Data-Driven Behavioral Analytics with Networks

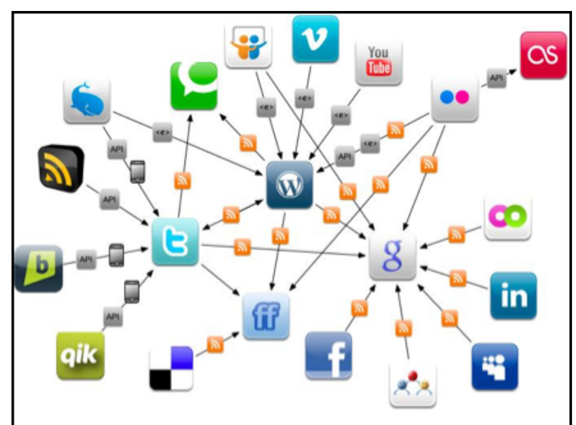
Meng Jiang
www.meng-jiang.com

Talk at XXX

What is Behavior (Analysis)?

- *Definition*. **Interactions** made by **individuals** in conjunction with **themselves** or their **environment**. (*Wikipedia*)
- *Significance*. What can we discover from behavioral data?
  - *Ex*. Given every phone call/message between military leaders, scientists, businesspersons, find…
- *Today*. The human behaviors are broadly recorded in an unprecedented level. Insights of sciences and society?

Physical World **Online Applications**



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

VI. Computational Models of Human Behavior

A fundamental understanding and predictive capability of human behavior dynamics from individuals to societies.

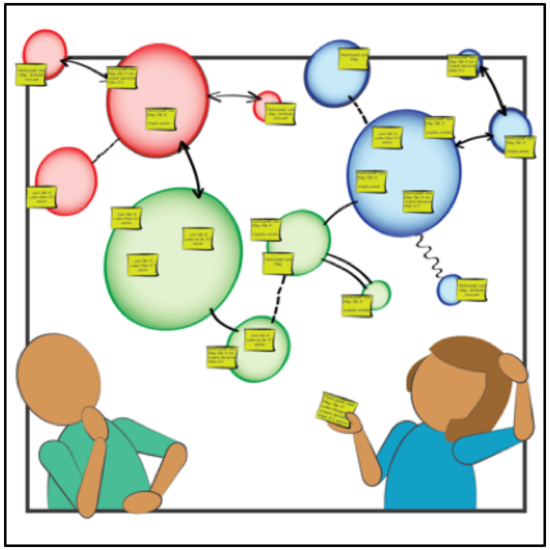

RESEARCH ARTICLES

Costly Punishment Across Human Societies

Joseph Henrich, Richard McElreath, Abigail ... Alexander Bolyanatz, Juan Camilo Cardenas, ... Natalie Henrich, Carolyn Lesorogol, Frank M...

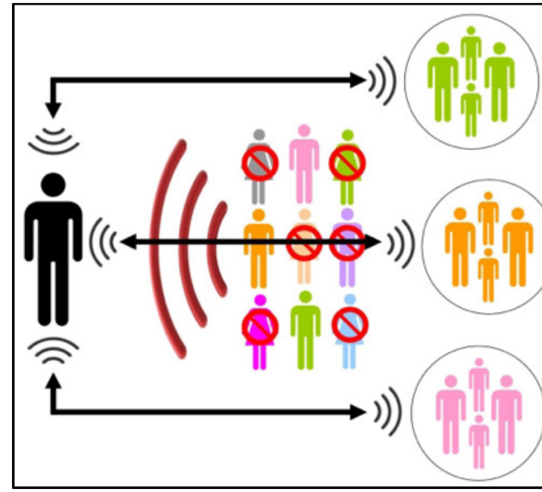
- ***Enabled capabilities***
  - *Predictive models supporting strategic, operational, and tactical decision making and planning*
  - *Real time cultural situational awareness*
  - *Immersive training and mission rehearsal*
  - *Cross cultural coalition building*
- ***Key research challenges:***
  - *Conflicting theories*
  - *Data management and fusion*
  - *Mathematical complexity*
  - *Validation of models*
- ***Measures of success***
  - *Early success of simple models*
  - *Success of social network analysis*
  - *Prediction of crowd tipping points*

Challenges in Behavioral Analysis



Content
(preference)

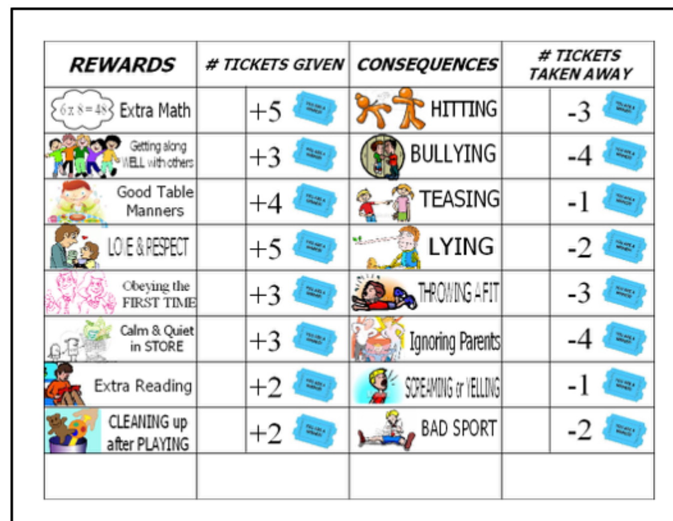
Social context
(influence)



Behavioral Analysis

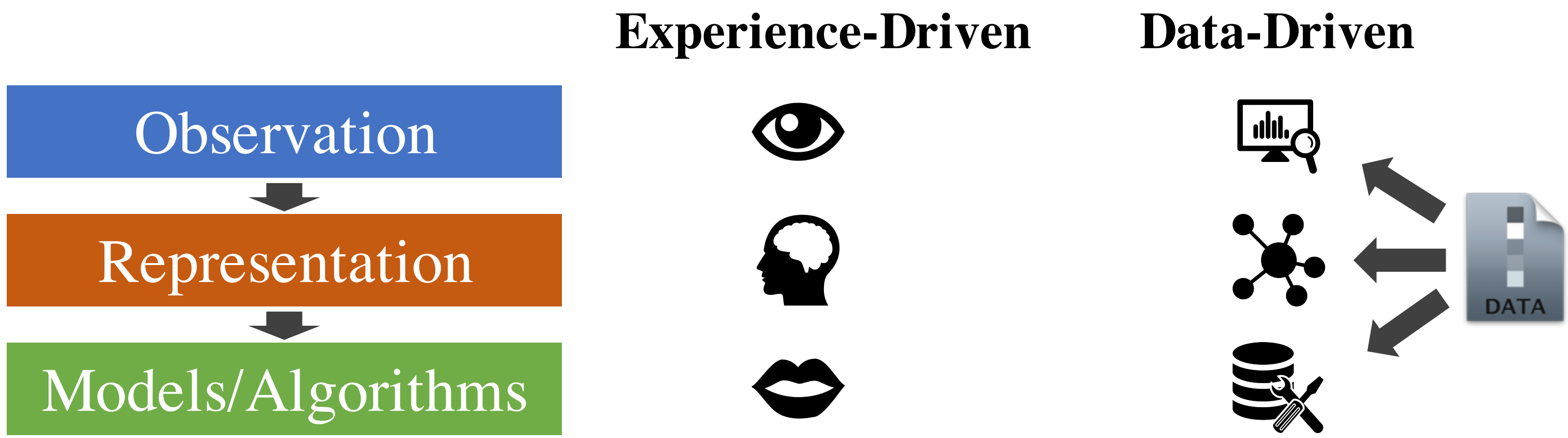
Spatiotemporal context

Intention
(suspiciousness)



REWARDS	# TICKETS GIVEN	CONSEQUENCES	# TICKETS TAKEN AWAY
Extra Math	+5	HITTING	-3
Getting along WELL with others	+3	BULLYING	-4
Good Table Manners	+4	TEASING	-1
LOVE & RESPECT	+5	LYING	-2
Obeying the FIRST TIME	+3	THROWING A FIT	-3
Calm & Quiet in STORE	+3	Ignoring Parents	-4
Extra Reading	+2	SCREAMING or YELLING	-1
CLEANING up after PLAYING	+2	BAD SPORT	-2

Methodology: Why Data-Driven?

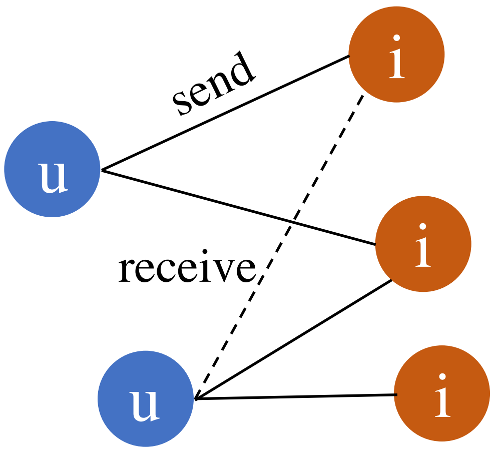


- *Applications*. Recommender systems, fraud/spam detection.
- *Representation*. **B**ehavior **N**etwork for interaction.
  - **Nodes:** users/authors, items (*e.g.*, products, tweets, papers), *etc*.
  - **Links:** (interaction) following, purchasing, tweeting, publishing, *etc*.
  - **Node attributes:** user profiles, item properties/features, *etc*.
  - **Link attributes:** similarity, distance, weight, *etc*.

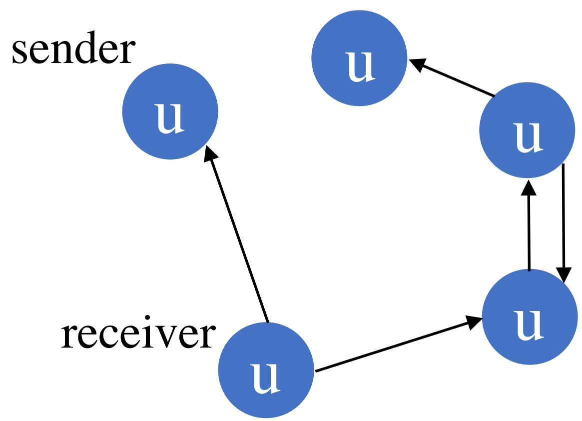
The 1st Day I Became a Data Mining Person

- April 20, 2011: **Tencent Weibo** visited Tsinghua University
  - Low *conversion rate* (< 6%): #retweets per feed request
  - Can we build a *tweet/item recommender system*?
  - **Given**

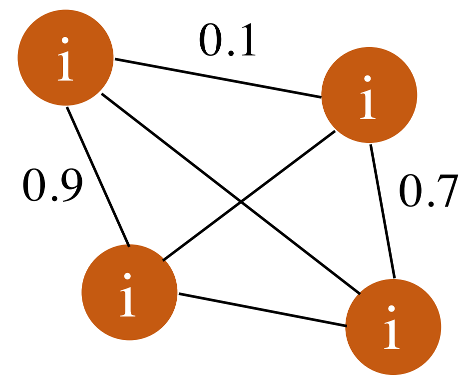
User-item behavior network



User-user social network



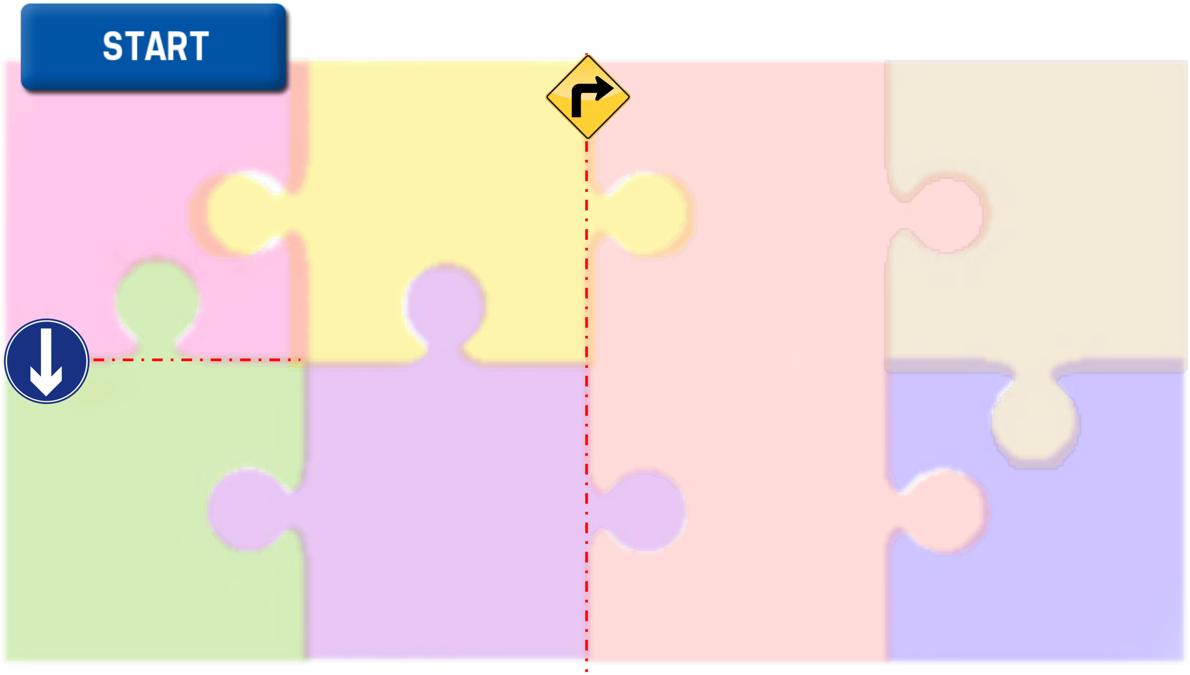
Content similarity (topic level) [Blei *et al.*]



  - **Predict** which tweet/item a user will retweet.

Roadmap

Toolbox

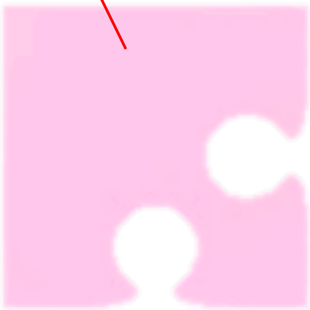


Data-Driven Behavioral Analytics with Networks

Roadmap

Toolbox

Behavior prediction



Related Work

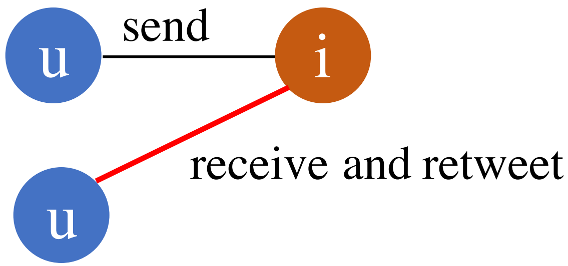
	Behavior	Content	Social	Trust
Collaborative filtering (CF) [Herlocker *et al*. TOIS; Koren KDD]	✔			
Content-based filtering with CF [Balabanovic *et al*.; Liu *et al*. CIKM;]	✔	✔		
SoRec [Ma *et al*. CIKM, TIS] SoReg [Ma *et al*. WSDM]	✔		✔	
Trust-based methods [Massa *et al*. RecSys; Jamali *et al*. KDD; Ma *et al*. SIGIR, TIST]	✔			✔

- **Q:** What are the **factors** of users' decisions on retweeting? Can we **observe** them from the data? How to **integrate** the information for accurate prediction?

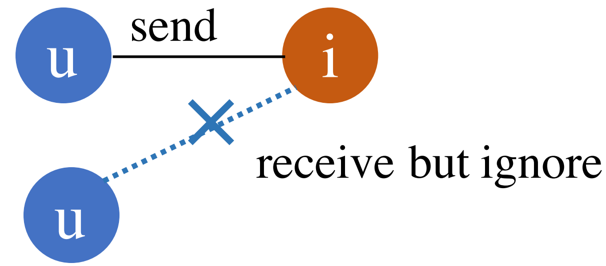
Social Contextual Factors

- Will Michelle Obama share this message?
- Please list your reasons.

