社交网络中用户行为的分析与预测

报告人：程学旗
合作者：沈华伟、黄俊铭等
中国科学院网络数据科学与技术重点实验室

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

大纲

- 背景
- 社交网络中的用户行为分析和预测
  - 用户兴趣估计
  - 信息传播范围预测
  - 影响力最大化
- 开放问题

人类行为分析与预测

- 人类行为分析与预测的意义
  - 作为一个科学问题从Watson(1913)开始研究已近一个世纪
  - 社会学、心理学、经济学的共同研究兴趣
  - 价值：理解、指导社会群体行为
- 人类行为分析与预测的困难
  - 人类行为的复杂性和多样性
  - 难以获得大规模样本以提供稳定统计

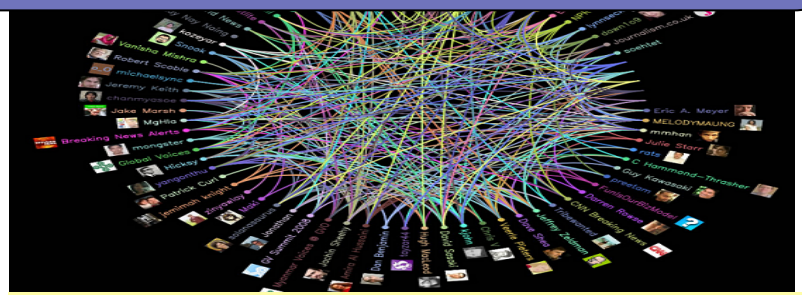
在线社交网络兴起

- 在线社交网络记录了大
  - 在线社交网络记录了人们
    - 关系(Facebook、Google+)、作(Linkedin)、照片(Flickr)
  - 在线社交网络飞速发展

社交网络的海量数据为人类行为分析和预测提供机遇

Facebook世界地图

近9亿用户，1400亿多条连接



5亿以上用户，每天数十亿条信息传播

基于大规模社交网络数据的行为分析

- 学术界广泛关注
  - KDD、SIGIR、WWW、NIPS、ICML、WSDM等顶级会议和期刊常有社交网络分析的讨论
  - WSDM等会议设立social day专门讨论社交网络
- 应用领域持续关注
  - 利用Twitter分析预测美国股市和大选
  - 根据Google搜索日志预警流感爆发
  - Amazon根据海量用户行为日志提供个性化推荐

从计算的角度看人类的行为

- 人类行为的可观测性（测不准？）
  - 如何观测隐藏的人类行为动机？
  - 对人类行为的观测本身是否会造成影响？
- 人类行为的可预测性（看不清？）
  - Barabasi: 93%的人类行为可以预测
  - 黑天鹅事件是大自然固有的随机属性还是仅仅因为学习规模不足？

从计算的角度看个人与群体的关系

- 社会因素的可计算性
  - 如何量化计算社交网络中的行为与影响？如何定量预测个体和群体行为？
  - 如何量化社会环境、心理因素、物理事件对行为的影响
- 同配性与影响力的争论
  - 朋友常常具有相似的行为，是因为交友过程的同配选择还是因为行为影响的趋同过程？
  - 影响力如何定性判断和定量度量？

本次报告关注的问题

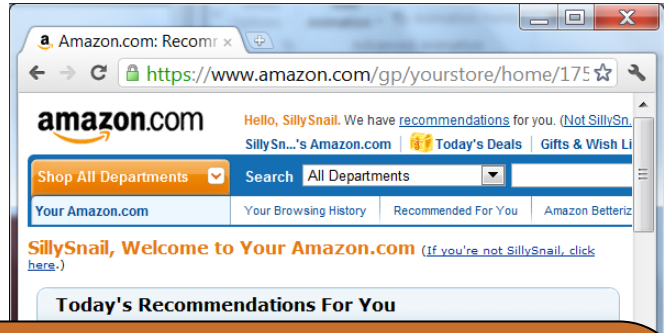
- 社交网络中个体与群体行为的典型问题
  - 准确估计用户兴趣
  - 理解信息传播机制并预测传播范围
  - 设计影响力最大化的传播策略

