社交媒体复杂行为分析与建模

清华大学 计算机科学与技术系
答辩人：蒋 朦 (2010310522)
指导老师：杨士强 教授

联系方式：mjiang89@gmail.com
个人网页：www.meng-jiang.com

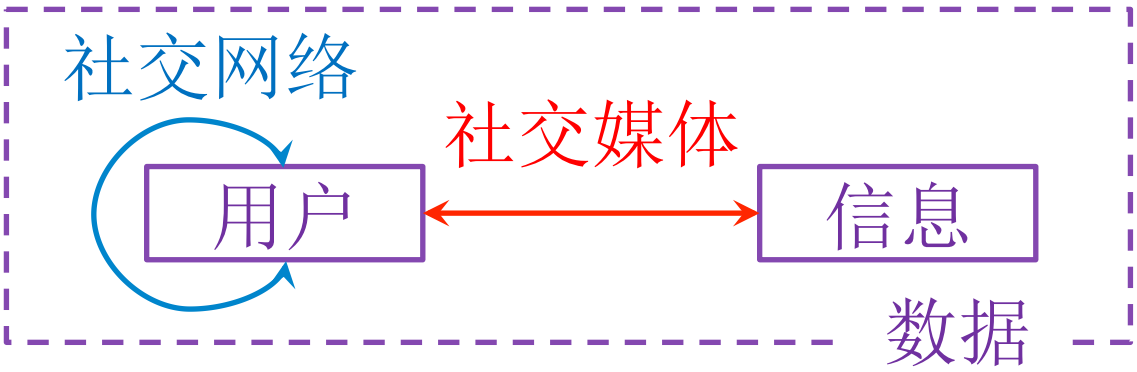
研究背景：社交媒体用户行为

- 社交媒体使研究用户行为成为可能

“用户-用户”交互

“用户-信息”交互

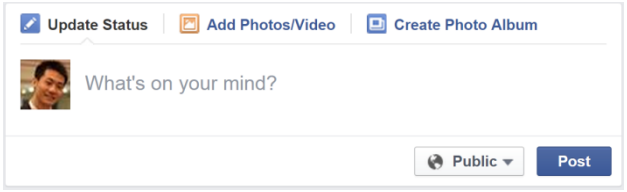
- 社交媒体用户行为比社交网络上更丰富



研究背景：用户行为和应用系统

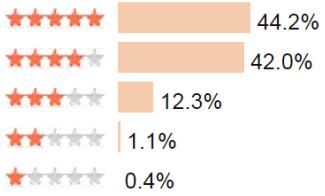
- 基于用户行为的应用系统创造巨大的市场价值

发布、转发文字和图片



新鲜事排序

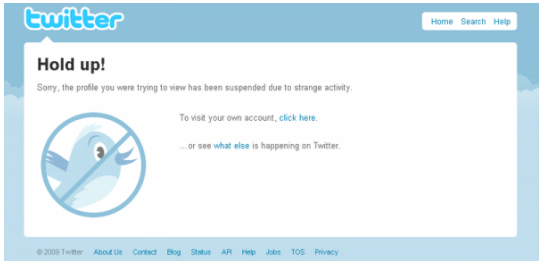
书籍、电影、音乐评分



推荐系统

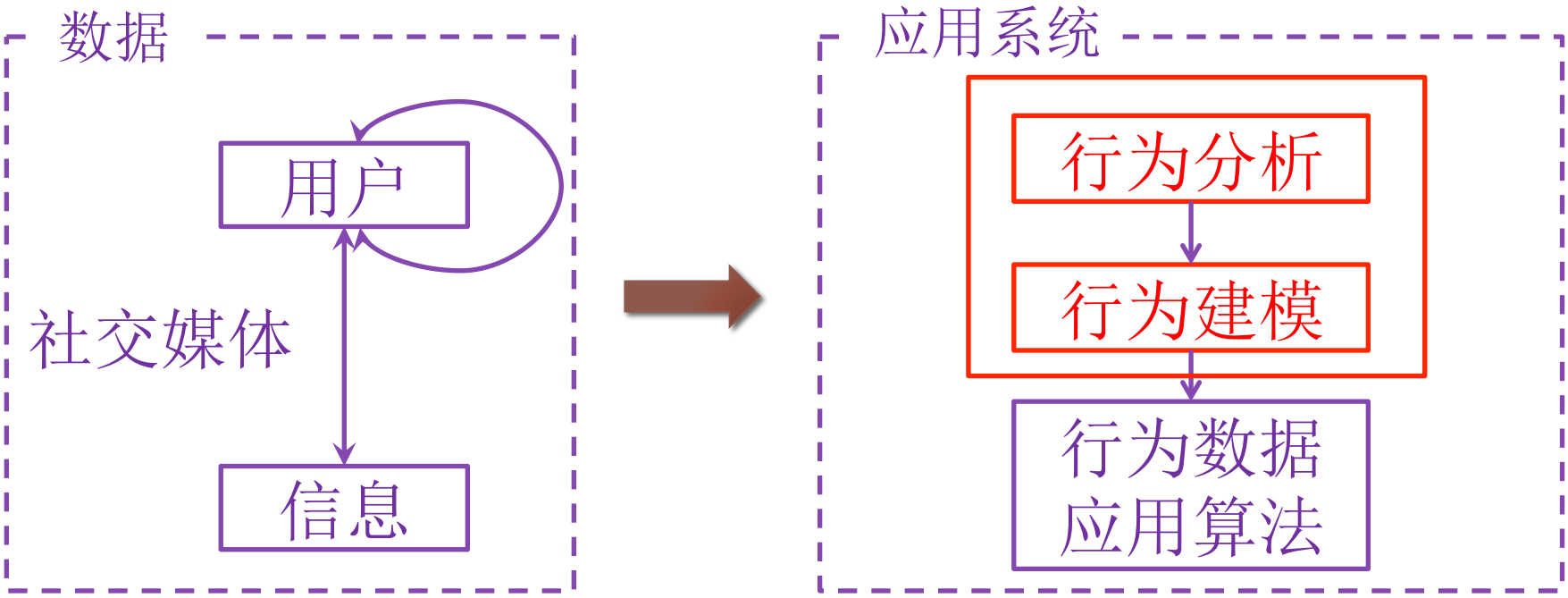
购买僵尸粉等欺诈行为

反欺诈系统：账号封禁



研究问题：社交媒体行为分析与建模

- 是 实现应用系统 的 第一步
- 是 社交媒体服务 的 基础
- 是 社交媒体数据处理 的 核心问题



相关工作

- 采纳信息行为预测
  - 基于内容过滤 [Balabanovic et al. ’97][Basu et al. AAAI’98]
  - 启发式协同过滤 [Herlock et al. SIGIR’99][Sarwar et al. WWW’01]
  - 社交关系同质性 [McPherson et al. ’01]
  - 社交影响力模式 [Leskovec et al. PAKDD’06]
  - 基于模型的协同过滤 [Yehuda KDD’08][Ma et al. CIKM’08][Ma et al. WSDM’11]
- 可疑行为检测
  - 重复内容的特征 [Jindal et al. WSDM’08]
  - 垃圾文本特征 [Lim et al. CIKM’10]
  - 脉冲现象 [Xie et al. KDD’12]
  - 情绪区分度 [Hu et al. ICDM’14]

研究对象：社交媒体的复杂行为

社交上下文

时空上下文

跨平台性

研究对象：社交媒体的复杂行为

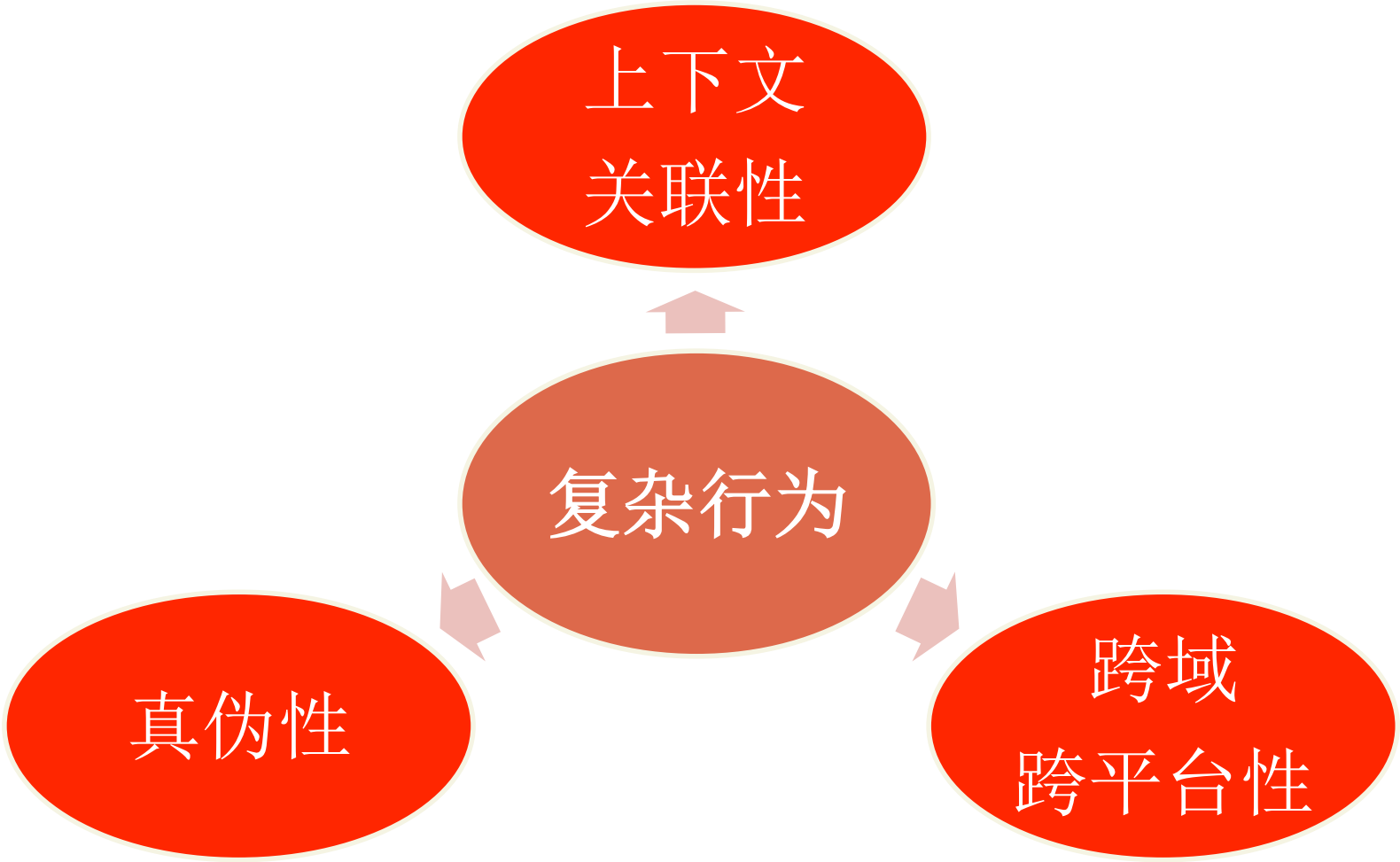
研究对象：社交媒体的复杂行为

真伪性

跨域

平台性

研究对象：社交媒体的复杂行为



研究路线

上下文关联行为的分析与建模

基于社交上下文的采纳信息行为模型

基于时空上下文的多面进化分析方法

跨域跨平台行为的分析与建模

跨域社交媒体的混合随机漫步算法

跨社交平台的半监督迁移学习算法

可疑行为的分析与建模

可疑社交用户同步行为的检测方法

可疑多面行为异常程度的衡量指标

研究路线

上下文关联行为的分析与建模

基于社交上下文的采纳信息行为模型

基于时空上下文的多面进化分析方法

跨域跨平台行为的分析与建模

跨域社交媒体的混合随机漫步算法

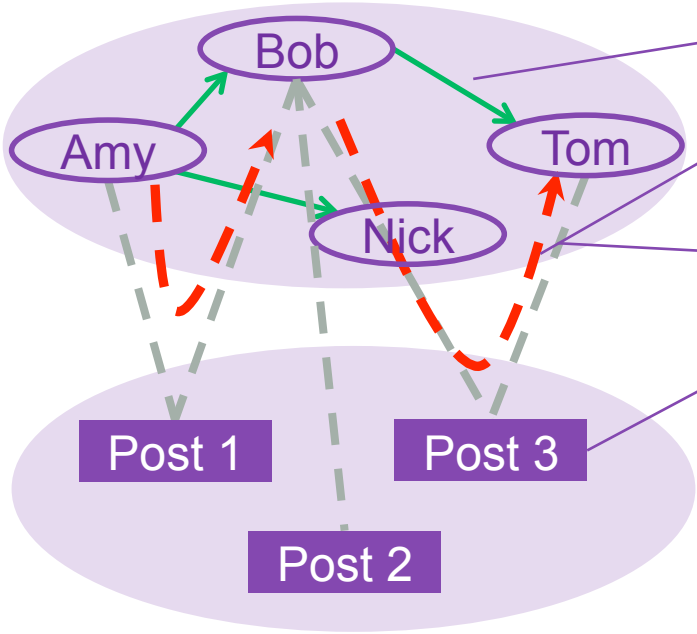
跨社交平台的半监督迁移学习算法

可疑行为的分析与建模

可疑社交用户同步行为的检测方法

可疑多面行为异常程度的衡量指标

基于上下文的行为预测问题和难点



(1) 社交关系

(2) “用户-用户”交互频度

(3) “用户-微博”采纳次数 过于稀疏

(4) 微博内容

传统行为预测方法

	(1) 社交关系	(2) 交互频度	(3) 采纳次数	(4) 微博内容
内容过滤 协同过滤	✗	✗	√	√
社交信用 社交影响	✗	√	√	✗
行为矩阵 低秩分解	√	✗	√	√

