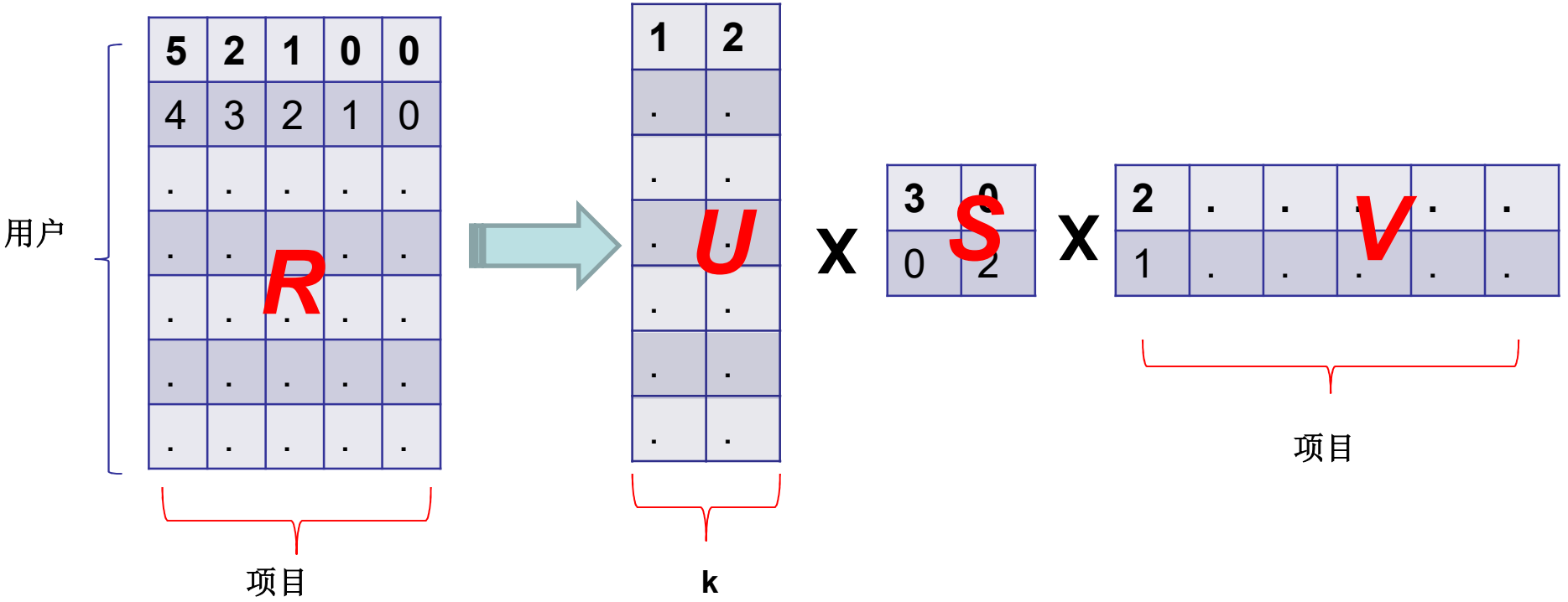
数据分析及实践
Analysis and Practice of the Data

第五章 推荐算法与社交网络

刘 淇
Email: qiliuql@ustc.edu.cn

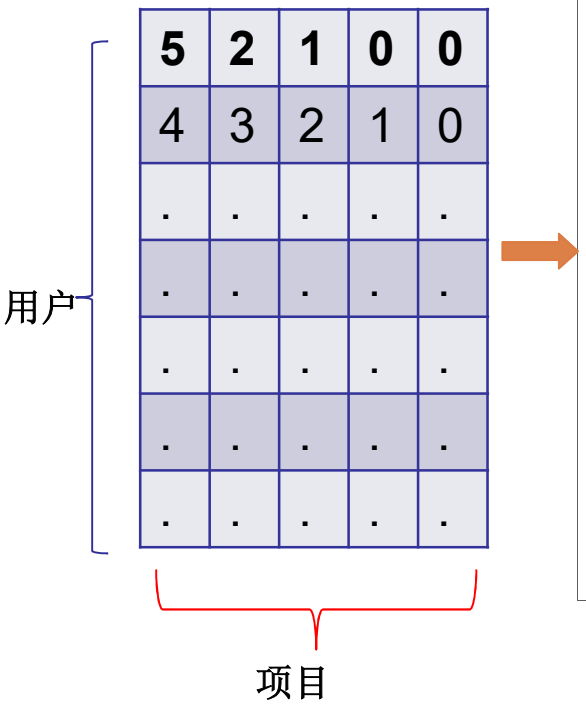
推荐算法设计

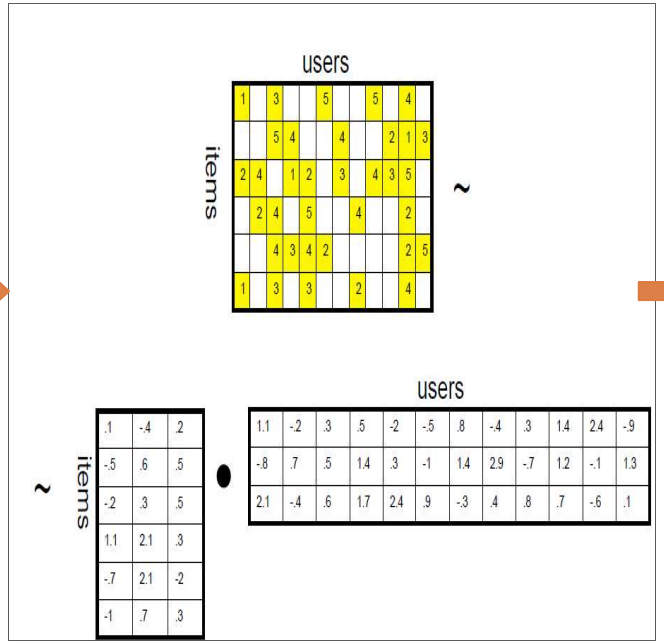
- 基于矩阵分解的协同过滤算法
  - 面向评分预测的模型
    - 奇异值分解(SVD)



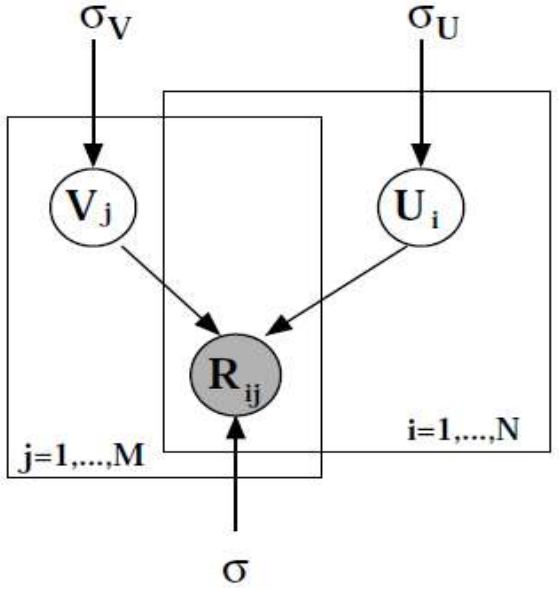
评分预测算法设计

- 基于矩阵分解的协同过滤算法
  - 面向评分预测的模型





基本分解思想



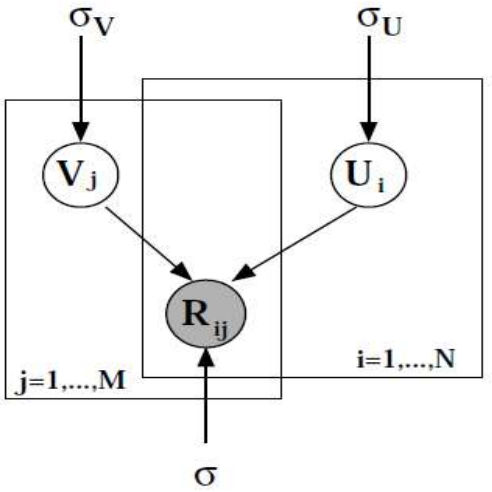
PMF（概率矩阵分解）

Probabilistic Matrix Factorization

评分预测算法设计

- 基于矩阵分解的协同过滤算法
  - PMF Solution
    - 目标函数

$$p(R|U,V,\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^TV_j,\sigma^2)\right]^{I_{ij}}$$



  - How to get U and V? --MAP
    - The log-posterior of user and item features over fix parameters

$$p(V|\sigma_V^2)=\prod_{j=1}^{M}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})$$

$$p(U|\sigma_U^2)=\prod_{i=1}^{N}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$$

$$p(U,V\mid R,\sigma^2,\sigma_U^2,\sigma_V^2)$$

$$\propto p(R\mid U,V,\sigma^2)*p(U\mid\sigma_U^2)*p(V\mid\sigma_V^2)$$

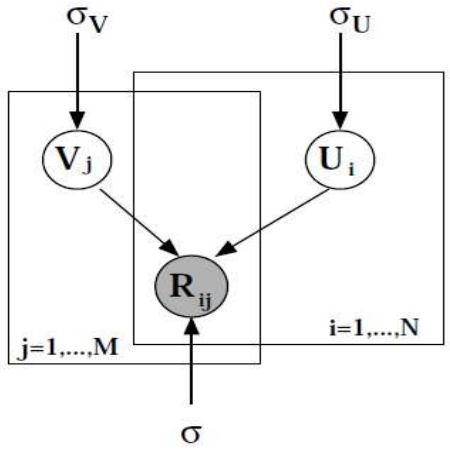
Likelihood!

Prior

评分预测算法设计

- 概率矩阵分解
  - MAP learning

$$\ln p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij}-U_i^T V_j)^2 - \frac{1}{2\sigma_U^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{M} V_j^T V_j - \frac{1}{2}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\right)\ln\sigma^2 + ND\ln\sigma_U^2 + MD\ln\sigma_V^2\right) + C, \quad (3)$$



  - Equivalent to minimize sum-of-squared-errors with quadratic regularization terms.

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij}-U_i^T V_j\right)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2$$

$$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$$

regularization term to avoid over fitting

评分预测算法设计

概率矩阵分解

$$E=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-U_i^TV_j\right)^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2+\frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2$$

1)Intialize U ,V with small, random values

2)repeat

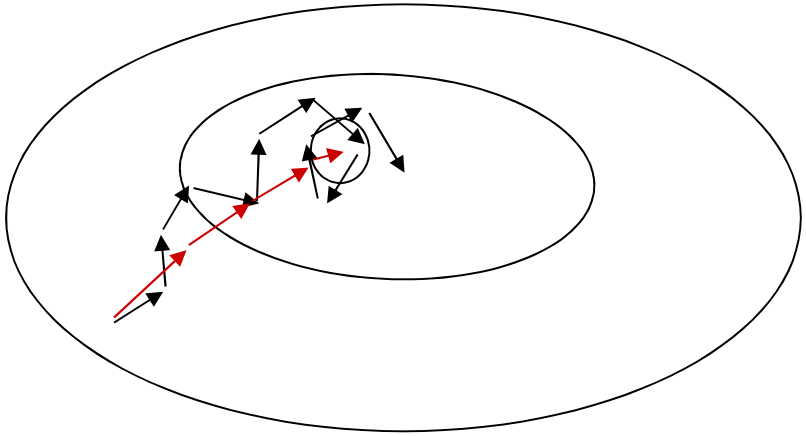
for each record in the training data

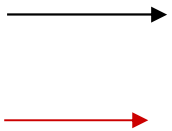
2.a) $U_i=U_i-a\frac{\partial E}{\partial U_i}=U_i-a\left(\sum_j I_{ij}(R_{ui}-U_i^TV_j)(-V_j)+\lambda_U U_i\right)$

