



清华大学  
Tsinghua University

# 大规模社交网络中的行为预测和异常检测

清华大学计算机系博士生 蒋 朦  
[www.meng-jiang.com](http://www.meng-jiang.com)





# 提纲：社交网络用户行为

- ❖ 为什么值得研究？
- ❖ 具有哪些特点？
- ❖ 带来哪些机遇和挑战？
- ❖ 有什么感悟？





# 提纲：社交网络用户行为

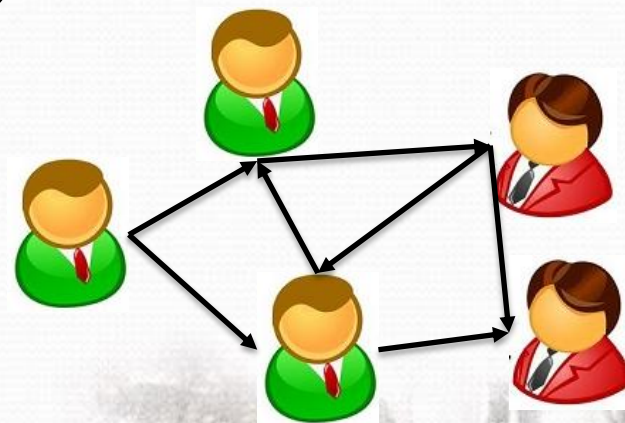
- ❖ 为什么值得研究？
- ❖ 具有哪些特点？
- ❖ 带来哪些机遇和挑战？
- ❖ 有什么感悟？





# 社交网络用户行为 为什么值得研究？

- ❖ “以人为本”、以用户为中心的应用 激发 用户行为
- ❖ 用户行为中潜在的需求 提出 研究问题和新的应用
- ❖ 注册、登录（“用户-应用”）
- ❖ 增加/取消“关注”、好友请求（“用户-用户”）

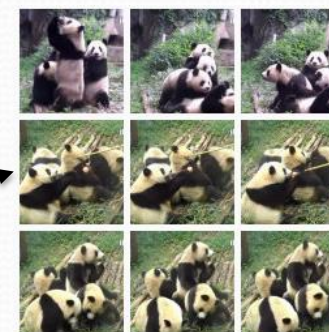
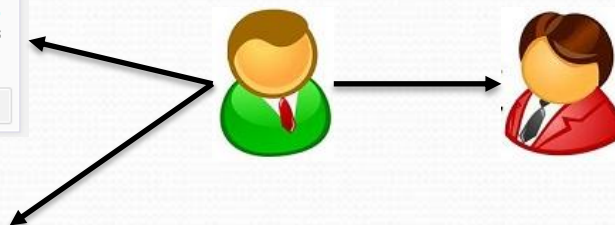


- ❖ 好友推荐、社团搜索、影响力分析
- ❖ 防御访问攻击、抵制骚扰行为



# 社交网络用户行为 为什么值得研究?

- ❖ “以人为本”、以用户为中心的应用 激发 用户行为
- ❖ 用户行为中潜在的需求 提出 研究问题和新的应用
- ❖ 发布、转发、评论，访问对方主页（“用户-信息”）



SocialBeta V

#摘要# LinkedIn是一家给企业提供解决方案的2B的公司，给企业提供三类解决方案：1. 招聘解决方案（Talent solution）；2. 营销解决方案（Marketing solution）；3. 销售解决方案（Sales solution）。



- ❖ 信息（微博）推荐、添加个人介绍、标签
- ❖ 多媒体信息包括图片、视频；增添群组
- ❖ 微博推广、舆论监督、抵制垃圾信息





# 提纲：社交网络用户行为

- ❖ 为什么值得研究？
- ❖ 具有哪些特点？
- ❖ 带来哪些机遇和挑战？
- ❖ 有什么感悟？



# 社交网络用户行为数据的四大特征

- ❖ 大规模（**Large-scale**）
- ❖ 富含关系属性（**Relational**）
- ❖ 多元异质性（**Heterogeneous**）
- ❖ 复杂的行为意图（**Complex**）





# 社交网络用户行为数据的四大特征

## ❖ 大规模 (Large-scale)

- 腾讯微博
- 1.2亿余注册用户，30亿余“关注”关系（2011年1月）
- 3.6亿余注册用户（2011年11月）
- 5.4亿余注册用户，1亿月活跃用户（2014年）
- 千万级的日产生微博数

## ❖ 富含关系属性 (Relational)

## ❖ 多元异质性 (Heterogeneous)

## ❖ 复杂的行为意图 (Complex)





# 社交网络用户行为数据的四大特征

- ❖ 大规模 (Large-scale)
- ❖ 富含关系属性 (Relational)
  - 有向图: Twitter, 微博等
  - 无向图: Facebook, 人人网等
  - 二分图: “用户-转发-微博”, “用户-添加-标签”等
  - 超图: “发布微信: 用户-设备-地点-照片-评论”等
- ❖ 多元异质性 (Heterogeneous)
- ❖ 复杂的行为意图 (Complex)



# 社交网络用户行为数据的四大特征

- ❖ 大规模 (Large-scale)
- ❖ 富含关系属性 (Relational)
- ❖ 多元异质性 (Heterogeneous)
  - 用户-用户 链接 (社交关系)
  - 用户-信息 链接 (微博、帖子、文章、标签、视频、群组等)
- ❖ 复杂的行为意图 (Complex)



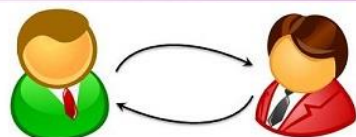
# 社交网络用户行为数据的四大特征

- ❖ 大规模 (Large-scale)
- ❖ 富含关系属性 (Relational)
- ❖ 多元异质性 (Heterogeneous)
- ❖ 复杂的行为意图 (Complex)
  - 正常意图：为了人际关系，为了获取信息，为了娱乐，为了炫耀
  - 异常意图：为了获取经济利益或商业价值

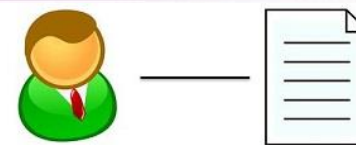




# 意图 + 链接 = 新应用



“用户-用户”链接



“用户-信息”链接



正常意图

影响力分析  
社区搜索、圈子发掘  
好友推荐、“关注”推荐



异常意图

僵尸粉检测  
可疑的好友请求、私信行为

商业推广  
真假新闻分类  
垃圾信息传播检测



# 提纲：社交网络用户行为

- ❖ 为什么值得研究？
- ❖ 具有哪些特点？
- ❖ 带来哪些机遇和挑战？
- ❖ 有什么感悟？



# “大规模”带来的问题：社交推荐系统

## ❖ 研究问题

- 用户每分钟都能接收到大量来自社交网络的信息
- 如何更好的推荐信息或是对新鲜事进行排序？
- 我们是否能够预测用户下一步会点击/分享/转发什么？

## ❖ 问题定义

- 预测缺失的“用户-信息”链接（“用户-信息”矩阵中的缺省值）

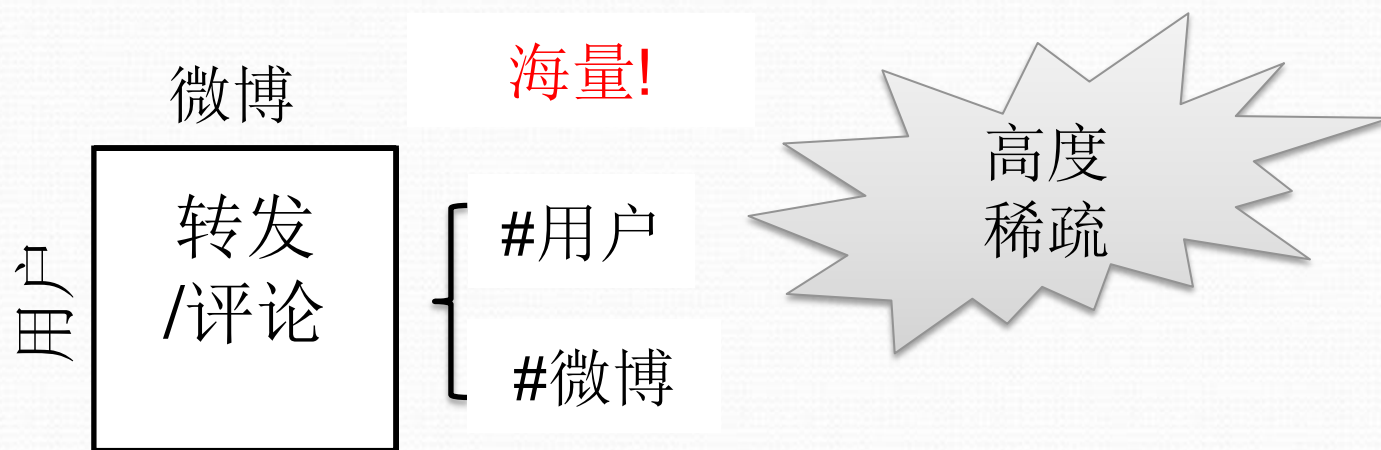


	User 1	User 2	...	User M
Tweet 1	1	1	...	1
Tweet 2	0	?	...	?
...	...	...	...	...
Tweet N	1	?	...	1