能否通过某人的朋友们的行为，准确估计其兴趣分布？

用户兴趣估计

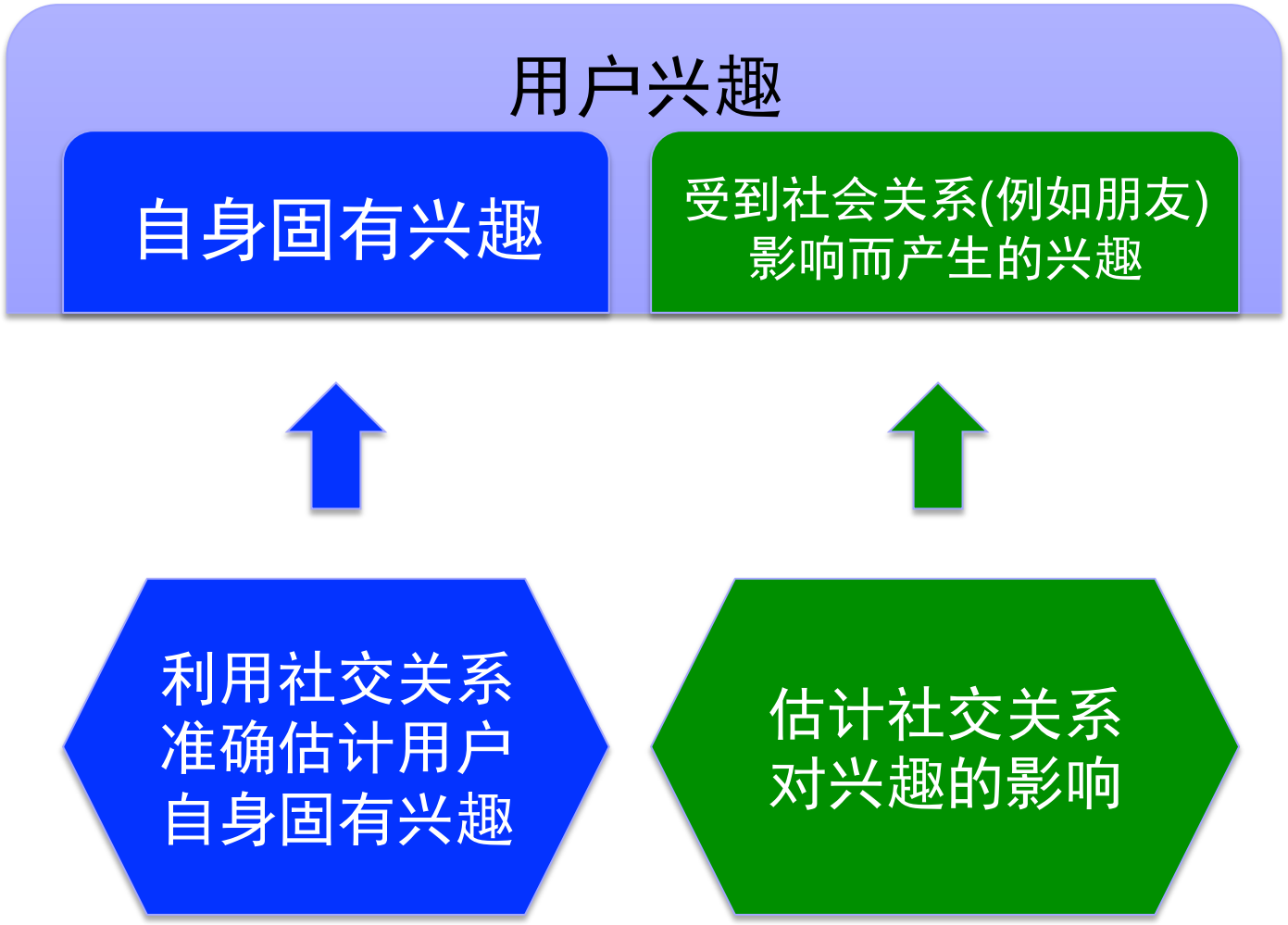
准确估计用户兴趣

- 信息过载催生个性化时代
  - 个性化搜索引擎
  - 个性化商品推荐
  - 个性化医疗方案
  - 个性化阅读

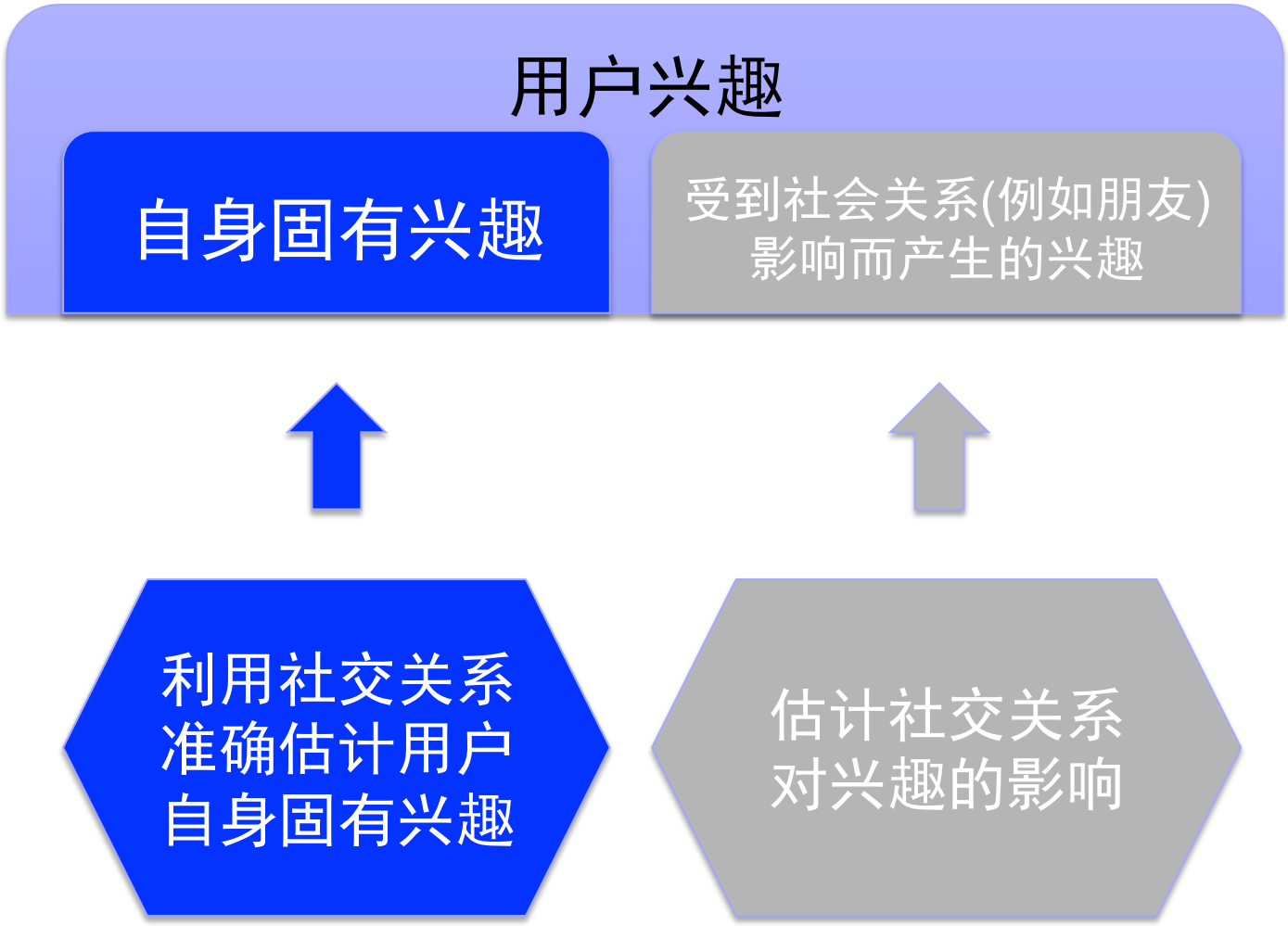


准确估计用户兴趣
是个性化的第一步

利用社交网络准确估计用户兴趣



利用社交网络准确估计用户兴趣

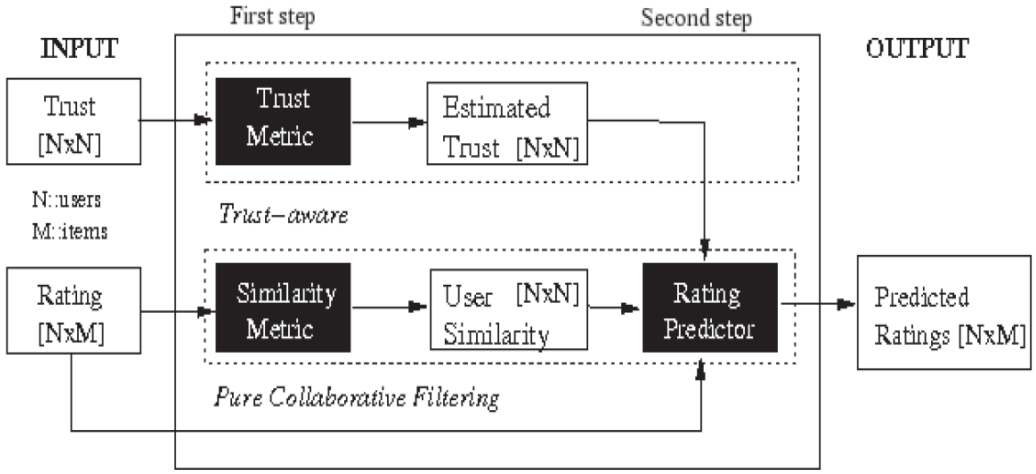


利用社交关系估计用户固有兴趣

- 认为相互信任的用户具有相似的兴趣
- 基于用户关系建立推荐系统
  - 提高推荐结果的准确性
    - 朋友的行为间接反映用户兴趣
  - 增强推荐结果的可解释性和可接受性
    - 展示结果时同时展示朋友的行为

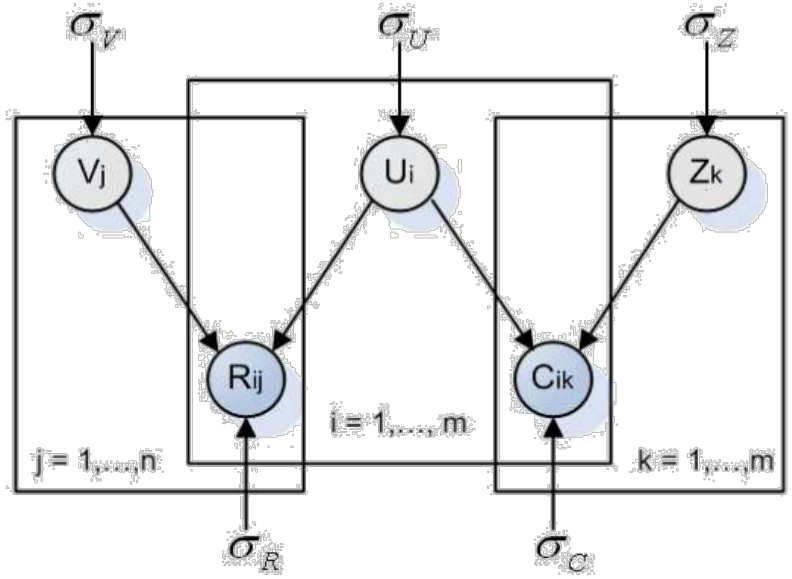
利用社交关系估计用户固有兴趣

- 认为相互信任的用户具有相似的兴趣
  - Trust-aware recommender system [Massa 2007]
    - 度量用户之间信任关系
    - 结合信任度矩阵和相似度矩阵预测用户兴趣



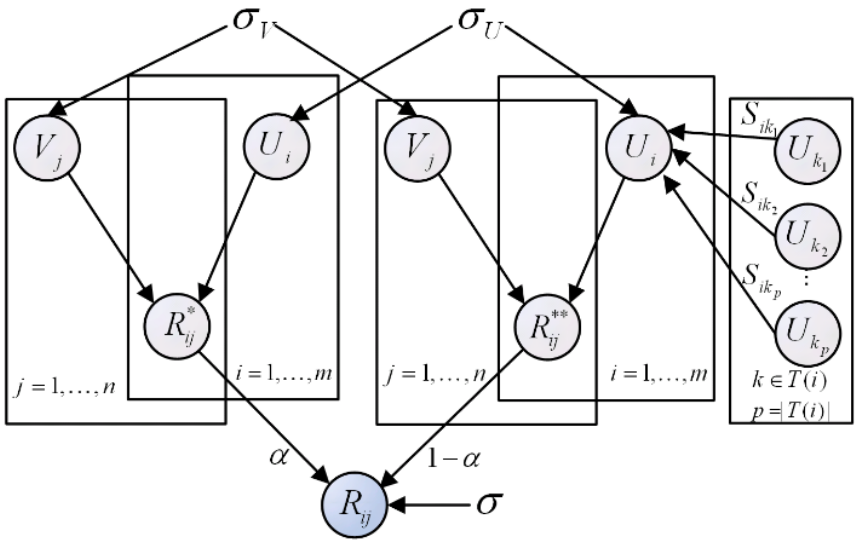
利用社交关系估计用户固有兴趣

- 认为相互信任的用户具有相似的兴趣
  - Trust-aware recommender system [Massa 2007]
  - SoRec [Ma 2008]
    - 基于概率化矩阵分解模型(PMF)
    - 约束用户的隐向量与朋友的隐向量应该相似



