UNCOVERING AND MODELING COMPLEX BEHAVIORS IN SOCIAL MEDIA

Meng Jiang

Department of Computer Science and Technology, Tsinghua University

Advisor: Professor Shiqiang Yang

Contact: mjiang89@gmail.com

Homepage: www.meng-jiang.com

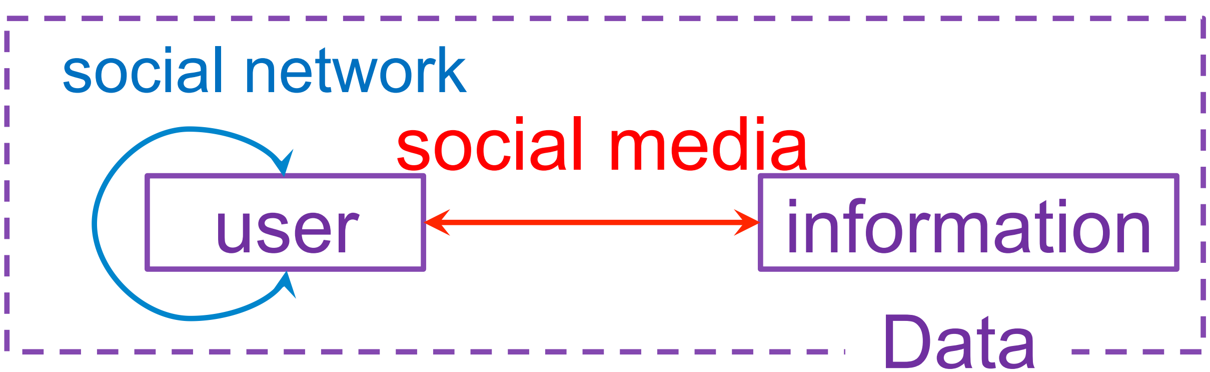
Background: Behaviors in Social Media

- Scientists can study user behaviors now!

"user-user" interaction

"user-information" interaction

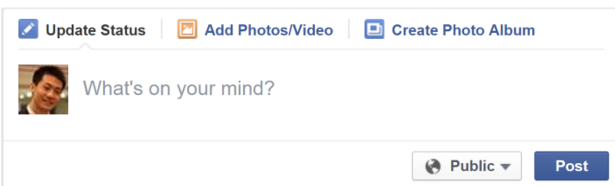
- We have richer behaviors in social media!



Background: Behavior-Oriented Systems

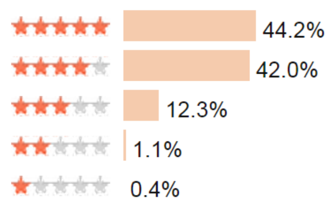
- Great marketing values!

Post, forward text/image



News feed ranking

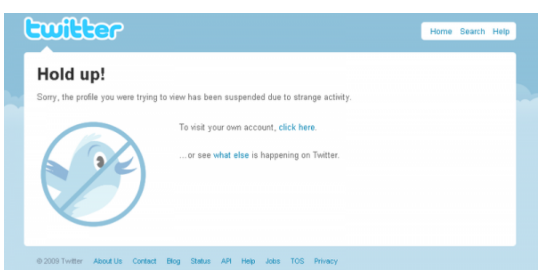
Give ratings to movies



Recommender systems

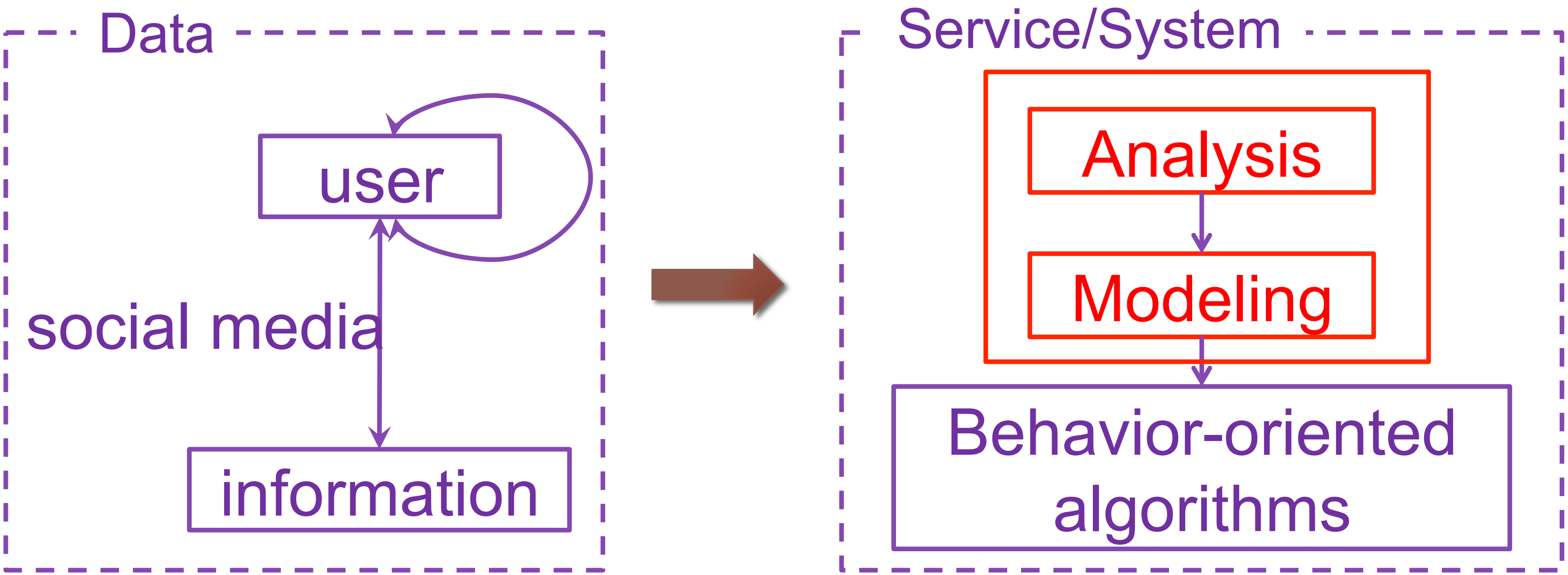
Zombie followers, fraud

Anti-spam, anti-fraud



Behavioral Analysis and Modeling

- The **1st step** to implement a behavioral system
- The **basis** of social media services
- The **key problem** of social data processing



Complex Behaviors in Social Media

Social contexts

Spatial-temporal contexts

Complex Behaviors in Social Media

Cross-domain

Cross-platform

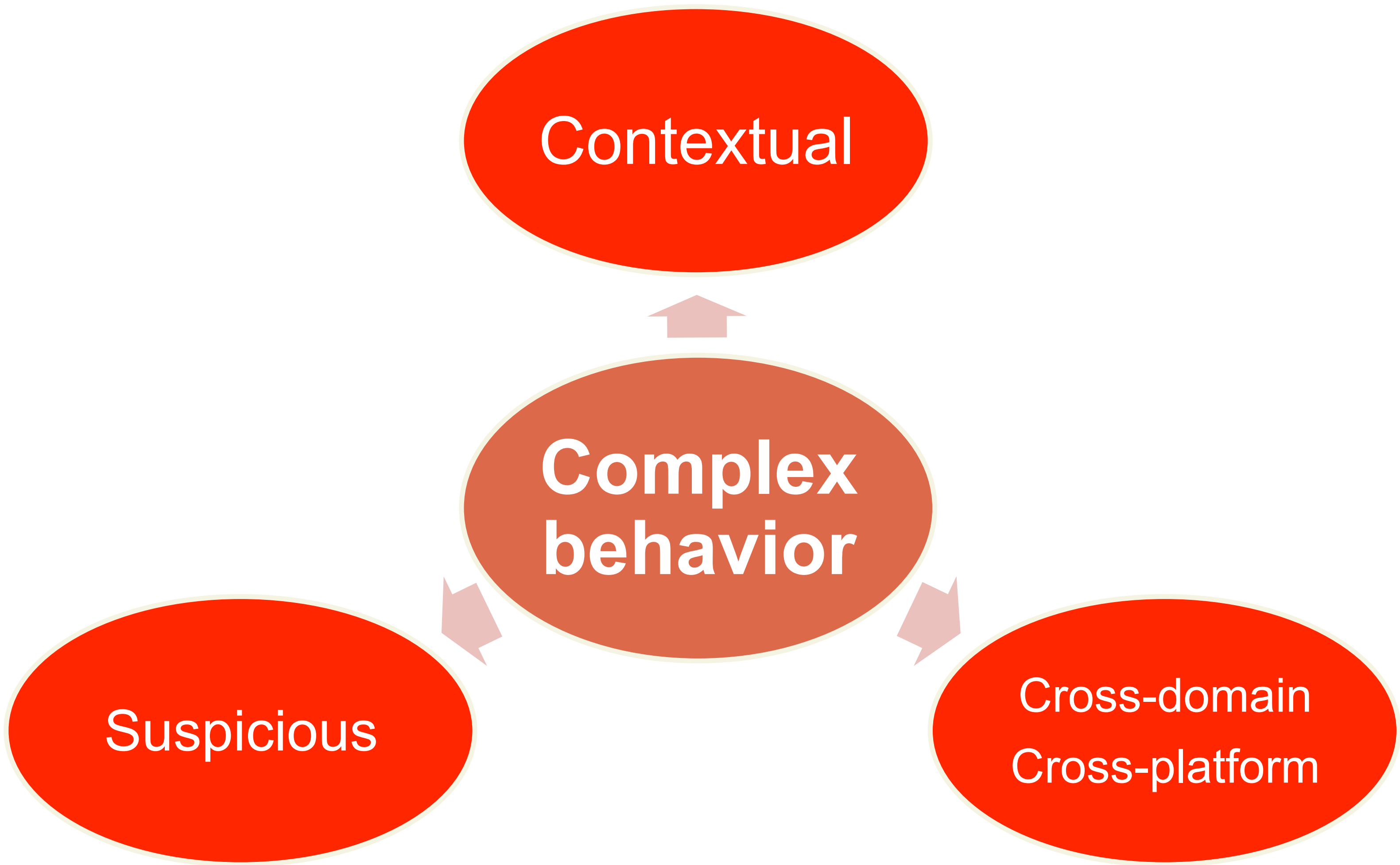
Complex Behaviors in Social Media

Suspicious

zombie follower

ill-gotten “Likes”

Complex Behaviors in Social Media



Related Works

- Information adopting behavior prediction
  - Content-based filtering [Balabanovic et al. ’97][Basu et al. AAAI’98]
  - Memory-based Collaborative Filtering [Herlock et al. SIGIR’99] [Sarwar et al. WWW’01]
  - Homophily [McPherson et al. ’01]
  - Social Influence [Leskovec et al. PAKDD’06]
  - Model-based CF [Yehuda KDD’08][Ma et al. CIKM’08][Ma et al. WSDM’11]
- Suspicious behavior detection
  - Duplicated content [Jindal et al. WSDM’08]
  - Spam text and image [Lim et al. CIKM’10]
  - Burst [Xie et al. KDD’12]
  - Sentiment difference [Hu et al. ICDM’14]

ROADMAP

Contextual behavior analysis & modeling

Social context-based behavioral model

Spatial and temporal context-based analysis

Cross-domain/platform behavior modeling

Cross-domain hybrid random walk algorithm

Cross-platform semi-supervised transfer learning

Suspicious behavior analysis & detection

Detecting synchronized suspicious users

Evaluating suspicious multi-faceted behaviors

ROADMAP

Contextual behavior analysis & modeling

Social context-based behavioral model

Spatial and temporal context-based analysis

Cross-domain/platform behavior modeling

Cross-domain hybrid random walk algorithm

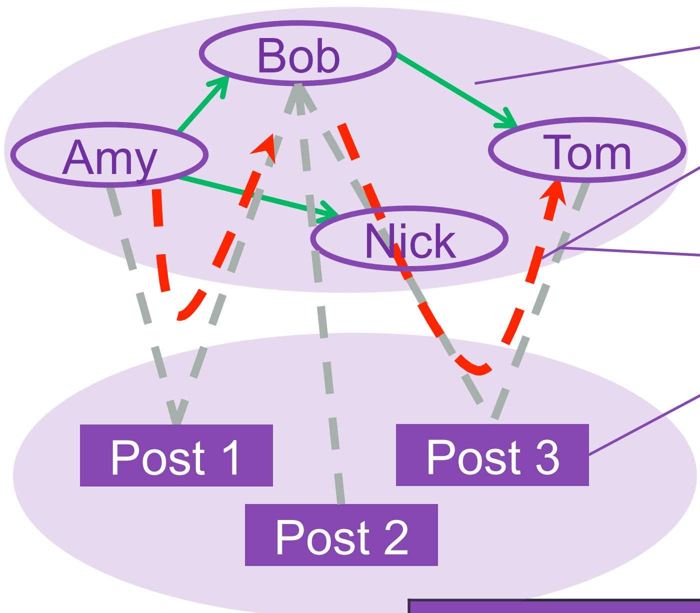
Cross-platform semi-supervised transfer learning

Suspicious behavior analysis & detection

Detecting synchronized suspicious users

Evaluating suspicious multi-faceted behaviors

P1 & C: Contextual Behavior Modeling



(1) Social relation

(2) "User-user" interaction frequency

(3) "User-post" interaction

Sparse!

