User Behavior Analysis in Social Networks: Pattern Discovery, Prediction and Anomaly Detection

Meng Jiang 蒋 朦
www.meng-jiang.com

My Group in Tsinghua

Professor Jianping Wu (Network)
Professor Maosong Sun (NLP)

Professor Yuanchun Shi (HCI)
Professor Shimin Hu (Graphics)

Professor Shiqiang Yang (Multimedia)
Professor Jiangtao Wen (Net-of-Things)
Professor Lifeng Sun (Multimedia)

Professor Wenwu Zhu (Multimedia)
Assistant Professor Peng Cui (DM & Multimedia)

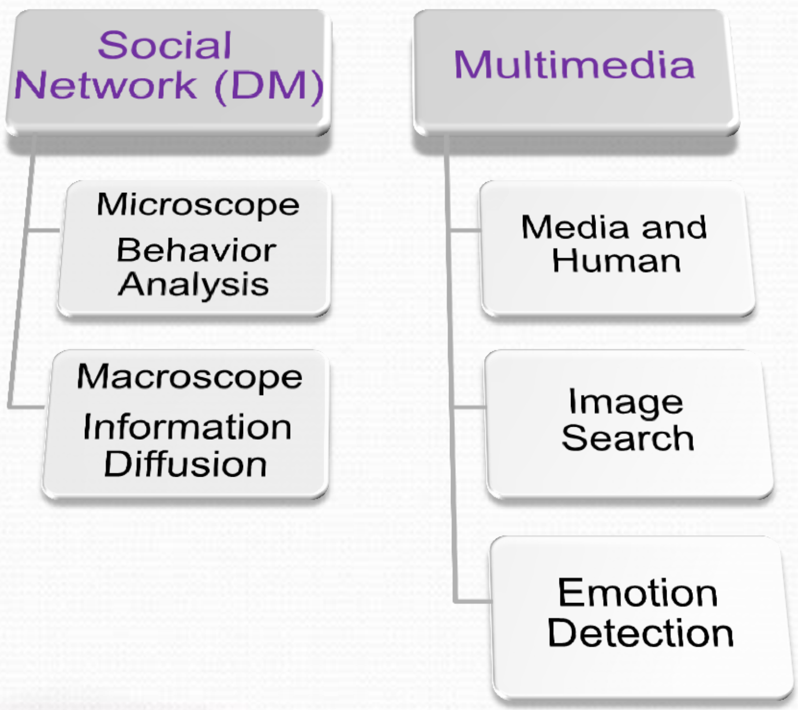
Department of Computer Science and Technology

Institute of HCI and Media Integration

Lab of Multimedia and Networking

Social Media Lab

Social-sensed Multimedia Computing



- Behavior analysis
  - Pattern discovery and prediction [KDD'14]
  - Social recommendation [CIKM'12, CIKM'12, TKDE'14]
  - Anomaly detection [PAKDD'14, WWW'14 Poster, KDD'14]
- Information diffusion
  - Item-level influence [AAAI'11, SIGIR'11]
  - Outbreak detection [KDD'13]
- Big social data
  - Heterogeneous hashing [KDD'13]

- Multimedia computing [TMM'09, DMKD'11, TIP'12, TMM'12, MM'12 Best, TOMCCAP'13, TOMCCAP'14]
- Image search [MMM'13 Best, TOIS'14, CVIU'14]
- Emotion detection [MM'13, ICMR'14]

Tsinghua University

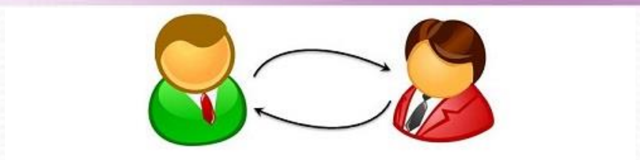
Social-sensed Multimedia Computing

- Social Network (DM)
  - Microscope Behavior Analysis
  - Macroscope Information Diffusion
- Multimedia
  - Media and Human
  - Image Search
  - Emotion Detection

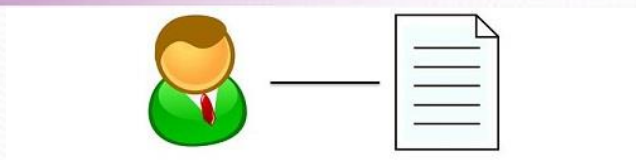
- **Behavior analysis**
  - **Pattern discovery and prediction [KDD'14]**
  - **Social recommendation [CIKM'12, CIKM'12, TKDE'14]**
  - **Anomaly detection [PAKDD'14, WWW'14 Poster, KDD'14]**
- Information diffusion
  - Item-level influence [AAAI'11, SIGIR'11]
  - Outbreak detection [KDD'13]
- Big social data
  - Heterogeneous hashing [KDD'13]

- Multimedia computing [TMM'09, DMKD'11, TIP'12, TMM'12, MM'12 Best, TOMCCAP'13, TOMCCAP'14]
- Image search [MMM'13 Best, TOIS'14, CVIU'14]
- Emotion detection [MM'13, ICMR'14]

My Research Interests



User-user link

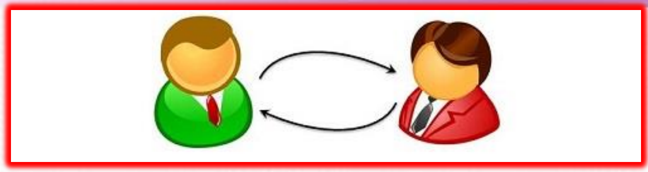


User-item link

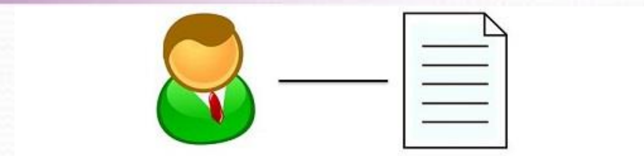
Normal use

Anomalous use

My Research Interests



User-user link



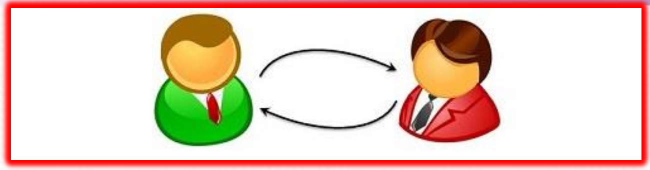
User-item link

Normal use

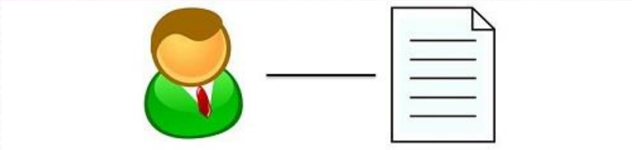
Influence mining
Community detection
Friend recommendation

Anomalous use

My Research Interests



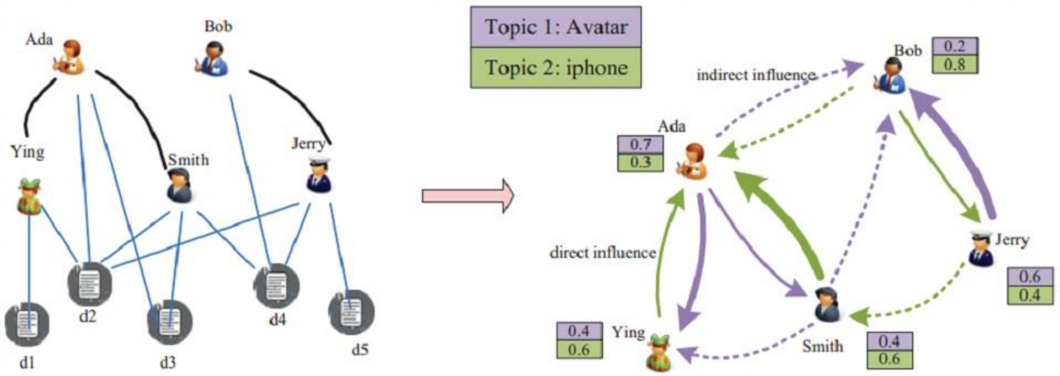
User-user link



User-item link

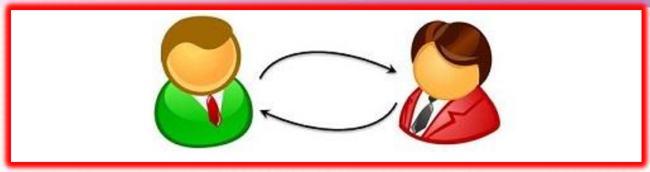
Normal use

Topic-level influence [Liu et al. CIKM 2010]

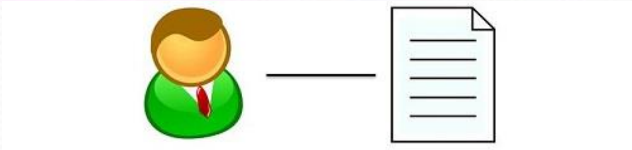


Anomalous use

My Research Interests



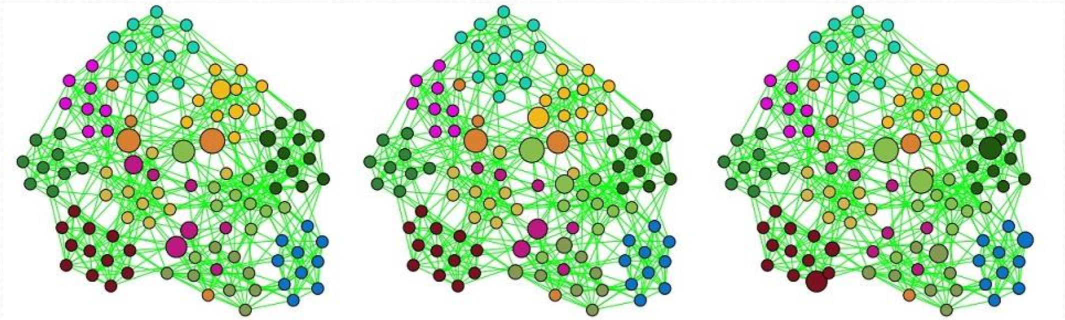
User-user link



User-item link

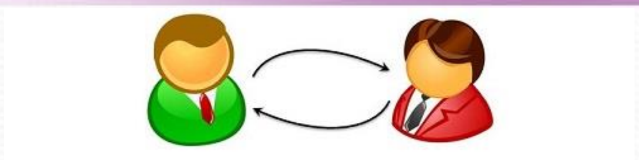
Normal use

User diversity [Lu et al. Multimed Tools Appl 2012]

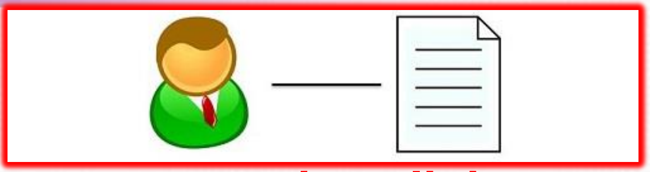


Anomalous use

My Research Interests



User-user link



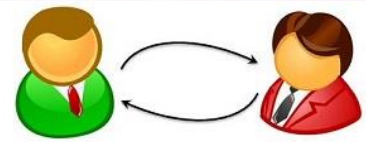
User-item link

Behavior prediction
Social recommendation
Behavior pattern discovery

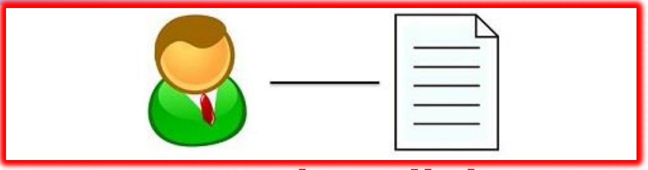
Normal use

Anomalous use

My Research Interests



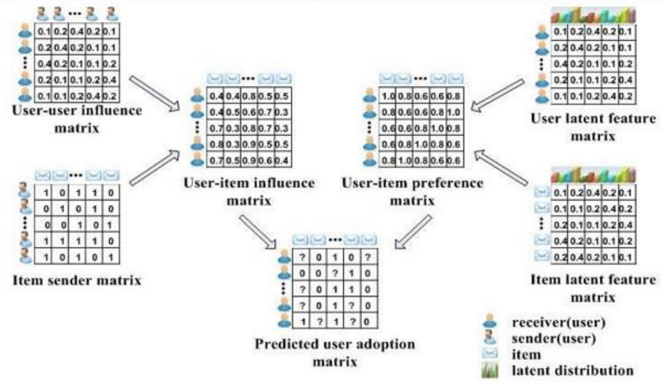
User-user link



User-item link

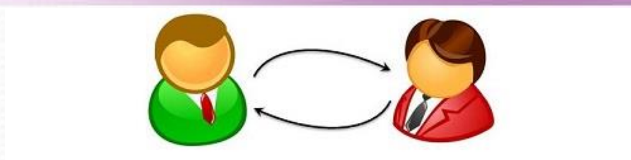
Normal use

Social contextual recommendation [Jiang et al. CIKM 2012]

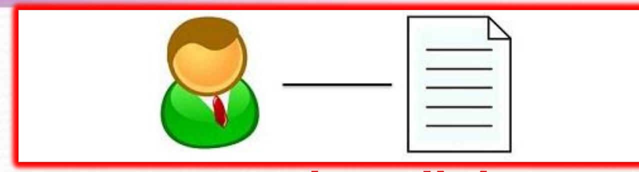


Anomalous use

My Research Interests



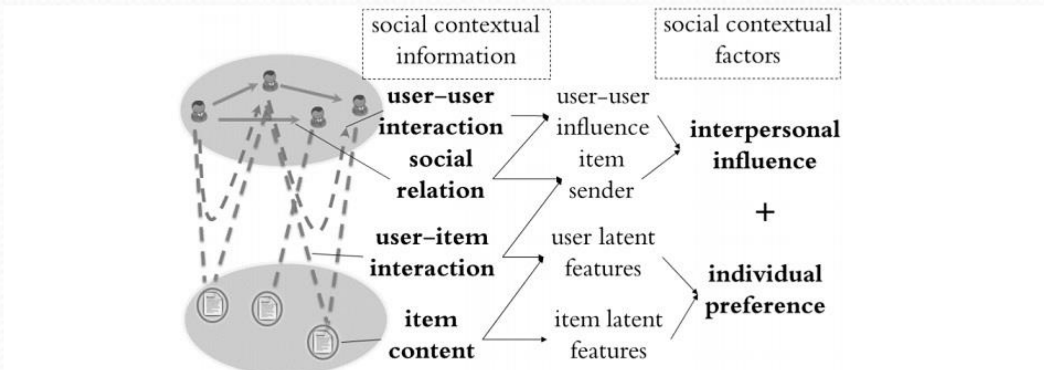
User-user link



User-item link

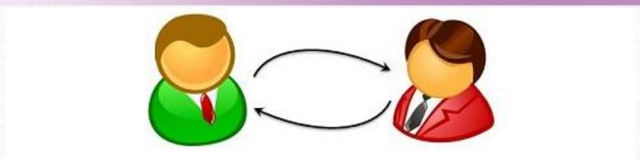
Normal use

Scalable recommendation with social context
[Jiang et al. TKDE 2014]

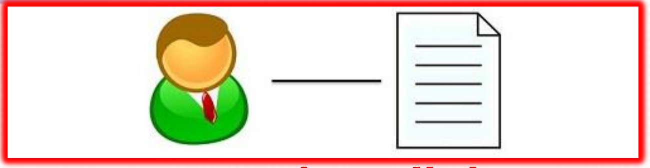


Anomalous use

My Research Interests



User-user link

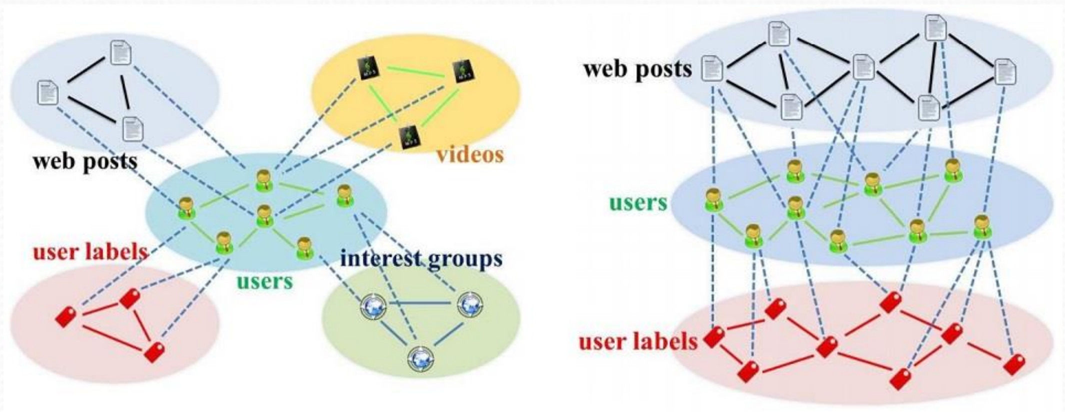


User-item link

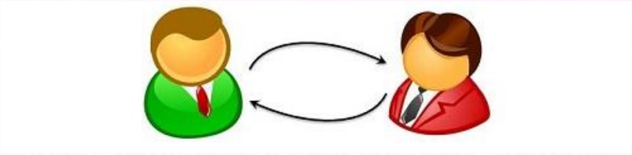
Normal use

Anomalous use

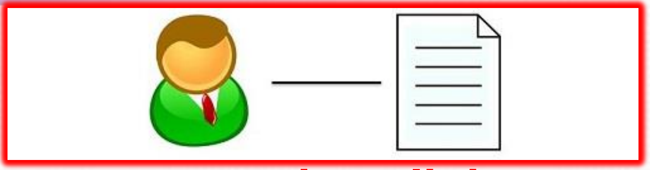
Cross-domain social recommendation [Jiang et al. CIKM 2012]



My Research Interests



User-user link

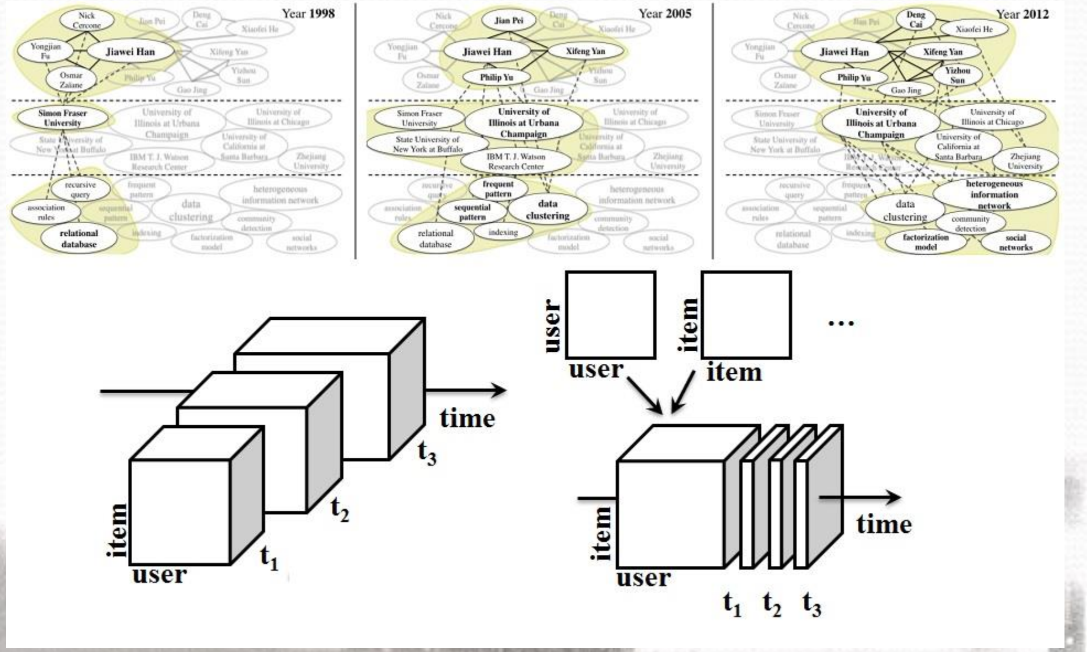


User-item link

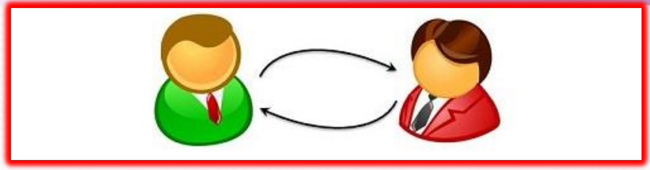
Normal use

Anomalous use

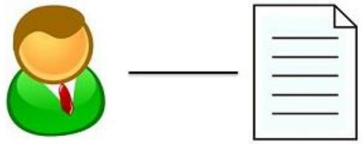
Behavior modeling and pattern discovery [Jiang et al. KDD 2014]



My Research Interests



User-user link



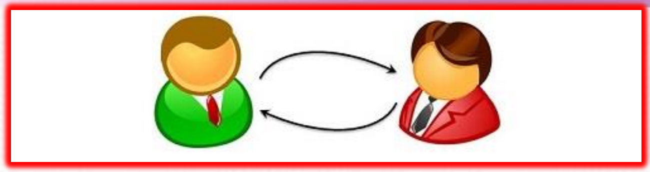
User-item link

Normal use

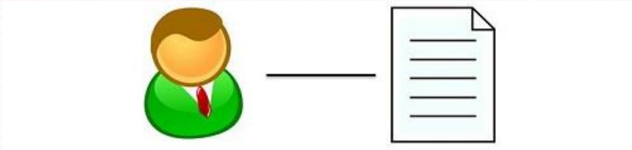
Anomalous use

Zombie follower detection
Suspicious behavior patterns

My Research Interests



User-user link

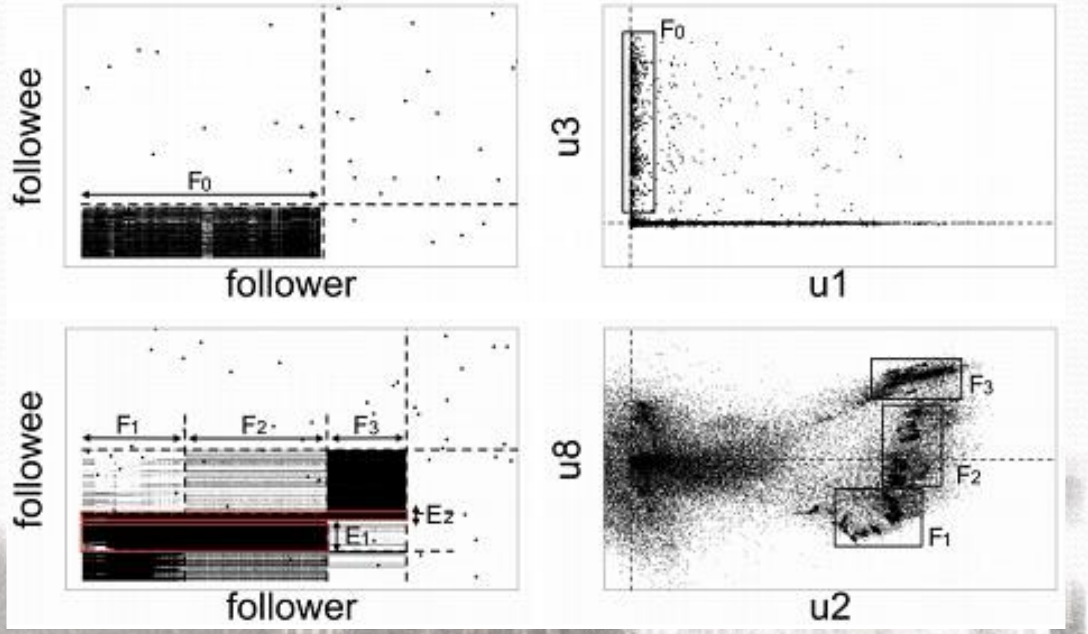


User-item link

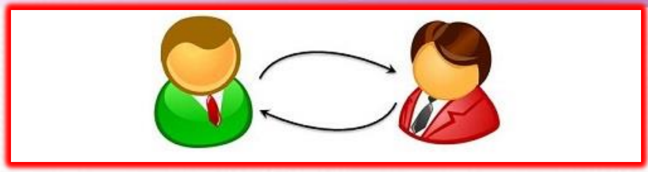
Normal use

Anomalous use

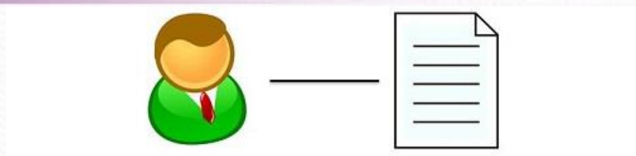
Dense bipartite core detection [Jiang et al. PAKDD 2014]



My Research Interests



User-user link



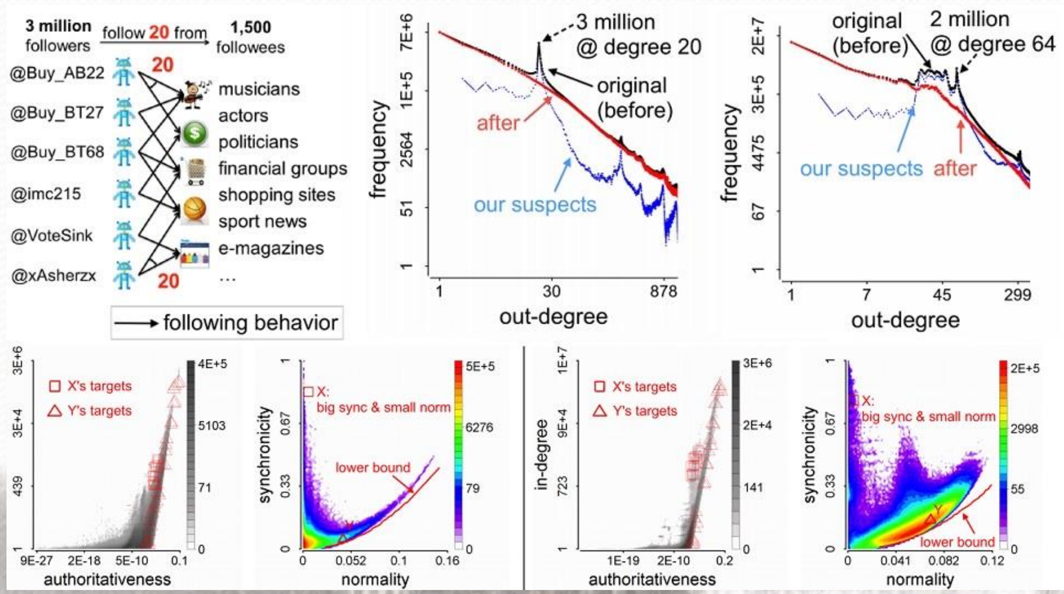
User-item link

Normal use

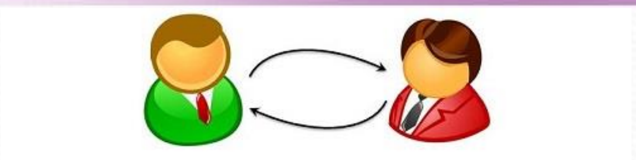
Anomalous use

Zombie follower detection

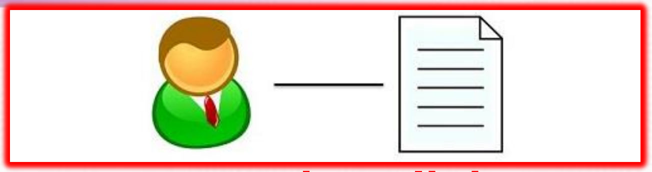
[Jiang et al. WWW 2014 Poster, KDD 2014]



My Research Interests



User-user link



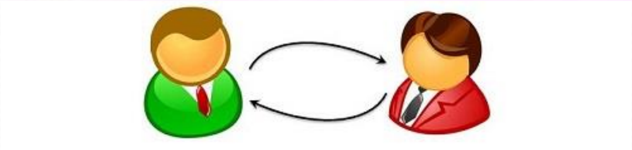
User-item link

Normal use

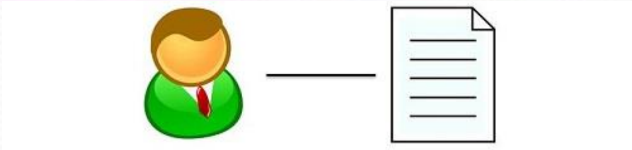
Anomalous use

Spam/fraud detection
Promoting behavior patterns
[On-going]

My Research Interests

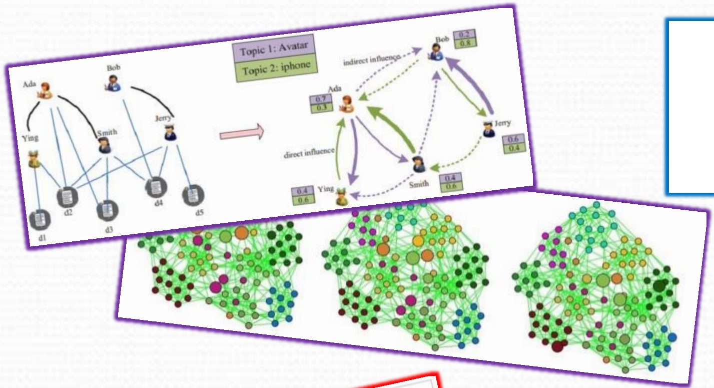


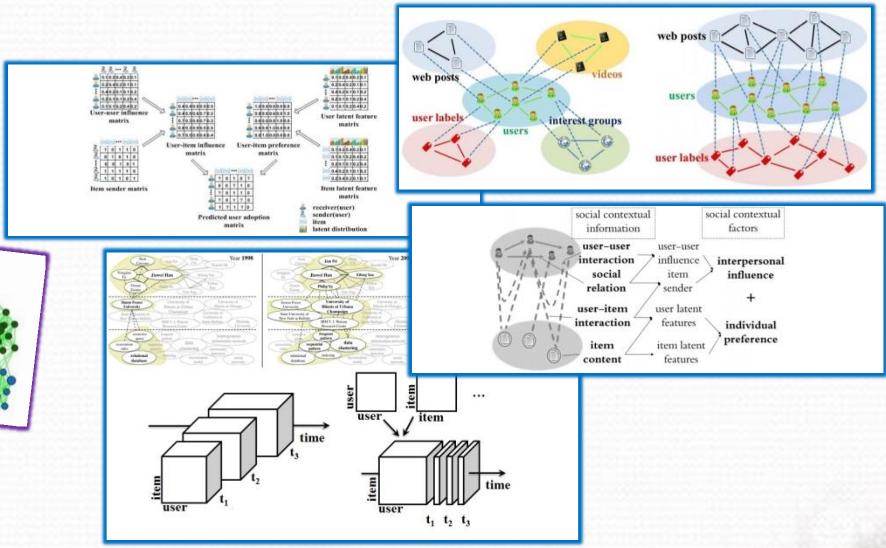
User-user link



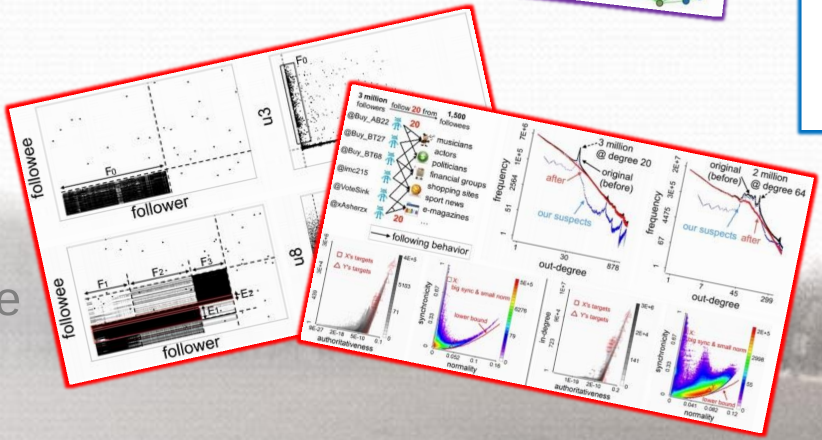
User-item link

Normal use





Anomalous use



Behavior analysis

Pattern discovery and prediction [KDD'14]