利用社交关系估计用户固有兴趣

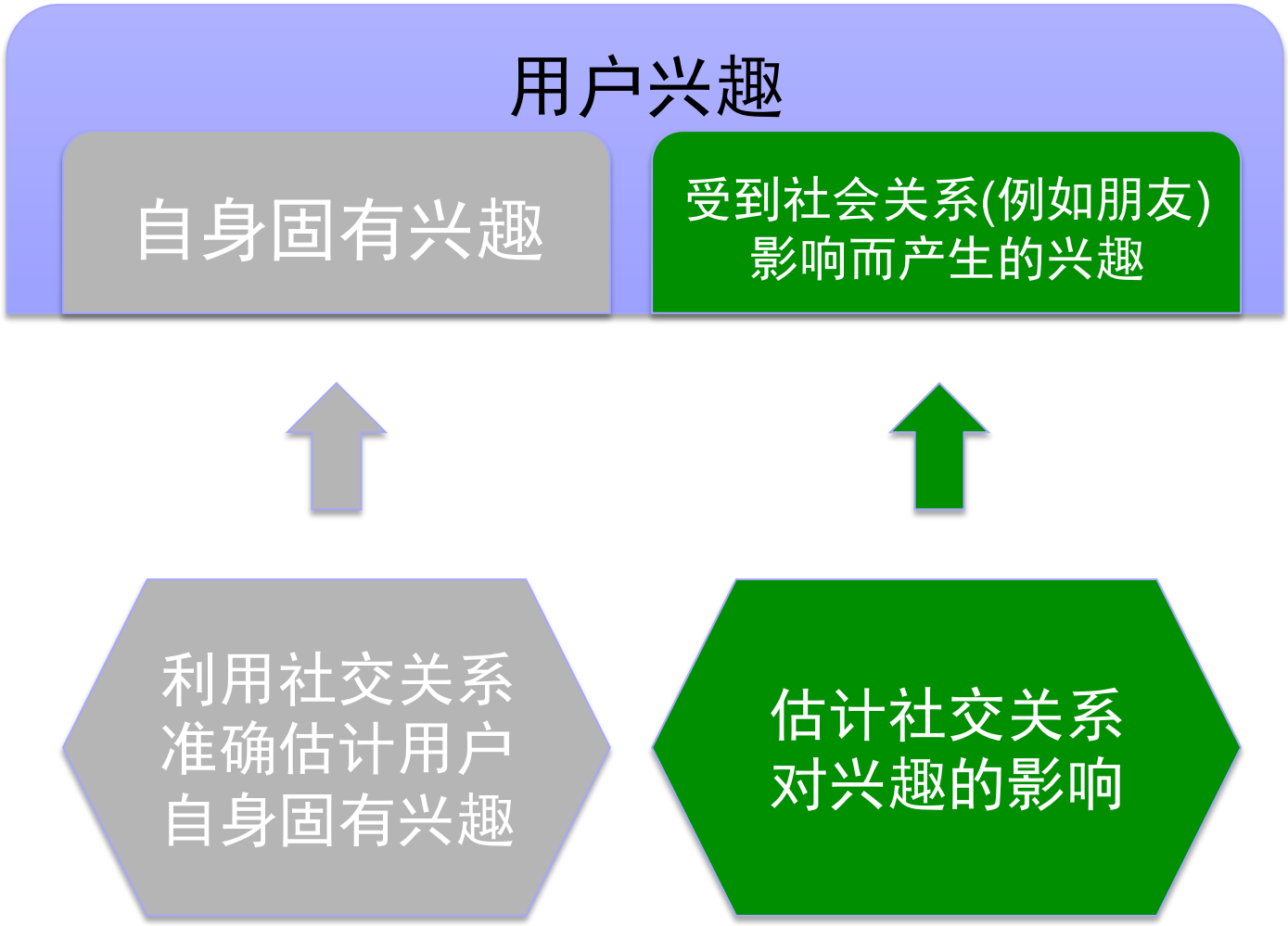
- 认为相互信任的用户具有相似的兴趣
  - Trust-aware recommender system [Massa 2007]
  - SoRec [Ma 2008]
  - Learning to rec with trust ensemble [Ma 2009]
    - 用户决策由反映自身兴趣的隐向量和好友的隐向量共同决定



利用社交关系估计用户固有兴趣

- 优点
  - 提高了兴趣估计的准确率，尤其是冷启动用户
  - 提高了推荐结果的可解释性与可接受性
- 局限
  - 用户之间的信任关系并不总能观测到

利用社交网络准确估计用户兴趣



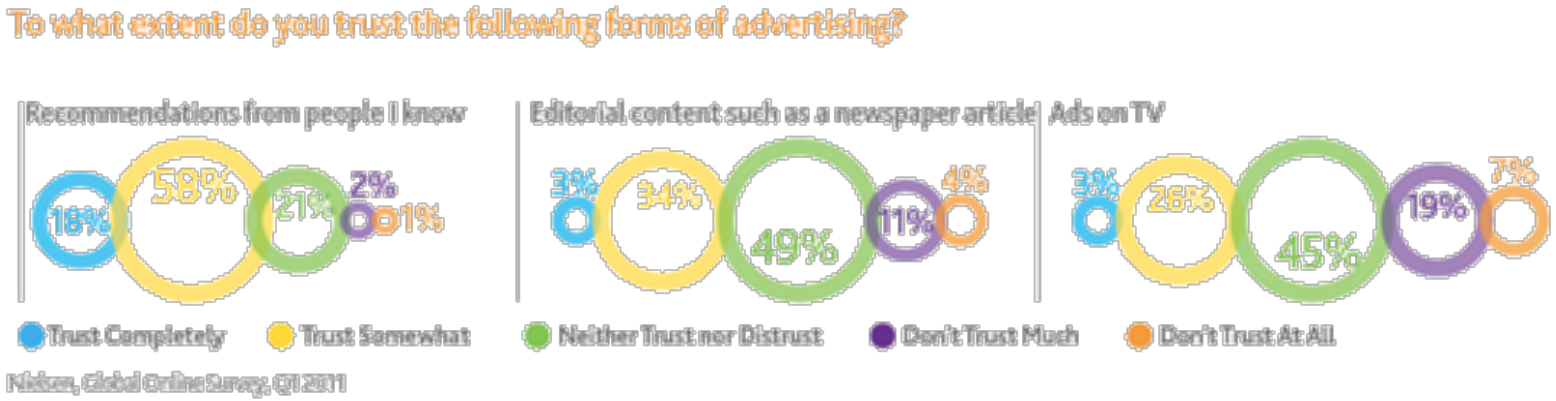
来自社会关系的影响是否存在

现象：朋友之间常常具有类似的兴趣

- 因为同质化？
- 因为影响力？

来自社会关系的影响是否存在

- 工业界的实证研究认为影响力是社交活动中的重要内容
  - 缺乏严格的学术讨论



社会推荐的影响

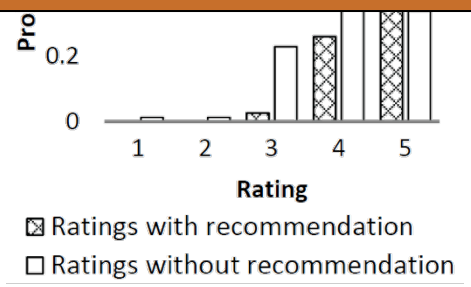
- 社会推荐
  - 经由社交关系口口相传的推荐
  - 集中反映社会关系的影响是否存在
- 社会推荐的影响
  - 传统讨论只关注先验影响
    - 相比于无推荐场景，用户更可能接受某一产品？研究已证实
  - 对兴趣的贡献是后验影响
    - 相比于无推荐场景，用户更可能对某一产品满意？缺乏相关研究
- 研究意义
  - 准确估计由于社会关系影响导致的用户兴趣