解决思路：社交上下文关联性因素

- 微博内容是否符合**兴趣偏好**？
- 发来微博的用户是否有**影响力**？

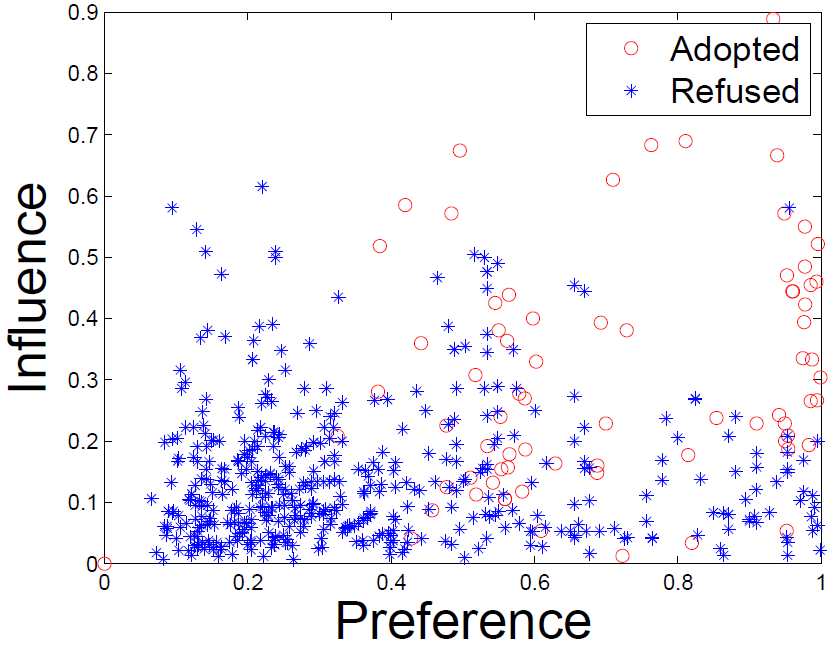
(0) 关注

(1) 发布

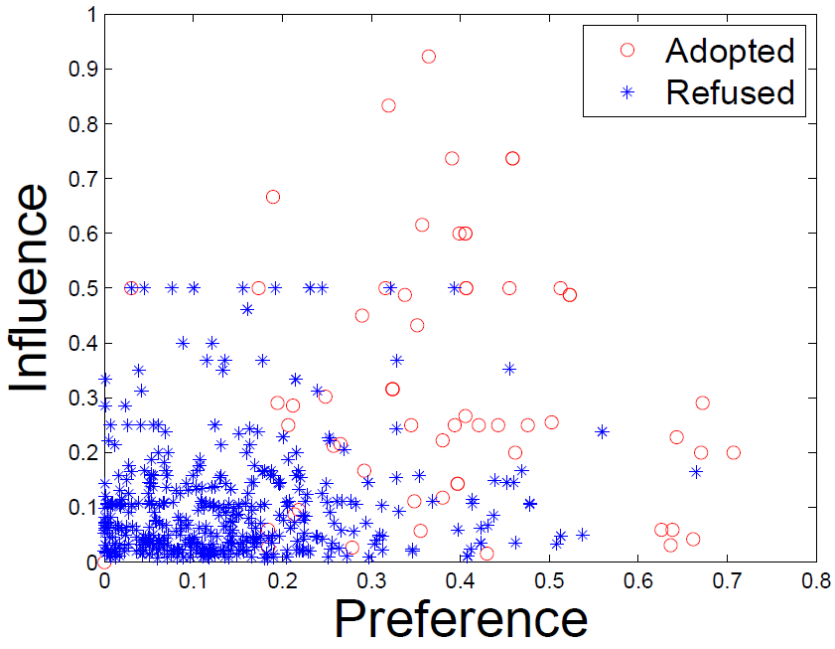
(2) 接收并阅读

(3) 是否转发？

解决思路：社交上下文关联性因素



人人网



腾讯微博

基于社交上下文行为模型ContextMF

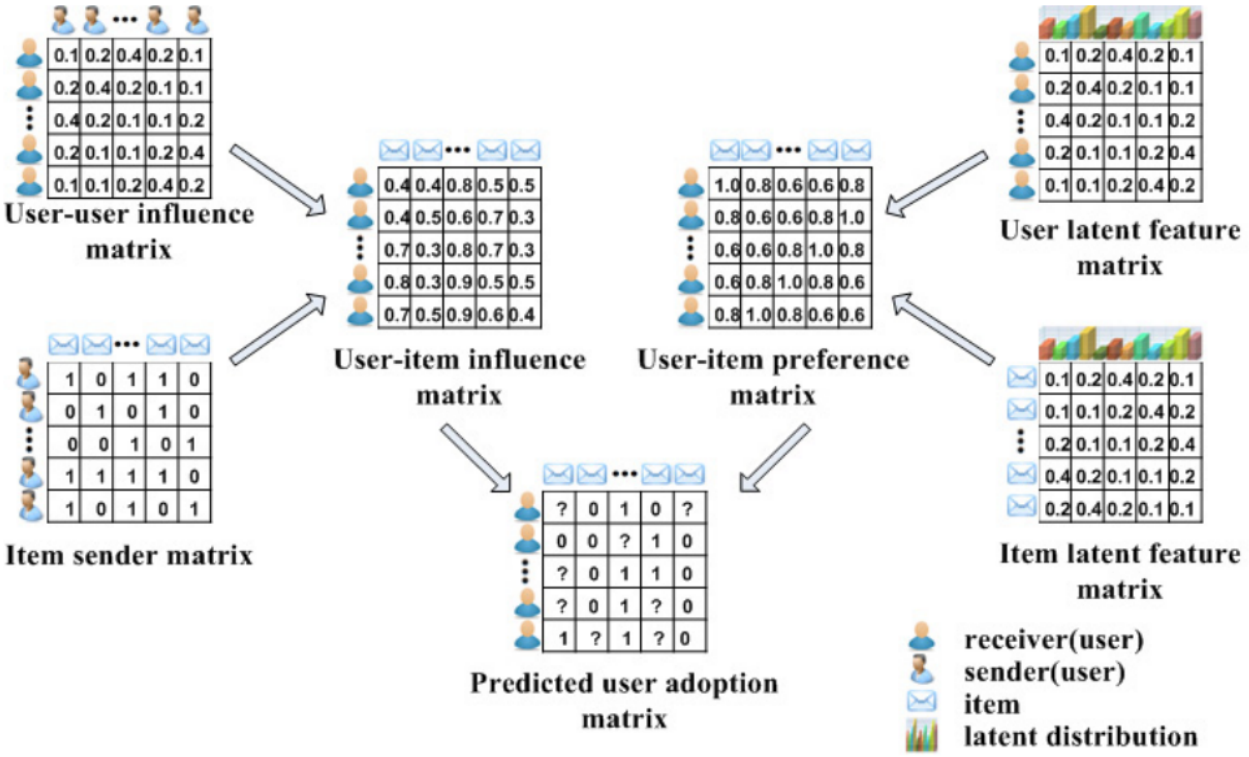
$$P(\mathbf{R}|\mathbf{S},\mathbf{U},\mathbf{V},\sigma_R^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}\mathcal{N}(\mathbf{R}_{ij}|\mathbf{S}_i\mathbf{G}_j^{\top}\odot\mathbf{U}_i^{\top}\mathbf{V}_j,\sigma_R^2)$$

用户-发送者影响力 ***S***

发送者 ***G***

用户特征向量 ***U***

微博特征向量 ***V***



行为观测矩阵 ***R***

ContextMF模型算法

- 最小化sum-of-squared errors function

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

- 梯度下降框架：Block coordinate descent scheme

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

性能评测：采纳信息行为预测

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**

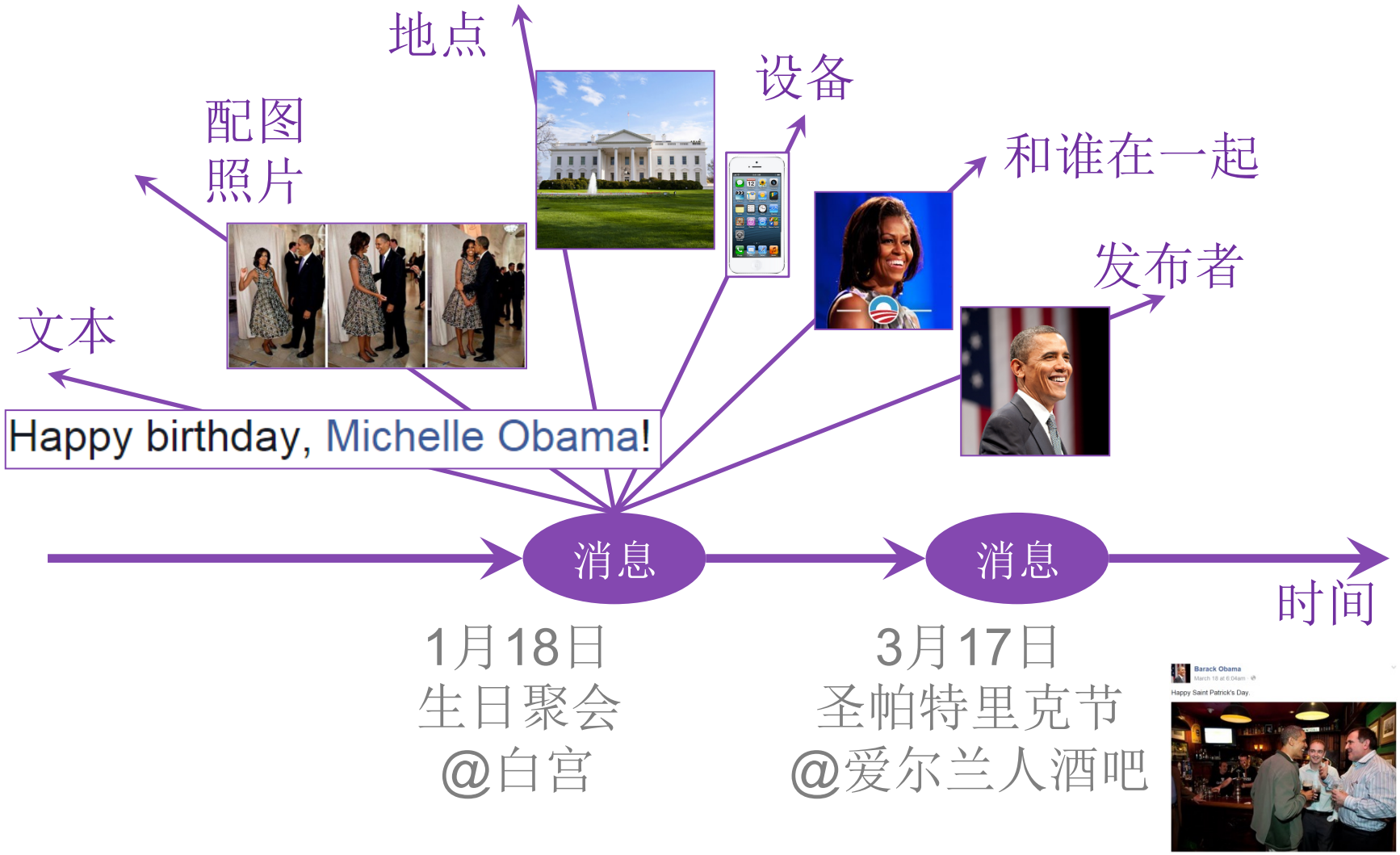
	人人网	腾讯微博
MAE	-19.1%	-24.2%
RMSE	-12.8%	-20.7%
Kendall's	+9.82%	+2.1%
Spearman's	+10.6%	+3.1%

小结

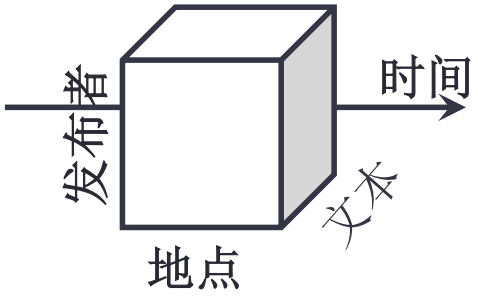
- 分析用户采纳信息行为的社交上下文关联特性
- 提出基于社交上下文的行为预测模型ContextMF
- 提升真实社交媒体中用户行为预测效果