Social recommendation [CIKM'12, CIKM'12, TKDE'14]

Anomaly detection [PAKDD'14, WWW'14 Poster, KDD'14]

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

Good user-item links

- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

Good user-item links

- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

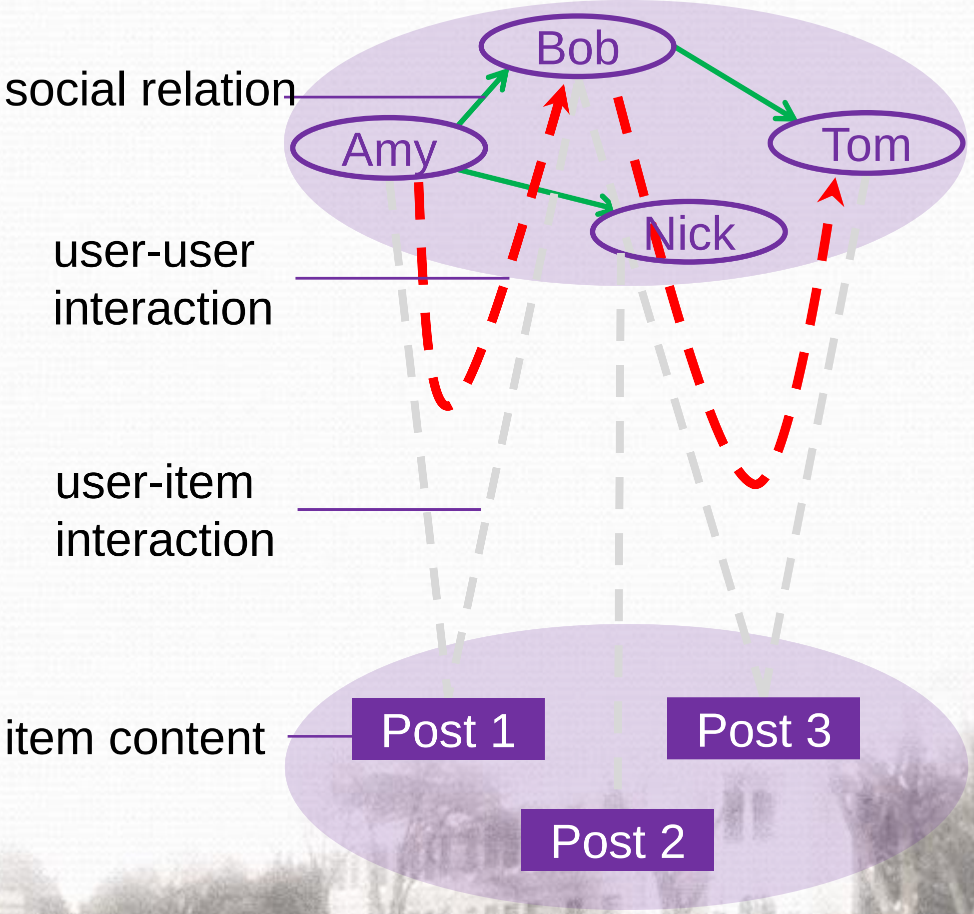
Social Recommendation

- Problem
  - Too many messages are generated and received every minute.
  - How to recommend posts/rank feeds in social networks?
- Challenges
  - **High sparsity** for user-item matrix factorization/completion

	User 1	User 2	…	User M
Post 1	**1**	**1**	**…**	**1**
Post 2	**0**	**?**	**…**	**?**
…	**…**	**…**	**…**	
Post N	**1**	**?**	**…**	**1**

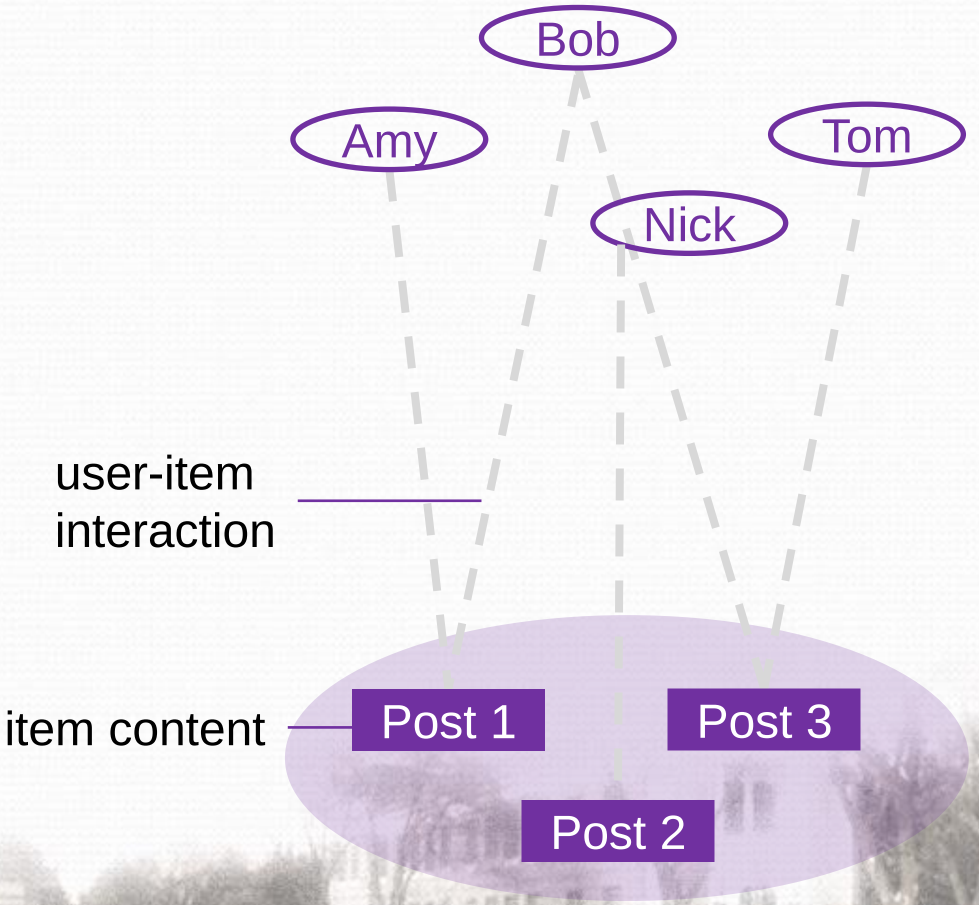
Social Contextual Information

- More knowledge! But …
- Can we use social relations (user-user links) to help infer missing user-item links?
- Can we fully use social contextual information on the networks?



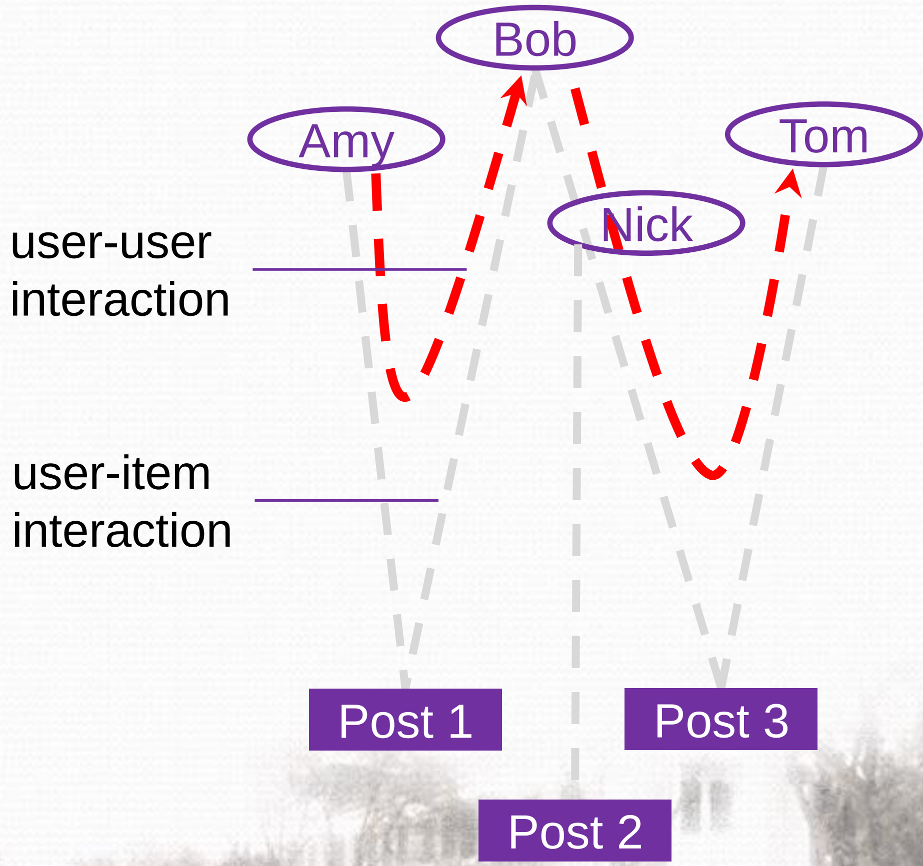
Related Works

- Content-based filtering
- Collaborative filtering



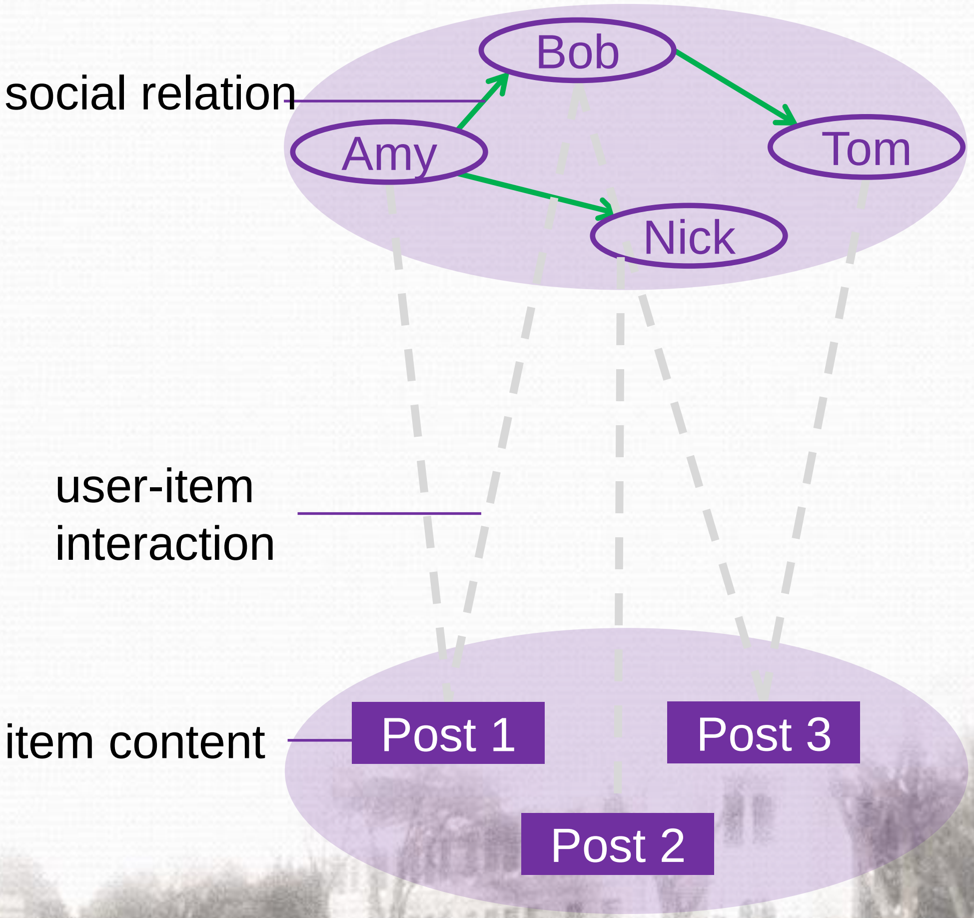
Related Works

- Content-based filtering
- Collaborative filtering
- Trust-based recommendation
- Influenced-based recommendation



Related Works

- Content-based filtering
- Collaborative filtering
- Trust-based recommendation
- Influenced-based recommendation
- Social recommendation with MF/Social Regularization



Related Works

	Social relation	User-user interaction	User-item interaction	Item content
Content-based & CF	×	×	√	√
Trust & Influence	×	√	√	×
SoRec & SoReg	√	×	√	√
?	√	√	√	√

- How to fully use contextual information for social recommendation?
- Answer: Understand **user intention of adopting messages**.

User Intention of Adopting Messages

Peng Cui : Is there anyone who call for paper via Renren? Hah! http://media.cs.tsinghua.edu.cn/~multimedia/cuipeng/IR_SI_Social Media.htm

2011-01-05 13:47 Reply | Share

Call for paper? About social media? Wow! **Personal Preference**

This is my best friend and co-author! **Interpersonal influence**

Meng Jiang : Support! //Peng Cui : Is there anyone who call for paper via Renren? Hah! http://media.cs.tsinghua.edu.cn/~multimedia/cuipeng/IR_SI_Social Media.htm

2011-01-05 14:05 Reply | Share

User Intention of Adopting Messages

Maosong Sun : KDD Summer School on Mining the Big Data will be held in Tsinghua. This is the first time for KDD to hold Summer School. Dean Xiaoyong Du, Dr. Hang Li and me are the Chairs. Today Jiawei Han(UIUC), Christos Faloutsos(CMU) and Bing Liu(UIC) gave lectures for 2 hours each.

2012-08-10 21:47 Retweet | Save | Reply

Amazing! Summer school! It is KDD!
Jiawei Han! Christos Faloutsos! Bin Liu!

Personal Preference

This is the Dean of my Department!
His research area is Artificial Intelligence and Data Mining!

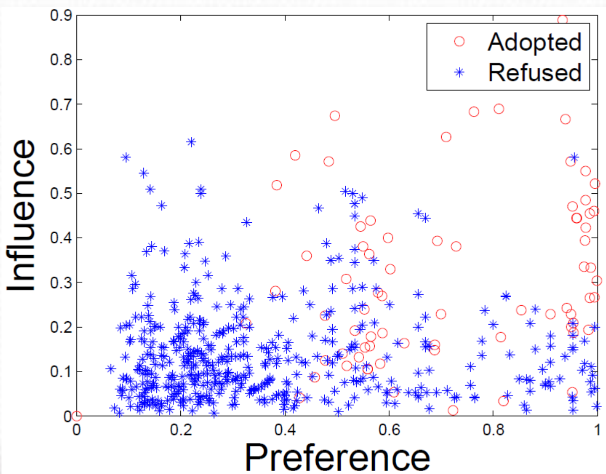
Interpersonal influence

Meng Jiang : Amazing! //Maosong Sun : KDD Summer School on Mining the Big Data will be held in Tsinghua. This is the first time for KDD to hold Summer School. Dean Xiaoyong Du, Dr. Hang Li and me are the Chairs. Today Jiawei Han(UIUC), Christos Faloutsos(CMU) and Bing Liu(UIC) gave lectures for 2 hours each.

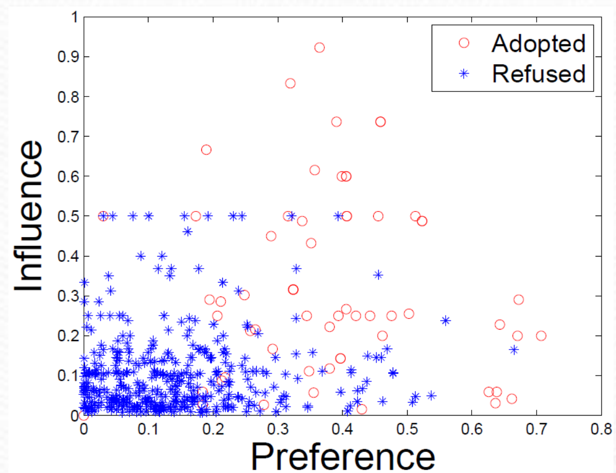
2012-08-11 09:35 Retweet | Save | Reply

User Intention of Adopting Messages

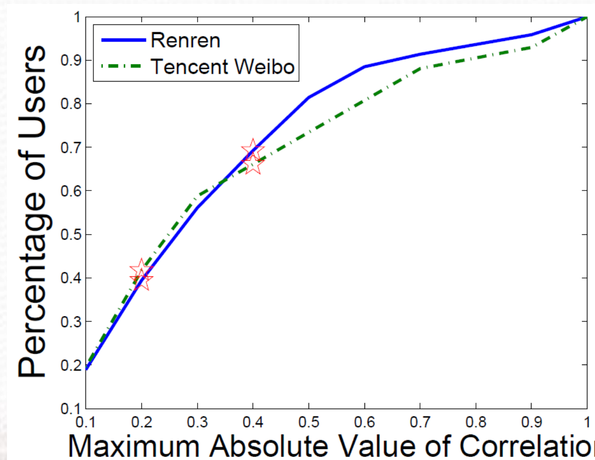
- Do we care the item content? Do we care the sender?
- Preference: topic-level user-item similarity
- Influence: user-sender interaction frequency



Renren

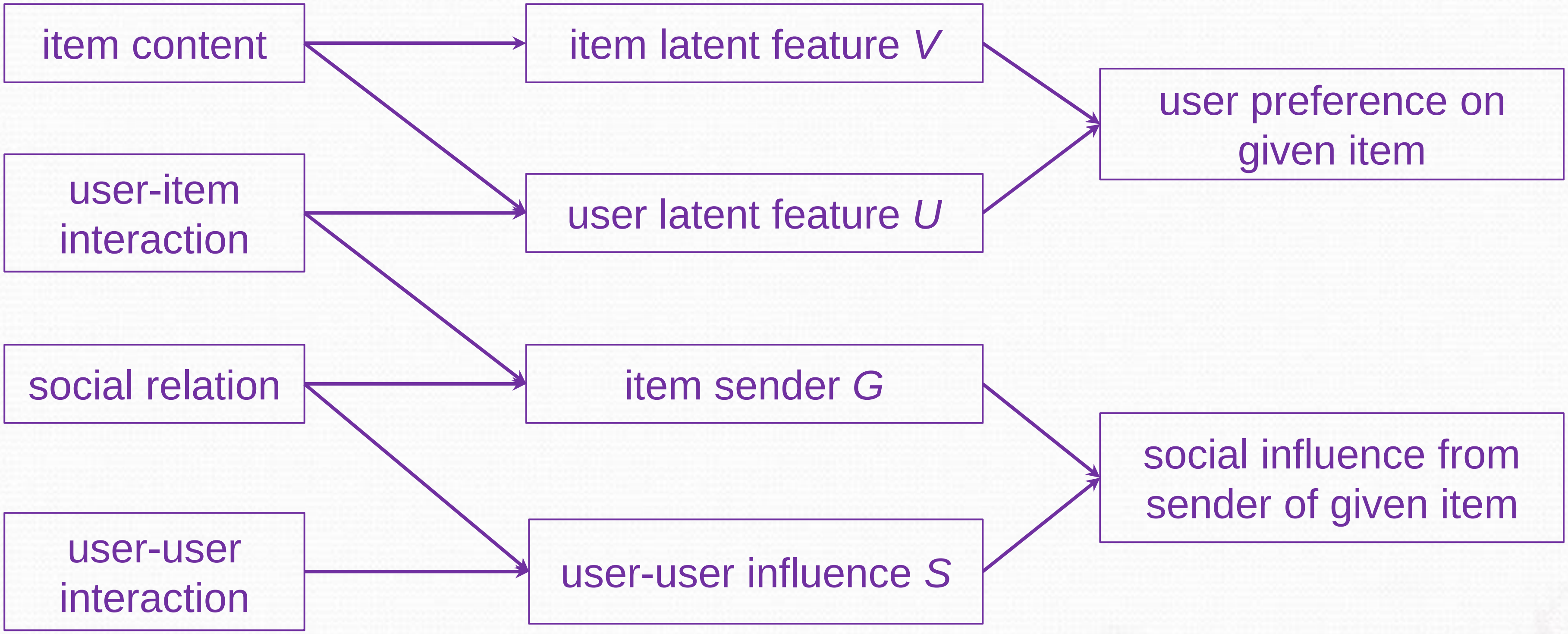


Tencent Weibo



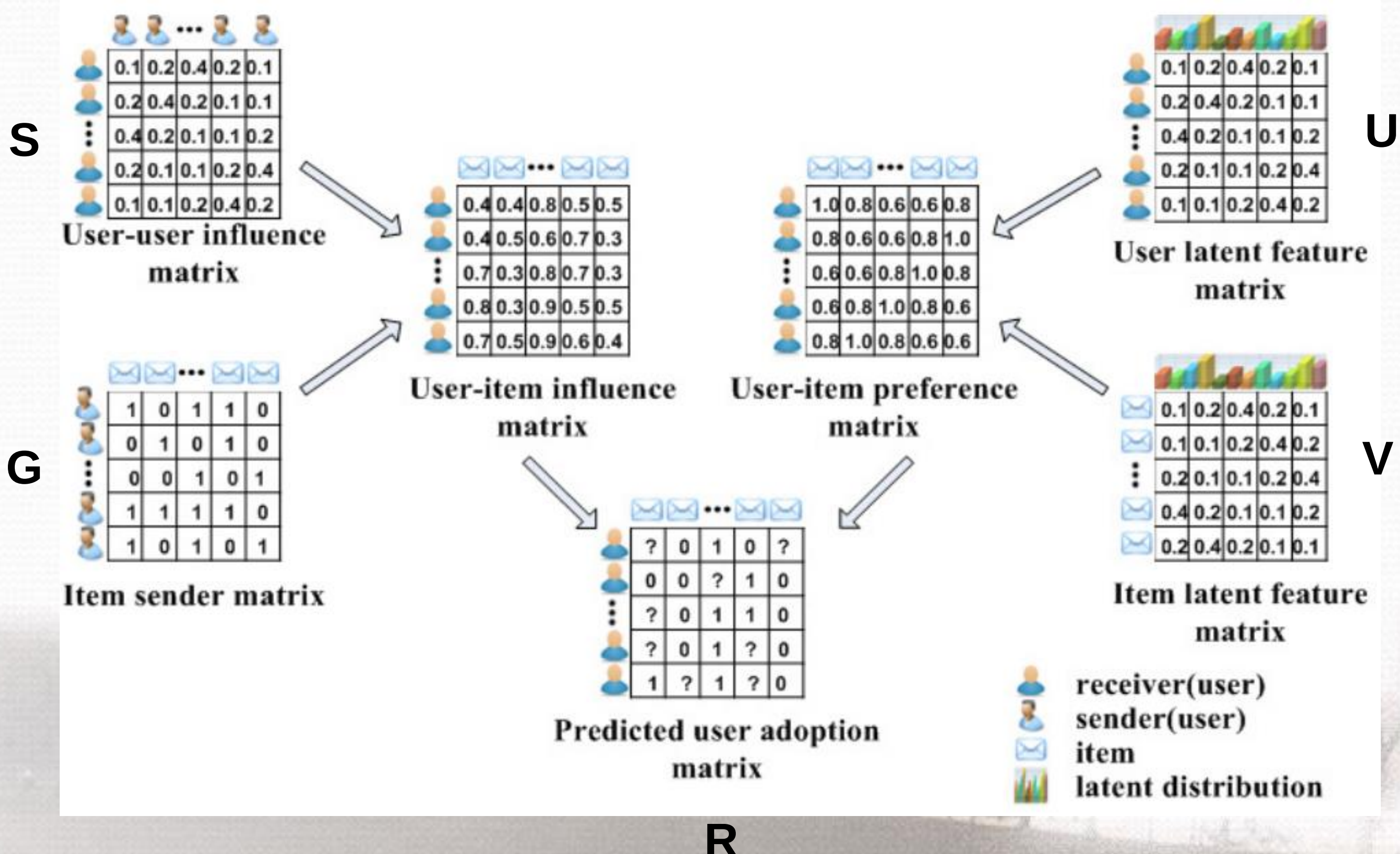
Correlation(Preference, Influence) is small.

Framework: Social Contextual Factors



Framework: Social Contextual Factors

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



ContextMF Algorithm

- Minimize sum-of-squared errors function

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

- Block coordinate descent scheme with gradients.

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

New Users and New Items Coming

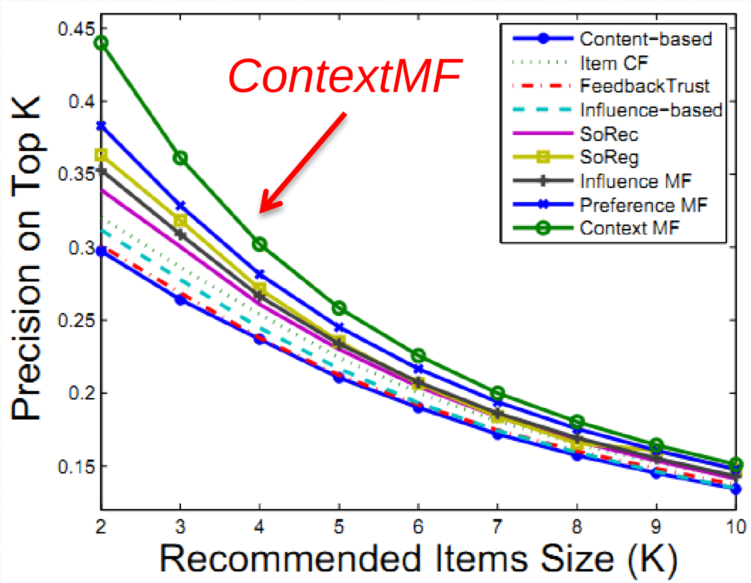
New data = Entries between…	**Old *N* items**	**New *ΔN* items**
Old *M* users	Done! $O(k^2L(M+N)^2)$	$\mathcal{J}_{\Delta V} = \|\|\mathbf{\Delta C} - \mathbf{\Delta V}^\top \mathbf{V}\|\|_F^2, \frac{\partial \mathcal{J}}{\partial \mathbf{\Delta V}} = -2\mathbf{V}\mathbf{\Delta C}^\top + O(\mathbf{\Delta V})$ $O(k^2L\Delta NN) << O(k^2LN(M+N))$
New *ΔM* users	$\mathcal{J}_{\Delta S} = \|\|\mathbf{\Delta F} - \mathbf{\Delta S}\|\|_F^2, \quad \frac{\partial \mathcal{J}}{\partial \mathbf{\Delta S}} = -2\mathbf{\Delta F} + O(\mathbf{\Delta S})$ $\mathcal{J}_{\Delta U} = \|\|\mathbf{\Delta W} - \mathbf{\Delta U}^\top \mathbf{U}\|\|_F^2, \quad \frac{\partial \mathcal{J}}{\partial \mathbf{\Delta U}} = -2\mathbf{U}\mathbf{\Delta W}^\top + O(\mathbf{\Delta U})$ $O(k^2L\Delta MM) << O(k^2LM(M+N))$	Sorry…

Effectiveness in Predicting Adoptions

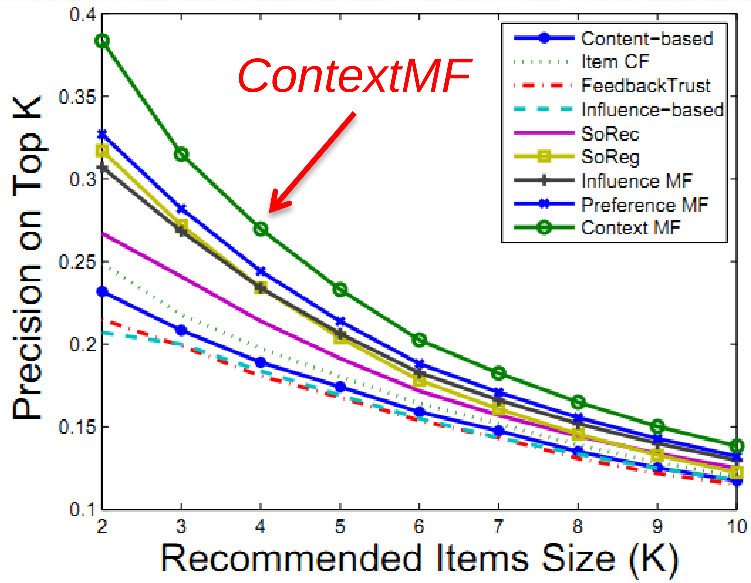
Method	MAE	RMSE	$\hat{\tau}$	$\hat{\rho}$
Renren Dataset				
Content-based [1]	0.3842	0.4769	0.5409	0.5404
Item CF [25]	0.3601	0.4513	0.5896	0.5988
FeedbackTrust [22]	0.3764	0.4684	0.5433	0.5469
Influence-based [9]	0.3859	0.4686	0.5394	0.5446
SoRec [19]	0.3276	0.4127	0.6168	0.6204
SoReg [20]	0.2985	0.3537	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%

Effectiveness in Predicting Adoptions



Renren



Tencent Weibo

	Renren	Tencent Weibo
Top-5 Precision	+21.7%	+12.3%
Top-10 Precision	+10.8%	+6.85%

Efficiency in Incremental Data

- Better than SoReg [Ma et al. WSDM 2011]
- A little bit worse than offline learning (re-training)
- Save time from hours to minutes when M,N ~ million and ΔM,ΔN ~ thousand

Dataset	RMSE (smaller is better)			ERR (bigger is better)			Time cost	
	$SoReg$	$\Delta ContextMF$	$ContextMF^{\Delta}$	$SoReg$	$\Delta ContextMF$	$ContextMF^{\Delta}$	$\Delta ContextMF$	$ContextMF^{\Delta}$
RΔM1000	0.342	0.263	0.257	0.555	0.610	0.636	172s	41.7h
RΔM10000	0.502	0.464	0.444	0.481	0.542	0.559	1610s	41.7h
TΔM1000	0.168	0.122	0.105	0.652	0.764	0.783	54.2s	2.42h
TΔM10000	0.342	0.333	0.317	0.534	0.611	0.651	531s	2.42h
RΔN1000	0.335	0.276	0.276	0.570	0.663	0.680	97.3s	41.7h
RΔN10000	0.546	0.478	0.465	0.514	0.587	0.609	941s	41.7h
TΔN1000	0.218	0.192	0.173	0.726	0.824	0.864	17.8s	2.42h
TΔN10000	0.427	0.376	0.355	0.658	0.720	0.751	160s	2.42h

Meng Jiang, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu and Shiqiang Yang. Social Contextual Recommendation. *The 21st ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.
Meng Jiang, Peng Cui, Fei Wang, Wenwu Zhu and Shiqiang Yang. Scalable Recommendation with Social Contextual Information. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2014.