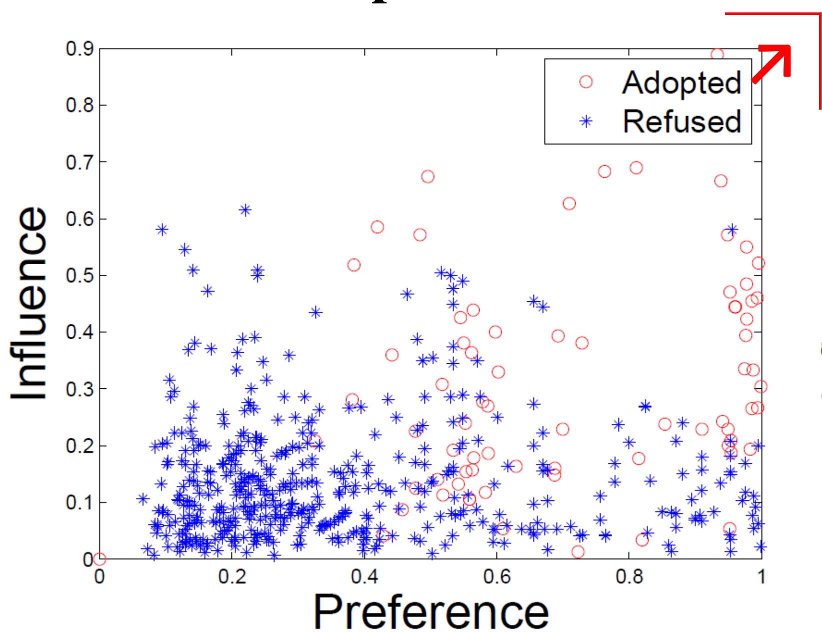
Social Contextual Factors



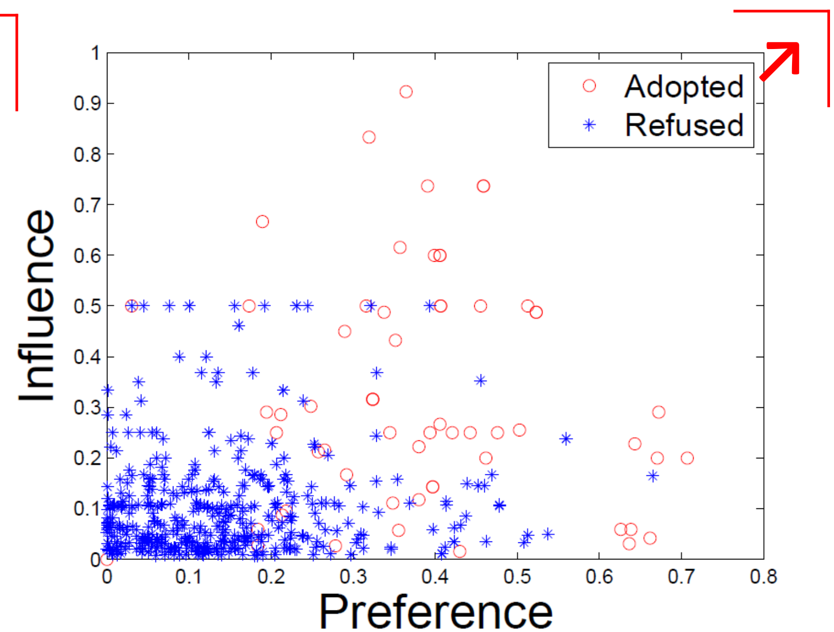
adopted



refused

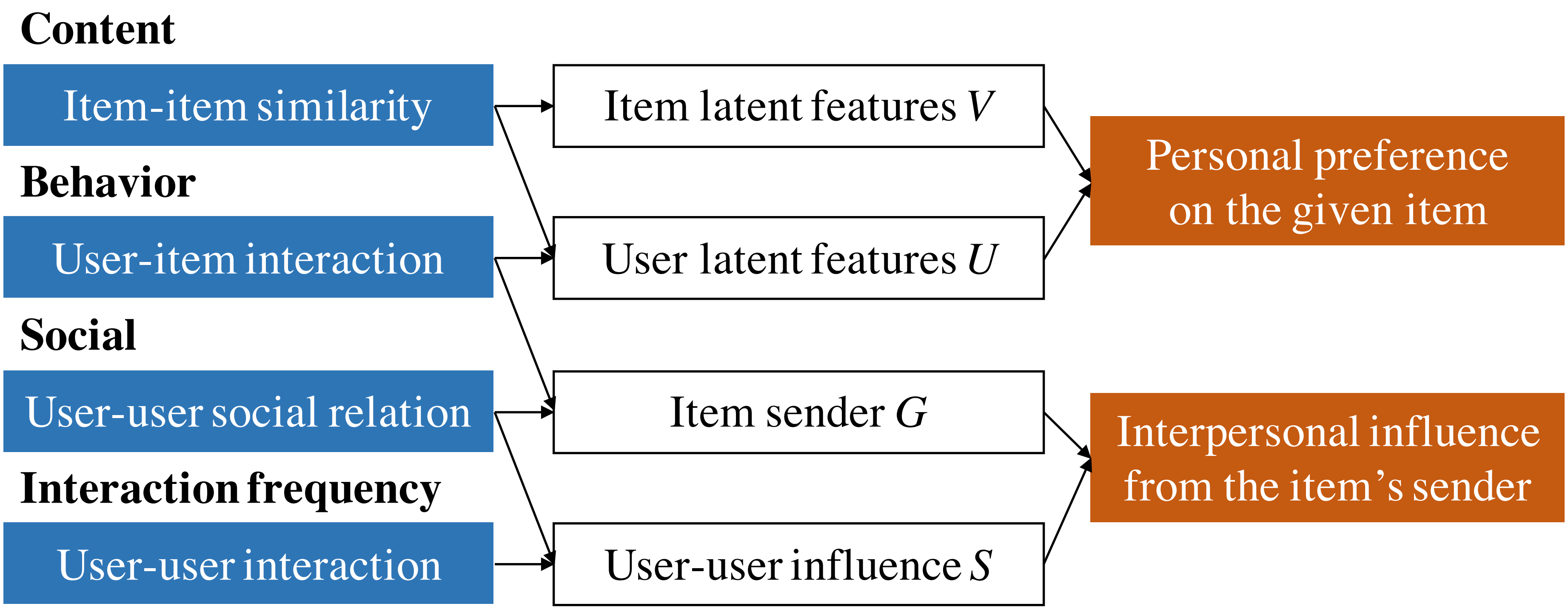


China's Facebook: Renren

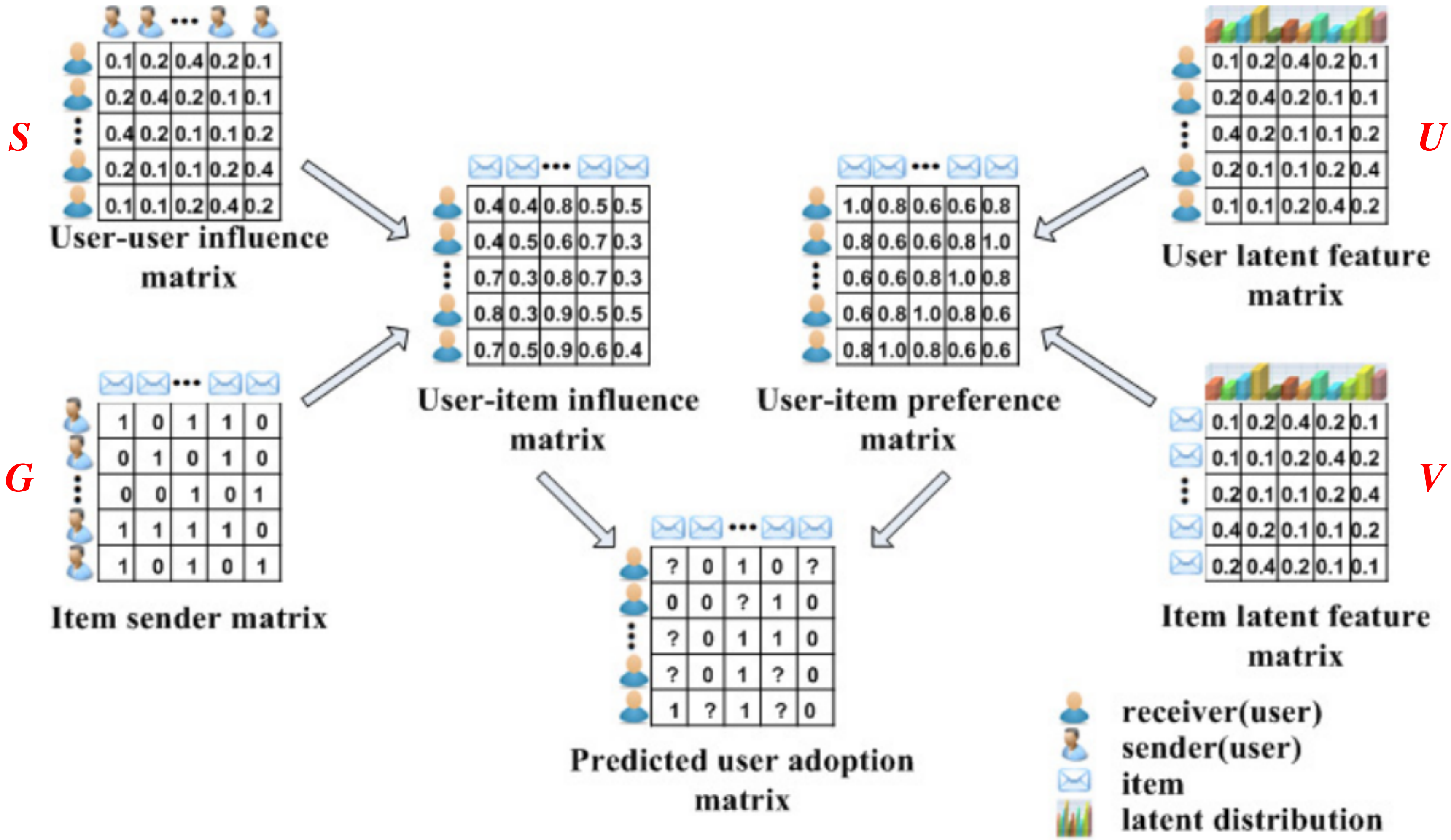


China's Twitter: Tencent Weibo

Modeling: From Contextual Information to Contextual Factors



ContextMF


S
User-user influence matrix
G
Item sender matrix
User-item influence matrix
User-item preference matrix
Predicted user adoption matrix
U
User latent feature matrix
V
Item latent feature matrix
receiver(user)
sender(user)
item
latent distribution

ContextMF

behavior influence preference

$$P(\mathbf{R}|\mathbf{S},\mathbf{U},\mathbf{V},\sigma_R^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}\mathcal{N}(\underline{\mathbf{R}_{ij}}|\underline{\mathbf{S}_i\mathbf{G}_j^\top}\odot\underline{\mathbf{U}_i^\top\mathbf{V}_j},\sigma_R^2)$$

behavior

interaction frequency/trust

item content

$$\begin{aligned}\mathcal{J}=&||\underline{\mathbf{R}-\mathbf{S}\mathbf{G}^\top\odot\mathbf{U}^\top\mathbf{V}}||_F^2+\alpha||\underline{\mathbf{W}-\mathbf{U}^\top\mathbf{U}}||_F^2\\&+\beta||\underline{\mathbf{C}-\mathbf{V}^\top\mathbf{V}}||_F^2+\gamma||\underline{\mathbf{S}-\mathbf{F}}||_F^2\\&+\delta||\mathbf{S}||_F^2+\eta||\mathbf{U}||_F^2+\lambda||\mathbf{V}||_F^2\end{aligned}$$

social relation

ContextMF

- Gradient descent method

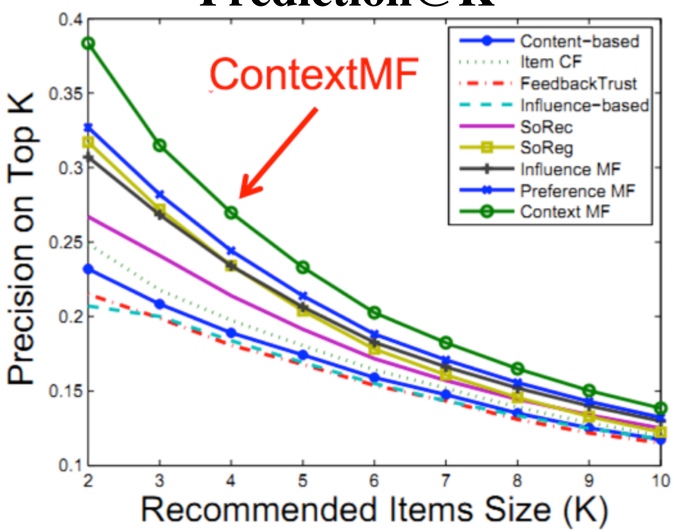
$$
\begin{aligned}
\frac{\partial \mathcal{J}}{\partial \mathbf{S}} &= 2\Big(-\mathbf{R}(\mathbf{G} \odot \mathbf{V}^\top \mathbf{U}) + (\mathbf{S}\mathbf{G}^\top \odot \mathbf{U}^\top \mathbf{V})\mathbf{G} \\
&\quad + \gamma(\mathbf{S} - \mathbf{F}) + \delta \mathbf{S}) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{U}} &= 2\Big(-\mathbf{V}\mathbf{R}^\top + \mathbf{V}(\mathbf{G}\mathbf{S}^\top \odot \mathbf{V}^\top \mathbf{U}) - 2\alpha \mathbf{U}\mathbf{W} \\
&\quad + 2\alpha \mathbf{U}\mathbf{U}^\top \mathbf{U} + \eta \mathbf{U}\Big) \\
\frac{\partial \mathcal{J}}{\partial \mathbf{V}} &= 2\Big(-\mathbf{U}\mathbf{R} + \mathbf{U}(\mathbf{S}\mathbf{G}^\top \odot \mathbf{U}^\top \mathbf{V}) - 2\beta \mathbf{V}\mathbf{C} \\
&\quad + 2\beta \mathbf{V}\mathbf{V}^\top \mathbf{V} + \lambda \mathbf{V}\Big)
\end{aligned}
$$

Experimental Results

Method	MAE	RMSE	$\hat{\tau}$	$\hat{\rho}$
Renren Dataset				
Content-based [1]	0.3842	0.4769	0.5409	0.5404
Item CF [25]	0.3601	0.4513	0.5896	0.5988
FeedbackTrust [22]	0.3764	0.4684	0.5433	0.5469
Influence-based [9]	0.3859	0.4686	0.5394	0.5446
SoRec [19]	0.3276	0.4127	0.6168	0.6204
SoReg [20]	0.2985	0.3537	0.7086	0.7140
Influence MF	0.3102	0.3771	0.6861	0.7006
Preference MF	0.3032	0.3762	0.6937	0.7036
Context MF	**0.2416**	**0.3086**	**0.7782**	**0.7896**
Tencent Weibo Dataset				
Content-based [1]	0.2576	0.3643	0.7728	0.7777
Item CF [25]	0.2375	0.3372	0.7867	0.8049
FeedbackTrust [22]	0.2830	0.3887	0.7094	0.7115
Influence-based [9]	0.2651	0.3813	0.7163	0.7275
SoRec [19]	0.2256	0.3325	0.7973	0.8064
SoReg [20]	0.1997	0.2962	0.8390	0.8423
Influence MF	0.2183	0.3206	0.8179	0.8258
Preference MF	0.2111	0.3088	0.8384	0.8453
Context MF	**0.1514**	**0.2348**	**0.8570**	**0.8685**

vs. SoReg [TIST'11]	Renren	Tencent Weibo
MAE	↓19.1%	↓24.2%
RMSE	↓12.8%	↓20.7%
Kendall's	↑9.82%	↑2.1%
Spearman's	↑10.6%	↑3.1%

Prediction@K

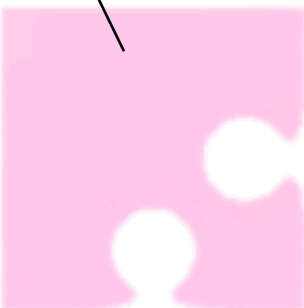


Impact

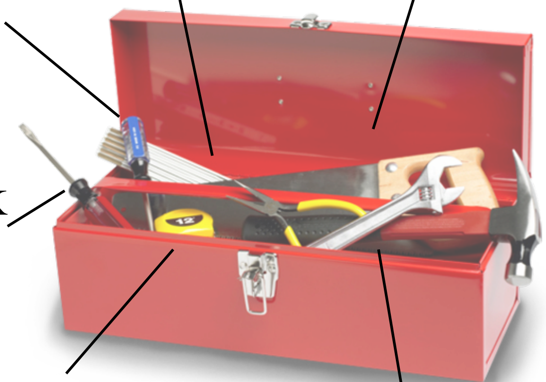
- **Deployed** in Weibo News Feed. Improved conversion rate from 5.78% to 8.27% (relatively **43%**).
- **M. Jiang**, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu and S. Yang. "Social Contextual Recommendation" in **CIKM'12**.
  - #citations = **149**
- **Can we transfer knowledge across domains/platforms?**
- **Cross-domain** behavior modeling. CIKM'12.
  - #citations = **52**
- **Cross-platform** behavior modeling. AAAI'16.

Roadmap

Behavior prediction



Toolbox



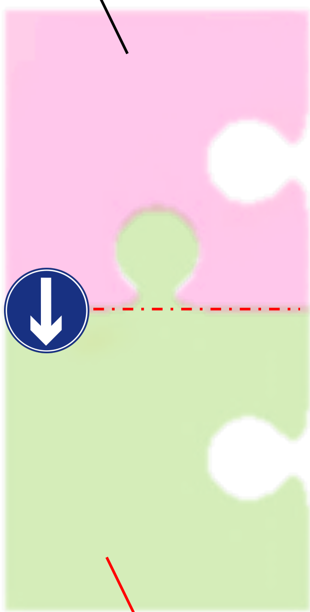
A More Serious Problem in Weibo

Experience-driven approaches: features of #followees, #hashtags, #URLs…

Roadmap

Toolbox

Behavior prediction

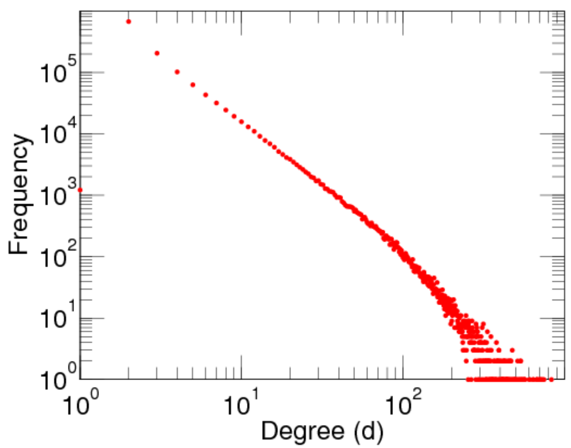
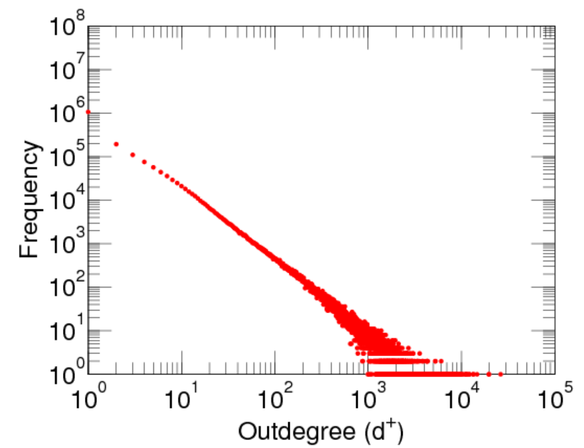
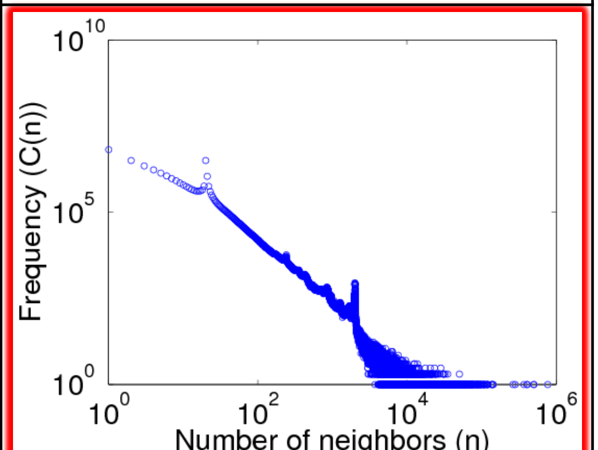