- 论文发表
  - ACM CIKM 2012 (长文, 接收率13.8%)
  - IEEE TKDE 2014 (长文)
  - 引用次数：**85**

时空上下文：多面性和动态性



基于时空上下文的行为建模问题



用户行为

空间 多面性

时间 动态性

目标

行为建模

模式发现

行为预测

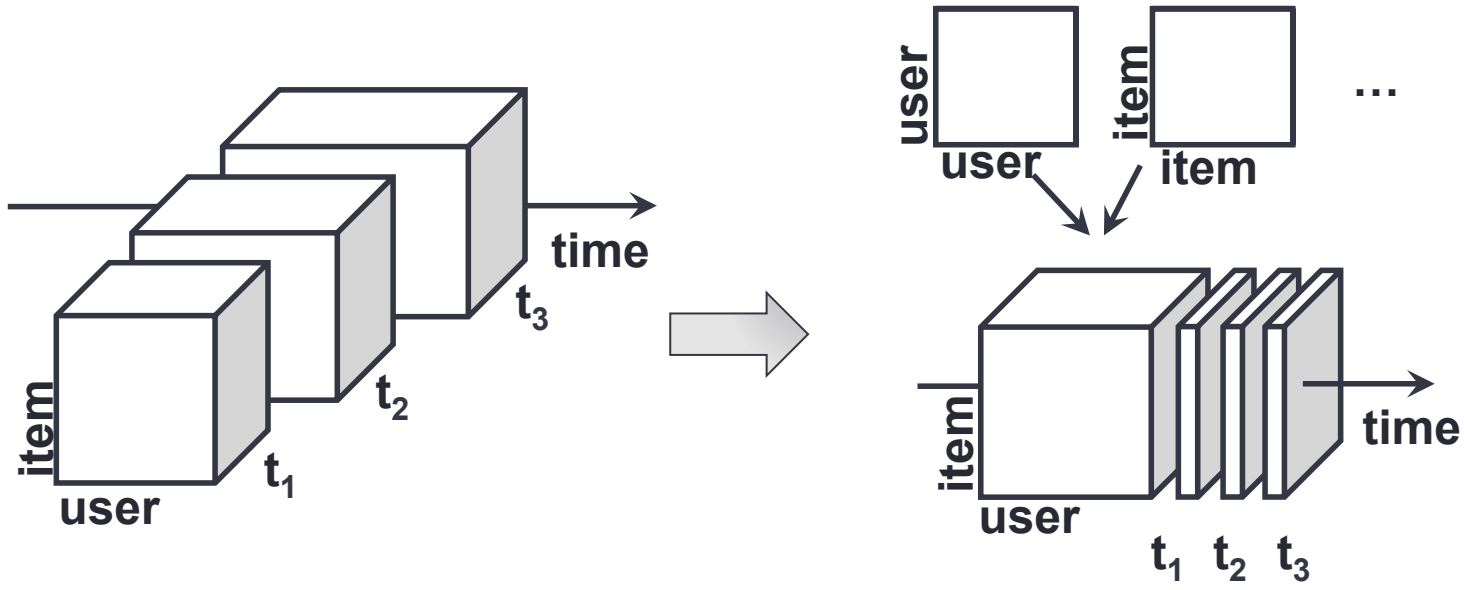
张量序列

思路

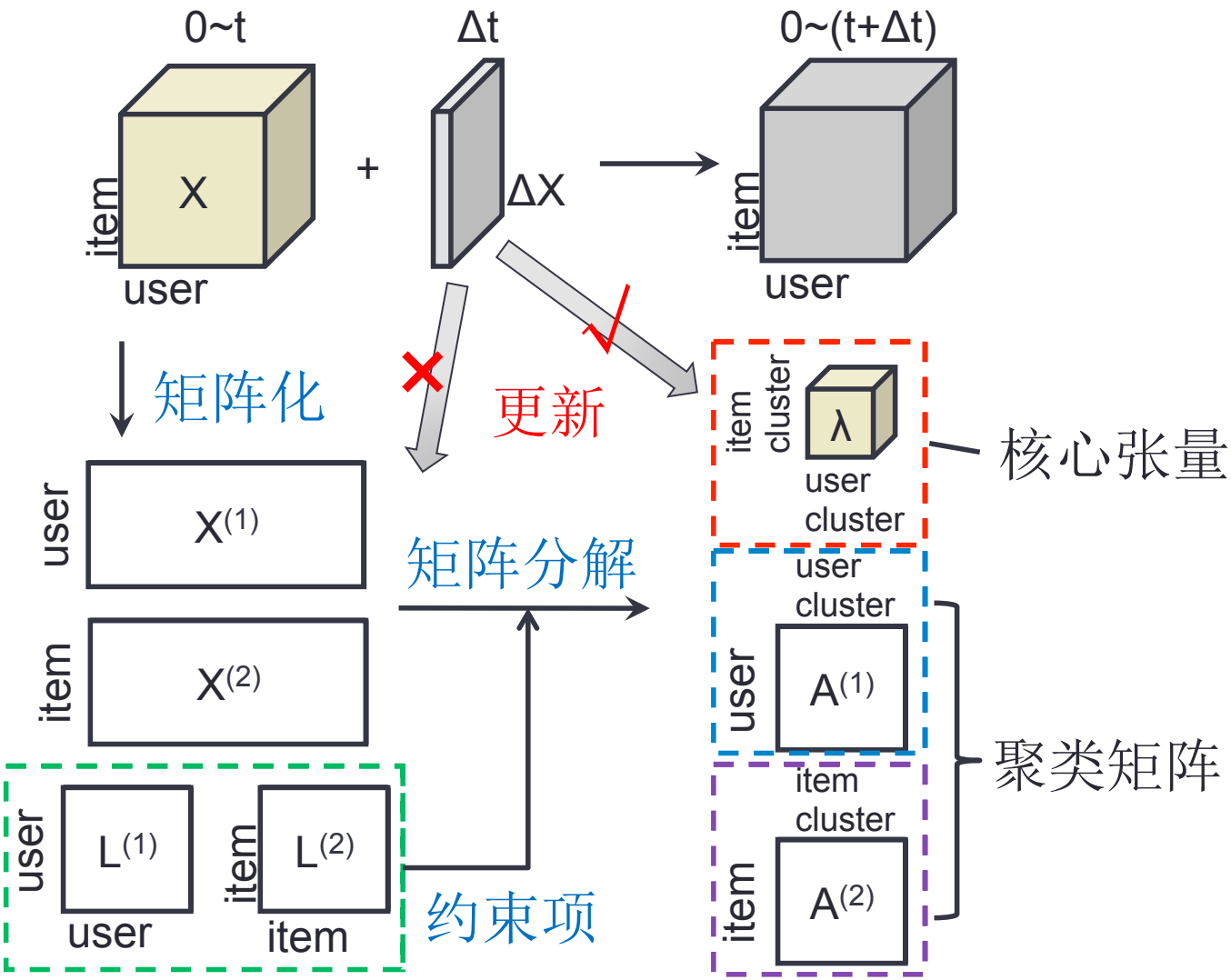
分解与填充

时空上下文建模的难点和解决方案

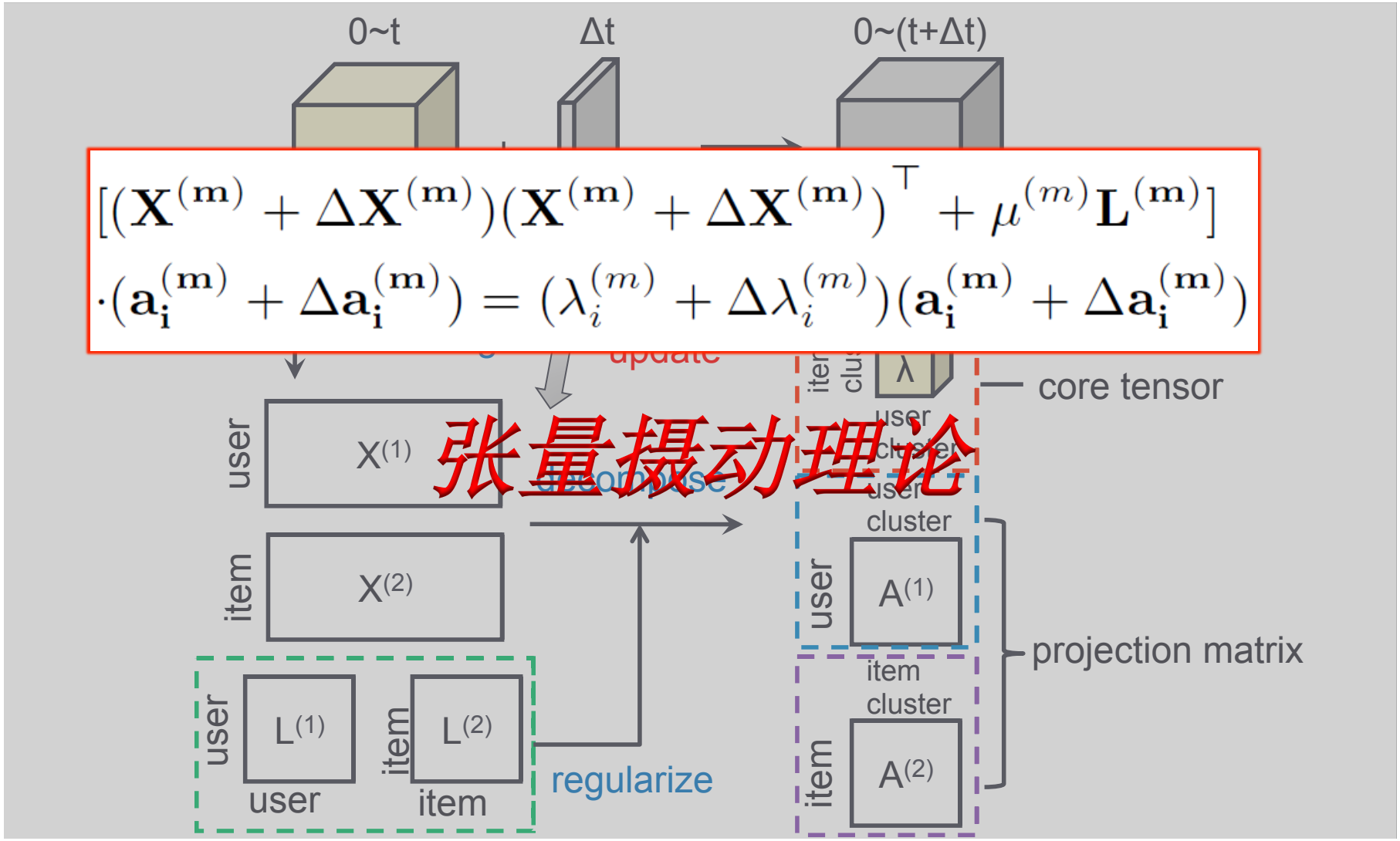
- 技术难点
  - 高稀疏度
  - 高复杂性
- 解决方案
  - 富足的辅助知识作为约束
  - 用增量数据更新聚类矩阵



基于张量摄动理论的多面进化分析方法FEMA



基于张量摄动理论的多面进化分析方法FEMA



FEMA算法

近似算法

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

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

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

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

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

and compute

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

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

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

and compute

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

end for

end for

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

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

收敛域的临界证明

核张量

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

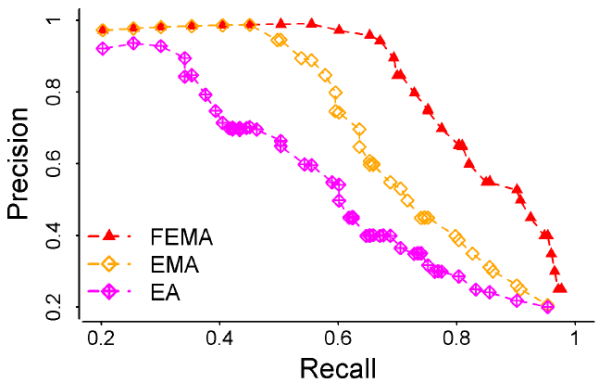
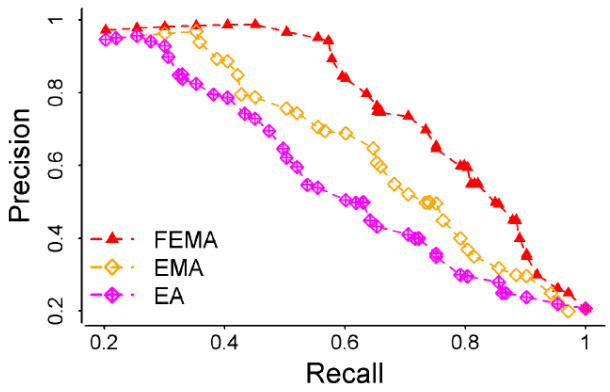
映射矩阵

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

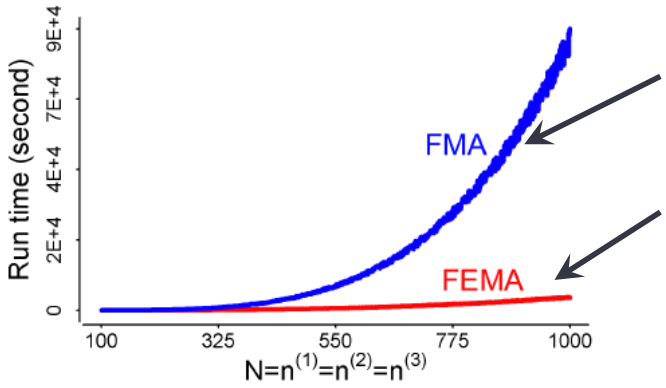
性能评测：预测学术行为和微博提及行为

预测 研究者-关键字
FEMA 利用 研究机构

预测 源用户@目标用户
FEMA 利用 微博内容

	微软学术搜索数据集		腾讯微博提及行为数据集	
	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
准确率 VS 召回率				

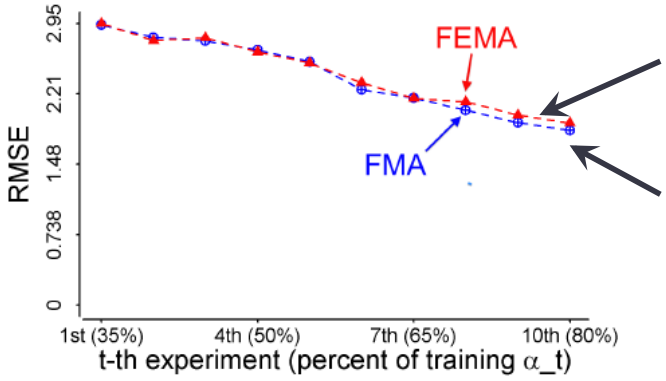
性能评测：效率高、损失小



重新分解
直接重计算矩阵

动态分析:
用 ΔX 更新 λ 和 a

Time vs Num. objects N



动态分析:
用 ΔX 更新 λ 和 a

重新分解
直接重计算矩阵

The loss is small.