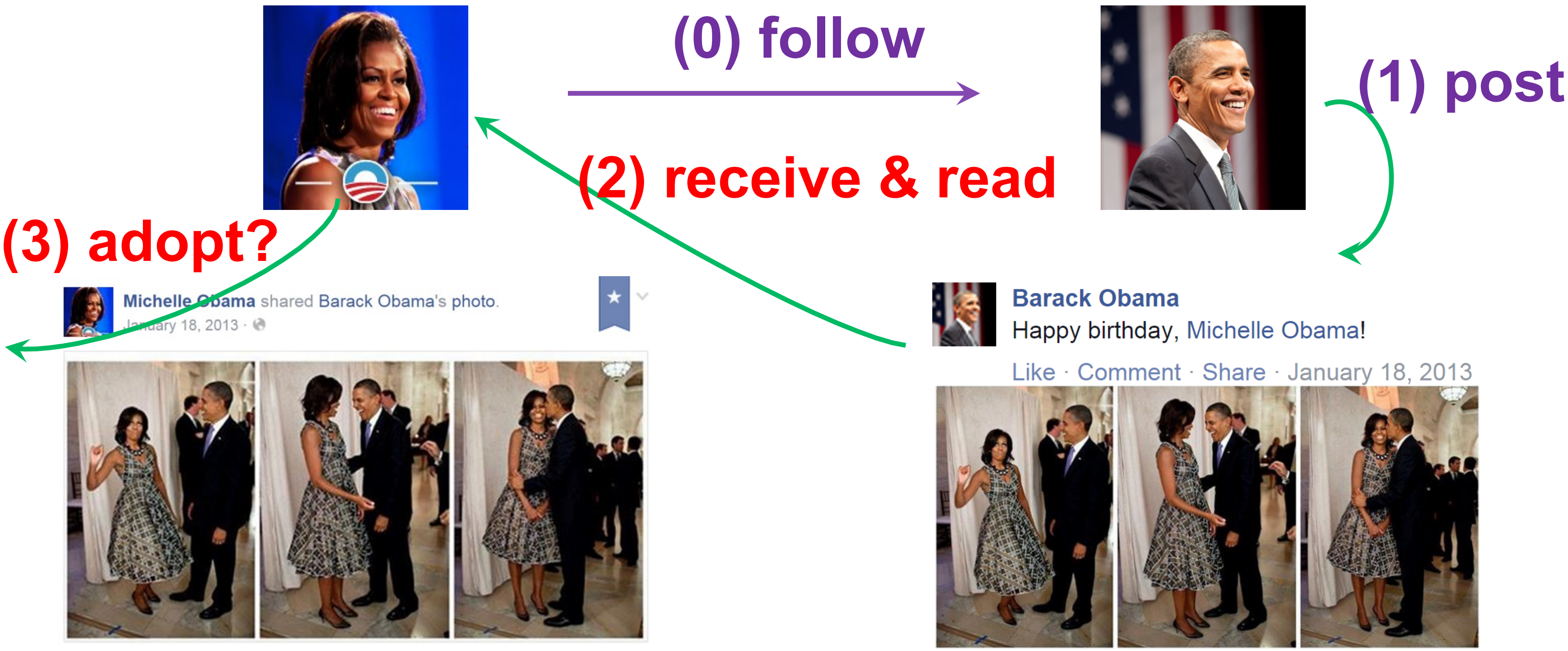
(4) Post content

Existing behavior models

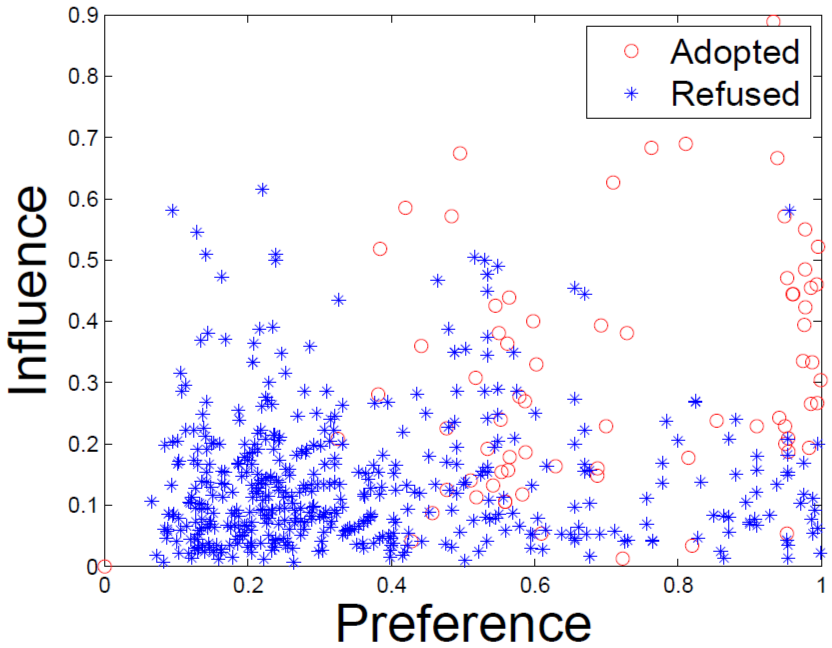
	(1) Social relation	(2) User-user	(3) User-post	(4) Post content
Content filtering & CF	✖	✖	√	√
Social trust & Influence	✖	√	√	✖
Low-rank matrix factorization	√	✖	√	√

Idea: Social Contextual Factors

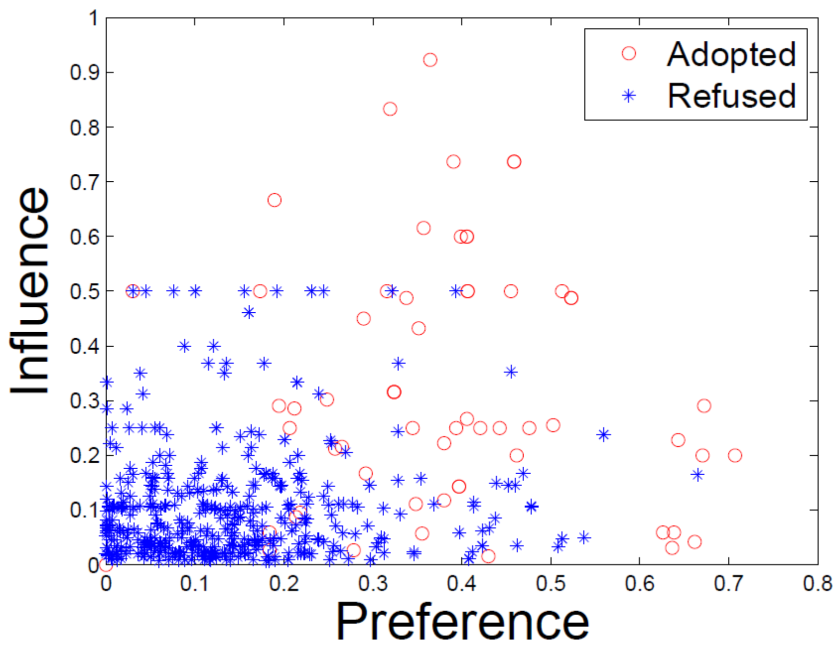
- Personal preference: Like the content?
- Interpersonal Influence: Who is the sender?



Idea: Social Contextual Factors



China's Facebook: Renren



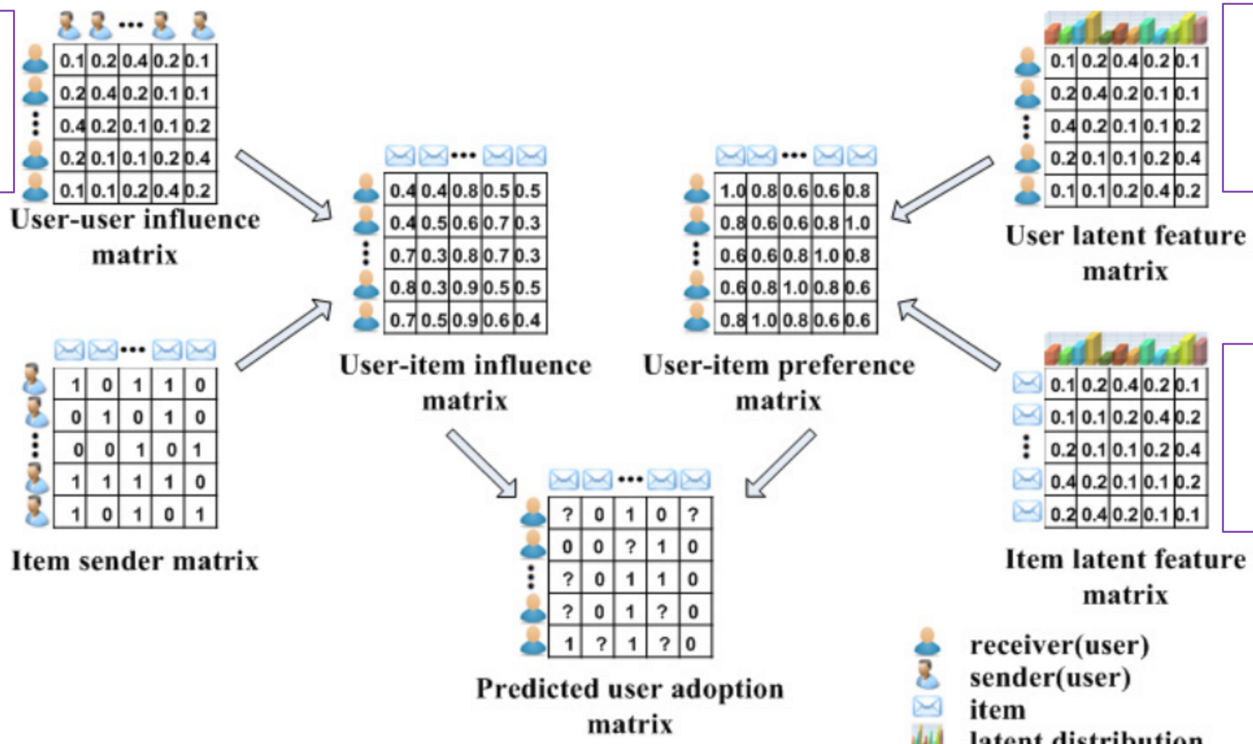
China's Twitter: Tencent Weibo

ContextMF: Social Contextual Model

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

User-sender influence ***S***

Sender ***G***



User latent features ***U***

Item latent features ***V***

Observed behaviors ***R***

ContextMF Algorithm

- Minimize sum-of-squared errors function

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

- Block coordinate descent scheme with gradients

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

Performance: Predicting Adoptions

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**

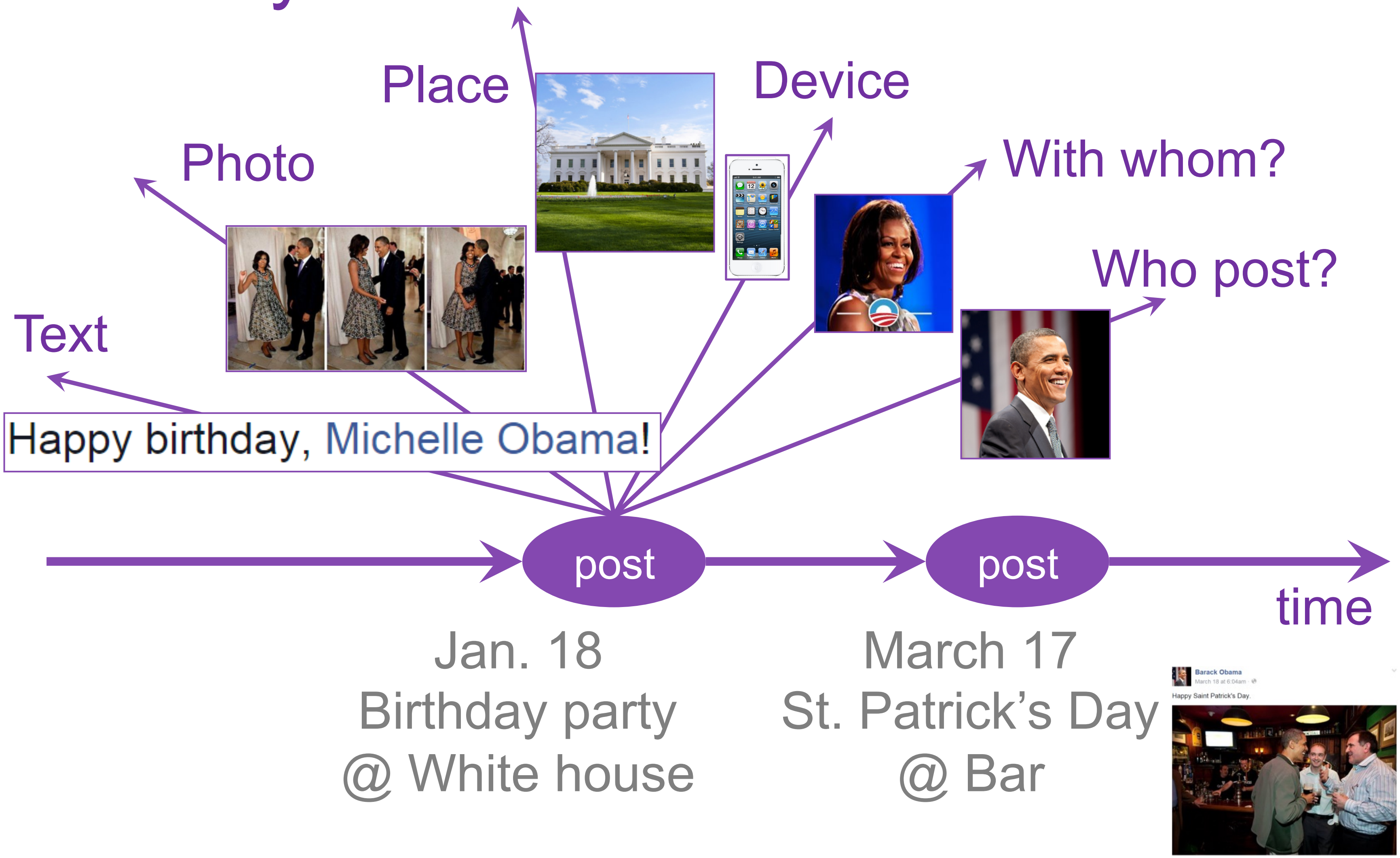
	Renren	Tencent Weibo
MAE	-19.1%	-24.2%
RMSE	-12.8%	-20.7%
Kendall's	+9.82%	+2.1%
Spearman's	+10.6%	+3.1%

Contributions

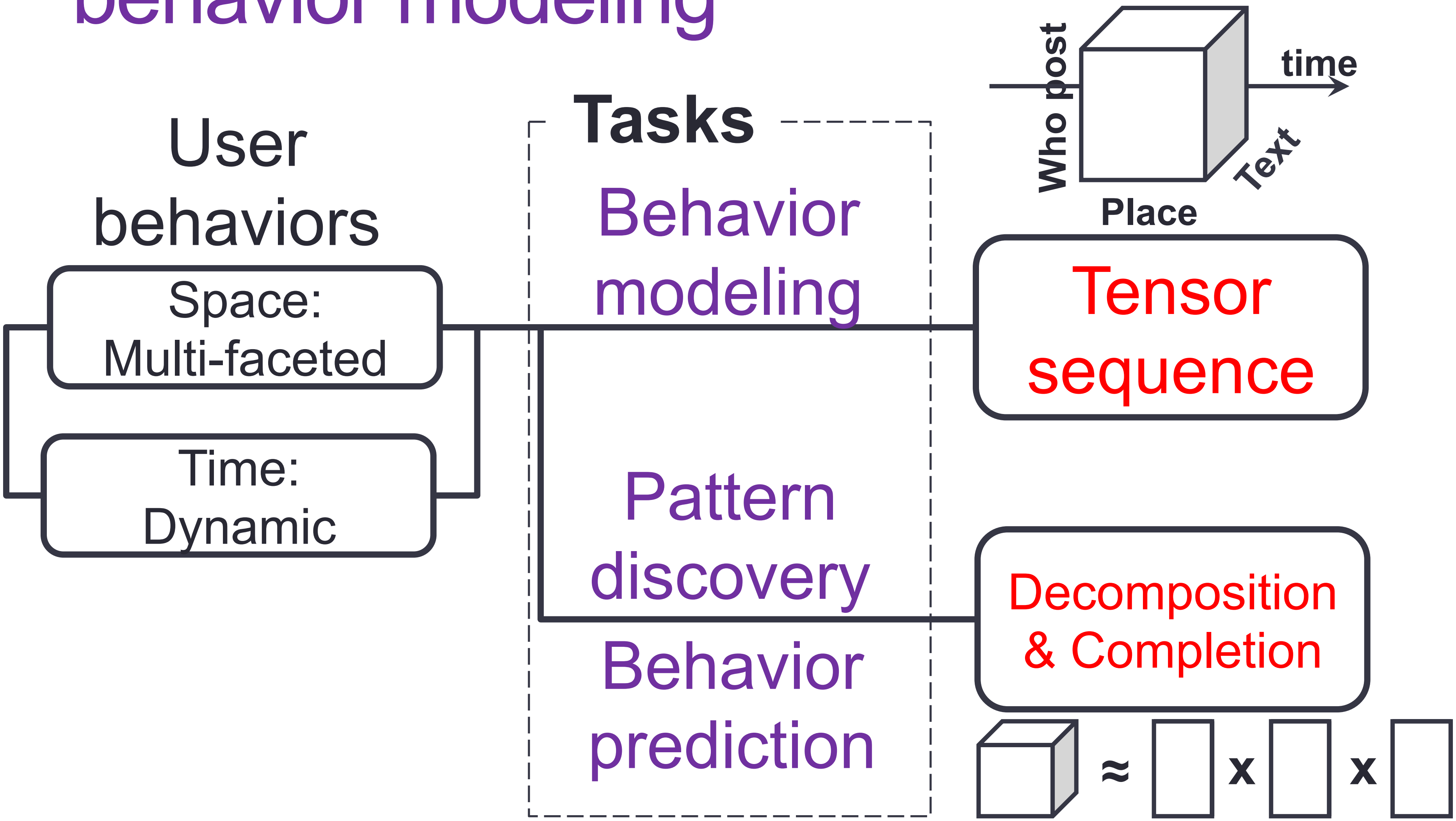
- Analyzed social contextual behaviors
- Proposed social context-based behavior prediction model ContextMF
- Improved behavior prediction performance in real social media