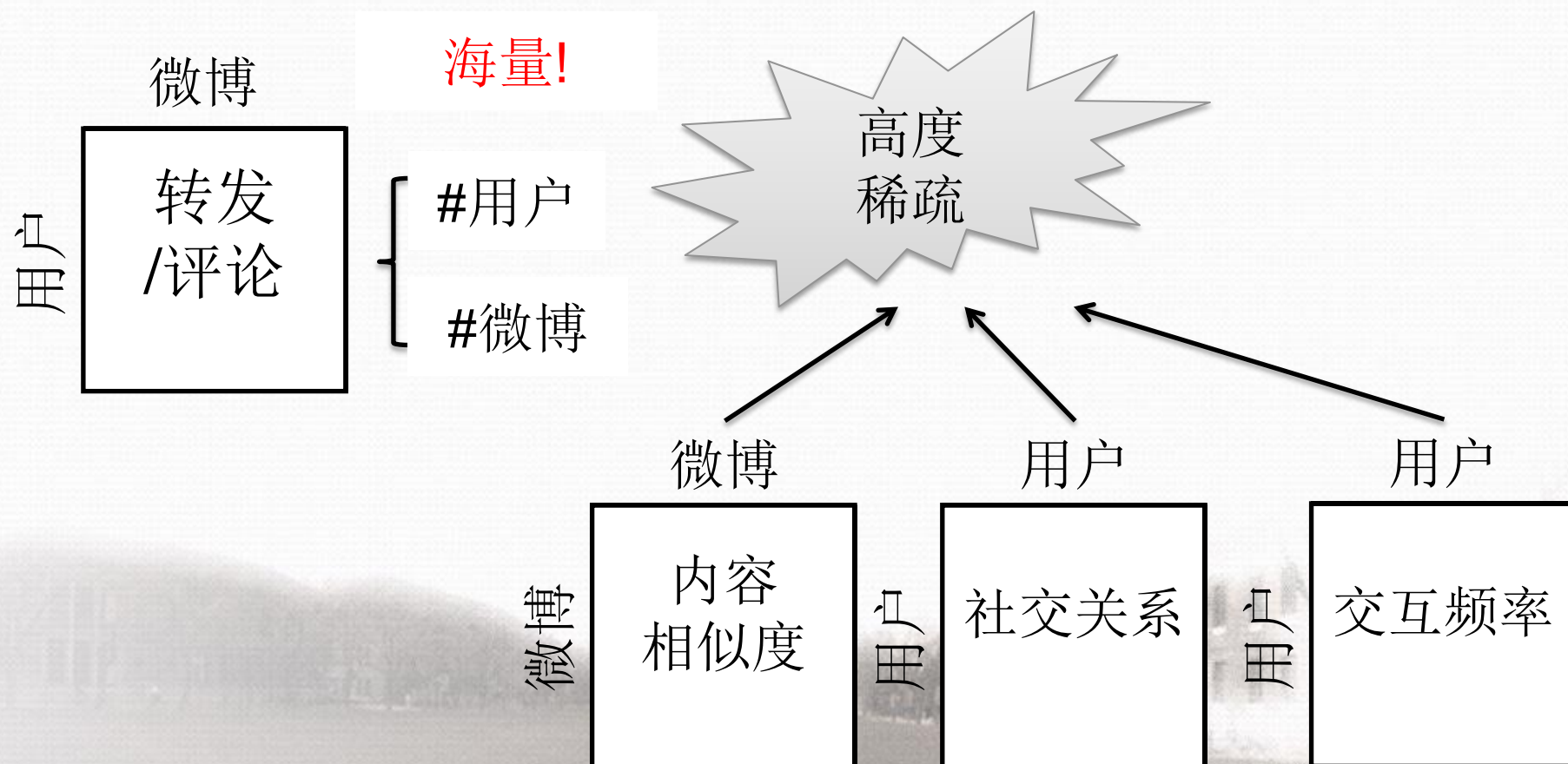
# “大规模”带来的问题：社交推荐系统





# “大规模”带来的问题：社交推荐系统

❖ 我们能否用其他已有链接帮助预测？

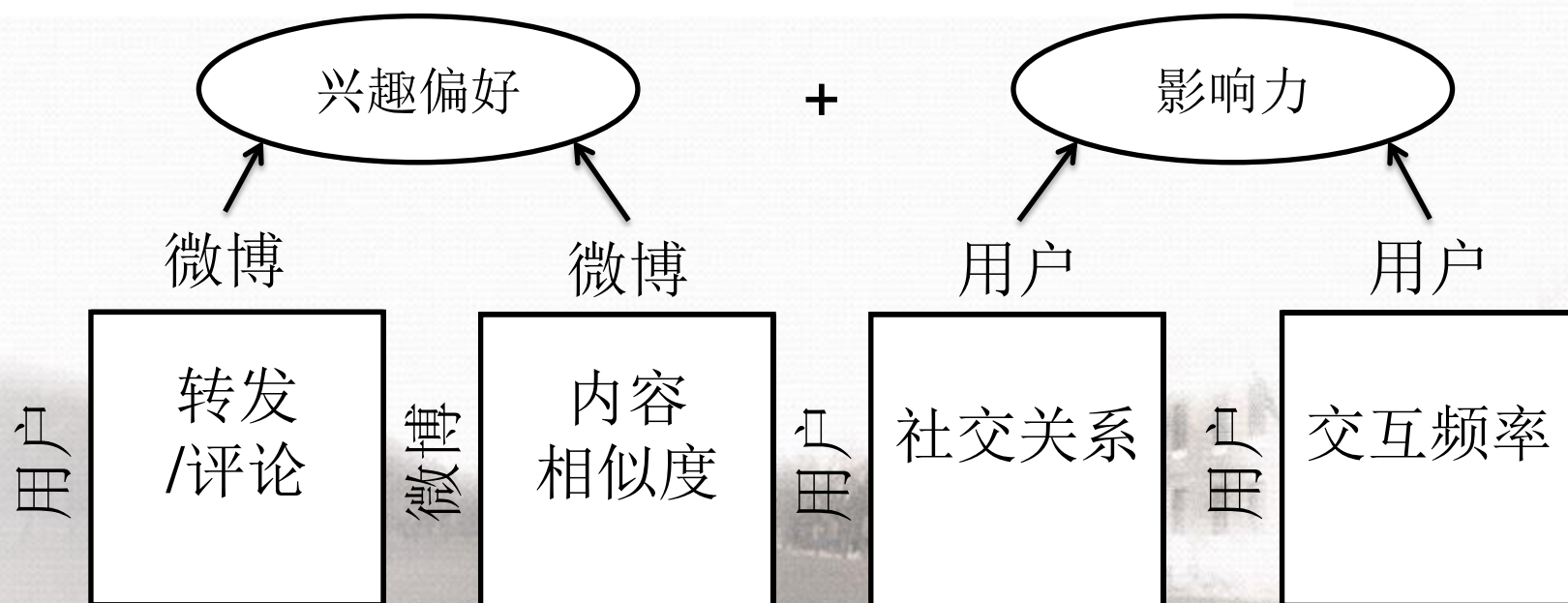




# “大规模”带来的问题：社交推荐系统

❖ 可以! 必须理解用户行为的意图!

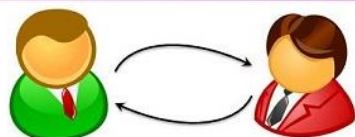
1. 收到某个消息
2. 内容是什么? 谁发布的?
3. 转发, 还是不转发.....