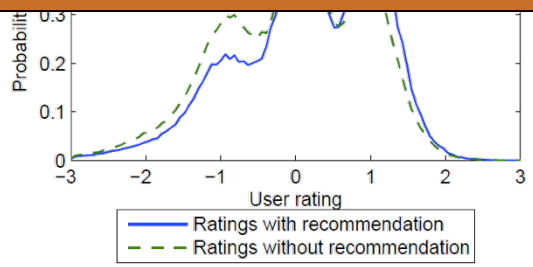
J. Huang, X.-Q. Cheng, H.-W. Shen, T. Zhou, X. Jin, Exploring Social Influence via Posterior Effect of Word-of-Mouth Recommendations, WSDM'12

社会推荐的影响

- 直观认识：后验影响不存在
  - 用户评分取决于个人观感，似乎与朋友推荐无关
- 实证发现：后验影响可能存在

问题：后验影响是否存在？即社会推荐是否直接影响用户后验评价？

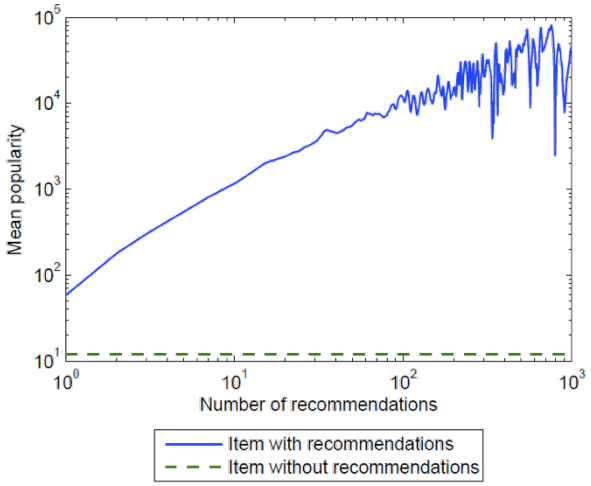


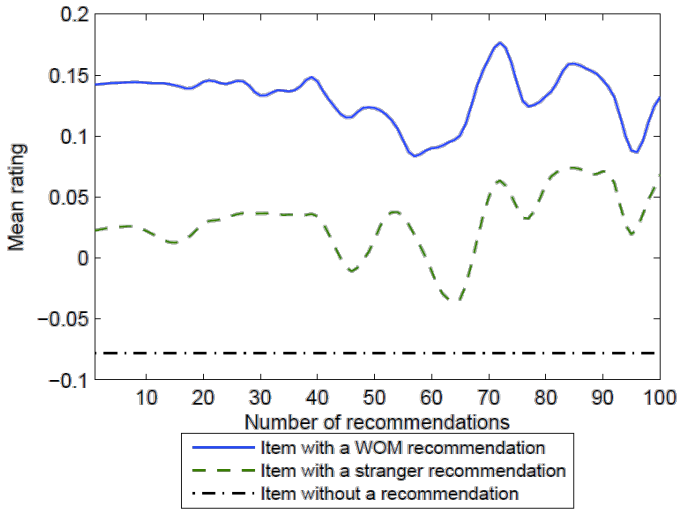


J. Huang, X.-Q. Cheng, H.-W. Shen, T. Zhou, X. Jin, Exploring Social Influence via Posterior Effect of Word-of-Mouth Recommendations, WSDM'12

社会推荐的影响

- 实证发现结果的两种可能解释
  - (1) 好的电影自然容易受到推荐(相关性)
    - 相关性有一定实证支持，但不足以完全解释实证现象

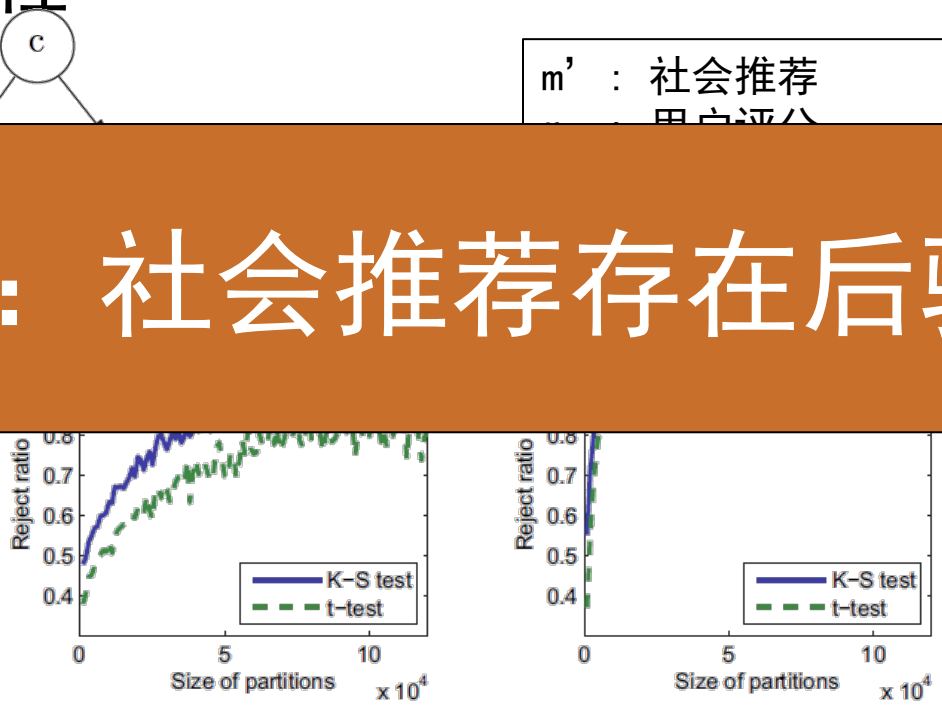




  - (2) 社会推荐直接影响用户对电影的评价(因果性)
    - 需要统计假设检验

社会推荐的影响

- 统计假设检验
  - 设计统计实验以区分“因果性”和“相关性”



结论：社会推荐存在后验影响

社会推荐的影响：量化与探索

- 后验影响力体现了推荐者的社会影响
  - 让接受推荐者更容易喜爱被推荐产品
- 什么样的人更具有社会影响？
  - 量化影响力

$$f_{u,v} = \frac{\sum_{i \in I_{uv}} \hat{x}_{v,i}}{|I_{uv}|} \qquad x_{u,i} = U^{+}_{u,i} - U^{-}_{u,i}$$

x：评分提升　　*U*：用户评价　　*I*：推荐集　　*f*：社会影响力

  - 寻找高影响力人群的特征
    - 社会属性：拥有大量关注者
    - 自身属性：有能力影响关注者

用户兴趣估计小结

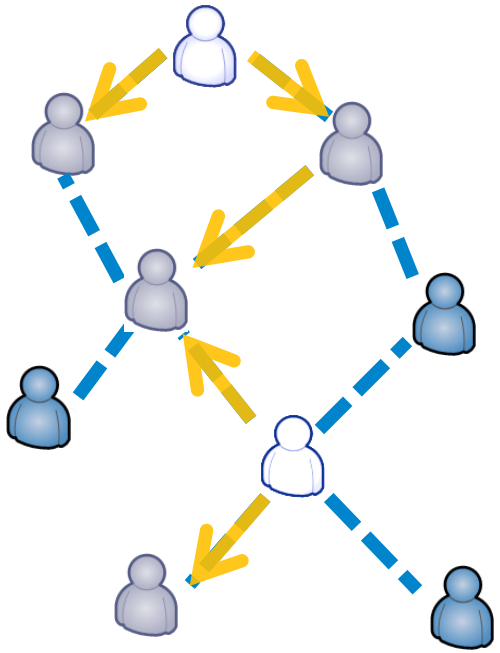
- 准确估计用户兴趣具有重要价值
- 用户兴趣包括自身固有兴趣和朋友影响的兴趣
  - 利用社交关系估计用户固有兴趣可提高准确度
- 朋友影响可以产生新的兴趣
- 定量测定朋友影响的兴趣，并发现高影响力人群的特征