2.b) $V_j=V_j-a\frac{\partial E}{\partial V_j}=V_j-a\left(\sum_i I_{ij}(R_{ui}-U_i^TV_j)(-U_i)+\lambda_V V_j\right)$

until convergence

Optimize: stochastic gradient descent





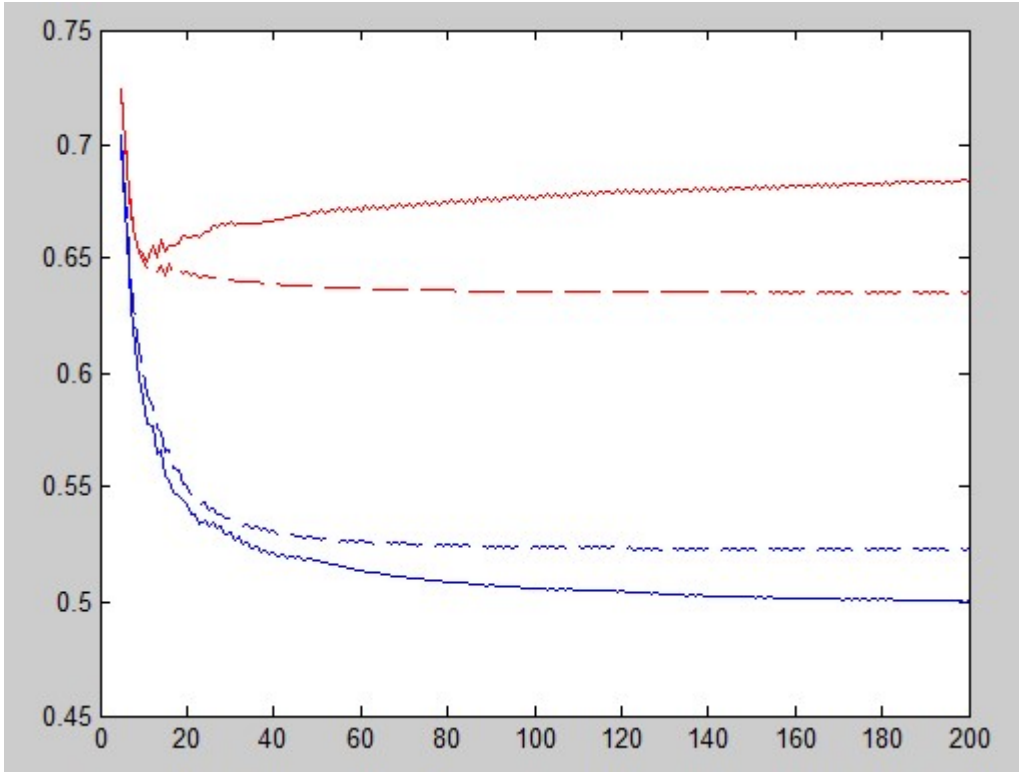
stochastic updates

full updates (averaged over all data-items)

评分预测算法设计

- 概率矩阵分解测试

Using Movielens dataset



SVD train_error

SVD test_error

PMF train_error

PMF test_error

排序预测算法设计

- 基于排序学习的协同过滤
  - 无显式的评分，只有隐式的购买行为，如何产生推荐？
    - 用一个模型预测产品之间的排序关系，把排在前边的产品推荐给用户----BPR模型
      - 低秩分解的时候去拟合项目pair的排序关系

$$\hat{x}_{ui} = \langle w_u, h_i \rangle = \sum_{f=1}^{k} w_{uf} \cdot h_{if}$$

$$\hat{x}_{uij} := \hat{x}_{ui} - \hat{x}_{uj}$$

目标函数：$p(i >_u j|\Theta) := \sigma(\hat{x}_{uij}(\Theta))$

where σ is the logistic sigmoid:

$$\sigma(x) := \frac{1}{1 + e^{-x}}$$

- 然后用分解的结果去预测未知的项目pair排序关系，由此生成排序序列

	A	B	C	D	E
U1	1	0	1	0	0
U2	1	0	0	1	1
用户	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.

X

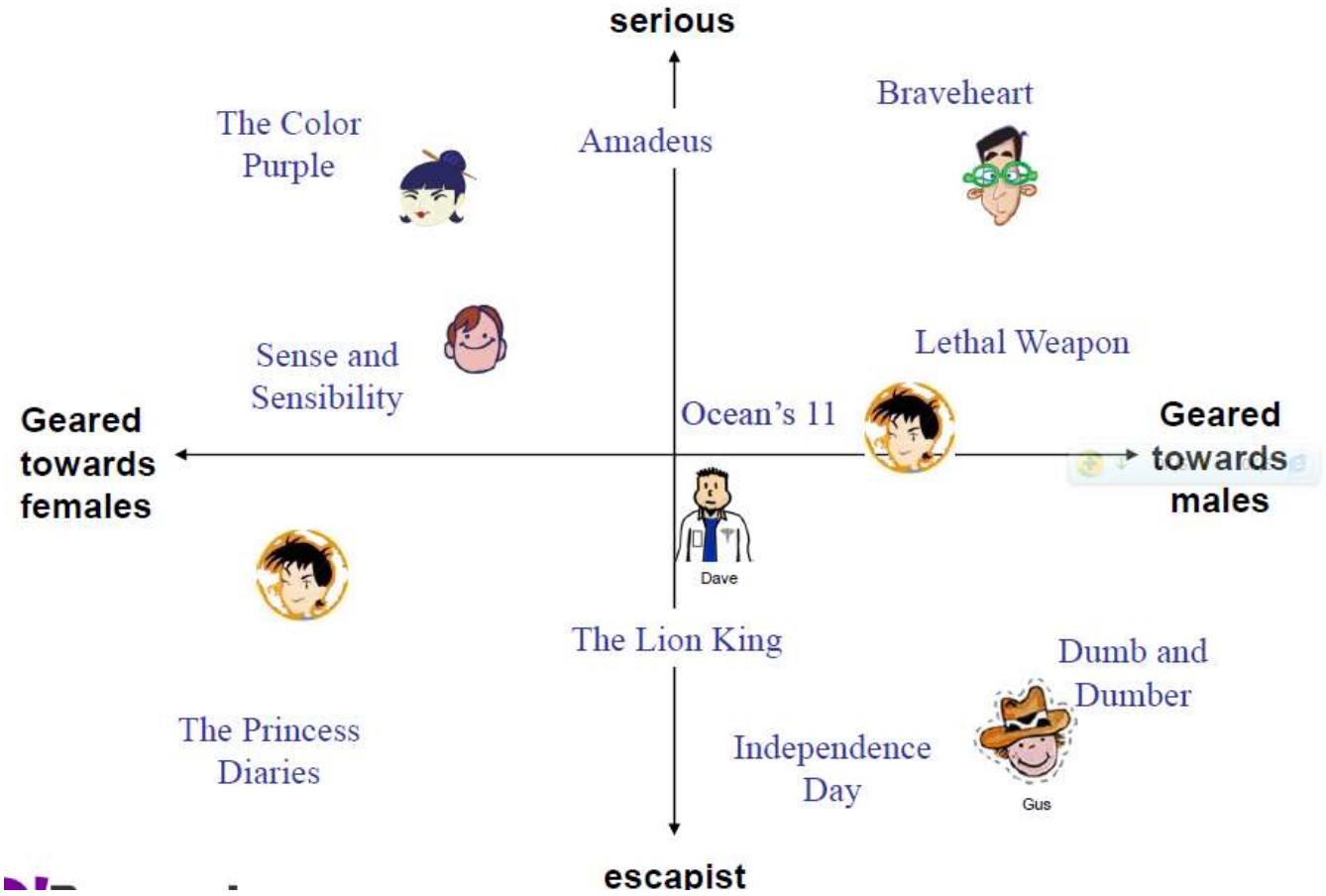
项目

项目pair的排序关系

(U1,A,B)	$A >_{U1} B$
(U1,A,D)	$A >_{U1} D$
(U1,A,E)	$A >_{U1} E$
(U1,C,B)	$C >_{U1} B$
(U1,C,D)	$C >_{U1} D$
(U1,C,E)	$C >_{U1} E$
(U2,A,B)	$A >_{U2} B$
.....	

推荐算法设计

- 基于矩阵分解的协同过滤算法



推荐算法设计

- 工程应用中的混合推荐（如点击率预测）方法
  - 成千上百万特征的提取和建模
    - Logistic Regression (LR)
    - Factorization Machines(FM、Field-aware FM)
    - GBDT(LR+GBDT)--Gradient Boosting Decision Tree

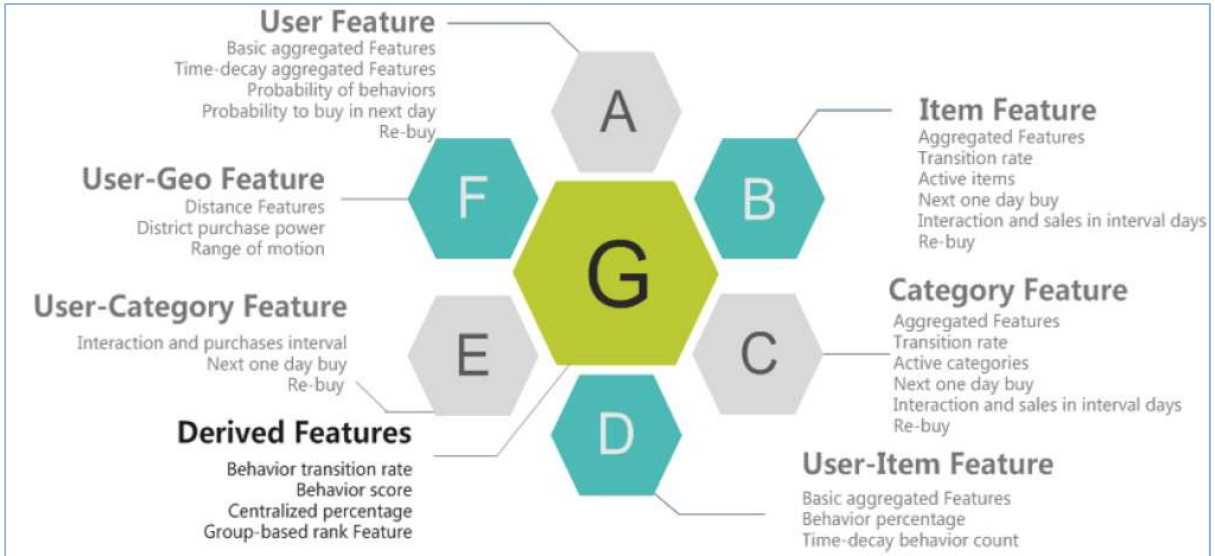


Fig. 1. The overall features architecture

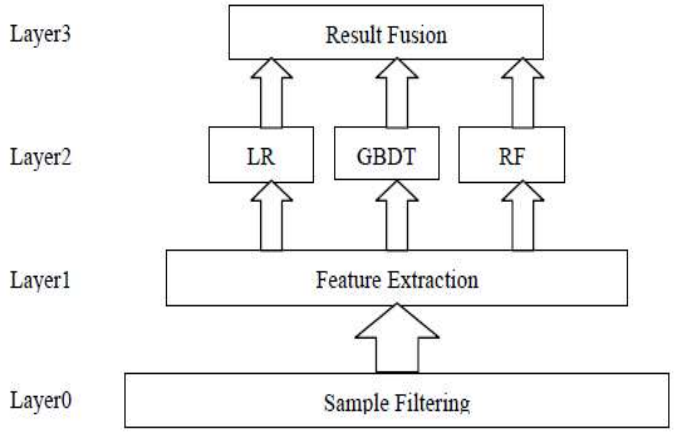
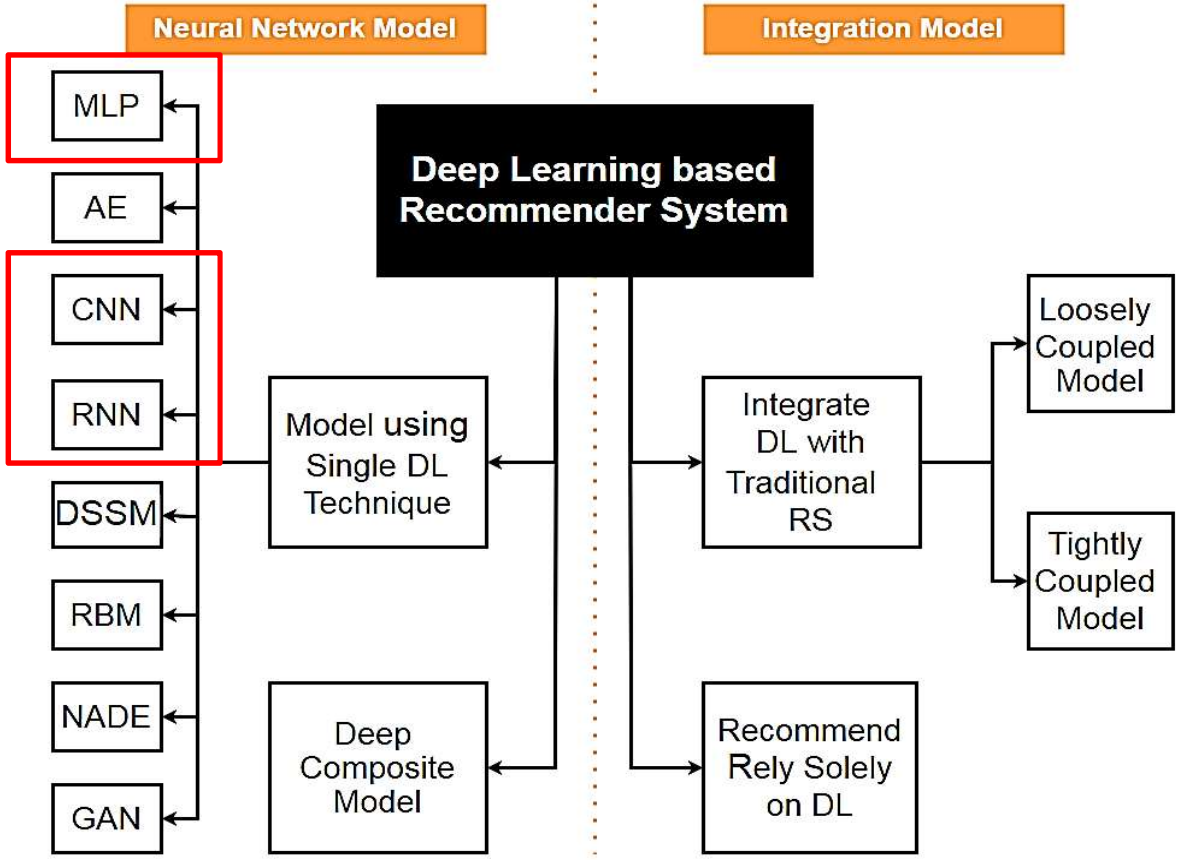