小结

- 分析基于时空上下文的行为建模问题：多面、动态
- 基于张量摄动理论的多面进化分析方法FEMA
- 提升在两个真实数据集上的预测效果和效率

- 论文发表
  - ACM SIGKDD 2014 (长文, 接收率14.6%)

研究路线

上下文关联行为的分析与建模

基于社交上下文的采纳信息行为模型

基于时空上下文的多面进化分析方法

跨域跨平台行为的分析与建模

跨域社交媒体的混合随机漫步算法

跨社交平台的半监督迁移学习算法

可疑行为的分析与建模

可疑社交用户同步行为的检测方法

可疑多面行为异常程度的衡量指标

跨域行为建模问题和难点

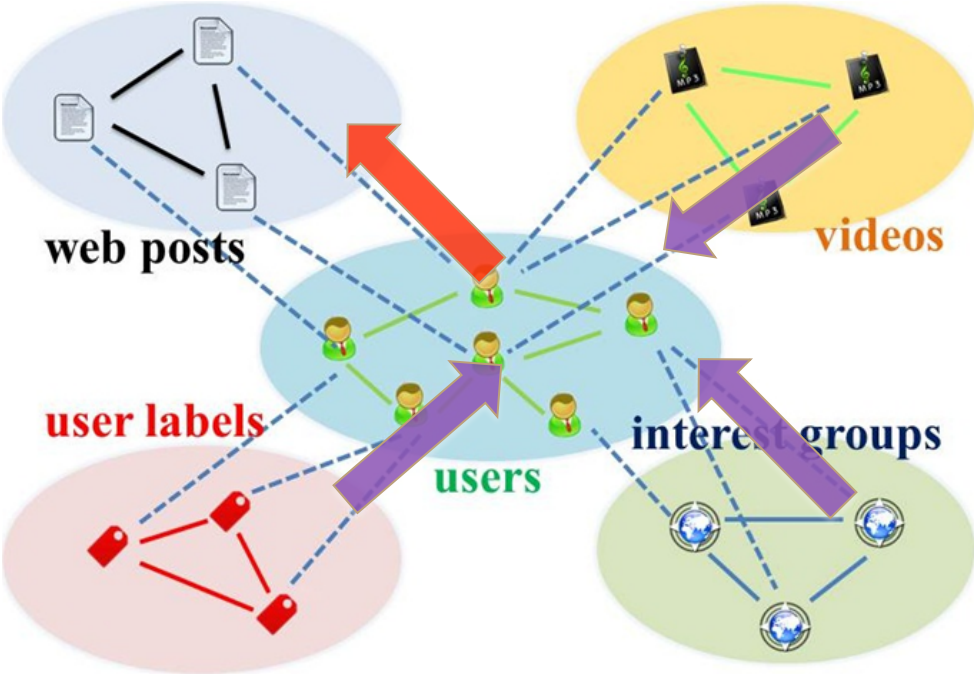
- 腾讯微博数据

域	规模	跨域链接（用户-信息 行为）	
		采纳（+）	拒绝（-）
用户	53.4K	—	—
微博	142K	1.47M (0.02%)	3.40M (0.04%)
社交标签	111	330K (5.57%)	—

- 多域社交媒体：微博、标签、视频、群组等
- **难点1：** 单一域（目标域）的**高稀疏度、冷启动**
- **难点2：** 多域社交媒体的**多元异构性**

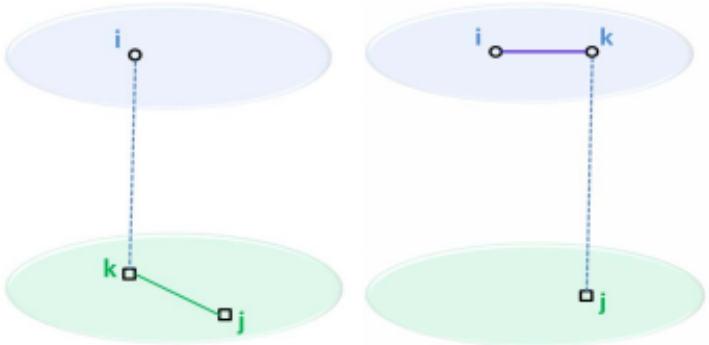
解决方案：社交域是纽带

- 重构社交媒体：社交域为中心的星状图
- 迁移学习
  - 辅助域优化社交域学习，从而优化目标域预测



混合随机漫步HybridRW

- 更新跨域链接权重



(a) Route 1 (b) Route 2

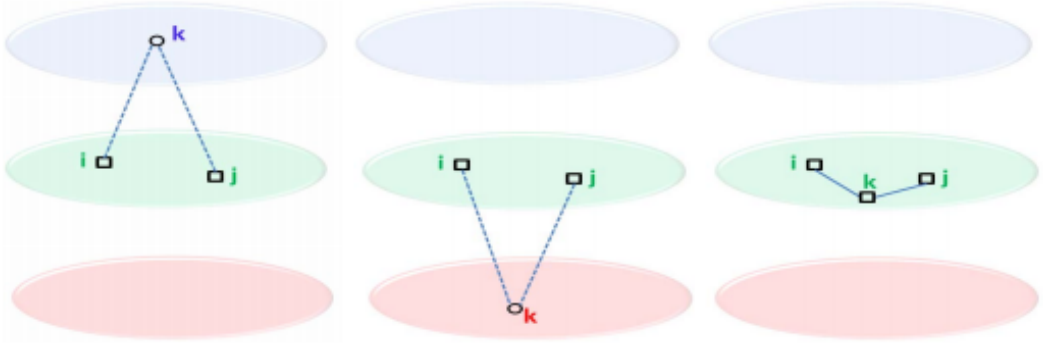
$$p_{ij}^{(\mathcal{UP})^+} = \delta \sum_{u_k \in \mathcal{U}} r_{ik}^{(\mathcal{U})} p_{kj}^{(\mathcal{UP})^+} + (1-\delta) \sum_{p_k \in \mathcal{P}} p_{ik}^{(\mathcal{UP})^+} r_{kj}^{(\mathcal{P})}$$

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

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

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

- 更新域内链接权重



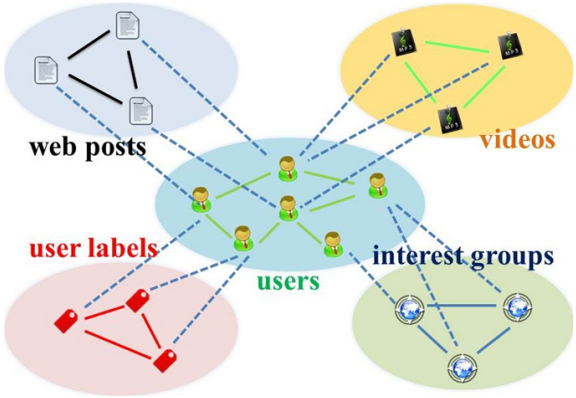
(a) Route 1 (b) Route 2 (c) Route 3

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

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

混合随机漫步HybridRW

- 高阶星状的多元异构图中随机漫步



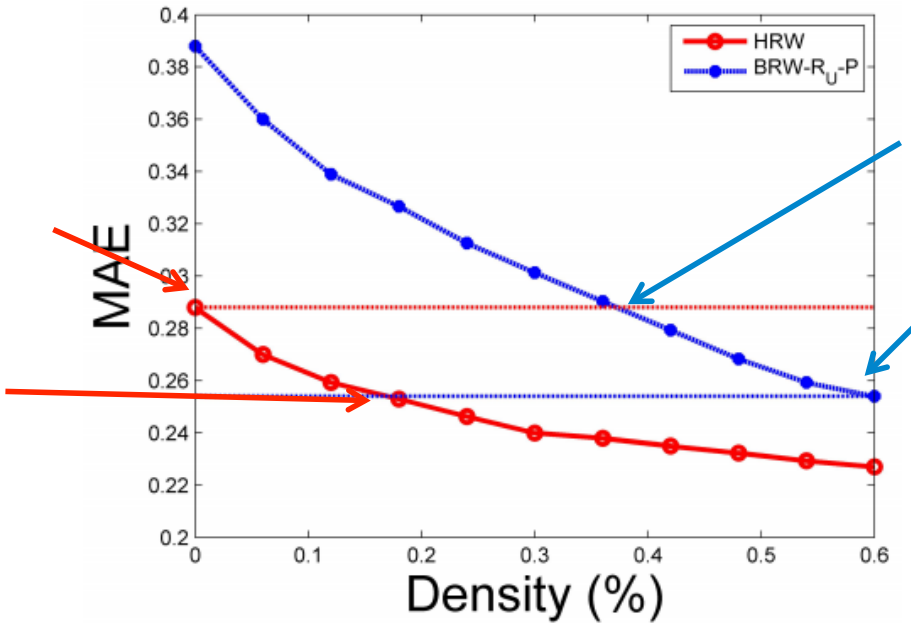
$$
\begin{aligned}
\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t+1) &= \delta_i \mathbf{R}^{(\mathcal{U})}(t)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t) + (1-\delta_i)\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^+}(t)\mathbf{R}^{(\mathcal{D}_i)} \\
\mathbf{P}^{(\mathcal{U}\mathcal{D}_i)^-}(t+1) &= \delta_i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

性能评测：冷启动用户行为预测

- 辅助域能提升冷启动用户的行为预测效果
  - 从标签域行为迁移学习能够节省>60%的微博域行为数据

0 微博域
100% 标签域

18% 微博域
100% 标签域



35% 微博域
0 标签域

60% 微博域
0 标签域

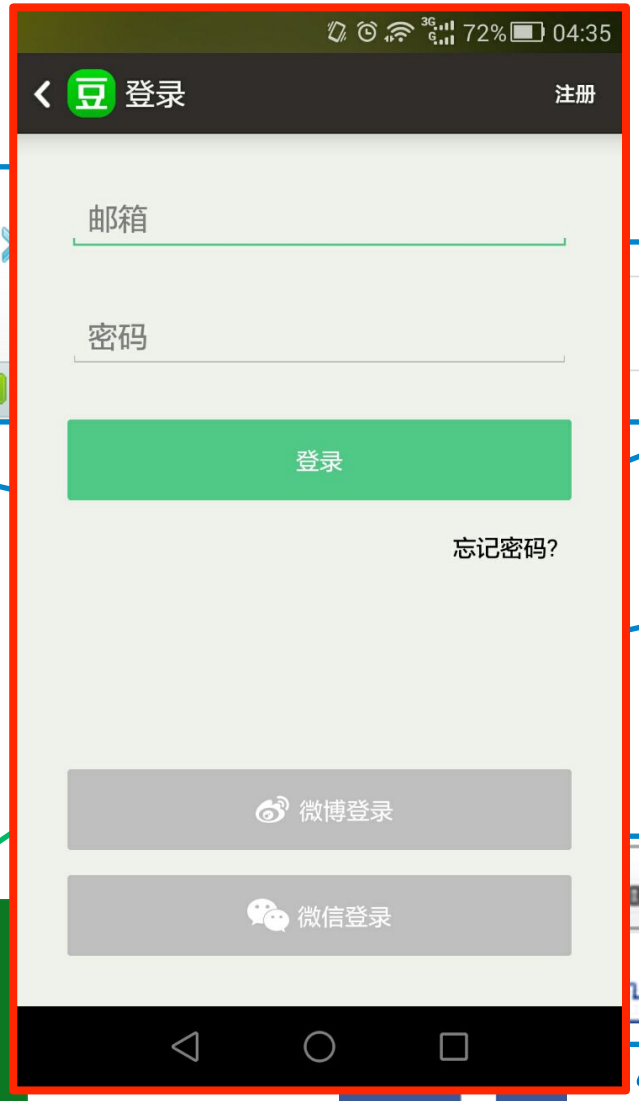
小结

- 分析跨域行为建模问题：以社交域关系为纽带
- 提出基于迁移学习的混合随机漫步算法HybridRW
- 提升目标域的行为预测效果，给出冷启动用户行为预测的有效解决方案

- 论文发表
  - ACM CIKM 2012 (长文, 接收率13.8%)
  - IEEE TKDE 2015 (已录用, 长文)
  - 引用次数： **32**

跨平台行为建模问题

选择社交标签



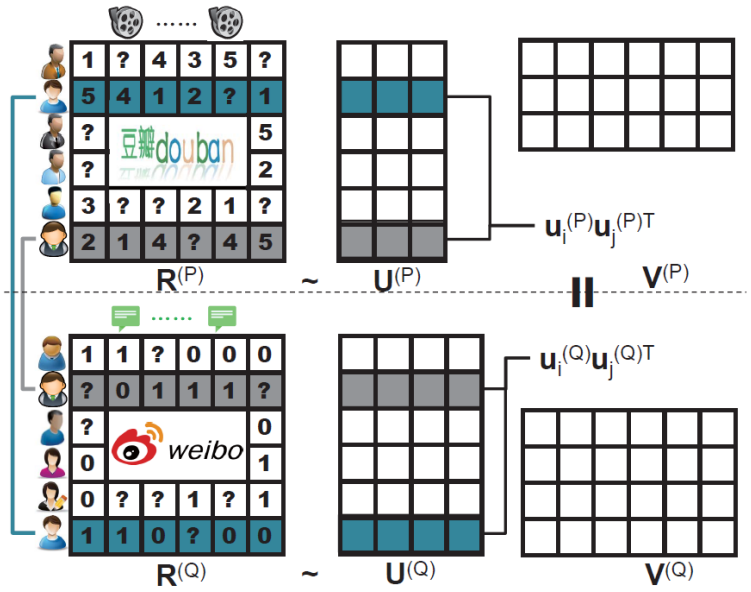
发布和转发微博

视频，音乐…

电影评分

“喜欢”页面、产品

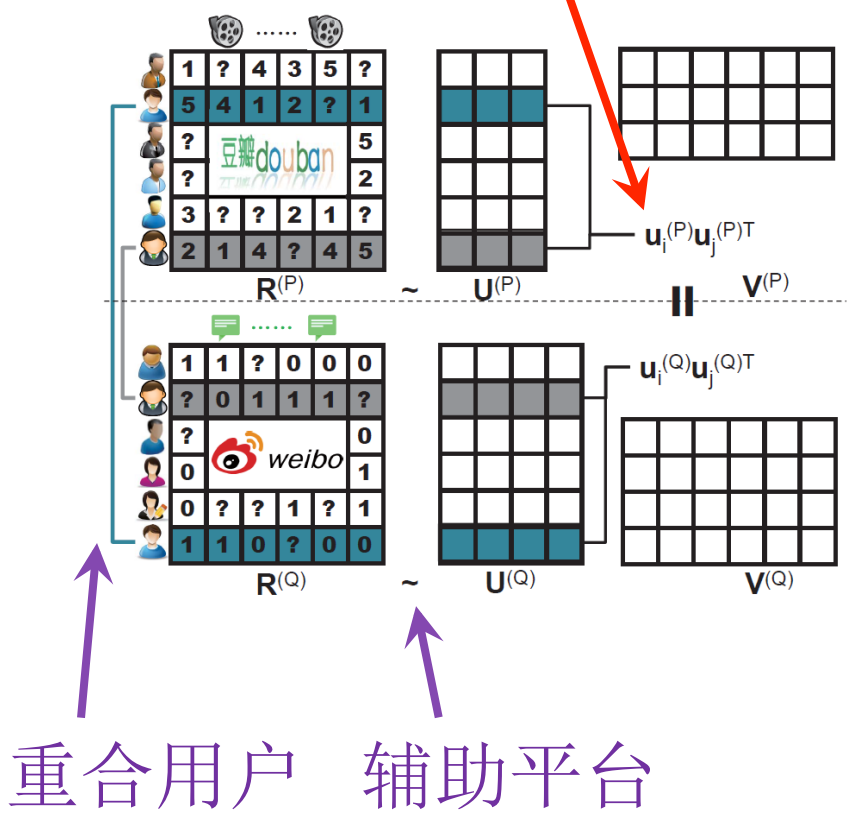
跨平台行为建模的技术难点



- 问题目标
  - 目标平台高稀疏度
- 解决思路
  - 辅助平台的用户行为
- 技术难点
  - 平台的多元异构性
  - 不同平台用户表征不同

跨平台行为建模的解决方案

重合用户相似度
约束跨平台用户表征



- 问题目标
  - 目标平台高稀疏度
- 解决思路
  - 辅助平台的用户行为
- 技术难点
  - 平台的多元异构性
  - 不同平台用户表征不同
- 解决方案
  - 平台间用户的部分重合

半监督迁移学习算法 (XPTrans)

- **输入**
  - 目标平台P和辅助平台Q;
  - 用户-信息行为矩阵$R^{(P)}$和$R^{(Q)}$;
  - 观测数据二值矩阵$W^{(P)}$和$W^{(Q)}$;
  - 重合用户的指示矩阵$W^{(P,Q)}$,
- **输出**
  - 用户聚类矩阵$U^{(P)}$和$U^{(Q)}$;
  - 信息聚类矩阵$V^{(P)}$和$V^{(Q)}$;
  - $R^{(P)}$中的缺失值

- **目标函数**

非监督项

目标平台 辅助平台

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

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

重合用户相似度

监督项

性能评测：评分预测

- 基线算法
  - CMF：不利用辅助平台行为信息
  - CBT：用“行为模式编码”作为平台间联系，也就不利用重合用户
  - XPTrans-Align：不同平台的重合用户采用统一的特征表示
  - **XPTrans**：不同平台的重合用户采用相似度匹配约束表征
- 跨平台迁移学习XPTrans预测最准确

	Q: Weibo tweet entity → P: Douban movie				Q: Douban book → 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%

小结

- 分析跨平台行为建模问题：以重合用户为纽带
- 提出半监督迁移学习算法XPTrans
- 跨平台迁移学习能够大幅度提升目标平台的行为预测效果

- 论文投稿中

研究路线

上下文关联行为的分析与建模

基于社交上下文的采纳信息行为模型

基于时空上下文的多面进化分析方法

跨域跨平台行为的分析与建模

跨域社交媒体的混合随机漫步算法

跨社交平台的半监督迁移学习算法

可疑行为的分析与建模

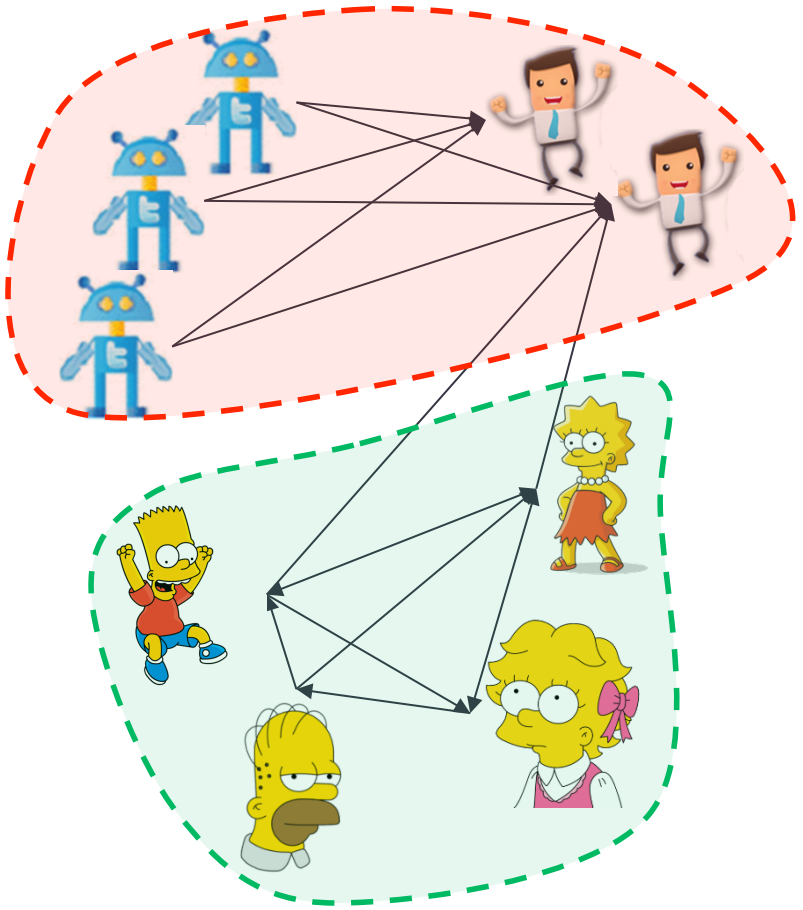
可疑社交用户同步行为的检测方法

可疑多面行为异常程度的衡量指标

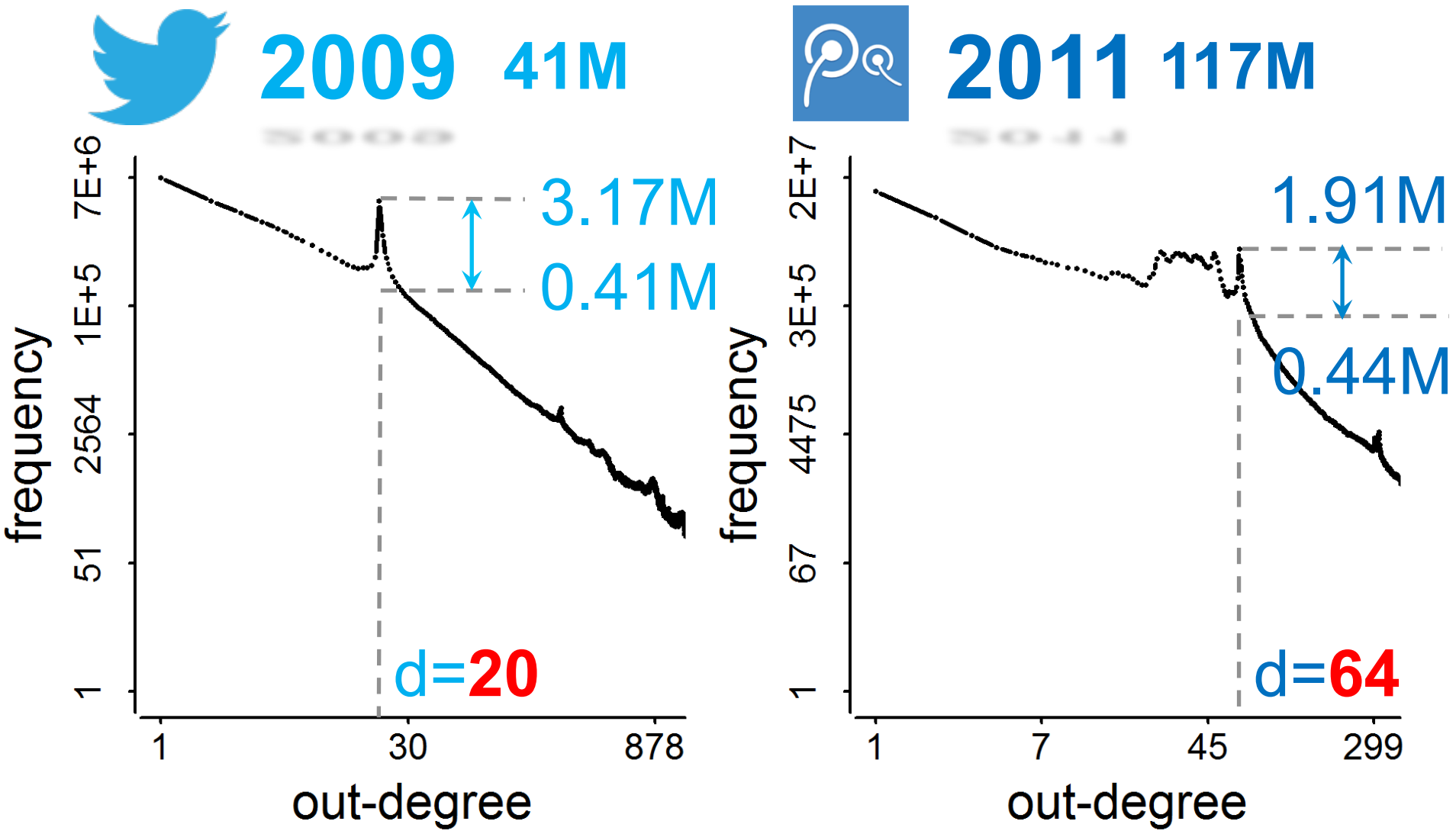
用户行为真伪性：检测僵尸粉问题

[www.buyfollowz.org]

[buymorelikes.com]