- Publications
  - ACM CIKM 2012 (Full. Acc. rate=13.8%.)
  - IEEE TKDE 2014 (Regular.)
  - Citation count: **85**

Spatial Temporal Contexts: Multi-faceted and Dynamic Behaviors

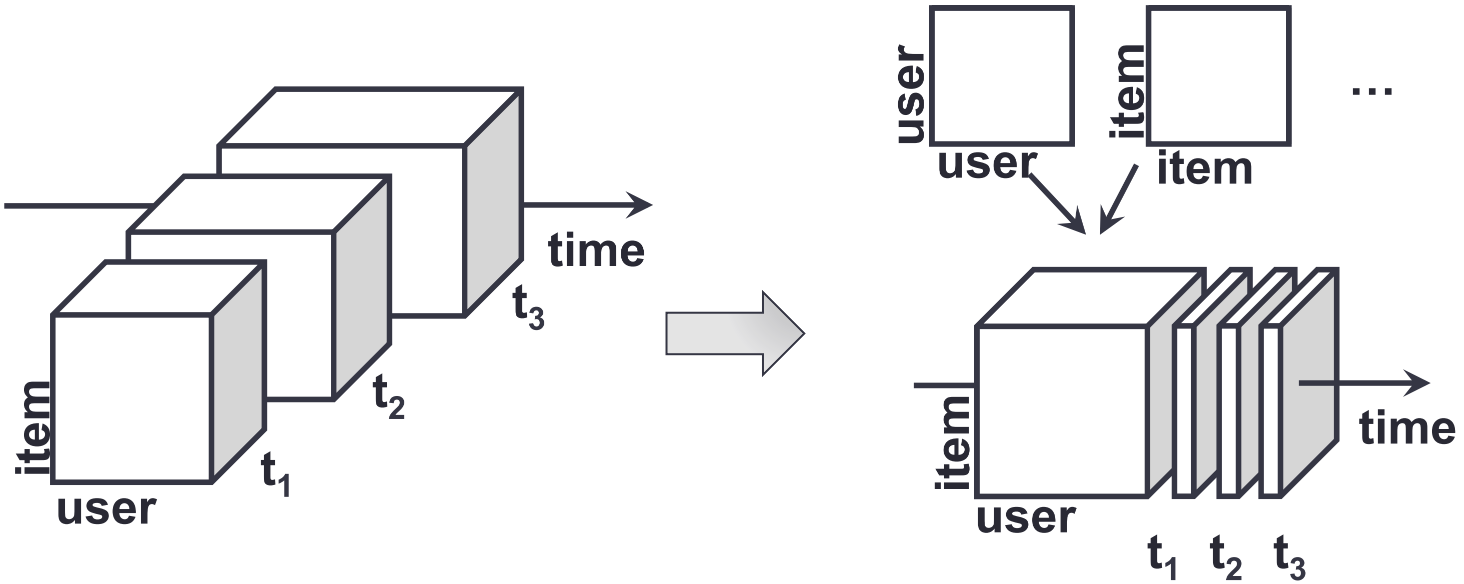


P2: Spatial temporal context-based behavior modeling

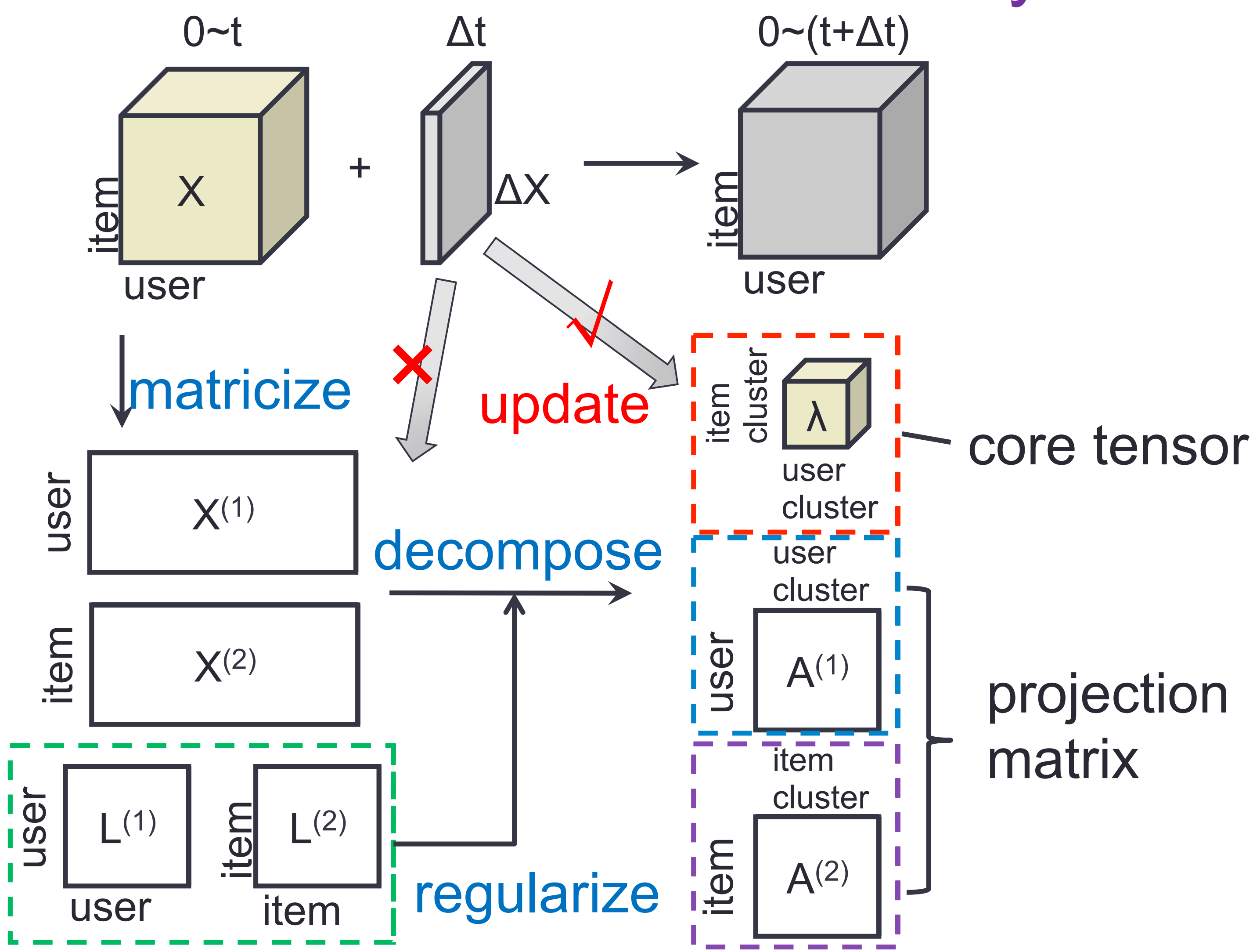


C & Idea

- Challenges
  - **High sparsity**
  - **High complexity**
- Ideas
  - Flexible regularization with auxiliary data
  - Incremental updates for projection matrix

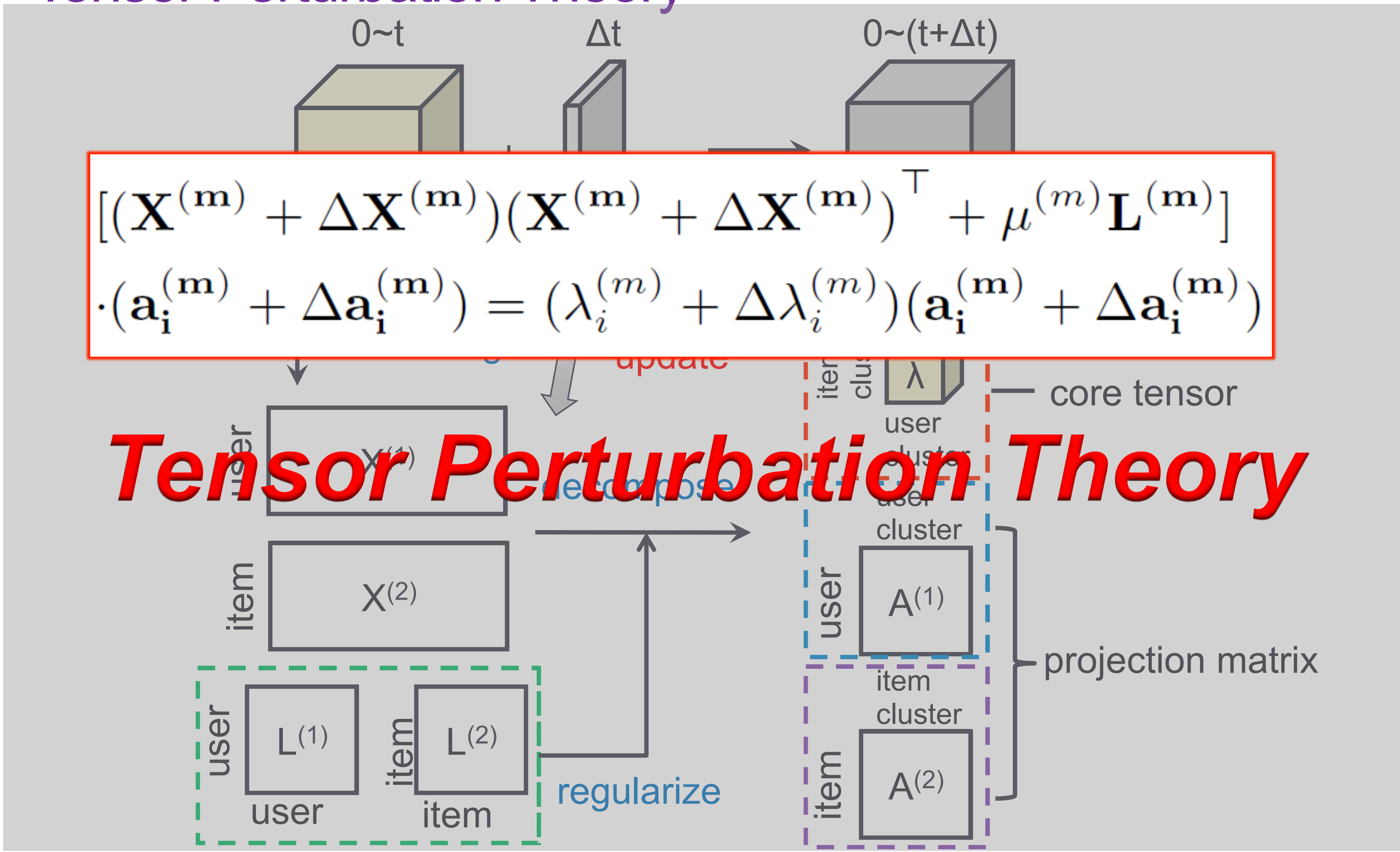


FEMA: Flexible Evolutionary Multi-faceted Analysis with Tensor Perturbation Theory



FEMA: Flexible Evolutionary Multi-faceted Analysis with Tensor Perturbation Theory

$$[(\mathbf{X}^{(\mathbf{m})} + \Delta\mathbf{X}^{(\mathbf{m})})(\mathbf{X}^{(\mathbf{m})} + \Delta\mathbf{X}^{(\mathbf{m})})^{\top} + \mu^{(m)}\mathbf{L}^{(\mathbf{m})}]$$
$$\cdot(\mathbf{a}_\mathbf{i}^{(\mathbf{m})} + \Delta\mathbf{a}_\mathbf{i}^{(\mathbf{m})}) = (\lambda_i^{(m)} + \Delta\lambda_i^{(m)})(\mathbf{a}_\mathbf{i}^{(\mathbf{m})} + \Delta\mathbf{a}_\mathbf{i}^{(\mathbf{m})})$$



Tensor Perturbation Theory

FEMA Algorithm

Approximation

Require: $\mathcal{X}_{\mathbf{t}}, \Delta\mathcal{X}_{\mathbf{t}}, \mathbf{A}_{\mathbf{t}}^{(\mathbf{m})}|_{m=1}^{M}, \lambda_{\mathbf{t}}^{(\mathbf{m})}|_{m=1}^{M}$

for $m = 1, \ldots, M$ **do**

for $i = 1, \ldots, r^{(m)}$ **do**

Compute $\Delta\lambda_{t,i}^{(m)}$ using

$$\boxed{\Delta\lambda_i^{(m)} = \mathbf{a}_i^{(\mathbf{m})\top}(\mathbf{X}^{(\mathbf{m})}\Delta\mathbf{X}^{(\mathbf{m})\top} + \Delta\mathbf{X}^{(\mathbf{m})}\mathbf{X}^{(\mathbf{m})\top})\mathbf{a}_i^{(\mathbf{m})}}$$

and compute

$$\lambda_{t+1,i}^{(m)} = \lambda_{t,i}^{(m)} + \Delta\lambda_{t,i}^{(m)};$$

Compute $\Delta\mathbf{a}_{\mathbf{t,i}}^{(\mathbf{m})}$ using

$$\boxed{\Delta\mathbf{a}_i^{(\mathbf{m})} = \sum_{j\neq i}\frac{\mathbf{a}_j^{(\mathbf{m})\top}(\mathbf{X}^{(\mathbf{m})}\Delta\mathbf{X}^{(\mathbf{m})\top} + \Delta\mathbf{X}^{(\mathbf{m})}\mathbf{X}^{(\mathbf{m})\top})\mathbf{a}_i^{(\mathbf{m})}}{\lambda_i^{(m)} - \lambda_j^{(m)}}\mathbf{a}_j^{(\mathbf{m})}}$$

and compute

$$\mathbf{a}_{\mathbf{t+1,i}}^{(\mathbf{m})} = \mathbf{a}_{\mathbf{t,i}}^{(\mathbf{m})} + \Delta\mathbf{a}_{\mathbf{t,i}}^{(\mathbf{m})} \text{ and } \mathbf{A}_{\mathbf{t+1}}^{(\mathbf{m})} = \{\mathbf{a}_{\mathbf{t+1,i}}^{(\mathbf{m})}\};$$

end for

end for

$\mathcal{Y}_{\mathbf{t+1}} = (\mathcal{X}_{\mathbf{t}} + \Delta\mathcal{X}_{\mathbf{t}})\prod_{m=1}^{M}\times_{(m)}\mathbf{A}_{\mathbf{t+1}}^{(\mathbf{m})\mathbf{T}};$

return $\mathbf{A}_{\mathbf{t+1}}^{(\mathbf{m})}|_{m=1}^{M}, \lambda_{\mathbf{t+1}}^{(\mathbf{m})}|_{m=1}^{M}, \mathcal{Y}_{\mathbf{t+1}}$

Bound Guarantee

core tensor

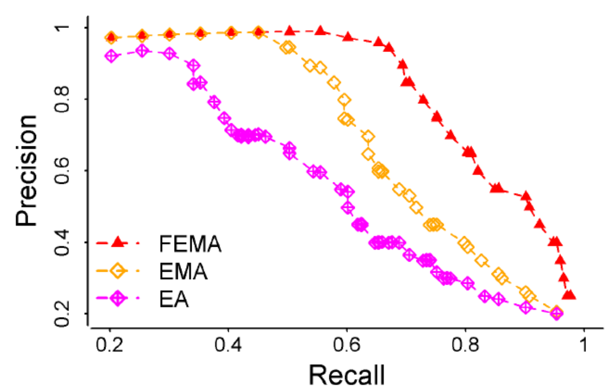
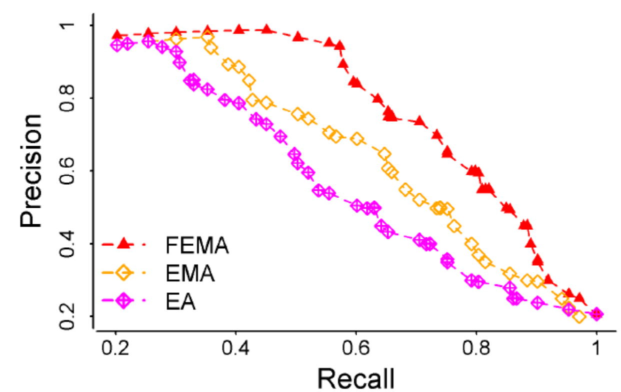
$$|\Delta\lambda_i^{(m)}| \leq 2(\lambda^{\max}_{\mathbf{X}^{(\mathbf{m})\top}\mathbf{X}^{(\mathbf{m})}})^{\frac{1}{2}}\|\Delta\mathbf{X}^{(\mathbf{m})}\|_2$$

projection matrix

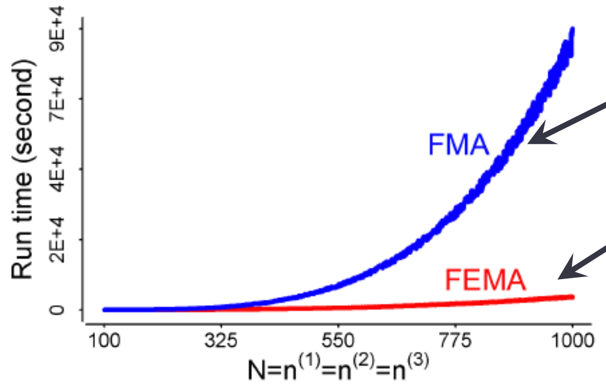
$$|\Delta\mathbf{a}_{\mathbf{i}}^{(\mathbf{m})}| \leq 2\|\Delta\mathbf{X}^{(\mathbf{m})}\|_2\sum_{j\neq i}\frac{(\lambda^{\max}_{\mathbf{X}^{(\mathbf{m})\top}\mathbf{X}^{(\mathbf{m})}})^{\frac{1}{2}}}{|\lambda_i^{(m)} - \lambda_j^{(m)}|}$$

Performance: Predicting academic behaviors and mentioning behaviors

Paper: author-(affiliation)-keyword **Tweet: who @-@ whom-(word)**

	Microsoft Academic Search		Tencent Weibo mentions “@”	
	MAE	RMSE	MAE	RMSE
FEMA X L	**0.735**	**0.944**	**0.894**	**1.312**
EMA X	0.794	1.130	0.932	1.556
EA X	0.979	1.364	1.120	1.873
Precision vs Recall				

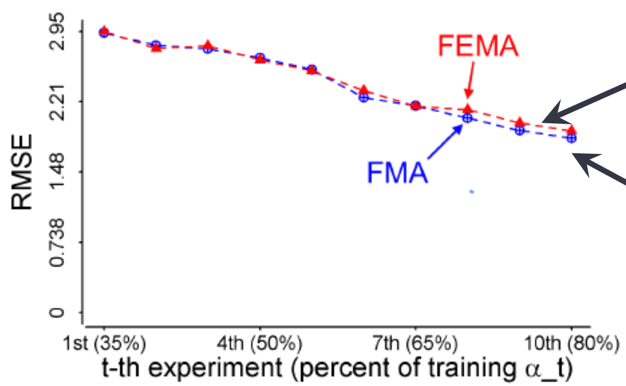
Performance: Efficiency & Small Loss



Time vs Num. objects N

Re-decomposition:
Re-compute projection matrices

Evolutionary analysis:
Use ΔX to update λ and a



The loss is small.

Evolutionary analysis:
Use ΔX to update λ and a

Re-decomposition:
Re-compute projection matrices

Contributions

- Analyzed spatial temporal context-based behaviors: multi-faceted/dynamic
- Proposed behavioral analysis method FEMA
- Improved prediction effects and efficiency on two real datasets

- Publication
  - ACM SIGKDD 2014 (Full. Acc. rate=14.6%.)

ROADMAP

Contextual behavior analysis & modeling

Social context-based behavioral model

Spatial and temporal context-based analysis

Cross-domain/platform behavior modeling

Cross-domain hybrid random walk algorithm

Cross-platform semi-supervised transfer learning

Suspicious behavior analysis & detection

Detecting synchronized suspicious users

Evaluating suspicious multi-faceted behaviors

P3 & C: Cross-domain Behavior Modeling

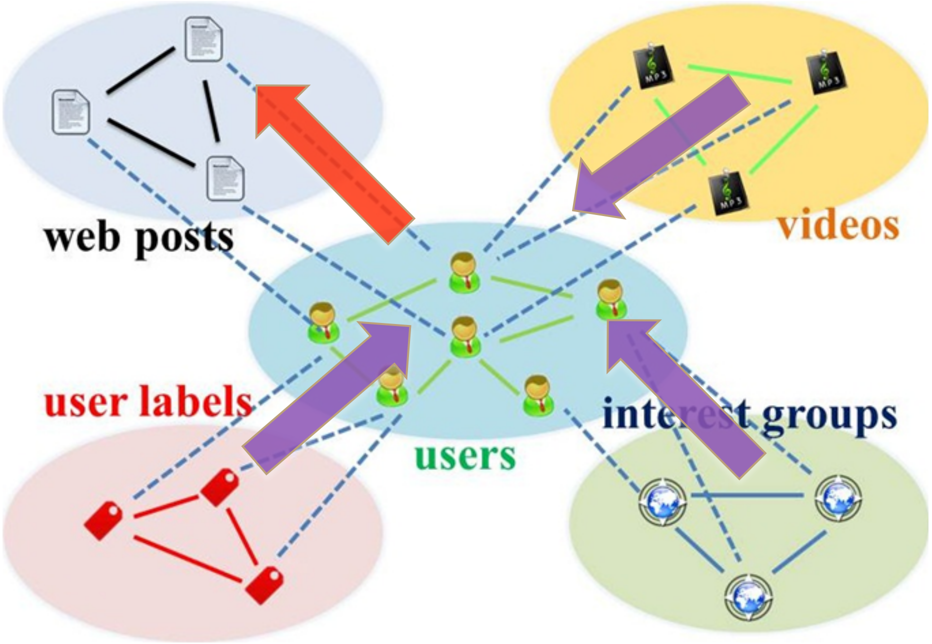
- Tencent Weibo data

Domain	Size	Cross-domain Link（User-Item）	
		Adopt（+）	Refuse（-）
User	53.4K	—	—
Weibo	142K	1.47M (0.02%)	3.40M (0.04%)
Label	111	330K（5.57%)	—

- Multi-domain social media: post, label, video…
- **C1**：Sparsity and cold start in target domain
- **C2**：Heterogeneity in multiple domains

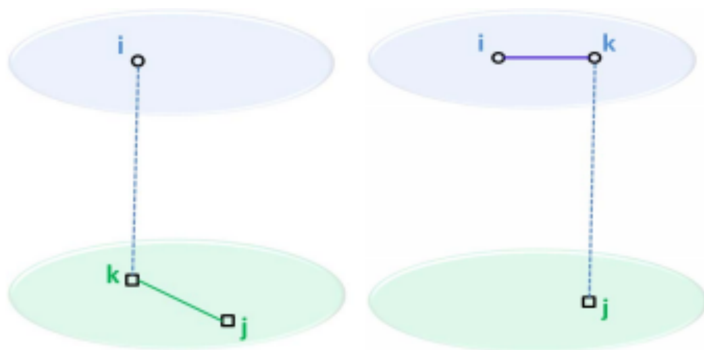
Idea: Social Domain as Bridge

- Reframe social media: Star-structured graph with social domain in the center
- Transfer learning
  - Auxiliary domain → Social domain → Target domain



Hybrid Random Walk Algorithm

- Update cross-domain link weights



(a) Route 1 (b) Route 2

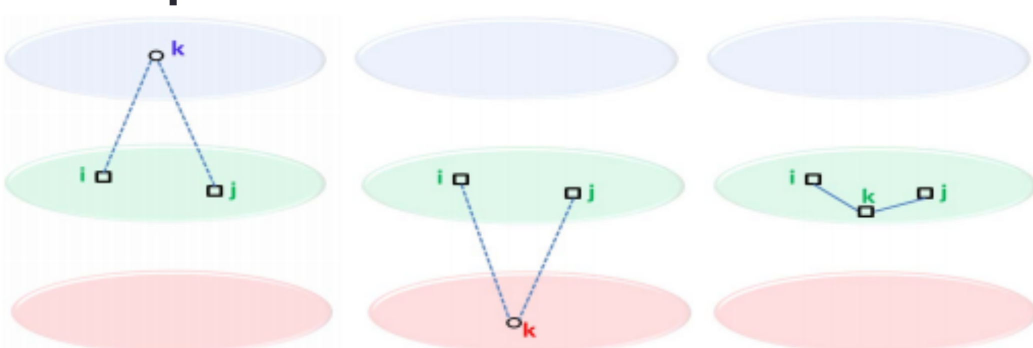
$$p_{ij}^{(\mathcal{UP})^+} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^+} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^+} r_{kj}^{(\mathcal{P})}$$

$$p_{ij}^{(\mathcal{UP})^-} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^-} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^-} r_{kj}^{(\mathcal{P})}$$

$$\mathbf{P}^{(\mathcal{UP})^+}(t+1) = \delta \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{UP})^+}(t) + (1-\delta)\mathbf{P}^{(\mathcal{UP})^+}(t)\mathbf{R}^{(\mathcal{P})}$$

$$\mathbf{P}^{(\mathcal{UP})^-}(t+1) = \delta \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{UP})^-}(t) + (1-\delta)\mathbf{P}^{(\mathcal{UP})^-}(t)\mathbf{R}^{(\mathcal{P})}$$

- Update within-domain link weights



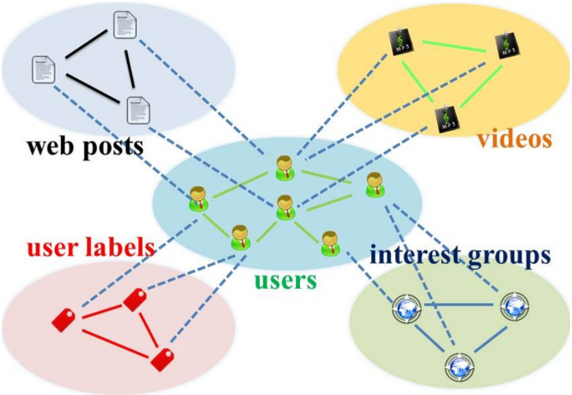
(a) Route 1 (b) Route 2 (c) Route 3

$$r_{ij}^{(\mathcal{U})} = \tau^{(\mathcal{P})}\Big(\mu \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^+} p_{jk}^{(\mathcal{UP})^+} + (1-\mu) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^-} p_{jk}^{(\mathcal{UP})^-}\Big) + \tau^{(\mathcal{T})} \sum_{t_k \in \mathcal{T}} p_{ik}^{(\mathcal{UT})^+} p_{jk}^{(\mathcal{UT})^+} + \tau^{(\mathcal{U})} \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} r_{kj}^{(\mathcal{U})}$$

$$\mathbf{R}^{(\mathcal{U})}(t+1) = \tau^{(\mathcal{P})}\Big(\mu \mathbf{P}^{(\mathcal{UP})^+}(t)\mathbf{P}^{(\mathcal{UP})^+}(t)^T + (1-\mu)\mathbf{P}^{(\mathcal{UP})^-}(t)\mathbf{P}^{(\mathcal{UP})^-}(t)^T\Big) + \tau^{(\mathcal{T})}\mathbf{P}^{(\mathcal{UT})^+}(t)\mathbf{P}^{(\mathcal{UT})^+}(t)^T + \tau^{(\mathcal{U})}\mathbf{R}^{(\mathcal{U})}(t)\mathbf{R}^{(\mathcal{U})}(t)^T$$

Hybrid Random Walk Algorithm

- On high-order star-structured graph



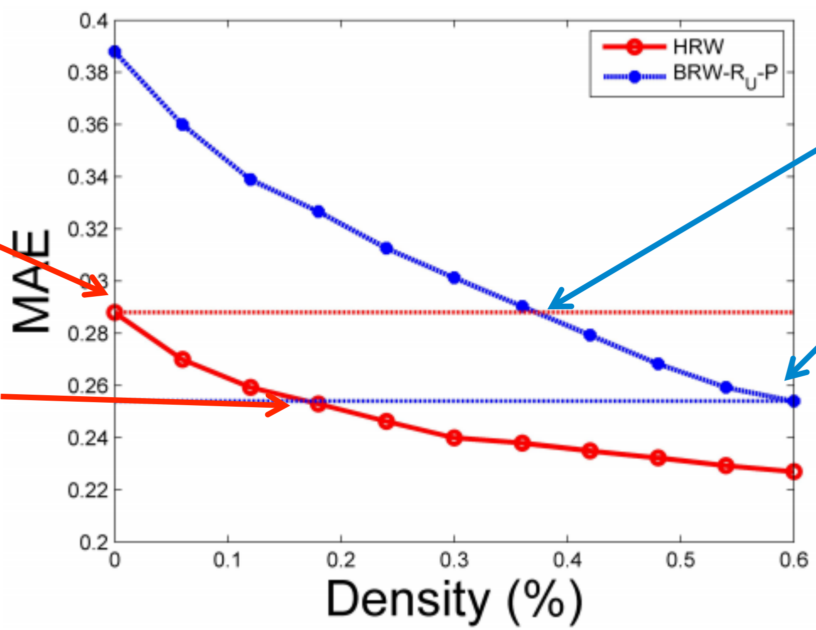
$$
\begin{aligned}
\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t+1) &= \delta_i \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t) + (1-\delta_i)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t)\mathbf{R}^{(\mathcal{D}_i)} \\
\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t+1) &= \delta_i \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t) + (1-\delta_i)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t)\mathbf{R}^{(\mathcal{D}_i)} \\
\mathbf{R}^{(\mathcal{U})}(t+1) &= \sum_{\mathcal{D}_i \in \mathcal{D}} \tau_i \mu_i \mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t)^T \\
&\quad + \sum_{\mathcal{D}_i \in \mathcal{D}} \tau_i (1-\mu_i)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t)^T \\
&\quad + \tau^{(\mathcal{U})}\mathbf{R}^{(\mathcal{U})}(t)\mathbf{R}^{(\mathcal{U})}(t)^T
\end{aligned}
$$