Visited **Prof. Christos Faloutsos** (CMU) from Aug 2012 to May 2013

Suspicious behavior detection

Out-Degree Distributions

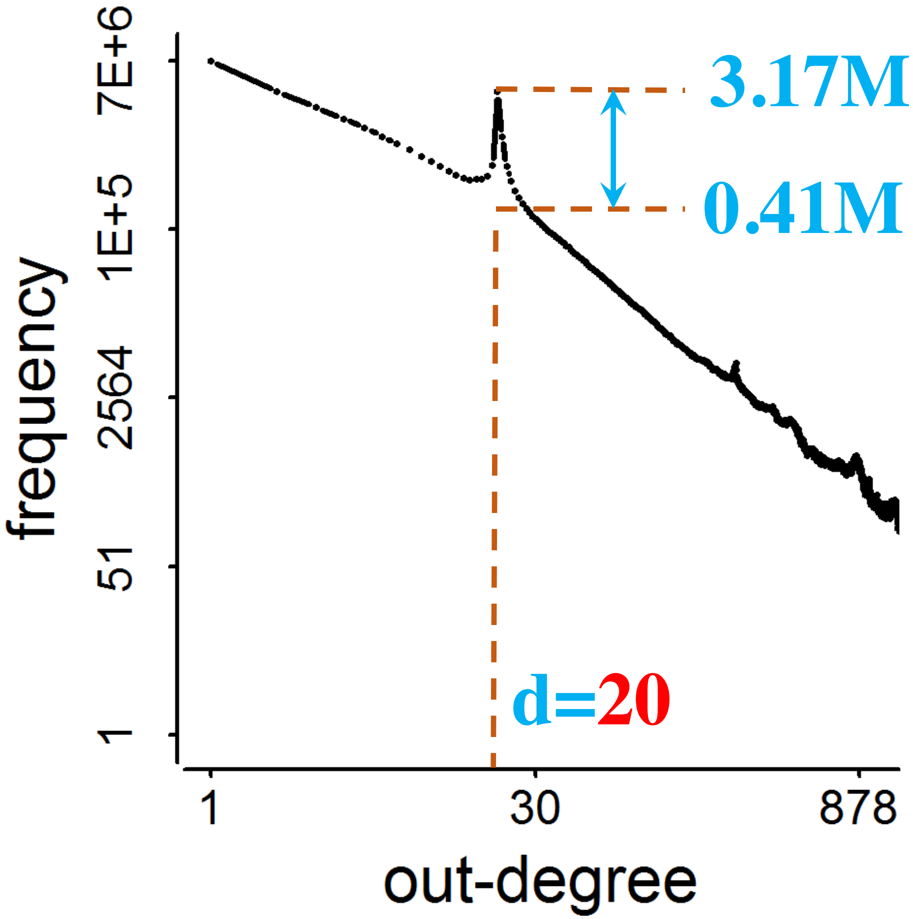
- **Power-law distribution** [Faloutsos *et al*. SIGCOMM; Broder *et al*. Computer Networks; Chung *et al*. PNAS]

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

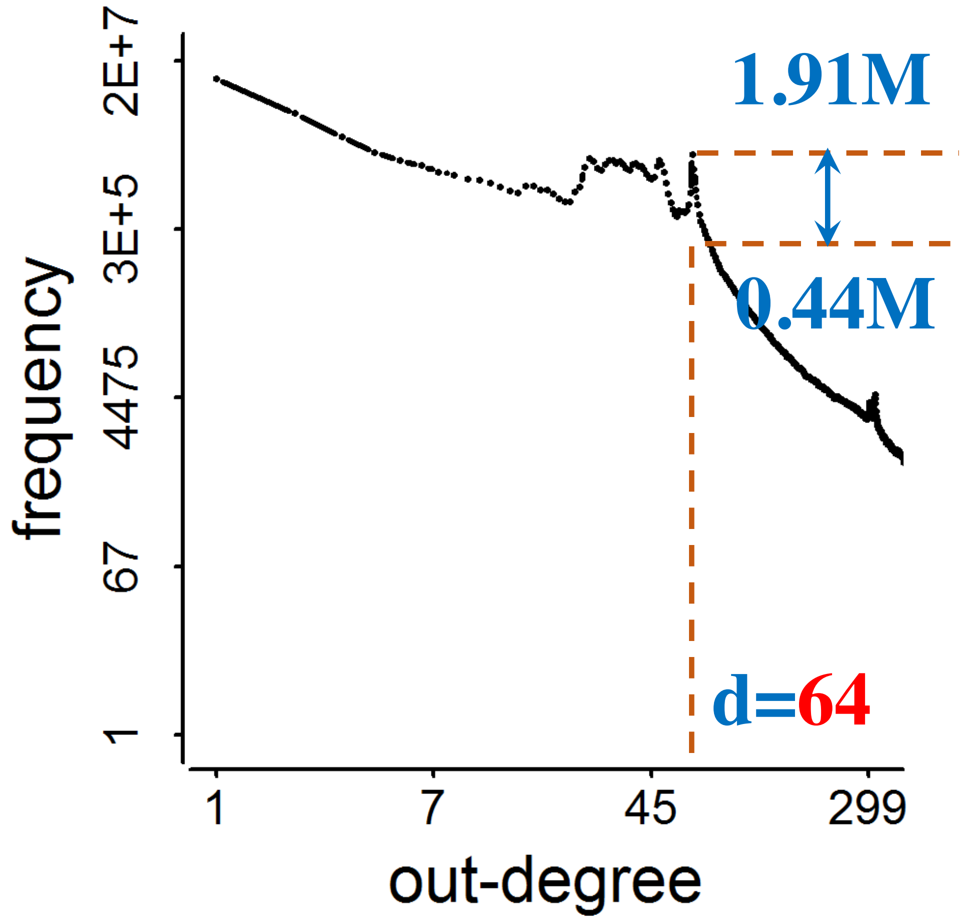
[konect.uni-koblenz.de/networks/]

Spikes!

2009 41M

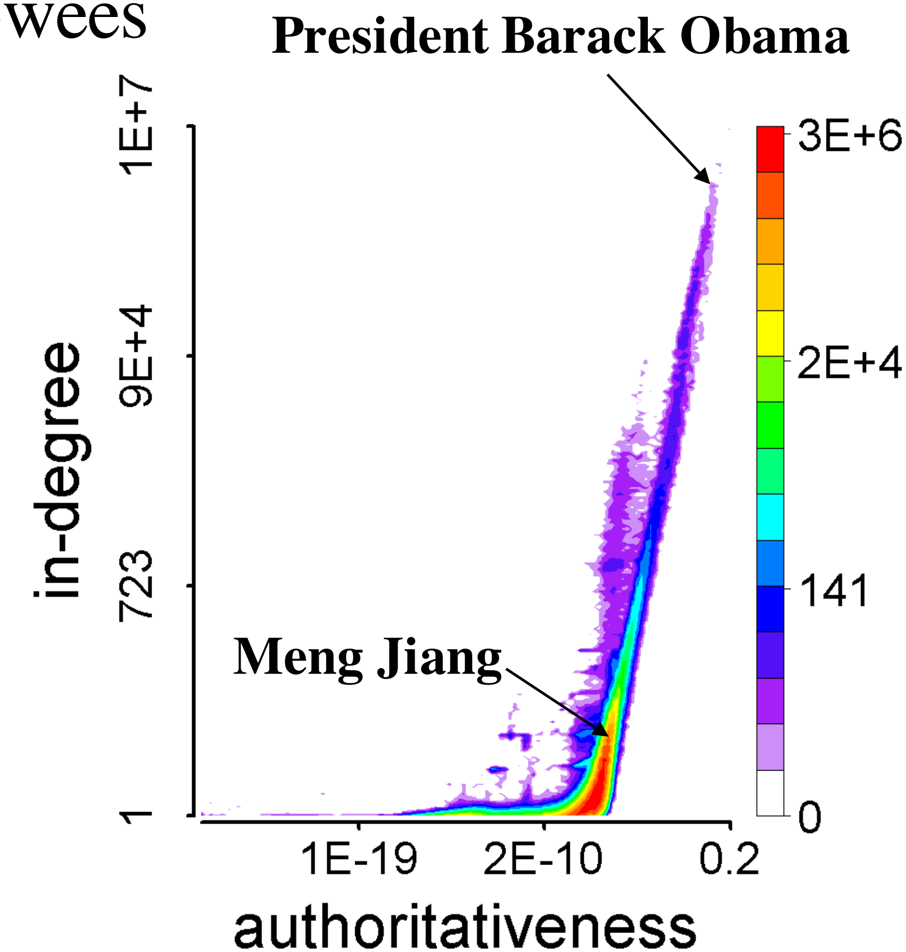


2011 117M



Observation: How They Behave

- Feature space of followees [Kleinberg. JACM]



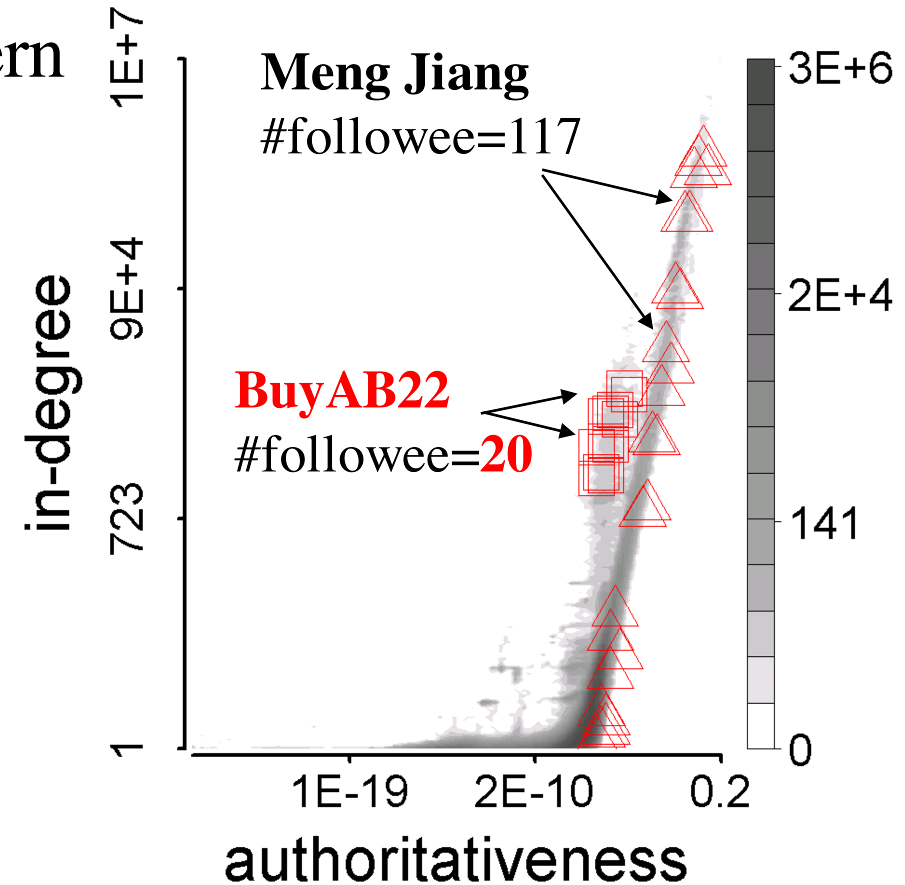
Observation: How They Behave

- Who are their followees?
- Their behavioral pattern
  - **Synchronized**

Similar with each other

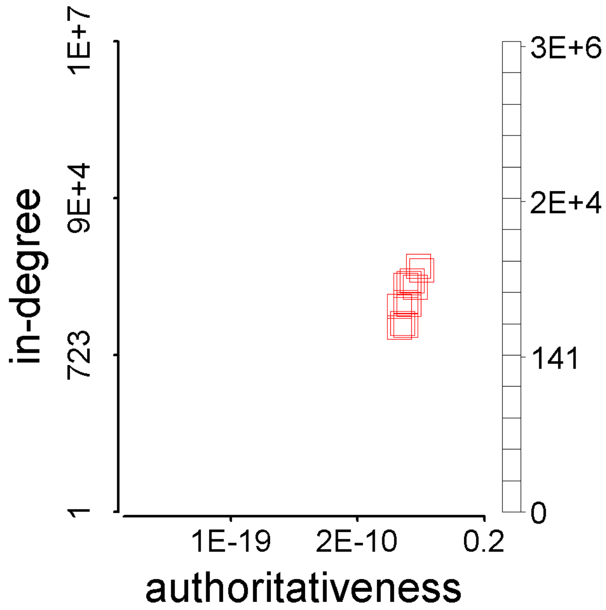
  - **Abnormal**

Different from the majority

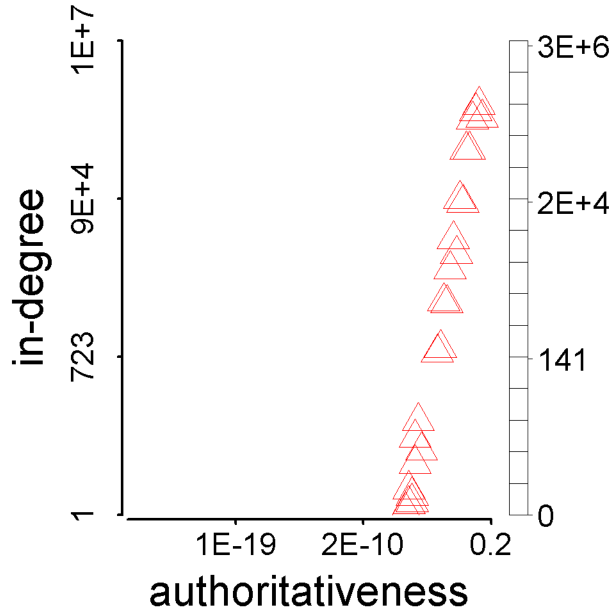


Represent Synchronicity

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

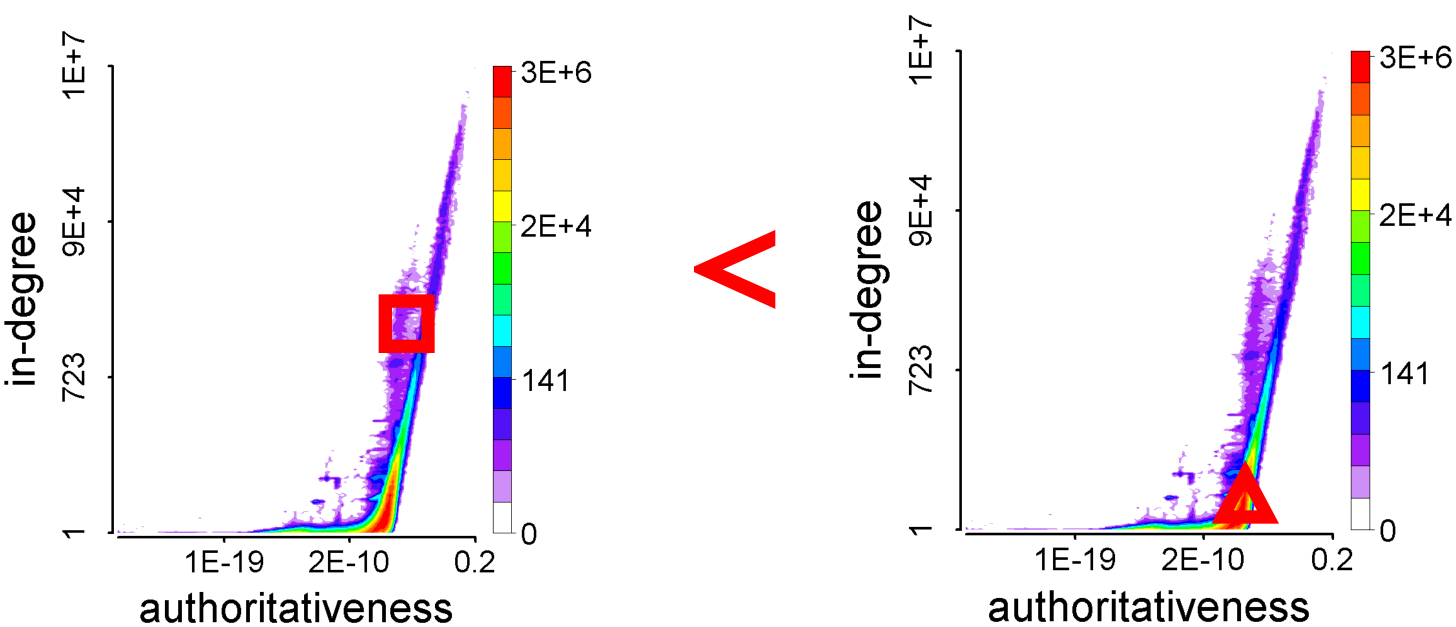


>

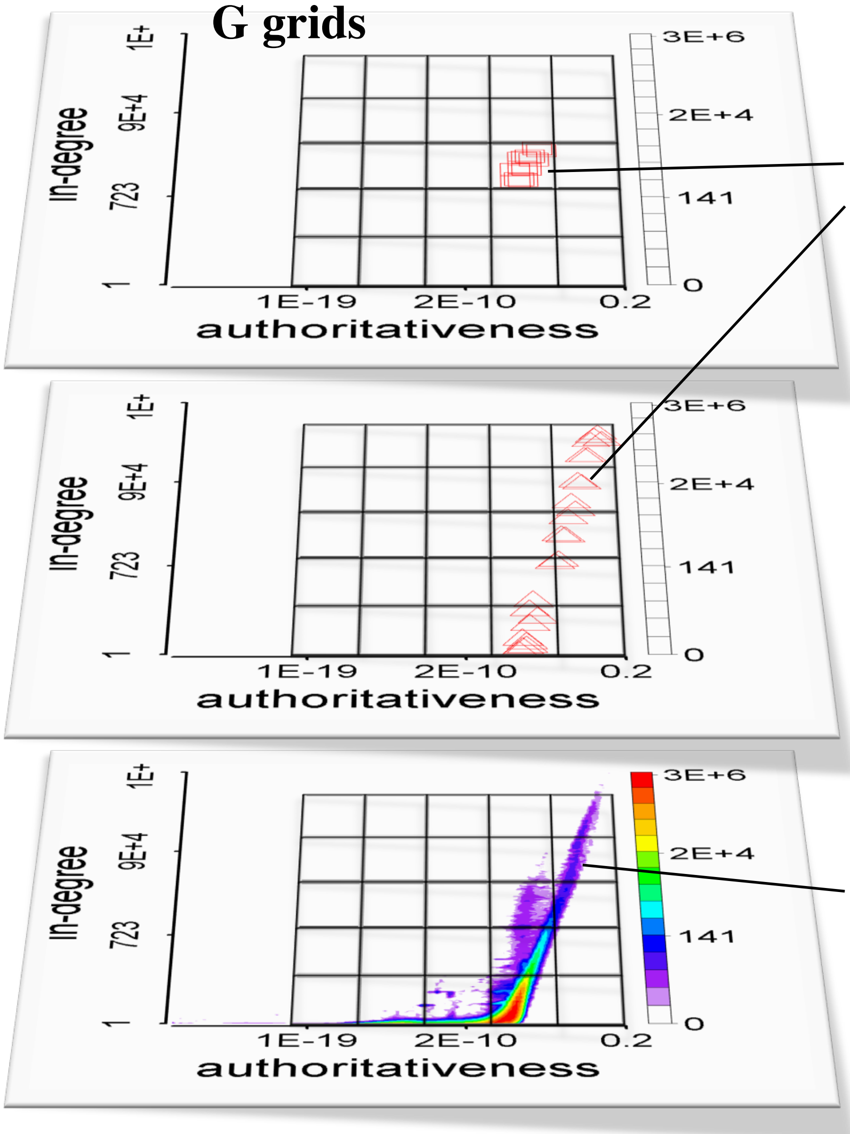


Represent Normality

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



Theorem: Synchronicity vs. Normality



$\mathbf{fp_g}$: #foreground points in grid **g**

$\sum \mathbf{fp_g} = \mathbf{F} = \mathbf{d(u)}$ (#followees of **u**)