技术难点：出度分布的异常尖峰



技术难点：传统检测方法的局限性

标签 (+1,-1)	出度	入度	微博数	微博中的url数量	微博hashtag数量

基于内容的特征

分类器

不足：受限于缺乏表象信息

解决方案：行为规律是关键

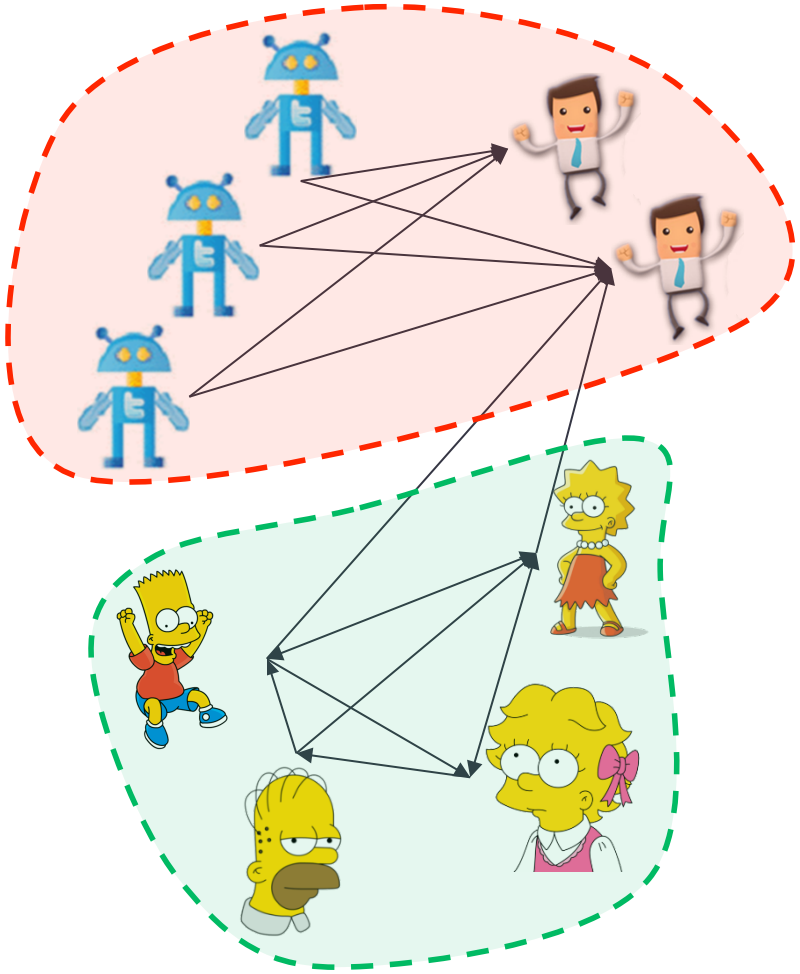
金钱利益驱动

内容

僵尸粉
可能会
表现得
内容可疑

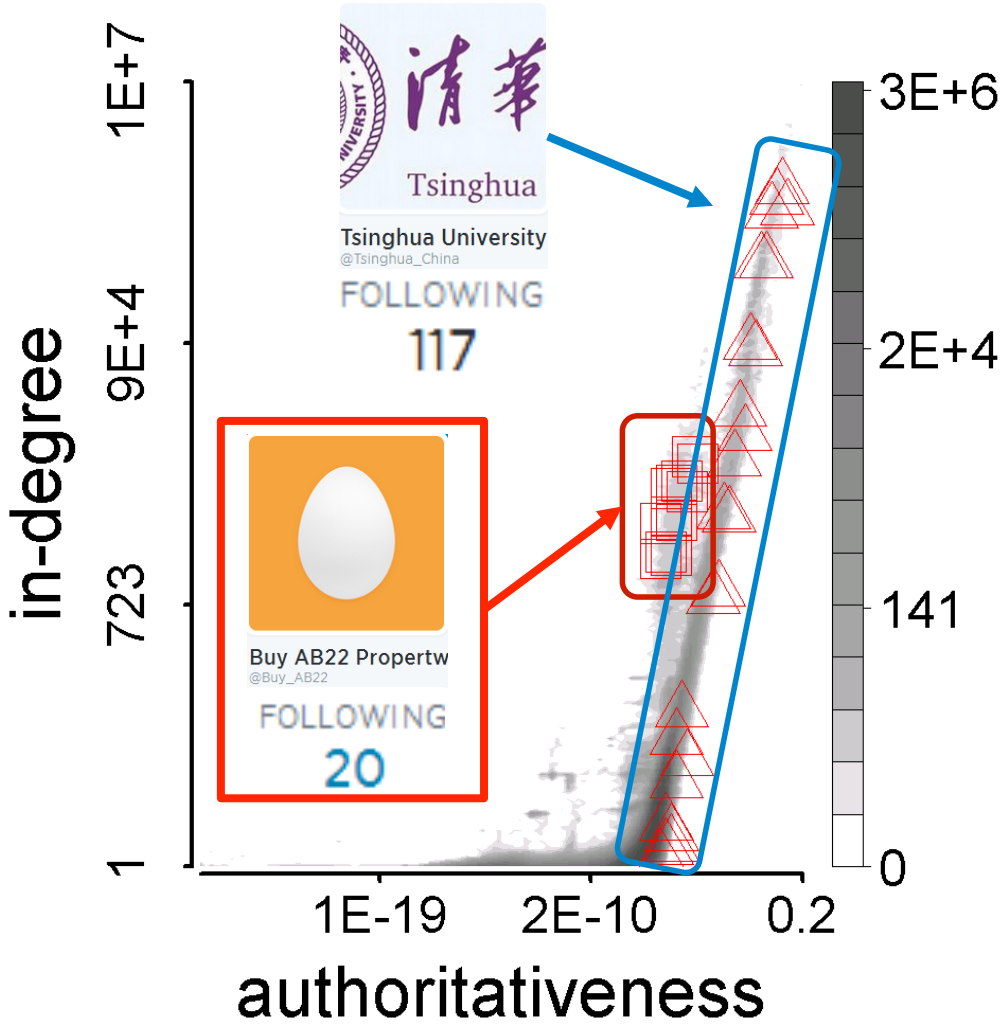
行为

僵尸粉
不得不
做出
可疑行为



僵尸粉的行为规律：所关注是谁？

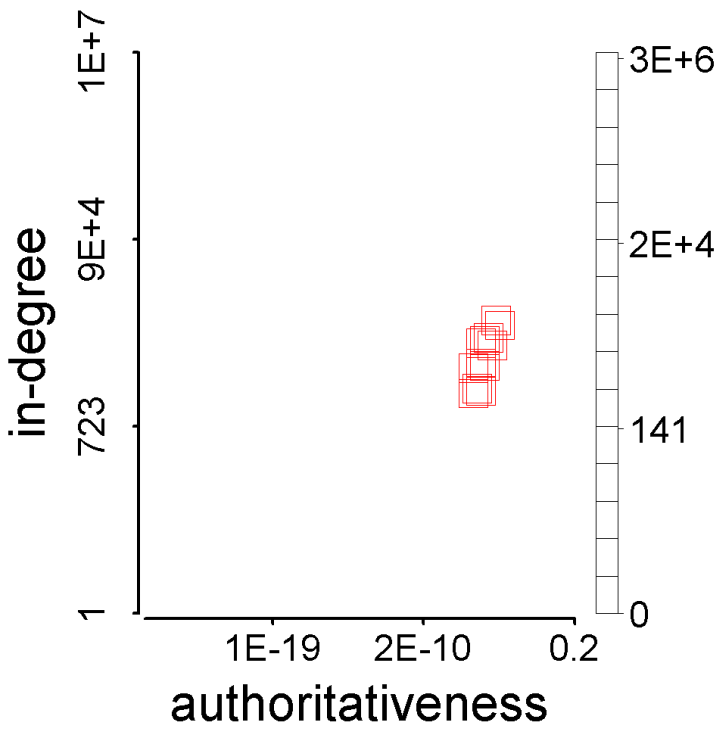
- 同步性
- 异常性

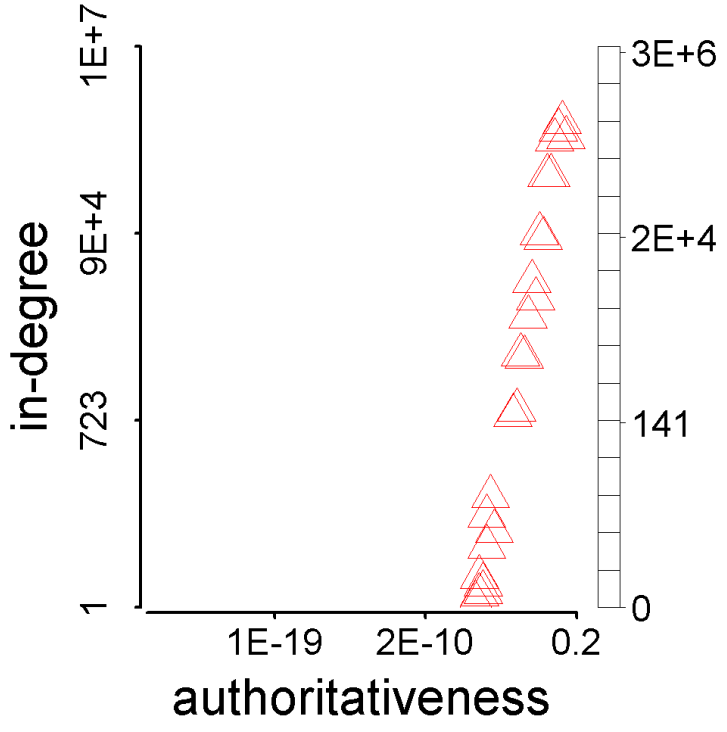


同步程度 和 正常程度

- 同步程度

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

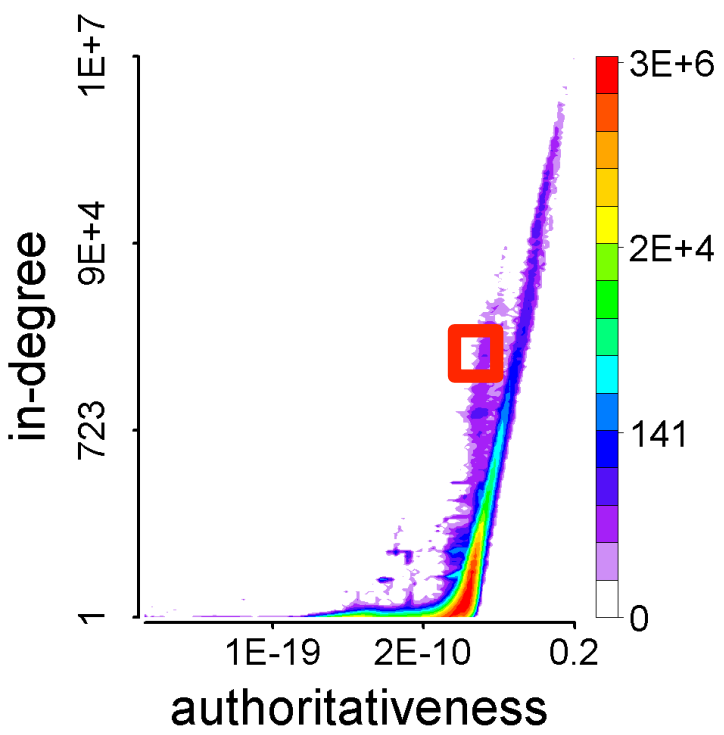


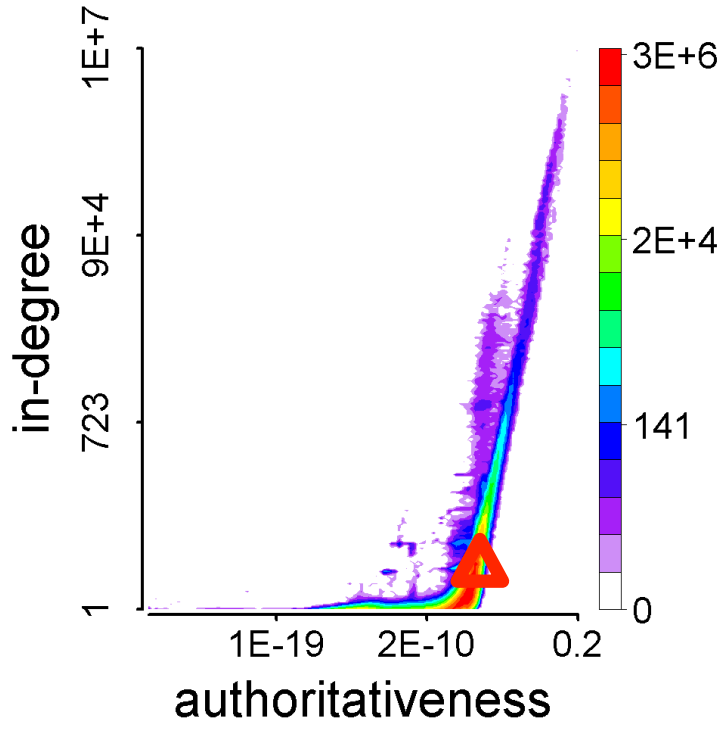


同步程度 和 正常程度

- 正常程度

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