社交网络上信息传播规律未知，如何探索传播行为并预测传播范围？

信息传播范围预测

社交网络上的信息传播

- 信息通过社交关系逐级传播
- 社交网络成为新兴媒体
  - Twitter、新浪微博等
- 能否在传播早期预测信息传播的范围？
  - 帮助用户发现热门话题
  - 识别早期谣言
  - 预估营销效果



信息传播范围预测

- 问题
  - 如何理解社交网络中信息传播行为机理
  - 如何准确预测信息传播的规模和范围
  - 如何促进或控制信息传播
- 价值
  - 建立高效的信息分发系统
  - 帮助商业系统提供信息服务并设计营销策略
  - 帮助监管和引导信息

信息传播范围预测

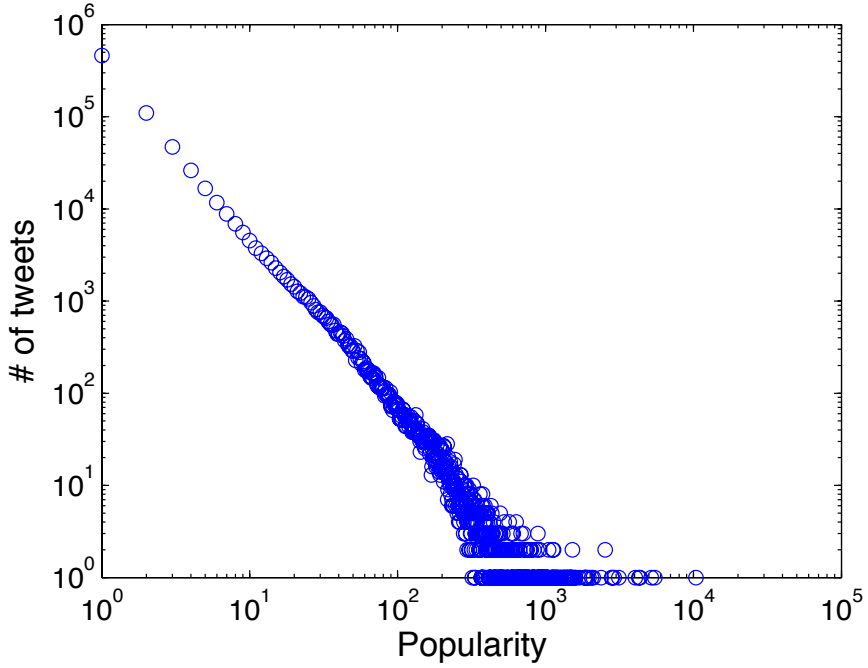
- 在信息传播的早期（即信息传播尚未完成时）准确预测单条信息通过传播到达的人群规模
  - 挖掘与信息传播规模相关的因素
  - 设计预测算法以准确地预测扩散范围

信息传播范围相关因素

- 传播范围实证研究
  - 转发树的深度大多小于6，转发树的结点规模服从幂律分布[Kwak 2010]
  - 推荐扩散网络呈现星形结构[Leskovec 2007b]
  - 弱连接对新信息的扩散起到了重要的作用[Bakshy 2012]
  - 规则网络中比在随机网络中扩散得快[Lü 2011]
  - 个体传播概率不是受该个体的接触邻居个数决定，而且受其接触邻居的连通分支个数决定[Ugander 2012]