Fig. 1. Overview of our hierarchical model.

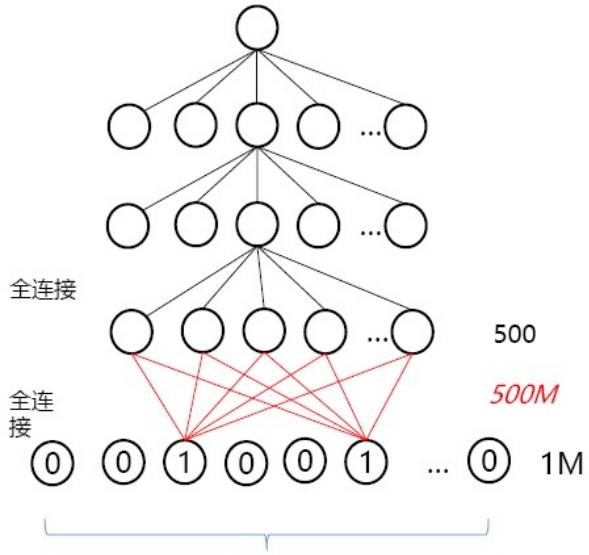
推荐算法设计

- 基于深度学习的推荐算法
  - 框架： 直接用深度学习来实现推荐功能　　集成DL到推荐系统中



推荐算法设计

- 基于深度学习的推荐算法
  - 细粒度、一致的特征表达
  - 可以建模非线性关系
    - 例如，用户-item之间，特征之间

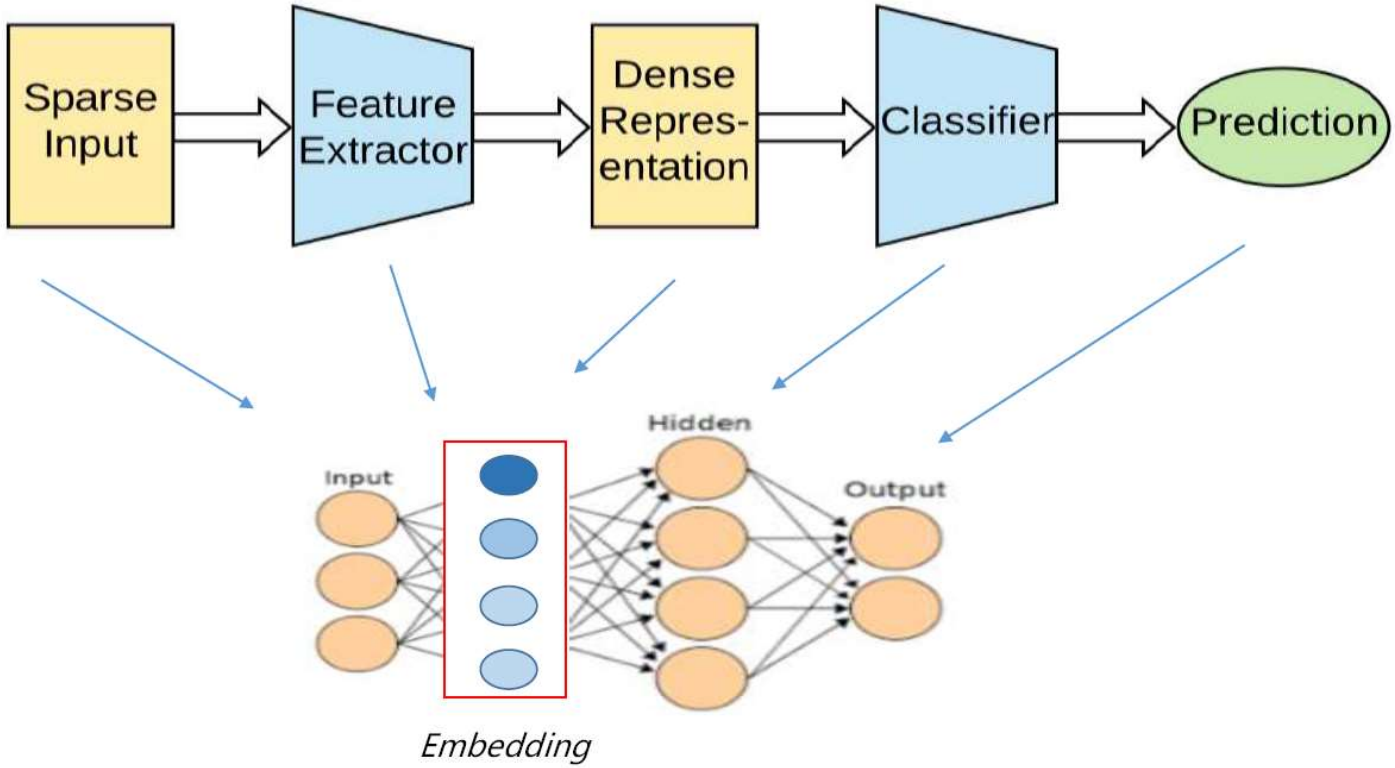


1. Embedding

2. Interaction （非线性关系的建模）

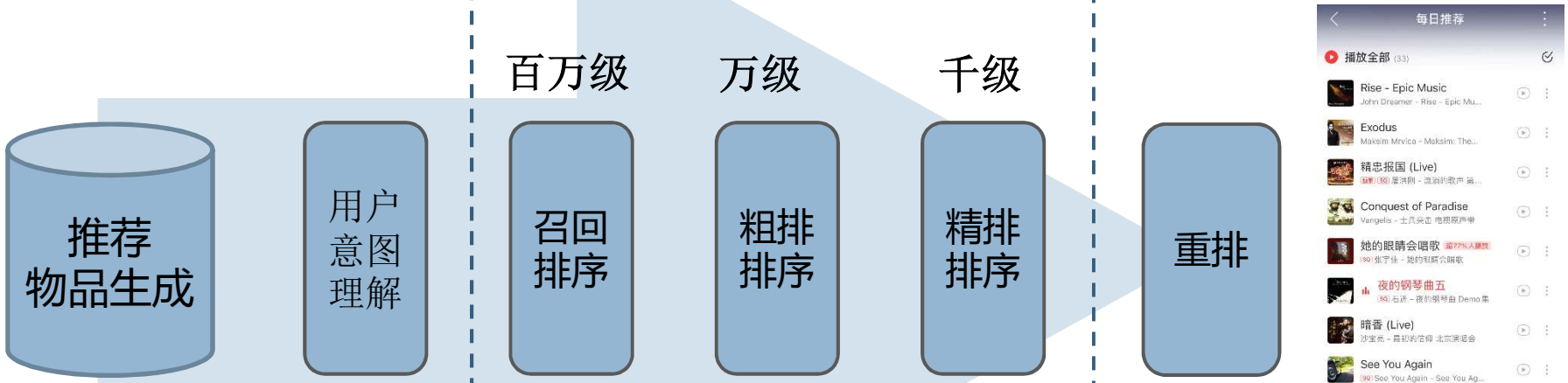
推荐算法设计

- 基于深度学习的推荐算法
  - 基本架构：Embedding+MLP



《深度学习推荐算法实践与挑战》何秀强

推荐系统架构



- 产生推荐内容池
- 用户意图理解

- 召回（候选集生成）
  - 多路召回
- 粗排（候选集排序）
  - 线性模型
- 精排（对Top-N结果排序）
  - DNN

- 重排
  - 多样性
  - 产品体验
  - 运营策略

推荐系统前沿研究方向

- 热门研究话题
  - Fairness in Recommender System
  - Explanation in Recommender System
  - Adversarial Attacks in Recommender System
  - Debias in Recommender System
  - Evaluation
- 结合前沿技术
  - Pre-training in Recommender Systems
  - AutoML in Recommender Systems
  - Federated Learning in Recommender System
  - Reinforcement Learning in Recommender System

部分参考论文

- Wu, Le, et al. "A Survey on Neural Recommendation: From Collaborative Filtering to Content and Context Enriched Recommendation." arXiv preprint arXiv:2104.13030 (2021).
- Zhang, Weinan, et al. "Deep Learning for Click-Through Rate Estimation." arXiv preprint arXiv:2104.10584 (2021).
- Guo, Qingyu, et al. "A survey on knowledge graph-based recommender systems." IEEE Transactions on Knowledge and Data Engineering (2020).
- Chen, Shuo, and Thorsten Joachims. "Modeling intransitivity in matchup and comparison data." Proceedings of the ninth acm international conference on web search and data mining. ACM, **2016**.
- Chen, S., & Joachims, T. (**2016**, August). Predicting Matchups and Preferences in Context. In KDD
- Ma, Hao, et al. "Sorec: social recommendation using probabilistic matrix factorization." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.
- Martin Ester，Recommendation in Social Networks，Tutorial at RecSys 2013
- Zhepeng (Lionel) Li, Xiao Fang, and Olivia R. Liu Sheng. **2017**. A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions. ACM Trans. Manage. Inf. Syst. 9, 1, Article 1 (October 2017), 26 pages. DOI: https://doi.org/10.1145/3131782
- **Zhang S, Yao L, Sun A. Deep Learning based Recommender System: A Survey and New Perspectives. arXiv:1707.07435**
- Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. **2017**. Recurrent recommender networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 495–503.