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

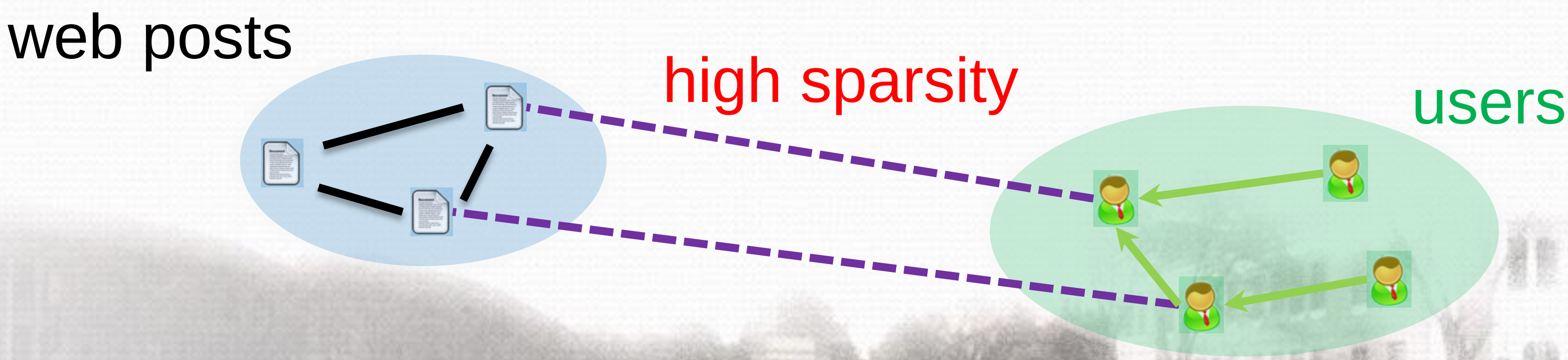
Good user-item links

- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

Social Recommendation

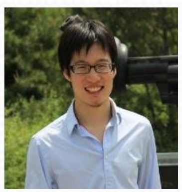
- Problem
  - Usually we train historical behavioral data (user-post links) to understand their interests. However, if history ~ 0?
  - How can we recommend posts for new(-registered) users?
- Challenges
  - **Cold-start and extremely high sparsity** in user-item domain



Social Recommendation

- We have social relation domain.
- We also have auxiliary knowledge in other domains.
- **User label domain**

Choose < 10 from 200+ labels like ‘iPhone fan’



Peng Cui
Haidian, Beijing
Company: Tsinghua

User labels (5)
Tsinghua, Ph.D., World Wide Web, Social Network, Social Media

Meng Jiang
Haidian, Beijing
University: Tsinghua

User labels (9)
Chinese food, World Wide Web, Social Network, Data Mining, Liverpool Football Club, NBA, Humors, Sports, Ph.D. Candidates

Social Recommendation

- We have social relation domain.
- We also have auxiliary knowledge in other domains.
- **Interest group domain**

Interest Groups (2)

Tsinghua University

I love sing!

Interest Groups (3)

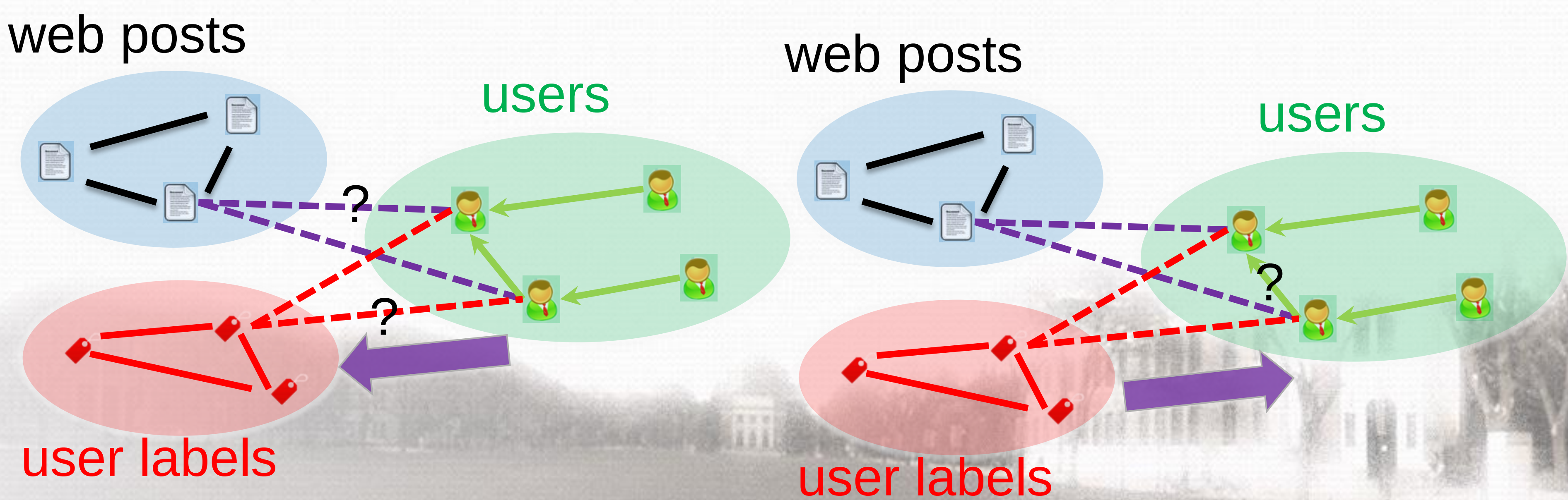
Tsinghua University

Social Media & Reputation

World Wide Web Team

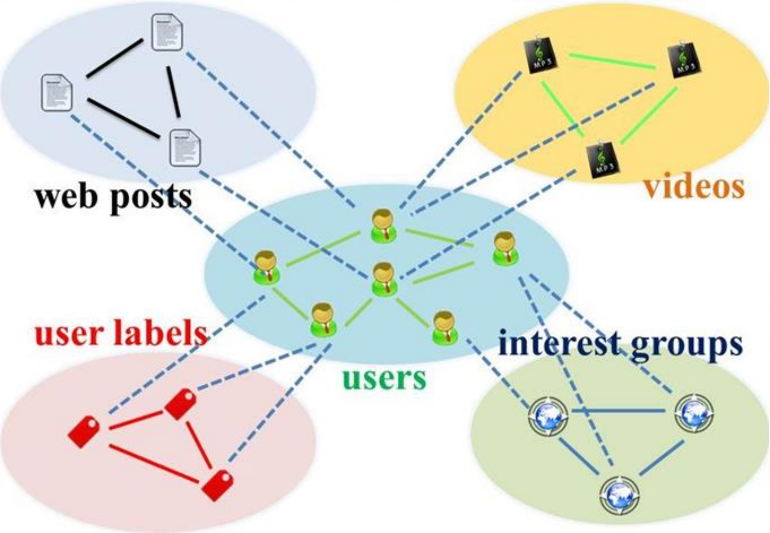
Social Recommendation

- How to construct social network with multiple domains?
  - We have user-post, user-label and user-group links.
  - No relations between item domains. No post-label links in nature.
  - Stronger social relations can help collaborate user-item links.
  - More collaborating in user-item links can strengthen the social relations.



Social Recommendation

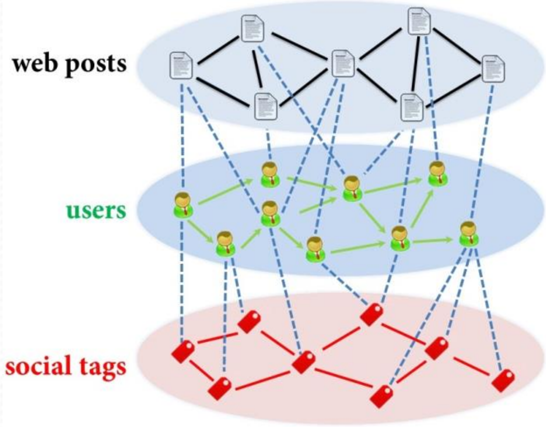
- Answer: A star-structured graph.
- Properties
  - Relational
    - Within-domain and cross-domain
  - Heterogeneous
    - Different types of item domains
  - Sparse
    - Different sparsity levels



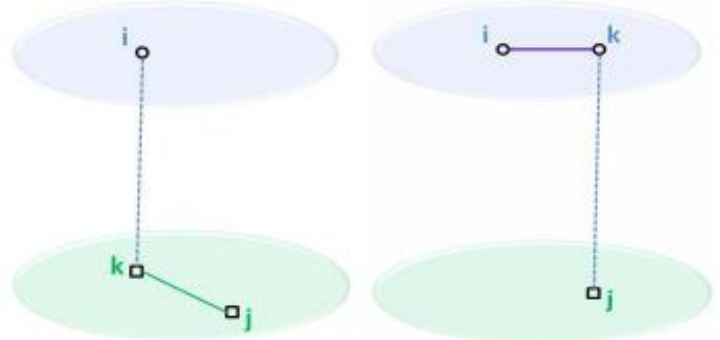
- Key idea
  - Use social relation domain as a bridge
- Method
  - Transfer learning + Random walk with restarts

Hybrid Random Walk

- On second-order star-structured graph



- Update cross-domain links



(a) Route 1 (b) Route 2

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

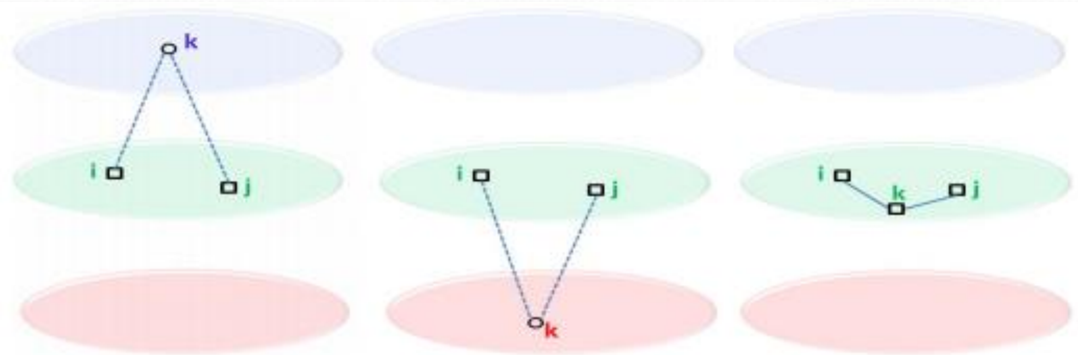
$$p_{ij}^{(\mathcal{UP})^-} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^-} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^-} r_{kj}^{(\mathcal{P})}$$

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

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

Social Recommendation

❖ Update within-domain links



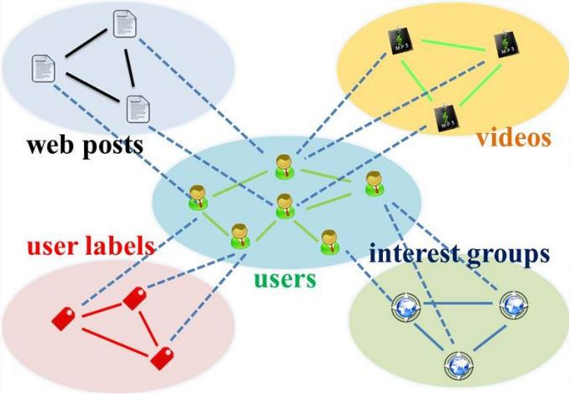
(a) Route 1 (b) Route 2 (c) Route 3

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

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

Social Recommendation

- On high-order star-structured graph



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

Data Set

- Tencent Weibo (January 2011)

Domain	Size	Cross-domain links	
		Accept	Refuse
User	53.4K	—	—
Web post	142K	1.47M (0.02%)	3.40M (0.04%)
User label	111	330K (5.57%)	—

Good to Transfer?

- Comparative Algorithms (RWR)
  - $W^{(P)}$: Use web post similarity?
  - $W^{(U)}$: Use social relation?
  - $R^{(U)}$: Update tie strength?
  - $W^{(T)}$: Use user label similarity?

Algorithm	$\mathbf{R}^{(\mathcal{U})}$	$\mathbf{W}^{(\mathcal{U})}$	$\mathbf{W}^{(\mathcal{P})}$	$\mathbf{W}^{(\mathcal{T})}$
HRW	√	√	√	√
BRW-R_U-P (TrustWalker)	√	√	√	×
BRW-R_U	√	√	×	×
BRW-W_U-P	×	√	√	×
BRW-W_U (ItemRank)	×	√	×	×
BRW-P	×	×	√	×

Good to Transfer?

❖ Compare with RWR models

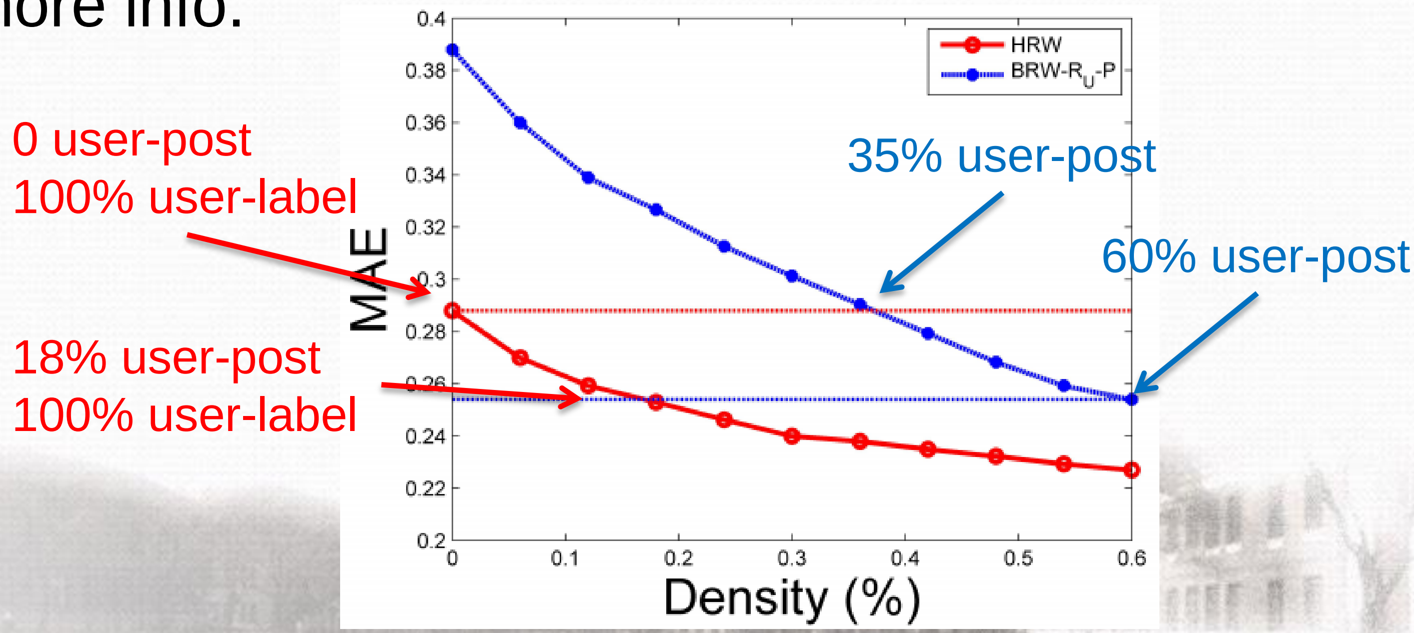
Algorithm	MAE	Precision	Recall	F1	Kendall's $\hat{\tau}$
HRW	**0.227±1.5e-3**	**0.711±1.3e-3**	0.921±1.4e-3	**0.802±1.1e-3**	**0.792±2.5e-3**
BRW-R_U-P (TrustWalker)	0.276±1.1e-3	0.657±7.6e-4	**0.935±9.8e-4**	0.772±7.6e-4	0.774±1.6e-3
BRW-R_U	0.282±5.3e-3	0.655±4.0e-3	0.921±1.2e-2	0.765±7.7e-3	0.725±2.8e-3
BRW-W_U-P	0.292±1.1e-3	0.666±7.0e-4	0.900±5.2e-4	0.765±6.6e-4	0.725±8.5e-4
BRW-W_U (ItemRank)	0.318±1.4e-3	0.671±1.5e-3	0.713±2.4e-3	0.691±1.2e-3	0.661±2.2e-3
BRW-P	0.438±2.6e-4	0.571±3.4e-4	0.499±4.2e-4	0.532±3.2e-4	0.606±2.3e-4

❖ Compare with Baselines

Algorithm	MAE	Precision	Recall	F1	Kendall's $\hat{\tau}$
HRW	**0.227±1.5e-3**	**0.711±1.3e-3**	0.921±1.4e-3	**0.802±1.1e-3**	**0.792±2.5e-3**
BRW-R_U-P (TrustWalker) [10]	0.276±1.1e-3	0.657±7.6e-4	0.935±9.8e-4	0.772±7.6e-4	0.774±1.6e-3
BRW-W_U (ItemRank) [8]	0.318±1.4e-3	0.671±1.5e-3	0.713±2.4e-3	0.691±1.2e-3	0.661±2.2e-3
MCF [5]	0.352±2.3e-4	0.592±1.8e-3	**0.951±6.0e-4**	0.730±1.3e-3	0.582±4.3e-4
CF [22]	0.506±3.4e-4	0.552±1.5e-3	0.589±7.2e-4	0.570±1.0e-3	0.540±5.2e-4

Good to Transfer? More Insights!

- If we do transfer (from user-label domain), we need only ~30% to reach the same performance.
- Advice: build more features to let new users provide more info.



Meng Jiang, Peng Cui, Fei Wang, Qiang Yang, Wenwu Zhu and Shiqiang Yang. Social Recommendation across Multiple Relational Domains. *The 21st ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

Good user-item links

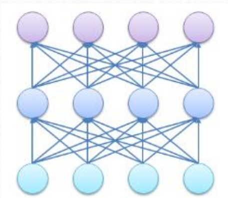
- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

Human Behavior Pattern

- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - **Multi-faceted**

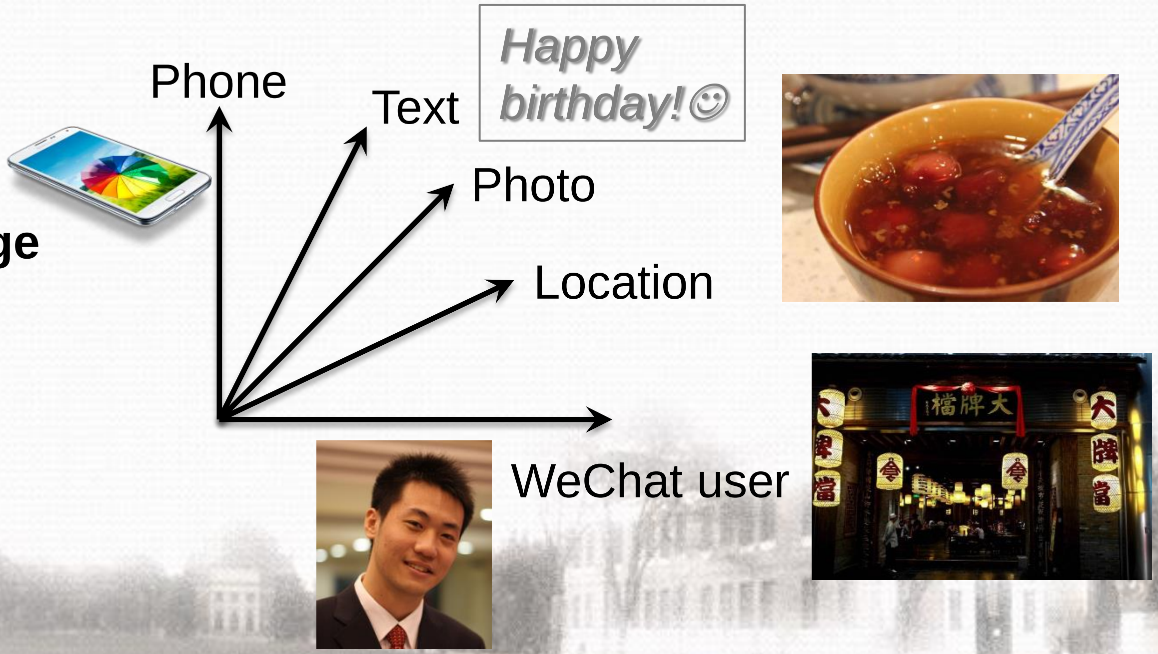
Write a paper



Human Behavior Pattern

- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - **Multi-faceted**

Post a WeChat message

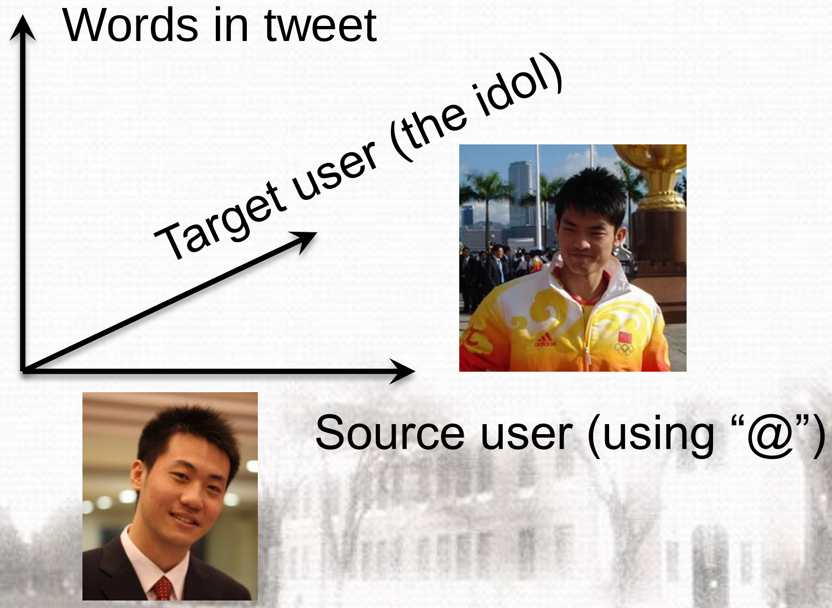


Human Behavior Pattern

- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - **Multi-faceted**

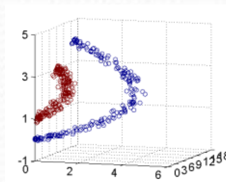
Congratulations to the Champion!

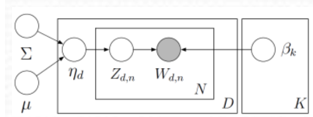
Mention your "idol"

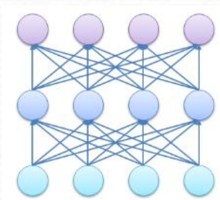


Human Behavior Pattern

- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - Multi-faceted
  - **Evolutionary**







Write a paper

time

Stanford University

Baidu 百度

Human Behavior Pattern

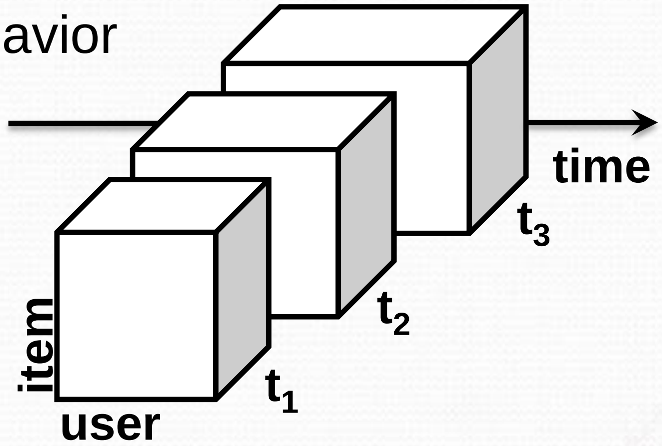
- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - Multi-faceted
  - **Evolutionary**

Post a WeChat message

Human Behavior Pattern

- Problem
  - Human behavior pattern discovery and prediction: significant for industry, national security and life experience.
  - Two basic characteristics of human behavior
  - **Multi-faceted**
  - **Evolutionary**

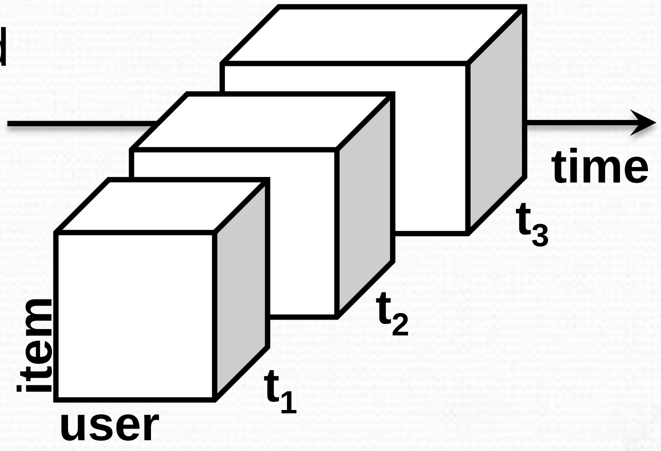
  - How to model human behavior?
  - **Tensor sequence.**
  - How to do pattern discovery and prediction?
  - **Tensor decomposition and completion.**



Human Behavior Pattern

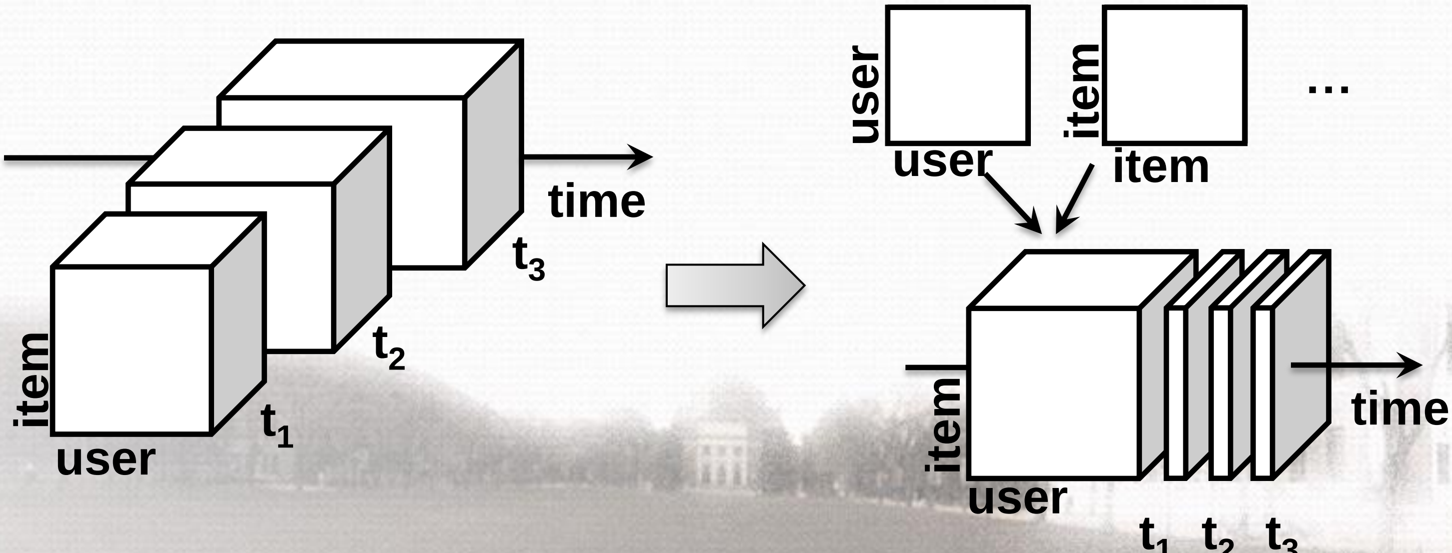
- Challenges in high-order tensor decomposition
  - **High sparsity**
  - Write a paper: #author * #affiliation * #keyword
  - "@idol": #(src) user * #(tgt) user * #word

  - **High complexity**
  - Long sequence of large tensors
  - Slow: decomposition at each time