Performance: Predicting Cold-start Behaviors

- Knowledge transfer from auxiliary domains improves cold-start users' behavior prediction
  - Using aux. (label) data, saving >60% tgt. (post) data

0 user-post
100% user-label

18% user-post
100% user-label



35% user-post
0 user-label

60% user-post
0 user-label

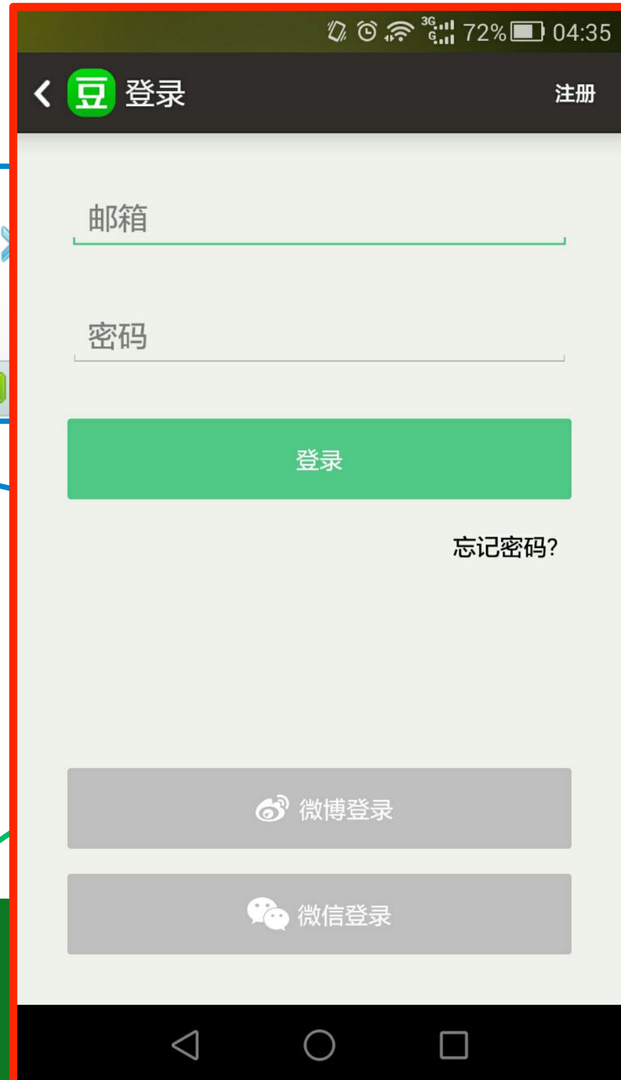
Contributions

- Analyzed cross-domain behavior modeling problem: Using social domain as bridge
- Proposed a novel Hybrid Random Walk method
- Improved behavior prediction in target domain and provided effective solutions to cold start

- Publication
  - ACM CIKM 2012 (Full paper. Acc. rate=13.8%.)
  - IEEE TKDE 2015 (to appear. Regular.)
  - Citation count: **32**

P4: Cross-platform Behavior Modeling

Social label



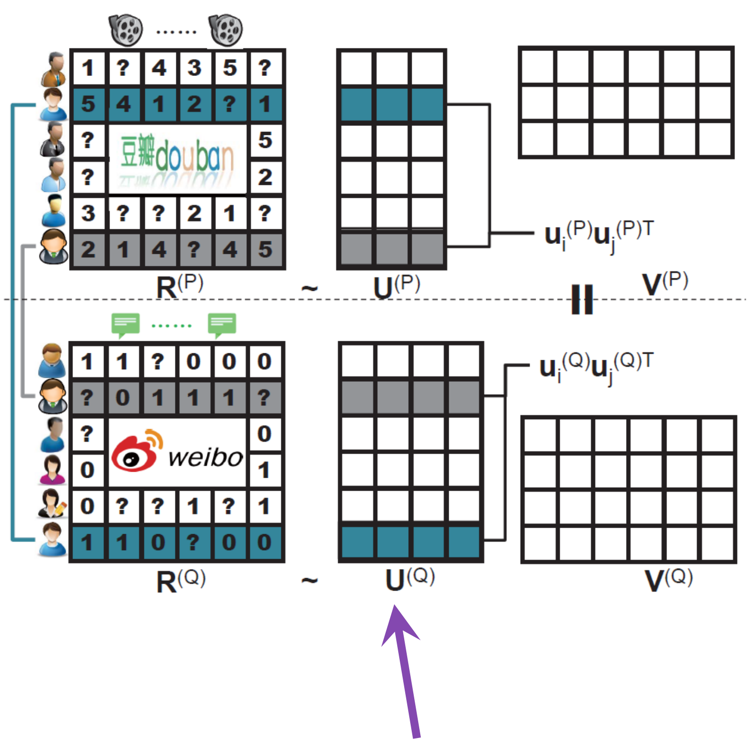
Tweet & Retweet

Video, Music …

Movie rating

“Like” message, page

C & Idea

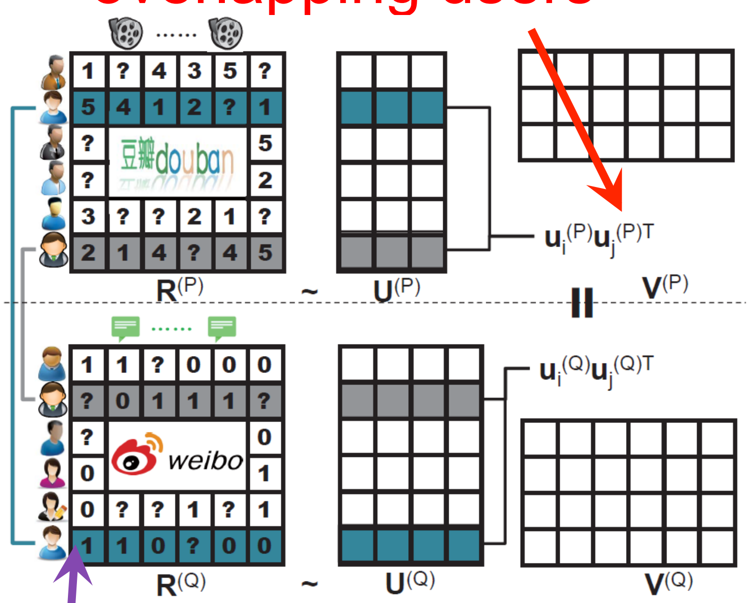


Auxiliary platform

- Goal: Solve high sparsity problem in target platform
- Solution: Using knowledge from auxiliary platform
- Challenges
  - Heterogeneity
  - Non-uniform representation

C & Idea

Constraints to user representations: Similarity between overlapping users



Overlaps

Auxiliary platform

- Goal: Solve high sparsity problem in target platform
- Solution: Using knowledge from auxiliary platform
- Challenges
  - Heterogeneity
  - Non-uniform representation
- Idea
  - Partially overlapping users across platforms

XPTrans: Semi-supervised Transfer

- Input
  - Tgt./Aux. platform <u>P</u>/<u>Q</u>;
  - Behavior data <u>$R^{(P)}$</u>/<u>$R^{(Q)}$</u>;
  - Observation <u>$W^{(P)}$</u>/<u>$W^{(Q)}$</u>;
  - Overlapping indicator <u>$W^{(P,Q)}$</u>,
- Output
  - User latent representation <u>$U^{(P)}$</u>/<u>$U^{(Q)}$</u>;
  - Item latent representation <u>$V^{(P)}$</u>/<u>$V^{(Q)}$</u>;
  - Missing values in <u>$R^{(P)}$</u>

- Objective function

Unsupervised term

Target platform

Auxiliary platform

$$\mathcal{J} = \sum_{i,j} W_{i,j}^{(P)} \left(R_{i,j}^{(P)} - \sum_{r} U_{i,r}^{(P)} V_{r,j}^{(P)} \right)^2 + \lambda \sum_{i,j} W_{i,j}^{(Q)} \left(R_{i,j}^{(Q)} - \sum_{r} U_{i,r}^{(Q)} V_{r,j}^{(Q)} \right)^2 + \mu \sum_{i_1,j_1,i_2,j_2} W_{i_1,j_1}^{(P,Q)} W_{i_2,j_2}^{(P,Q)} \left(A_{i_1,i_2}^{(P)} - A_{j_1,j_2}^{(Q)} \right)^2$$

$$A_{i_1,i_2}^{(P)} = \sum_{r=1}^{r_P} U_{i_1,r}^{(P)} U_{i_2,r}^{(P)};\ A_{j_1,j_2}^{(Q)} = \sum_{r=1}^{r_Q} U_{j_1,r}^{(Q)} U_{j_2,r}^{(Q)}$$

Overlapping user similarity

Supervised term

Performance: Behavior Prediction

- Baselines
  - CMF: NOT using knowledge in auxiliary platform
  - CBT: NOT using overlapping user but sharing "Codebook" pattern
  - XPTrans-Align: Latent representation of overlaps as constraints
  - **XPTrans:** Overlapping users' latent similarity as constraints
- XPTrans improves accuracy by 21%

	Q: Weibo tweet entity $\rightarrow$ P: Douban movie				Q: Douban book $\rightarrow$ P: Weibo tag			
	RMSE		MAP		RMSE		MAP	
	$P \cap Q$	$P \setminus Q$	$P \cap Q$	$P \setminus Q$	$P \cap Q$	$P \setminus Q$	$P \cap Q$	$P \setminus Q$
CMF [24]	0.779	1.439	0.805	0.640	0.267	0.429	0.666	0.464
CBT [10]	0.767	1.290	0.808	0.676	0.261	0.419	0.675	0.477
XPTRANS-ALIGN	0.757	1.164	0.811	0.702	0.256	0.411	0.681	0.487
XPTRANS	**0.715**	**0.722**	**0.821**	**0.820**	**0.236**	**0.374**	**0.705**	**0.533**
vs CBT	↓6.8%	↓44.0%	↑1.62%	↑21.3%	↓9.6%	↓10.8%	↑4.5%	↑11.7%
vs XPTRANS-ALIGN	↓5.5%	↓38.0%	↑1.3%	↑16.8%	↓8.0%	↓9.0%	↑3.6%	↑9.4%

Contributions

- Analyzed cross-platform behavior modeling: Using overlapping users as bridge
- Proposed semi-supervised transfer learning method XPTrans
- Improved behavior prediction in target platform

- Not published yet

ROADMAP

Contextual behavior analysis & modeling

Social context-based behavioral model

Spatial and temporal context-based analysis

Cross-domain/platform behavior modeling

Cross-domain hybrid random walk algorithm

Cross-platform semi-supervised transfer learning

Suspicious behavior analysis & detection

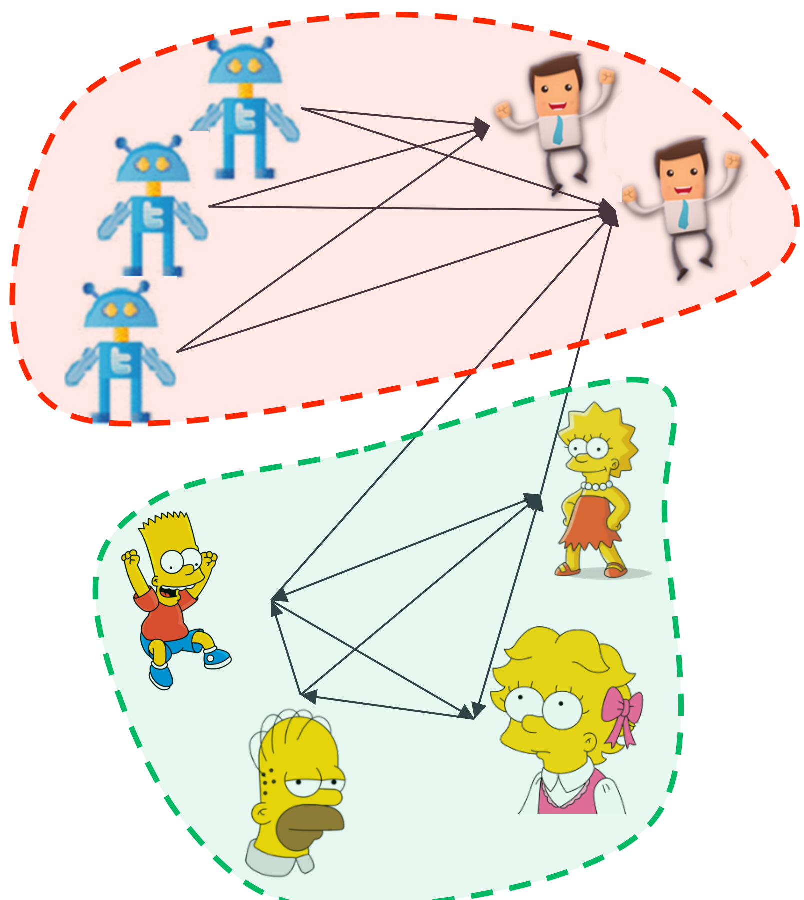
Detecting synchronized suspicious users

Evaluating suspicious multi-faceted behaviors

P5: Suspicious Zombie Followers

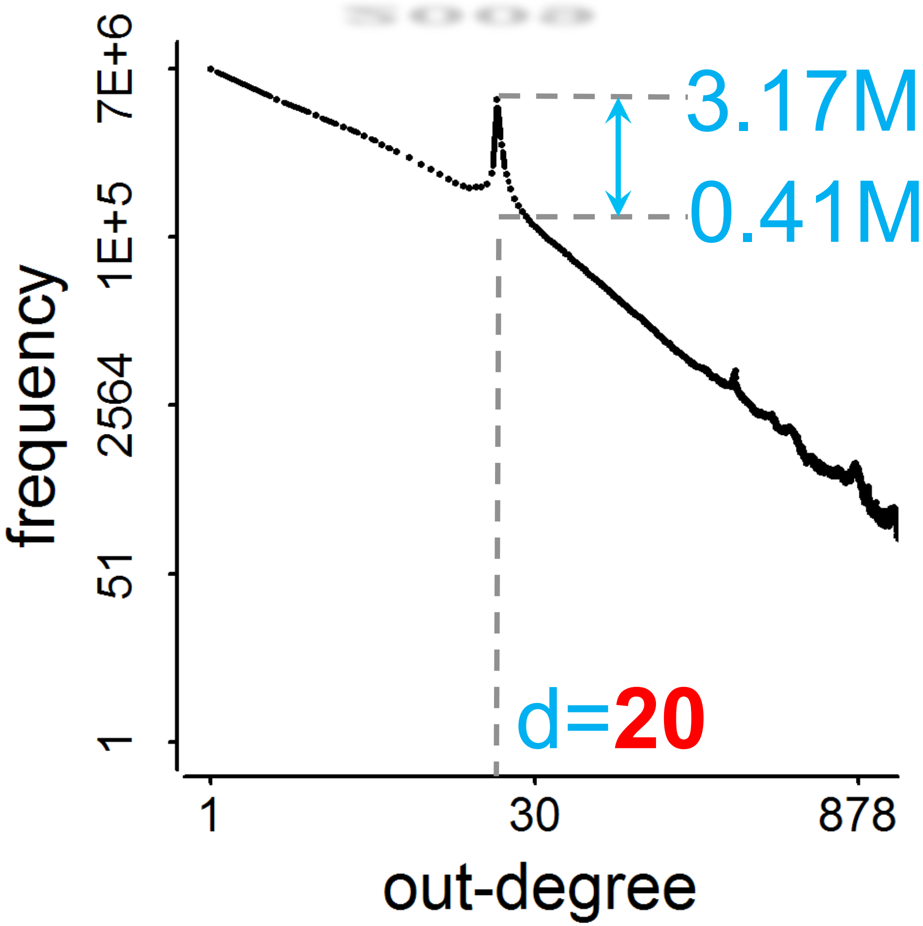
[www.buyfollowz.org]