目录

社交网络概述

- 社交网络
  - 即社交网络服务（SNS，Social Networking Services），是指以一定社会关系或共同兴趣为纽带、以各种形式为在线聚合的用户提供沟通、交互服务的互联网应用。这种以人与人关系为核心的方式建立的社会关系网络映射在互联网上就形成了以用户为中心、以人为本的互联网应用。

社交网络概述

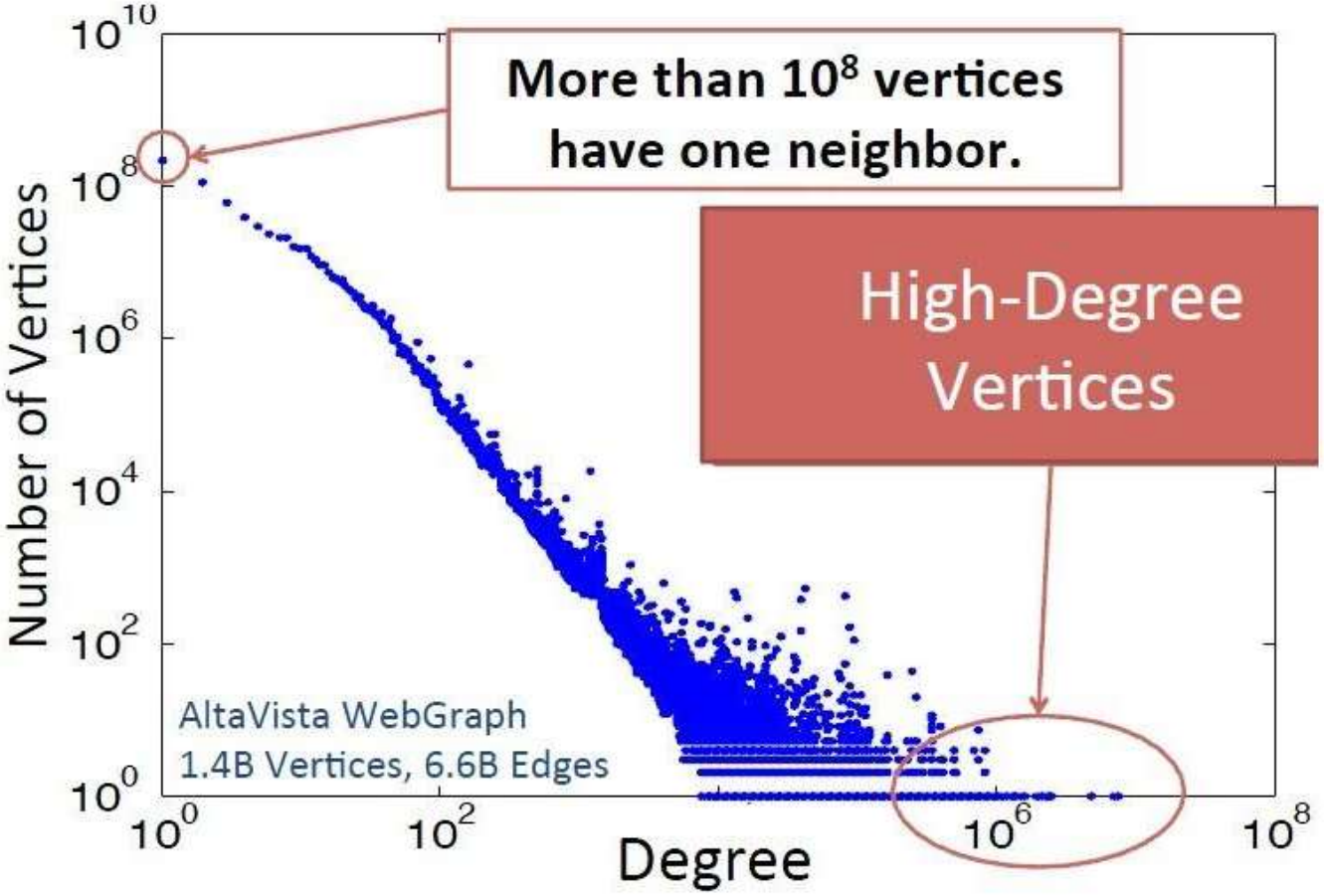
社交网络概述

- 社交网络特点
  - 开放、以用户为中心
  - 以社会关系或共同兴趣为纽带
  - 内容多元化
    - 文本、图片、视频、表情
  - 信息即时更新与传播迅速
  - 互动性
    - 评论
    - 回复
    - 转发

社交网络概述

- 社交网络性质
  - 幂律分布：对数空间下呈现出线性关系

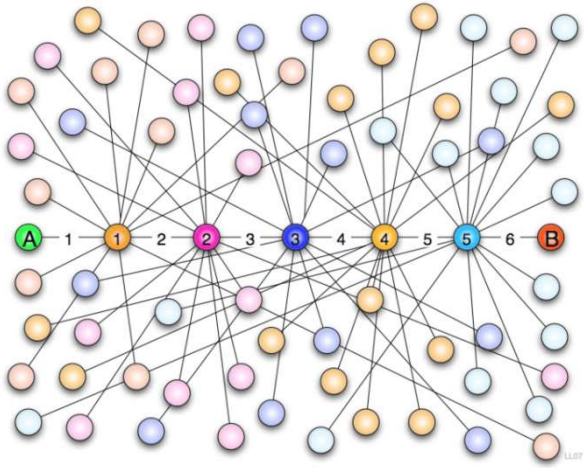
$$y = cx^{-r}$$



社交网络概述

- 小世界现象 Small World phenomenon
  - 小世界现象指世界上的每个人之间都可以通过很短的社会关系链联系起来。小世界网络，引伸了小世界现象，不仅是社会关系网络，可以指任何网络。
  - 小世界网络就是对这种现象（也称为小世界现象）的数学描述。用数学中图论的语言来说，小世界网络就是一个由大量顶点构成的图，其中任意两点之间的平均路径长度比顶点数量小得多。

- 六度分隔理论
  - 六度分隔理论: 假设世界上所有互不相识的人只需要很少中间人就能建立起联系。
  - 后来1967年哈佛大学的心理学教授斯坦利·米尔格拉姆根据这概念做过一次连锁信实验，尝试证明平均只需要5个中间人就可以联系任何两个互不相识的美国人，见右图。



目录

结点重要性

- 社交网络结点重要性
  - 用户影响力排名
  - 用户活跃度排名
  - 热门话题排名
  - 热门关键字排名

- 一般网络中的重要性评估
  - 网站排名或网页排名
    - 根据一些客观的、公正的衡量标准，对网页的重要性或权威性进行排序如网页按访问流量排名，按权威性排名（PageRank）等

结点重要性

微信排名

微信传播力指数：wci指数

- 总阅读数R、总点赞数Z、发布文章数N、该账号当期最高阅读数Rmax、该账号最高点赞数Zmax。
- 采用指标：总阅读数R、平均阅读数R/N，最高阅读数Rmax，总点赞数Z，平均阅读数Z/N 最高点赞数Zmax 点赞率Z/R。

WCI（100%）						
WCI（R）(65%)			WCI（Z）(35%)			
R≥1000	R/N	Rmax	Z	Z/N	Zmax	Z/R
60%	30%	10%	60%	30%		
WCI（100%）						
WCI（R）(65%)			WCI			
R<1000	R/N	Rmax	Z	Z/N		
60%	30%	10%	60%	30%		

（当总阅读数 R<1000 时，不考虑 Z/R 这一指标，Zmax 权重调为 10%）

WCI计算公式：

$$WCI=\begin{cases}[65\%*ln(60\%*R+30\%*\frac{R}{N}*5+10\%*R_{max}*3)+35\%*\ln(60\%*Z+30\%*\frac{Z}{N}*4.5 \\ +5\%*Z_{max}*2+5\%*\frac{Z}{R}*1000)]^2*10 & if(1000\le R) \\ [65\%*ln(60\%*R+30\%*\frac{R}{N}*5+10\%*R_{max}*3)+35\%*\ln(60\%*Z+30\%*\frac{Z}{N}*4.5 \\ +10\%*Z_{max}*2)]^2*10 & if(R<1000)\end{cases}$$

ln（X=0）或 lnX<0 时取值为 0。

最优的发微信策略是什么？

5.03-5.09 全国省级团组织微信公号综合影响力 排行榜

中国青年报出品 微信公号 zqbcyol

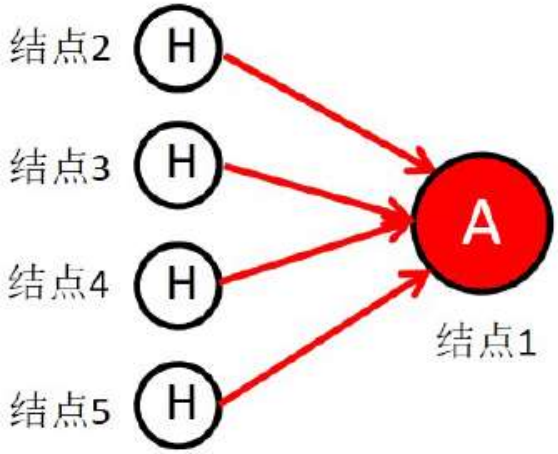
		发布	阅读	头条	平均	点赞	WCI
	ngsh	7/14	317307	171692	22665	2893	1153
	lth	7/14	187039	117170	13360	2024	1042
3	青春山东 sdgqt2014	5/25	198778	47743	7951	3609	1010
4	青春浙江 zhejiangtsw	7/19	172694	85054	9089	2127	998
5	江苏共青团 js_gqt	7/35	176379	70941	5039	1553	937
6	共青团陕西省委员会 sxgqtwx	6/29	115988	27748	4000	425	845
7	青春上海 qingchunshanghai	7/33	81177	29168	2460	563	791
8	津彩青春 tjtuanwei	5/15	33917	9253	2261	1739	787

HITS(Hyperlink-Induced Topic Search)

- HITS算法 康奈尔大学(Cornell University) 的Jon Kleinberg提出
  - 权威(Authority)度：网页被高质量枢纽页指向的能力
    - 页面所有入链对应网页的枢纽度之和
  - 枢纽(Hub)度:网页指向高质量权威页面的能力
    - 页面所有出链指向的页面的权威度之和