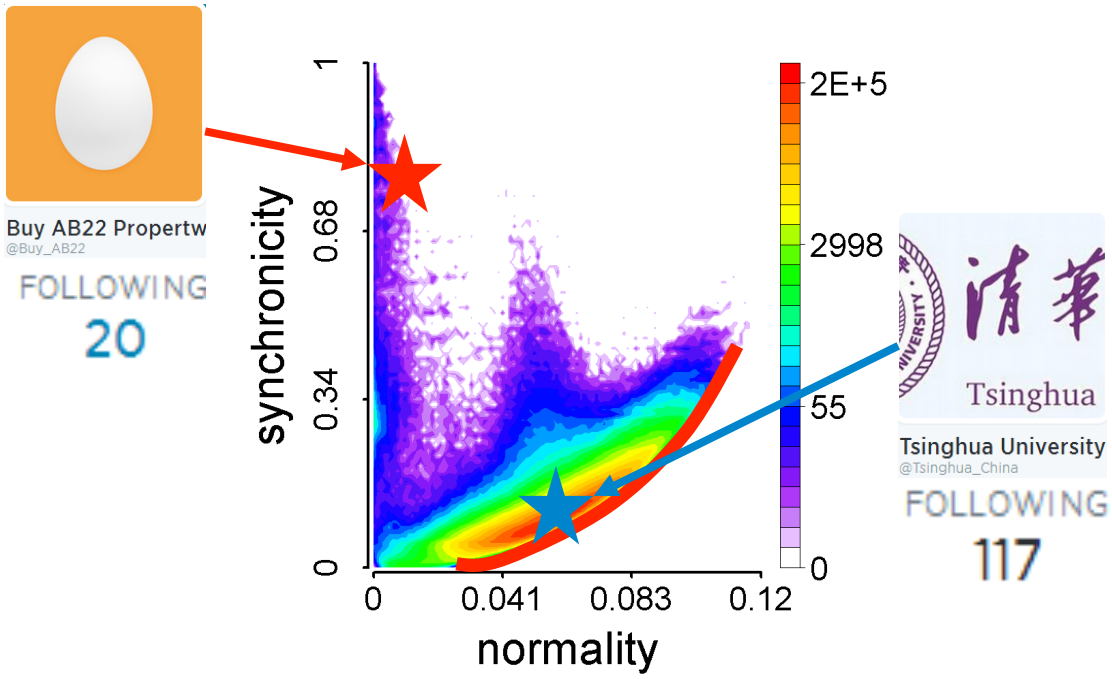
定理：给定正常程度时同步程度的下限

- 对于任意的分布，在“同步-正常程度图”上存在**二次曲线的下限**

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

同步程度

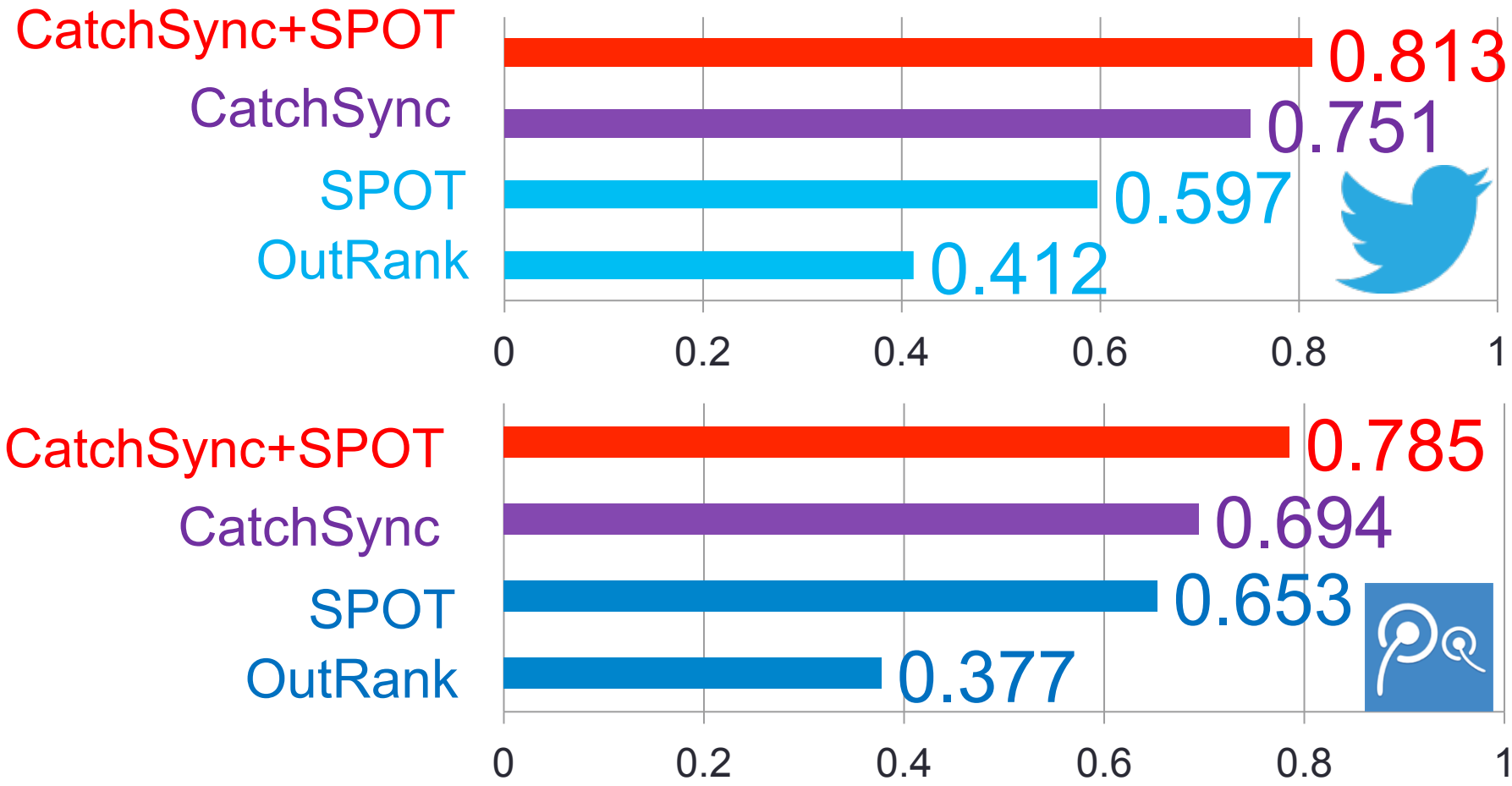
正常程度



CatchSync:
基于距离的异常检测

性能评测：检测人工标记的采样用户

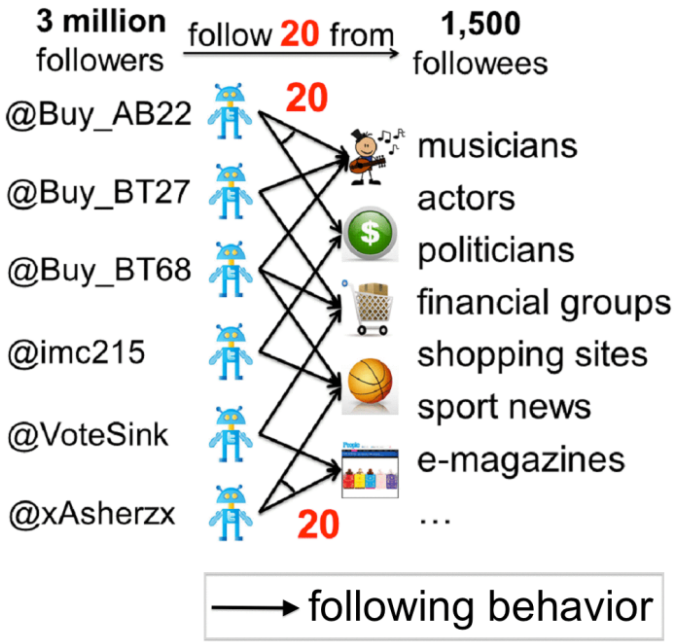
- 准确度提升：行为方法CatchSync与内容方法SPOT互补

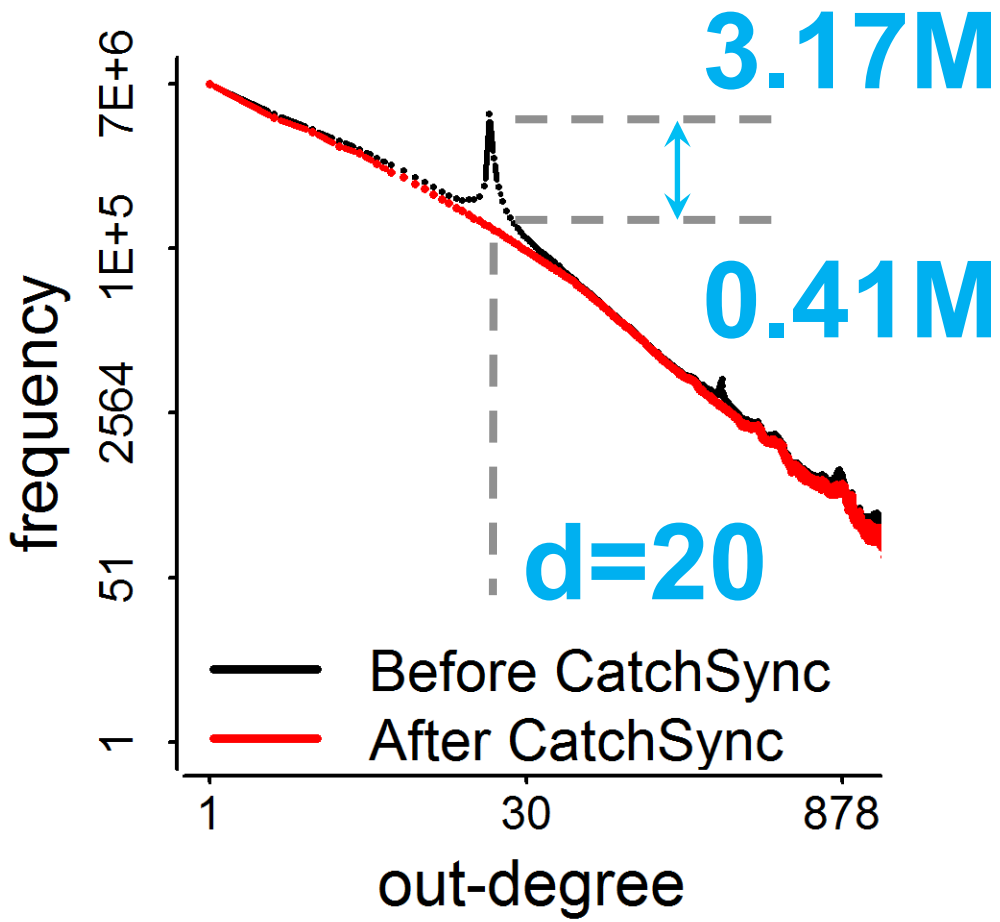


性能评测：还原被扭曲的出度分布

2009

41M





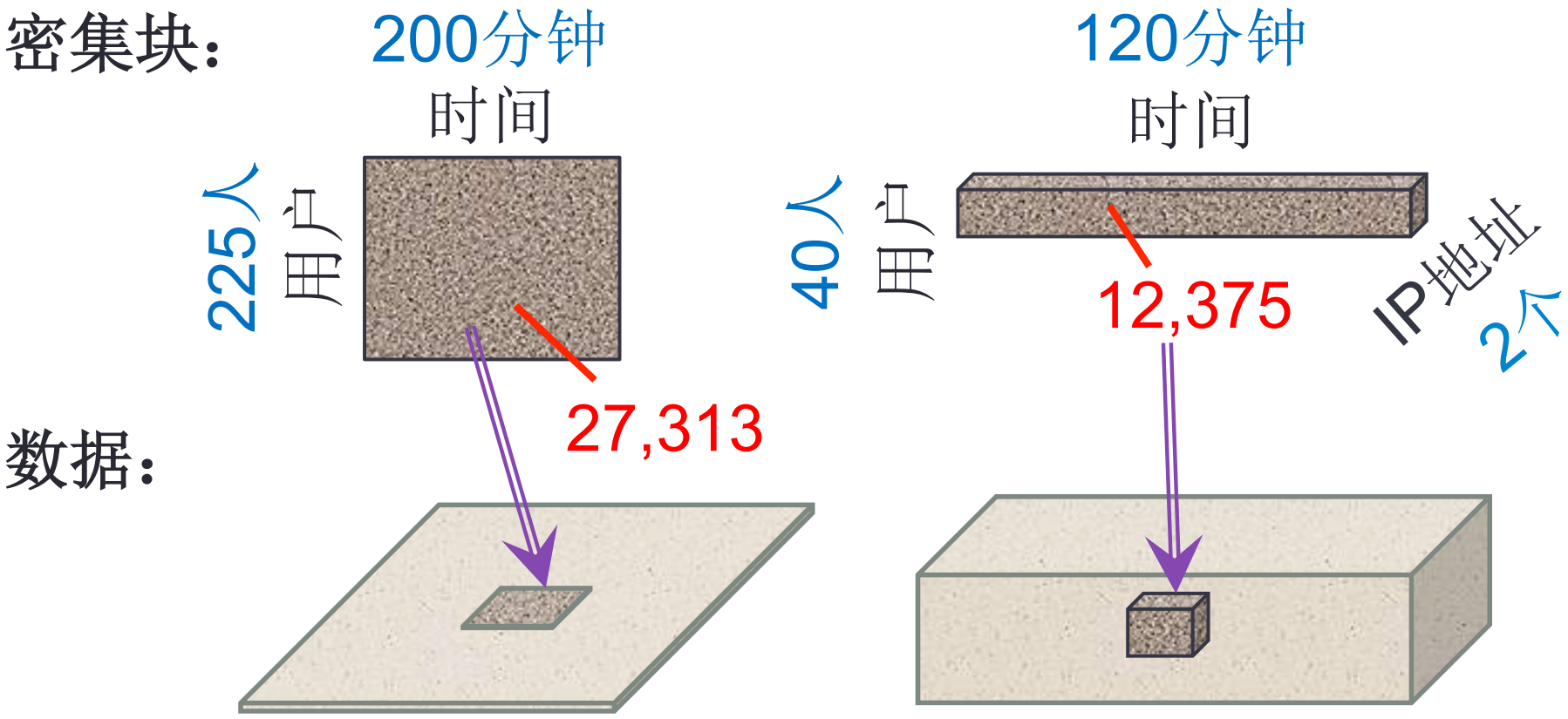
小结

- 分析僵尸粉行为的同步性和异常性规律
- 提出快速大规模图挖掘算法CatchSync
- 成功还原顺滑、幂律的出度分布

- 论文发表
  - ACM SIGKDD 2014 (长文, 接收率14.6%)
    - 最佳论文最终列表 (大陆地区首次获得该荣誉)
  - ACM TKDD 2015 (已录用, 特邀)
  - 引用次数： **18**

衡量跨维度可疑行为的问题：信息操纵

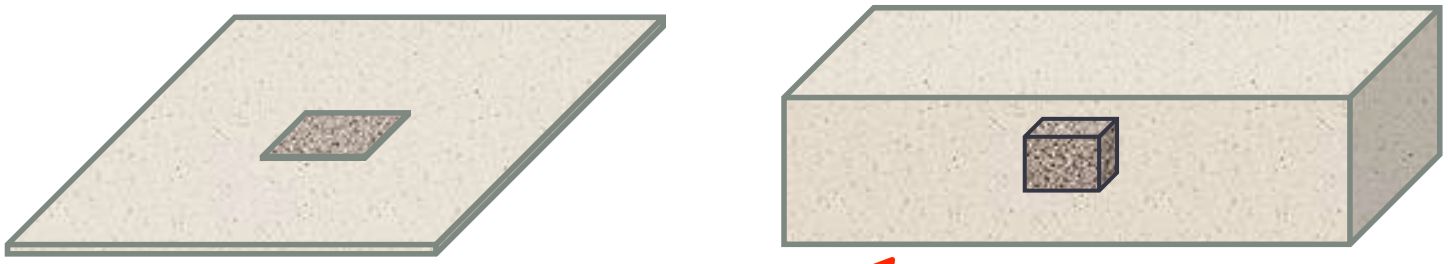
- 一个2维，一个3维，哪一个更可疑？



衡量跨维度可疑行为的解决思路

- 衡量跨维度？

密集块+数据：



发生概率

0.9%

0.05%

更可疑！优先寻找！

解决方案：“可疑程度”指标

- 定义 跨维度的可疑程度 为 数据在*Erdos-Renyi-Poisson*模型下 该块的存在概率的负对数似然估计

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

- 公式
- 局部搜索
- CrossSpot

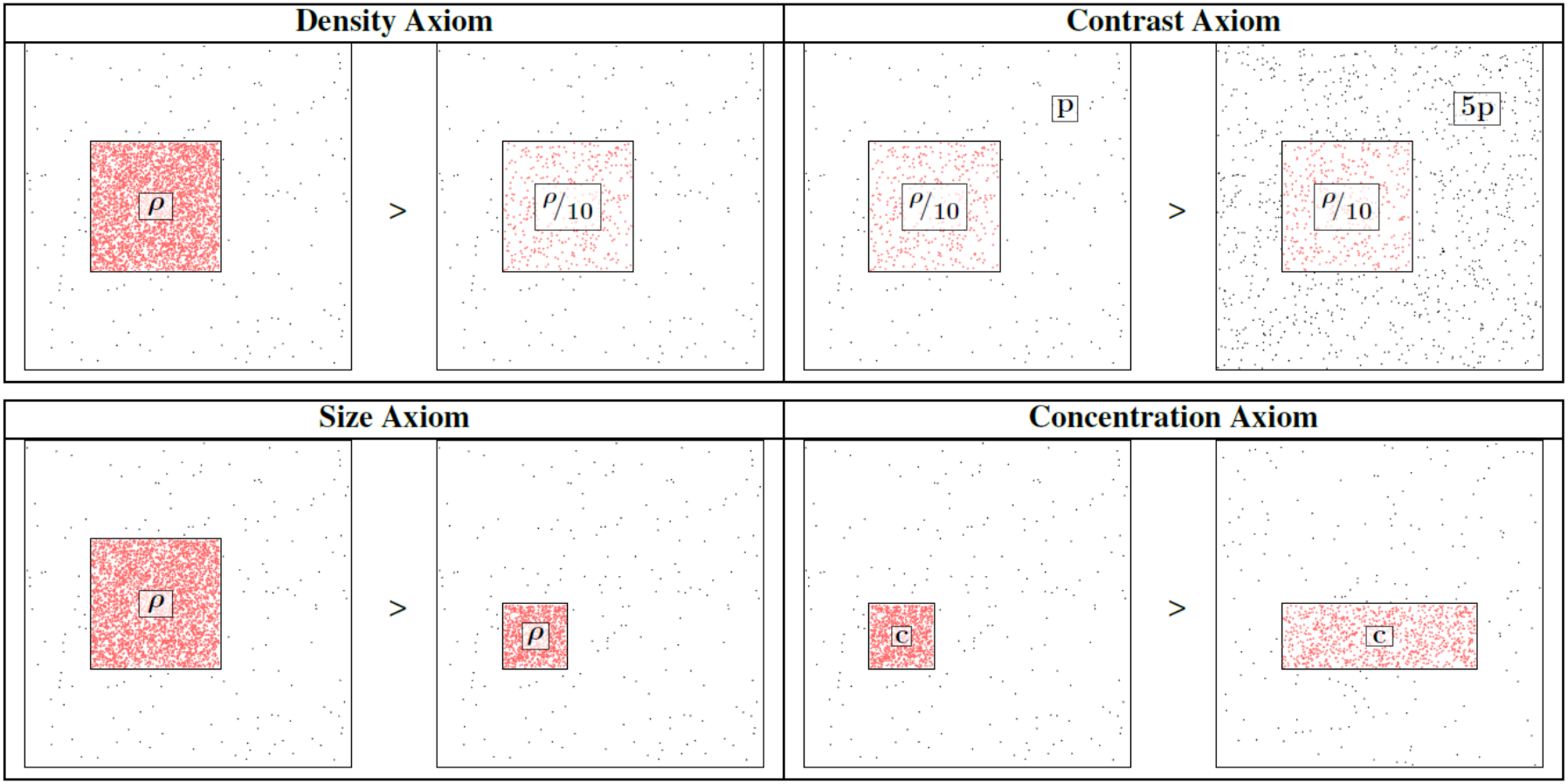
Lemma *Given an $n_1 \times \cdots \times n_K$ block of mass c in $N_1 \times \cdots \times N_K$ data of total mass C, the suspiciousness function is*

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

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

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

验证指标：密度、对比、尺寸、浓度



“可疑程度”和CrossSpot的优势

- 量化密集块状子数据的异常性
- 允许块状子数据维度不同
- 满足一系列公理

	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				

性能评测：检测不同维度的仿真块

- 仿真实验
  - 1,000×1,000×1,000 of 10,000 随机数据
  - 仿真块#1: 30×30×30 of 512　　3维
  - 仿真块#2: 30×30×1,000 of 512　　2维
  - 仿真块#3: 30×1,000×30 of 512　　2维
  - 仿真块#4: 1,000×30×30 of 512　　2维

	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**

三个庞大的真实数据集

数据集	维度				规模
微博转发	用户	原始微博	IP	时间 (分)	转发数
	29.5M	19.8M	27.8M	56.9K	211.7M
话题热度	用户	Hashtag	IP	时间 (分)	微博数
	81.2M	1.6M	47.7M	56.9K	276.9M
网络包 (LBNL)	源IP	目标IP	端口	时间 (秒)	包数
	2,345	2,355	6,055	3,610	230,836

性能评测：操纵微博话题的热度

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!

小结

- 提出用发生概率量化跨维度用户行为的可疑程度
- 快速局域搜索算法CrossSpot

- 论文发表
  - IEEE ICDM 2015 (短文, 接收率18.2%)

总结

社交媒体复杂行为分析、建模与应用算法

上下文关联性

1. 基于社交上下文的采纳信息行为模型
2. 基于时空上下文的多面进化分析方法

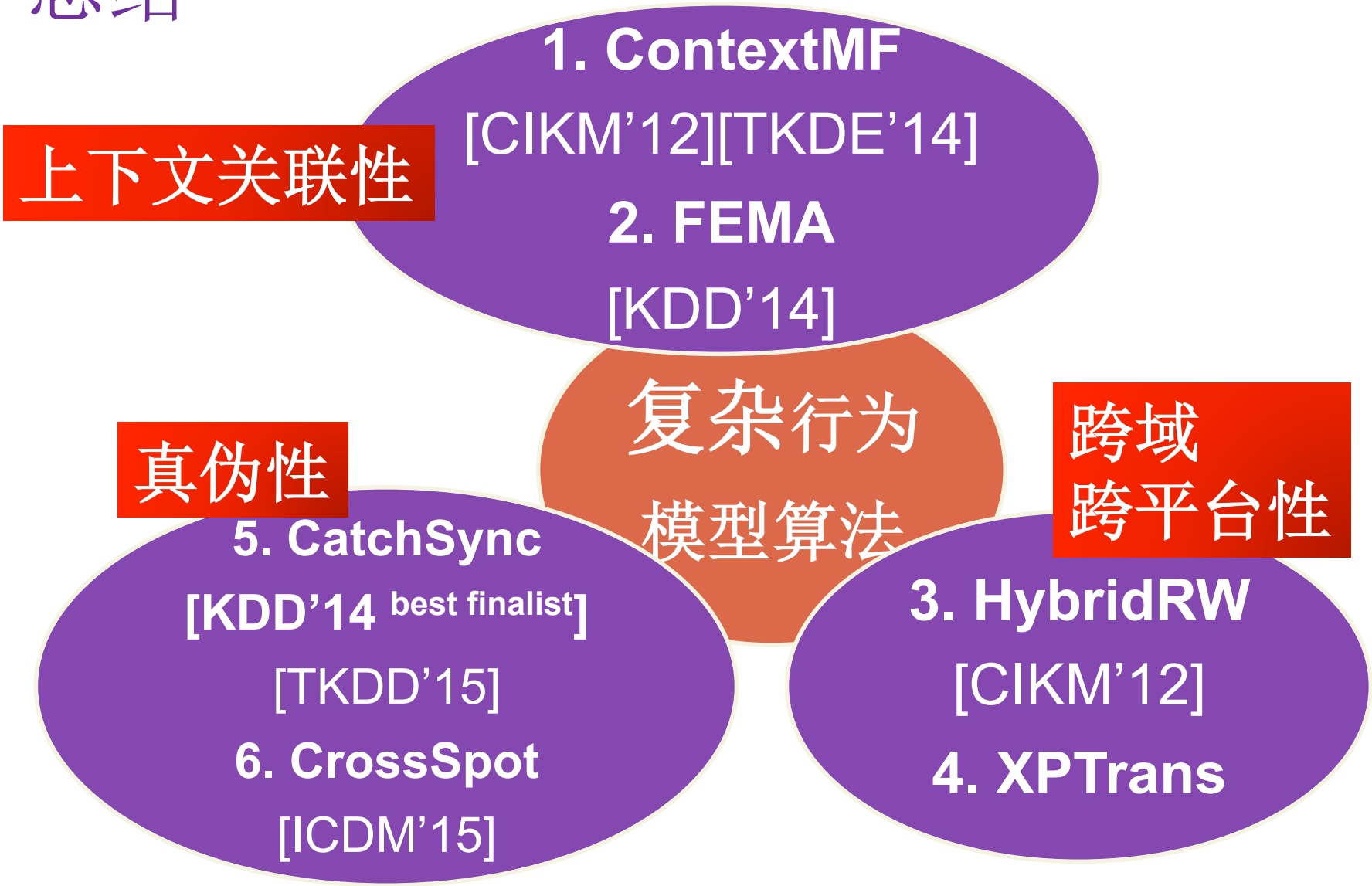
跨域跨平台性

3. 跨域社交媒体的混合随机漫步算法
4. 跨社交平台的半监督迁移学习算法

真伪性

5. 可疑社交用户同步行为的检测方法
6. 跨维度行为可疑程度的衡量指标

总结


上下文关联性
1. ContextMF
[CIKM'12][TKDE'14]
2. FEMA
[KDD'14]
复杂行为
模型算法
跨域
跨平台性
3. HybridRW
[CIKM'12]
4. XPTrans
真伪性
5. CatchSync
[KDD'14 best finalist]
[TKDD'15]
6. CrossSpot
[ICDM'15]

研究成果

- 第一作者论文 **10/12** 篇
  - 3×IEEE/ACM Trans. (3×Regular)：TKDE*2, TKDD
  - 7×Top Conf. (5×Full)： SIGKDD*2, CIKM*2, PAKDD
- 有影响的工作
  - “大规模有向图中的同步行为检测”： SIGKDD’14 **最佳论文最终列表**, 大陆地区首次获得此荣誉
  - “社交上下文推荐”： CIKM’12 和TKDE’14 (引用： **85**)
  - 个人总引用次数： **290**
- **奖励**： 国家奖学金、搜狐研发奖学金

参与项目

- 国家973计划：“网络可视媒体的有效搜索与服务”，2011CB302206
- 自然科学基金重大国际（地区）合作研究项目：“社会化多媒体计算理论与关键技术研究”，61210008
- 自然科学基金面上项目：“基于社交访问行为与传播特性的在线视频内容部署与传输方法研究”，61272231
- 国家科技重大专项：“大型网络应用及服务平台方案设计及示范”，2012ZX01039001-003
- 科技部国家科技合作项目：“基于社交网络中媒体内容的品牌监测合作研究”，2013DFG12870
- 自然科学基金面上项目：“跨域异构媒体信息的社会化推荐关键技术研究”，61370022
- 自然科学基金青年科学基金：“网络信息感知的视频语义分析与检索”，61303075

期刊论文

- **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. 11 citations till 09/2015.)
- 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.

会议论文

- **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 09/2015.)

会议论文（续）

- **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 09/2015.)
- Lu Liu, Jie Tang, Jiawei Han, **Meng Jiang** and Shiqiang Yang. “Mining Topic-Level Influence in Heterogeneous Networks”, in CIKM 2010.

在投论文

- **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).

感谢您来参加!