[buymorelikes.com]

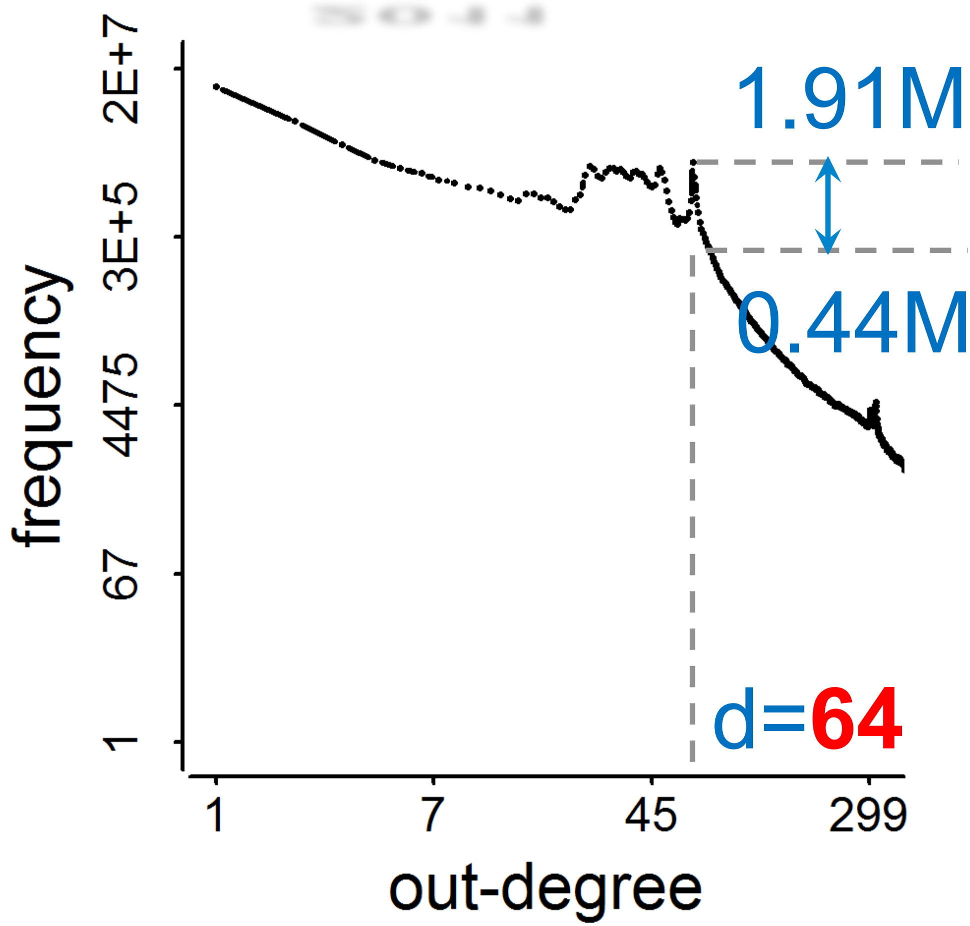


C1: Small Out-degree, Big Spikes

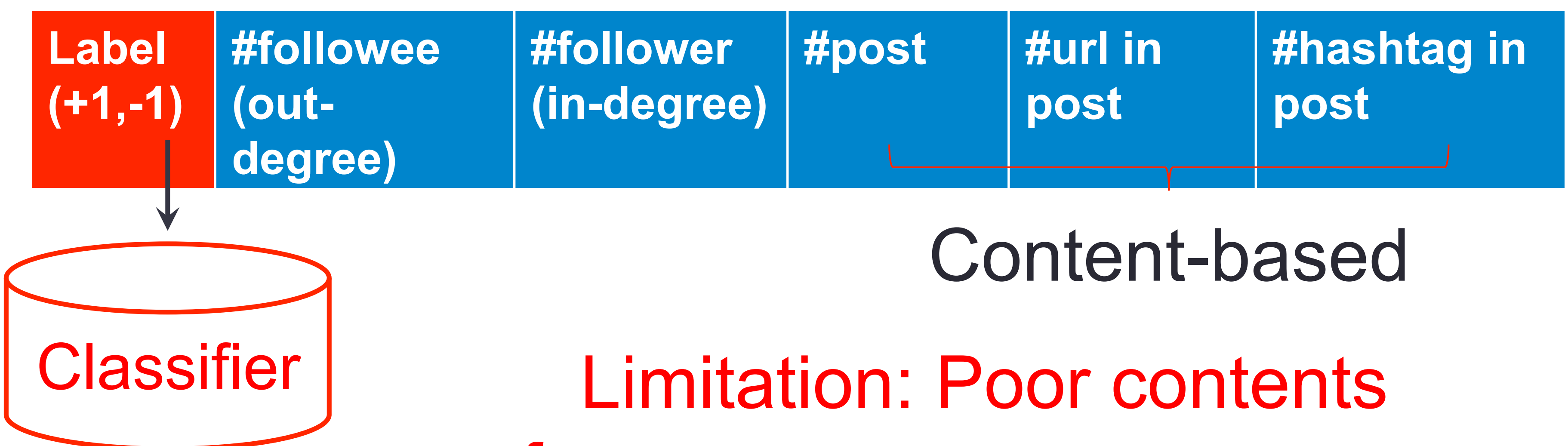
2009 **41M**



2011 **117M**



C2: Limitations of Existing Methods



Limitation: Poor contents for unnecessary appearance

Idea: Behavior Pattern is the Key

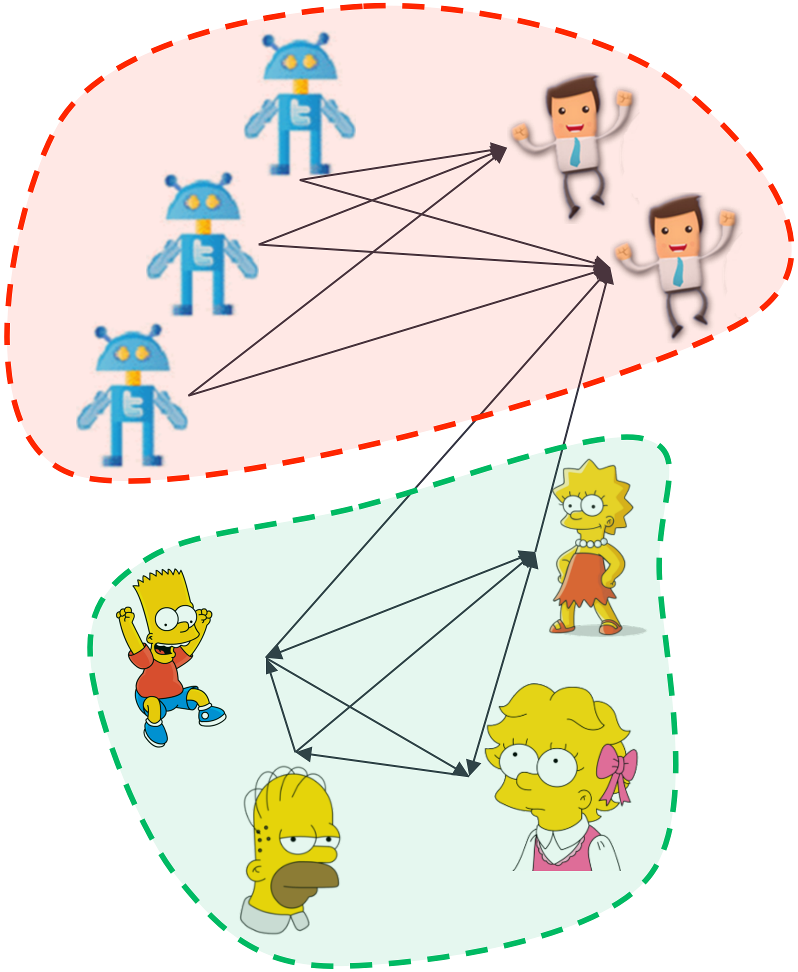
Monetary Incentive

Content

May show suspicious appearance

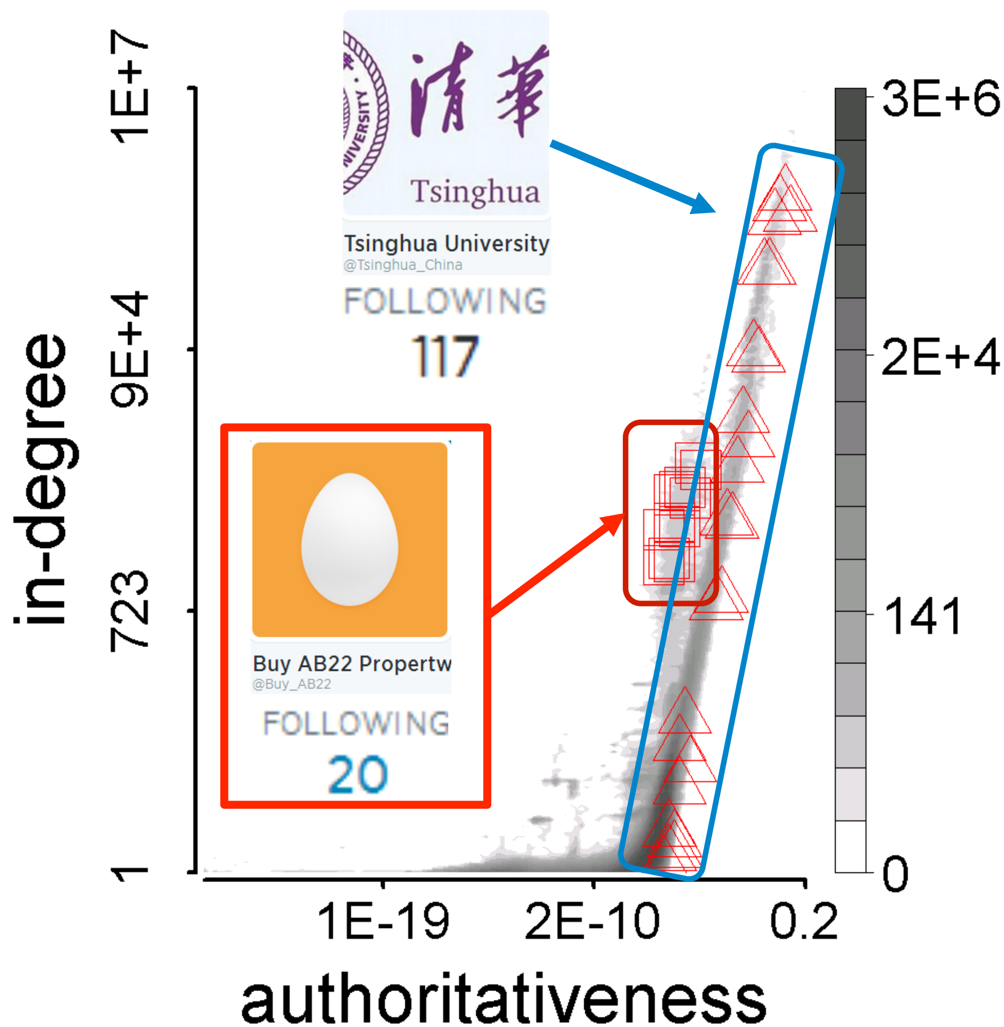
Behavior

Have to behave suspicious



Behavior Pattern: Whom do Zombie Followers Connect to?

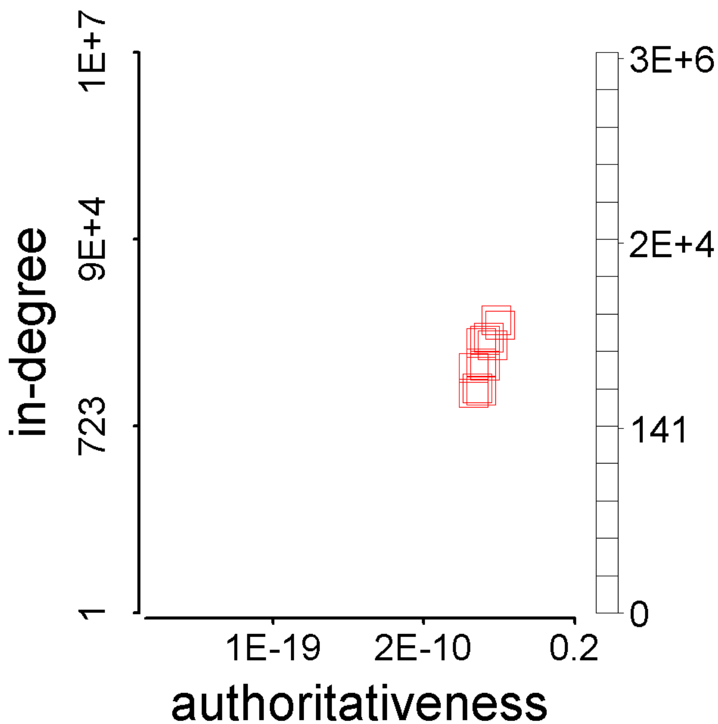
- Synchronized
- Abnormal

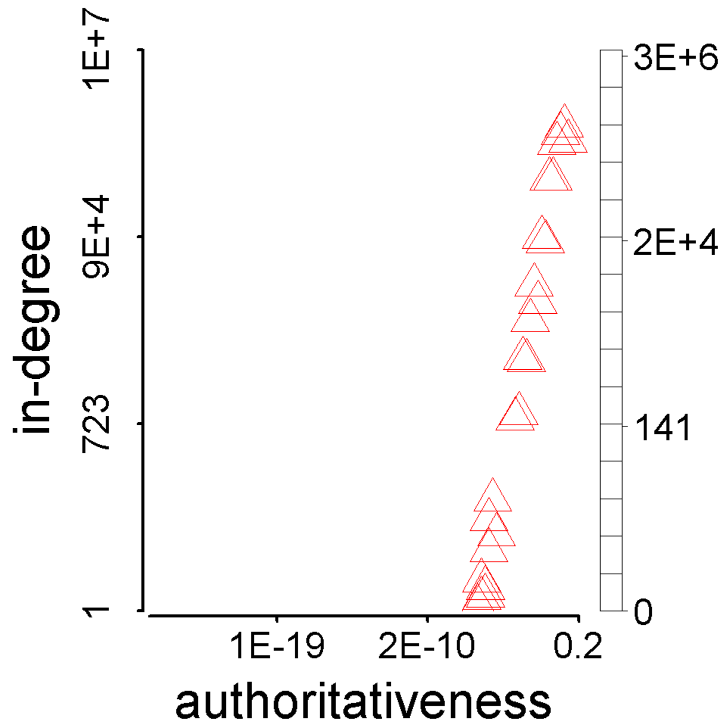


Synchronicity & Normality

- Synchronicity

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

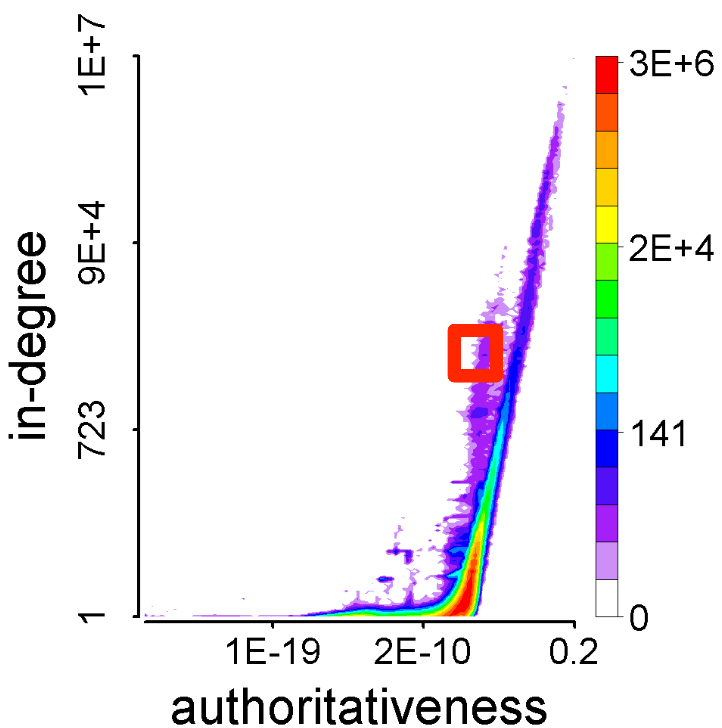


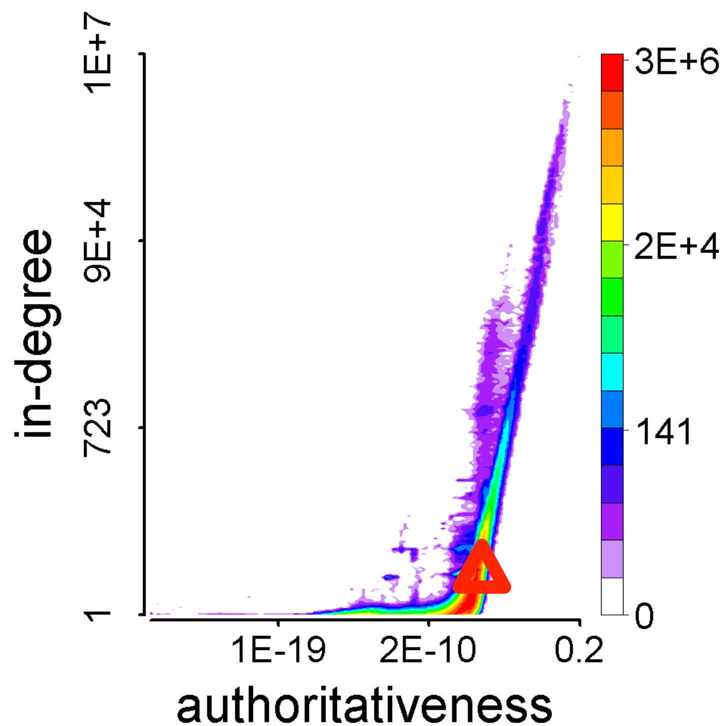


Synchronicity & Normality

- Normality

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





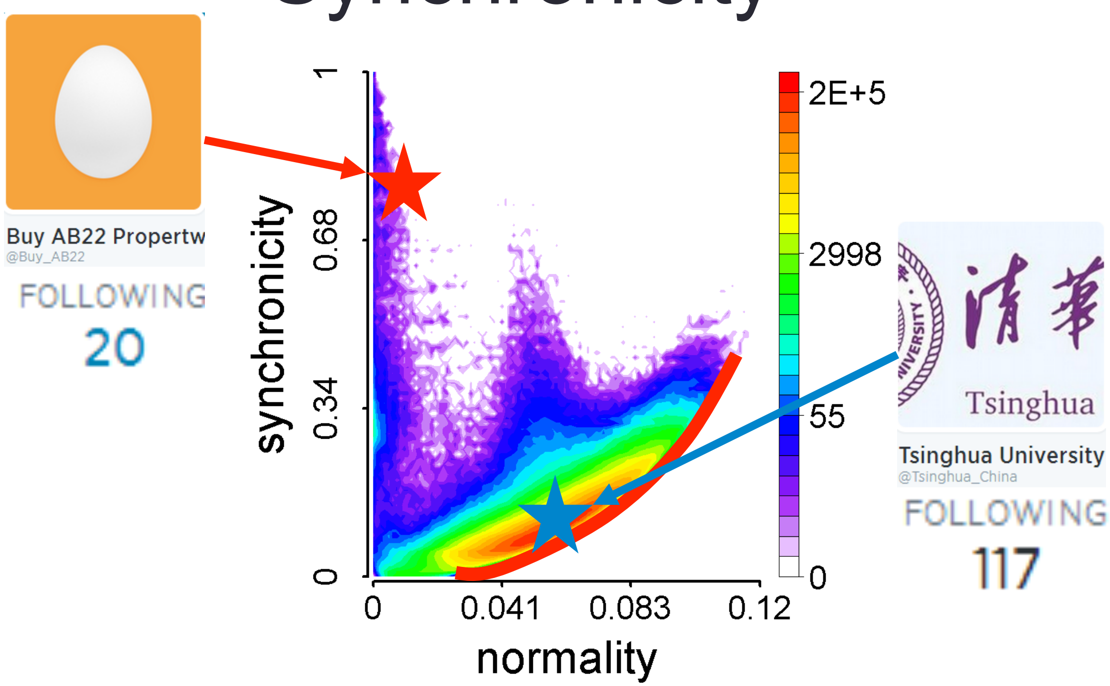
Theorem: Given Normality, We have Parabolic Lower Bound of Synchronicity

- For any distribution, “Synchronicity-Normality” plot has a parabolic lower limit

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

Synchronicity

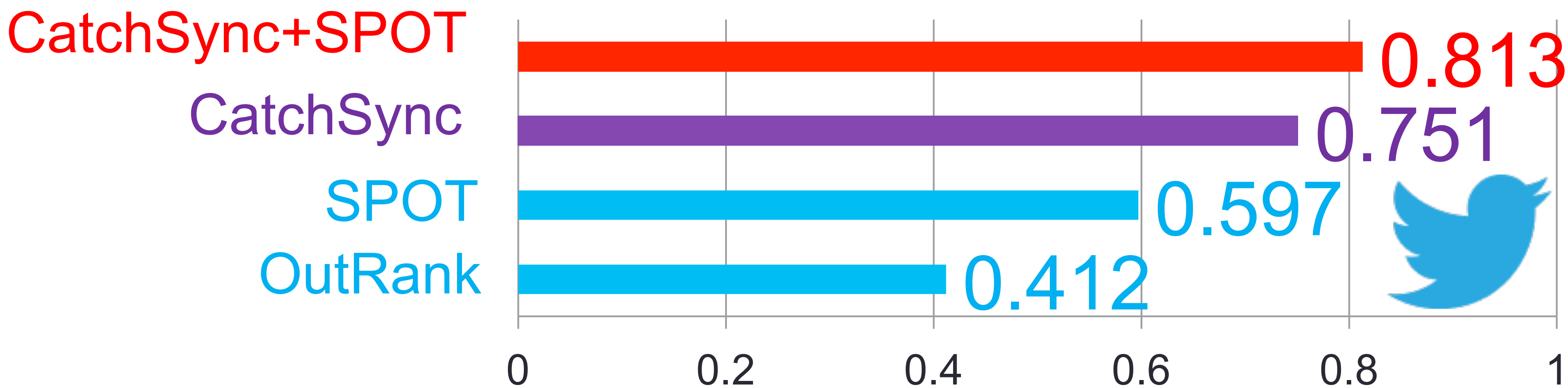
Normality

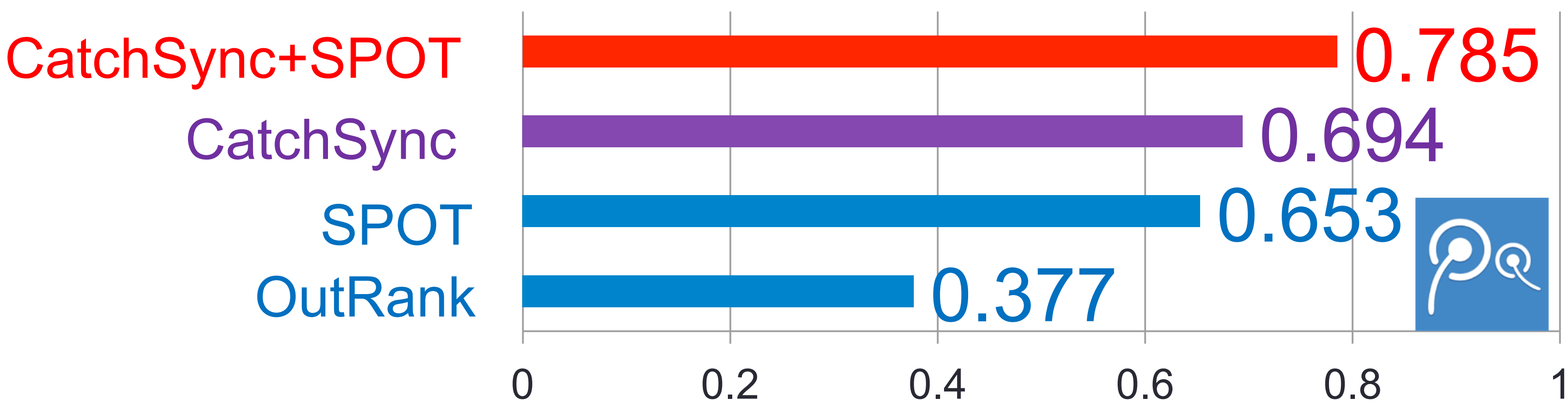


CatchSync:
Distance-based anomaly detection

Performance: Detecting Labelled Suspicious Accounts

- Improving detection accuracy: Complementary between Behavior-based CatchSync and Content-based SPOT