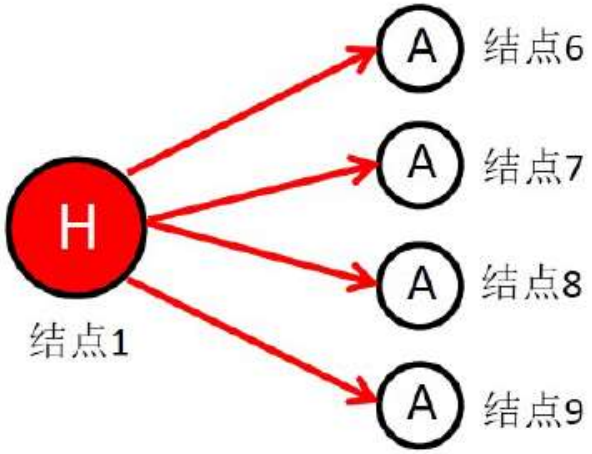
枢纽结点　　权威结点

$$A(1) = H(2)+H(3)+H(4)+H(5)$$



枢纽结点　　权威结点

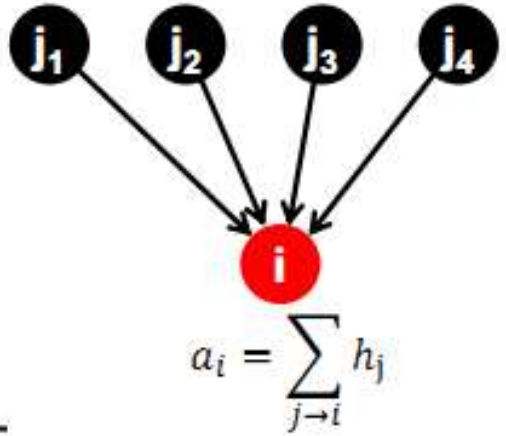
$$H(1) = A(6)+A(7)+A(8)+A(9)$$



迭代

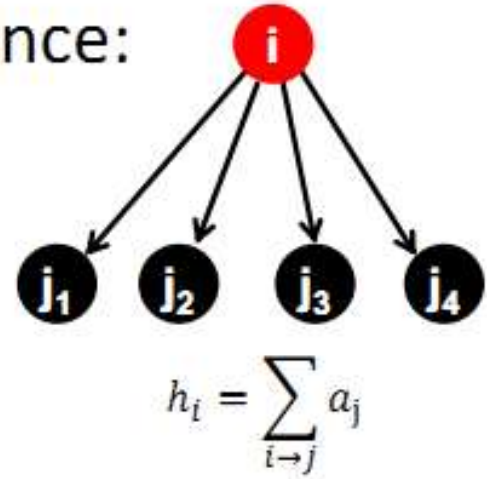
HIST

- **Each page *i* has 2 scores:**
  - Authority score: a_i
  - Hub score: h_i



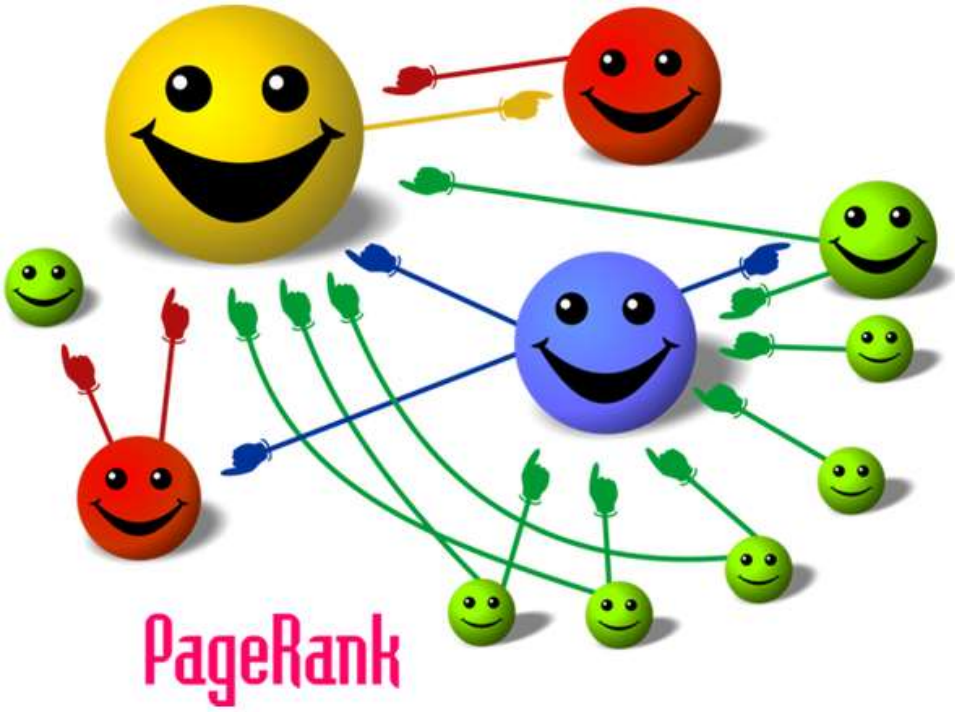
HITS algorithm:

- Initialize: $a_j = 1/\sqrt{n},\ h_i = 1/\sqrt{n}$
- Then keep iterating until convergence:
  - $\forall i$: Authority: $a_i = \sum_{j \to i} h_j$
  - $\forall i$: Hub: $h_i = \sum_{i \to j} a_j$
  - $\forall i$: Normalize:
    $\sum_i a_i^2 = 1, \sum_j h_j^2 = 1$



PageRank

- 数据挖掘十大经典算法（2006 ICDM）
  - C4.5
  - K-Means
  - SVM
  - Aprior
  - EM
  - PageRank
  - AdaBoost
  - KNN
  - Naive Bayes
  - CART



PageRank

1998，由Google的创始人拉里·佩奇和谢尔盖·布林提出

PageRank算法

- 为每个网页计算一个PageRank值，用来衡量网页的重要性
- 核心假设：
  - 如果指向一个网页的页面越多，则该网页的重要性越大。
  - 如果指向一个网页的页面越重要，则该网页的重要性也越大。
- 迭代计算
  - 输入：n个结点，节点i的出链个数 d_i
  - 输出：每一个结点i的重要性(PageRank值) r_i

(1) 初始化每个结点的重要性为1/n

(2) for t = 1, 2, 3, . . .

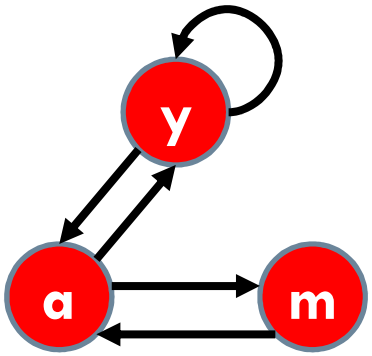
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

直到前后两次所PageRank值几乎不变为止

算法：PageRank

- 举例
  - 如右图的网络链接图，
  - 节点数n=3
  - 令d=0.85，PageRank公式为

$$r_j = 0.85 * \sum_{i \to j} \frac{r_i}{d_i}$$



	y	a	m
y	½	½	0
a	½	0	½
m	0	1	0

iteration / Node	1	2	3	4	…	…	…
y	1/3	1/3	5/12	3/8	…	….	**2/5**
a	1/3	1/2	1/3	11/24	…	…	**2/5**
m	1/3	1/6	1/4	1/6	…	…	**1/5**

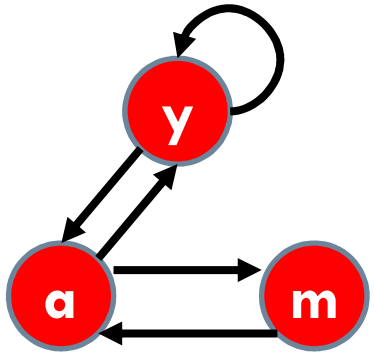
算法：PageRank

- 疑问

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

  - 公式会不会收敛
  - 收敛效果是不是我们想要的
  - 收敛结果是否合理

	y	a	m
y	½	½	0
a	½	0	½
m	0	1	0



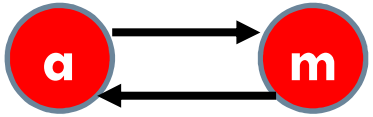
算法：PageRank

- 疑问

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

  - 公式会不会收敛
  - 收敛效果是不是我们想要的
  - 收敛结果是否合理

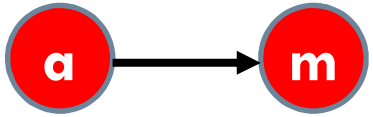
- 举例1



  - Ra = 1 0 1 0
  - Rm = 0 1 0 1

  iteration 0, 1, 2, ...

- 举例2

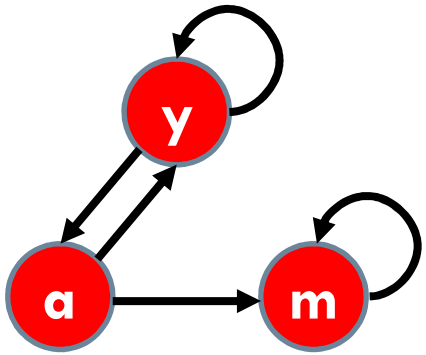


  - Ra = 1 0 0 0
  - Rm = 0 1 1 1

  iteration 0, 1, 2, ...

算法：PageRank

- 问题
  - Sprider traps
    - 一个节点的出边都指向自己
    - 一个节点集的出边都只指向这个节点集内的节点
    - 导致这个Spider traps吸收了所有的权威性

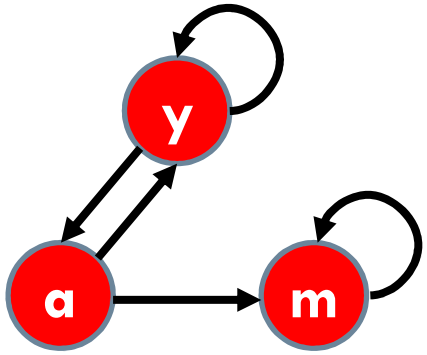


	y	a	m
y	½	½	0
a	½	0	½
m	0	0	1

问题：Spider traps

- 计算公式：

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$



	y	a	m
y	½	½	0
a	½	0	½
m	0	0	1

- 例子：

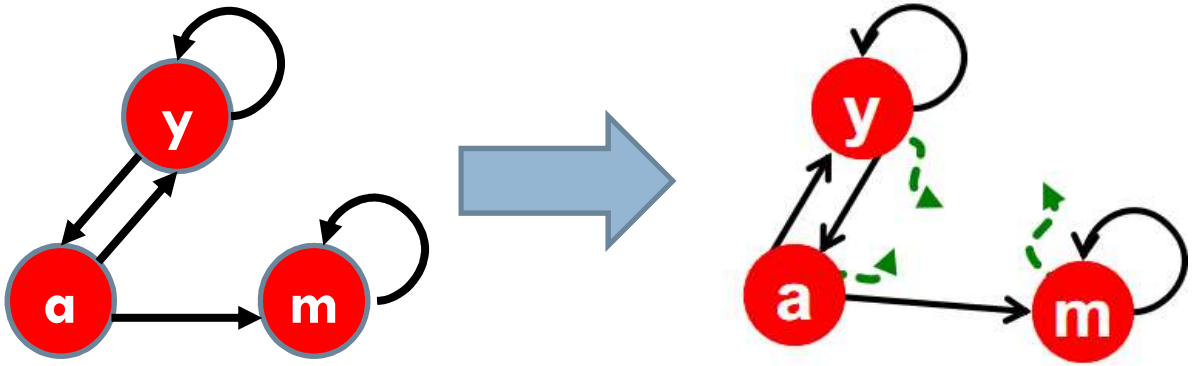
R_y = 1/3 2/6 3/12 5/24 0
R_a = 1/3 1/6 2/12 3/24 … 0
R_m = 1/3 3/6 7/12 16/24 ... 1
Iteration 0, 1, 2, …

算法：PageRank

- 随机游走模型——解决spider traps问题
  - 理解：用户在任意的网络结点开始游走到下一个结点时，有可能是点击了这个网页内的某个链接，也有可能是直接从地址栏里面输入了要给网页地址。
  - 因此一个网页被访问一部分来源于其他网页内指向它链接（概率为d），一部分来源于用户直接输入其网页地址的可能性（概率为1-d)。
  - 所以一个用户的权威值计算公式为：

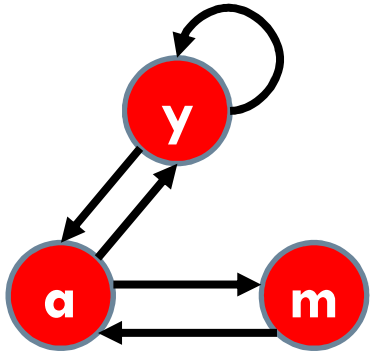
$$r_j = d * \sum_{i \to j} \frac{r_i}{d_i} + (1-d)\frac{1}{n}$$

  - 以上公式不断迭代，直到收敛。稳定状态下的各结点概率分布即为每个结点的权威值。



算法：PageRank

- 举例
  - 如右图的网络链接图，
  - 节点数n=3
  - 令d=0.85，PageRank公式为



	y	a	m
y	½	½	0
a	½	0	½
m	0	1	0

$$r_j = 0.85 * \sum_{i \to j} \frac{r_i}{d_i} + (1 - 0.85)\frac{1}{3}$$

iteration / Node	1	2	3	4	…	…	n
y	1/3	0.3333	0.3935	0.4006	…	….	**0.3817**
a	1/3	0. 4750	0.4313	0.4186	…	…	**0.3988**
m	1/3	0.2519	0.2333	0.2279	…	…	**0.2195**

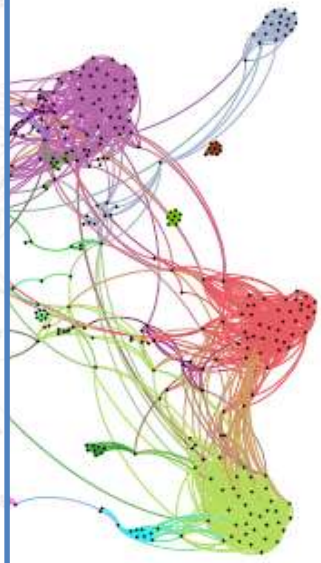
目录

- 社交网络概述
- 结点重要性
- 社团挖掘
- 社交影响力分析
- 信息源检测
- 符号网络
- 社交推荐
- 网络嵌入

社团挖掘

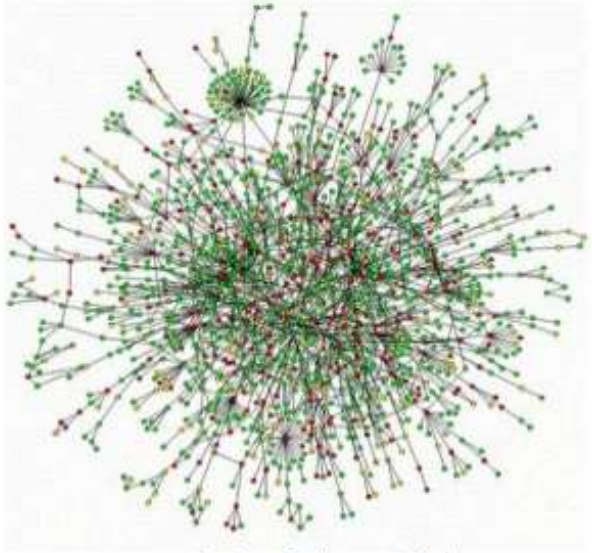
什么是社团(community)