Given normality

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

find minimal synchronicity

$$\mathbf{s} = \sum(\mathbf{fp_g}/\mathbf{F})\,(\mathbf{fp_g}/\mathbf{F}) = \sum \mathbf{f_g}^2$$

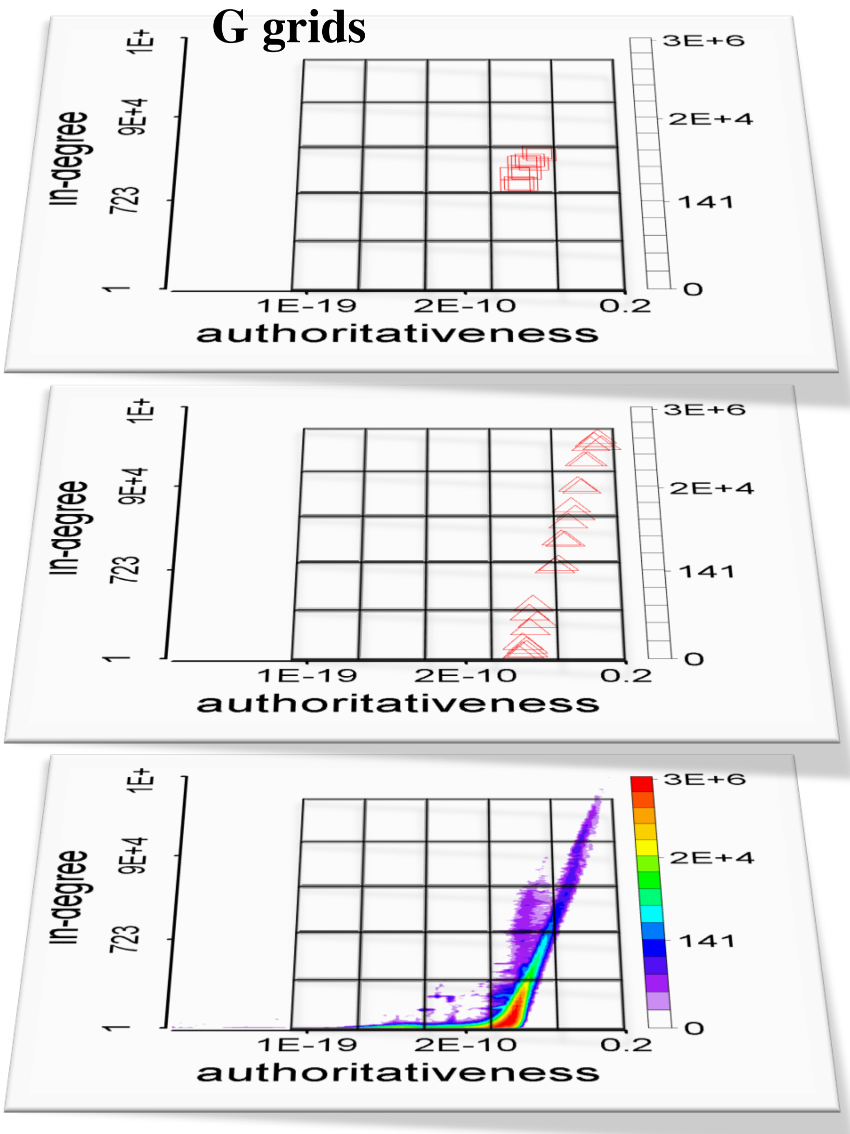
where

$$\sum \mathbf{f_g} = \mathbf{1},\ \sum \mathbf{b_g} = \mathbf{1}$$

$\mathbf{bp_g}$: #background points in grid **g**

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

Theorem: Synchronicity vs. Normality



Solution.

Lagrange multiplier:

minimize $s(f_g) = \sum f_g^{\,2}$

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

Lagrange function:

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

Gradients:

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

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

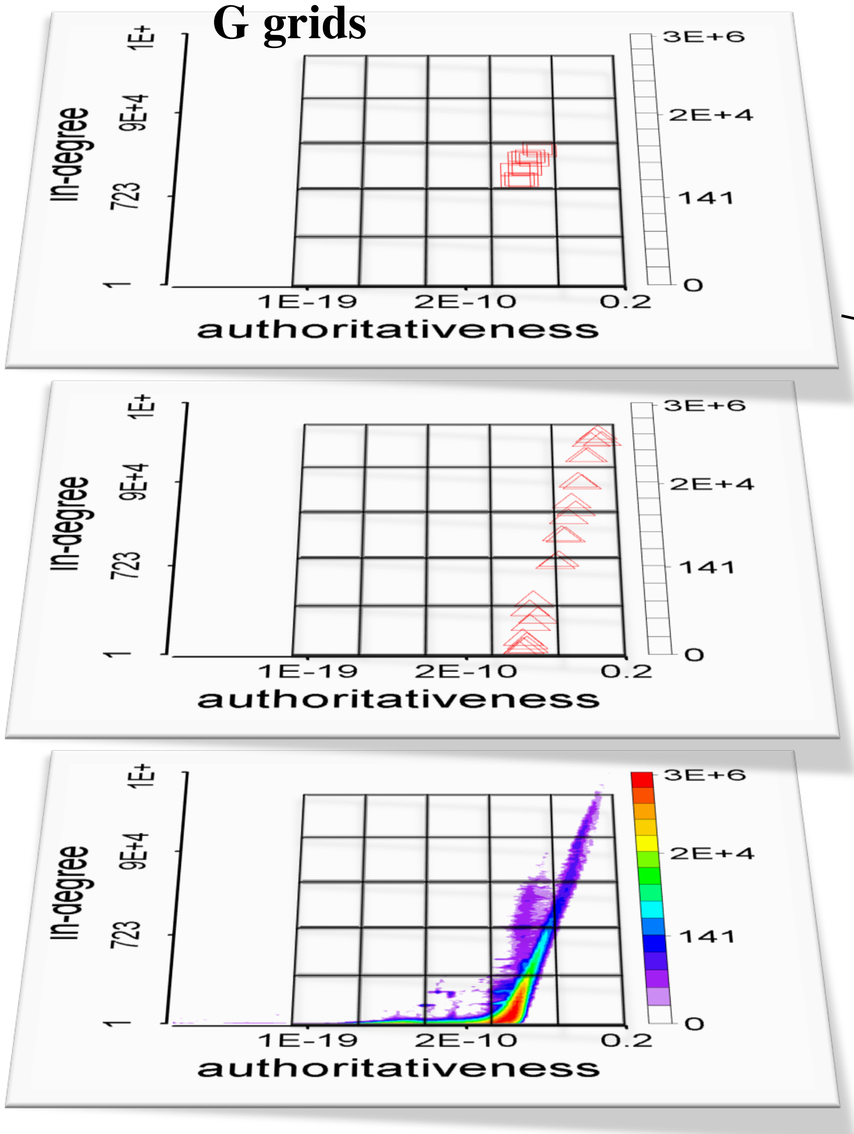
where $s_b = \sum b_g^{\,2}$.

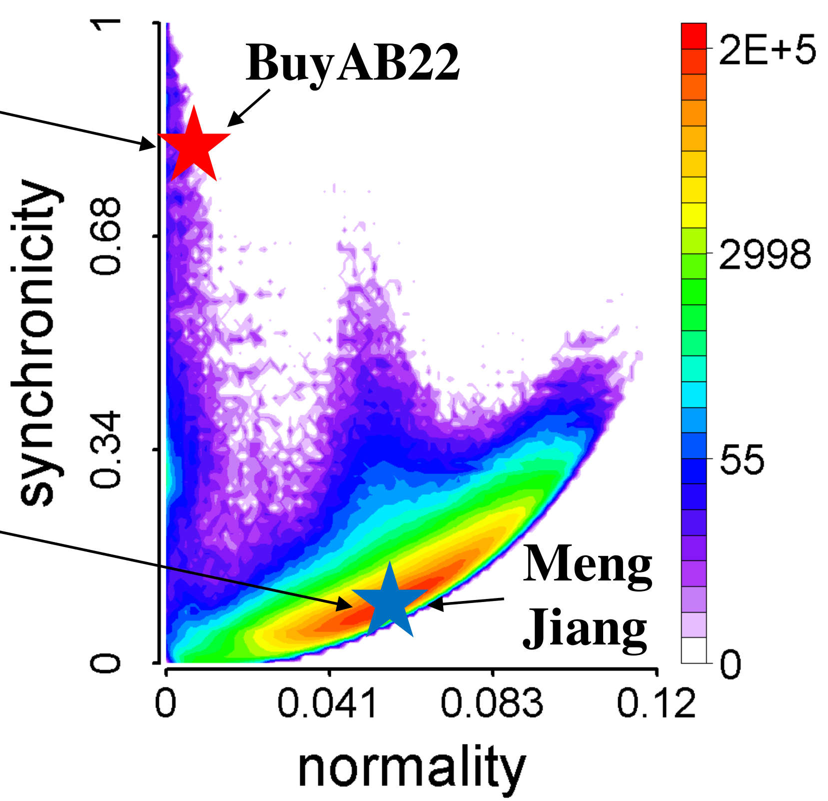
Therefore,

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

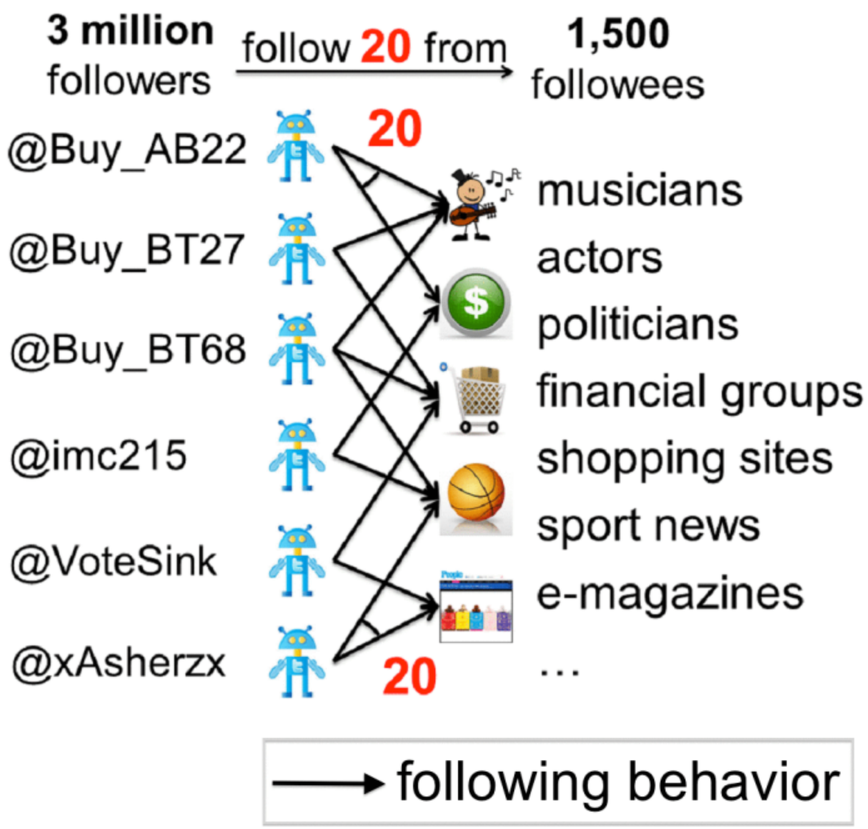
CatchSync

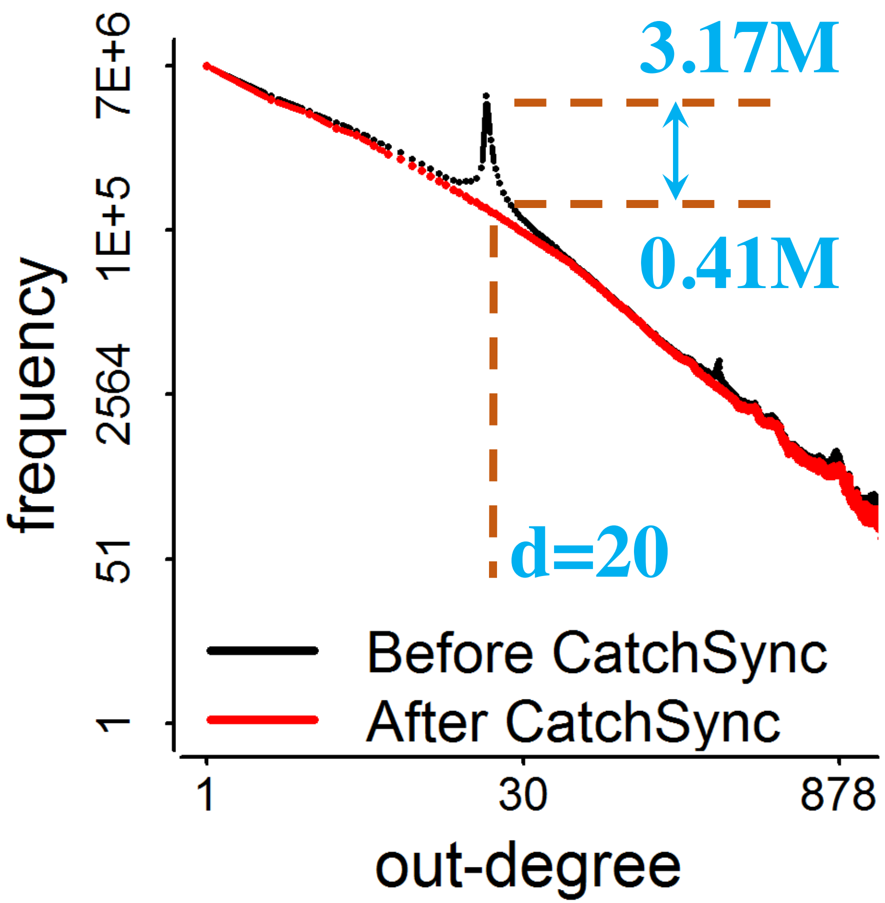
$$s_{min} = \frac{-G\,\mathbf{n}^2 + 2\,\mathbf{n} - s_b}{1 - G\,s_b}$$





Experimental Results





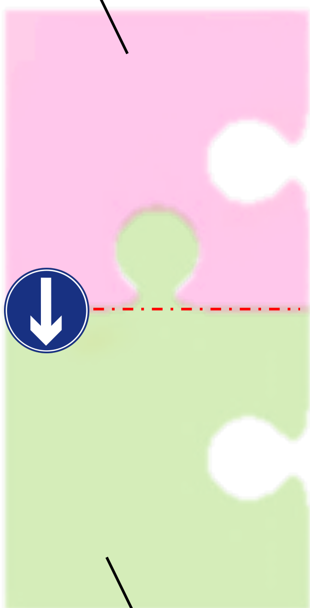
Impact

- **M. Jiang**, P. Cui, A. Beutel, C. Faloutsos and S. Yang. "CatchSync: Catching Synchronized Behavior in Large Directed Graphs" in **KDD'14 Best Paper Finalist**, Aug 2014. (#citations = **36**)
- Taught in
  - **CMU 15-826:** Multimedia Databases and Data Mining
  - **UMich EECS 598:** Graph Mining and Exploration at Scale
  - **ASONAM'16 Tutorial:** "Identifying Malicious Actors on Social Media" by S. Kumar, F. Spezzano, V.S. Subrahmanian
- Deployed in Weibo? Unfortunately, in July 2014…

- Smart enough? First proposed **Camouflage** in PAKDD'14.
  - #citations = **26**
  - Cited by *KDD'16 Best Research Paper*: the authors (B. Hooi *et al*.) provided theoretical bound to prevent the camouflage.

