcal{A}}$.



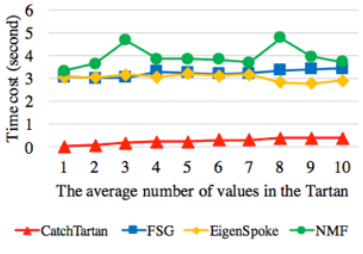
(g) F1 score vs n_d.



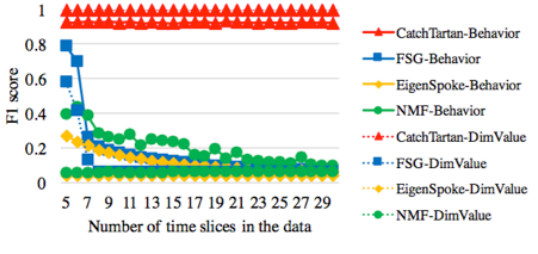
(h) Time cost vs n_d.



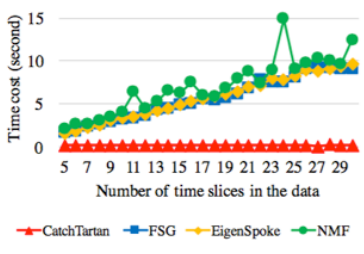
(i) F1 score vs ρ.



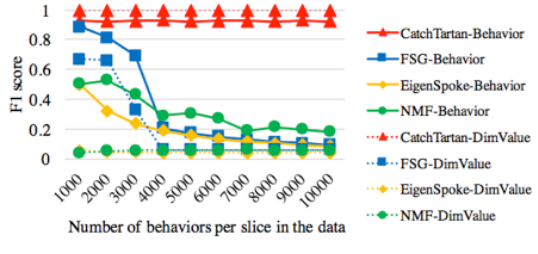
(j) Time cost vs ρ.



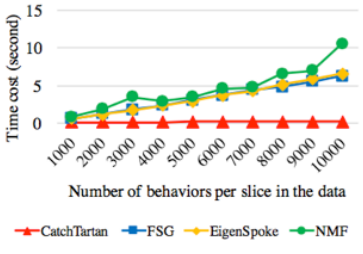
(k) F1 score vs T.



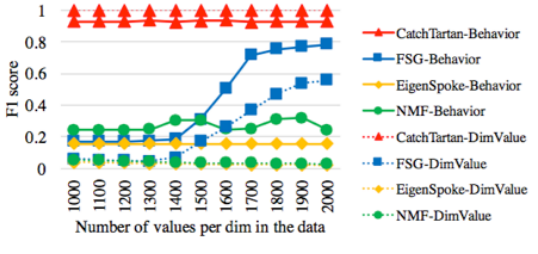
(l) Time cost vs T.



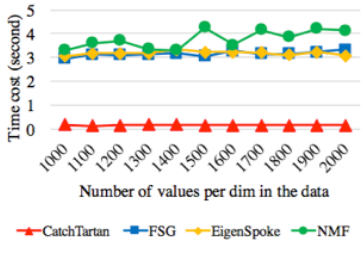
(m) F1 score vs $E^{(t)}$.



(n) Time cost vs $E^{(t)}$.



(o) F1 score vs N_d.



(p) Time cost vs N_d.

Summary

- Novel representations
  - Behavior: “two-level matrix” vs. tensor
  - Behavioral summary: “Tartan” vs. dense block
- A new summarization algorithm
  - Principled-scoring and Parameter-free: Objective function based on Minimum Description Length
  - Scalable: Greedy search for local optimum
- Effectiveness, discovery and efficiency

THANK YOU!

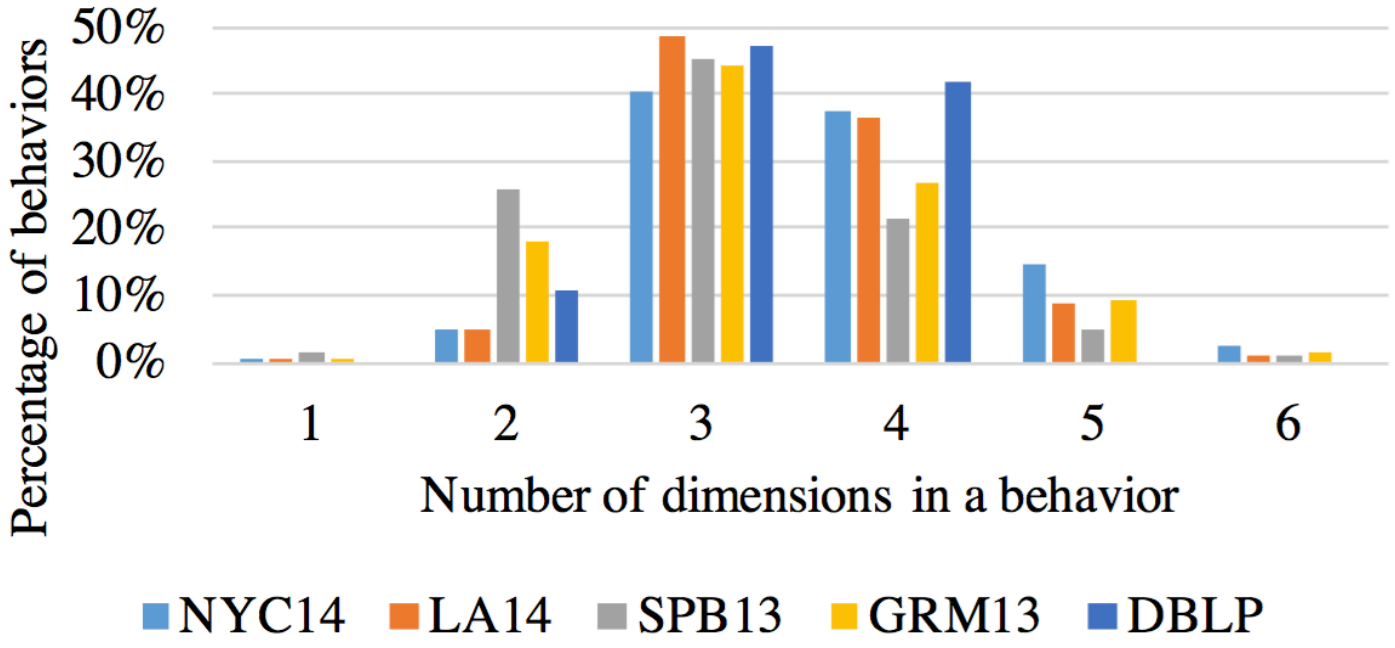
CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors

www.meng-jiang.com

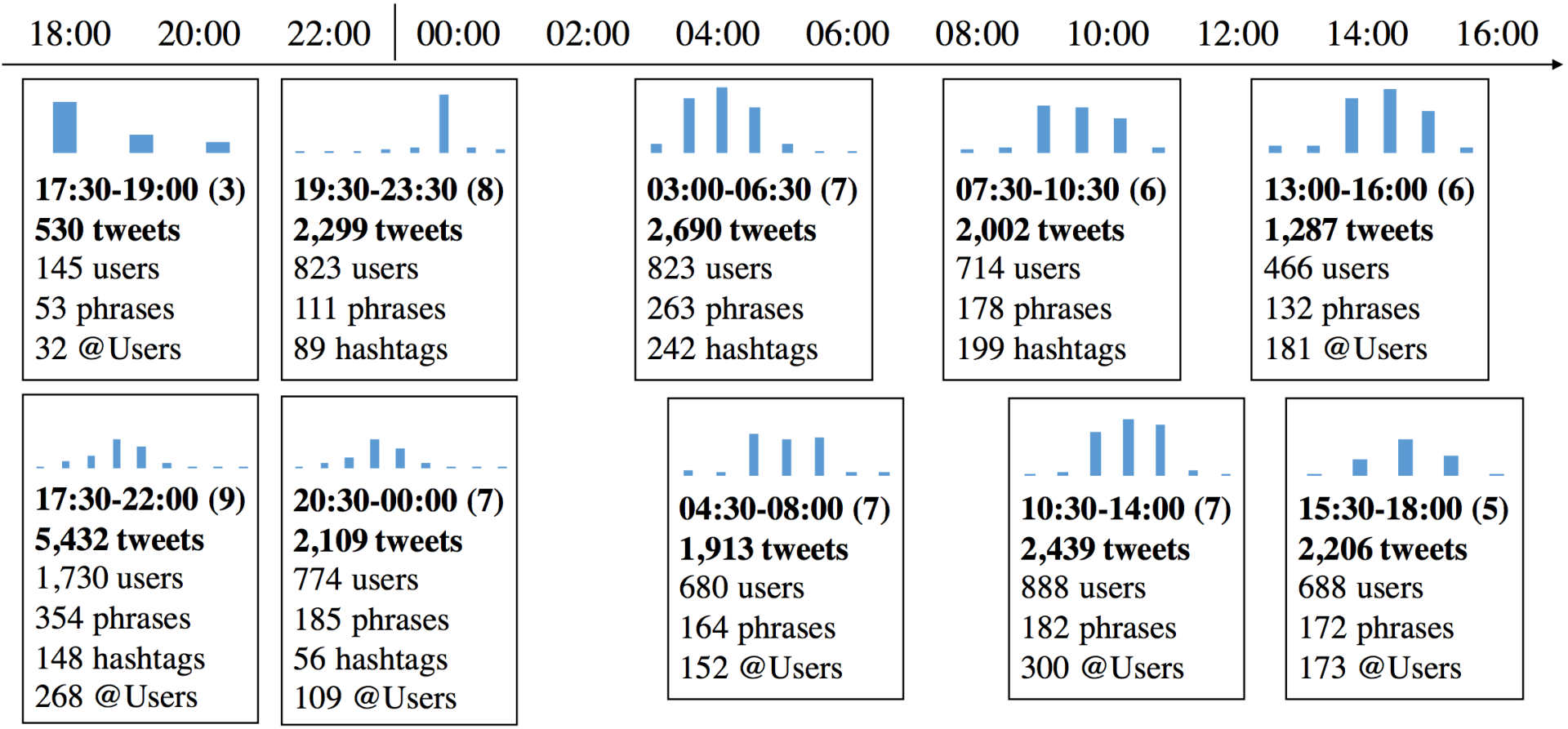
The Distributions in Real Data

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

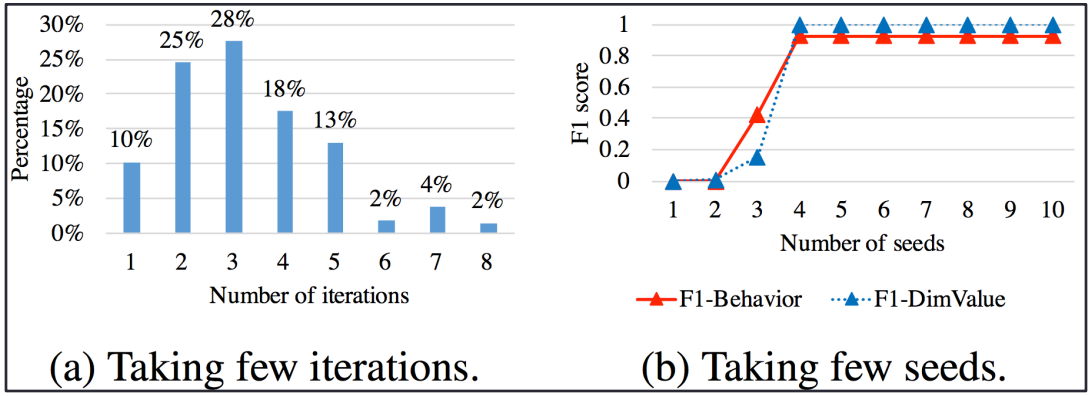


Qualitative Analysis: Grammy 2013

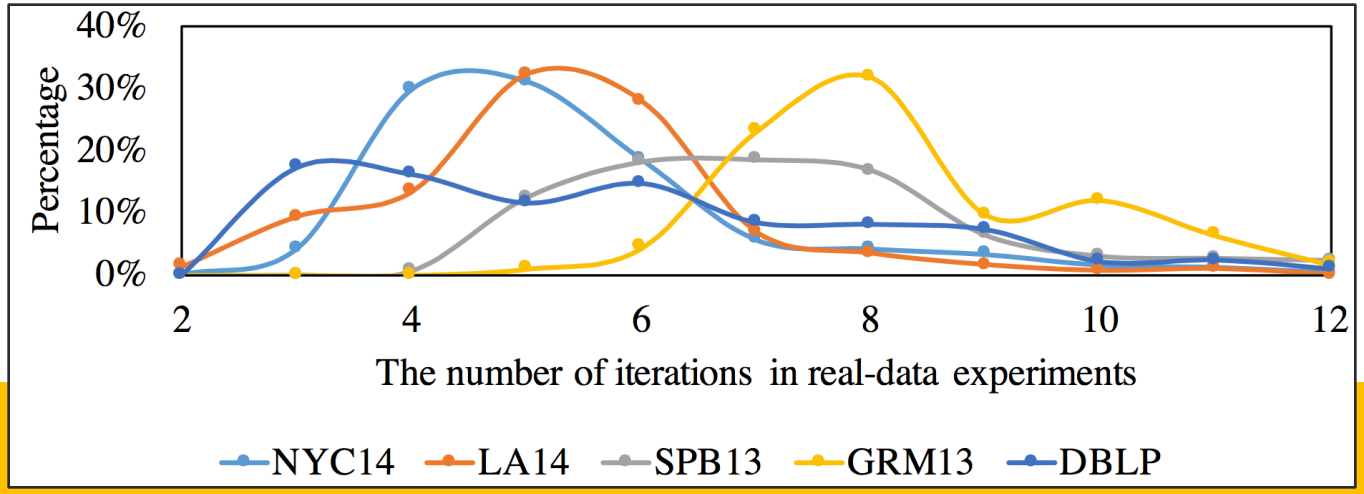


Convergence

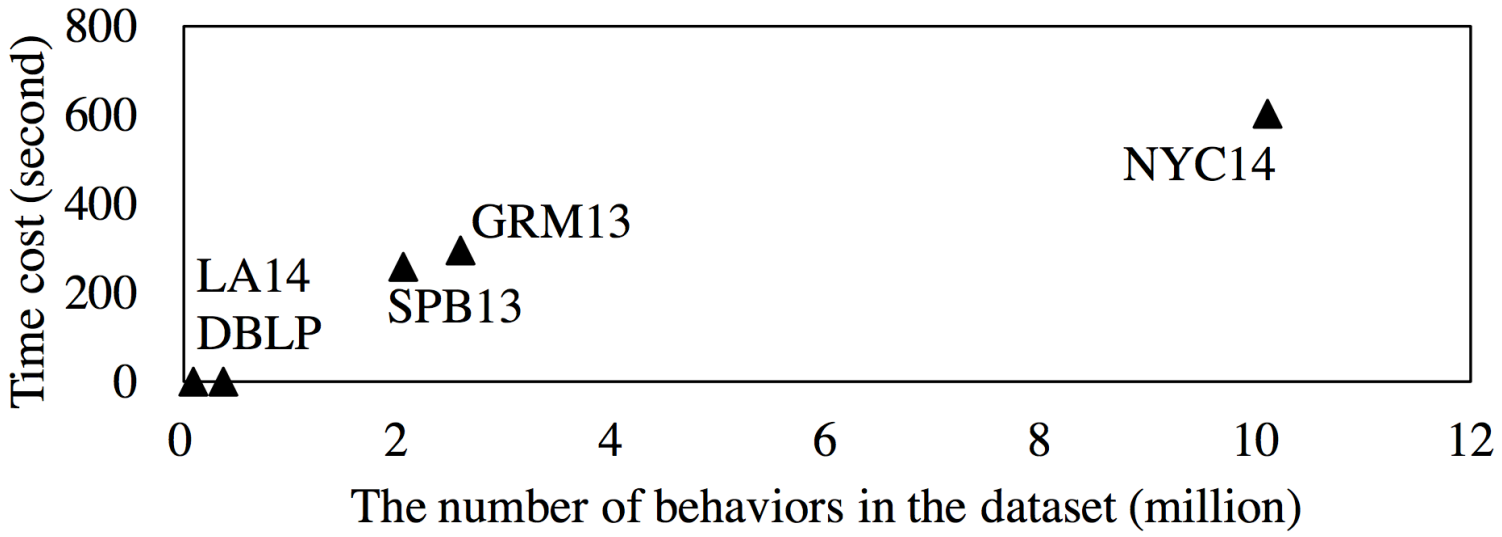
- Synthetic test



- Real-data test



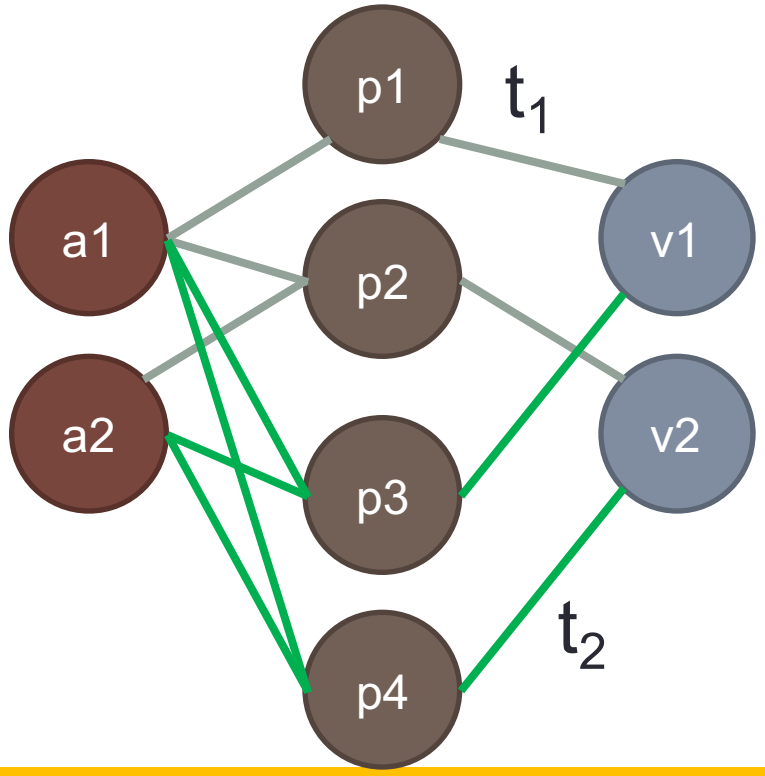
Efficiency and Tartan Distributions



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%

Things Related with "Two-Level" Matrix

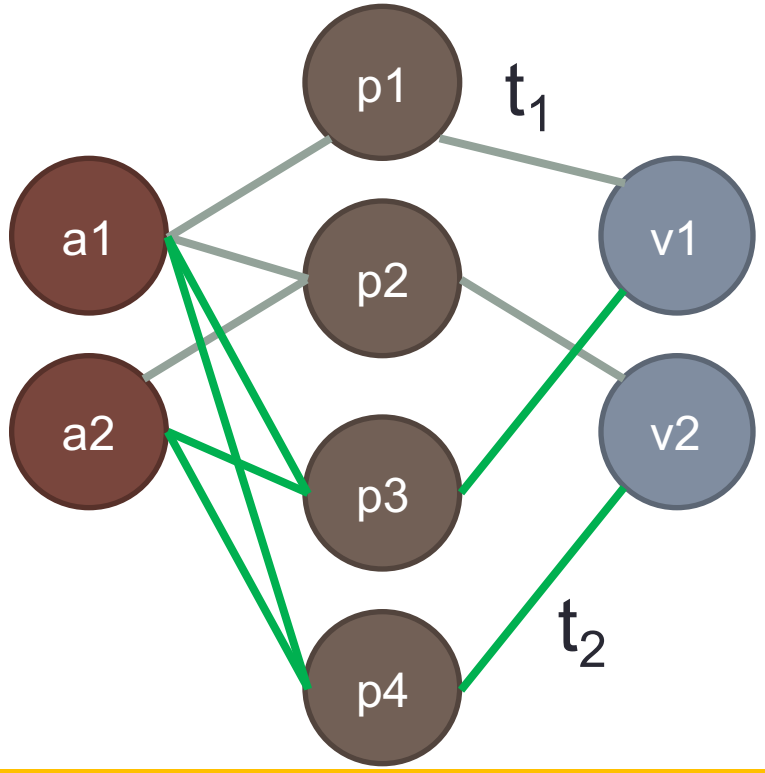
- Time-evolving heterogeneous networks
  - Bipartite one-to-many graph



		Author		Venue		
Papers	t_1	1		1		p_1
		1	1		1	p_2
		...	...	...	...	...
	t_2	1	1	1		p_3
		1	1		1	p_4
		...	...	...	...	...
		a_1	a_2	v_1	v_2	

Things Related with "Two-Level" Matrix (cont.)

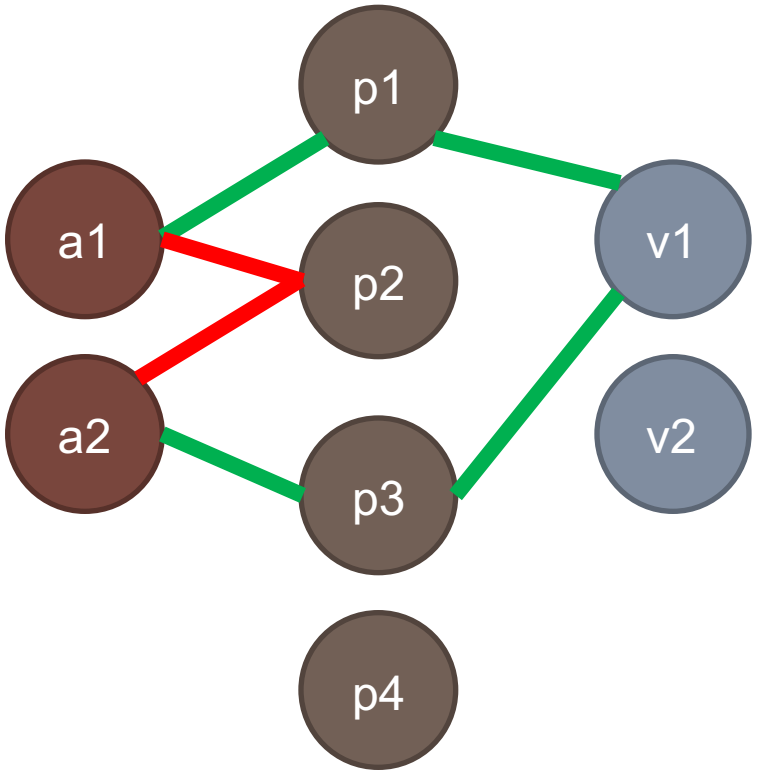
- Time-evolving heterogeneous networks
  - Relationships



		Author		Paper		Venue	
Relationsihps	t_1	1		1			
		1			1		
		...	...	...	...	...	...
	t_2						
		...	...	...	...	...	...
		a_1	a_2	p_1	p_2	v_1	v_2

Things Related with "Two-Level" Matrix (cont.)

- The Meta-Path similarity metric



Author – Paper – Author

Author – Paper – Venue – Paper – Author

		Author		Venue		
Papers	t_1	1		1		p_1
		1	1		1	p_2
		...	...	...	...	...

		Author		Venue		
Papers	t_1	1		1		p_1
		1	1		1	p_2
		...	...	...	...	...
	t_2	1	1	1		p_3
		1	1		1	p_4
		...	...	...	...	...
		a_1	a_2	v_1	v_2	