- 尚无统一数学定义，普遍认为，社区是内部连接紧密、与外部连接的相对疏松的顶点集合
- 复杂网络三大特征：幂率分布、小世界、强社区结构

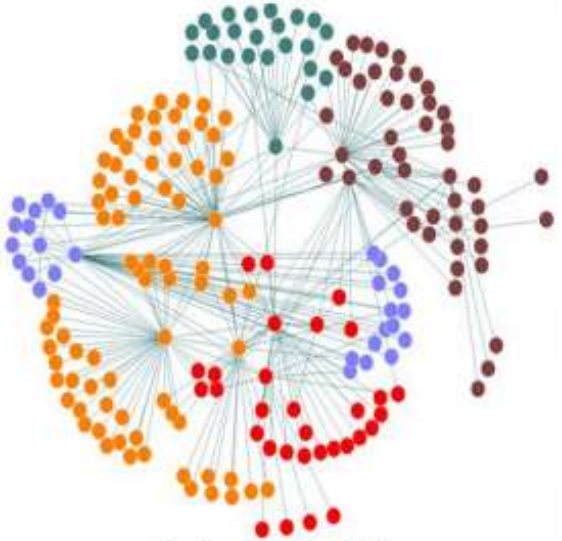


社团广泛存在

- 社会里有很多自然形成的社团：家庭，工作圈和朋友圈，城镇，国家。
- 因特网的发展导致很多虚拟在线社团的产生。
- 社团存在于很多网络系统中，生物学、计算机科学、工程、经济、政治等。



蛋白质交互网络

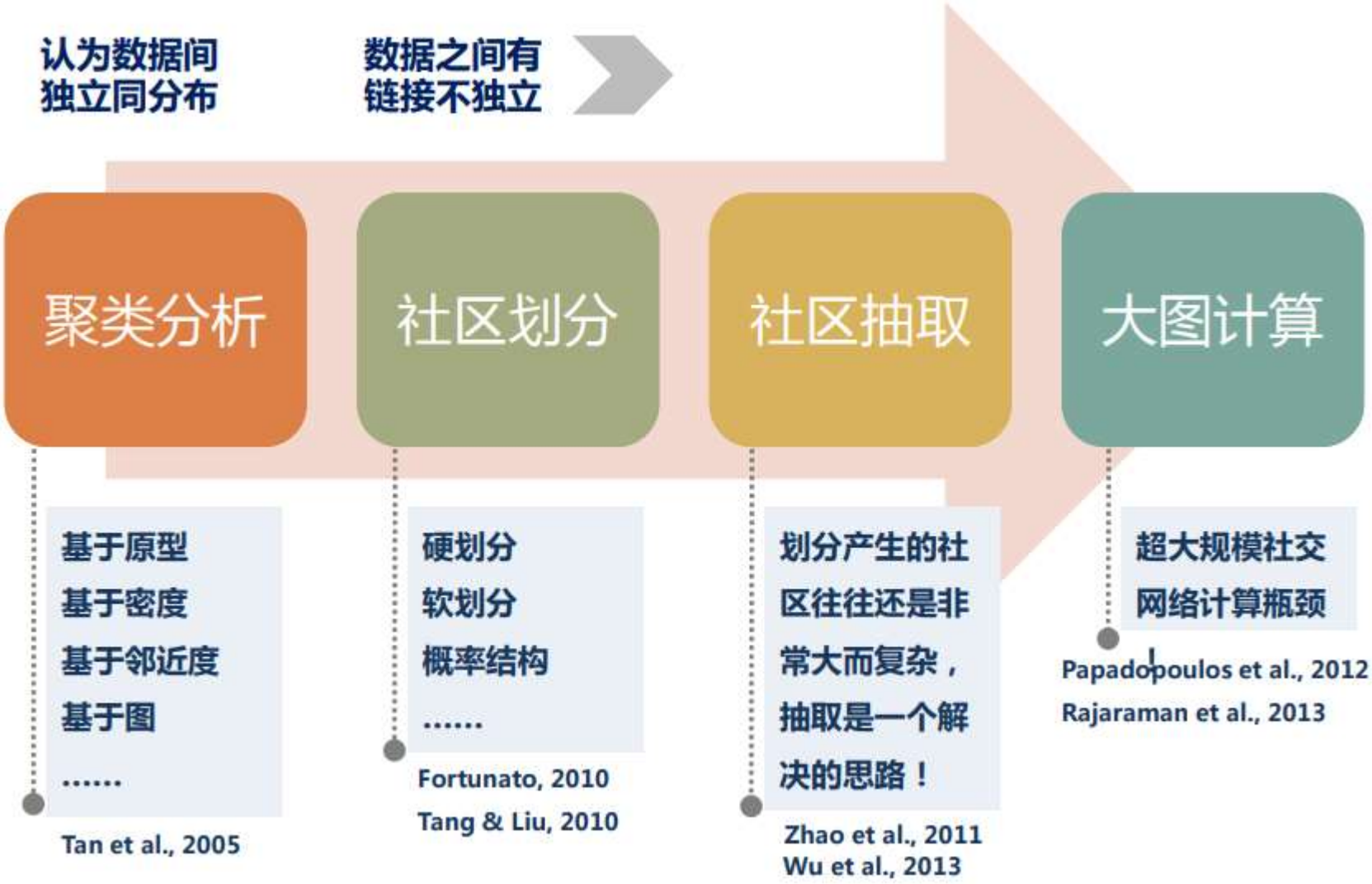


社会关系网络

社团挖掘的重要性

- 学术方面
  - 社团挖掘在社会网络分析、数据挖掘、统计学、机器学习、空间数据库技术、生物学和市场学等领域有广泛的应用。
- 实用价值
  - 商品的精准营销
  - SNS中好友的推荐
  - 公共安全和舆情控制
  - 控制疾病传播
  - ……

社团挖掘：技术发展脉络


认为数据间
独立同分布
数据之间有
链接不独立
聚类分析
社区划分
社区抽取
大图计算
基于原型
基于密度
基于邻近度
基于图
……
Tan et al., 2005
硬划分
软划分
概率结构
……
Fortunato, 2010
Tang & Liu, 2010
划分产生的社区往往还是非常大而复杂，抽取是一个解决的思路！
Zhao et al., 2011
Wu et al., 2013
超大规模社交网络计算瓶颈
Papadopoulos et al., 2012
Rajaraman et al., 2013

社团挖掘： Girvan-Newman算法

- 假设：社团之间所存在的少数几个连接应该是社团间通信的瓶颈，是社团间通信时通信流量的必经之路。如果我们考虑网络中某种形式的通信并且寻找到具有最高通信流量（比如最小路径条数）的边，该边就应该是连接不同社团的通道。
- 边介数(Edge Betweenness)
  - 一条边的边介数为任意两个节点间的最短路径通过该边的次数。在图G中，给定一条边e，假设图G中任意两个节点间的最短路径的集合为P，并且P中有m条最短路径经过边e，那么边e的边介数为m。

根据上面的假设，社团间的边有很高的边介数。

社团挖掘： Girvan-Newman算法

- Girvan-Newman算法的流程

1. 计算网络中每条边的边介数；
2. 删除边介数最高的边；
3. 重新计算所有边的边介数；
4. 重复第二步和第三步，直到所有的边都被删除。

自上而下

但是在不知道社团数目的情况下，该算法不能判断算法终止位置。因此Newman等人引入了模块度(Modularity)。

社团挖掘：模块度(Modularity)

- 模块度不仅仅作为优化的目标函数提出，它也是目前最流行的用来衡量社团划分质量的标准之一，它的提出被称作社团挖掘研究历史上的里程碑。
- 定义：所谓模块度(Modularity)是指网络中连接社团内部顶点的边所占的比例与另外一个随机网络中连接社团内部顶点的边所占比例的期望值相减得到的差值。

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\delta(C_i, C_j)$$

- 式中Q是模块度；A是网络图的邻接矩阵；P_{ij}是随机网络中节点i与节点j间边的条数；m是图的边数；C_i是节点i所隶属的社团；δ当C_i和C_j为同一社团时值为1，否则为0。

社团挖掘：模块度(Modularity)