Roadmap

Behavior prediction



Social contexts

Intentions

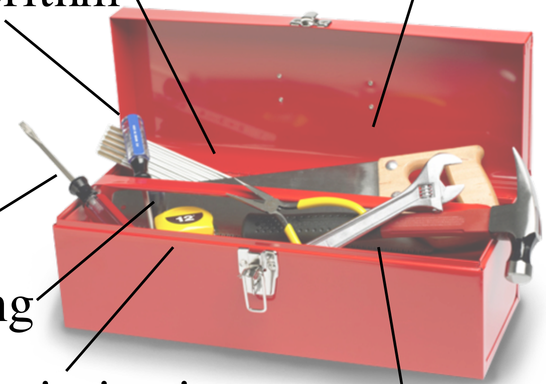
Suspicious behavior detection

Toolbox

Graph-based outlier detection

Community detection

HITS algorithm



SVD

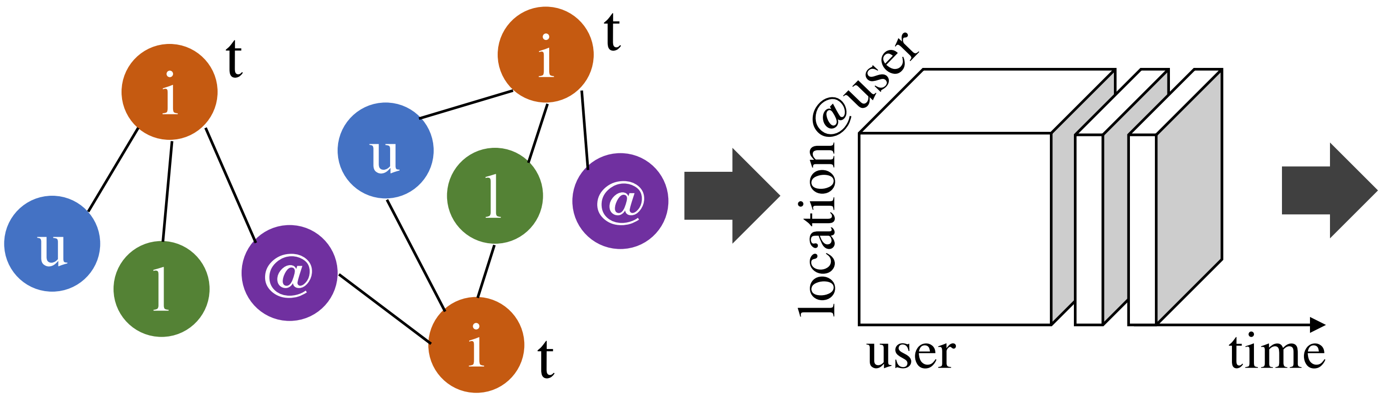
Graph mining

Optimization

Power-law distribution

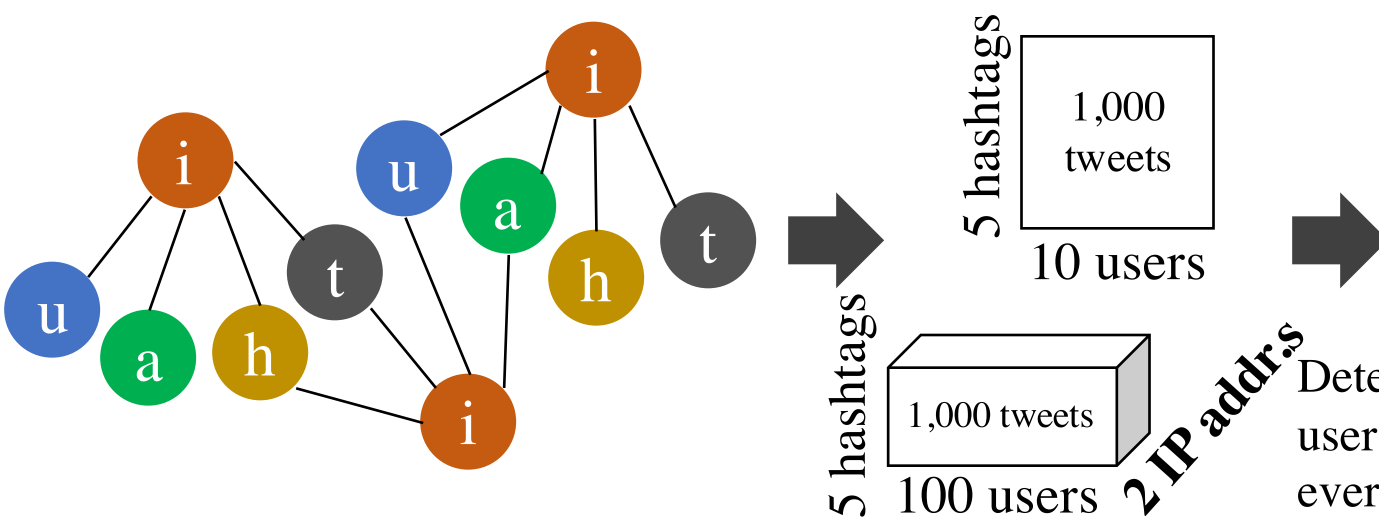
Modeling Spatiotemporal Contexts

- Behavior prediction with multi-dimensional data (KDD'14)



Evolutionary analysis by Incremental Tensor Decomposition

- Suspicious behavior detection across dimensions (ICDM'15)

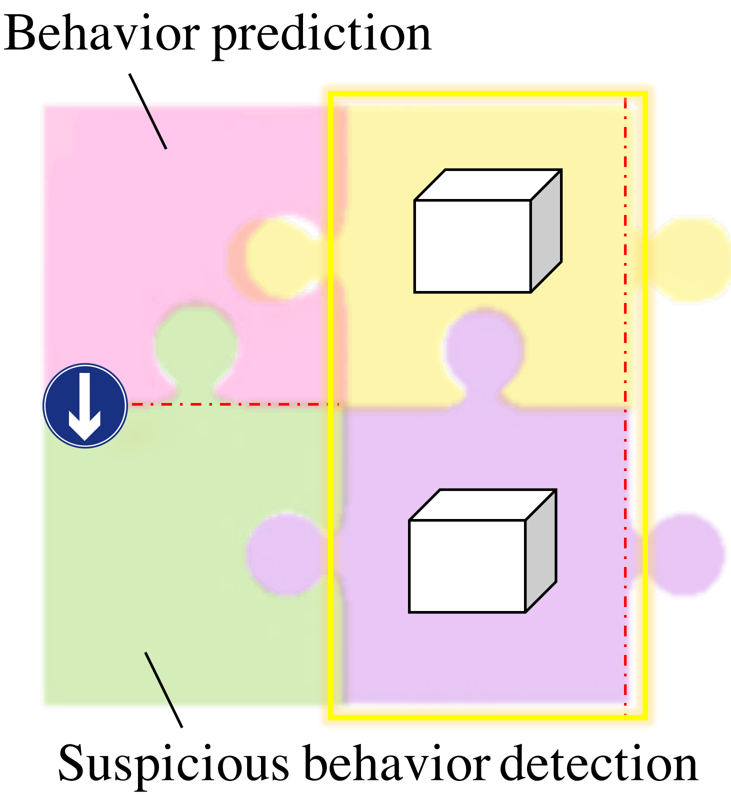


Define a principled suspiciousness metric:

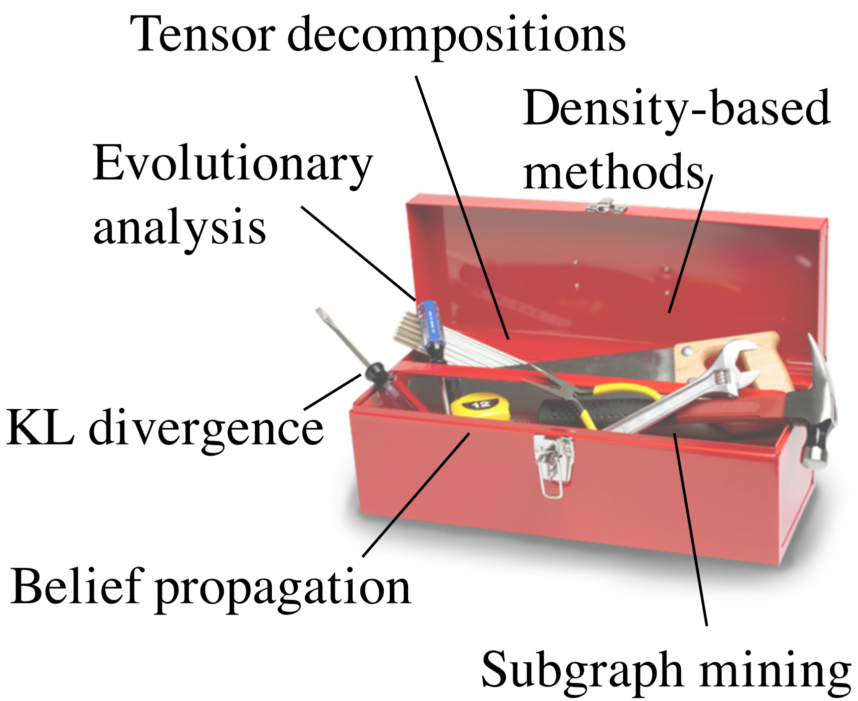
- Dimension
- Size
- Density
- Mass

Detected **5.94M tweets** (582 users, **3 hashtags**, 294 IPs, every minute in 40d)

Roadmap

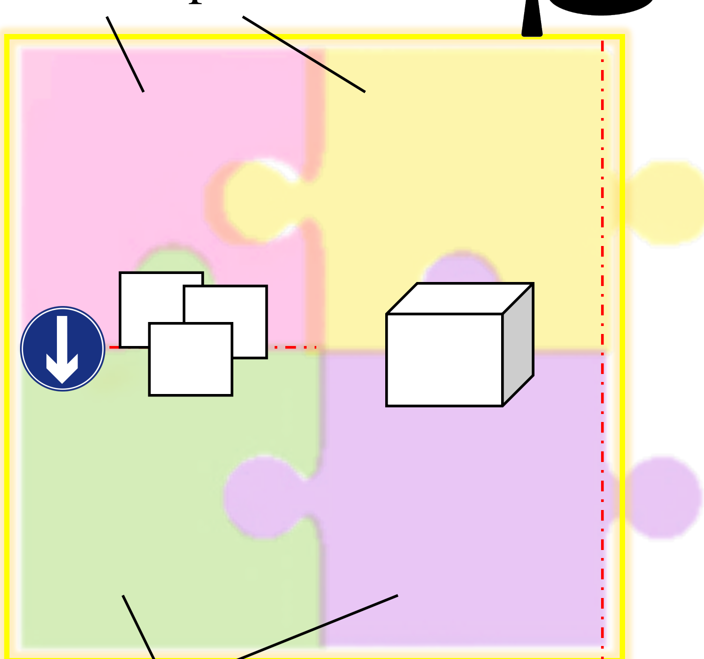


Toolbox



Roadmap Toolbox

Behavior prediction



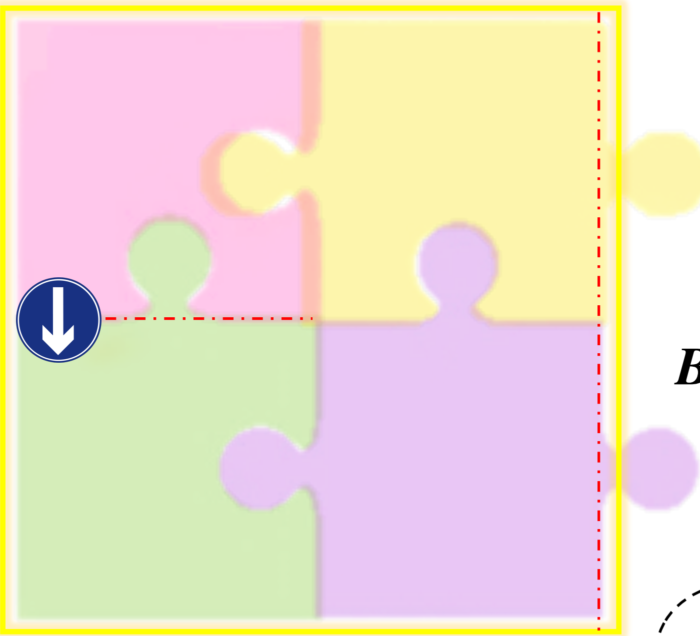
Suspicious behavior detection

Ph. D., Tsinghua University, Beijing (2015)
“Modeling Complex Behaviors in Social Media”
Dissertation Award

Roadmap

Toolbox

T1: Mining behavior networks with social, spatiotemporal contexts



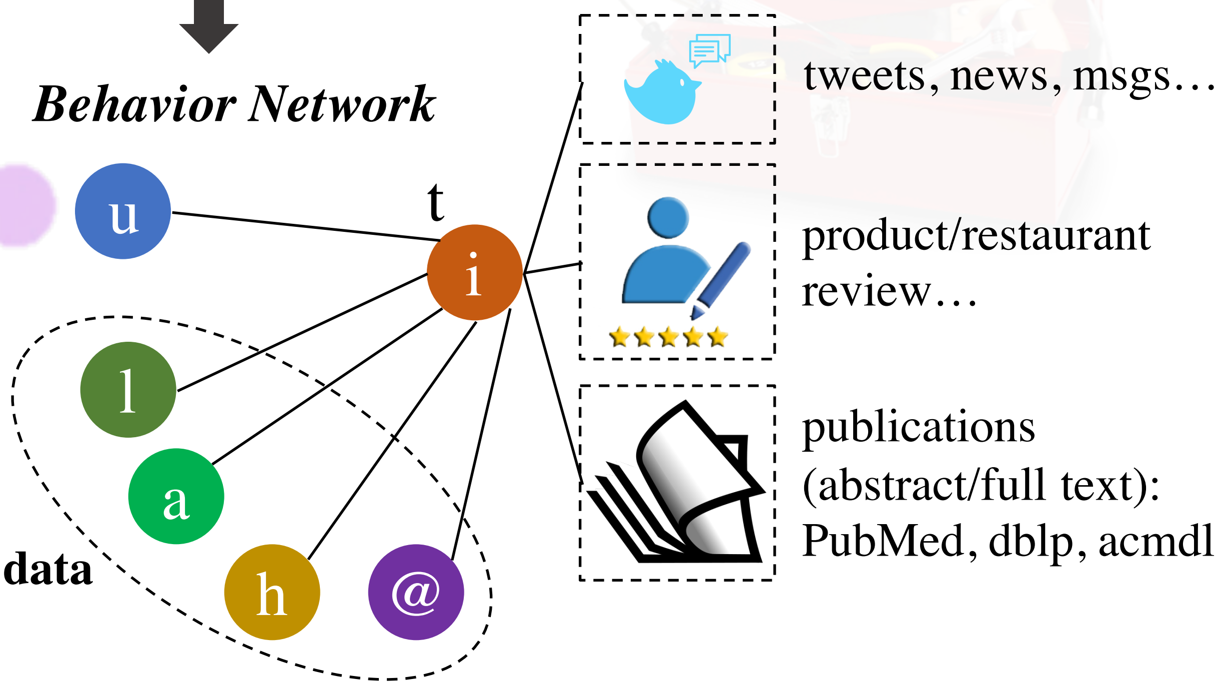
Information Network (entities, attributes, relationships)

Integration

Structuring (text mining)

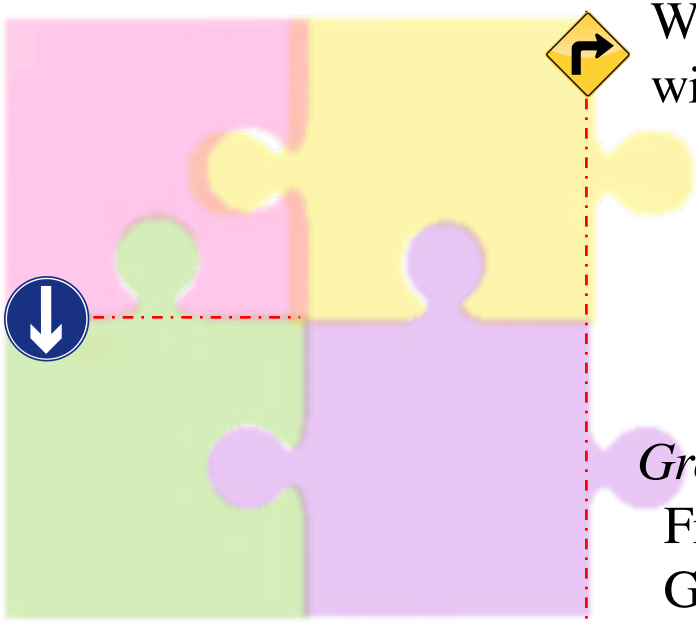
Rich unstructured text data

Behavior Network



Structured data

Roadmap Toolbox



Worked as Postdoctoral Research Associate
with **Prof. Jiawei Han** (UIUC) since Aug 2015

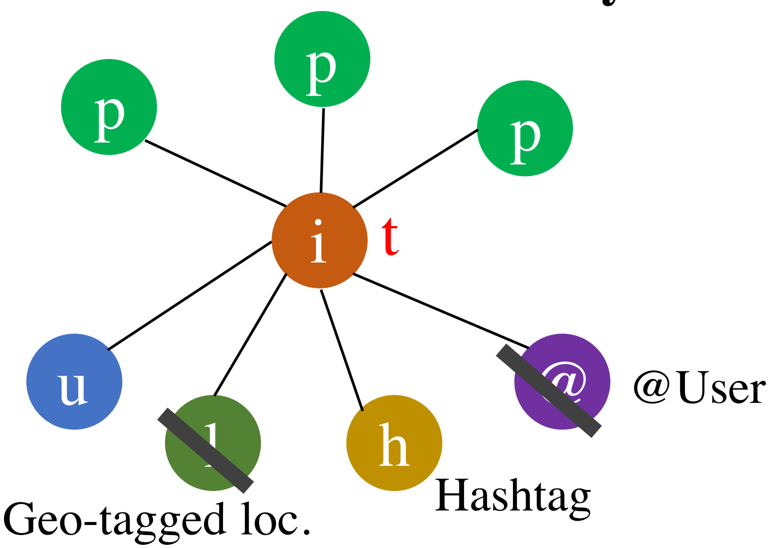
Group's Strength:

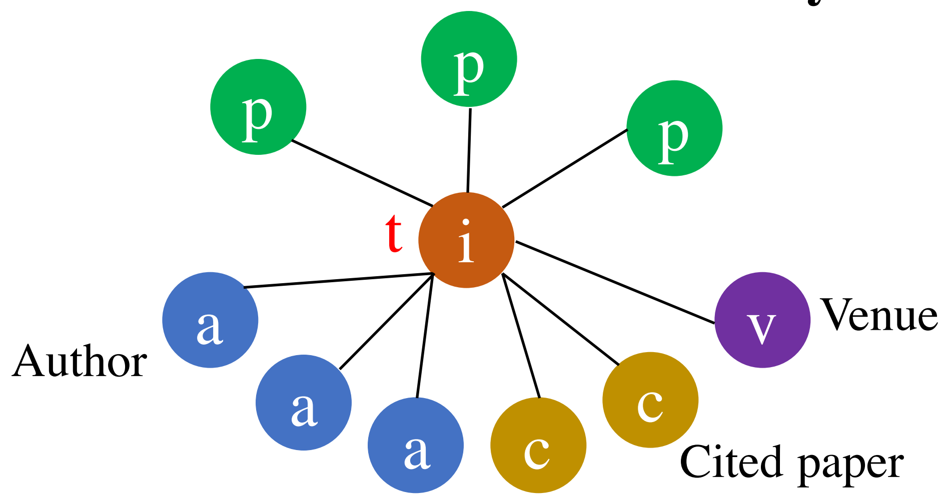
Frequent pattern mining (-2003)
Graph pattern mining (-2007)
Mining heterogeneous information network (-2013)
Automatic Text Mining (2014-):

- **Phrase mining (SegPhrase)** [Liu *et al*. SIGMOD]
- Entity recognition and typing [Ren *et al*.]
- Concept hierarchy discovery [Wang *et al*.; Liu *et al*.]