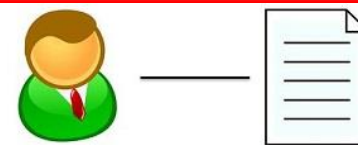




# “大规模”带来的问题：社交推荐系统



“用户-用户”链接



“用户-信息”链接

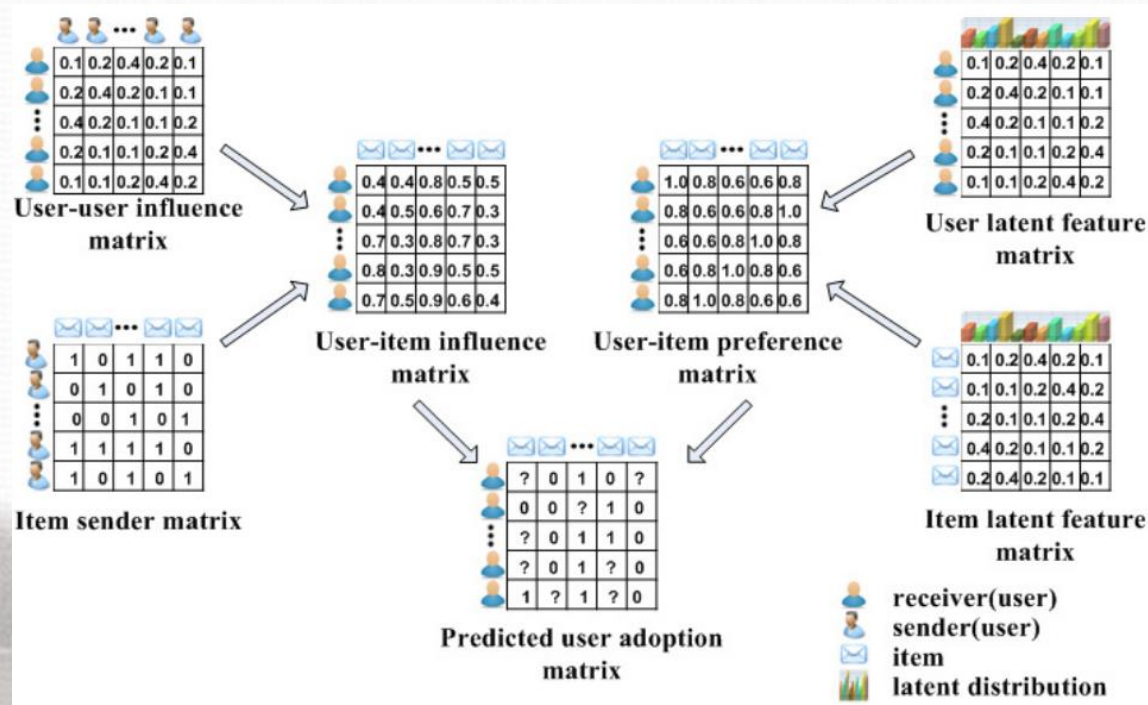


正常意图



异常意图

基于上下文信息的社交推荐算法 [Jiang et al. CIKM 2012]



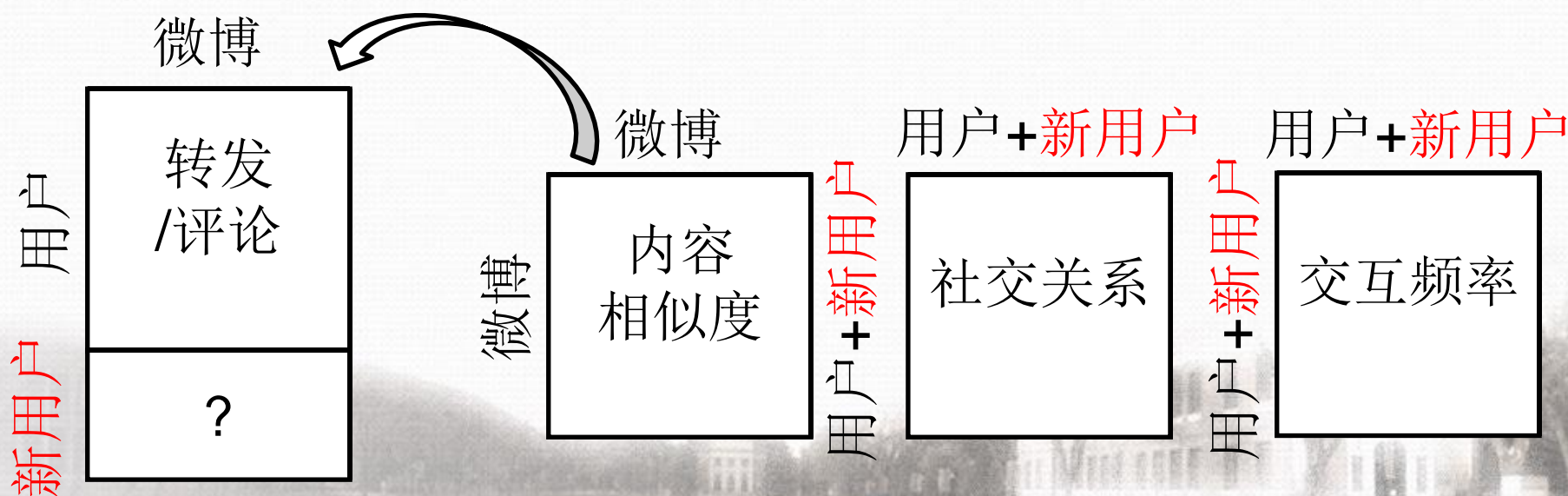


# “大规模”带来的问题：可扩展的社交推荐算法

## ❖ 研究问题

- 如何处理新来的用户？如何处理新来的微博？
- 我们能不能用已有用户和微博的模型结果来快速给出结果？

## ❖ 问题定义（新来的用户）



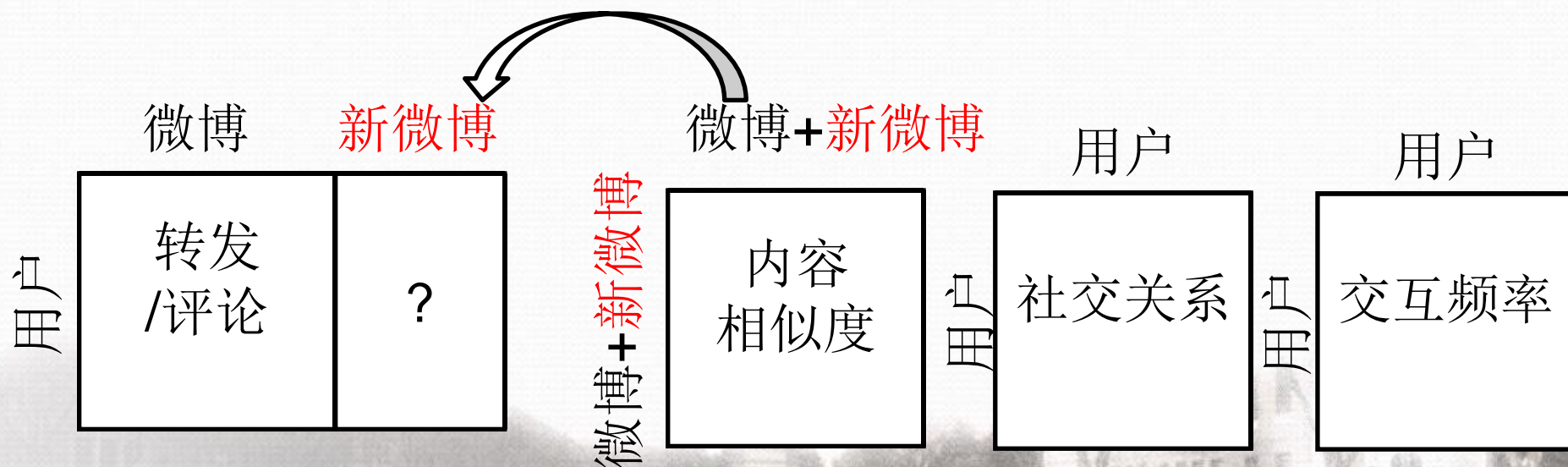


# “大规模”带来的问题：可扩展的社交推荐算法

## ❖ 研究问题

- 如何处理新来的用户？如何处理新来的微博？
- 我们能不能用已有用户和微博的模型结果来快速给出结果？

## ❖ 问题定义（新来的微博）







# “大规模”带来的问题：冷启动问题

## ❖ 研究问题

- 我们已经成功解决（新用户，已有微博）和（已有用户，新微博）的情形。那么（**新用户，新微博**）呢？如果新来用户从未分享过微博呢？

	微博	新微博
用户	√	√
新用户	√	?

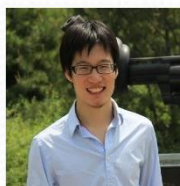


# “关系性”和“多元异质性”带来的机遇： 解决冷启动问题

❖ 我们在其他领域有着富裕知识可供迁移学习。

## ❖ 标签领域

从200多社交标签（如“iPhone迷”）中选择不超过10个



崔鹏

北京 海淀  
公司：清华

### 标签(5)

清华，博士，万维网，  
社交网络，社交媒体



蒋朦

北京 海淀  
公司：清华

### 标签 (9)

中国菜，万维网，社交网络，数据挖掘，  
利物浦，NBA，幽默，体育，博士生



# “关系性”和“多元异质性”带来的机遇： 解决冷启动问题

- ❖ 我们在其他领域有着富裕知识可供迁移学习。
- ❖ 群组领域



群组(2)



清华大学



艺术协会



群组(3)



清华大学



社交媒体



万维网社区



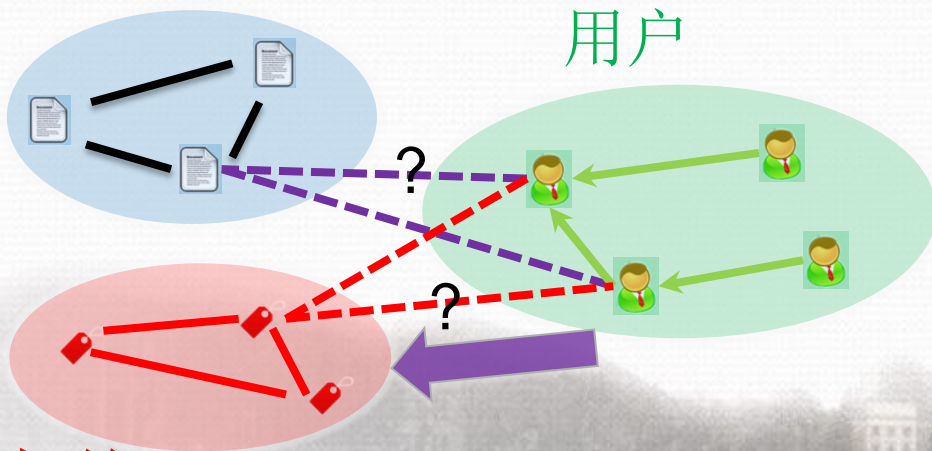


# “关系性”和“多元异质性”带来的机遇： 解决冷启动问题

❖ 含有多种关系域的社交网络应该如何建模？

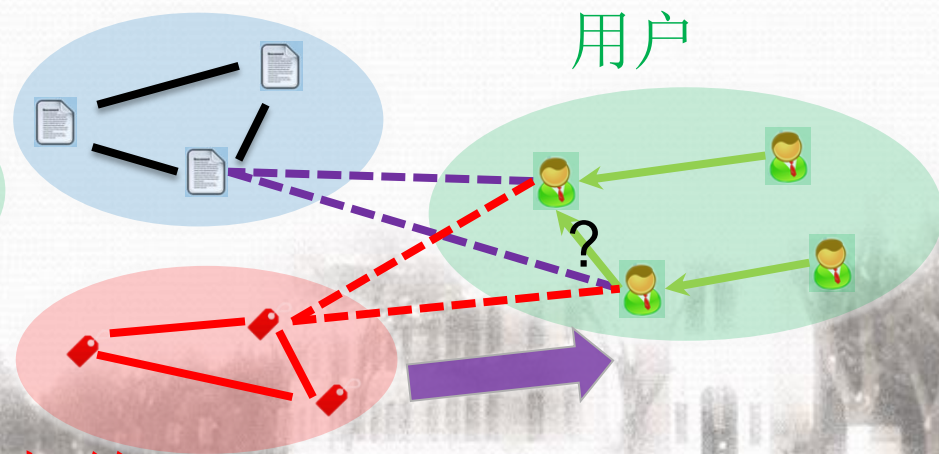
- 我们有“用户-微博”，“用户-标签”和“用户-群组”链接
- 非用户的不同类信息之间并无直接链接，如微博和标签并无自然联系
- 更强的社交关系会让每一种“用户-信息”链接都产生协作效果
- 更频繁地“用户-信息”协作能增强“用户-用户”的社交关系

微博



标签

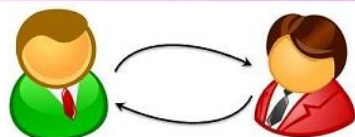
微博



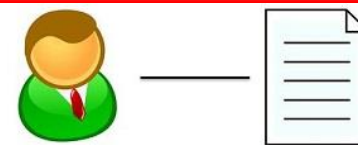
标签



# “关系性”和“多元异质性”带来的机遇： 解决冷启动问题



“用户-用户”链接



“用户-信息”链接

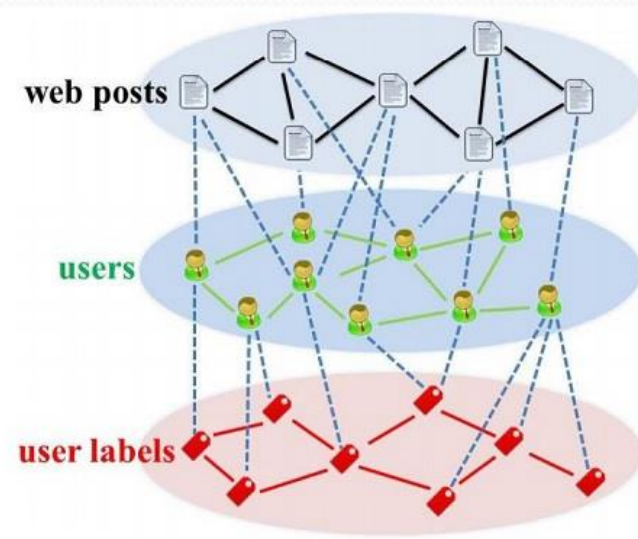
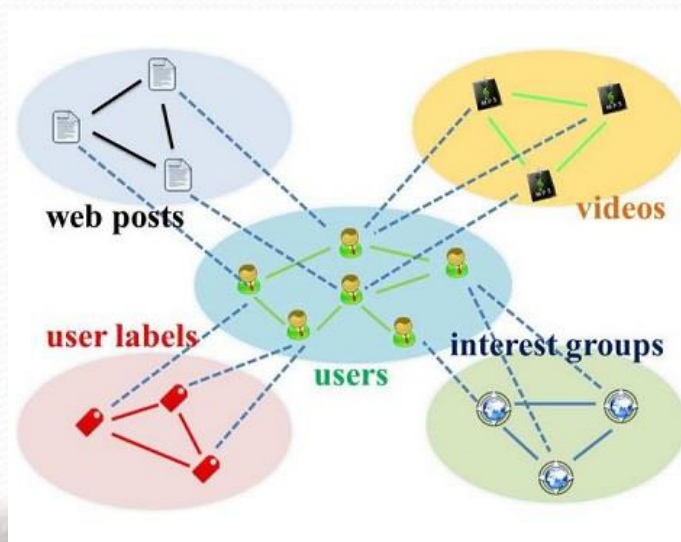
跨域的社交推荐算法 [Jiang et al. CIKM 2012]



正常意图



异常意图

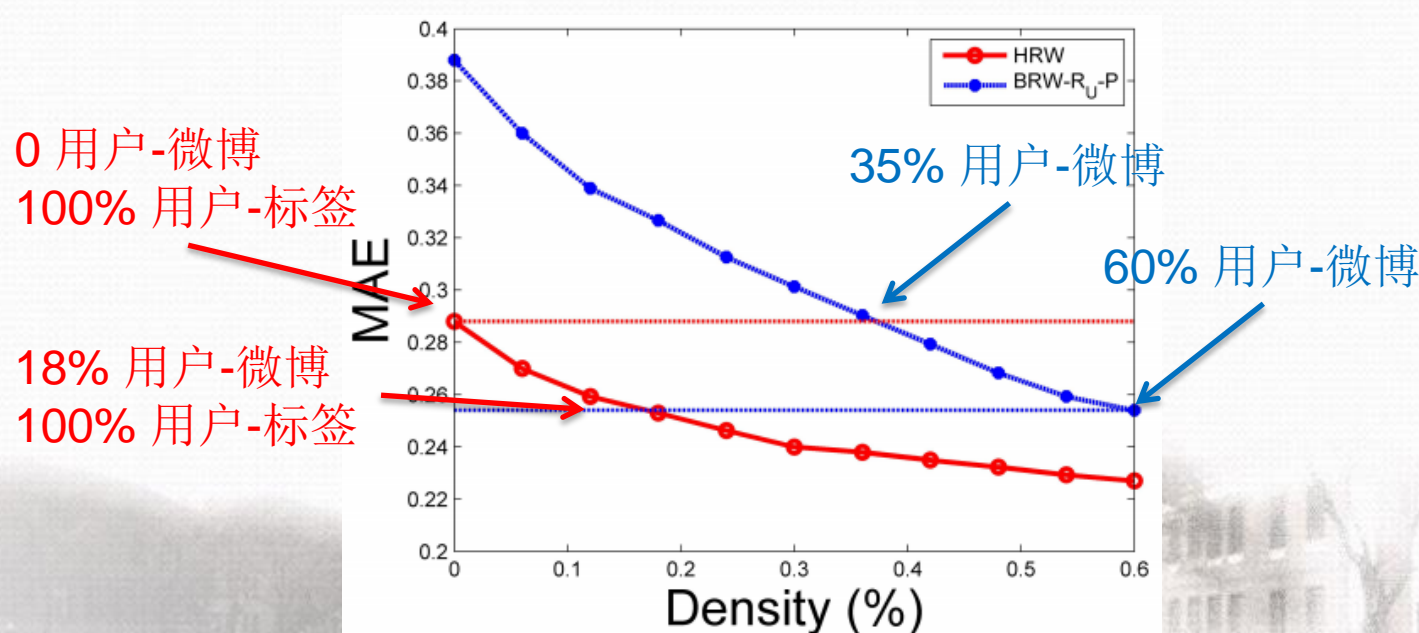






# “关系性”和“多元异质性”带来的机遇： 解决冷启动问题

- ❖ 如果从“用户-标签”域中迁移学习知识来预测“用户-微博”域中的链接，只需要训练数据量的30%就能达到原效果
- ❖ 建议：增添更多应用让新用户提供更多信息吧！







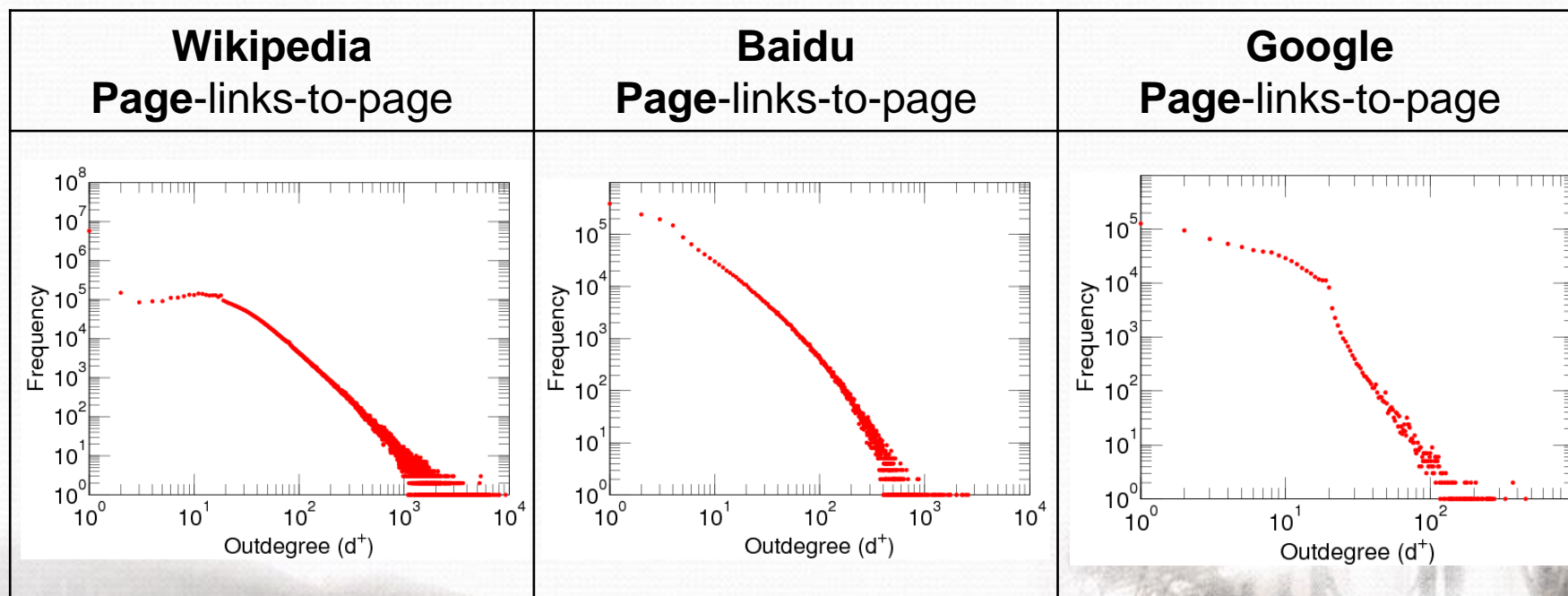
# 社交网络用户行为数据的四大特征

- ❖ 大规模 (Large-scale)
- ❖ 富含关系属性 (Relational)
- ❖ 多元异质性 (Heterogeneous)
- ❖ 复杂的行为意图 (Complex)



# “大规模”的统计模式：出度分布

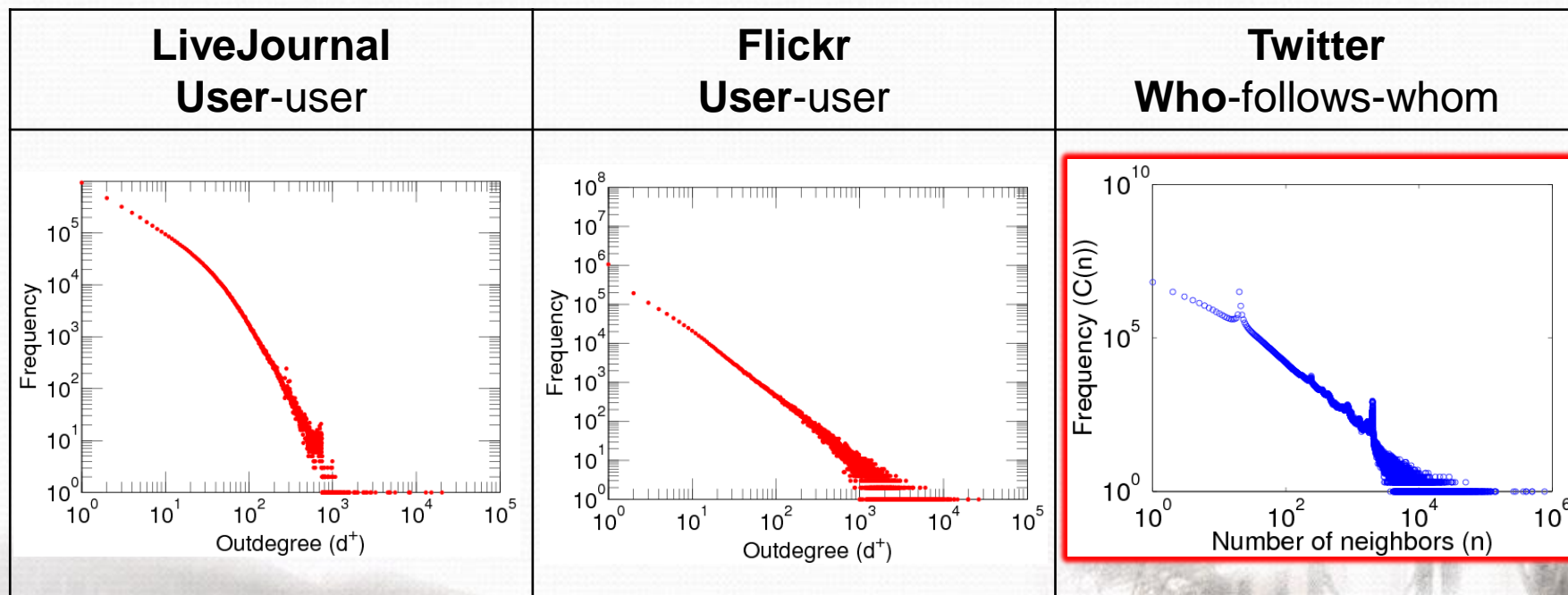
## ❖ 幂律分布（有向图）





# “大规模”的统计模式：出度分布

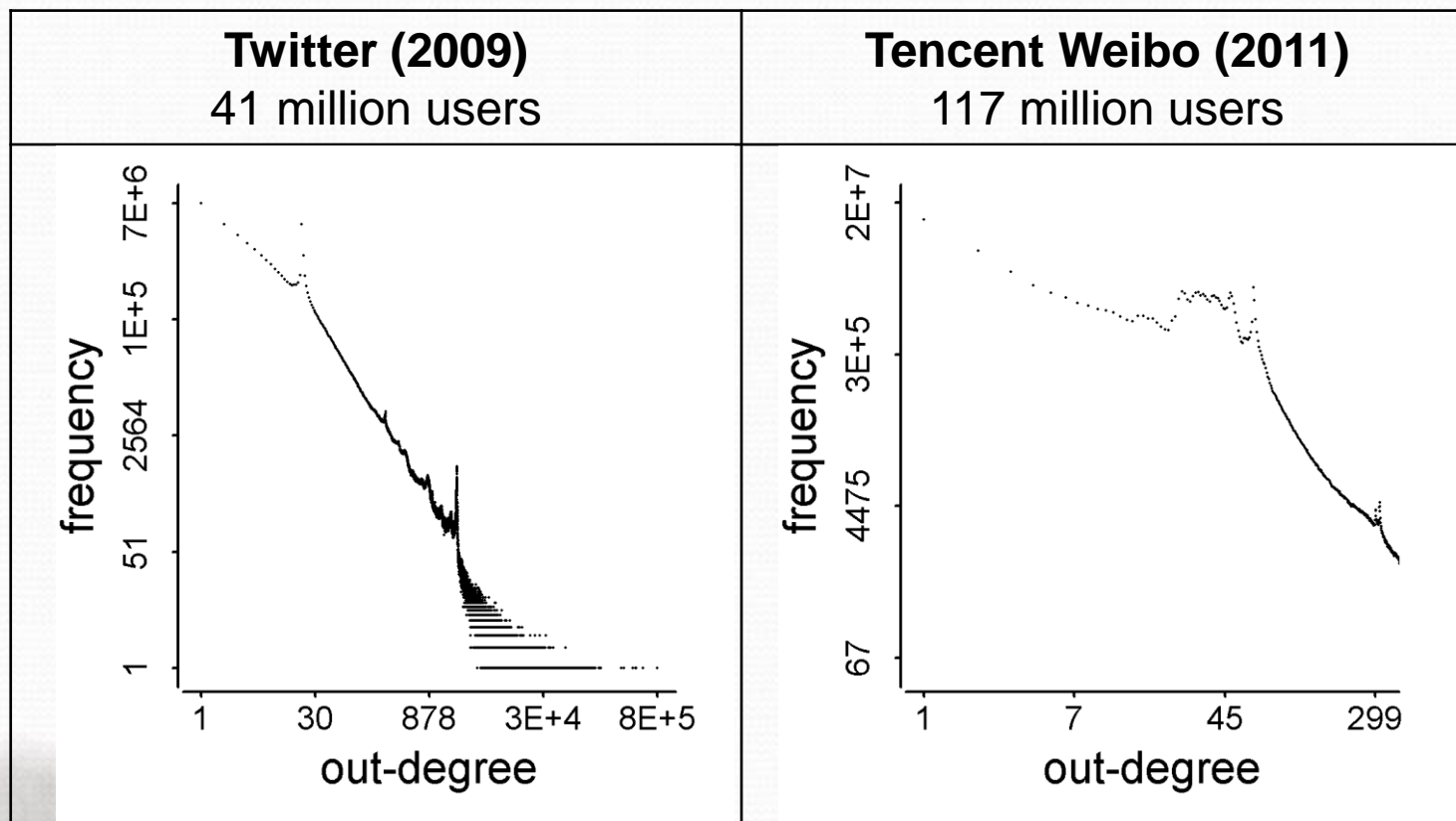
## ❖ 幂律分布（有向图-社交网络）







# 我们的世界里.....很复杂





# “大规模”和“复杂意图”的挑战：僵尸粉检测

## ❖ 挑战

- **可扩展性**: 如何在拥有亿万节点和边的社交网络图中抓住僵尸粉？  
我们可以解释出度分布中的尖峰吗？

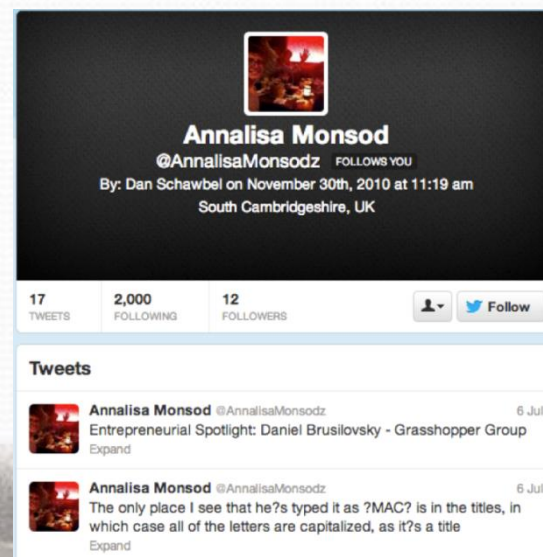




# “大规模”和“复杂意图”的挑战：僵尸粉检测

## ❖ 挑战

- 可扩展性: 如何在拥有亿万节点和边的社交网络图中抓住僵尸粉？我们可以解释出度分布中的尖峰吗？
- **伪装能力**: 假冒个人信息，没有或很少发布微博，有其他的貌似正常的行为



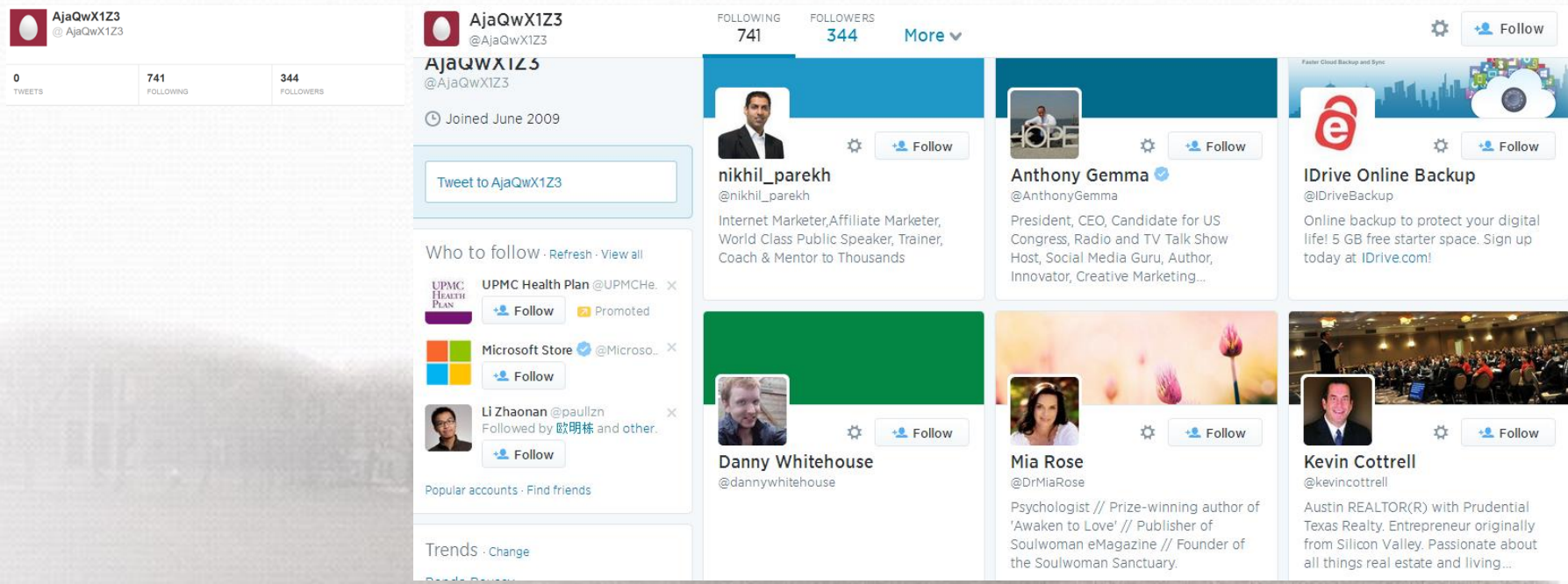




# “大规模”和“复杂意图”的挑战：僵尸粉检测

## ❖ 挑战

- 可扩展性: 如何在拥有亿万节点和边的社交网络图中抓住僵尸粉？我们可以解释出度分布中的尖峰吗？
- **伪装能力**: 假冒个人信息，没有或很少发布微博，有其他的貌似正常的行为






# “大规模”和“复杂意图”的挑战：僵尸粉检测

## ❖ 挑战

- 可扩展性: 如何在拥有亿万节点和边的社交网络图中抓住僵尸粉？我们可以解释出度分布中的尖峰吗？
- **伪装能力**: 假冒个人信息，没有或很少发布微博，有其他的貌似正常的行为

**Buy AB22 Propertwee**  
@Buy\_AB22


0 TWEETS	20 FOLLOWING	2 FOLLOWERS
-------------	-----------------	----------------


**Buy AB22 Propertwee**  
@Buy\_AB22


FOLLOWING  
20


FOLLOWERS  
2


More


**B.J. Mendelson**  
@BJMendelson  
I just post silly stuff that I found on the Internet. None of it is mine.


**someecards**  
@someecards  
Welcome to the Twitter feed of somewhat acclaimed humor sites, Someecards.com and HappyPlace.com. You'll love not unfollowing us!


**Steven Johnson**  
@stevenjohnson  
Author. (Latest: Future Perfect.) TV host. (How We Got To Now, on PBS/BBC soon.) Startup creator. (FEED, outside.in.) Dad. (Three boys.)...

**adventuregirl**  
@adventuregirl  
Hi Everyone! I'm Stef Michaels- an avid lifestyles journalist, TV personality, adventurer. Co-founder of KEEN Digital Summit. Contributor to Yahoo! Travel.

**People magazine**  
@peoplemag  
PEOPLE.com is the No. 1 site for celebrity news! Tweet your questions to our customer service team @Peoplemag\_Help

**ashton kutcher**  
@aplusk  
I make stuff, actually I make up stuff, stories mostly, collaborations of thoughts, dreams, and actions. That's me.

**Paul Pierce**  
@paulpierce34  
The one and only Truth. Founder of The @TruthonHealth.

**Mashable**  
@mashable  
News, resources, inspiration and fun for the connected generation. Tweets by

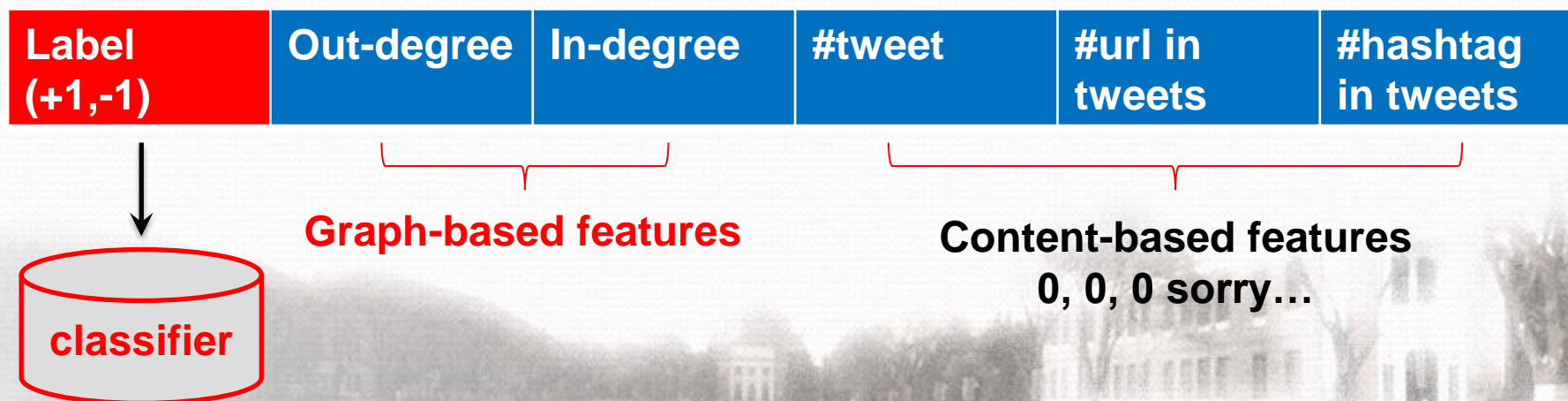


# “大规模”和“复杂意图”的挑战：僵尸粉检测

## ❖ 挑战

- 可扩展性: 如何在拥有亿万节点和边的社交网络图中抓住僵尸粉？我们可以解释出度分布中的尖峰吗？
- 伪装能力: 假冒个人信息，没有或很少发布微博，有其他的貌似正常的行为

## ❖ 已有工作

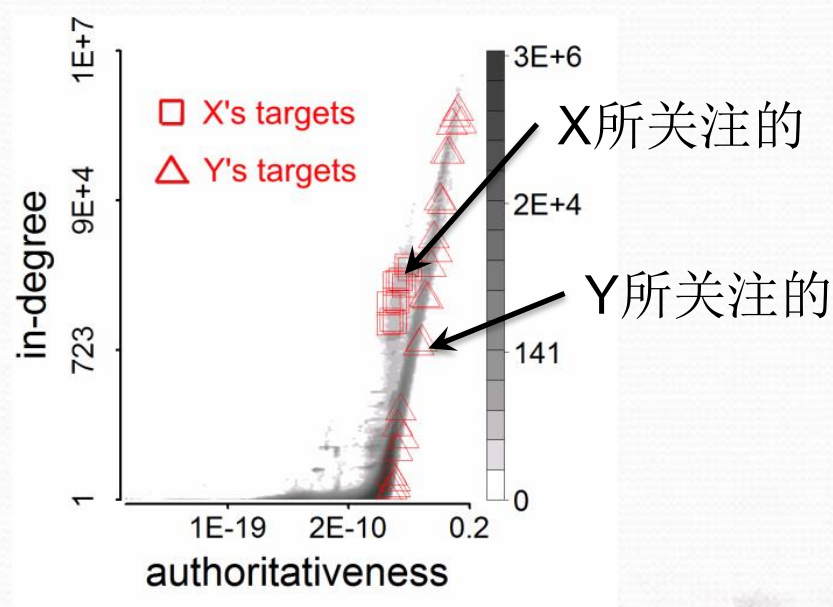
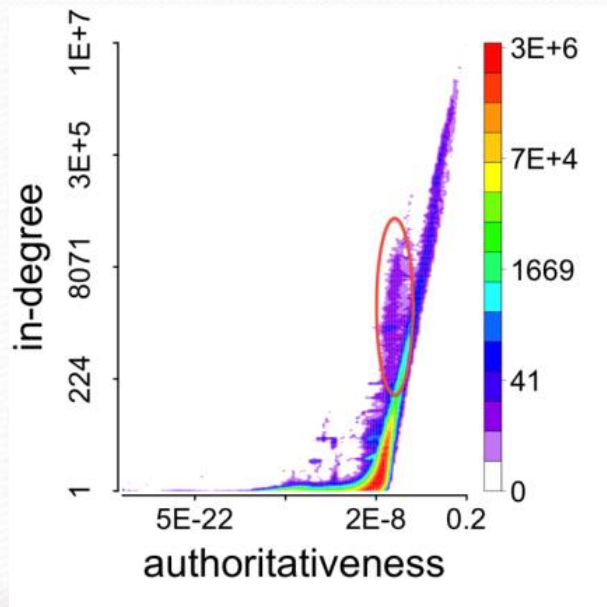






# 比较僵尸粉和正常用户

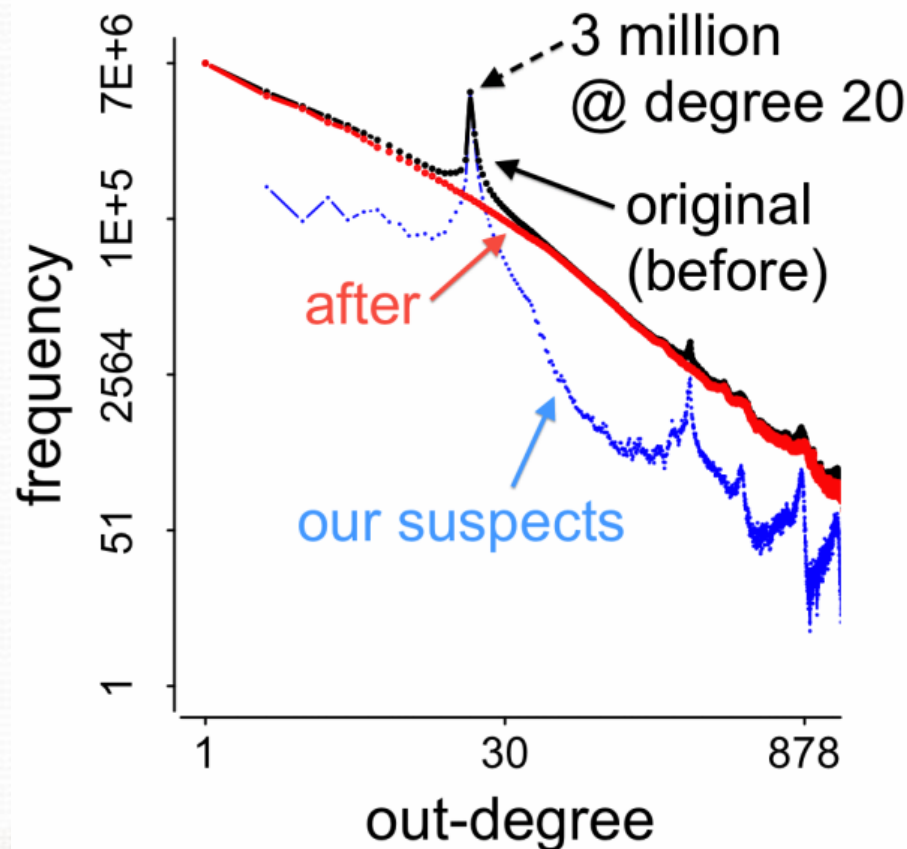
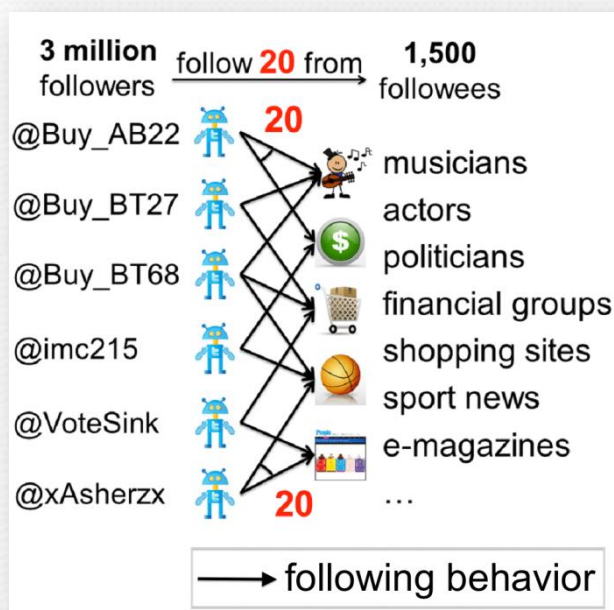
- ❖ X = @Buy\_AB22: 关注了20个人的僵尸粉
- ❖ Y = 关注了20个人的正常用户



- ❖ 可疑行为: 和同组的人极相似, 和全网大多数用户很不同

# 我们找到的是异常用户吗？

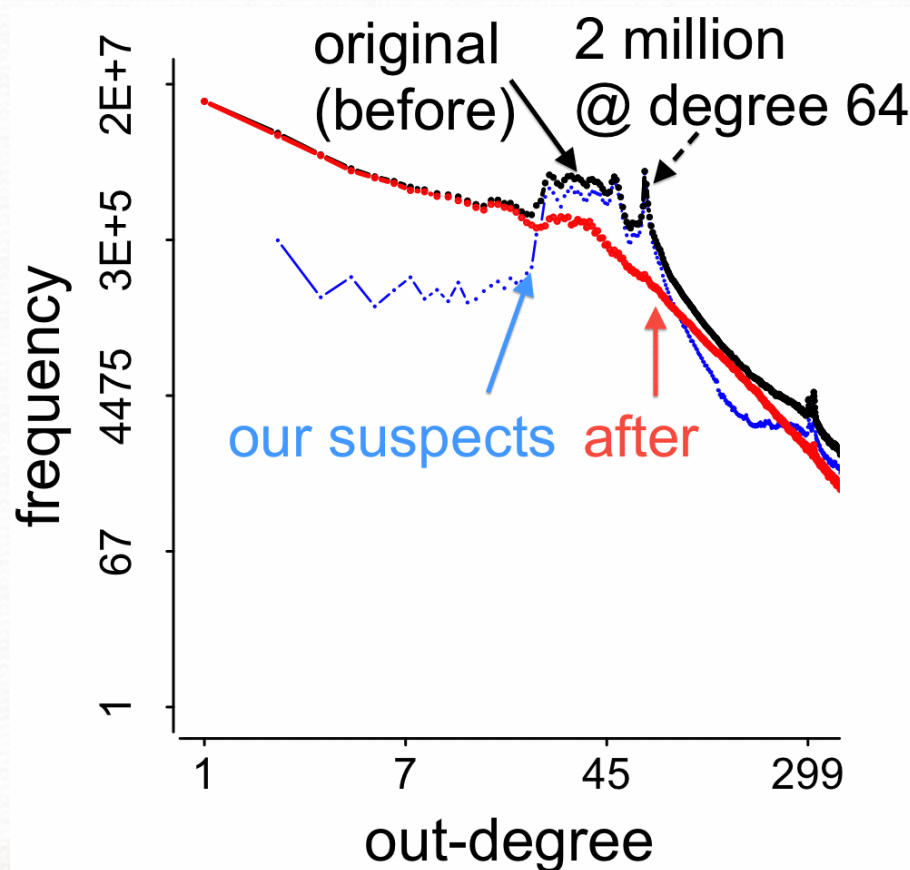
## ❖ Twitter





# 我们找到的是异常用户吗？

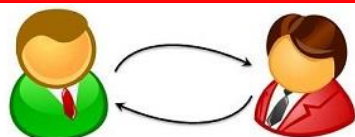
## ❖ 腾讯微博



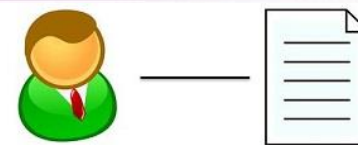




# “大规模”和“复杂意图”的挑战：僵尸粉检测



“用户-用户”链接

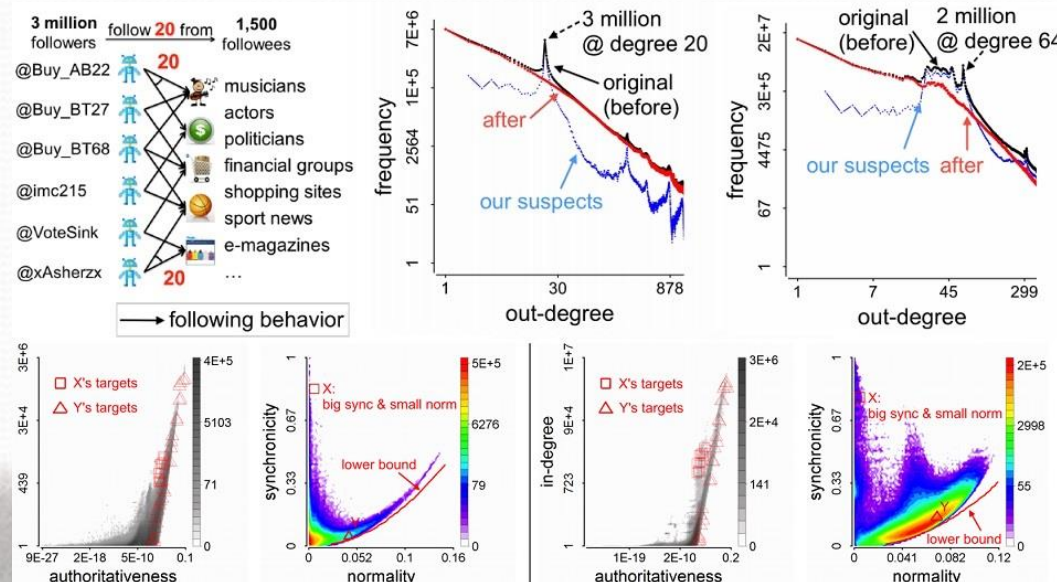


“用户-信息”链接



正常意图

## 僵尸粉检测 [Jiang et al. KDD 2014]



异常意图



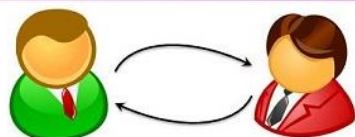
# 提纲：社交网络用户行为

- ❖ 为什么值得研究？
- ❖ 具有哪些特点？
- ❖ 带来哪些机遇和挑战？
- ❖ 有什么感悟？

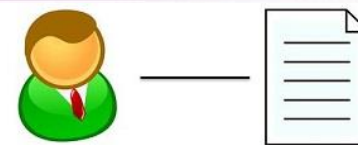




# 感悟：意图 + 链接 = 新应用



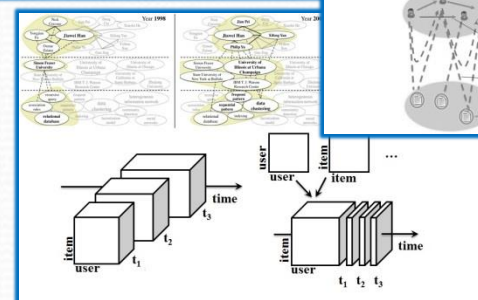
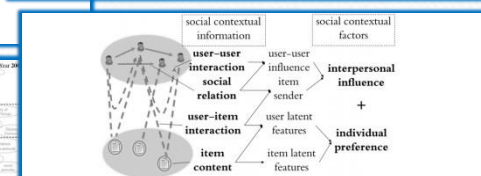
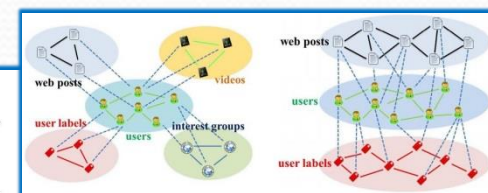
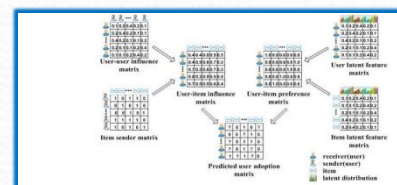
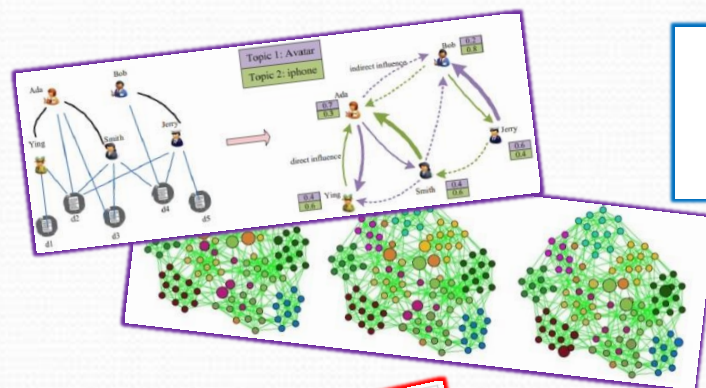
“用户-用户”链接



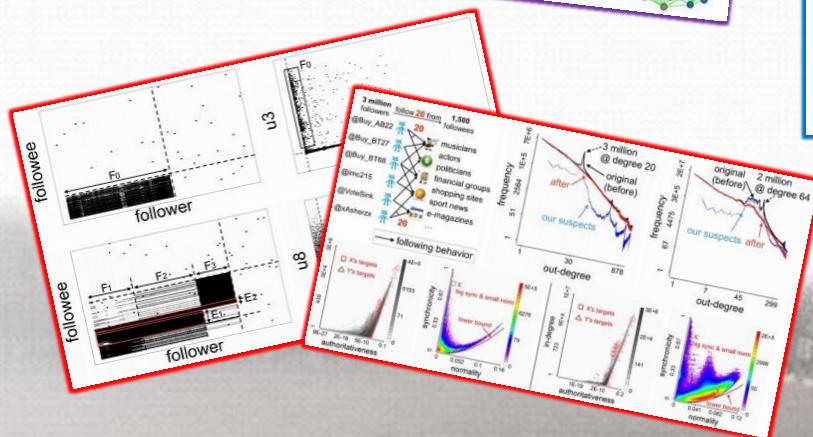
“用户-信息”链接



正常意图



异常意图







# 参考文献

- ❖ **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.



# 致谢

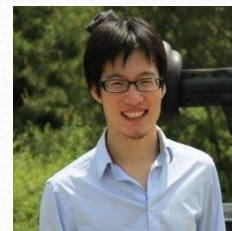
## ❖ Tsinghua University



Shiqiang Yang



Wenwu Zhu



Peng Cui



Lu Liu

## ❖ Carnegie Mellon University



Christos Faloutsos



Alex Beutel

## ❖ IBM T. J. Watson Research Center



Fei Wang

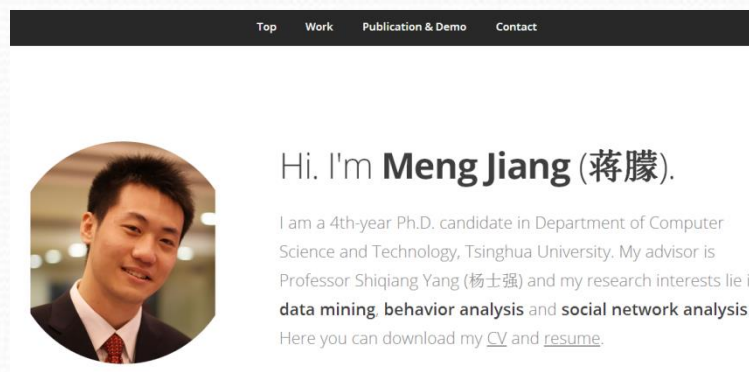




# 谢谢大家!

欢迎访问我的个人主页:

<http://www.meng-jiang.com>



♥ 交友 ♥ 讨论 ♥ 合作