- 通过层层推导，对无向图可以得到如下数学表达式

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right]$$

- 式中Q是模块度(Modularity)；
- m是图的边数；
- l_c是社团C中所有内部边的条数；
- d_c 是社团C中所有顶点的度之和；

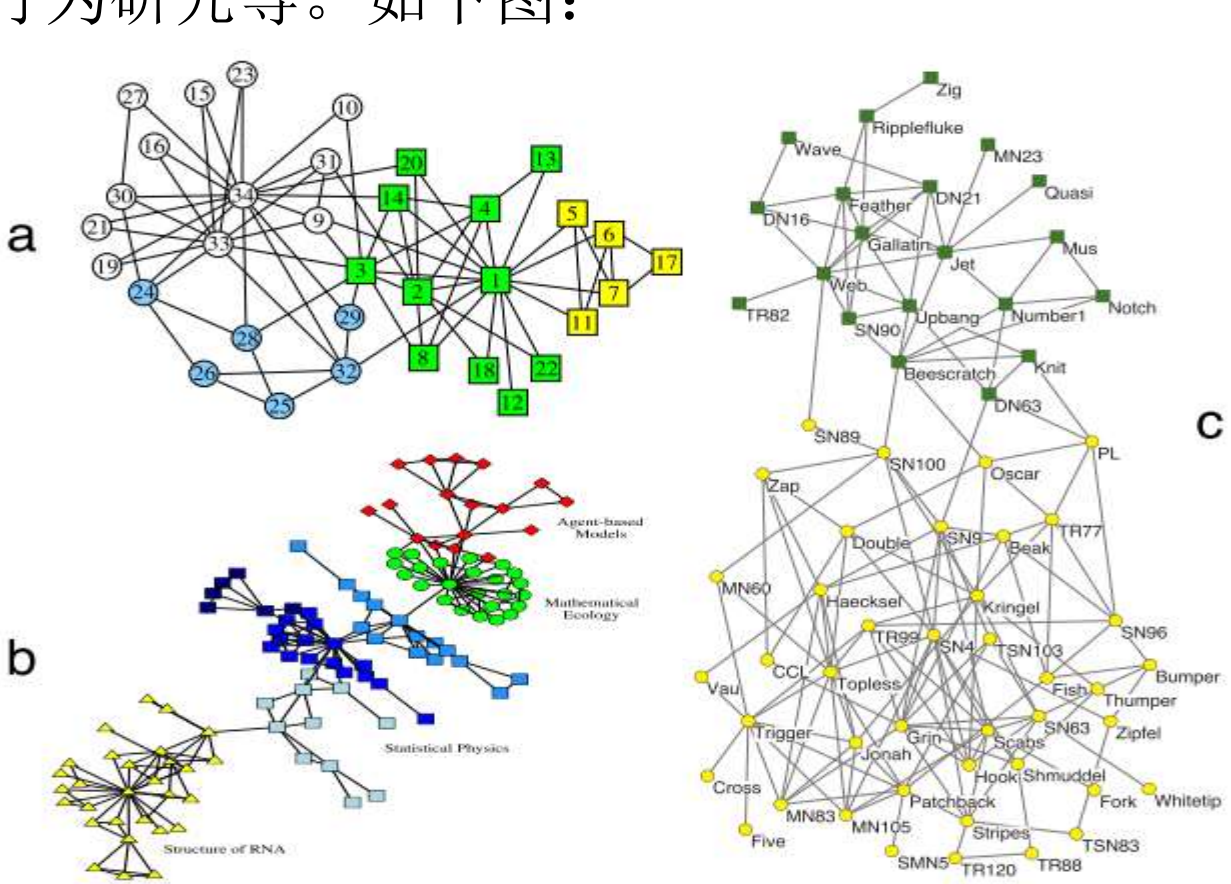
社团挖掘： Girvan-Newman算法

1. 计算网络中每条边的边介数；
2. 删除边介数最高的边，计算此时的模块度；
3. 重新计算所有边的边介数；
4. 重复第二步和第三步，直到所有的边都被删除。

- 当算法执行完毕，比较所有中间过程的模块度的值，取最大模块度的中间划分结果，作为社团挖掘最终的划分结果。

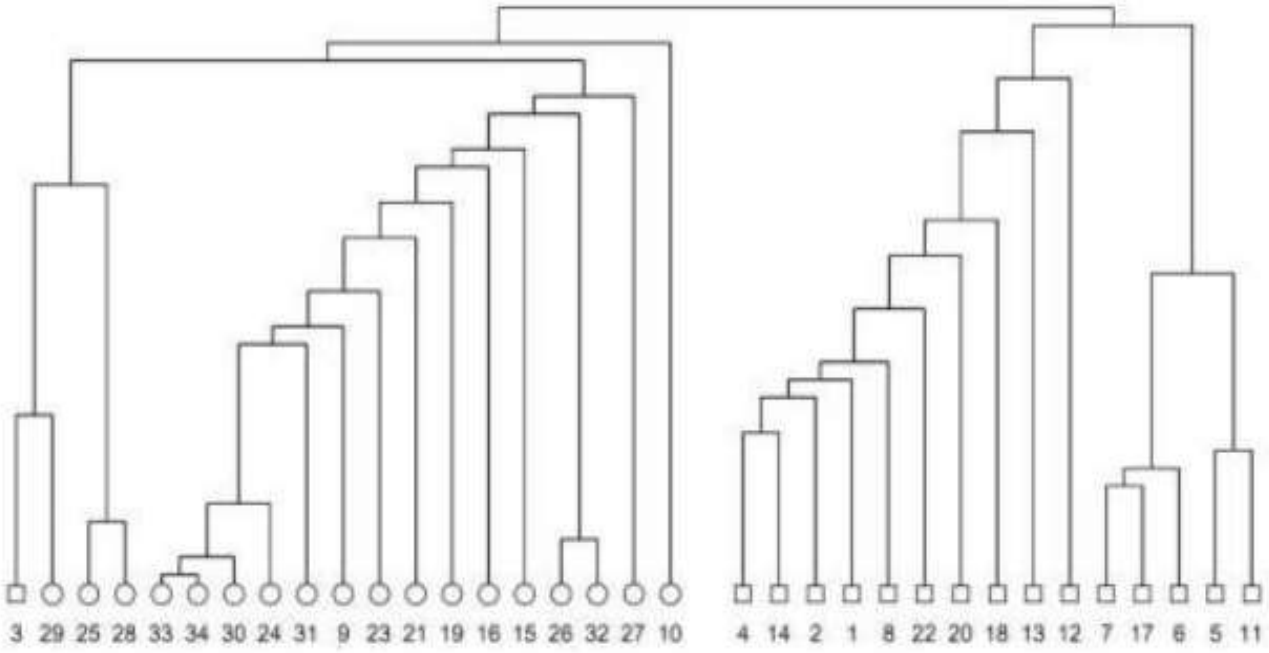
社团挖掘： Girvan-Newman算法

- 比较经典的社团研究案例包括空手道俱乐部(karate club),科学家合作网络(Collaboration network) 和斑马群体(zebras)的社交行为研究等。如下图：



社团挖掘： Girvan-Newman算法

- 其中，著名的空手道俱乐部社团已经成为通常检验社团挖掘算法效果的标准之一。下图是Girvan-Newman算法在空手道俱乐部上的运行结果。



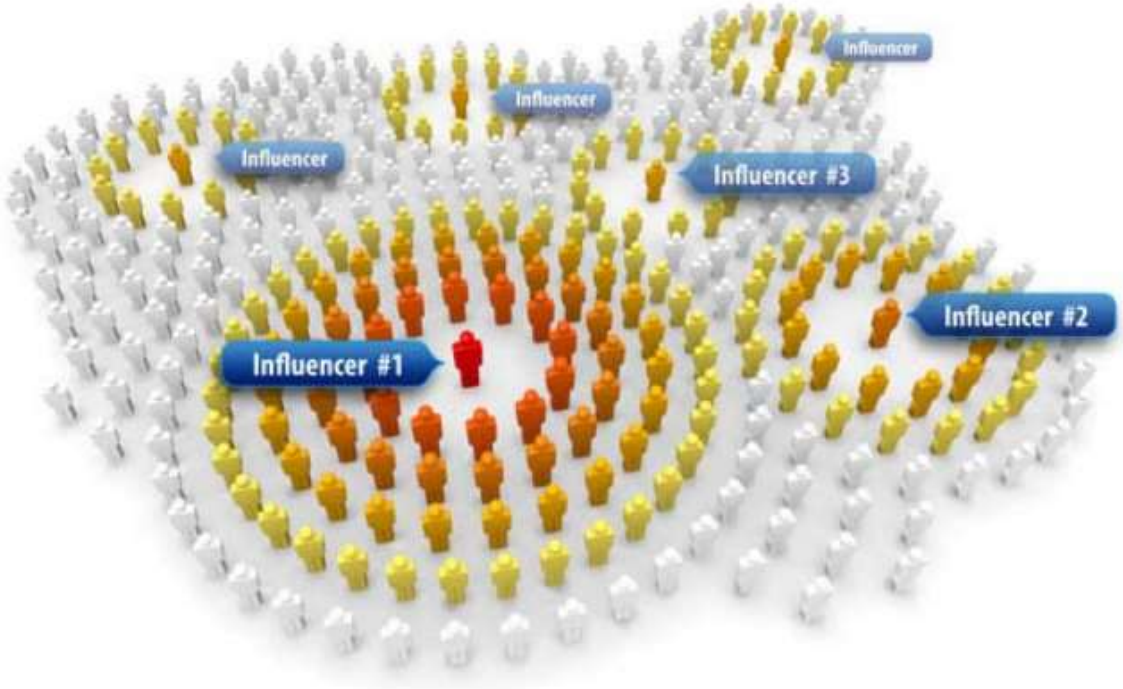
社团挖掘： Girvan-Newman算法

- Girvan-Newman算法准确率较高，但算法复杂度较大。
- Girvan-Newman算法复杂度为$O(m^2n)$,其中n是网络的节点数，m是网络的边数，而对于稀疏网络，复杂度约为$O(n^3)$。
- 对模块度的算法优化多种多样，从贪心到模拟退火等应有尽有。

目录

社交影响力定义

- 如用户买了一部新手机，在新浪微博平台发布一条微博“Note3的屏好大哦”，其好友看到后，将会成为三星手机的潜在用户，这就是社交影响力，即在社交网络中，一个用户的行为会对其好友的决策造成影响。



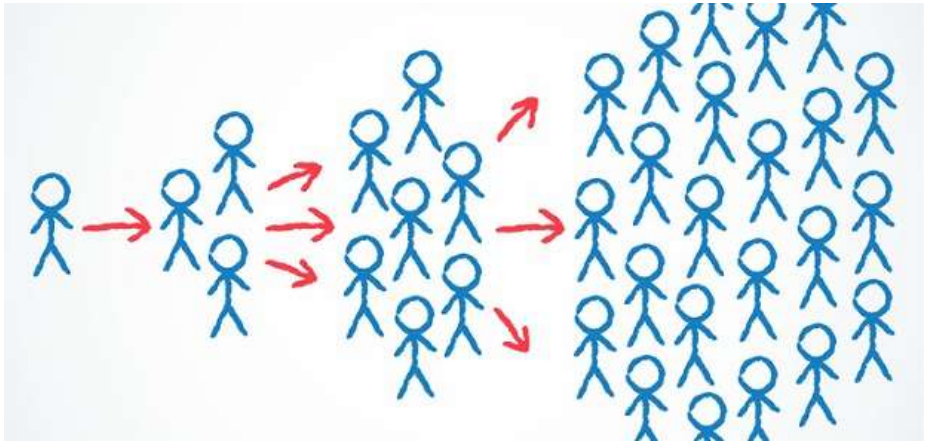
社交影响力传播举例


Beijing is great
Beijing is great
Beijing is great
Beijing is great
Beijing is great
Beijing is great
Beijing is great

社交影响力的应用-病毒营销

- 病毒营销(word of mouth)是由欧莱礼媒体公司（O‘Reilly Media）总裁兼CEO提姆·奥莱理（Tim O’Reilly）提出，其讯息传递策略是通过公众将信息廉价复制，告诉给其它受众，从而迅速扩大自己的影响。

传播快，成本低



病毒营销案例

- Hotmail
  - 在每一封免费发出的信息底部附加一个简单标签：”Get your private, free email at http://www.hotmail.com”
  - 1年半时间里，就吸引了1200万注册用户；每天超过15万新用户的速度发展；在网站创建的12个月内，Hotmail只花费很少的营还不到其直接竞争者的3%。
- 凡客体

病毒营销案例

- 三星的大手笔事件营销：奥斯卡史上最昂贵的自拍照
  - 2014，艾伦·德杰尼勒斯带领梅丽尔·斯特里普、茱莉亚·罗伯茨和布拉德·皮特等

病毒营销案例

- 奥巴马总统竞选

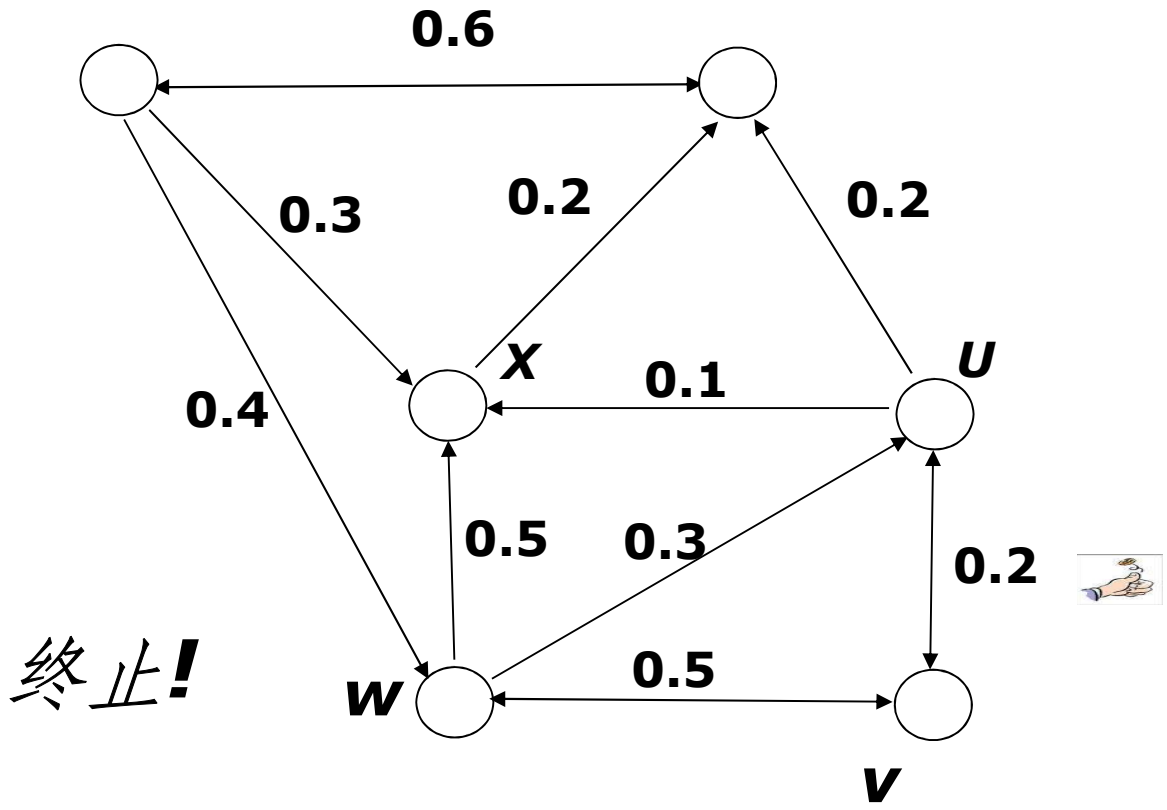
15万名奥巴马支持者在Facebook安装了“奥巴马201[影响力扩散]个程序，间接得到这些支持者数百万的Facebook好友信息。

- 江南Style的传播

《江南style》被放到youtube上后，被**汤姆·克鲁斯**等明星在社交帐[病毒营销]后引来模仿[病毒营销]了新一轮的推荐、恶搞，最终将这首歌推向“全球神曲”的地位。