Bring Phrases to Behavior Modeling

- Tweeting behavior
  - Event **summary**
- Paper-publishing behavior
  - Research trend **summary**



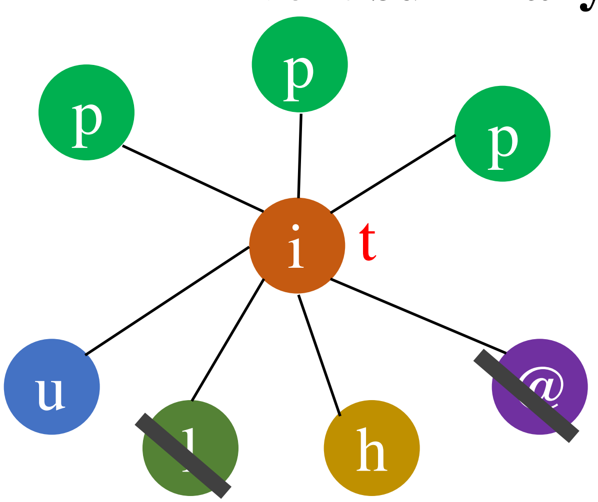


20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl

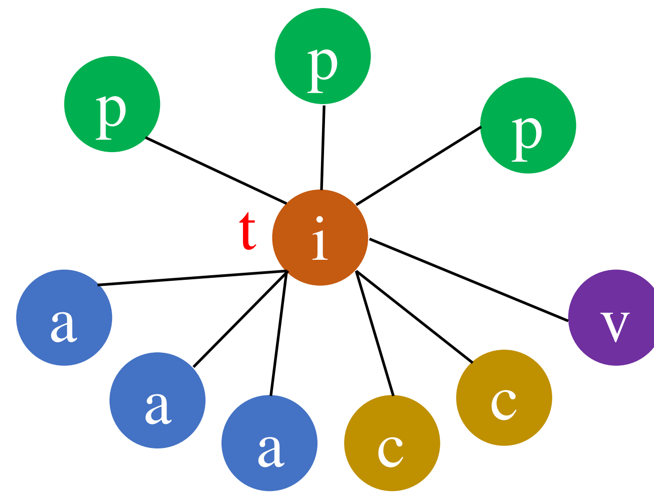
2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Tensor Fails

- Tweeting behavior
  - Event **summary**



- Paper-publishing behavior
  - Research trend **summary**

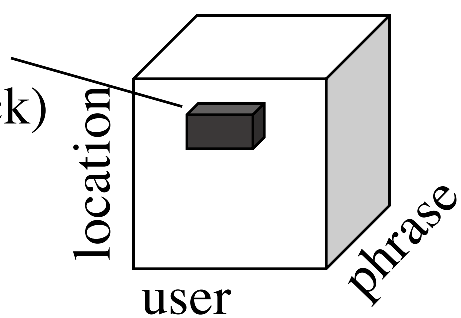


Q: How to represent and summarize **dynamic multi-contextual** behaviors?

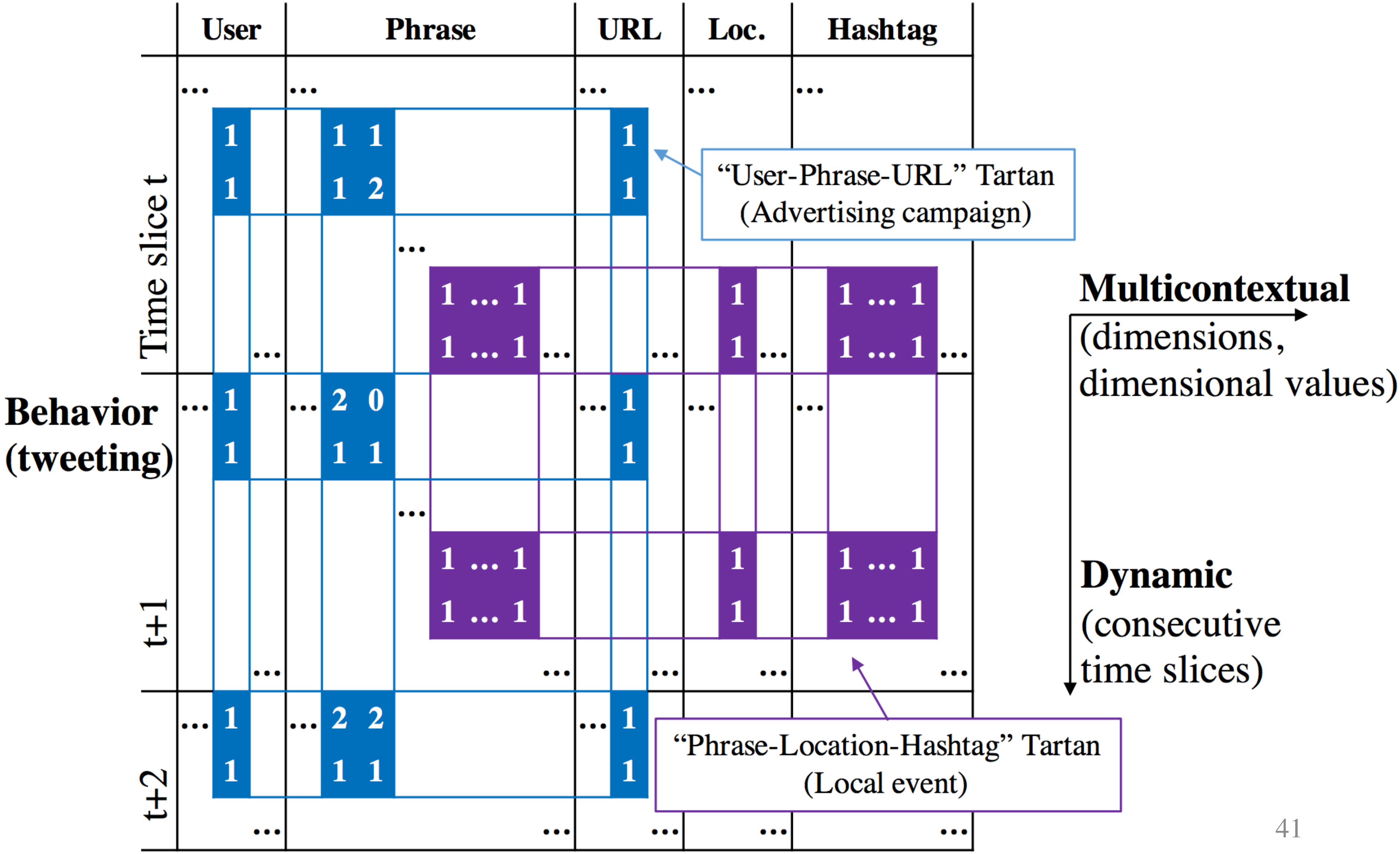
A set of values in dimensions (*one-guaranteed value*, *empty value*, *multi-values*)

✔ ✘ ✘

summary
(dense block)



Two-Level Matrix and "Tartan"

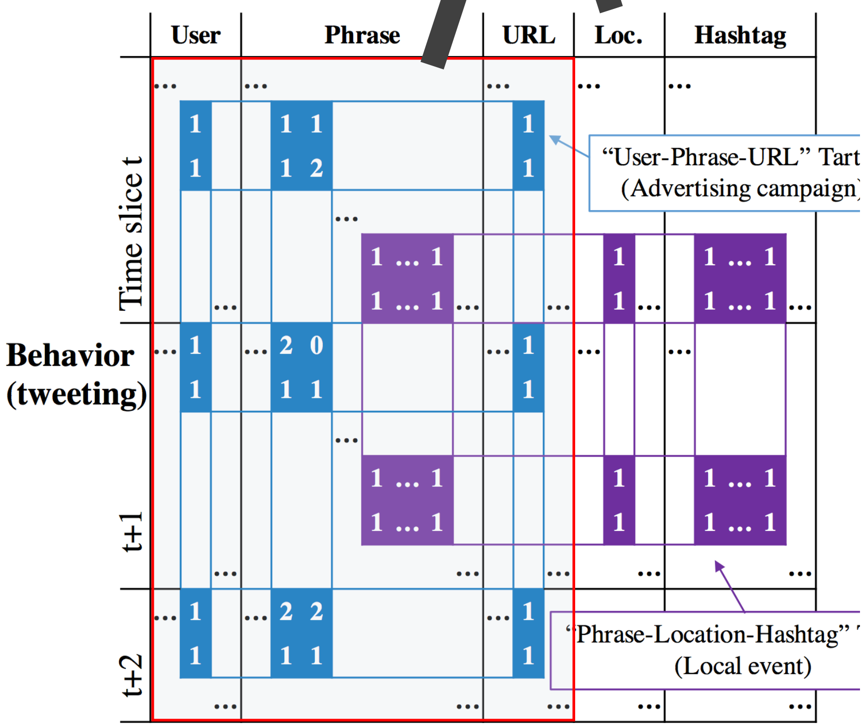


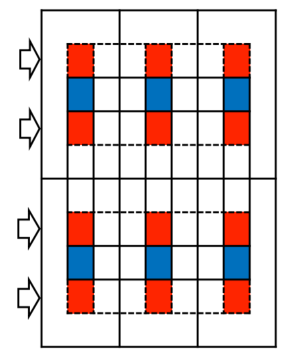
CMU Tartans

Optimize with MDL Principle

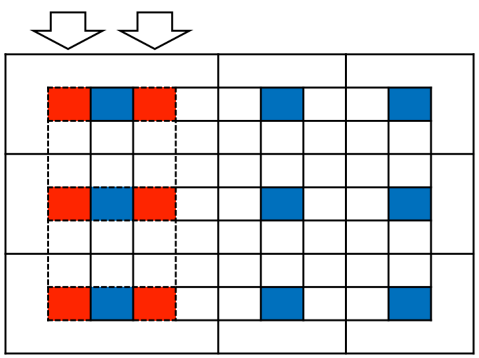
- *Maximize* the number of bits by encoding the Tartan

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

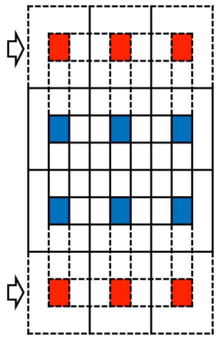




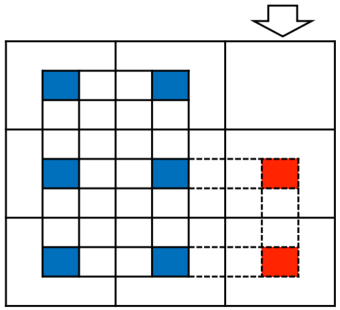
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Experimental Results

- DM/ML research trend summaries with 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

Experimental Results

- Event summaries with Super Bowl 2013 tweets

Time axis: 16:30, 17:00, 17:30, 18:00, 18:30, 19:00, 19:30, 20:00, 20:30, 21:00, 21:30, 22:00

Tweets	Keywords	Dim 1	Dim 2	Dim 3	Dim 4	Size	Tartan
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 (3,325)	**phrase** **226**	hashtag (0)	URL (0)	**3,397 tweets**	**Tartan #1: (1 dim)** 16:30-17:30
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 (196)	**phrase** **4**	**RT @user** **1**	**URL** **1**	**196 tweets**	**Tartan #2: (3 dims)** 17:00-18:00
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 (213)	**phrase** **21**	**RT @user** **3**	URL (0)	**215 tweets**	**Tartan #3: (2 dims)** 18:30-19:30
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 (617)	**phrase** **11**	**RT @user** **4**	**URL** **4**	**617 tweets**	**Tartan #4: (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** **2**	**phrase** **55**	**hashtag** **17**	URL (0)	**166 tweets**	**Tartan #5: (3 dims)** 20:00-21:00
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 (650)	**phrase** **69**	**hashtag** **11**	URL (0)	**653 tweets**	**Tartan #6: (2 dims)** 21:00-22:00
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 (1942)	**phrase** **248**	hashtag (0)	URL (0)	**1,950 tweets**	**Tartan #7: (1 dim)** 22:00-23:30

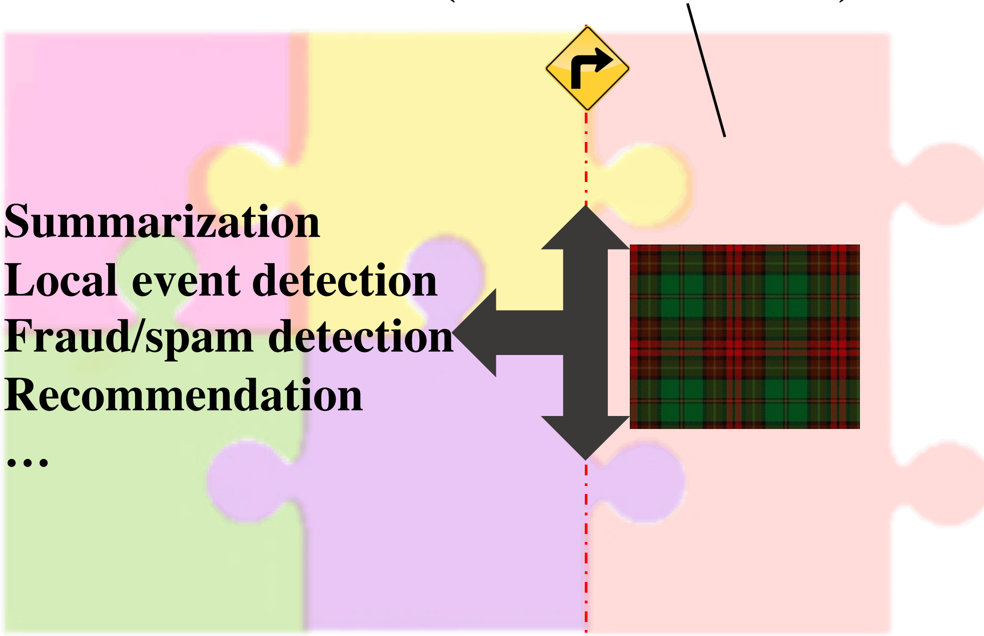
Fun Fact

- **M. Jiang**, C. Faloutsos and J. Han. "CatchTartan: Representing and Summarizing Dynamic Multicontextual Behaviors" in **KDD'16 Oral**. (Acceptance rate = **8.9%**)
- The **1st conference paper** Prof. Jiawei Han and Prof. Christos Faloutsos co-authored, though they have been predicted to be co-authors for long [Sun *et al*. ASONAM'11, WSDM'12, KDD'12].

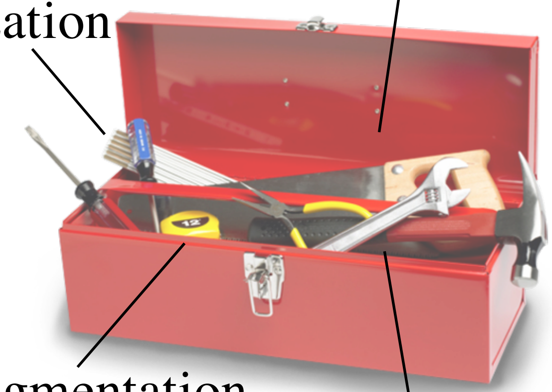
Roadmap

Toolbox

Representing and summarizing multi-dimensional (multi-contextual) data

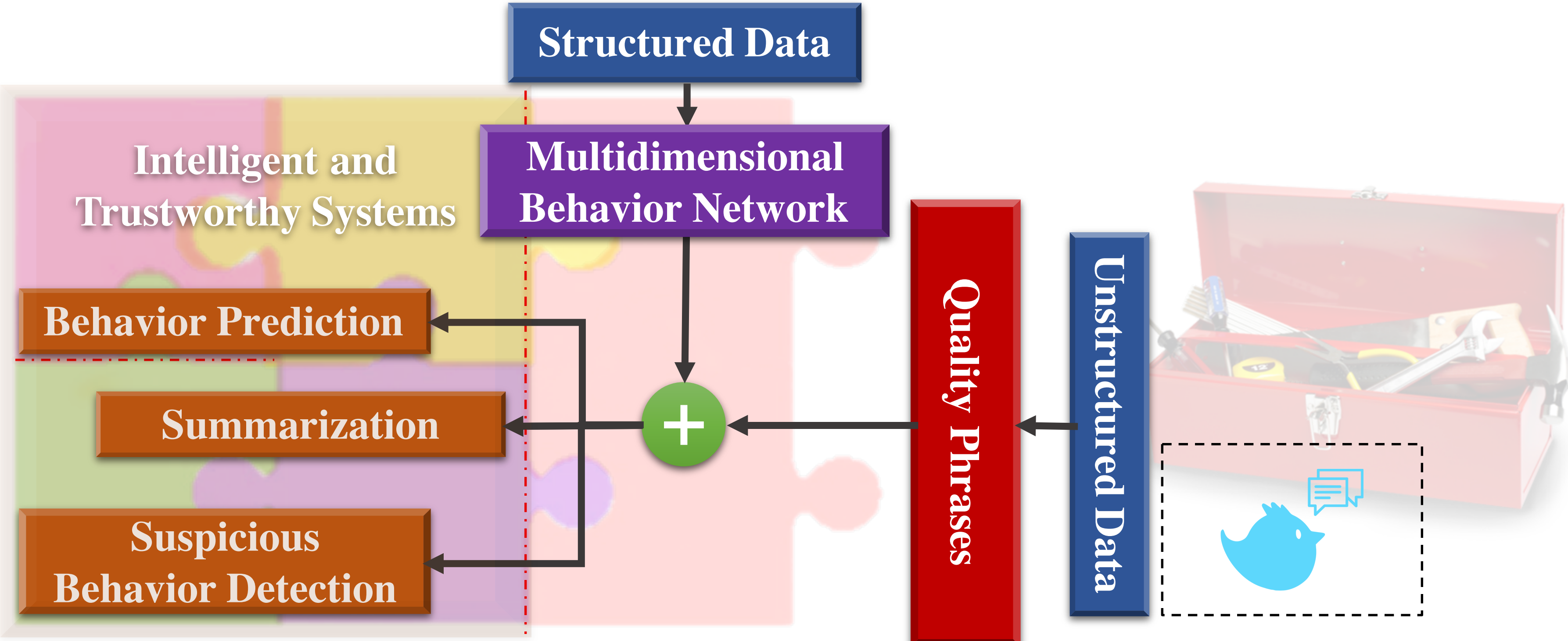


Summarization
Local event detection
Fraud/spam detection
Recommendation
…

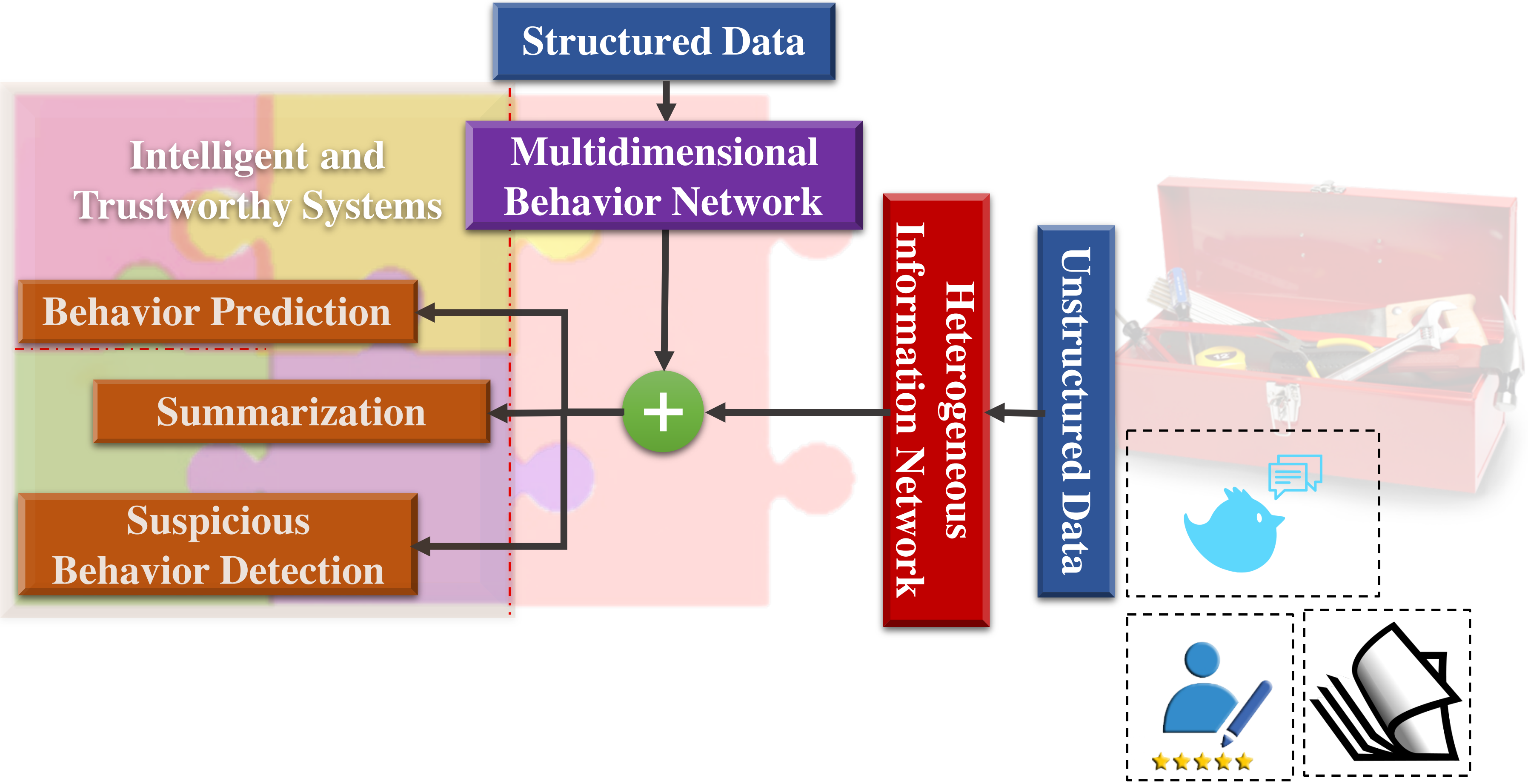


Grant. "**NSF III: Small: Multi-Dimensional Structuring, Summarizing and Mining of Social Media Data**", NSF IIS 16-18481 (**08-01-2016 to 07-31-2019, $500,000**). Jiawei Han, PI.
Wrote 8/15 pages of the proposal in Oct-Nov 2015.
Major supported member.