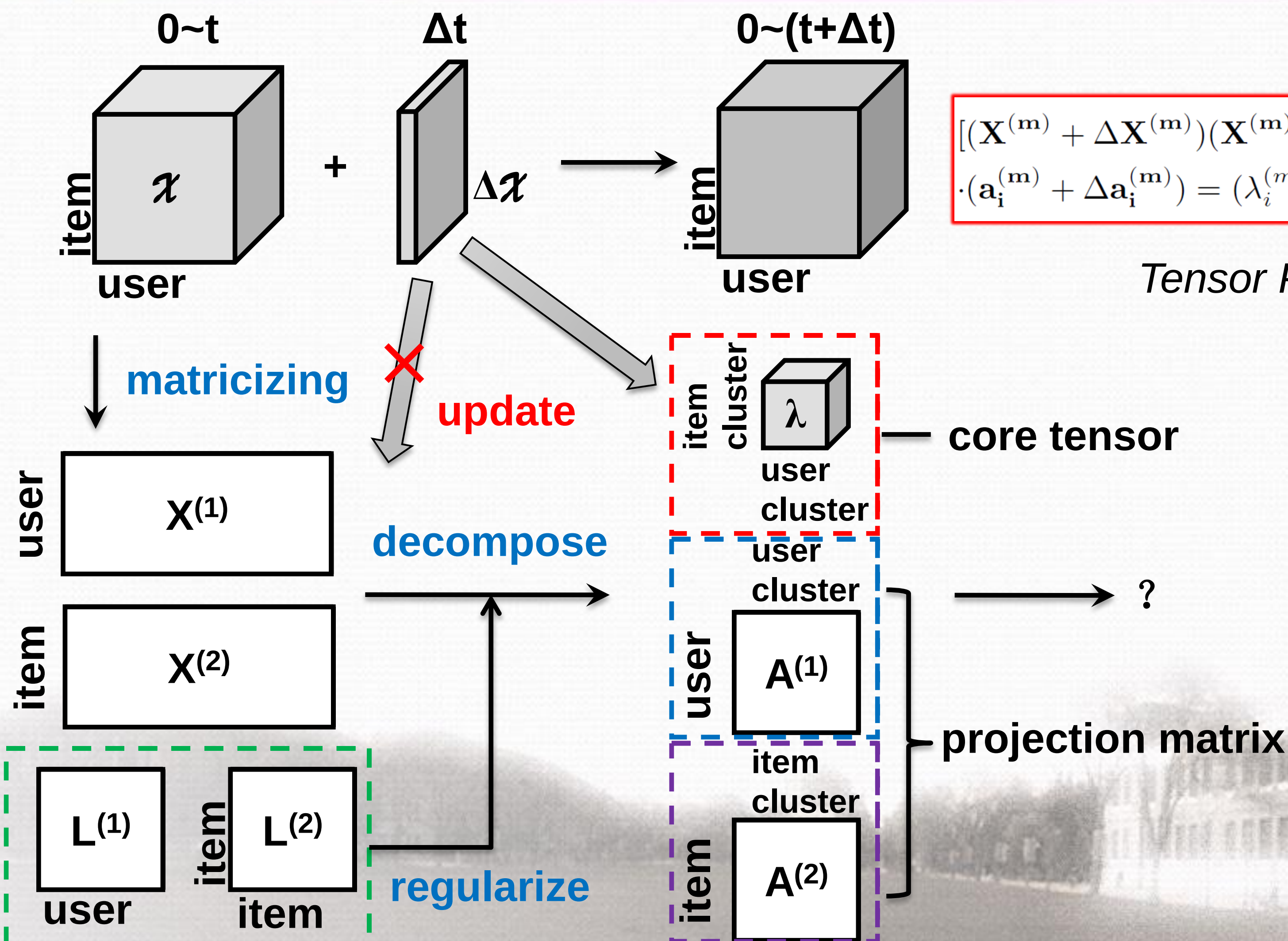


Human Behavior Pattern

- High sparsity: auxiliary knowledge as regularizers
  - Author - affiliation - keyword + co-authorship (author-author)
  - Src user - tgt user - word + social (user-user)
- High complexity: update decomposition results (projection matrix) with new coming piece of data



FEMA: Flexible Evolutionary Multi-faceted Analysis



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

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

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

    and compute

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

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

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

    and compute

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

  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

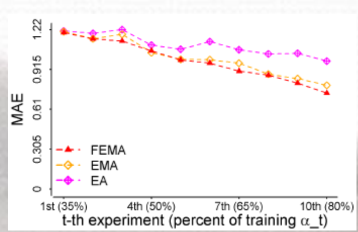
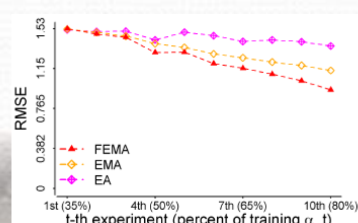
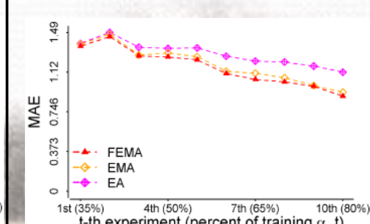
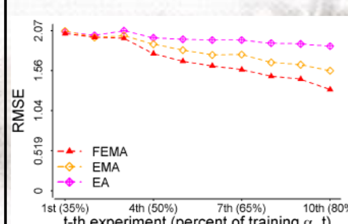
$$|\Delta\lambda_i^{(m)}| \leq 2(\lambda^{\max}_{\mathbf{X}^{(\mathbf{m})\top}\mathbf{X}^{(\mathbf{m})}})^{\frac{1}{2}}\|\Delta\mathbf{X}^{(\mathbf{m})}\|_2$$

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

Leveraging Multi-faceted Information

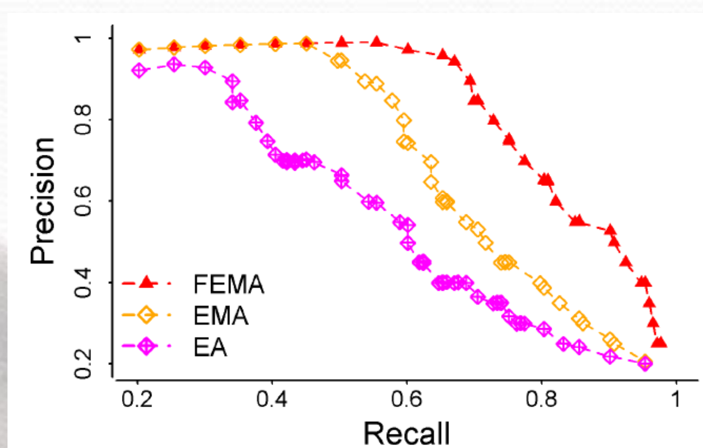
"Who"-"What keyword"?
Use "Where" (affiliation)?

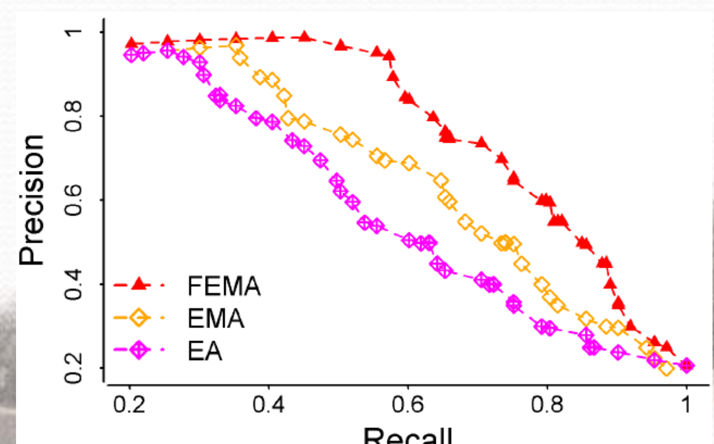
"Who"-"@Whom"?
Use "What" (tweet word)?

	MAS (Data mining subset)		Tencent Weibo (January, 2011)	
	MAE	**RMSE**	**MAE**	**RMSE**
FEMA $\mathcal{X}$ **L**	**0.735**	**0.944**	**0.894**	**1.312**
EMA $\mathcal{X}$	0.794	1.130	0.932	1.556
EA **X**	0.979	1.364	1.120	1.873
Cross Validation				

Leveraging Multi-faceted Information

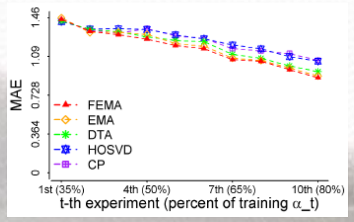
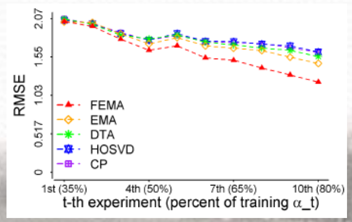
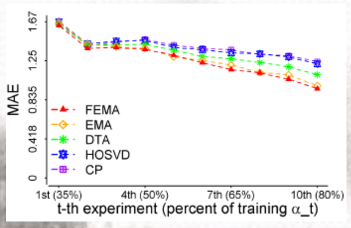
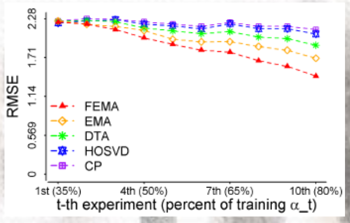
	MAS (Data mining subset)		Tencent Weibo (January, 2011)	
	MAE	**RMSE**	**MAE**	**RMSE**
FEMA $\mathcal{X}$ L	**0.735**	**0.944**	**0.894**	**1.312**
EMA $\mathcal{X}$	0.794	1.130	0.932	1.556
EA X	0.979	1.364	1.120	1.873





Leveraging Flexible Regularizations

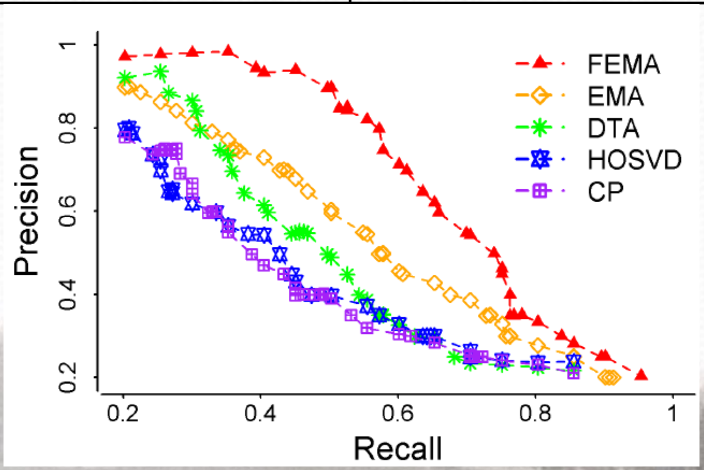
“Who”-“Where”-“What keyword”? “Who”-“@Whom”-“What”?

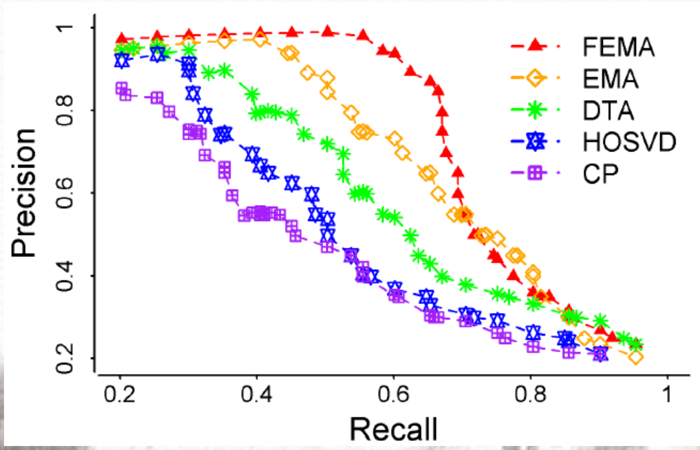
	MAS (Data mining subset)		Tencent Weibo (January, 2011)	
	MAE	**RMSE**	**MAE**	**RMSE**
FEMA $\mathcal{X}$ L	**0.893**	**1.215**	**0.954**	**1.437**
EMA $\mathcal{X}$	0.909	1.466	0.986	1.698
DTA	0.950	1.556	1.105	1.889
HOSVD	1.047	1.618	1.220	2.054
CP	1.055	1.612	1.243	2.117
Cross Validation				

Leveraging Flexible Regularizations

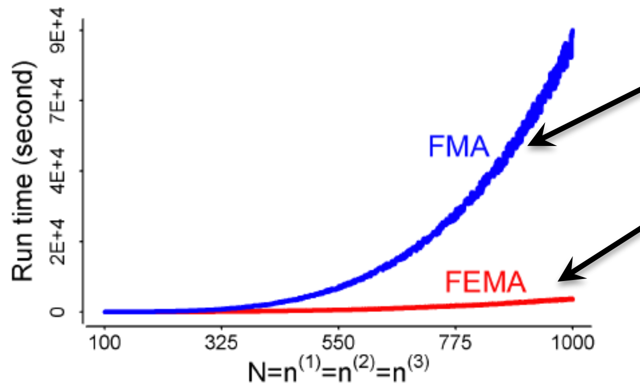
"Who"-"Where"-"What keyword"? **"Who"-"@Whom"-"What"?**

	MAS (Data mining subset)		Tencent Weibo (January, 2011)	
	MAE	**RMSE**	**MAE**	**RMSE**
FEMA $\mathcal{X}$ L	**0.893**	**1.215**	**0.954**	**1.437**
EMA $\mathcal{X}$	0.909	1.466	0.986	1.698





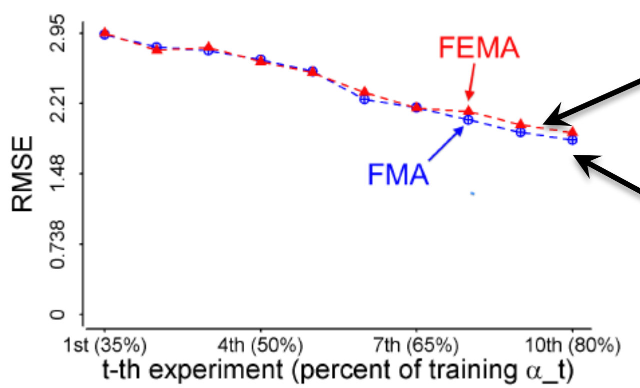
Efficiency, Loss and Parameters



Time vs Num. objects N

Re-decompose
updated matrices

Evolutionary analysis:
update λ and a with ΔX

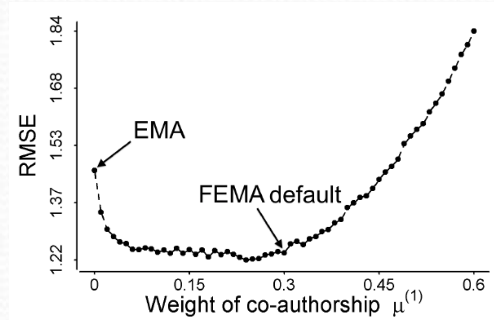


The loss is small.

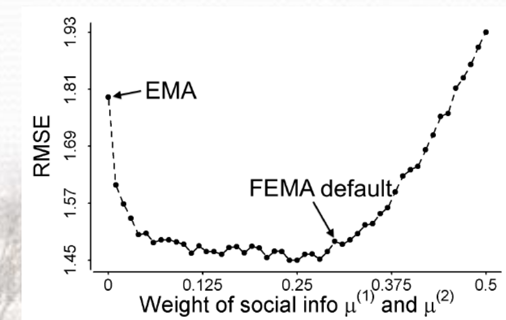
Evolutionary analysis:
update λ and a with ΔX