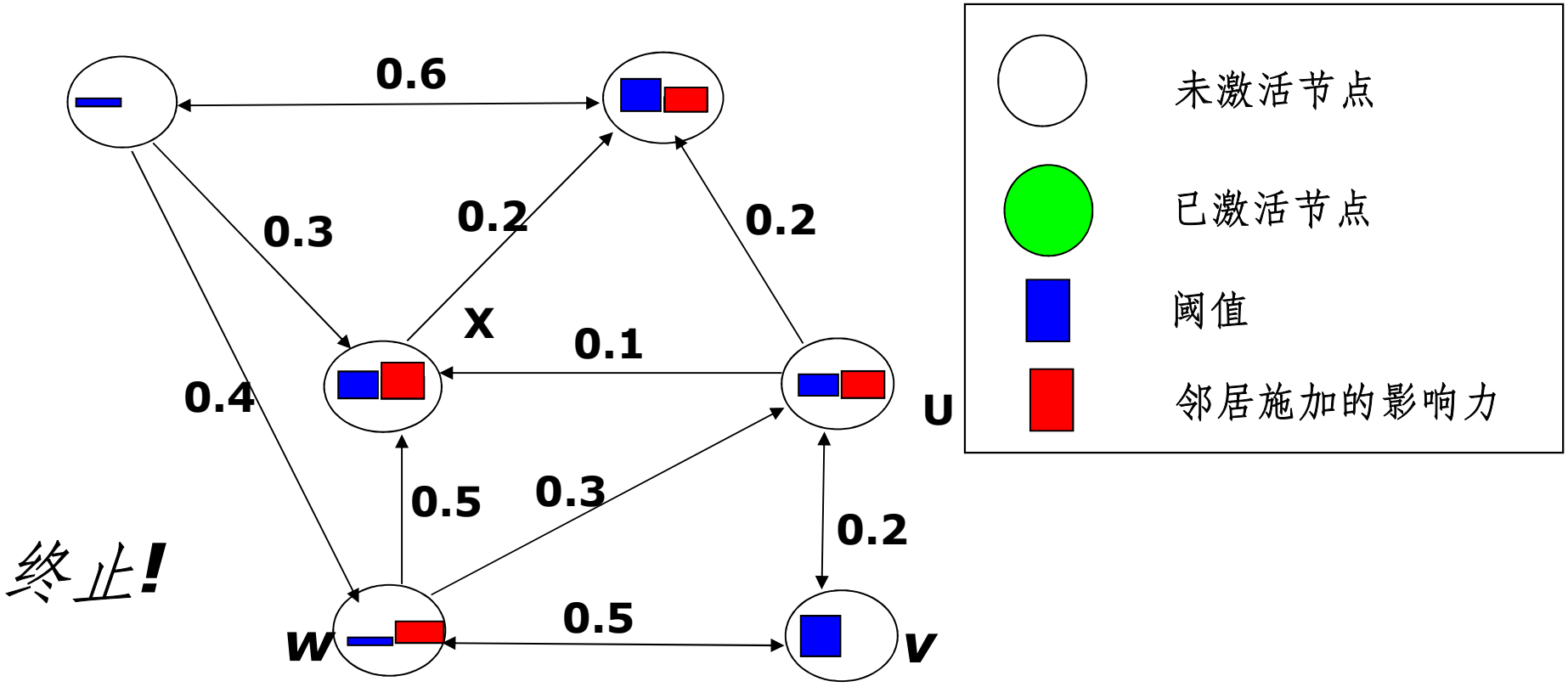
影响力传播模型

- 独立级联(Independent Cascade)模型



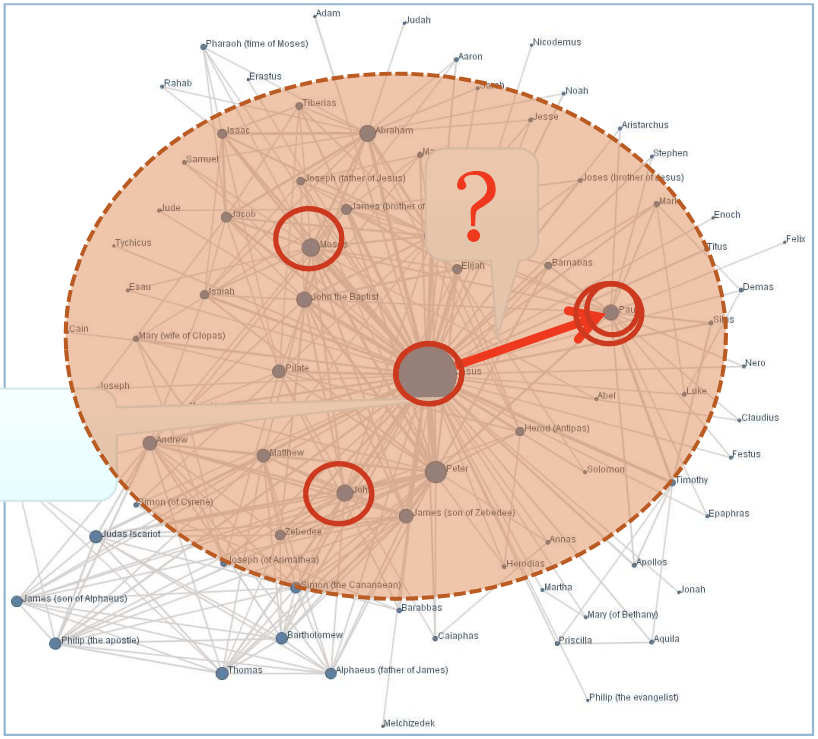
影响力传播模型（续）

- 线性阈值(Linear Threshold)模型



影响力最大化问题定义

- 给定影响力传播模型和图G=<V,E>，选取k个节点，使得选出的节点集合的影响力最大

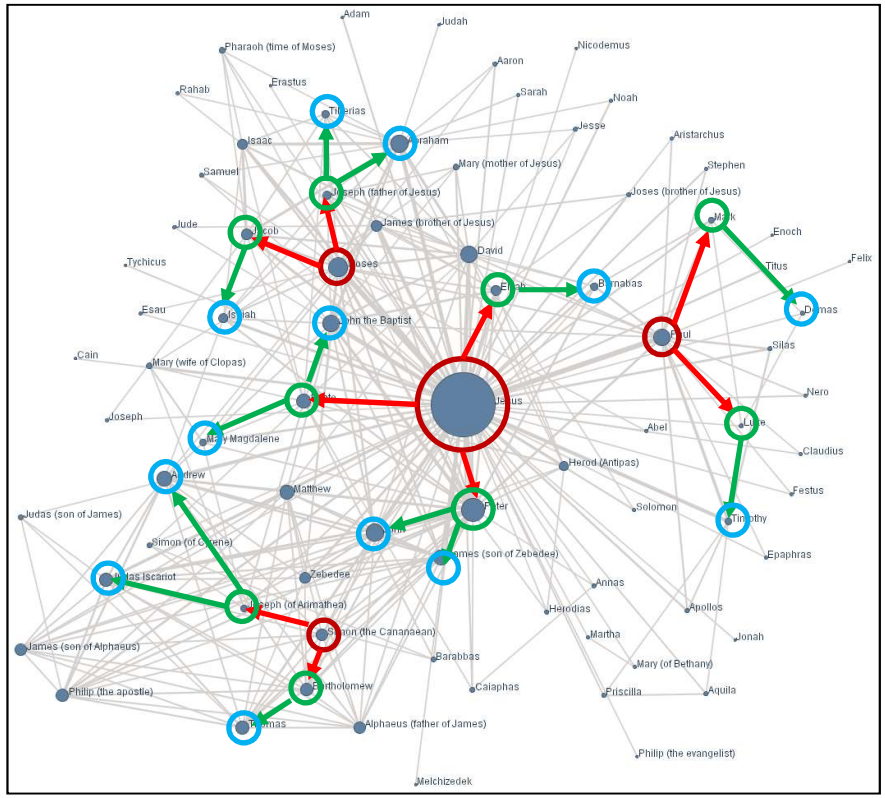


影响力？

影响力最大化问题定义

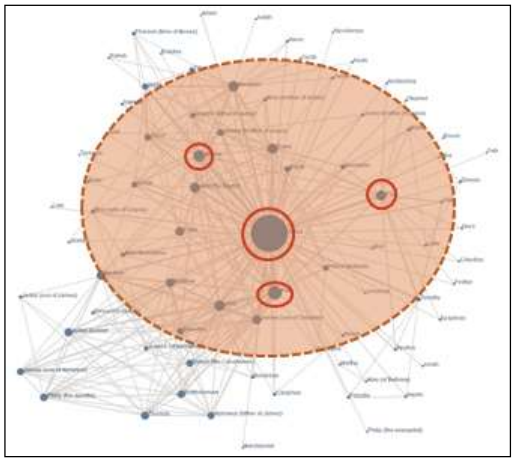
- 给定影响力传播模型和图G=<V,E>，选取k个节点，使得选出的节点集合的影响力最大

迭代次数	激活点数
0	4
1	9
2	13
...	...
总影响力	4+9+13+...



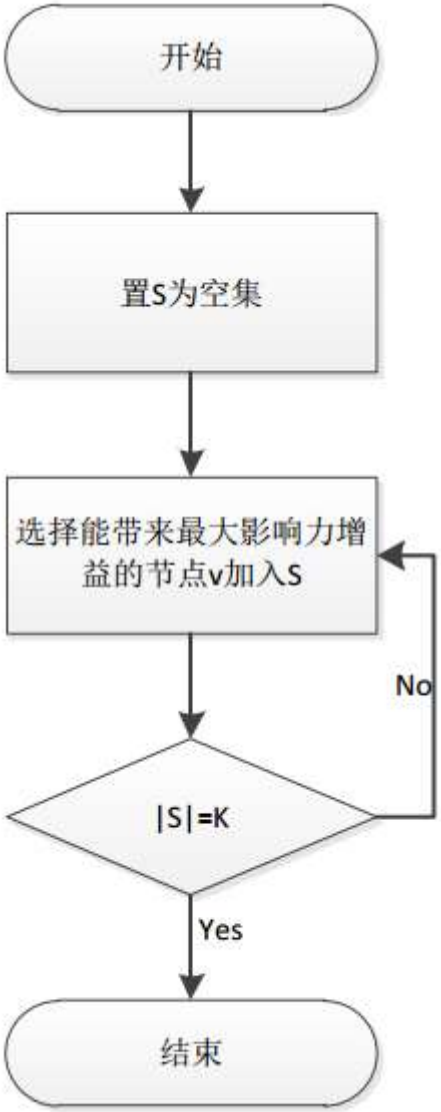
贪心算法

- 算法流程图（右）



优点：保证63%最优！(Submodularity)

缺点：需要大量蒙特卡罗模拟，速度慢



其他算法

- Baselines
  - Degree---选取度最大的K个节点
  - PageRank---选取PageRank值最大的K个节点
- CELF (Leskovec et al. KDD 2007)
  - 利用次模性质为每个节点维护一个上界，以减少计算量
- DegreeDiscount (Wei Chen et al. KDD 2009)
  - 按照度降序选择节点，每选中一个节点，将其邻居节点的度减小一定值
- PMIA (Wei Chen et al. KDD 2010)
  - 利用最大影响力路径，用树结构代替图，加速影响力计算
- IRIE (Jung Kyomin et al. ICDM 2012)
  - 将影响力排序和影响力估计结合起来
- UBLF (Zhou et al. ICDM 2013)
  - 在CELF的基础上，利用上界函数减少CELF算法初始化阶段的计算量

Highlights from KDD 2019

- Test of Time Award
  - **Cost-effective outbreak detection in networks** by **Jure Leskovec**, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Van Briesen, and Natalie Glance
  - **Jure Leskovec**: KDD 2005 Best Paper & 2016 Test of Time Award
- KDD 2019 Dissertation Award
  - Tim Althoff, Stanford, advised by **Jure Leskove**

https://cs.stanford.edu/people/jure/

KDD CUP 2021

https://ogb.stanford.edu/kddcup2021/

OGB

Get Started Updates KDD Cup Datasets Leaderboards Paper Team

OGB-LSC @ KDD Cup 2021

— Large – Scale Challenge —

An illustrative overview of the three OGB-LSC datasets is provided below.