Roadmap


Structured Data
Multidimensional Behavior Network
Intelligent and Trustworthy Systems
Behavior Prediction
Summarization
Suspicious Behavior Detection
Quality Phrases
Unstructured Data

Roadmap


Structured Data
Multidimensional Behavior Network
Intelligent and Trustworthy Systems
Behavior Prediction
Summarization
Suspicious Behavior Detection
+
Heterogeneous Information Network
Unstructured Data

Roadmap

T3: Integrating behavior networks with rich information networks: Principles and Models

Data-Driven Behavioral Analytics with Networks

T1: Mining behavior networks with social, spatiotemporal contexts

T2: Structuring information networks from behavioral content

Attributed Network Construction

- *Automatic* Attribute discovery: Given a class (*e.g.*, $Country)
  - **Feature** as a characteristic (*e.g.*, “population”)
    - Value: the feature value (*e.g.*, $Digit or NULL)
  - **Relationship** with another class (*e.g.*, “prime minister”)
    - Value: the other class (*e.g.*, $Person.Politician.PrimeMinister)
- Google’s [VLDB’14, WWW’16] based on **fact-seeking** queries
  - Challenge 1: (Class, Attribute name, **Attribute value**)
  - Challenge 2: Just text documents (news, tweets, etc.). **NO query**.

“canada prime minister”, “trudeau age”, “united states president”, “obama age”, “mexico prime minister” …

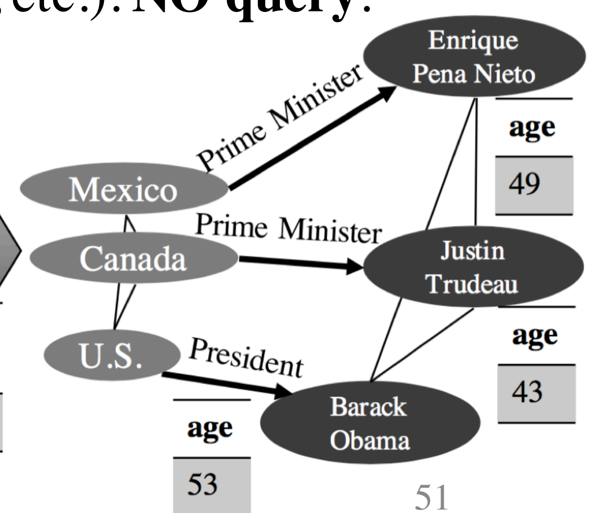
Unfortunately, we don’t have the query data.

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico’s Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…

Fortunately, we have large text corpus.

	Canada	**Mexico**
Prime Minister	Justin Trudeau	Enrique Pena Nieto

	Justin Trudeau	**Enrique Pena Nieto**	**Barack Obama**
age	43	49	53



Data-Driven: Meta Pattern Mining

- **Meta Pattern:** a sequence of *class symbols*, *words*, *phrases* and *punctuation marks* that appear contiguously in the text, and serves as a whole semantic unit.

News:

Text	Meta patterns
…he's gotten older and grayer, and he's been eclipsed at an Asian economic forum here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie… He's also the youngest leader at the Asia Pacific Economic Cooperation forum, six years the junior of **Mexico's Enrique Pena Nieto, 49**, … **Obama, 53**, who becomes the elder statesman…	**1. $Person, $Digit,** **2. $Location.Country Prime_Minister $Person.Politician.PrimeMinister**

Tweets:

Text	Meta patterns
…Protestors march to **Gordon Square** for **12** -year-old **Tamir Rice**…	**1. protestors march to $Location.Square** **2. $Digit -year-old $Person.Victim**

PubMed abstract:

Text	Meta patterns
… **Endocarditis** caused by **Streptococcus pneumoniae**… **Pericarditis** due to **Neisseria meningitidis** …	**$Cardiovasular_Diseases caused by $Bacteria** **$Cardiovasular_Diseases due to $Bacteria**

MetaPAD Framework

Integrated Data-Driven Text Mining

Meta Pattern Mining

Attribute Extraction from Meta Patterns

… Canada Prime Minister Justin Trudeau …
… Barack Obama , 53, …

Quality phrase mining (SegPhrase, SIGMOD'15)

… Canada **Prime_Minister Justin_Trudeau** …
… **Barack_Obama** , 53, …

Entity recognition and typing with distant supervision (ClusType, KDD'15)

… **$Location** Prime_Minister **$Person** …
… **$Person** , **$Digit** , …

Fine-grained typing (PLE, KDD'16)

… **$Country** Prime_Minister **$PrimeMinister** …
… **$President**, **$Digit** , …

MetaPAD Framework

Integrated Data-Driven Text Mining

↓

Meta Pattern Mining

↓

Attribute Extraction from Meta Patterns

Quality Meta-Pattern Classifier

- **Frequency**

  "prime_minister $PrimeMinister" vs "young $PrimeMinister"

- **Completeness**

  "$Country prime_minister $PrimeMinister" vs "$Country prime_minister"

- **Informativeness**

  "$Person 's brother , $Person ," vs "$Person and $Person"

- **Coverage**

  "$Person 's signature healthcare law": only "Barack Obama"

- **Classifier: Random forest**

MetaPAD Framework

Integrated Data-Driven Text Mining

Meta Pattern Mining

Attribute Extraction from Meta Patterns

…xxx **$Country** Prime_Minister **$PrimeMinister** xxx…
…xxx **$President** , **$Digit** , xxx…

Quality Meta-Pattern Classifier

$Location Prime_Minister **$Person**
$Person, **$Digit** ,
$Country Prime_Minister **$PrimeMinister**
$President , **$Digit** ,

Synonym Meta-Pattern Detection
(1) Shared instances (2) J.W. similar words

$Location Prime_Minister **$Person**
$Location PM **$Person**
Prime_Minister **$Person** of **$Location**

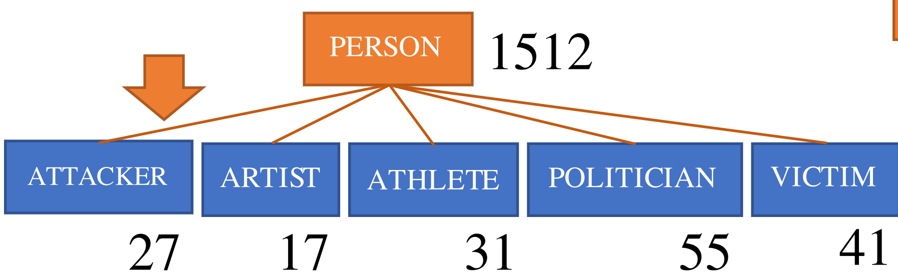
$Person , **$Digit** ,
$Person , a **$Digit** -year-old
$Person , age **$Digit**

Re-typing for Appropriate Granularity

$Country Prime_Minister **$PrimeMinister**
$Person , **$Digit** ,

Top-Down Re-Typing for Granularity

_____, a $Digit -year-old

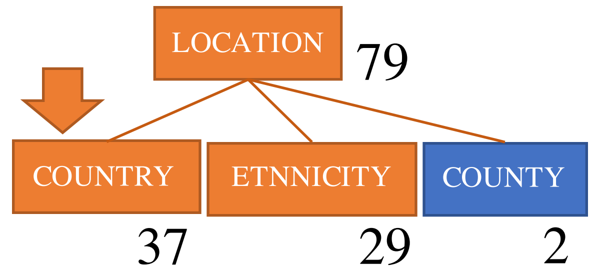


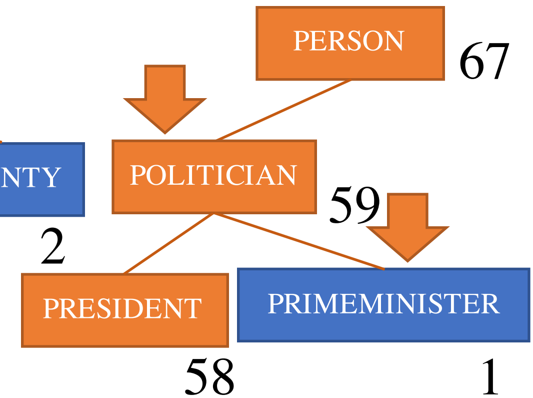
Graininess:

α=(27+17+…+41)/1512

small (< 0.8), stop going down

_____ President _____

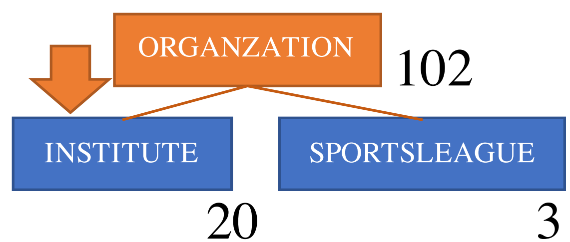


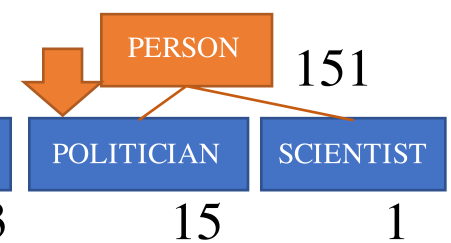


Support:

β=29/max(37,29,2)

big (>0.1), keep $Ethnicity





Similar for Bottom-Up…

Experimental Results

Class=$PERSON (METAPAD: **10,361** names, **4,839** pairs)			Class=$COUNTRY (METAPAD: **1,132** names, **3,930** pairs)		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**	**Name**:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
Mr.	Dr.	(-year-old,$DIGIT)	president	president	(ambassador,$COUNTRY)
Dr.	Mr.	(president,$ORGANIZATION)	people	government	(president,$PRESIDENT)
president	president	(spokesman,$ORGANIZATION)	government	war	(visit,$PERSON)
wife	director	(director,$ORGANIZATION)	capital	border	(dead,$DIGIT)
-year-old	spokesman	(wife,$PERSON)	visit	volcano	(prime minister,$PRIMEMINISTER)
death	chief	(chairman,$ORGANIZATION)	economy	sanctions	(senator,$SENATOR)
coach	professor	(governor,$USSTATE)	prime minister	ambassador	(embassy,$COUNTRY)
love	head	(spokeswoman,$ORGANIZATION)	part	earthquake	(condemn,$ORGANIZATION)
son	coach	(leader,$ORGANIZATION)	leaders	capital	(district judge,$PERSON)
...	...	...	...	...	...
code case homicide	staff sergeant	(told reporters,$WEEKDAY)	nuclear dossier	volcano eruption	(protests,$NEWSAGENCY)
snow pants	army chief	(board member,$ORGANIZATION)	similar box	security	(-magnitude earthquake,$DIGIT)
fellow director	basketball coach	(hack,$COMPANY)	episcopal oversight	parliament	(second biggest,$ORGANIZATION)
Class=$INSTITUTE (METAPAD: 402 names, 198 pairs)			**Class=$BASKETBALLPLAYER (METAPAD: 58 names, 40 pairs)**		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**	**Name**:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
professor	professor	(professor,$PERSON)	guard	forward	(points,$DIGIT)
students	students	(law professor,$PERSON)	star	points guard	(center,$TEAMNAME)
president	graduate	(political science professor,$PERSON)	game	game	(freshman,$SPORTSLEAGUE)
campus	law professor	(student,$PERSON)	forward	freshman	(forward,$TEAMNAME)
law professor	campus	(grad,$PERSON)	career	center	(point guard,$TEAMNAME)
graduate	degree	(signee,$PERSON)	teammate	get better	(all-star,$SPORTSLEAGUE)
director	dean	(economics professor,$PERSON)	point guard	basketball player	(games,$DIGIT)
study	faculty	(basketball coach,$PERSON)	points	full highlights	(rebounds,$DIGIT)
researchers	expert	(finance professor,$PERSON)	season	jumper	(ast,$DIGIT)
...	...	...	...	...	...
foul	commitment	(class,$YEAR)	understudy	retirement	(PG,$TEAMNAME)
socialism speech	dorm	(superintendent,$PERSON)	birthday boy	shoes	(career earnings,$DIGIT $DIGITUNIT)
good summary	program	(-year-old student,$DIGIT $PERSON)	injury meme	suspended without pay	(sue,$PERSON)

Experimental Results

Class=$Location; Value Type=$Month,$Day,$Year		
#	Meta Patterns	
1	$Location $Month $Day, $Year	
2	$Country, $Weekday, $Month $Day, $Year	
3	$Location on $Month $Day, $Year	
#	Entity	Attribute Value
1	Pearl Harbor	December 7, 1941
2	Green Bay	Sunday, Jan 11, 2015
3	Malta[1]	Friday, Nov 27, 2015
...	...	...
5862	Beijing[2]	October 11, 2013
5863	Finland[3]	April 8, 2015

Class=$Organization; Name="ceo"		
#	Meta Patterns	
1	$Organization CEO $Person	
2	$Company CEO $Businessperson	
3	$Organization's $Person	
#	Entity	Attribute Value
1	Apple	Tim Cook
2	Facebook	Mark Zuckerberg
3	Hewlett-Packard	Carly Fiorina
...	...	...
765	Boston Medical Center	Kate Walsh
766	Association of Private Sector Colleges and Universities	Steve Gunderson

Class=$Person; Name="-year-old"[7]		
#	Meta Patterns	
1	$Digit-year-old $Person	
2	$Person, $Digit,	
3	$Person, a $Digit-year-old	
#	Entity	Attribute Value
1	Tamir Rice	12
2	Bobbi Kristina Brown	21
3	Michael Brown	18
...	...	...
4993	Jay Nixon	58
4994	Xanana Gusmao	68

Class=$Person; Name="president"; Value Type=$Ethnicity		
#	Meta Patterns	
1	$Ethnicity President $President	
2	$Ethnicity leader $President	
3	$Ethnicity government of President $President	
#	Entity	Attribute Value
1	Vladimir Putin	Russian
2	Francois Hollande	French
3	Raul Castro	Cuban
...	...	...
254	Mohammed Morsi	Egyptian
255	Klaus Iohannis	Romanian

[1]Commonwealth Heads of Government Meeting. [2]UCI World Tour of Beijing. [3]Finnish parliamentary election begins.

Experimental Results

F1 score	WPB ('10, 100M)	CNA ('97-'10, 200M)	APR ('15, 200M)	TWT ('15, 1GB)
Total (vs Biperpedia -q)	↑67.7%	↑48.3%	↑189.5%	↑208.0%
w/ Meta pattern classifier	↑30.1%	↑27.0%	↑127.1%	↑195.6%
w/ Granularity	↑20.8%	↑15.6%	↑17.3%	↑3.1%
w/ Integrated text mining techs	↑13.8%	↑9.3%	↑13.0%	↑0.8%

$Cardiovasular_Diseases due to $Bacteria
$Cardiovasular_Diseases caused by $Bacteria

$Bacteria	$Cardiovascular_Diseases
Streptococcus pneumoniae	Endocarditis
Neisseria meningitidis	Pericarditis
Haemophilus paraphrophilus	Endocarditis
Proteus	Endocarditis
Listeria monocytogenes	Pericarditis
Corynebacterium	Endocarditis
Actinomyces	Endocarditis
Coxiella	Endocarditis
Pasteurella pneumotropica	Endocarditis
Cardiobacterium	Endocarditis

$Enzymes_and_Coenzymes inhibitor $Chemical

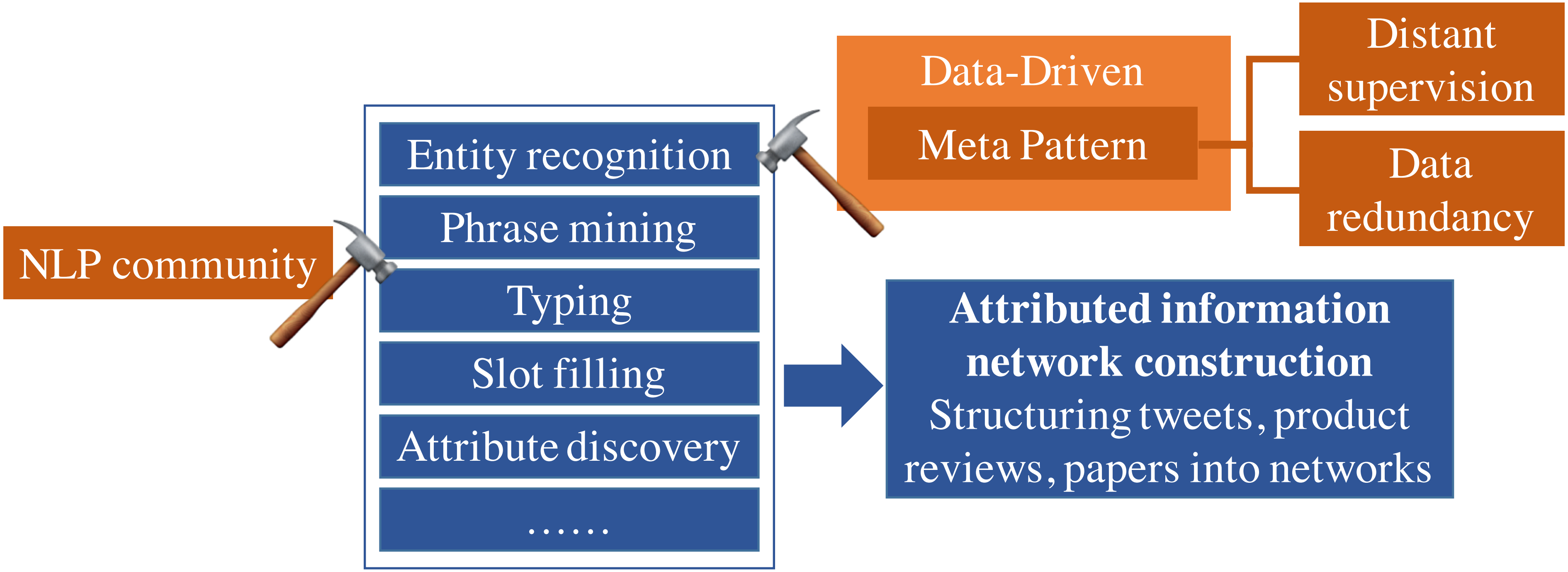
$Chemical	$Enzymes_and_Coenzymes
chelerythrine	protein kinase C
fondaparinux	Factor Xa
calphostin C	protein kinase C
bisindolylmaleimide	protein kinase C

$Diagnosis : $Digit +/- $Digit kg/m ($Digit)

$Diagnosis	$Digit $Digit $Digit
BMI	(31.0 , 6.4 , 2)
BMI	(26 , 4 , 2)
body mass index	(27 , 6 , 2)

Submission and Insights

- **M. Jiang**, J. Shang, T. Cassidy (ARL-ALC), L. M. Kaplan (ARL-ALC), T. P. Hanratty (ARL-APG), J. Han. "MetaPAD: Meta Pattern-driven Attribute Discovery in Massive Text Corpora". Submitted to *WSDM 2017*.

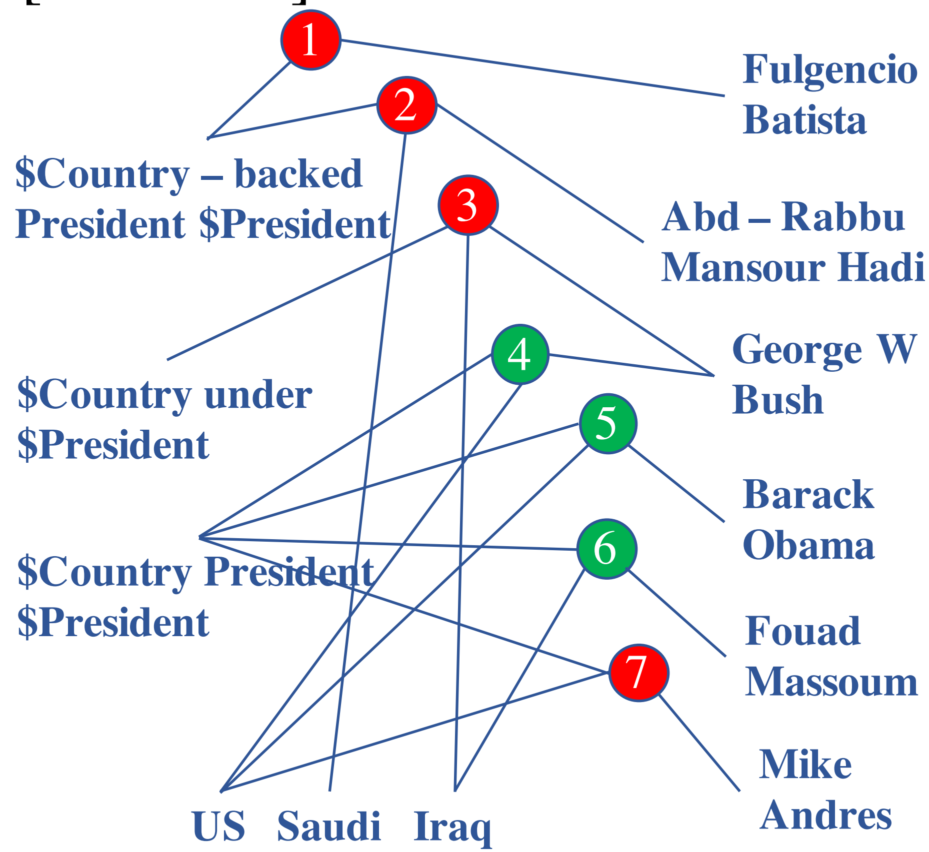


Finding Truth when Structuring

- Modeling "source" reliability [Gao *et al*.]

($Country, president, $President) : 0.829

Meta Pattern	Acc. (FP/P)
$Country 's President $President	0.984 (1/61)
President $President of $Country	1.000 (0/24)
$Country 's President $President ,	1.000 (0/16)
" $Country President $President	1.000 (0/7)
…	…
President $President said $Country	0.833 (1/6)
$Country President $President	0.807 (16/83)
$Country , President $President	0.650 (7/20)
$Country - backed President $President	0.500 (3/6)
$Country under $President	0.500 (1/2)

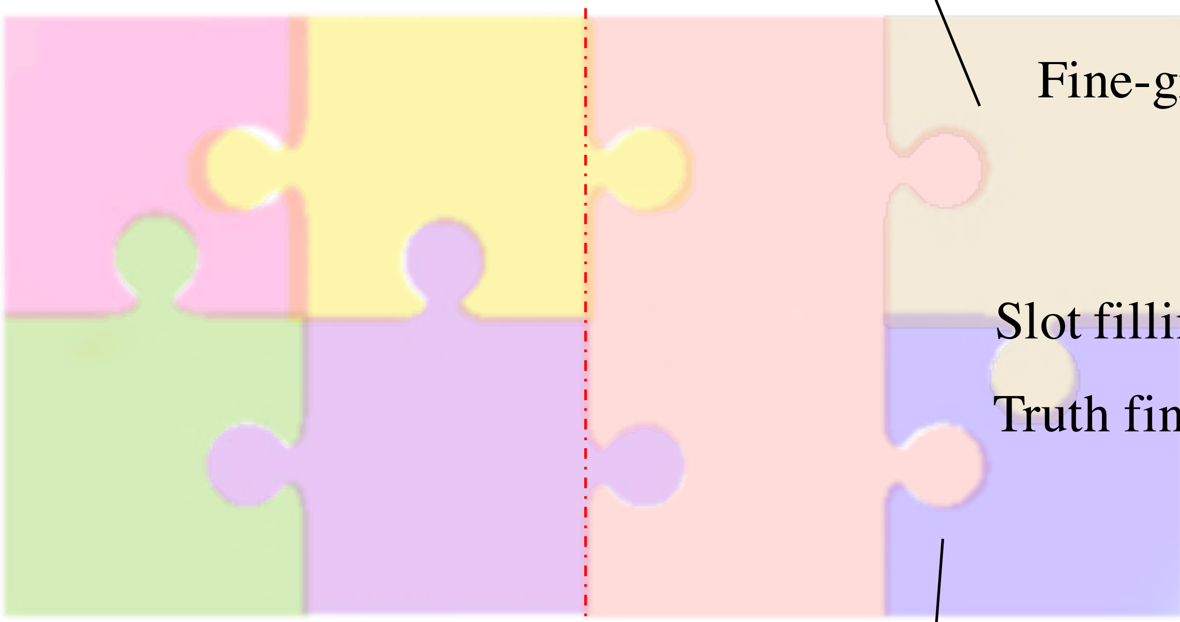


1. …Fidel Castro and his brother Raul led winning a revolution toppling ***US - backed President Fulgencio Batista*** .
2. …control of the country and at reinstating ***Saudi - backed President Abd - Rabbu Mansour Hadi*** .
3. …was profoundly forward - leaning and outspoken about the importance of invading ***Iraq under George W Bush*** .
7. …better delivering on those expectations , " McDonald 's ***US President Mike Andres*** said in the announcement .

Roadmap

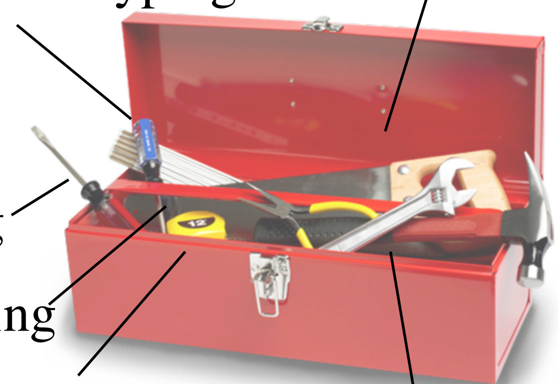
Toolbox

Automatic Attributed Information Network Construction with Meta Pattern Mining



Fine-grained typing

Meta Patten



Slot filling

Truth finding

Phrase mining

Entity recognition

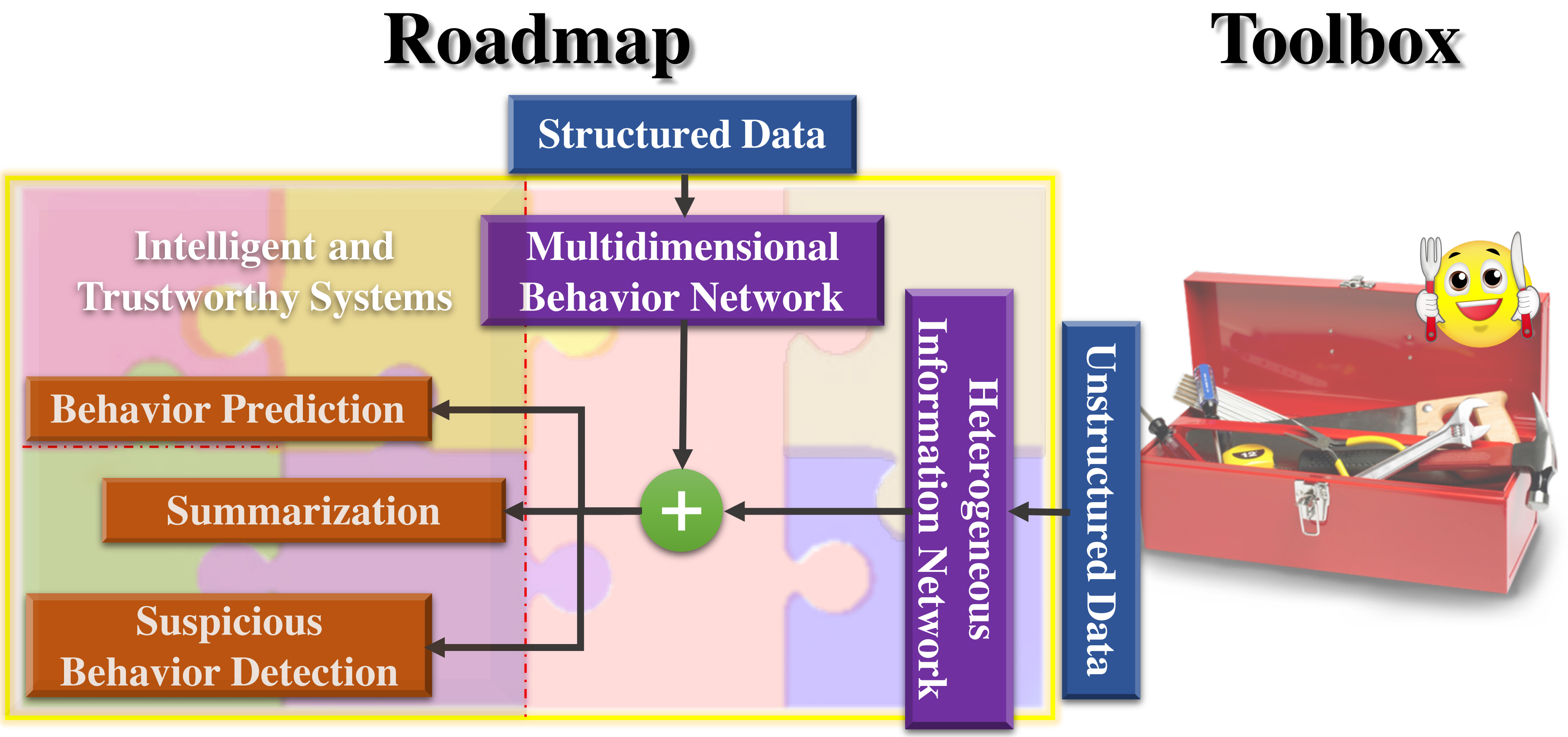
Constructing Networks with Trustworthy Information

Conclusion. **Data-Driven Behavioral Analytics with Networks**

T1: **Mining behavior networks with social, spatiotemporal contexts**

T2: **Structuring information networks from behavioral content**

T3: **Integrating behavior networks with rich information networks**



Acknowledgement

NSF
NSFC
National Natural Science Foundation of China
Carnegie Mellon University
University of Maryland 1856
Microsoft Research 微软亚洲研究院
Tencent 腾讯

References

D. Blei, A. Ng, and M. Jordan. "Latent dirichlet allocation." JMLR, 2003.

J. Herlocker, J. Konstan, L. Terveen, J. Riedl. "Evaluating collaborative filtering recommender systems." ACM TOIS, 2004.

Y. Koren, R. Bell, C. Volinsky. "Matrix factorization techniques for recommender systems." Computer, 2009.

Y. Koren. "Factorization meets the neighborhood: A multifaceted collaborative filtering model." KDD, 2008.

Y. Koren. "Collaborative filtering with temporal dynamics." CACM, 2010.

M. Balabanovic and Y. Shoham. "FAB: Content-based, collaborative recommendation." CACM, 1997.

N. Liu and Q. Yang. "Eigenrank: A ranking-oriented approach to collaborative filtering." SIGIR, 2008.

N. Liu, M. Zhao, and Q. Yang. "Probabilistic latent preference analysis for collaborative filtering." CIKM, 2009.

References

H. Ma, H. Yang, M. Lyu, and I. King. "Sorec: Social recommendation using probabilistic matrix factorization." CIKM, 2008.

H. Ma, T. Zhou, M. Lyu, and I. King. "Improving recommender systems by incorporating social contextual information." ACM TOIS, 2011.

H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. "Recommender systems with social regularization." WSDM, 2011.

J. Leskovec, A. Singh, and J. Kleinberg. "Patterns of influence in a recommendation network." PAKDD, 2006.

P. Massa and A. Paolo. "Trust-aware recommender systems." RecSys, 2007.

M. Jamali and E. Martin. "TrustWalker: A random walk model for combining trust-based and item-based recommendation." KDD, 2009.

H. Ma, I. King, and M. Lyu. "Learning to recommend with social trust ensemble." SIGIR, 2009.

H. Ma. I. King, and M. Lyu. "Learning to recommend with explicit and implicit social relations." ACM TIST, 2011.

References

M. Faloutsos, P. Faloutsos, and C. Faloutsos. “On power-law relationships of the internet topology.” SIGCOMM, 1999.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner. “Graph structure in the web.” Computer Networks, 2000.

F. Chung and L. Lu. “The average distances in random graphs with given expected degrees.” PNAS, 2002.

J. Kleinberg. “Authoritative sources in a hyperlinked environment.” JACM, 1999.

H. Kwak, C. Lee, H. Park, and S. Moon. “What is Twitter, a social network or a news media?” WWW, 2010.

B. Hooi, H.A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. “Fraudar: Bounding graph fraud in the face of camouflage.” KDD, 2016.

C. Aggarwal and J. Han. “Frequent pattern mining.” Springer, 2014.

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. “FreeSpan: Frequent pattern-projected sequential pattern mining.” KDD, 2000.

References

X. Yan and J. Han. "gspan: Graph-based substructure pattern mining." ICDM, 2003.

X. Yan and J. Han. "CloseGraph: Mining closed frequent graph patterns." KDD, 2003.

Y. Sun, J. Han, X. Yan, P.S. Yu, and T. Wu. "PathSim: Meta path-based top-k similarity search in heterogeneous information networks." VLDB, 2011.

Y. Sun, Y. Yu, and J. Han. "Ranking-based clustering of heterogeneous information networks with star network schema." KDD, 2009.

Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. "RankClus: Integrating clustering with ranking for heterogeneous information network analysis." EDBT, 2009.

Y. Sun, R. Barber, M. Gupta, C. Aggarwar, and J. Han. "Co-author relationship prediction in heterogeneous bibliographic networks." ASONAM, 2011.

A. El-Kishky, Y. Song, C. Wang, C.R. Voss, and J. Han. "Scalable topical phrase mining from text corpora." VLDB, 2014.

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. "Mining quality phrases from massive text corpora." SIGMOD, 2015.

References

X. Ren, A. El-Kishky, C. Wang, F. Tao, C.R. Voss, and J. Han. "Effective entity recognition and typing by relation phrase-based clustering." KDD, 2015.

X. Ren, W. He, M. Qu, C.R. Voss, H. Ji, and J. Han. "Label noise reduction in entity typing by heterogeneous partial-label embedding." KDD, 2016.

C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. "A phrase mining framework for recursive construction of a topical hierarchy." KDD, 2013.

E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos. "ParCube: Sparse parallelizable tensor decompositions." PKDD, 2012.

D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. "VOG: Summarizing and understanding large graphs." SDM, 2014.

R. Gupta, A. Halevy, X. Wang, S.E. Whang, and F. Wu. "Biperpedia: An ontology for search applications." VLDB, 2014.

M. Yahya, S. Whang, R. Gupta, and A. Halevy. "ReNoun: Fact extraction for nominal attributes." EMNLP, 2014.

A. Halevy, N. Noy, S. Sarawagi, S.E. Whang, and X. Yu. "Discovering structure in the universe of attribute names." WWW, 2016.

References

Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation." SIGMOD, 2014.

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. "A confidence-aware approach for truth discovery on long-tail data." VLDB, 2014.

F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation." KDD, 2015.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. "A survey on truth discovery." KDD Explorations Newsletter, 2016.

S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. "Modeling truth existence in truth discovery." KDD, 2015.

S. Kumar, R. West, and J. Leskovec. "Disinformation on the Web: Impact, characteristics, and detection of Wikipedia hoaxes." WWW, 2016.

S. Kumar, F. Spezzano, and V.S. Subrahmanian. "Identifying malicious actors on social media." ASONAM, 2016. (tutorial)

Thank you!

Data-Driven Behavioral Analytics with Networks

Meng Jiang

www.meng-jiang.com