Re-decompose
updated matrices

Insensitive to regularization weight

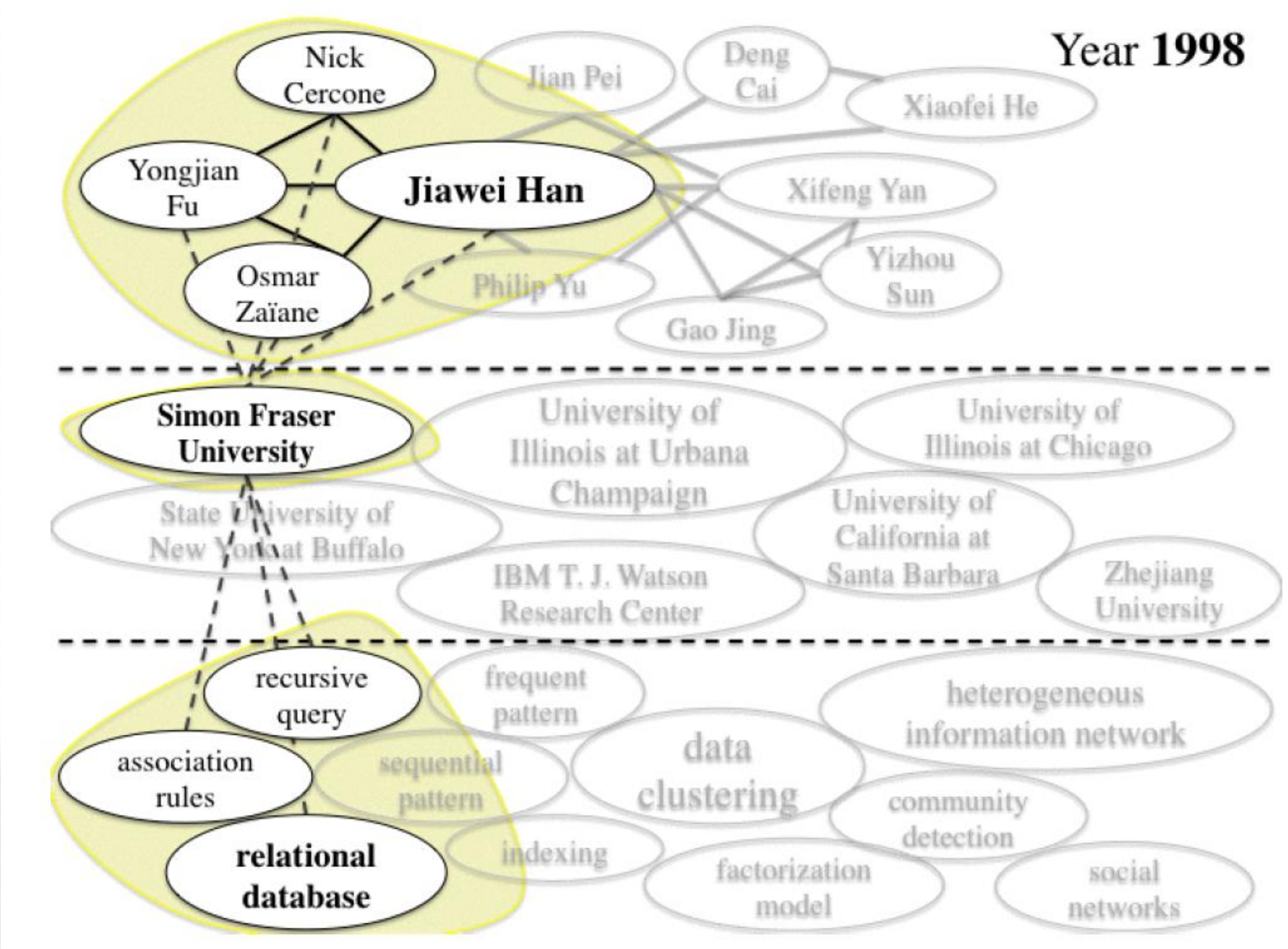


(a) Academic data MAS

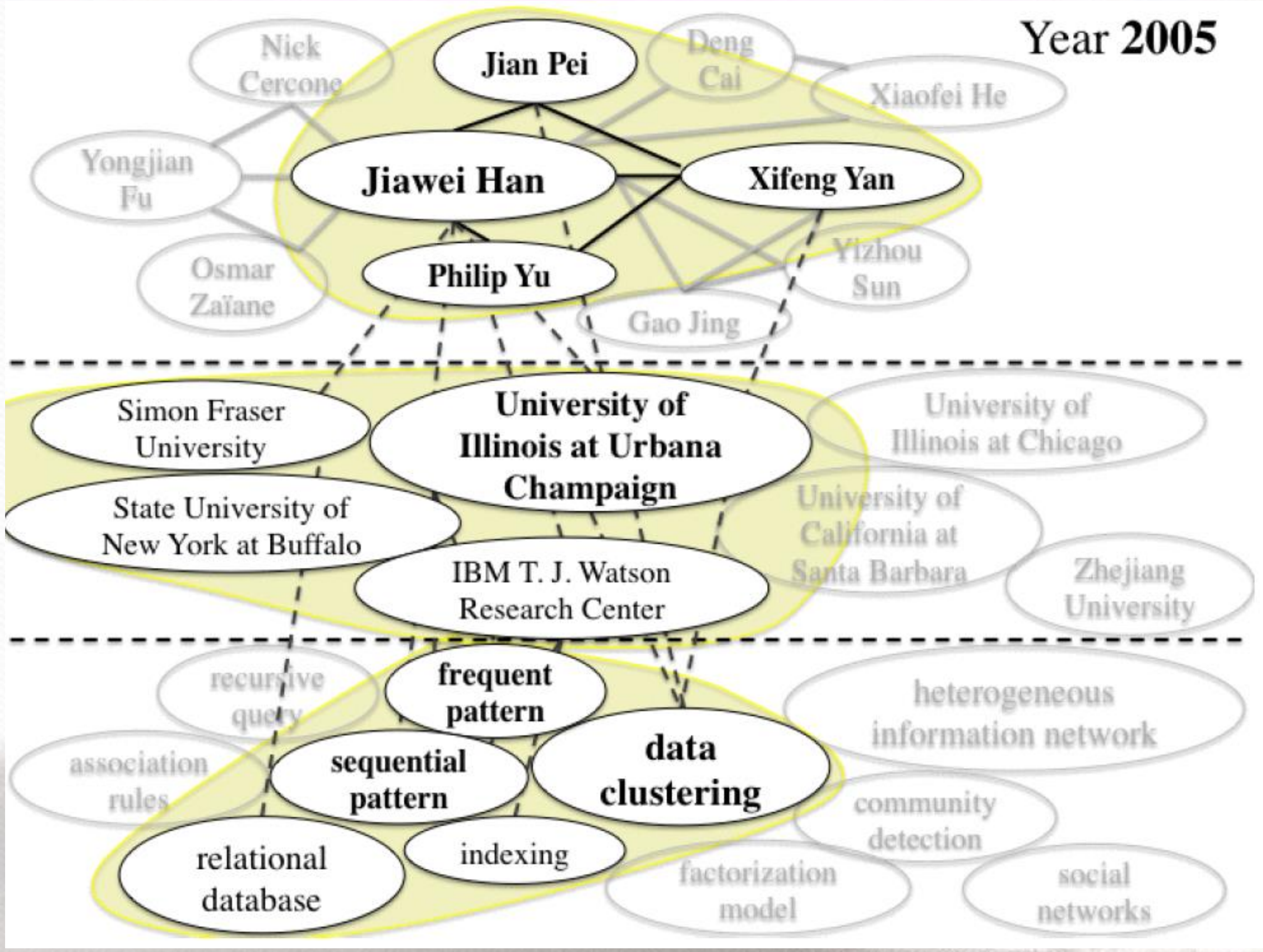


(b) Tweet data WEIBO

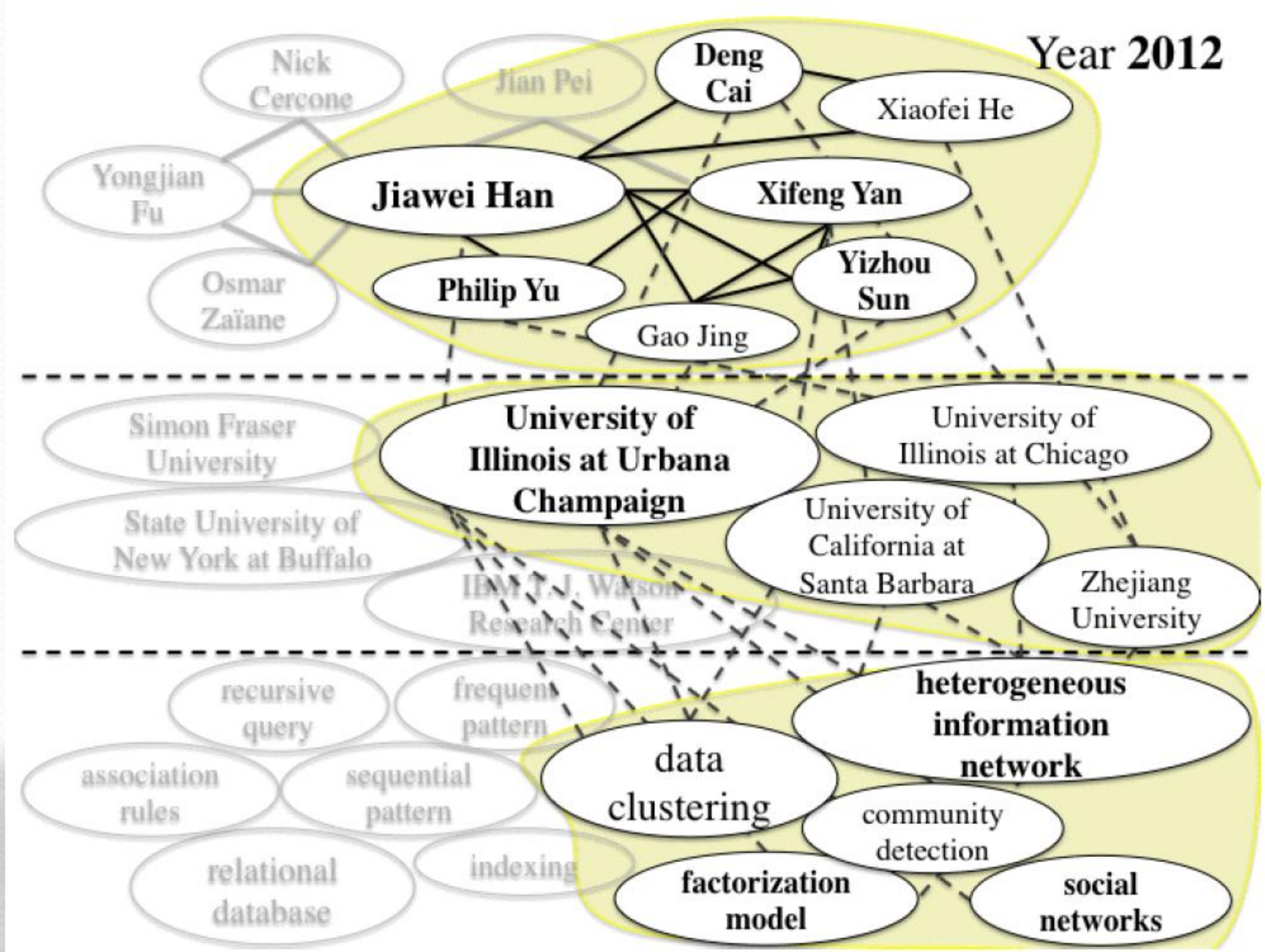
First Cluster in MAS DM set



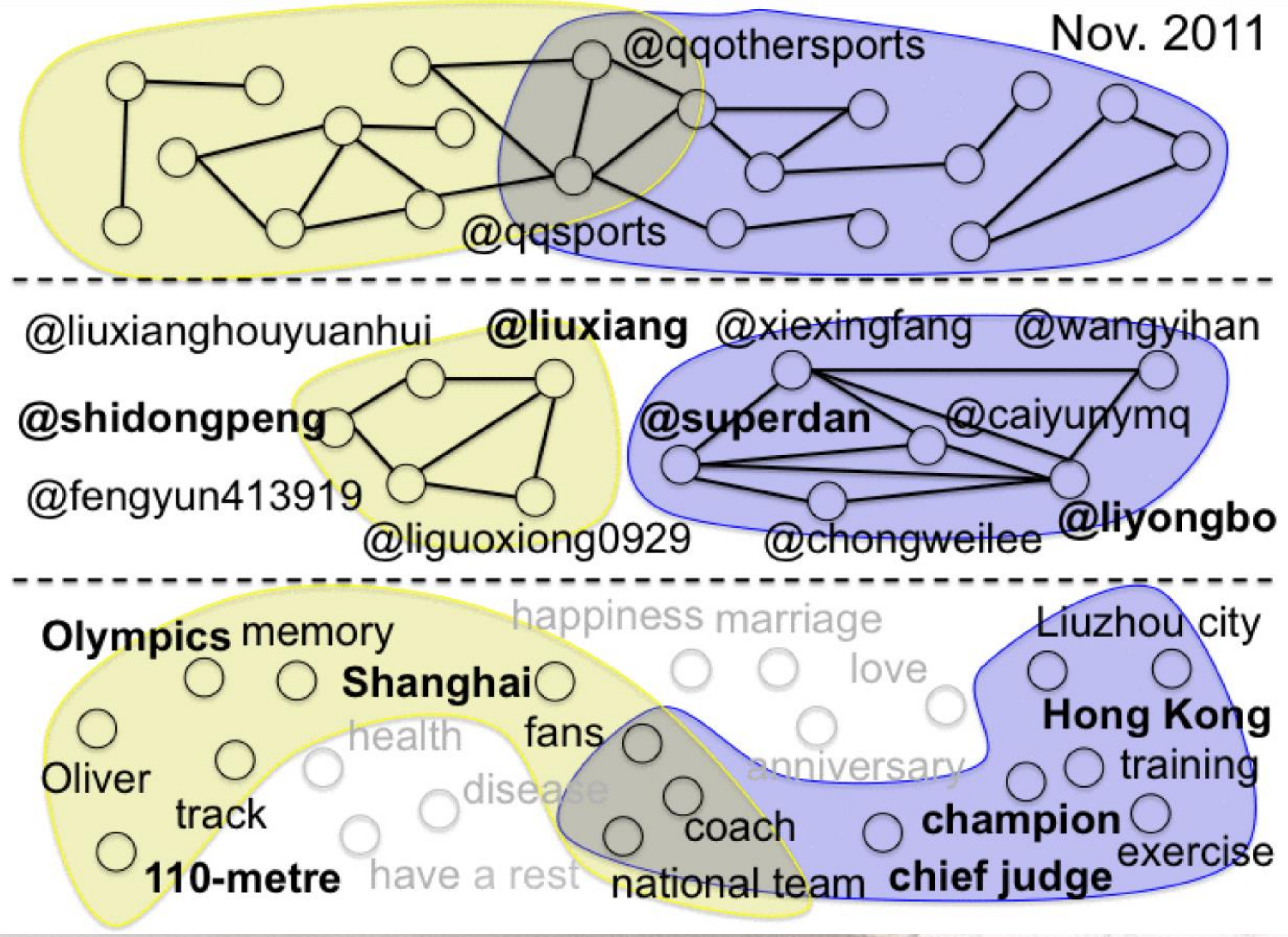
First Cluster in MAS DM set



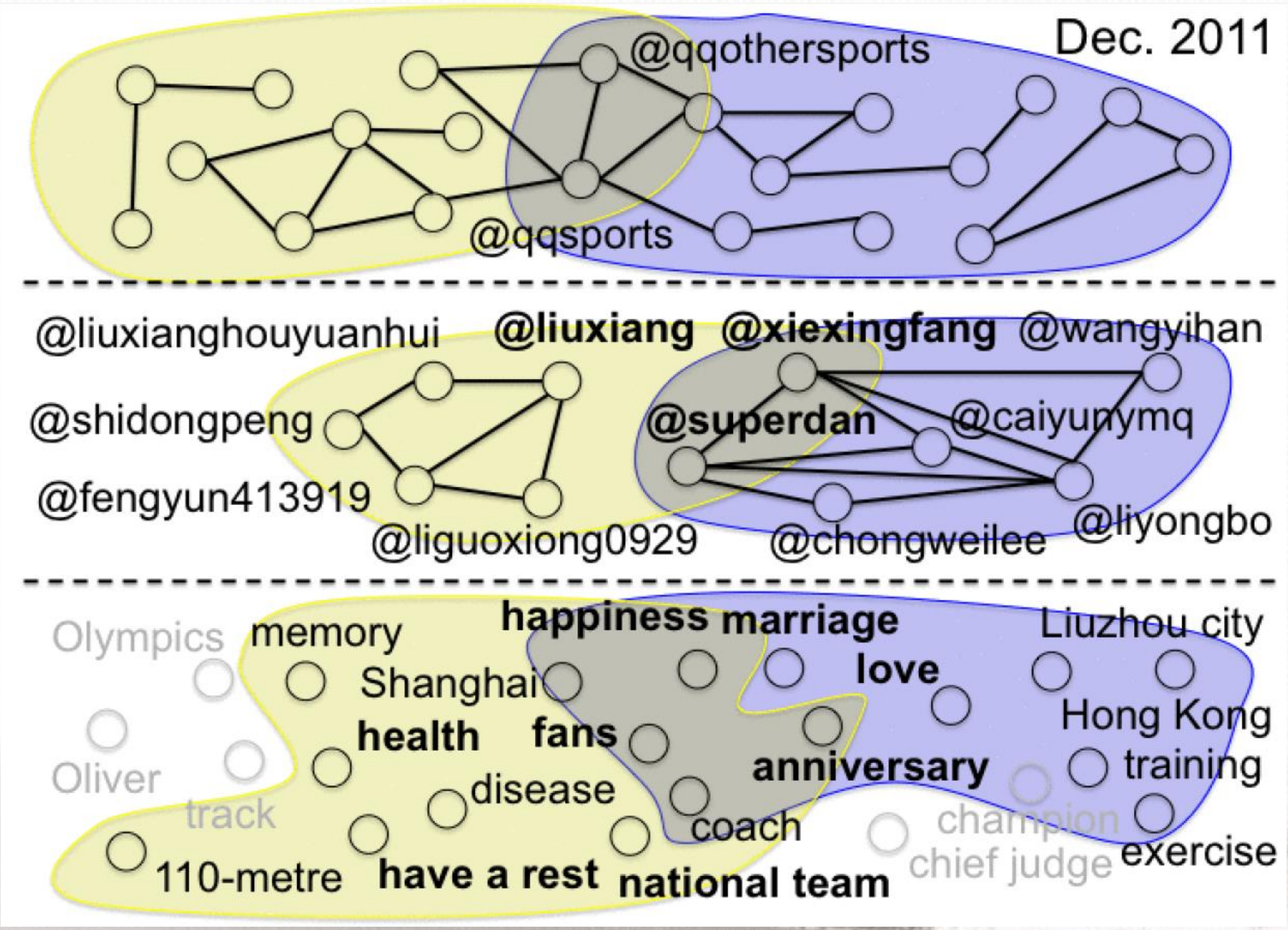
First Cluster in MAS DM set



First Two Clusters in Tencent Weibo set



First Two Clusters in Tencent Weibo set



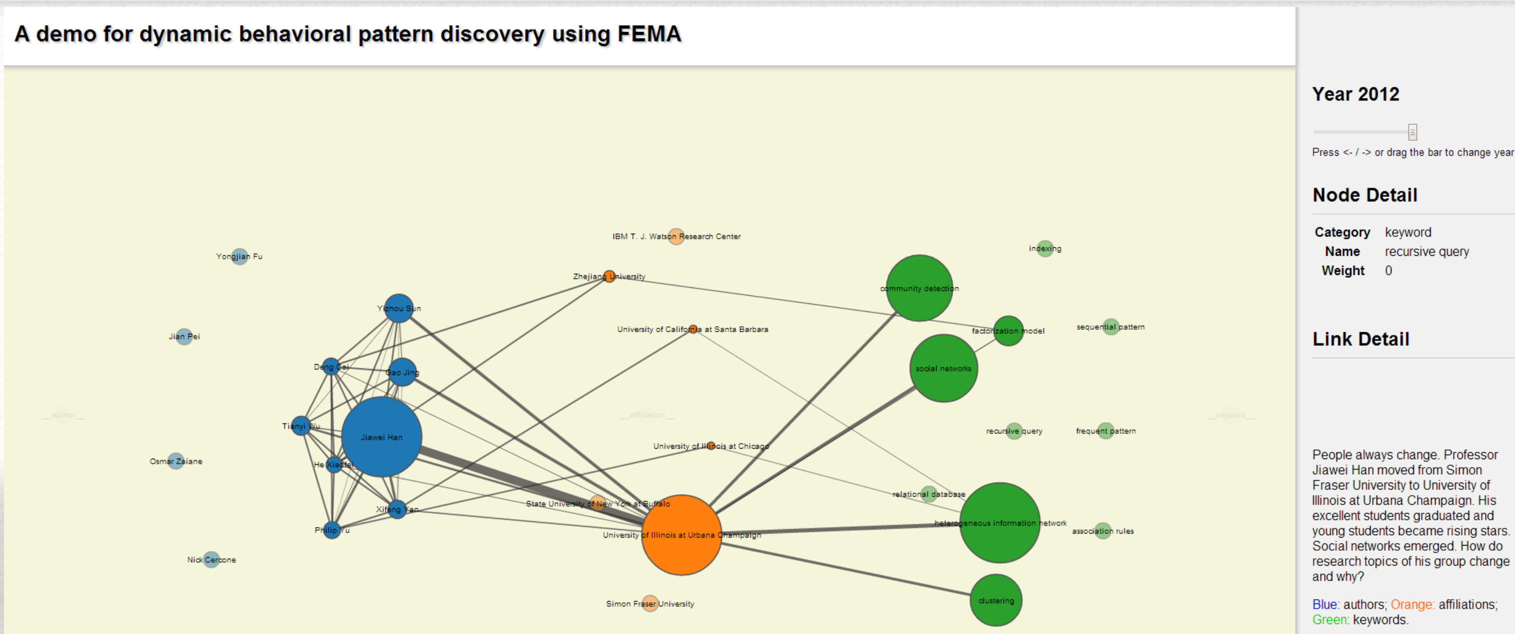
Demo: www.meng-jiang.com

FEMA: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery. (KDD 2014)

Meng Jiang, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu and Shiqiang Yang.

How do research topics of Professor Jiawei Han's group change? What happened in the data mining communities and why?

How do Weibo users mention (using "@") their idols (@superdan, etc.) in their tweets? What are the words in their tweets?



Meng Jiang, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu and Shiqiang Yang. FEMA: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery. *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

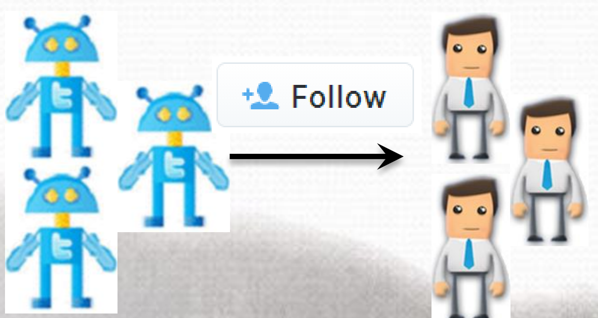
Good user-item links

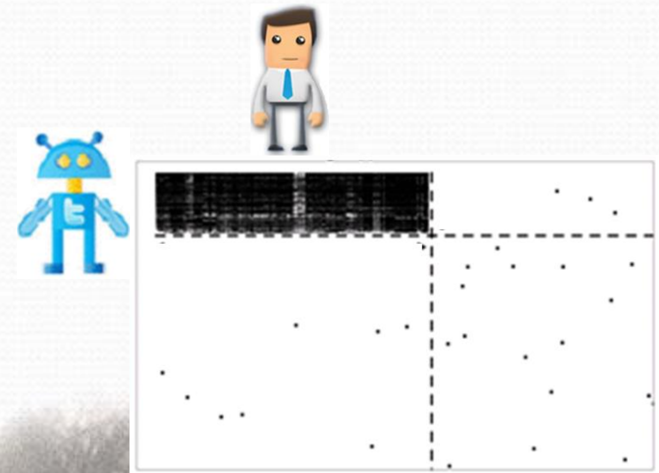
- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

Strange Behavior/Connectivity Patterns

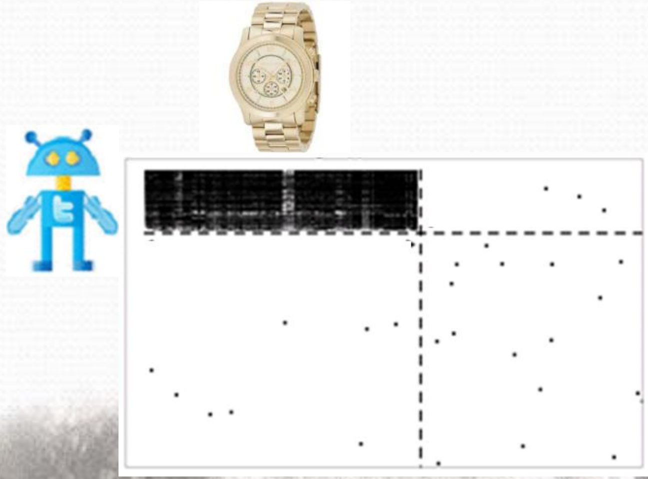
- Problem
  - Strange behavior on large graph: dense bipartite cores
  - **Twitter/Weibo “who-follows-whom”**





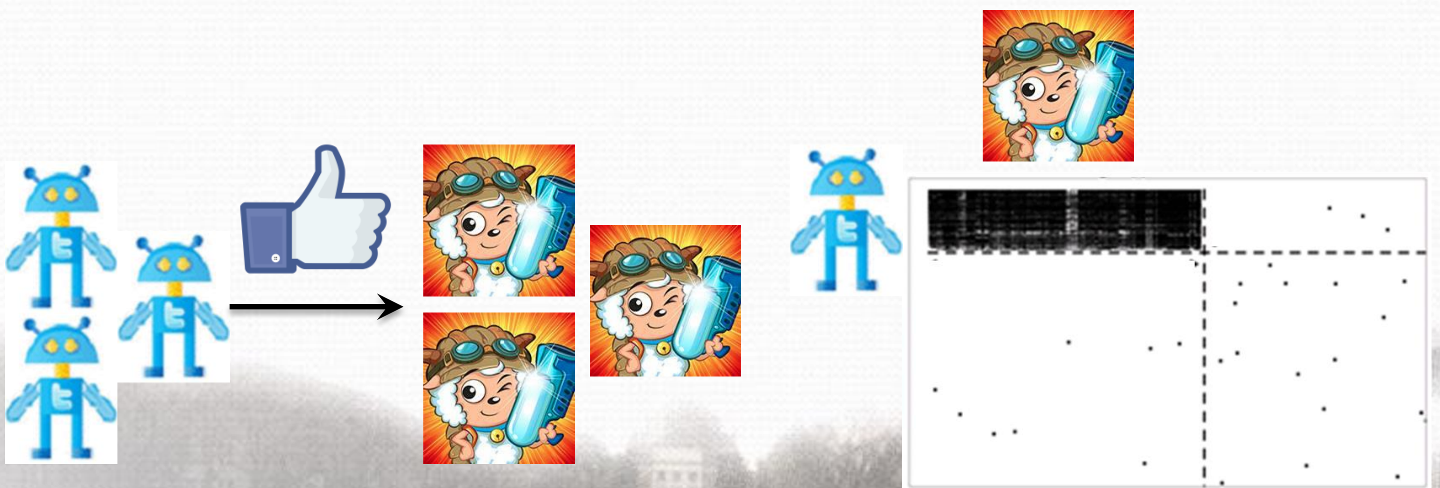
Strange Behavior/Connectivity Patterns

- Problem
  - Strange behavior on large graph: dense bipartite cores
  - Twitter/Weibo “who-follows-whom”
  - **eBay “who-reviews-what”**

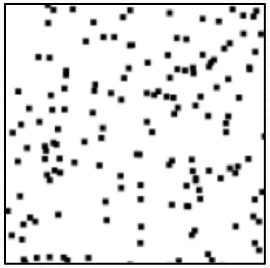
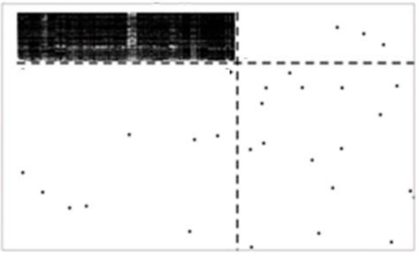
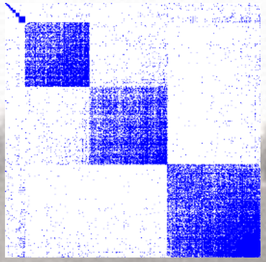


Strange Behavior/Connectivity Patterns

- Problem
  - Strange behavior on large graph: dense bipartite cores
  - Twitter/Weibo “who-follows-whom”
  - eBay “who-reviews-what”
  - **Facebook “who-likes-page”**

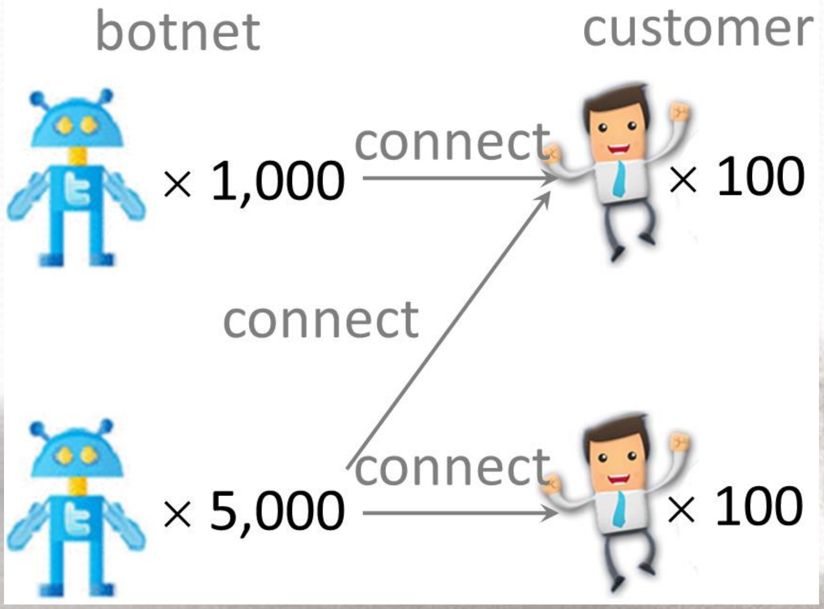


Strange Behavior/Connectivity Patterns

- Problem
  - Strange behavior on large graph: dense bipartite cores
  - Twitter/Weibo “who-follows-whom”
  - eBay “who-reviews-what”
  - Facebook “who-likes-page”
  - **“Lockstep” behavior**
  - Given  reordering 
  - Usually we have 

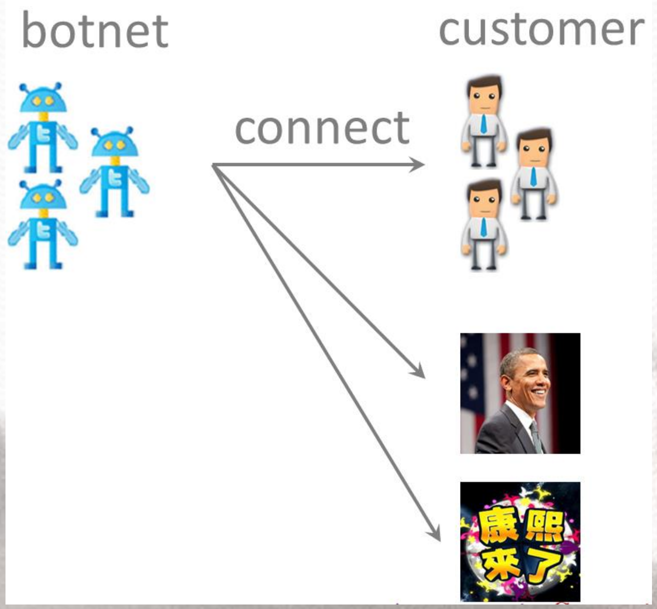
Strange Behavior/Connectivity Patterns

- Challenges
  - **“Lockstep” behavior**
  - **Overlapping**



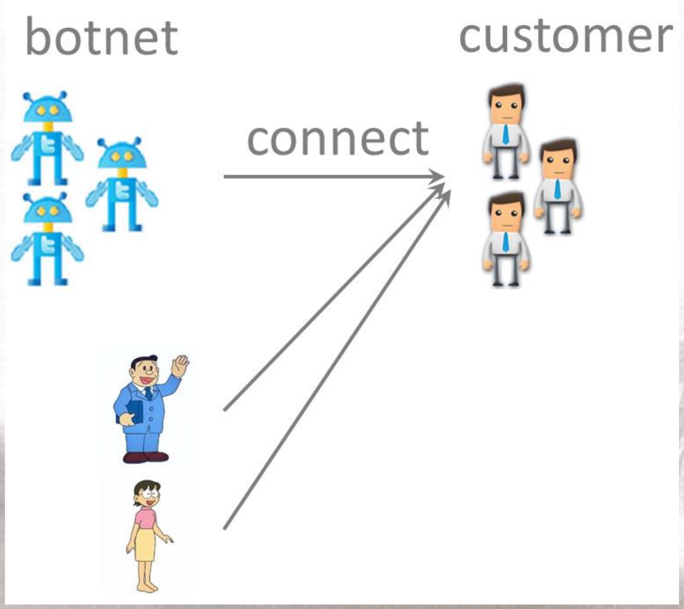
Strange Behavior/Connectivity Patterns

- Challenges
  - **"Lockstep" behavior**
  - Overlapping
  - **Camouflage**

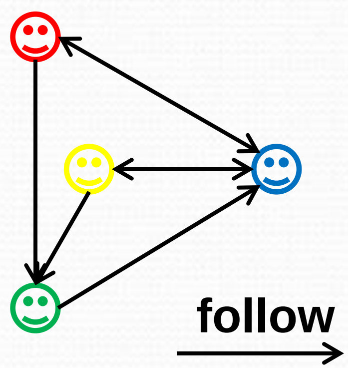


Strange Behavior/Connectivity Patterns

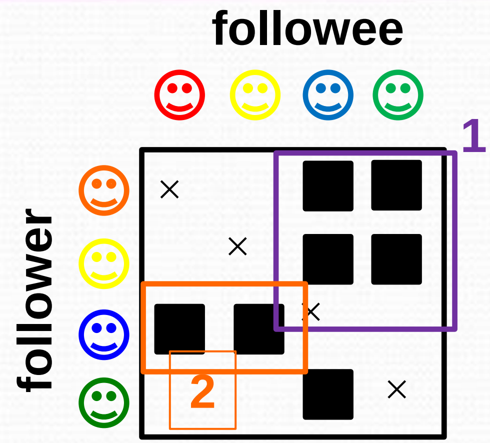
- Challenges
  - **"Lockstep" behavior**
  - Overlapping
  - Camouflage
  - Fame



SVD Reminder

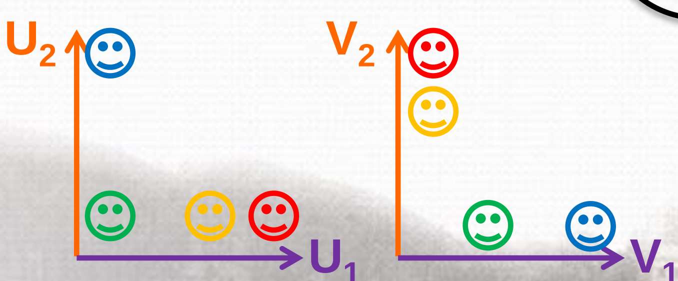


Graph Structure



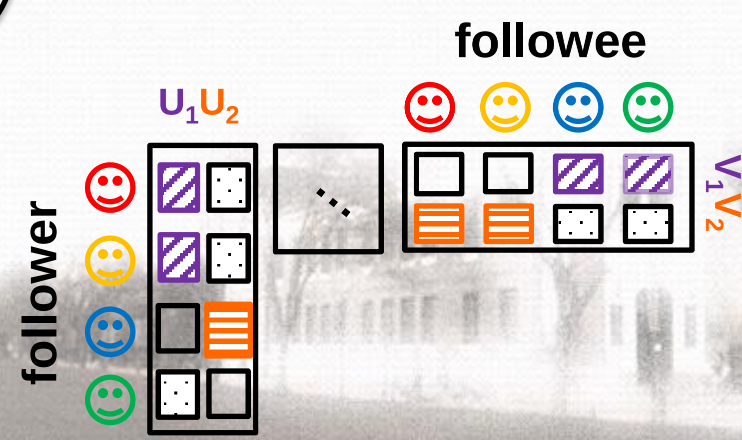
Adjacency Matrix

Pairs of singular vectors:



"Spectral Subspace Plot"

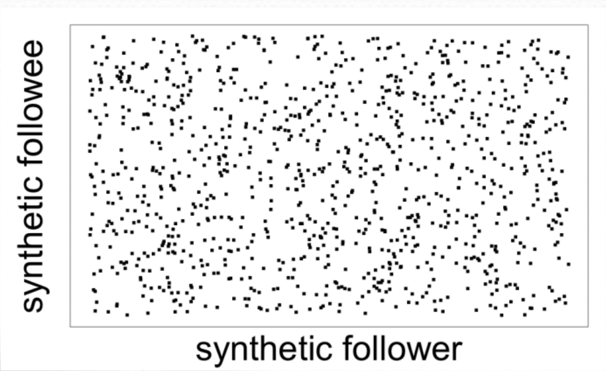
SVD: $A=USV^T$



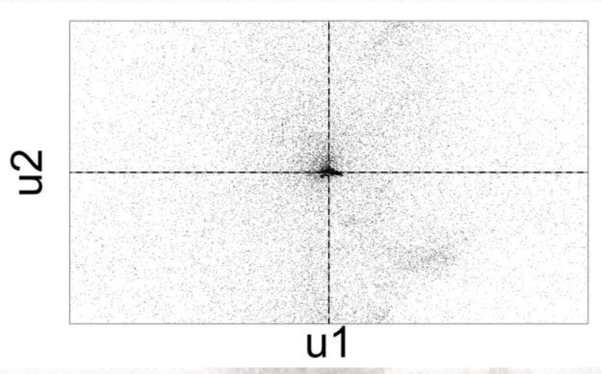
Lockstep and Spectral Subspace Plot

- Case #0: No lockstep behavior in random power law graph of 1M nodes, 3M edges
- Random ⟷ "Scatter"

Adjacency Matrix



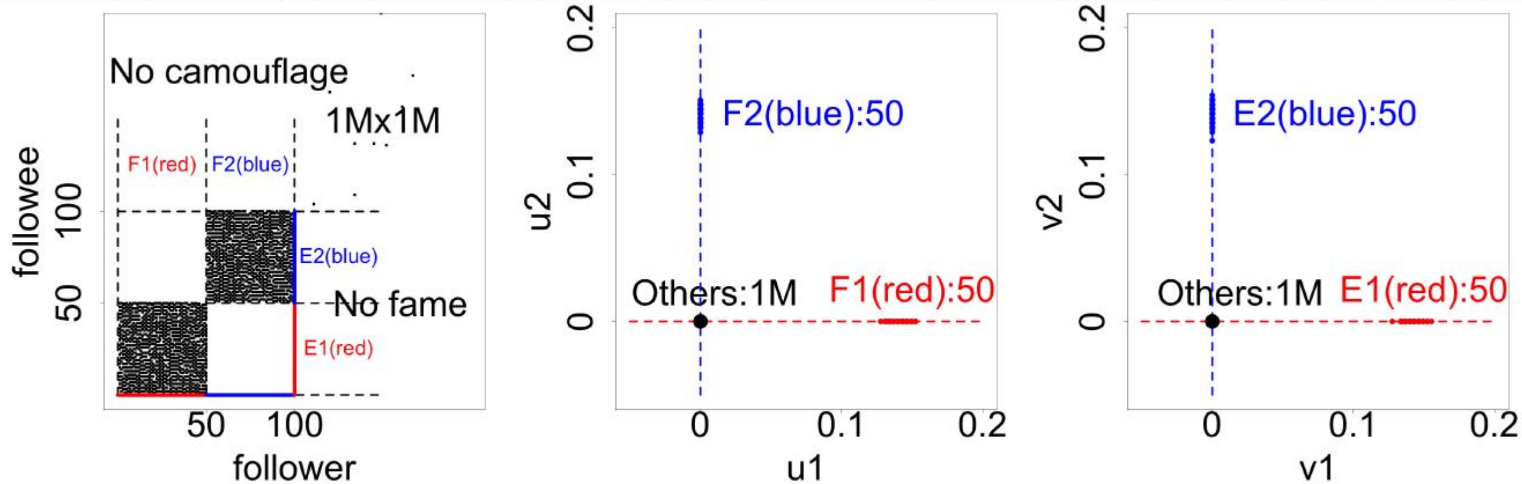
Spectral Subspace Plot



Lockstep and Spectral Subspace Plot

- Case #1: non-overlapping lockstep
- "Blocks" ⟷ "Rays"

Adjacency Matrix Spectral Subspace Plot

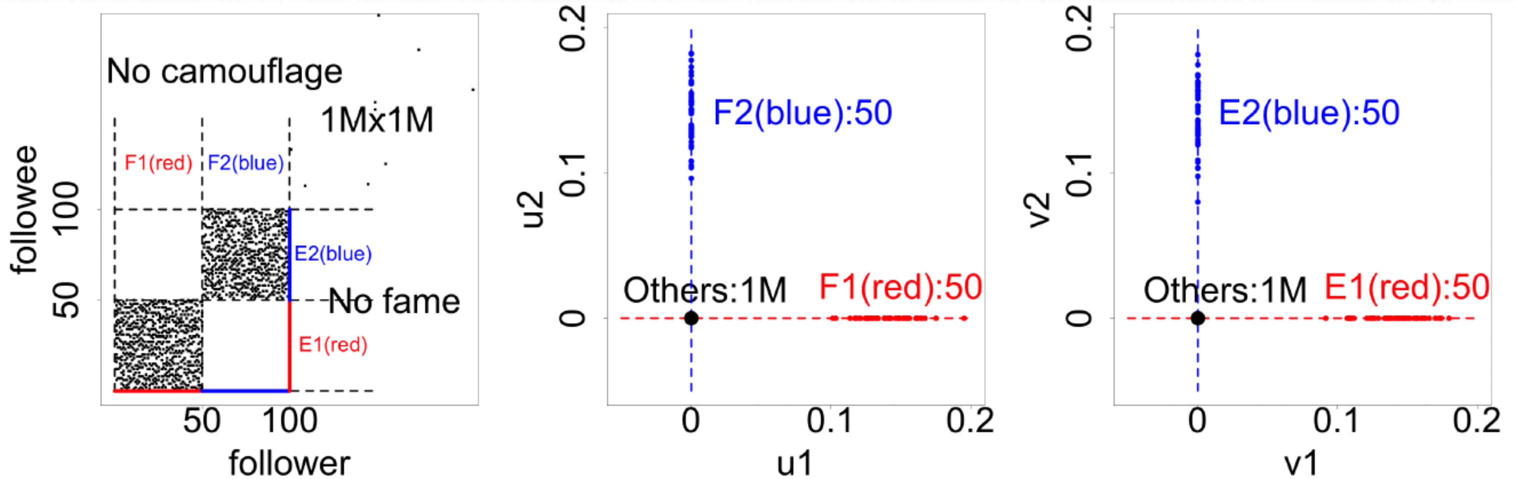


Rule 1 (short "rays"): two blocks, high density (90%), no "camouflage", no "fame"

Lockstep and Spectral Subspace Plot

- Case #2: non-overlapping lockstep
- "Blocks; low density" ⟷ Elongation

Adjacency Matrix Spectral Subspace Plot

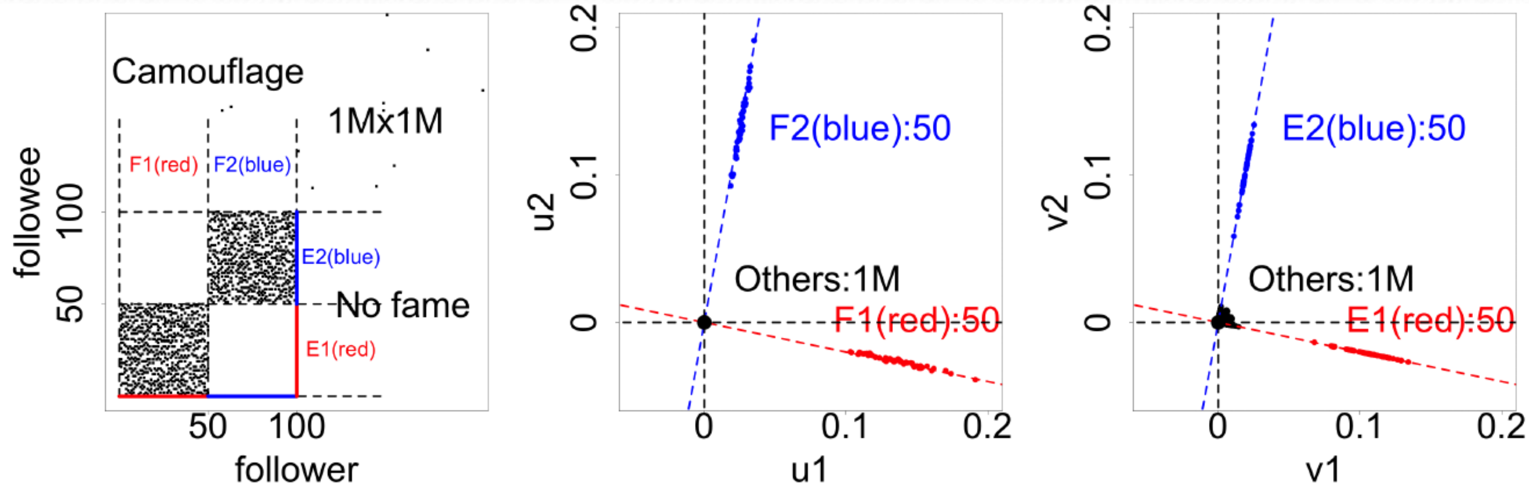


Rule 2 (long "rays"): two blocks, low density (50%), no "camouflage", no "fame"

Lockstep and Spectral Subspace Plot

- Case #3: non-overlapping lockstep
- "**Camouflage**" (or "Fame") ⟷ Tilting "Rays"