信息传播范围难以预测

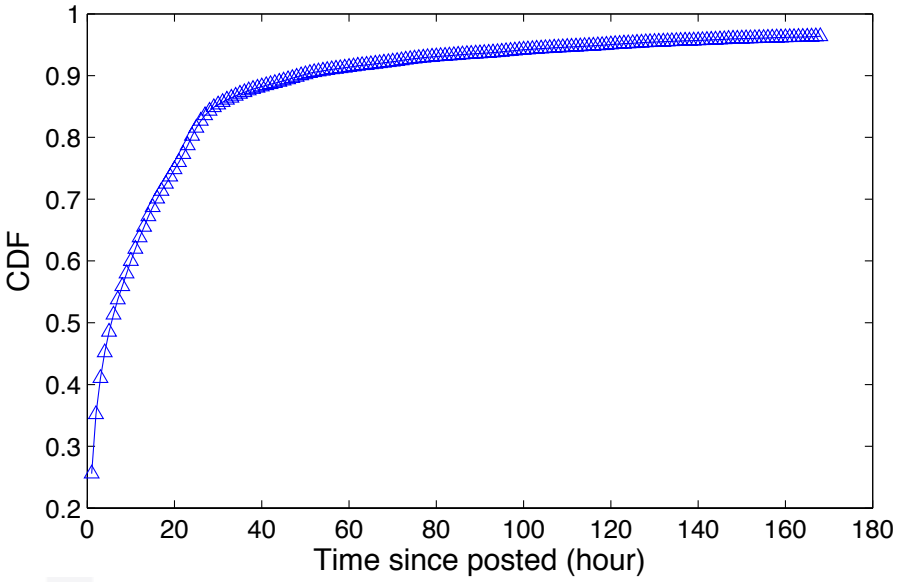
(1) 信息严重异质化，绝大部分信息传播范围小，极少信息传播范围极广

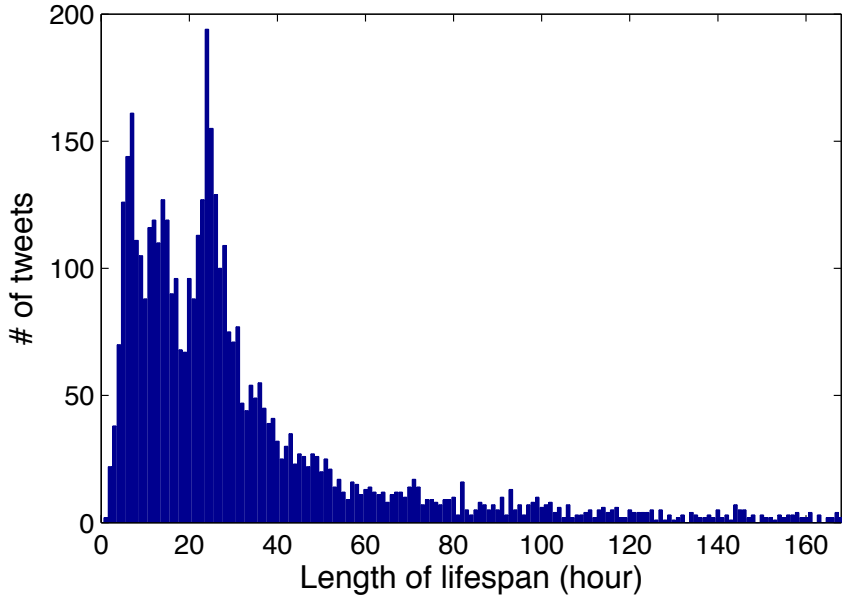


P. Bao, H.-W. Shen, J. Huang, X.-Q. Cheng, Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo, WWW'13

信息传播范围难以预测

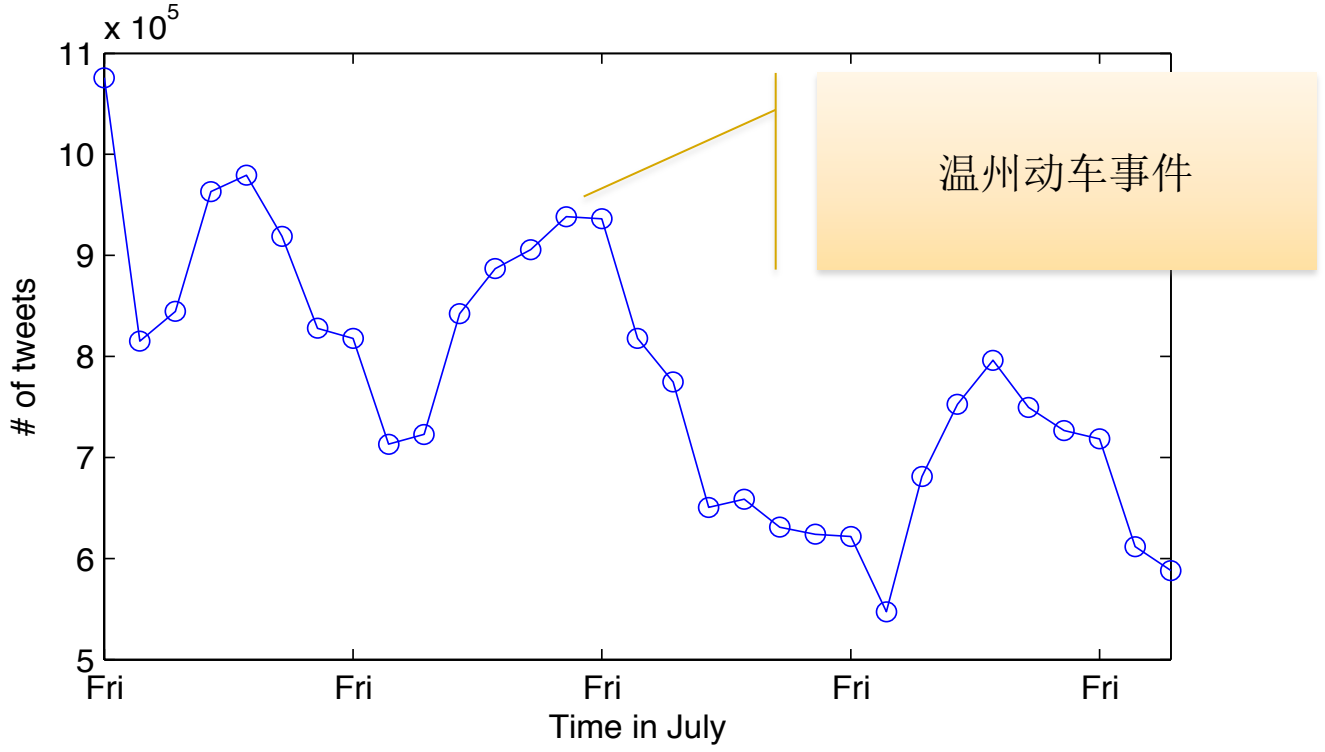
(2) 信息更新速度快，平均一条微博80%的转发发生在最初24小时





信息传播范围难以预测

(3) 很多高转发微博来自缺乏规律的突发事件



信息传播范围难以预测

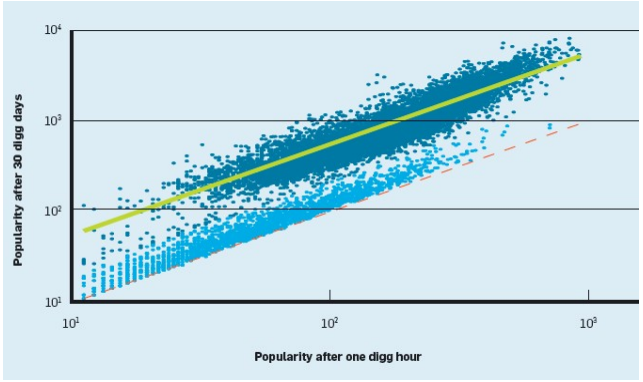
(4) 内容价值难以度量

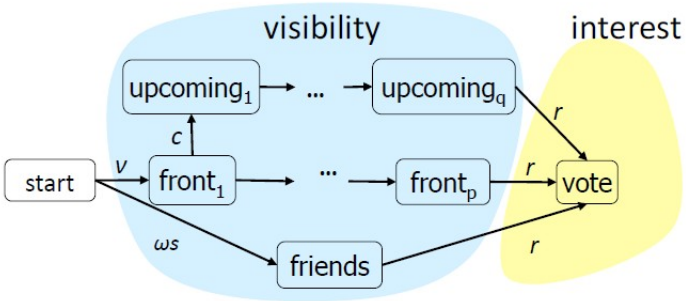
(5) 外部因素(如传统媒体)的影响难以探测

(6) 用户内在周期性和无序性

信息传播范围预测算法

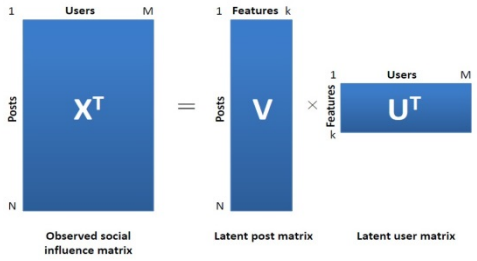
- 时序相关性[SzaBo 2010]
  - 早期传播规模与最终传播规模线性相关
- 可见性和内容“有趣”程度[Lerman 2010]
  - 用户行为模型
  - 估计内容的“有趣”程度

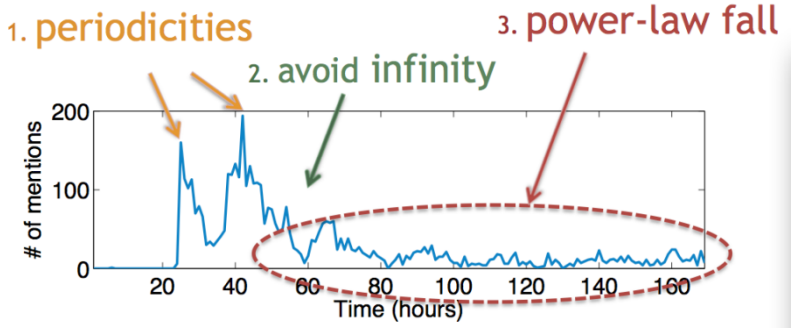




信息传播范围预测算法

- 基于矩阵分解[Cui 2011]
  - 传播行为因子分解为用户向量与内容向量的内积
- 基于特征空间[Hong 2011]
  - 形式化为经典分类问题
- 基于时序模型[Matsubara 2012]
  - Periodical;Avoid infinity;Power-law decay





现有算法的不足

- 现有算法均不考虑网络结构因素
  - 社交网络结构对信息传播影响大
  - 稀疏网络与稠密网络的传播效果不同
  - 单一人群(例如仅在同学圈传播)与多个人群的传播效果不同