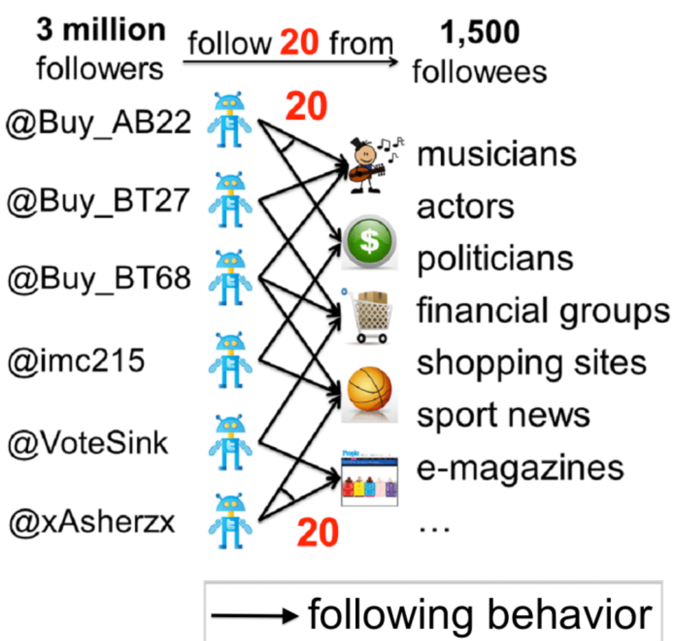


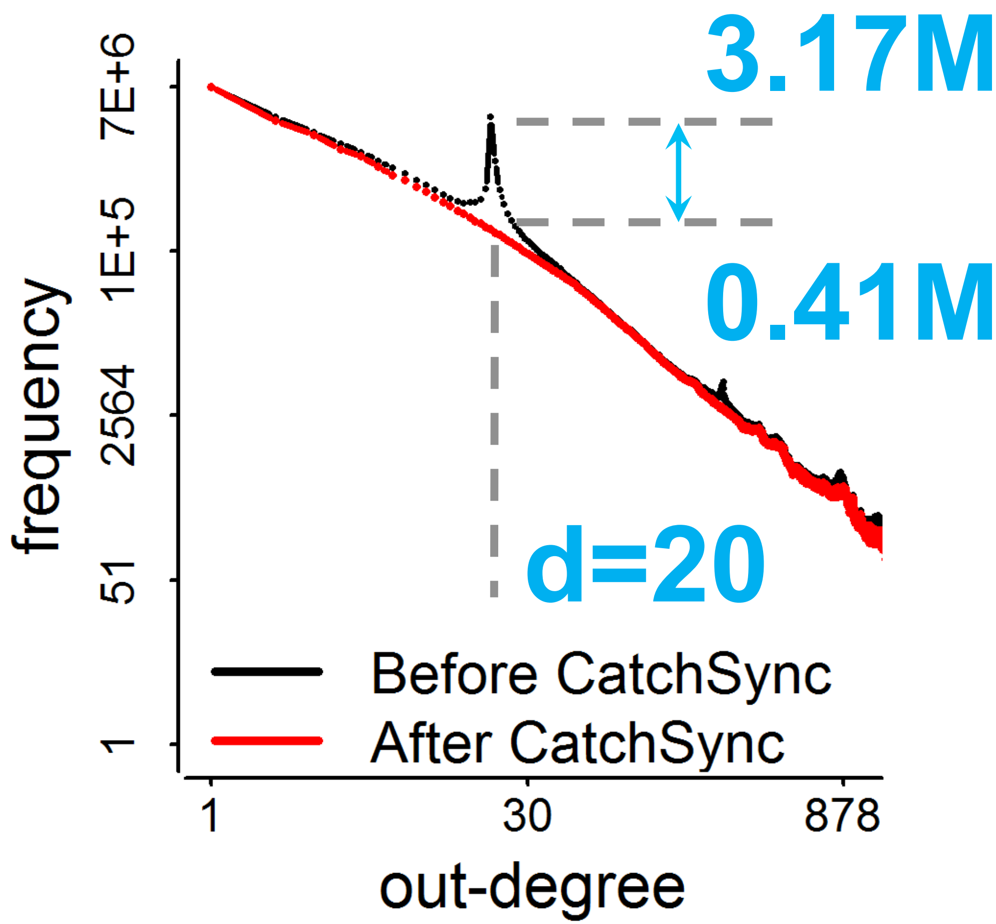


Performance: Recovering Distorted Out-degree Distribution

2009

41M





Contributions

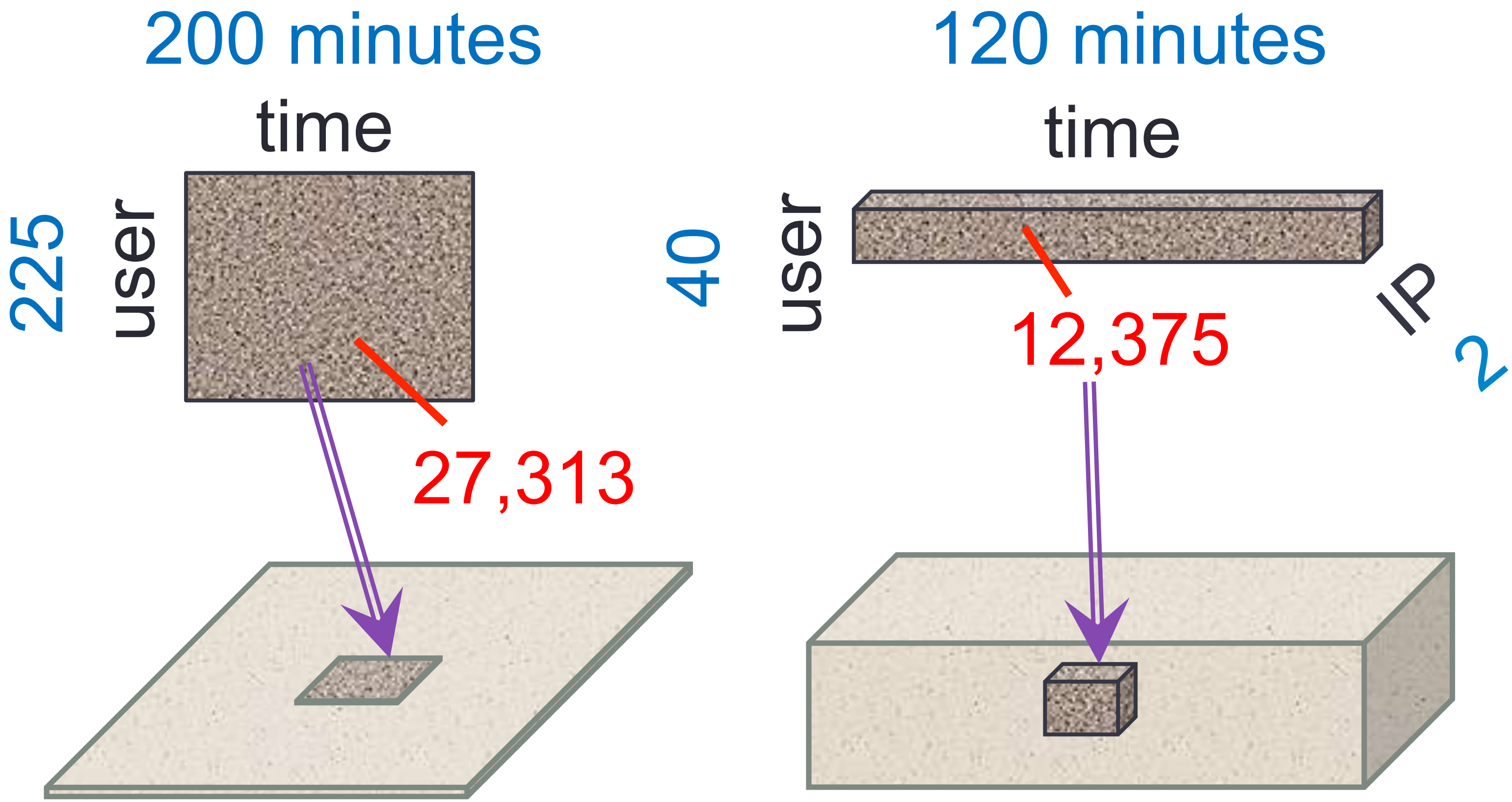
- Analyzed synchronized and abnormal behavioral patterns of zombie followers
- Proposed CatchSync: Detecting synchronized behaviors in large-scale graphs
- Recovered out-degree distribution to smooth and power-law-like
- Publication
  - ACM SIGKDD 2014 (Full. Acc. rate=14.6%. **Best paper finalist.**)
  - ACM TKDD 2015 (to appear. Special issue.)
  - Citation count: **18**

P6 & C: Measuring Suspicious Behaviors in Multi-modal Data

- 2 modes, and 3 modes, which is more suspicious?

Dense block:

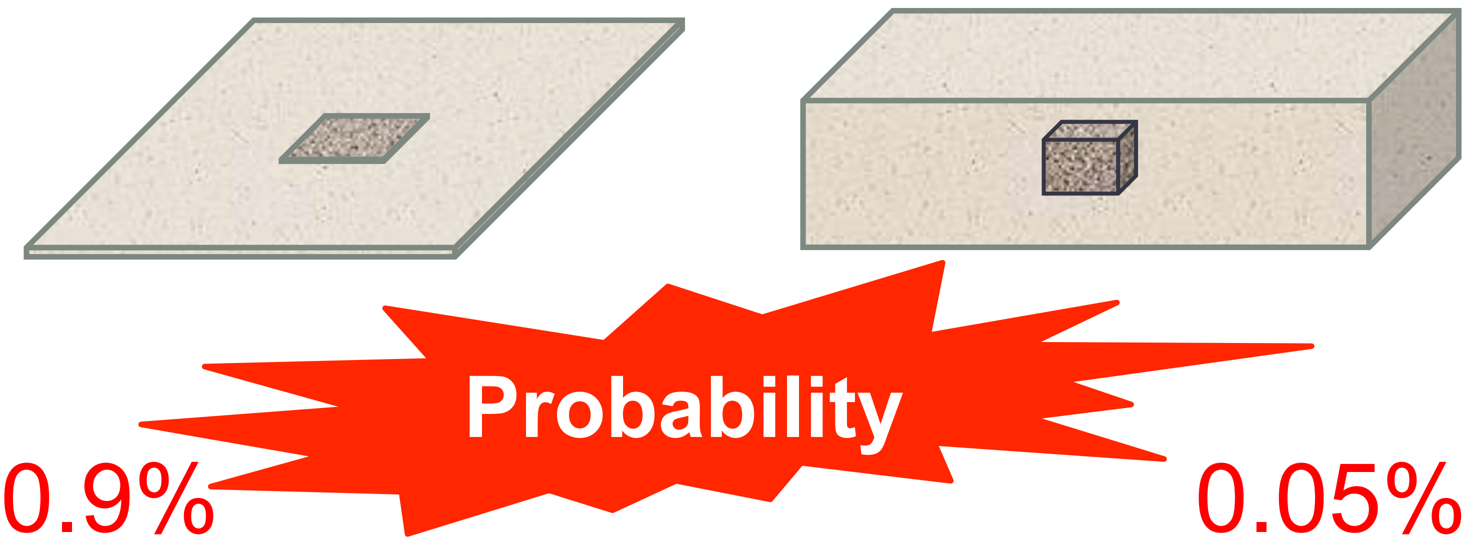
Data:



Idea: Multi-modal Suspiciousness

- Measuring multi-modal dense blocks

Dense block+Data:



More suspicious
Priority for detecting

Metric: "Suspiciousness"

- Suspiciousness = Negative log likelihood of block's probability under *Erdos-Renyi-Poisson*

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

- Local search
- CrossSpot

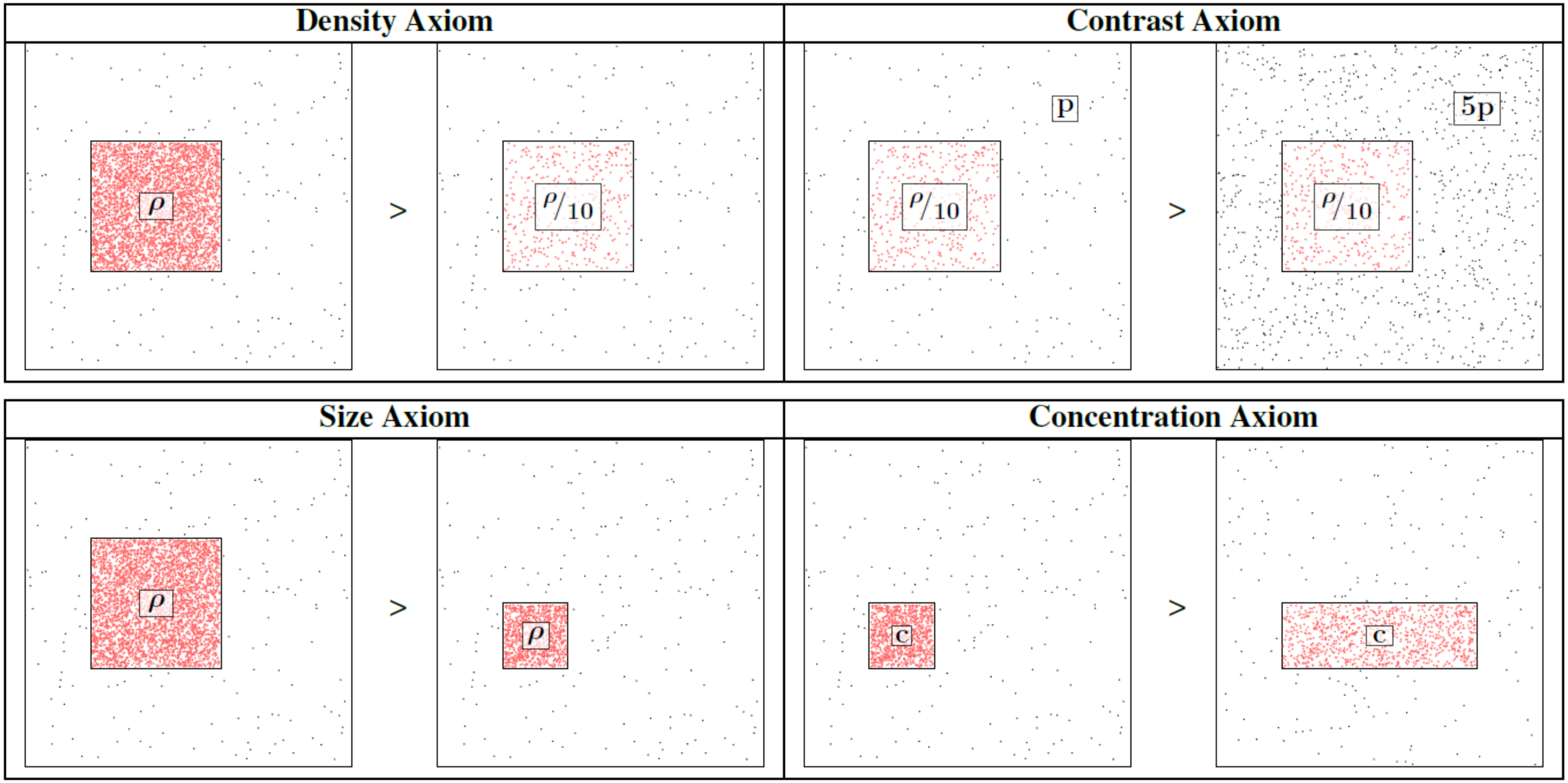
Lemma *Given an $n_1 \times \cdots \times n_K$ block of mass c in $N_1 \times \cdots \times N_K$ data of total mass C, the suspiciousness function is*

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

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

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

Satisfying Axioms



Advantage: “Suspiciousness”+CrossSpot

- Scoring dense blocks
- Targeting multi-modal data
- Satisfying axioms

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

Performance: Synthetic Data

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

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

Three Real Datasets

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

Manipulating Popular Trends

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

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

Contributions

- Cross modes with probability: Proposed a novel metric “suspiciousness” for multi-modal behaviors
- CrossSpot: Proposed local search algorithm for suspicious behaviors

- Publication
  - IEEE ICDM 2015 (Short. Acc. rate=18.2%.)

Summary

Complex Behaviors in Social Media: Analysis, Models and Applications

Contextual behavior

1. Social contexts for information adoption
2. Spatial temporal contexts for evolutionary analysis

Cross-domain/platform

3. Hybrid random walk for cross-domain modeling
4. Semi-supervised transfer for cross-platform modeling

True/False, Honest/Suspicious

5. Detecting synchronized behaviors
6. Measuring suspiciousness of multi-modal user behaviors

Summary

Contextual

1. ContextMF

[CIKM'12][TKDE'14]

2. FEMA

[KDD'14]

Models & Algorithms

Cross-domain
Cross-platform

3. HybridRW

[CIKM'12]
[TKDE'15]

4. XPTrans

Suspicious

5. CatchSync

[KDD'14 best finalist]

[TKDD'15]

6. CrossSpot

[ICDM'15]

Achievements

- **10/12** papers as the 1st author
  - 3×IEEE/ACM Trans. (3×Regular)
  - 7×Top Conf. (5×Full, 1×Short, 1×Poster)
- Selected papers
  - "Catching sync…": SIGKDD'14 **best paper finalist**
  - "Social context…": CIKM'12 & TKDE'14 (Cited by **85**)
- Total citation count: 290
- Awards
  - National Scholarship
  - Sohu research scholarship

Journal Papers

- **Meng Jiang**, Peng Cui, Fei Wang, Wenwu Zhu and Shiqiang Yang. "Social Recommendation with Cross-Domain Transferable Knowledge", in IEEE TKDE 2015. (to appear. Regular. IF=1.815.)
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. "Catching Synchronized Behaviors in Large Networks: A Graph Mining Approach", in ACM TKDD 2015. (to appear. Full. IF=1.147.)
- **Meng Jiang**, Peng Cui, Fei Wang, Wenwu Zhu and Shiqiang Yang. "Scalable Recommendation with Social Contextual Information", in IEEE TKDE 2014. (Regular. IF=1.815. 10 citations till 08/2015.)
- Lu Liu, Feida Zhu, **Meng Jiang**, Jiawei Han, Lifeng Sun and Shiqiang Yang. "Mining Diversity on Social Media Networks", in Multimedia Tools and Applications 2012.

Conference Papers

- **Meng Jiang**, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang and Christos Faloutsos. “A General Suspiciousness Metric for Dense Blocks in Multimodal Data”, in IEEE ICDM 2015. (Short. Acc. Rate=18.2%.)
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. “CatchSync: Catching Synchronized Behavior in Large Directed Graph”, in ACM SIGKDD 2014. (Full. **Best paper finalist**. Acc. rate=14.6%. 9 citations till 09/2015.)
- **Meng Jiang**, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu and Shiqiang Yang. “FEMA: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery”, in ACM SIGKDD 2014. (Full. Acc. rate=14.6%.)
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. “Inferring Strange Behavior from Connectivity Pattern in Social Networks”, in PAKDD 2014. (Full. Acc. rate=10.8%. 10 citations till 08/2015.)

Conference Papers (cont.)

- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. "Detecting Suspicious Following Behavior in Multimillion-Node Social Networks", in WWW 2014. (Poster. 9 citations till 09/2015.)
- **Meng Jiang**, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu and Shiqiang Yang. "Social Contextual Recommendation", in CIKM 2012. (Full. Acc. rate=13.4%. 74 citations till 09/2015.)
- **Meng Jiang**, Peng Cui, Fei Wang, Qiang Yang, Wenwu Zhu and Shiqiang Yang. "Social Recommendation across Multiple Relational Domains", in CIKM 2012. (Full. Acc. rate=13.4%. 32 citations till 08/2015.)
- Lu Liu, Jie Tang, Jiawei Han, **Meng Jiang** and Shiqiang Yang. "Mining Topic-Level Influence in Heterogeneous Networks", in CIKM 2010.

Submitted Papers

- **Meng Jiang**, Peng Cui, and Christos Faloutsos. "Suspicious Behavior Detection: Current Trends and Future Directions", to IEEE Intelligent Systems Magazine Special Issue on Online Behavioral Analysis and Modeling (IS, submitted).
- **Meng Jiang**, Peng Cui, Nicholas Jing Yuan, Xing Xie, and Shiqiang Yang. "Little is Much: Bridging Cross-Platform Behaviors Through Small Overlapped Crowds", to AAAI Conference on Artificial Intelligence (AAAI, submitted).
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. "Inferring Lockstep Behavior from Connectivity Pattern in Large Graphs", to Knowledge and Information Systems (KAIS, accepted with minor revision).

THANK YOU!