Adjacency Matrix   Spectral Subspace Plot

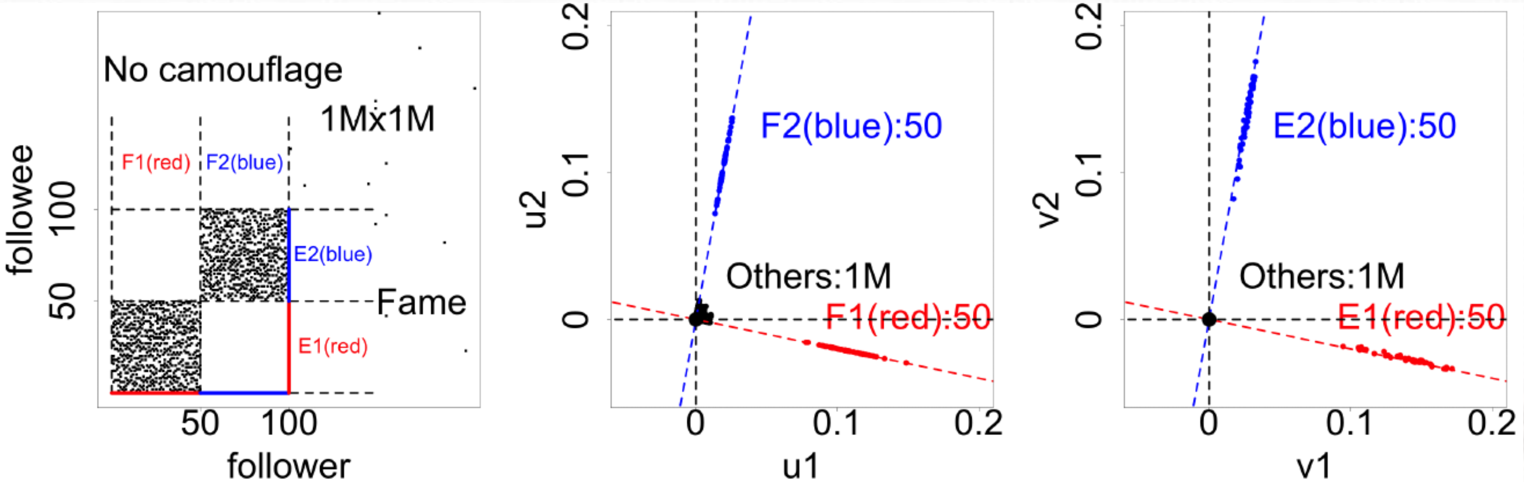


Rule 3 (tilting "rays"): two blocks, with "camouflage", no "fame"

Lockstep and Spectral Subspace Plot

- Case #3: non-overlapping lockstep
- "Camouflage" (or "**Fame**") ⟷ Tilting "Rays"

Adjacency Matrix Spectral Subspace Plot



Rule 3 (tilting "rays"): two blocks, no "camouflage", with "fame"

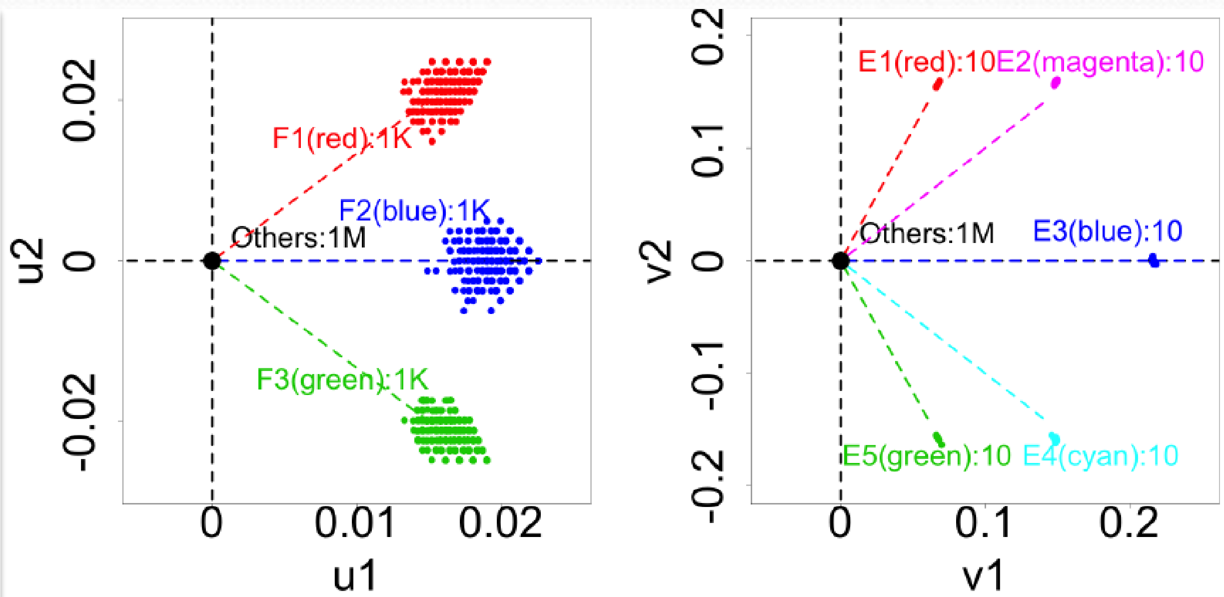
Lockstep and Spectral Subspace Plot

- Case #4: **?** lockstep
- **"?"** ⟷ "Pearls"

Adjacency Matrix

?

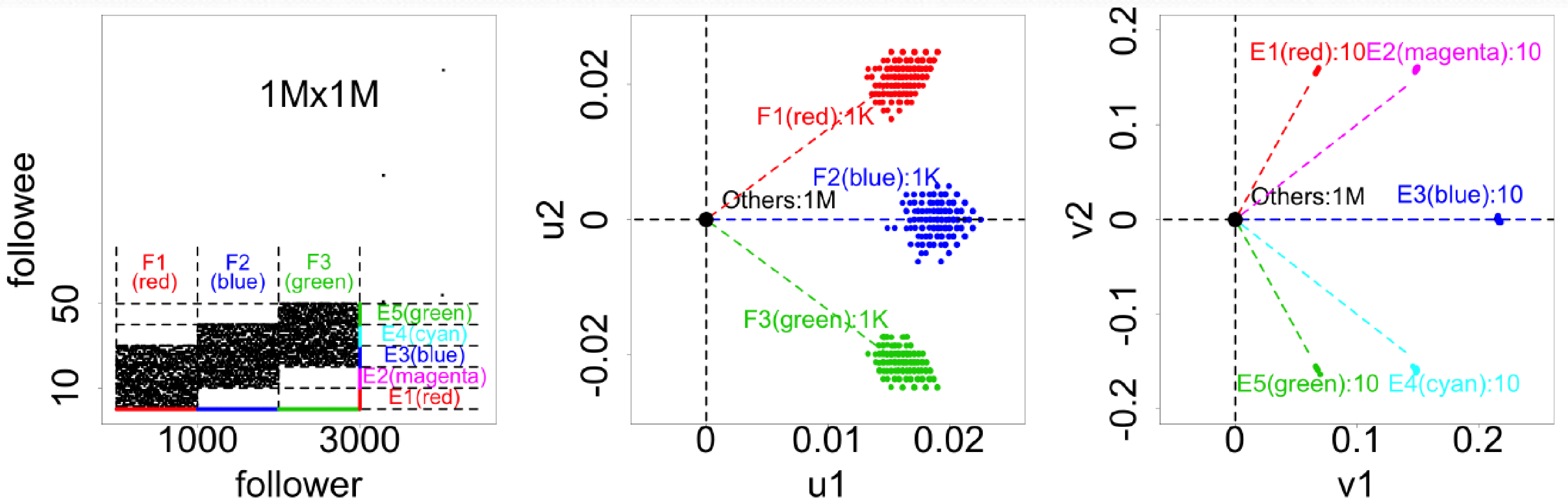
Spectral Subspace Plot



Lockstep and Spectral Subspace Plot

- Case #4: **overlapping** lockstep
- **"Staircase"** ⟷ "Pearls"

Adjacency Matrix　　　Spectral Subspace Plot

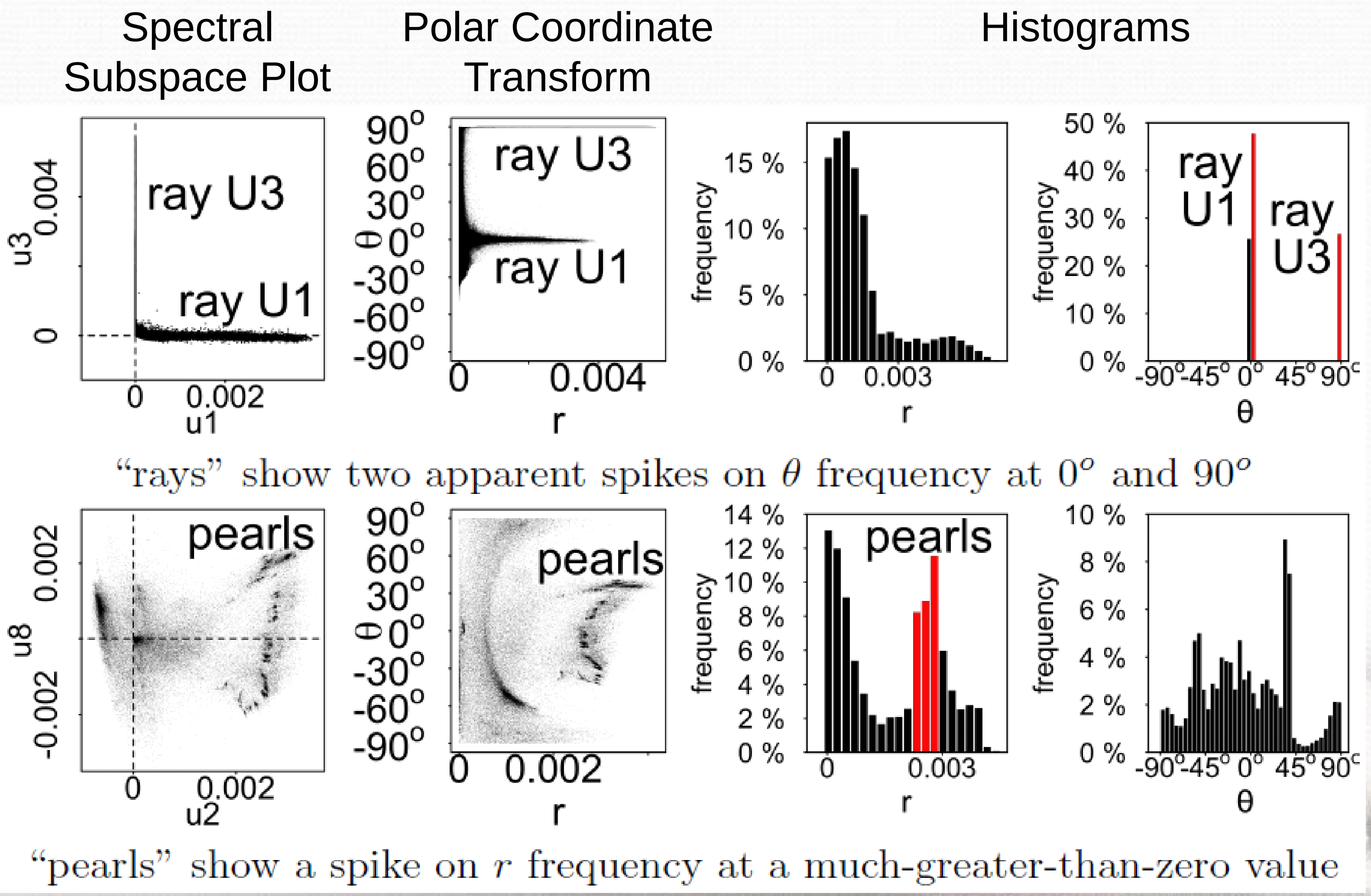


Rule 4 ("pearls"): a "staircase" of three partially overlapping blocks.

Algorithm

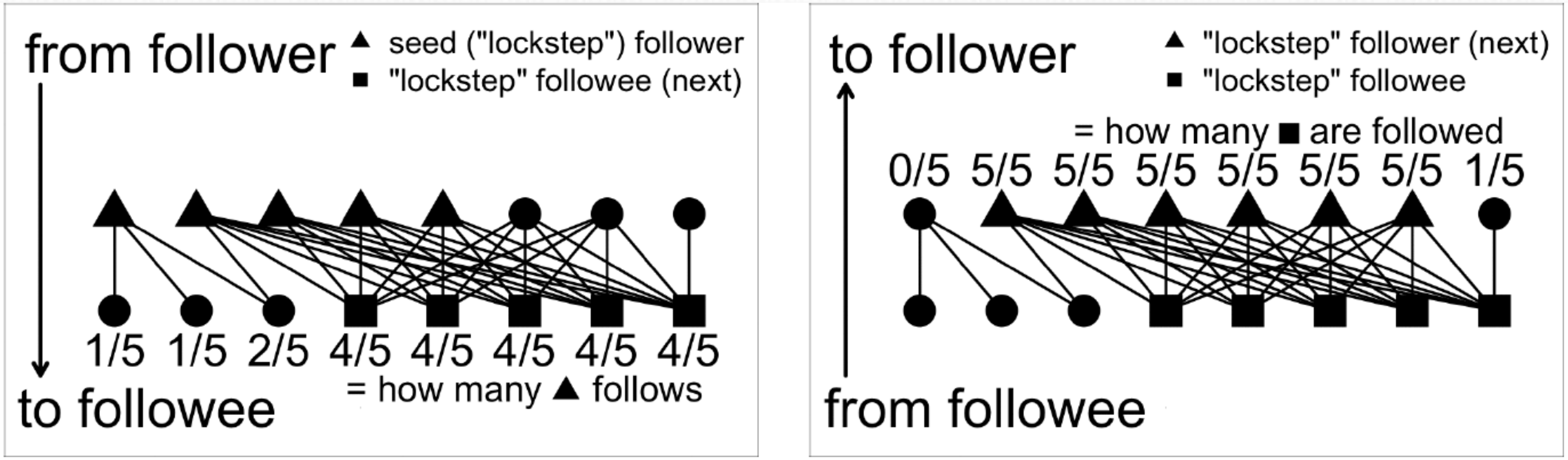
- Step 1: Seed selection
  - Spot “Rays” and “Pearls”
  - Catch seed followers
- Step 2: Belief Propagation
  - Blame followees with strange followers
  - Blame followers with strange followees

Automatically Spot "Rays" and "Pearls"



BP-based Algorithm

- Blame followees with strange followers
- Blame followers with strange followees

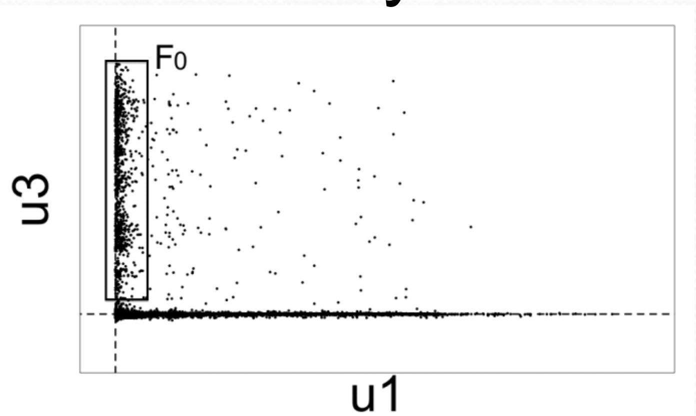


(a) select "lockstep" followees: from (seed) followers to followees

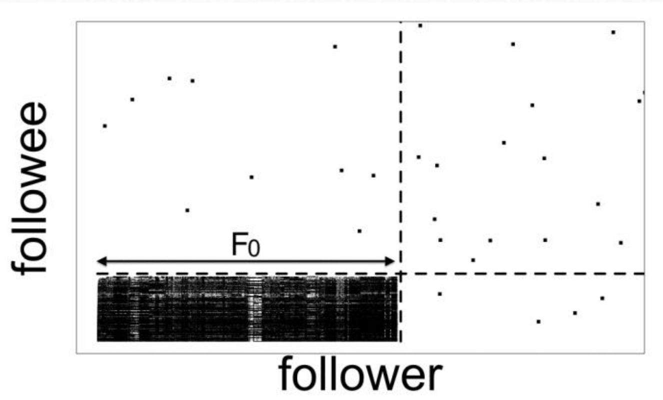
(b) select "lockstep" followers: from followees to followers

Real Data

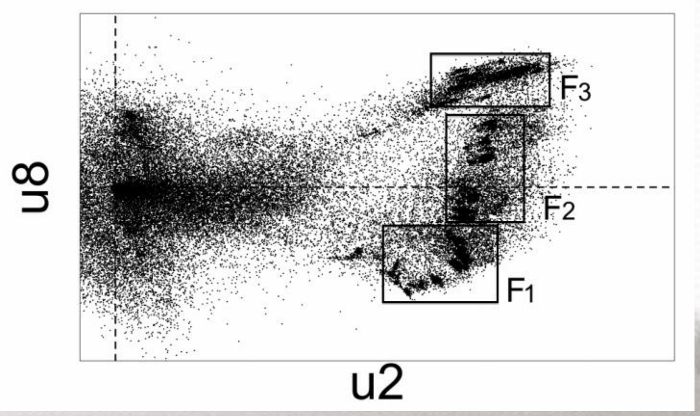
“Rays”



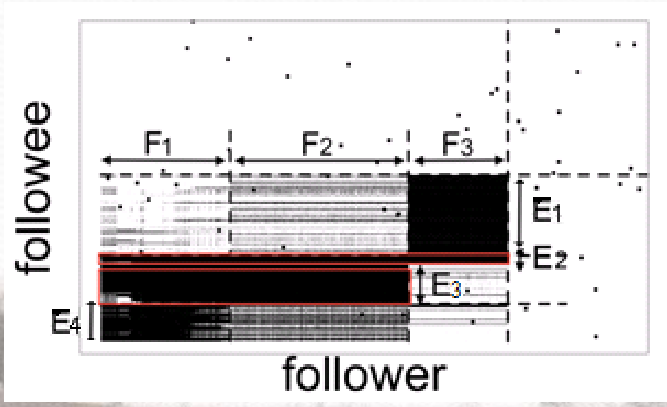
“Block”



“Pearls”

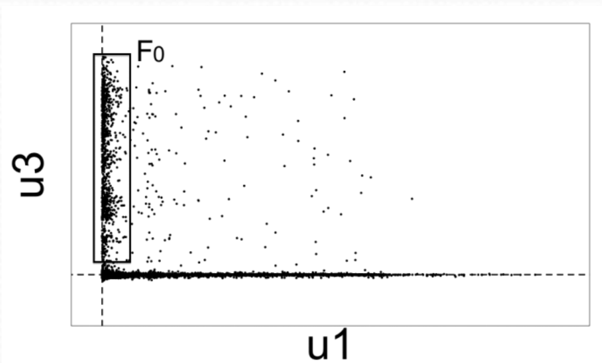


“Staircase”

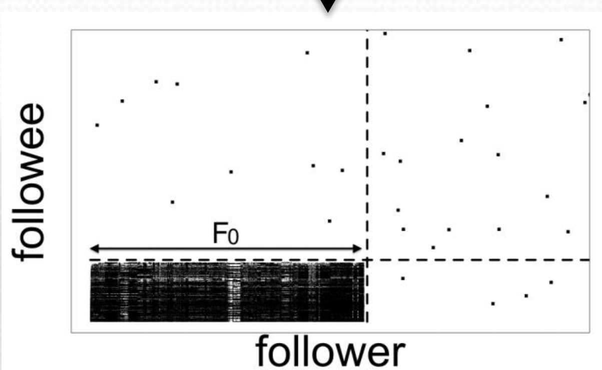


Real Data

“Rays”



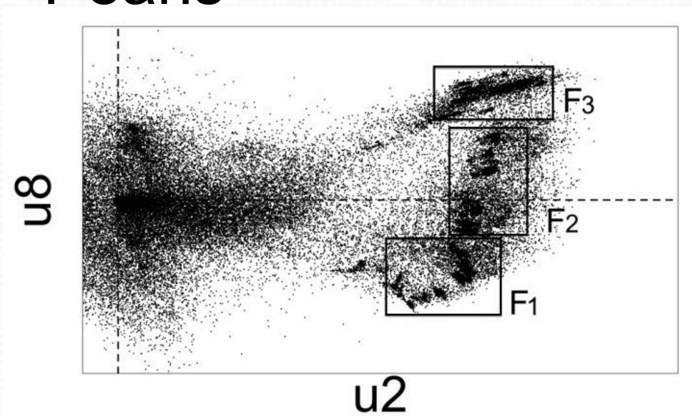
“Block”



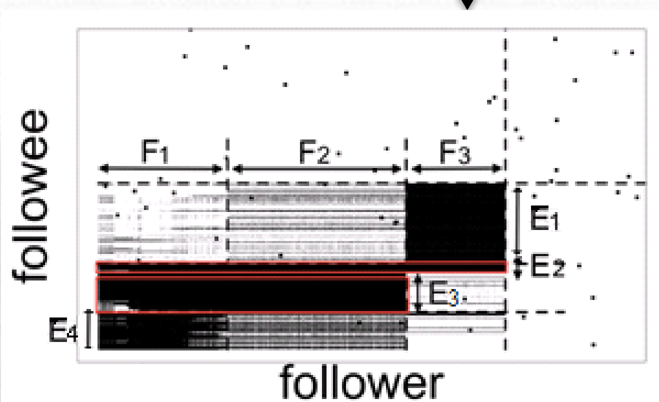
	“ray” F_0	“p
Num. seeds	100	
Size of block	$83,208 \times 30$	3, 1
Density	81.3%	[illegible]
Camouflage	0.14%	[illegible]
Fame	0.05%	[illegible]
Out-degree	231±109	[illegible]
In-degree	2.0±1.4	[illegible]

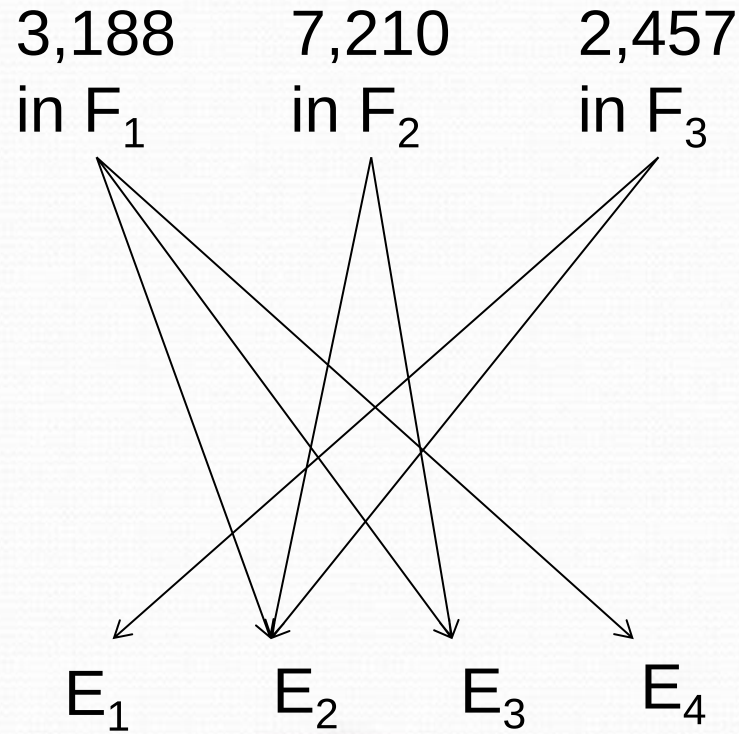
Real Data

“Pearls”



“Staircase”

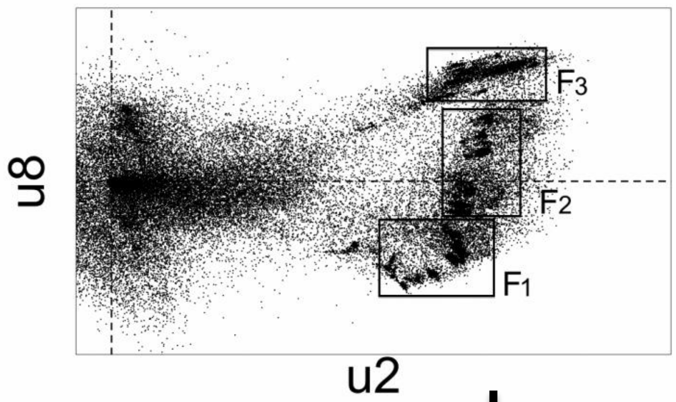




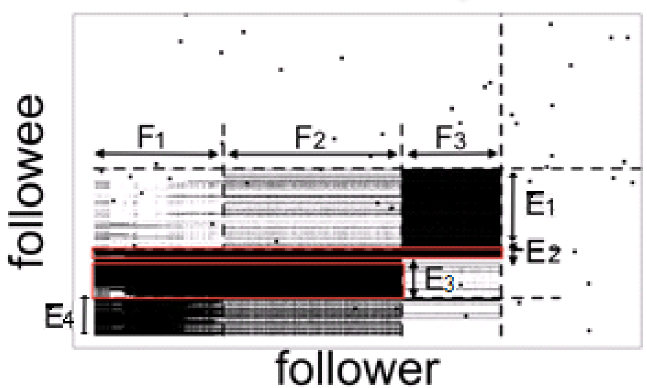
“F-E”	F_1-…	F_2-…	F_3-…
Density	91.3%	92.6%	89.1%

Real Data

"Pearls"



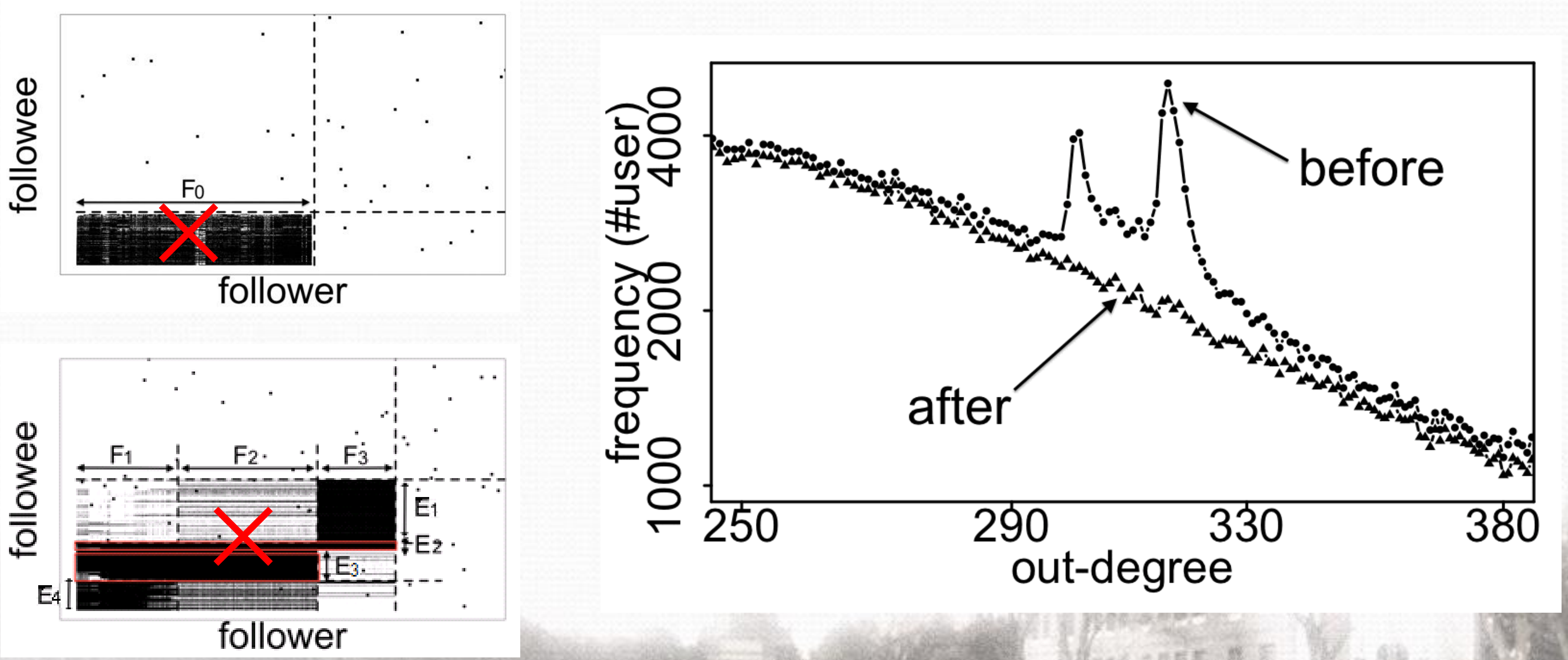
"Staircase"



"pearl" F_1	"pearl" F_2	"pearl" F_3	"pearl" Total
1,239	107	990	—
$3,188 \times 135$	$7,210 \times 79$	$2,457 \times 148$	$10,052 \times 270$
91.3%	92.6%	89.1%	43.1%
0.06%	0.10%	0.05%	0.07%
1.93%	1.94%	1.72%	1.73%
310±7	312±7	304±5	310±7
9±6	10±6	17±13	12±9

Real Data

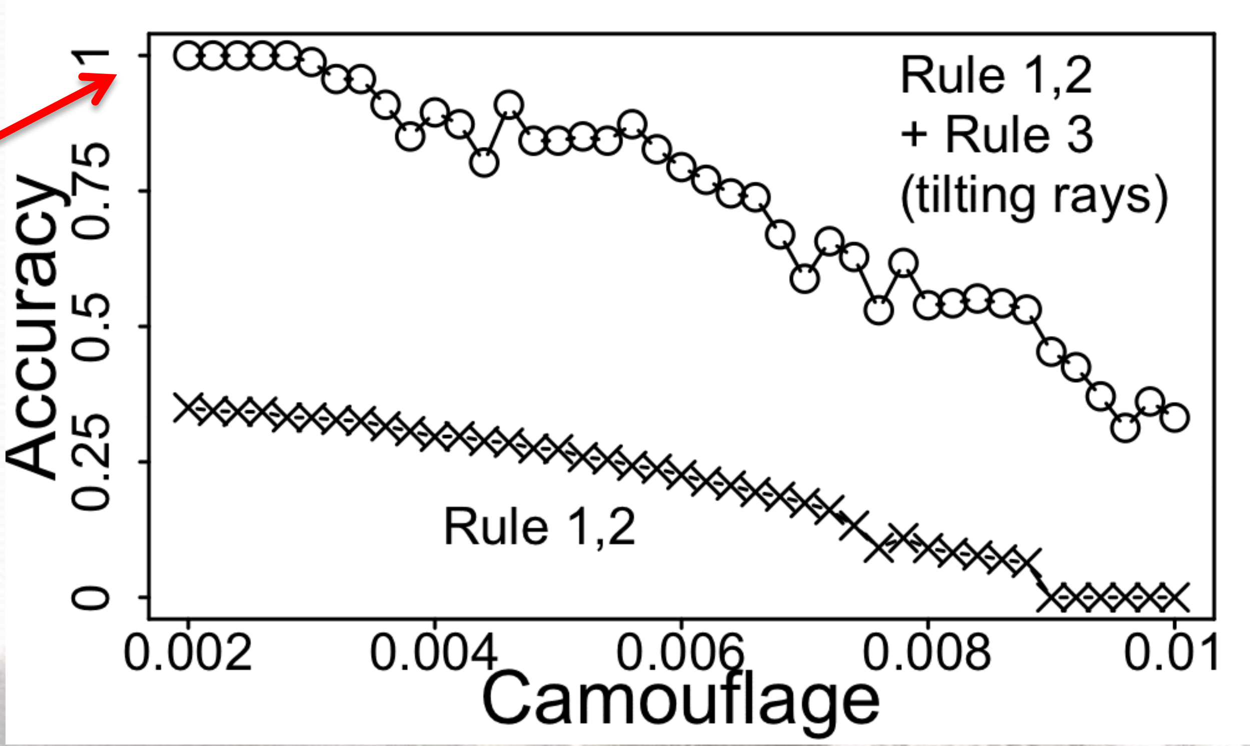
❖ Spikes on the out-degree distribution



Synthetic Data

❖Inject lockstep behavior with "**camouflage**"

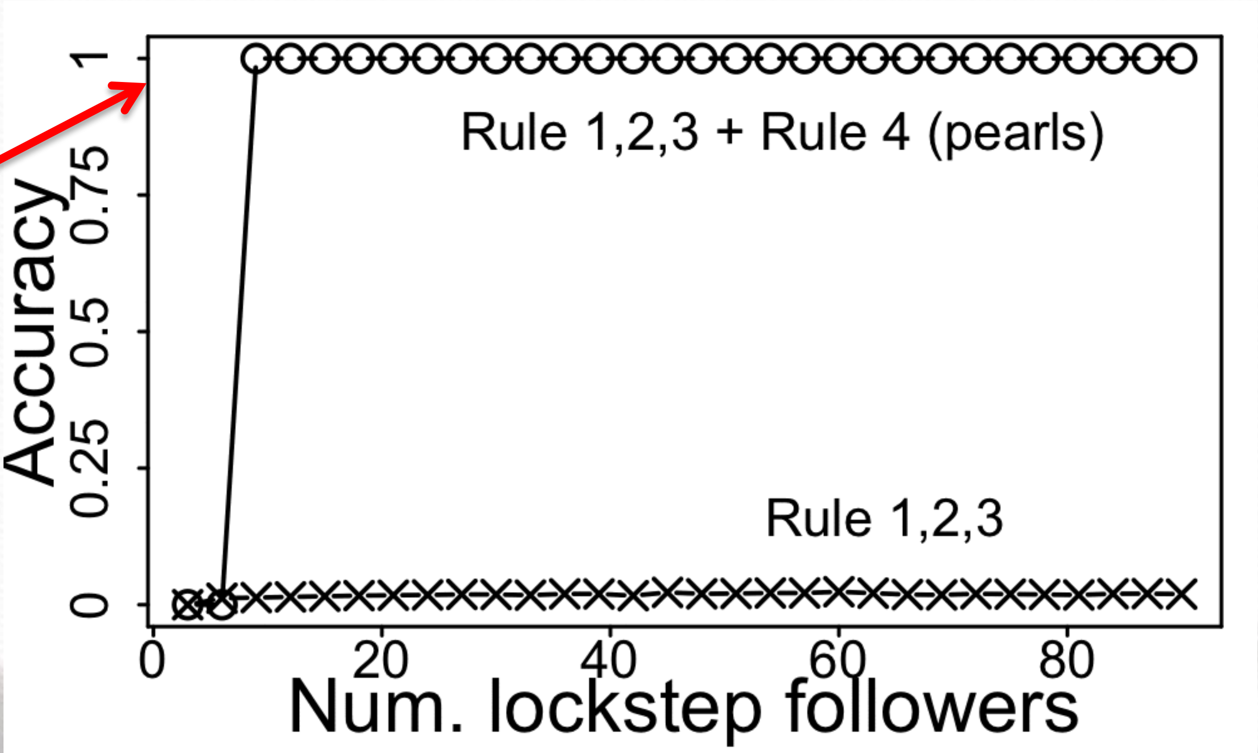
perfect



Synthetic Data

- Inject **overlapping** lockstep behavior

perfect



Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. Inferring Strange Behavior from Connectivity Pattern in Social Networks.*The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2014.