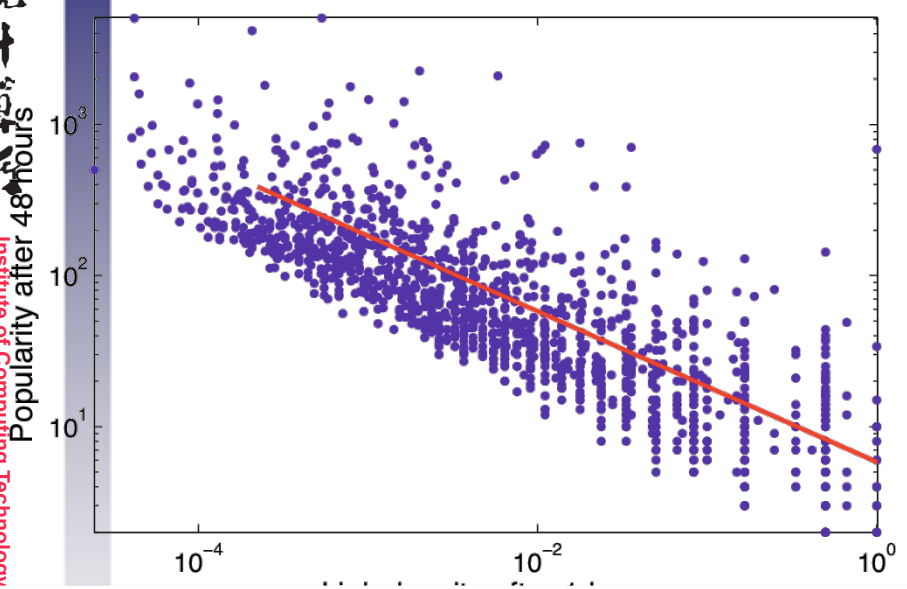
利用结构多样性预测信息传播范围

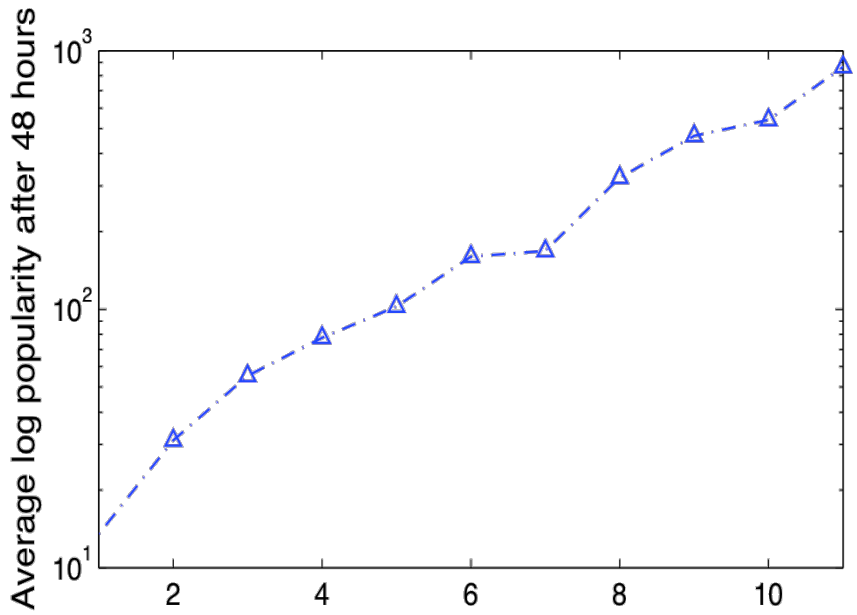
- 考察社交网络上早期转发者构成的子网
  - 一条微博的最初转发用户之间的相互关注关系
- 定义结构特征
  - 连边密度
  - 传播深度：源节点到子网任意节点的最长路径
- 直观认识
  - 若初始传播囿于较小的圈子，其邻居重合度较高，不利于后期大范围扩散。反之，连边密度小，传播深度浅，这样的初始传播圈子具有多样性，信息在早期即扩散到多个圈子，预计后期扩散范围大。

P. Bao, H.-W. Shen, J. Huang, X.-Q. Cheng, Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo, WWW'13

利用结构多样性预测信息传播范围

- 分析结构多样性与信息传播范围的相关性





早期传播者的子网结构特征对
信息最终传播范围有强指示作用

P. Bao, H.-W. Shen, J. Huang, X.-Q. Cheng, Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo, WWW'13

利用结构多样性预测信息传播范围

- 建立预测模型
  - 基于结构特征建立回归模型
  - 预测单条微博转发量的对数
- 实验验证
  - 以RMSE和MAE评价预测准确度
  - 相对不使用结构特征的现有算法有显著改进

Table : Prediction error of three approaches.

Primitive type	RMSE	MAE
Baseline	0.77	0.57
with link density	0.63	0.45
with diffusion depth	**0.61**	**0.43**

信息传播范围预测小结

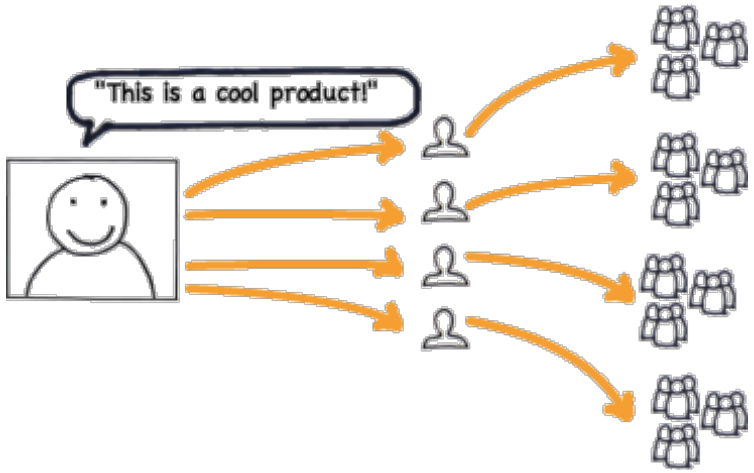
- 早期预测社交网络上的信息传播范围具有重要意义
- 信息传播影响因素多，难度较大
- 现有算法不考虑结构特征，难以精确求解
- 我们建模早期传播者的结构特征，显著提升预测精度
- 我们发现结构多样性有利于信息传播
  - 较早传播到多样化人群中的微博有较高的传播潜力

若已知社交网络上的信息传播模型，如何选择作为信息源的初始节点以最大化信息传播范围？

影响力最大化

病毒营销

- 某商家计划免费送出50份试用品，希望试用者使用后向亲朋宣传产品体验，这些朋友可能继续向他们的朋友宣传该产品。为了让尽可能多的人了解产品，商家应如何选择50位试用者？
- 假设传播符合经典扩散模型
  - 线性阈值模型
  - 独立级联模型



影响力最大化

- 给定信息传播模型，寻找k个初始用户作为信息源，使得信息扩散范围最大 [Kempe 2003]
  - 在独立级联模型和线性阈值模型上均为NP-hard
  - 信息扩散范围具有单调性和子模性，贪婪

如何设计算法快速准确地寻找最优的初始用户集

影响力最大化算法

- 基于随机抽样的算法
  - 贪婪算法 [Kempe 2003]
    - 从空集开始搜索初始节点集
    - 每轮寻找一个节点加入初始节点集，该节点最大化初始节点集的扩散范围
    - 待选节点集的扩散范围通过对社交网络上的信息传播过程的蒙特卡罗抽样来估计
  - CELF [Leskovec 2007]
    - 利用函数子模性降低候选节点数目
    - 速度提高近700倍
  - CELF++ [Goyal 2011a]
    - 利用单轮蒙特卡罗模拟同时估计两组集合
  - NewGreedy [Chen 2009]
    - 每轮使用同一批蒙特卡罗抽样
- 基于拓扑结构的算法
  - DegreeDiscount [Chen 2009]
    - 依据节点的一阶邻居进行影响力规模估计并惩罚overlap节点
  - PMIA [Chen 2010a]
    - 忽略传播概率较小的路径
    - 强可扩展性，解性能良好且更加稳定，解精度缺乏保障，内存开销增加
  - LDAG [Chen 2010b]
    - 对每个节点计算局部DAG以估计影响力范围
  - SIMPATH [Goyal 2011b]
    - 利用在节点集的邻近区域枚举简单路径的方法估计节点集影响力

影响力最大化算法的问题

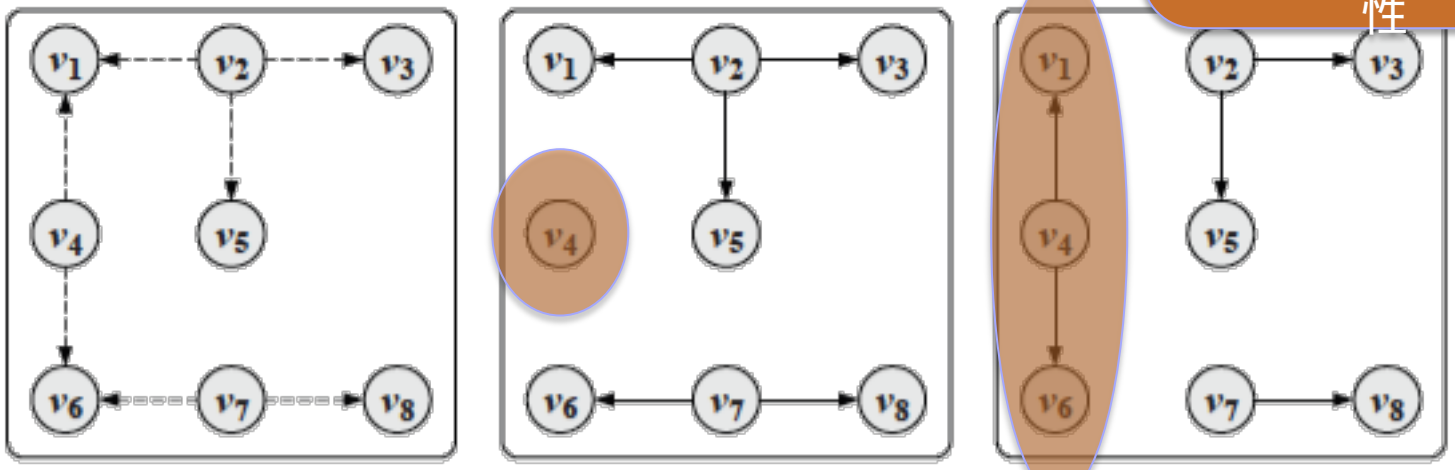
	基于随机抽样的算法	基于拓扑结构的算法
优势	精确 通过蒙特卡洛模拟精确估计扩散范围，保证获得不劣于最优解(1-1/e)性能的次优解	快速 只关注静态拓扑结构，计算速度快
劣势	慢速 需要大量蒙特卡罗模拟计算，时间开销大	不精确 对信息扩散范围估计不准，求解精度没有保证