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

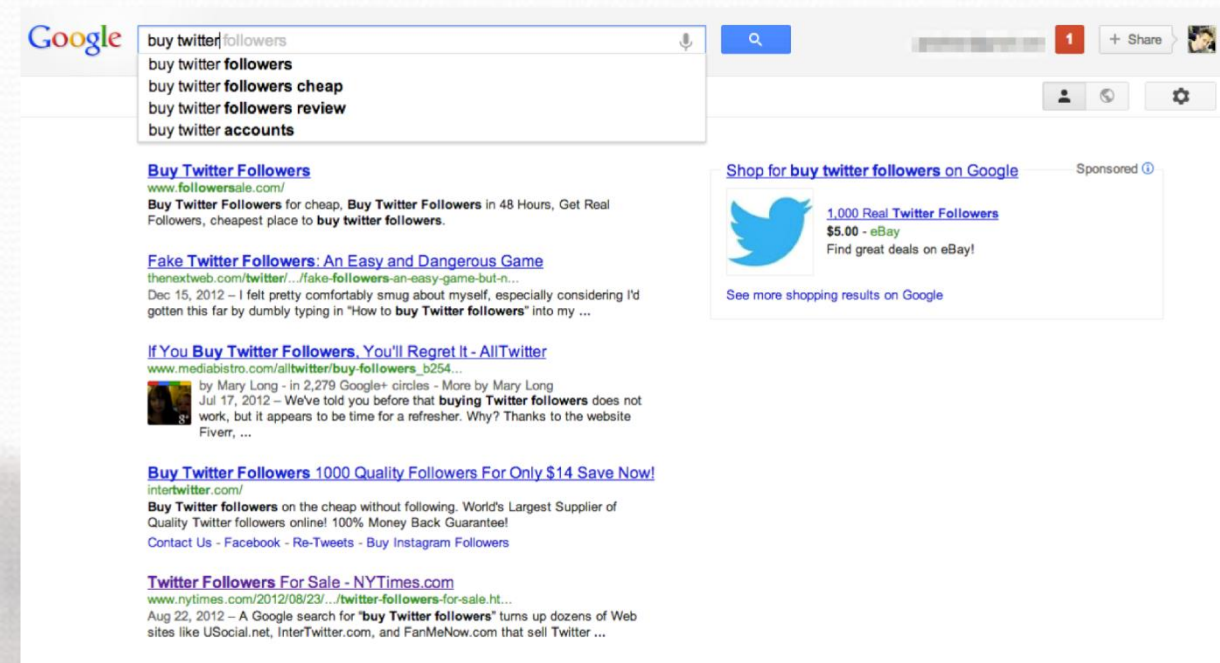
Good user-item links

- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Bad user-user links

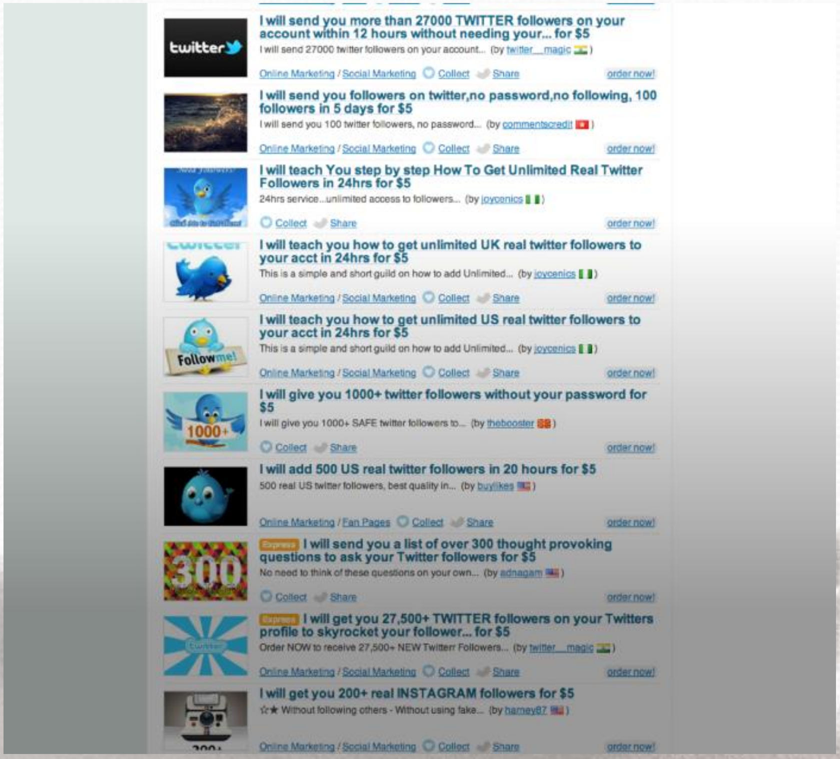
Buy Twitter Followers

- **(Fake) friends with (Real) benefits**
- **I paid $5 for 4,000 Twitter followers, and here's what I found**
- https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3



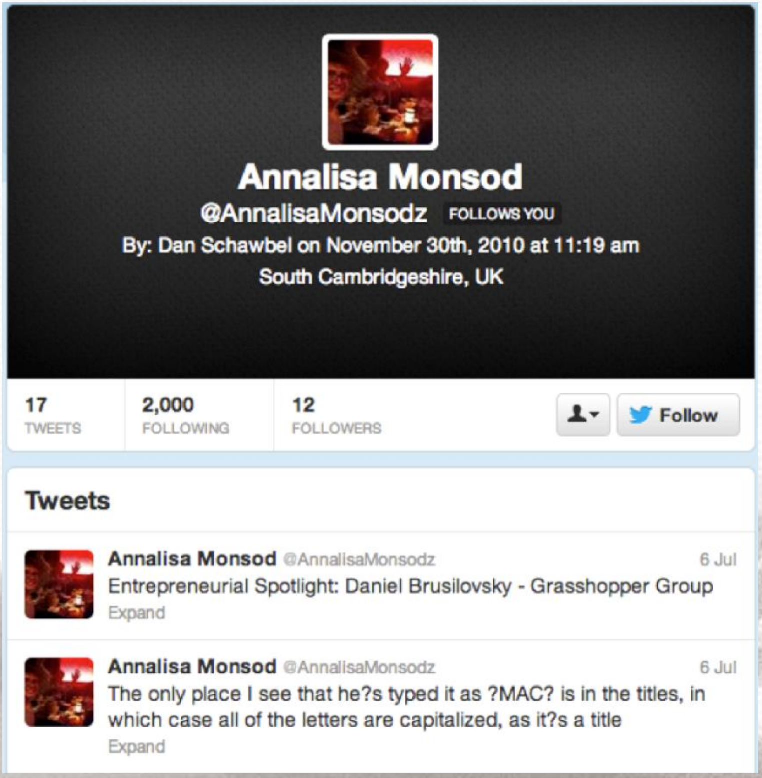
Buy Twitter Followers

- **(Fake) friends with (Real) benefits**
- **I paid $5 for 4,000 Twitter followers, and here's what I found**
- https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3



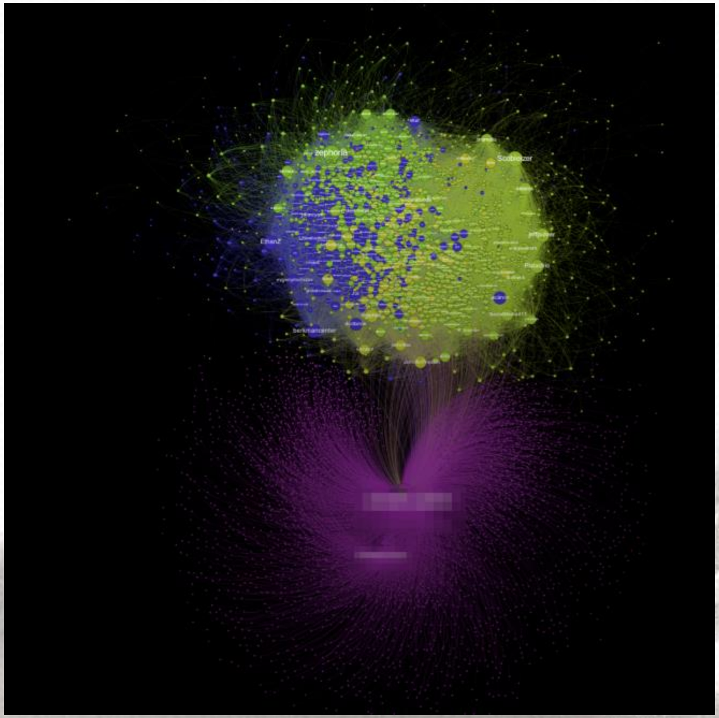
Buy Twitter Followers

- **(Fake) friends with (Real) benefits**
- **I paid $5 for 4,000 Twitter followers, and here's what I found**
- https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3



Buy Twitter Followers

- **(Fake) friends with (Real) benefits**
- **I paid $5 for 4,000 Twitter followers, and here's what I found**
- https://medium.com/i-data/fake-friends-with-real-benefits-eec8c4693bd3

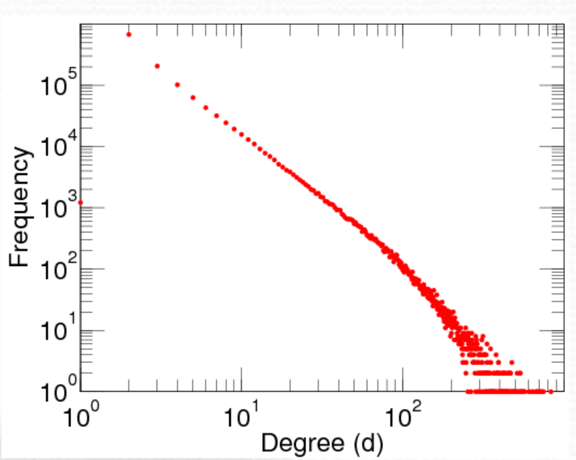
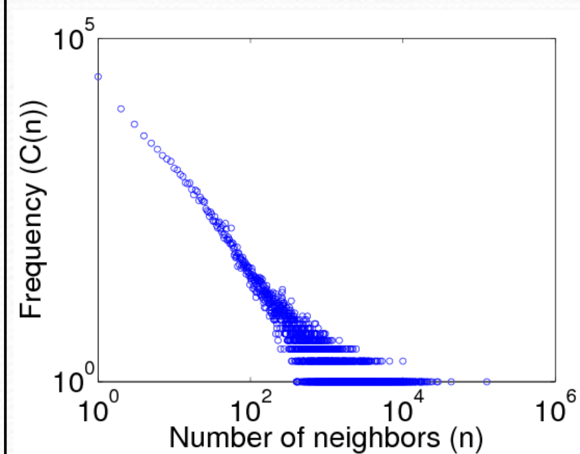
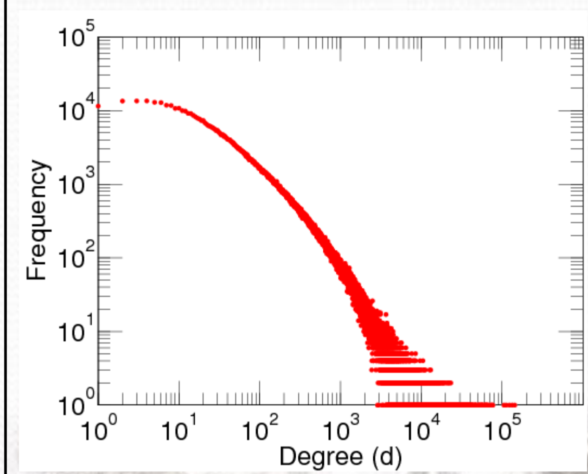


Zombie Follower Detection

- Challenges
  - **Scalability**: How to catch zombie followers from large graphs of millions of nodes and billions of edges?

Out-degree Distribution

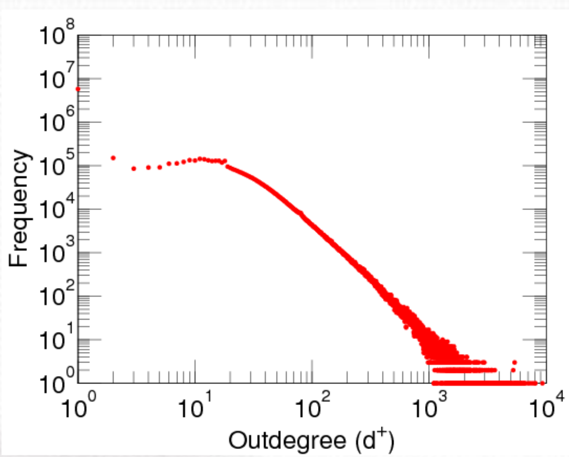
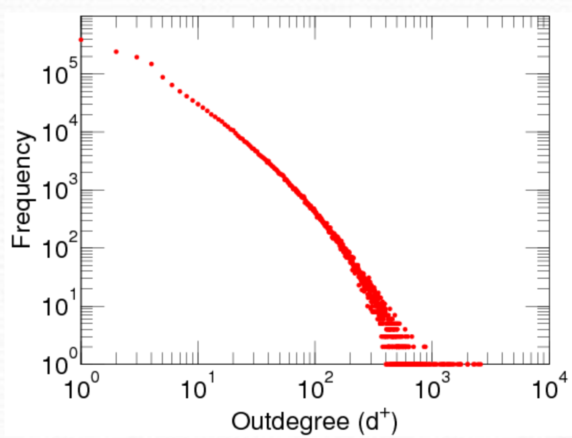
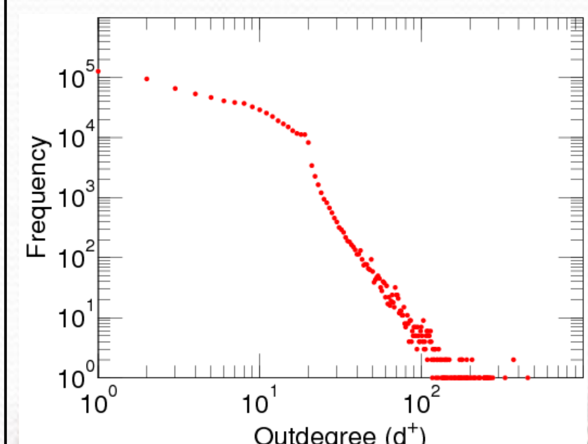
- Power-law distribution (bipartite graph)

DBLP **Author**-publication	**Epinions** **User**-rates-product	**Delicious** **User**-tag
		

http://konect.uni-koblenz.de/networks/

Out-degree Distribution

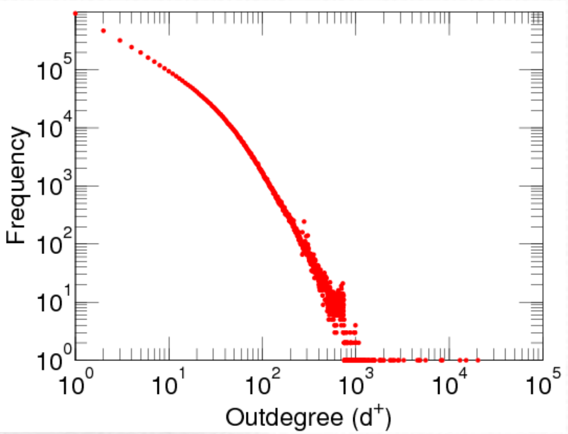
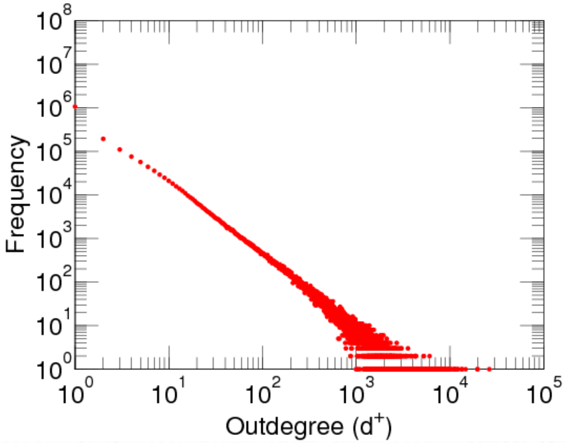
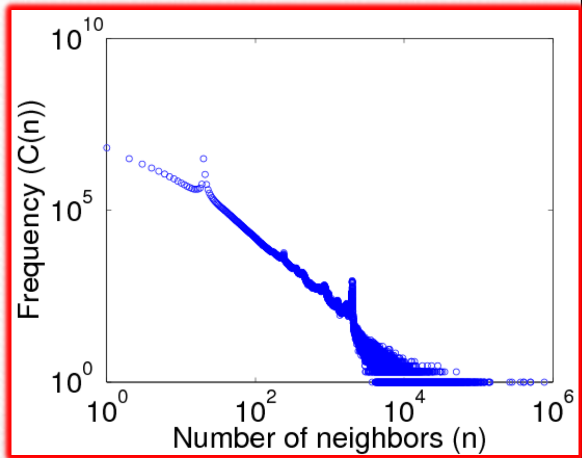
- Power-law distribution (directed graph)

Wikipedia **Page**-links-to-page	**Baidu** **Page**-links-to-page	**Google** **Page**-links-to-page
		

http://konect.uni-koblenz.de/networks/

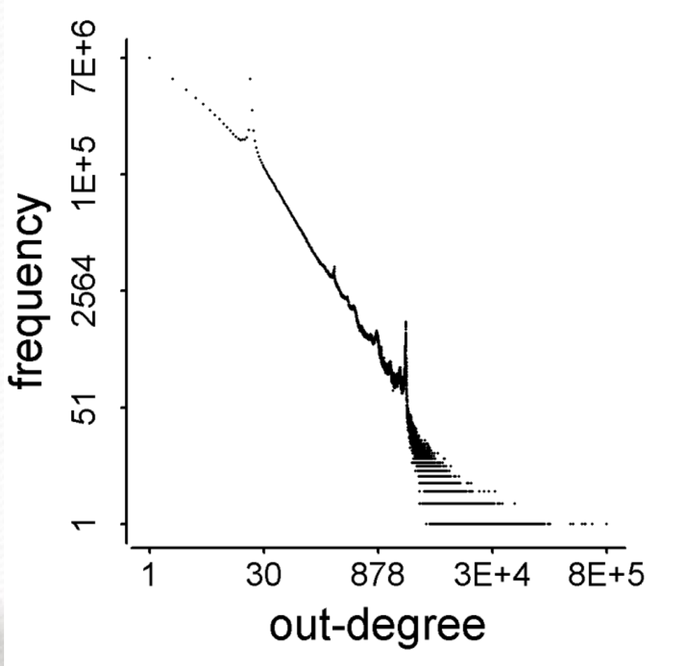
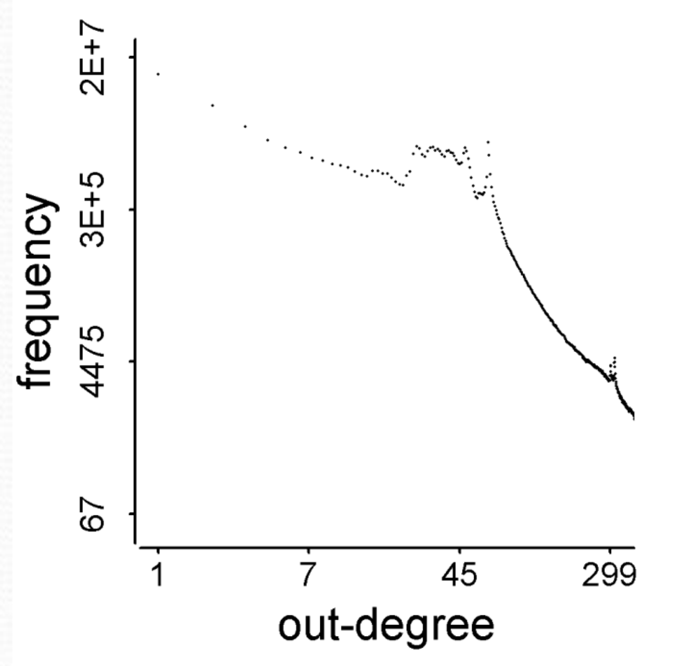
Out-degree Distribution

- Power-law distribution (directed graph - social network)

LiveJournal **User**-user	**Flickr** **User**-user	**Twitter** **Who**-follows-whom
		

http://konect.uni-koblenz.de/networks/

What We Have…

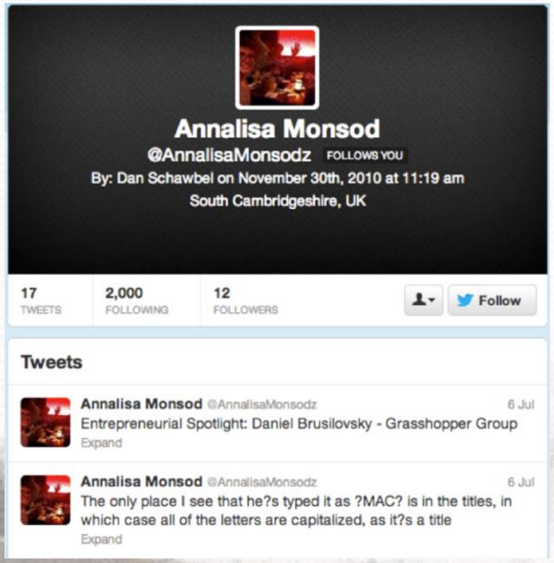
Twitter (2009) 41 million users	**Tencent Weibo (2011)** 117 million users
	

Zombie Follower Detection

- Challenges
  - **Scalability**: How to catch zombie followers from large graphs of millions of nodes and billions of edges? **Can we explain the spikes on out-degree distributions?**

Zombie Follower Detection

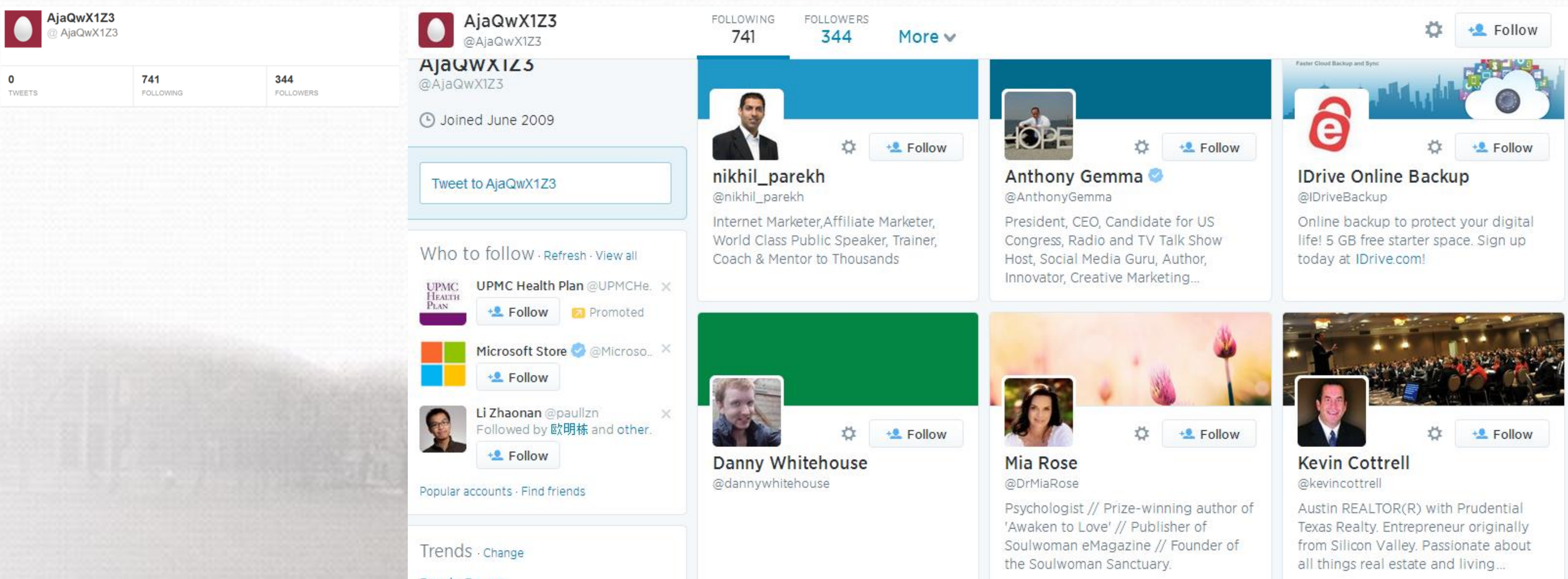
- Challenges
  - Scalability: How to catch zombie followers from large graphs of millions of nodes and billions of edges? Can we explain the spikes on out-degree distributions?
  - **Camouflage**: fake profile, no or little content, extra performance



Tsinghua University

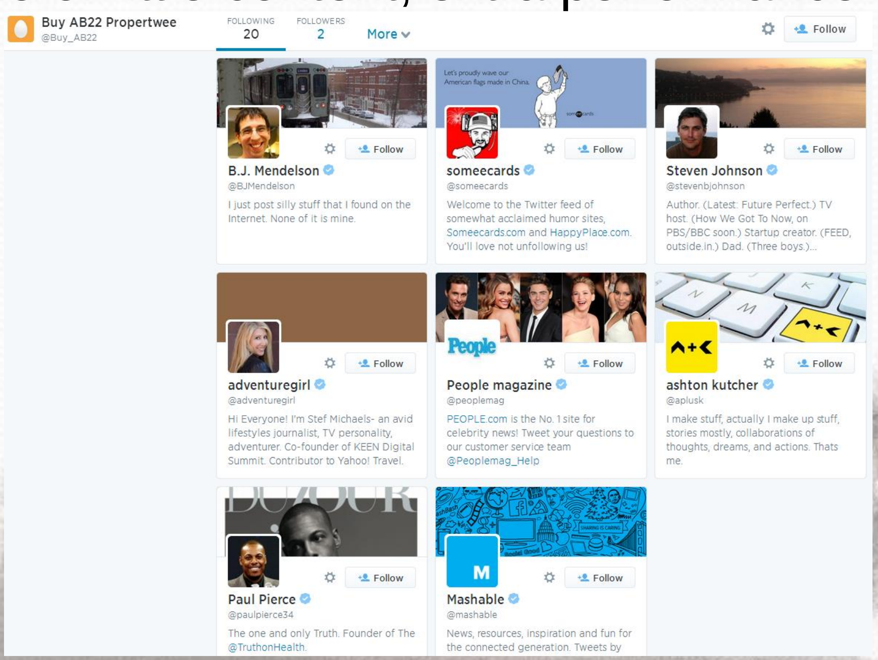
Zombie Follower Detection

- Challenges
  - Scalability: How to catch zombie followers from large graphs of millions of nodes and billions of edges? Can we explain the spikes on out-degree distributions?
  - **Camouflage**: fake profile, no or little content, extra performance



Zombie Follower Detection

- Challenges
  - Scalability: How to catch zombie followers from large graphs of millions of nodes and billions of edges? Can we explain the spikes on out-degree distributions?
  - **Camouflage**: fake profile, no or little content, extra performance



Zombie Follower Detection

- Challenges
  - Scalability: How to catch zombie followers from large graphs of millions of nodes and billions of edges? Can we explain the spikes on out-degree distributions?
  - Camouflage: fake profile, no or little content, extra performance
- Previous approaches

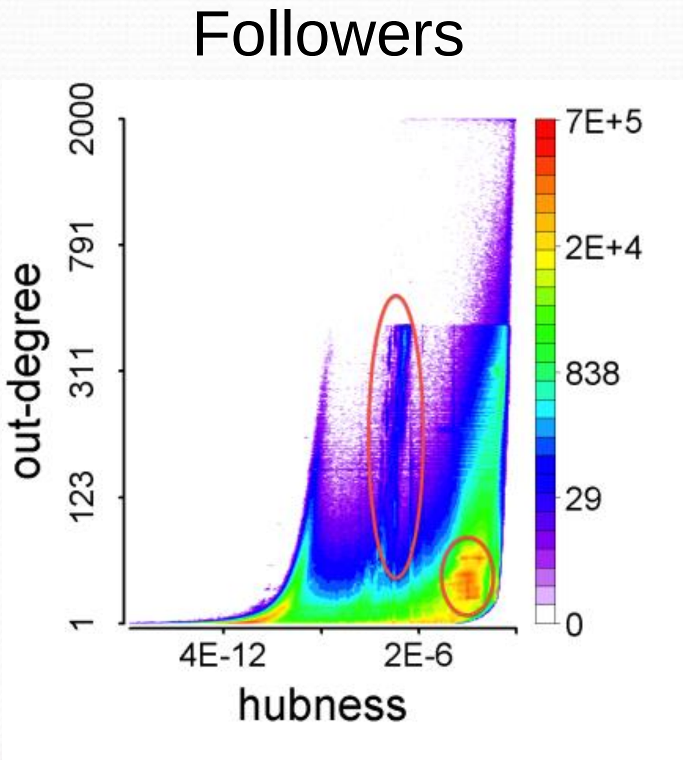
Label (+1,-1)	Out-degree	In-degree	#tweet	#url in tweets	#hashtag in tweets

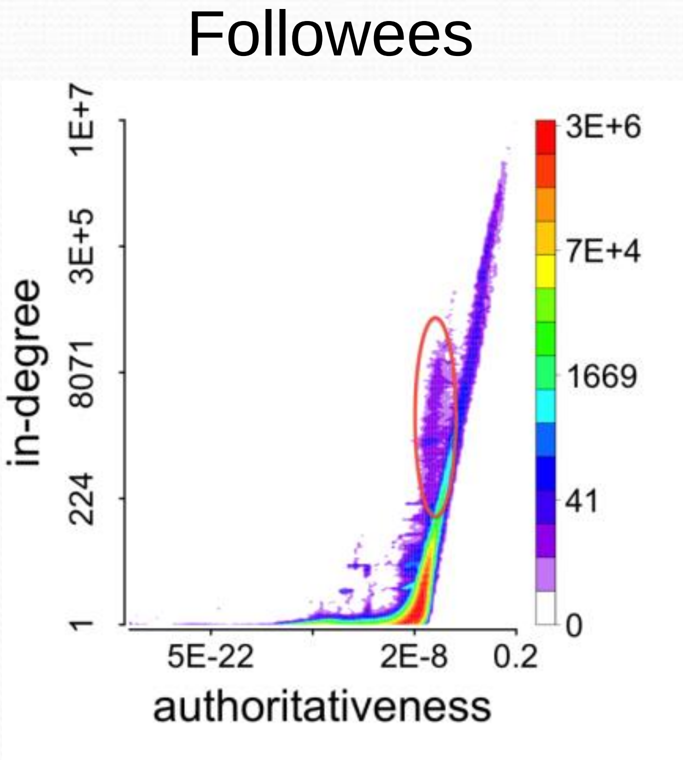
classifier

Graph-based features

Content-based features
0, 0, 0 sorry…

Graph-based Feature Space

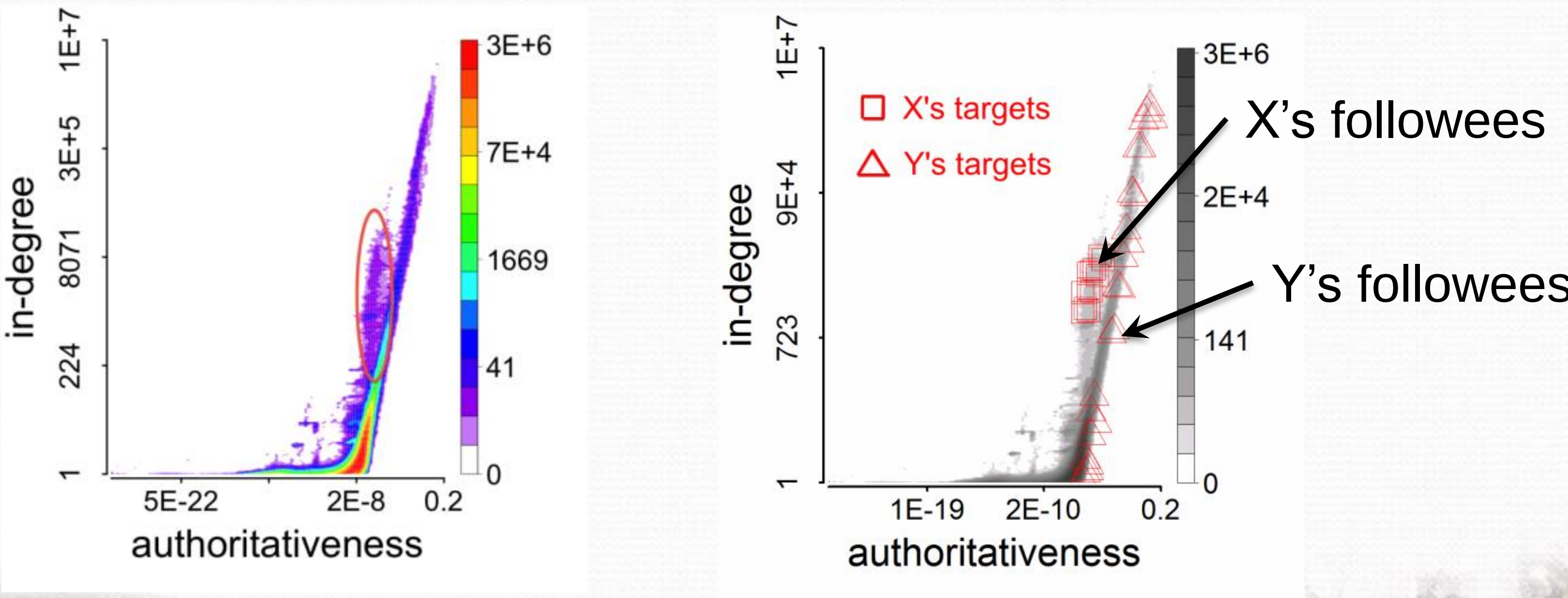




We spot some outliers, but are they suspicious?
Answer: **suspicious behavior patterns** create suspicious graph structure

Compare Zombie Follower and Normal User

- X = @Buy_AB22: a zombie follower with 20 followees
- Y = a random user with 20 followees



- Suspicious behavior: similar with each other, different from normal.

Suspicious Behavior Pattern

- Behavior-based features
- Synchronicity: similarity between one's followees

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

- Normality: similarity between one's followee and everyone

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

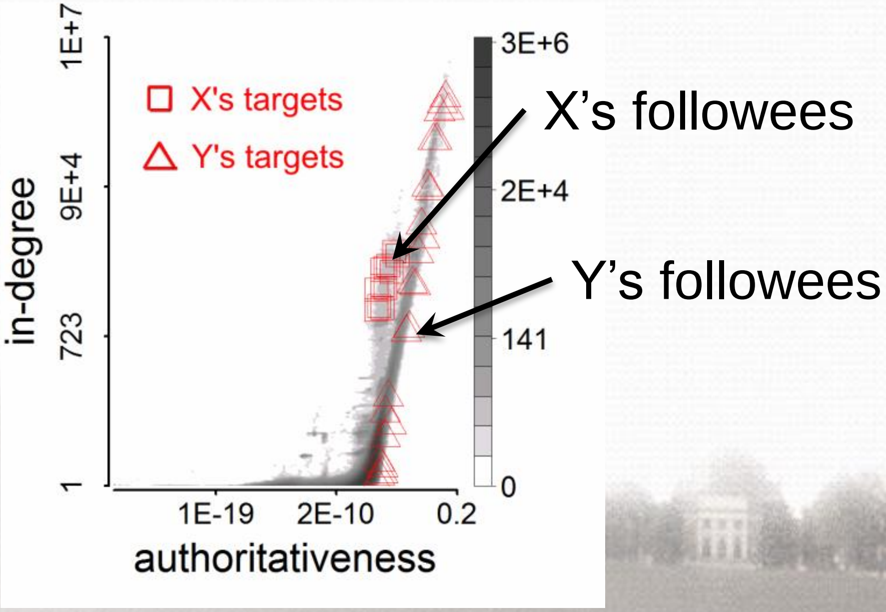
Feature vector

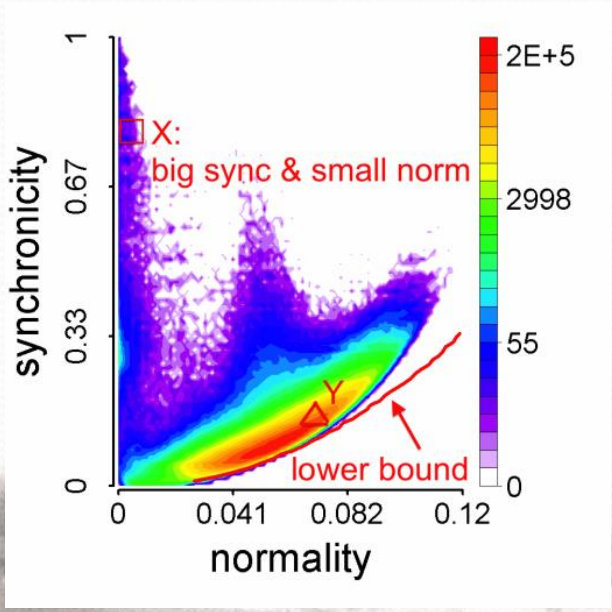
- **Theorem** *For any distribution, there is a parabolic lower limit in the synchronicity-normality plot.*

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

Synchronicity-Normality Plot

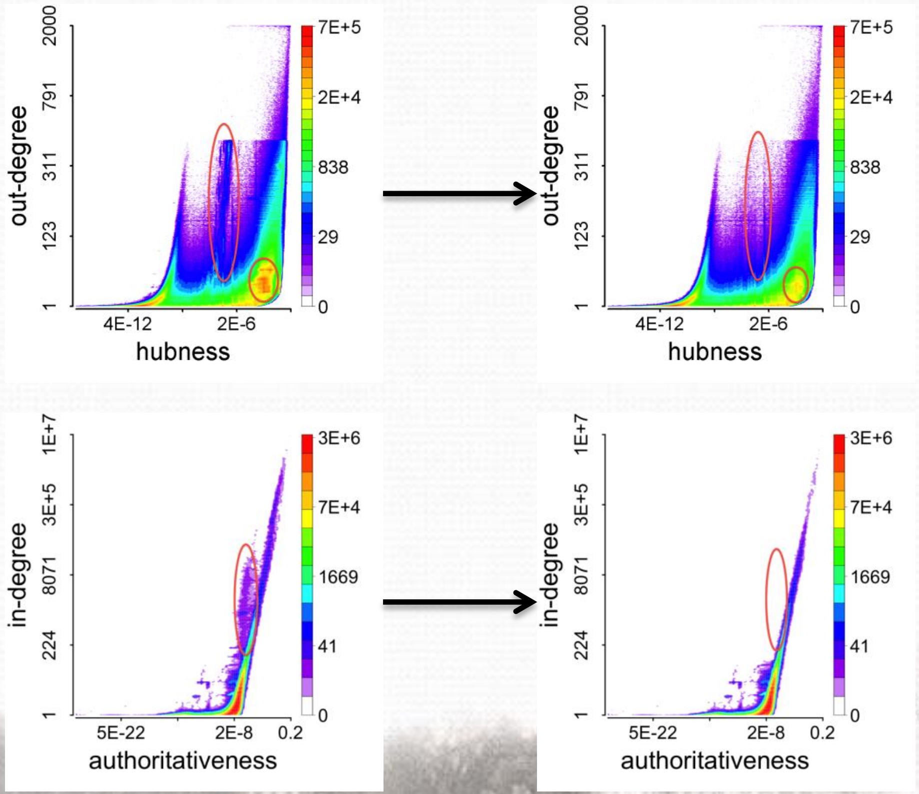
- Use *CatchSync* to catch Zombie follower
  - **Synchronized** and **abnormal** behavior
  - **Big** synchronicity and **small** normality; away from the densest
  - Last step: develop a distance-based anomaly detection method!





Goal 1: Do we catch the anomalies?

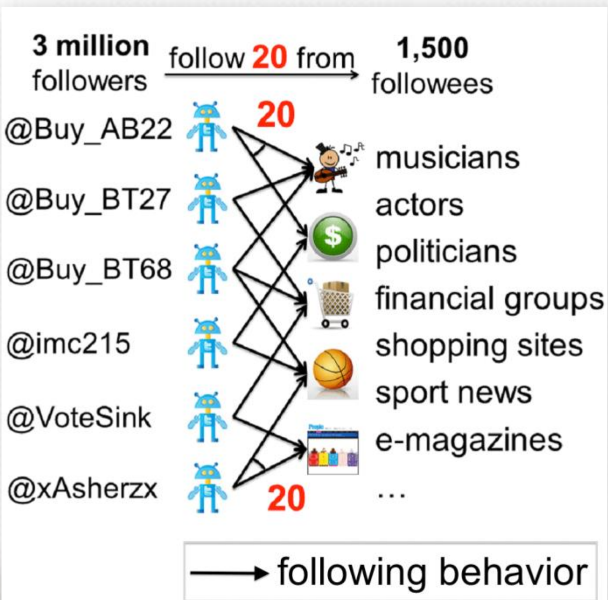
- Graph-based feature space
- Followers

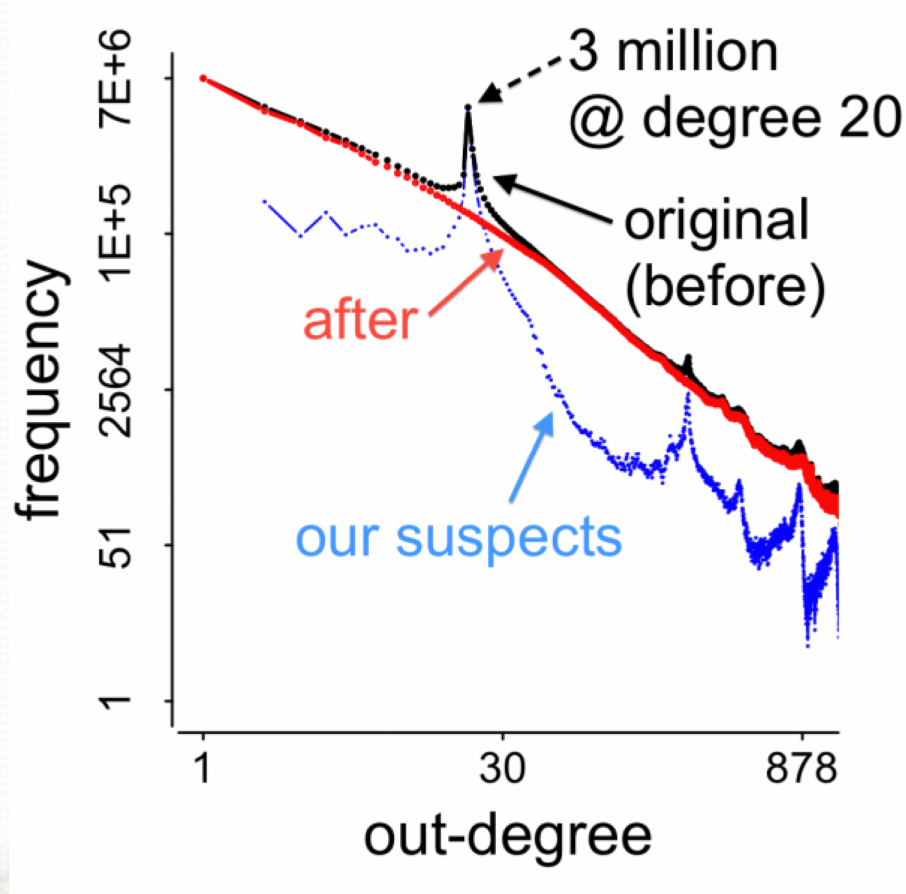


- Followees

Goal 1: Do we catch the anomalies?

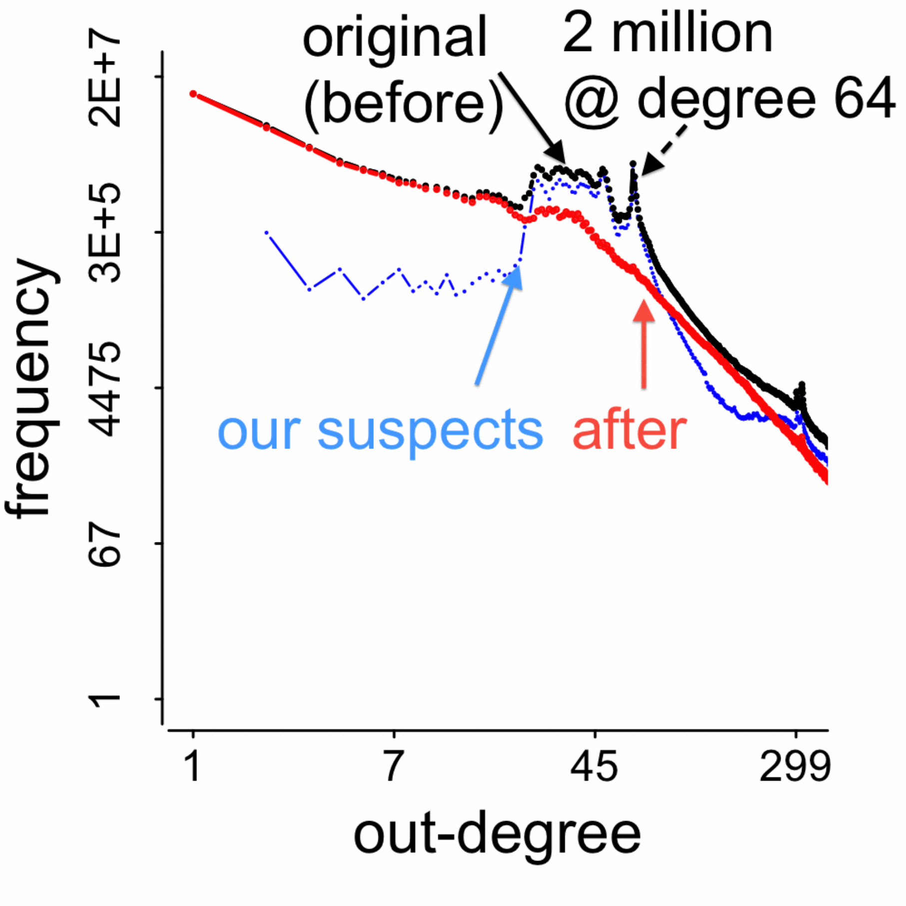
- Out-degree distribution
- Twitter





Goal 1: Do we catch the anomalies?

- Out-degree distribution
- Tencent Weibo

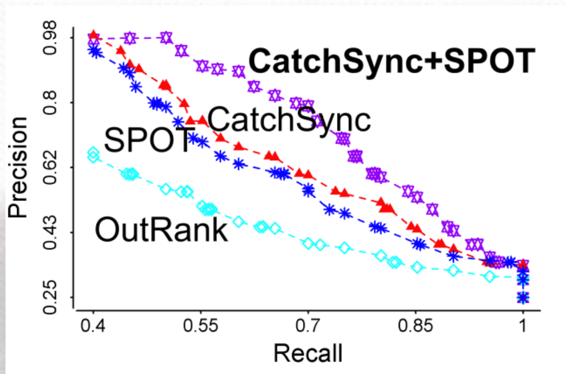


Goal 2: Do we catch suspicious users?

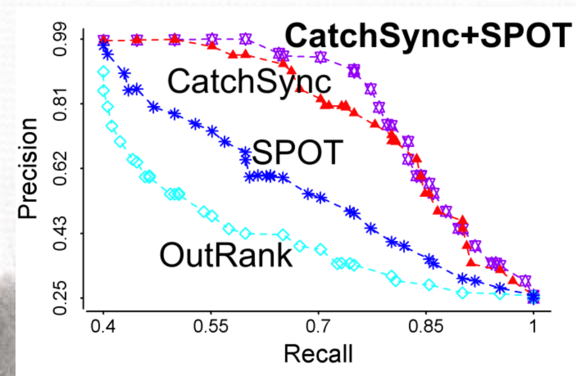
- Manually label 1,000 users as suspicious or not by reading their pages (profiles, tweets, followers, followings, etc.)
- Better than content-based method SPOT
- **Complementary** with SPOT

	TWITTERSG	WEIBOSG
CATCHSYNC	0.751	0.694
OUTRANK	0.412	0.377
SPOT	0.597	0.653
CATCHSYNC+SPOT	**0.813**	**0.785**

Twitter

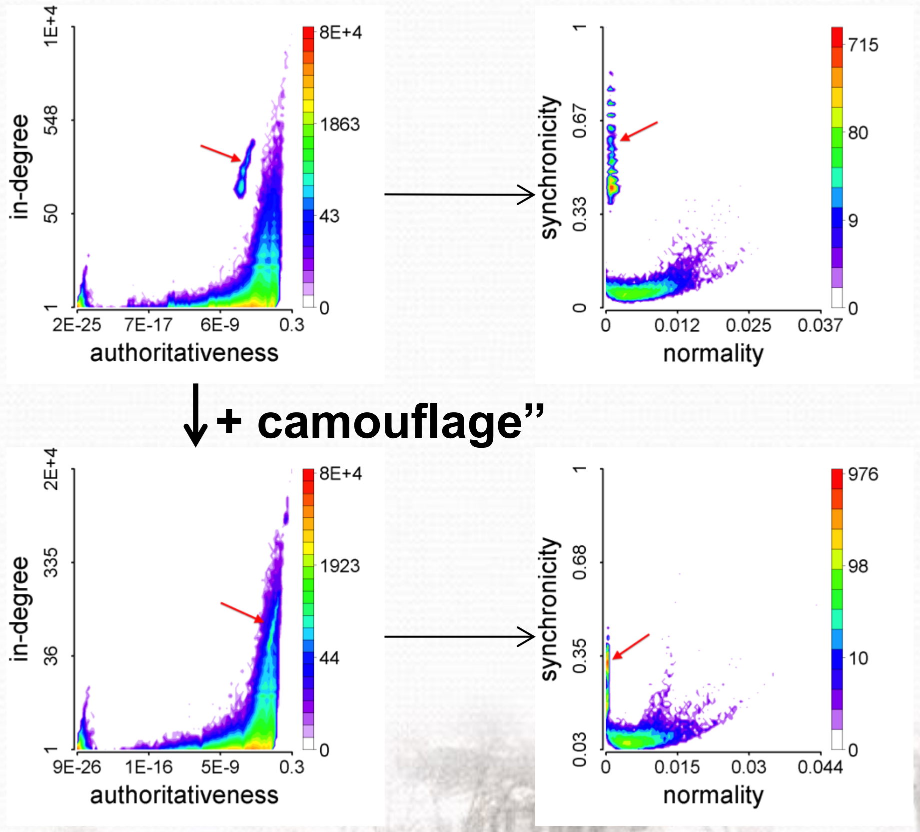


Tencent Weibo



Goal 3: Is CatchSync anti-camouflage?

- Camouflage
  - +10%, +50%
  - +random, +popular



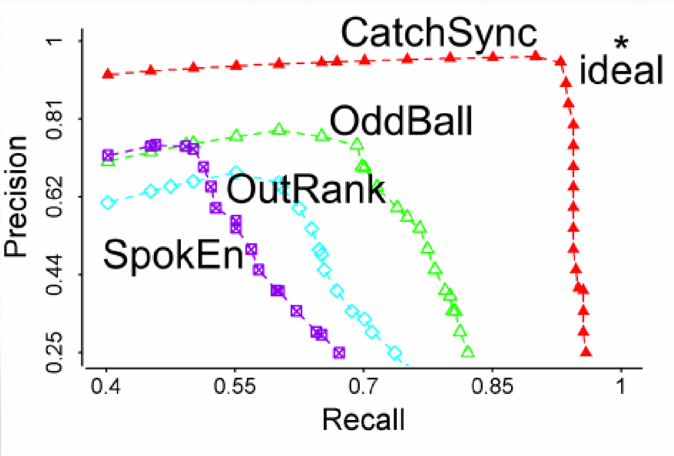
Hide in feature space **Show in SN-plot**

Goal 3: Is CatchSync anti-camouflage?

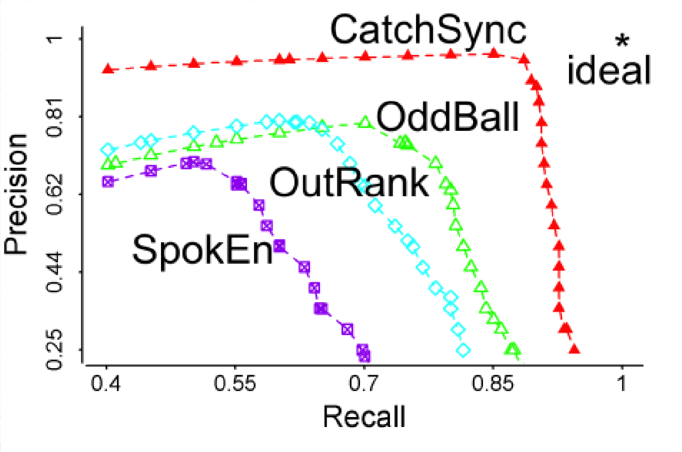
Accuracy:

Synthetic graph	SYNTH-3M	SYNTH-3M-RAND		SYNTH-3M-POP	
Camouflage (d_{camou})	None (0)	10%	50%	10%	50%
CATCHSYNC	**0.956**	**0.910**	**0.764**	**0.885**	**0.792**
ODDBALL	0.755	0.702	0.525	0.657	0.433
OUTRANK	0.725	0.678	0.516	0.694	0.392
SPOKEN	0.677	0.586	0.470	0.553	0.351

Random "Camouflage":



Popular "Camouflage":



Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. CatchSync: Catching Synchronized Behavior in Large Directed Graphs. *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. Detecting Suspicious Following Behavior in Multimillion-Node Social Networks. *The 23rd international conference on World Wide Web companion (WWW)*, 2014. *(Poster)*

Outline

- Social contextual recommendation [CIKM'12+TKDE'14]
- Cross-domain social recommendation [CIKM'12]
- Behavior discovery and prediction [KDD'14]

- Dense bipartite core detection [PAKDD'14]
- Zombie follower detection [KDD'14]

Good user-item links

Bad user-user links

References

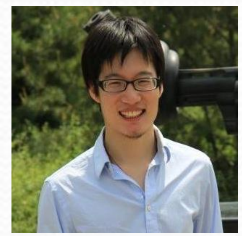
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. CatchSync: Catching Synchronized Behavior in Large Directed Graphs. *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.
- **Meng Jiang**, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu and Shiqiang Yang. FEMA: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery. *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. Inferring Strange Behavior from Connectivity Pattern in Social Networks.*The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2014.
- **Meng Jiang**, Peng Cui, Alex Beutel, Christos Faloutsos and Shiqiang Yang. Detecting Suspicious Following Behavior in Multimillion-Node Social Networks. *The 23rd international conference on World Wide Web companion (WWW)*, 2014. *(Poster)*
- **Meng Jiang**, Peng Cui, Fei Wang, Wenwu Zhu and Shiqiang Yang. Scalable Recommendation with Social Contextual Information. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2014.
- **Meng Jiang**, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu and Shiqiang Yang. Social Contextual Recommendation. *The 21st ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.
- **Meng Jiang**, Peng Cui, Fei Wang, Qiang Yang, Wenwu Zhu and Shiqiang Yang. Social Recommendation across Multiple Relational Domains. *The 21st ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.
- Lu Liu, Feida Zhu, **Meng Jiang**, Jiawei Han, Lifeng Sun, Shiqiang Yang. Mining Diversity on Social Media Networks. *Multimedia Tools and Applications*, 2012.
- Lu Liu, Jie Tang, Jiawei Han, **Meng Jiang**, Shiqiang Yang. Mining Topic-Level Influence in Heterogeneous Networks. *The 19th ACM International Conference on Information and Knowledge Management (CIKM)*, 2010.

Acknowledgements

- Tsinghua University

Shiqiang Yang

Wenwu Zhu



Peng Cui

Lu Liu

- Carnegie Mellon University

Christos Faloutsos

Alex Beutel

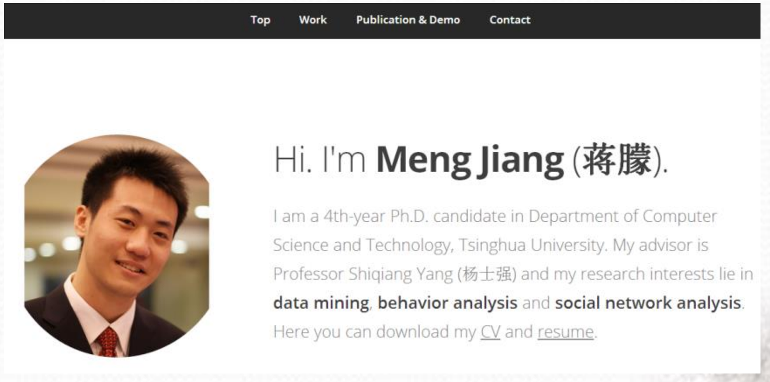
- IBM T. J. Watson Research Center

Fei Wang

Thank you!

Welcome to visit my homepage:

http://www.meng-jiang.com



❤ New friends ❤ Discussions ❤ Collaborations