能否设计算法同时获得高速度与高精度？

设计快速准确的影响力最大化算法

- 在随机抽样算法的基础上降低计算需求
  - 现有算法不保证子模性
    - 每一轮使用全新的蒙特卡罗抽样
    - 需要大量抽样以概率保证子模性

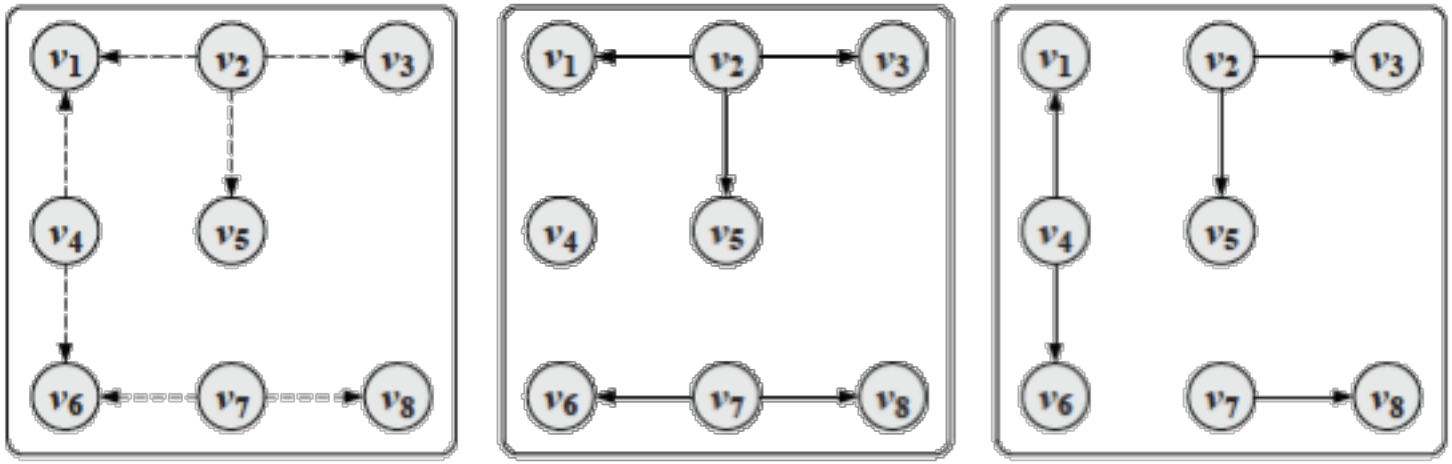
首轮v_4边际效应为1个节点(v_4)，次轮v_4边际效应为2个节点(v_4,v_6)，不满足边际效应递减的子模性



原网络　　首轮蒙特卡罗　　次轮蒙特卡罗

设计快速准确的影响力最大化算法

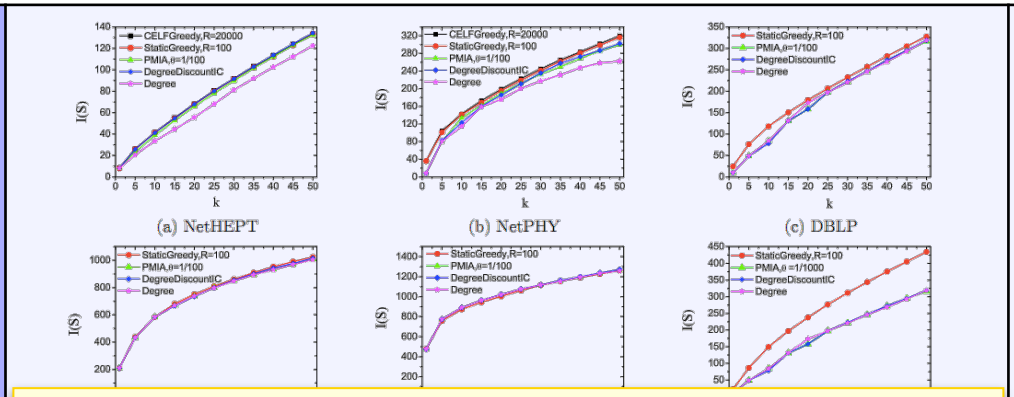
- 在随机抽样算法的基础上降低计算需求
  - 我们提出静态抽样算法StaticGreedy保证子模性
    - 每轮均使用同一组蒙特卡罗快照
    - 不必通过大规模抽样保证子模性，大大降低计算开销



原网络　　首轮蒙特卡罗　　次轮蒙特卡罗

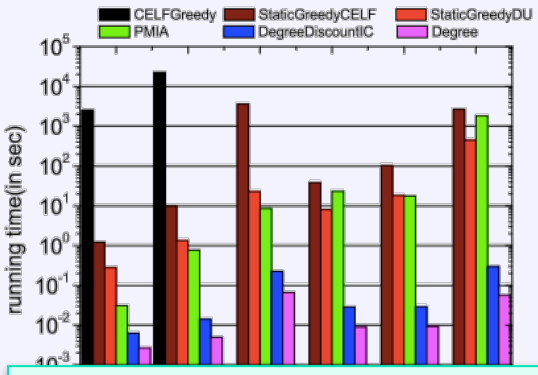
实验结果：兼具速度和精度优势

UIC



StaticGreedy的精度保证

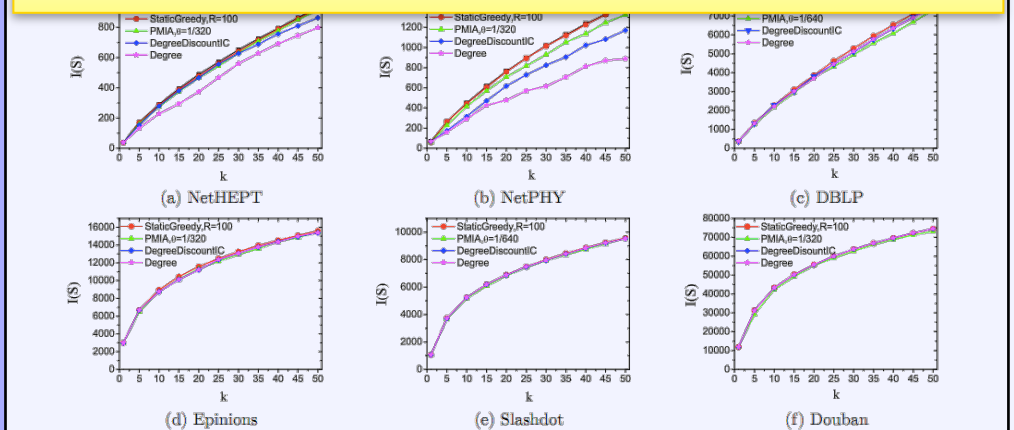
- 在NetHEPT, NetPHY上，StaticGreedy的精度始终与Greedy一致
- 在其它数据集中精度始终优于其它算法，或与它们一致

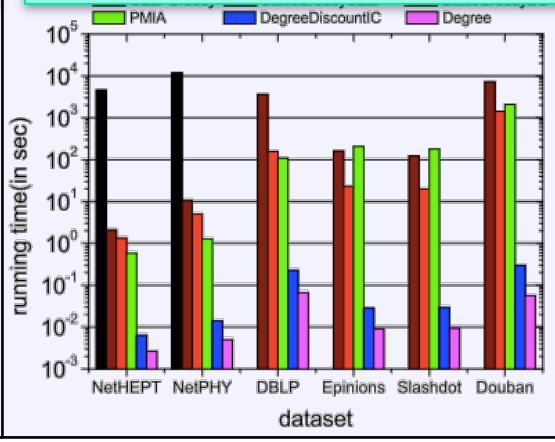


StaticGreedy的计算开销

- 与目前最好的启发式算法PMIA相当
- 在大规模数据集上表现甚至更优

WIC





影响力最大化小结

- 已知社交网络和传播模型，寻找最大化传播范围的初始节点
- 现有算法遭遇计算精度和速度的两难困境
- 我们提出静态抽样算法同时达到高精度和高速度

人类行为分析和预测任重道远

更多开放问题有待解决

量化问题：实证分析的难点

- 用户兴趣估计
  - 用户内在兴趣与社会关系影响的交互作用
  - 推荐系统对用户兴趣的引导和反作用
  - 推荐系统对社交网络的影响
  - 随时间变化的影响力
- 信息传播范围预测
  - 预测信息传播的具体轨迹
  - 信息传播的干预策略
- 影响力最大化
  - 多个信息的竞争性传播
  - 限制影响力与网络结构的鲁棒性

谢谢！

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences