



网络科学研究历程与梦想

Linyuan Lü (吕琳媛)

E-mail: linyuan.lv@hznu.edu.cn

**阿里巴巴复杂科学研究中心
杭州师范大学**



你的梦想是什么？

时常听到很多人想当年，

想当年啊，我想
成为一个画家



想当年啊，我想
成为一个作家



想当年啊，我想
成为一个发明家



梦想还是要有的，万一实现了呢

——老马



学习科研经历

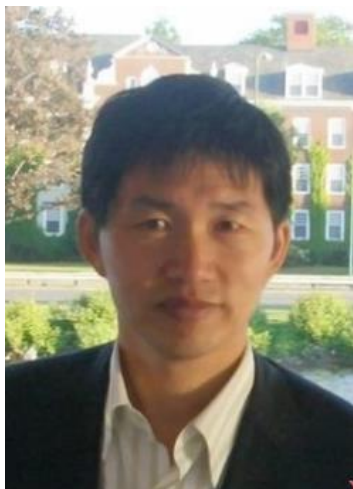
- 2002/09-2006/07 北京师范大学，管理学院，管理科学，学士（导师：王有贵）
- 2006/09-2008/07 北京师范大学，管理学院，系统分析与集成，硕士（导师：王有贵）
- 2008/09-2012/03 瑞士弗里堡大学，物理系，理论物理，博士（导师：张翼成）
- 2012/06-今 杭州师范大学，阿里巴巴商学院，特聘教授，杭州市特聘专家，阿里巴巴复杂科学研究中心副主任

从本科到硕士的转折

0.0001

对你意味着什么？

从硕士到博士的转折



北京师范大学系统科学学院，教授 博士生导师 世界经济学会（WEA）发起会员，非平衡社会科学（NESS）主要成员，中美交流项目富布莱特学友会成员。研究领域包括债务经济学，复杂经济学，拟化经济学。

机会永远留给有准备的人，努力的人！



张翼成

瑞士弗里堡大学教授。1984年毕业于罗马大学，获博士学位。先后在美国布克海文（Brookhaven）国家实验室、意大利核物理研究院（INFN）、丹麦Nordita大学从事物理学前沿研究工作。他是欧洲物理经济学论坛(Econophysics)的创始人及主编。张翼成教授与两位合作者Mehran Kardar和Giorgio Parisi于1986年提出的KPZ方程，开创了复杂科学的一个分支，它的数学解终于由Martin Hairer攻下并获2014费尔兹奖。他的研究十多年来转移到金融与信息经济，少数者博弈模型成功描述金融市场，而近期工作专注于互联网经济理论与应用。

博士期间

做什么？

怎么做？



Emergence of product differentiation from consumer heterogeneity and asymmetric information

L. Lü^{1,2}, M. Medo^{3,a}, Y

¹ Department of Systems

² Lab of Information Eco
610054 Chengdu, P.R. C

³ Physics Department, Ur

⁴ Institute for Scientific I

Received 8 Apr
Published online

Eur. Phys. J. B **71**, 565–57
DOI: 10.1140/epjb/e2009-0

Regular Article

The role of a

Linyuan Lü, M. Medo^a,

¹ Physics Department, Ur

² Lab of Information E
610054 Chengdu, P.R. C.....

Received 15 January 2009 / Received in final form 3 July 2009

Published online 26 September 2009 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2009



g, P.R. China
ogy of China,

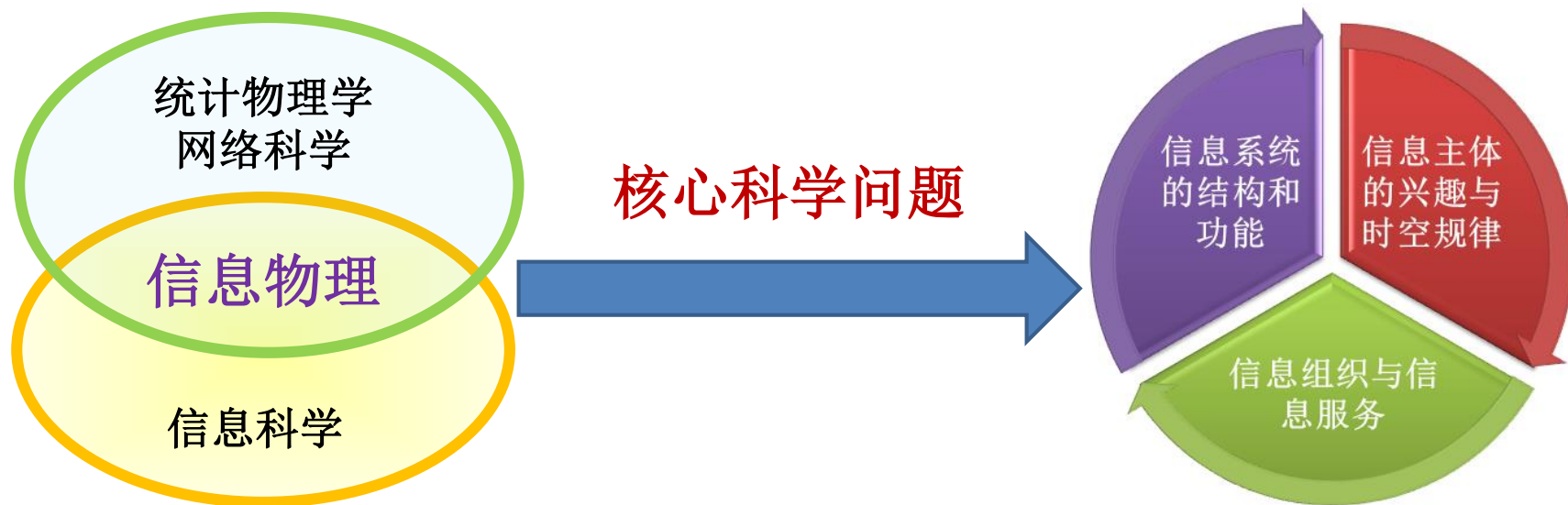
flag 2008

EUROPEAN
PHYSICAL JOURNAL B

ions

and Technology of China,

研究方向调整为复杂网络信息挖掘



EU FP7 欧盟第七框架项目

973: 社会网络与信息传播的基础问题研究

理解信息的产生与组织，
进行更好的导航、推荐、
预测和控制

链路预测，节点排序、个性化推荐

链路预测，节点排序、个性化推荐的关系图

Rank of LP

1	U2-U4
2	U1-U3
3	U4-U5
4	U1-U5
5	U3-U5

Prediction—complete partial information

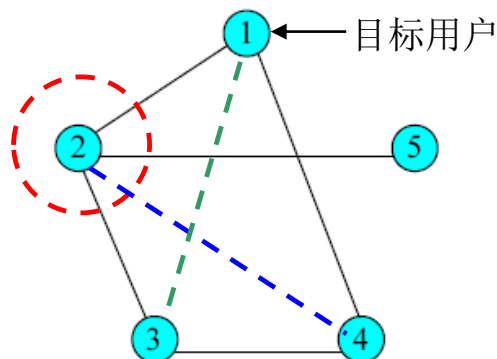
Sample networks:

Prediction of **existed yet unknown links**, e.g. *food webs, protein-protein interaction networks, metabolic networks*.

Evolving networks

Prediction of **future links**, e.g. *online friendship networks*

问：哪两个人会在将来成为朋友？



问：谁是最有影响力的节点？

Ranking—understand the significant part based on known information

Rank of LP

1	U2-U4
2	U1-U3
3	U4-U5
4	U1-U5
5	U3-U5

问：给用户1推荐谁为朋友？

Recommendation—
guide the information
generating
Subproblem of LP

科研方法：工欲善其事，必先利其器

- 爱因斯坦有一句名言：“方法比知识更重要”；
- 数学家拉普拉斯——认识一位天才的研究方法，对于科学的进步并不比发现本身更少用处，科学研究的方法经常是极富兴趣的部分；
- 教育家蔡元培比喻：“科学知识是点石成金的金，最终有限；科学方法则是点石成金的手指，可以产生无穷的金。”

方法的重要性胜过知识体系。

修炼“基本功”：如何完成一篇科研工作

- **定义“问题”**：能够形成理论和现实对话的、有适当的复杂性，且能引发科学共同体和大众感兴趣的问题域在哪？
- **找到“巨人”**：前人对这个问题已有哪些研究？哪些是具有代表性的关键理论？找到关键文献。
- **搭好“云梯”**：前人是如何阐释待研究的问题（为何没有做出完满的回答，或者为何没有作答，即理论背景）
- **站上“肩膀”**：在前人研究基础上搭建有针对性地回答研究问题的新的理论框架，并检验研究问题及研究框架是否适当。
- **一览众山**：表述创新的理论和方法，并与他人交流分享，从中可能碰撞出新的问题。

勤于思考
的习惯



文献检索
的能力



学习前人的
理论和方法



理论创新能力



写作和表
达能力

收益



20.033



9.674



11.470



5.578

- 博士顺利毕业，找到不错的职位。
- 发表学术论文50余篇，引用2700余次。
- 两篇论文入选全球Top-1%高引用论文。
- 出版专著一册，获得第四届中国大学出版社图书奖一等奖。

链路预测

什么是链路预测

Problem Description

- Given a network $G(V, E)$, V is the set of nodes, E is the set of links.
- Estimating the likelihood of the existence of a link between two nodes, based on the observed topology.
- Sort the nonexistent links in descending order according to their likelihoods, the top ones are most likely to exist.

Data

- **Training set** as known information
- **Probe set** for testing



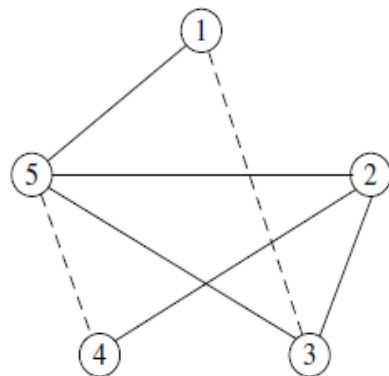
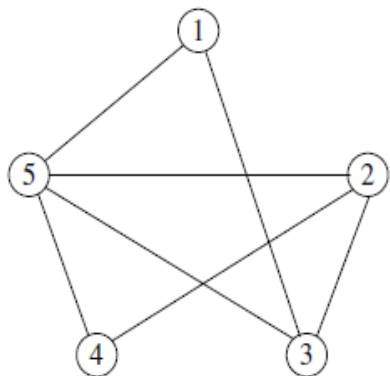
Metric

- AUC (area under the receiver operating characteristic curve)
 - Probability that a randomly chosen link in the probe set (PL) has higher likelihoods than a randomly chosen nonexistent link (NL).
 - Independently testing n times, and n_1 PL > NL, n_2 PL = NL, then
- Precision & Ranking Score

$$AUC = \frac{n_1 + 0.5n_2}{n}$$



示例



问题描述： 已知边集合 $E=\{\text{左图所有7条实线}\}$ ，测试集 $E^P=\{(1,3),(4,5)\}$ ，训练集 $E^T=\{(1,5),(2,3),(2,4),(2,5),(3,5)\}$ ，不存在的边集合为 $U-E=\{(1,2),(3,4),(1,4)\}$ 。此时链路预测就是要基于 E^T 的信息预测剩下的五条未观测到的链接 $\{(1,2),(1,4),(3,4),(1,3),(4,5)\}$ 出现的可能性。

方法： $s_{12}=0.4$ ， $s_{13}=0.5$ ， $s_{14}=0.6$ ， $s_{34}=0.5$ ， $s_{45}=0.6$

评价： 在计算AUC的时候需要将测试边的分数值 $\{(1,3),(4,5)\}$ 与不存在边的分数值 $\{(1,2),(3,4),(1,4)\}$ 进行比较。一共有六对，即

对测试边 (1,3)

$s_{12}=0.4 < s_{13}=0.5$ AUC+1
 $s_{14}=0.6 > s_{13}=0.5$ AUC+0
 $s_{34}=0.5 = s_{13}=0.5$ AUC+0.5

对测试边 (4,5)

$s_{12}=0.4 < s_{45}=0.6$ AUC+1
 $s_{14}=0.6 = s_{45}=0.6$ AUC+0.5
 $s_{34}=0.5 < s_{45}=0.6$ AUC+1

$$\text{AUC} = (3 \times 1 + 2 \times 0.5) / 6 = 2/3 \approx 0.67$$

链路预测的方法

- 基于节点属性相似性的方法
 - 物以类聚，人以群分，两个节点的属性越相似，两个节点之间就越有可能产生联系。
- 基于网络结构相似性的方法
 - 网络中如果两个节点之间相似性(或者相近性)越大，它们之间存在链接的可能性就越大。
 - CN, AA, RA, LP, Katz, LRW, SRW, RWR,
- 基于似然分析的方法
 - 基本思路是根据网络结构的产生和组织方式以及目前已经观察到的链路，计算网络的似然，并认为真实的网络都是使得网络似然最大的那种情况，然后再根据网络似然最大化计算每一对未连接的节点产生链接的可能性。
 - 代表算法：层次结构模型、随机分块模型、闭路模型
- 基于机器学习的方法

基于节点局部信息的相似性指标

名称	定义	名称	定义
共同邻居(CN)	$s_{xy} = \Gamma(x) \cap \Gamma(y) $	大度节点不利指标(HDI)	$s_{xy} = \frac{ \Gamma(x) \cap \Gamma(y) }{\max\{k(x), k(y)\}}$
Salton指标 ^[27]	$s_{xy} = \frac{ \Gamma(x) \cap \Gamma(y) }{\sqrt{k(x) \times k(y)}}$	LHN-I指标 ^[22]	$s_{xy} = \frac{ \Gamma(x) \cap \Gamma(y) }{k(x) \times k(y)}$
Jaccard指标 ^[28]	$s_{xy} = \frac{ \Gamma(x) \cap \Gamma(y) }{ \Gamma(x) \cup \Gamma(y) }$	优先链接指标(PA) ^[31]	$s_{xy} = k(x) \times k(y)$
Sorenson指标 ^[29]	$s_{xy} = \frac{2 \Gamma(x) \cap \Gamma(y) }{k(x) + k(y)}$	Adamic-Adar指标(AA) ^[32]	$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lg k(z)}$
大度节点有利指标(HPI) ^[30]	$s_{xy} = \frac{ \Gamma(x) \cap \Gamma(y) }{\min\{k(x), k(y)\}}$	资源分配指标(RA) ^[33]	$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}$

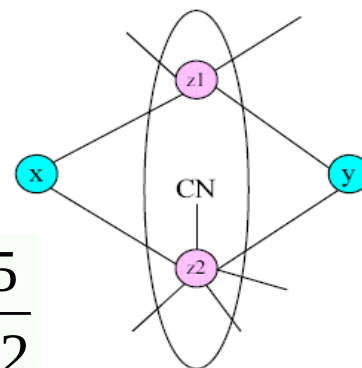
共同邻居的角色

Resource Allocation (RA) [Eur. Phys. J. B 71, 623-630 (2009)]

- 节点x可以通过邻居传递一些资源给节点y，他们的共同邻居扮演传播者的身份

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}$$

$$s_{xy} = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$



Local naïve Bayes model (LNB) [EPL 96 (2011) 48007]

- 定义一个角色函数，区分共同邻居节点的贡献是正的还是负的

$$r'_{xy} = \sum_{w \in O_{xy}} f(k_w) \log(s R_w).$$

$$s = \frac{M}{M^T} - 1 \quad R_w = \frac{N_{\Delta w} + 1}{N_{\wedge w} + 1}.$$

$$r_{xy}^{\text{LNB-CN}} = |O_{xy}| \log s + \sum_{w \in O_{xy}} \log R_w, \quad f(k_w) = 1$$

$$r_{xy}^{\text{LNB-AA}} = \sum_{w \in O_{xy}} \frac{1}{\log k_w} (\log s + \log R_w), \quad f(k_w) = \frac{1}{\log k_w}$$

$$r_{xy}^{\text{LNB-RA}} = \sum_{w \in O_{xy}} \frac{1}{k_w} (\log s + \log R_w), \quad f(k_w) = \frac{1}{k_w}$$

基于全局信息的相似性指标

- Average commute time (ACT)

(D. J. Klein, M. Randic, J. Math. Chemistry, 12 (1993) 81-95)

(F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, IEEE Trans. Knowl. Data. Eng. 19 (2007) 355)

$$n_{xy} = V(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+) \quad \text{where } V \text{ is the total degree}$$

- L^+ is pseudoinverse of the Laplacian matrix. ($L=D-A$)

- Cosine based on the Pseudoinverse of the Laplacian matrix

$$s_{xy} = \cos^+(x, y) = l_{ij}^+ / \sqrt{l_{ii}^+ \cdot l_{jj}^+}$$

- SimRank

$$s_{xy} = C \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} s_{ab}}{|\Gamma(x)| \cdot |\Gamma(y)|} \quad \text{where } C \text{ is decay factor}$$

- Random walk with restart (RWR)

- A random walker starting from node i , will iteratively moves to a random neighbor with probability c and return to node i with probability $1-c$.

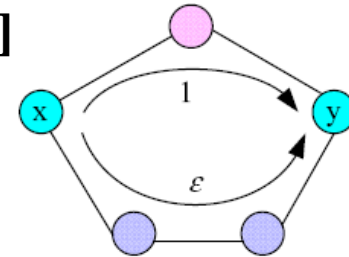
$$\vec{s}_i = cP^T \vec{s}_i + (1-c)\vec{e}_i \quad \longrightarrow \quad \vec{s}_i = (1-c)(I - cP^T)^{-1} \vec{e}_i$$

半局部指标

- Local Path (LP) [Phys. Rev. E 80, 046122 (2009)]

$$S = A^2 + \varepsilon A^3$$

- where A is the adjacent matrix, ε is a free parameter



- Local random walk (LRW) [EPL 89, 58007 (2010)]

$$s_{ij}(n) = \pi_{ij}(n) \cdot \frac{k_i}{2E} + \pi_{ji}(n) \cdot \frac{k_j}{2E}$$

- π_{ij} the probability that after n steps the walker happens to arrive at j .

- Superposed random walk (SRW) [EPL 89, 58007 (2010)]

- Reset the initial resource to node i at each time step

$$s'_{ij}(n) = \sum_{l=1}^n [s_{ij}(l)]$$



CN（局部），LP（半局部），KATZ（全局） 指标的比较

Topological features

- Giant component
- $\langle d \rangle$ average shortest distance

Networks	N	M	$\langle k \rangle$	$\langle d \rangle$	C	r	H
PPI	2375	11693	9.847	4.59	0.388	0.454	3.476
NS	379	941	4.823	4.93	0.798	-0.082	1.663
Grid	4941	6594	2.669	15.87	0.107	0.003	1.450
PB	1222	16717	27.360	2.51	0.360	-0.221	2.970
INT	5022	6258	2.492	5.99	0.033	-0.138	5.503
USAir	332	2126	12.807	2.46	0.749	-0.208	3.464

Accuracy

- Optimal parameter

Nets	PPI	NS	Grid	PB	INT	USAir
CN	0.915	0.983	0.627	0.924	0.653	0.958
LP	0.970	0.988	0.697	0.940	0.943	0.960 ^a
Katz	0.972	0.988	0.952	0.936	0.975	0.956

Time complexity

- In microsecond

Nets	PPI	NS	Grid	PB	INT	USAir
CN	10690	253	5161	31112	6711	2208
LP	543589	1638	11344	2873403	27641	93892
Katz	8073316	27479	69961063	1051528	72550935	17603

应用

- 通过添加目标节点与已标签节点间的虚拟连接提高局部标签网络的分类精度，可以应用于判断一篇学术论文的类型，判断一本书的政治态度（中立，自由派和保守派），或者判断一个手机用户是否产生了更换运营商(如从移动到联通)的念头
 - Q.-M. Zhang, M.-S. Shang, L. Lü, Int. J. Mod. Phys. C 21, 813 (2010).
- 帮助挖掘网络演化的驱动机制
 - 刘宏鲲，张效莉，曹莨，汪秉宏，周涛，“中国城市航空网络航线连接机制分析”，中国科学G，2009年第39卷，第7期，935-942;刘宏鲲，吕琳媛，周涛，利用链路预测推断网络演化机制，中国科学G 41, 816–823 (2011)
- 帮助对网络演化模型进行定量化评估。给定一个网络的演化模型，可以计算该模型下网络出现的似然，似然值越大模型越好。
 - W.-Q. Wang, Q.-M. Zhang, T. Zhou, EPL 98, 28004 (2012)
- 实现网络重构
 - R. Guimera, M. Sales-Pardo, PNAS 106, 22073 (2009).



- 链路预测还可以帮助更好的实现股票的预测
 - 可视图法将股票指数的时间序列直方图映射为一个网络，然后用局部随机游走LP方法预测网络的链接，再将网络结构映射为股票指数的时间序列直方图，由此预测出未来的股票指数信息。
 - Chen S, et al. arXiv: 1403.1713, 2014.
 - Lacasa L, et al. PNAS, 105(13):4972–4975, 2008.
- 生物医学：用于指导生物实验以提高实验成功率；帮助寻找疾病可能的致病基因。
- 社交网络：可以用于社交网络中的朋友推荐和敌友关系的预测；应用标签信息和链路预测的方法还可以在合著网络上对科学家进行分类并推荐合作者。
- 电力网络：社团划分以及电网恢复策略。
- 其他.....

综述： L. Lü, T. Zhou, Link prediction in complex networks: A survey, *Physica A* 390, 1150 (2011).

关于链路预测的论文

- T. Zhou, **L. Lü**, Y.-C. Zhang, Predicting Missing Links via Local Information, *Eur. Phys. J. B* 71, 623-630 (2009).
- **L. Lü**, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex network, *Phys. Rev. E* 80, 046122 (2009).
- **L. Lü**, T. Zhou, Link Prediction in Weighted Networks: The Role of Weak Ties, *EPL* 89, 18001 (2010).
- W.-P. Liu, **L. Lü***, Link Prediction based on Local Random walk, *EPL* 89, 58007 (2010).
- Q.-M. Zhang, M.-S. Shang, **L. Lü***, Similarity-Based Classification in Partially Labeled Networks, *Int. J. Mod. Phys. C* 21, 813 (2010).
- **L. Lü**, T. Zhou, Link prediction in complex networks: A survey, *Physica A* 390, 1150 (2011).
- Z. Liu, Q.-M. Zhang, **L. Lü***, T. Zhou, Link prediction in complex networks: a local naïve Bayes model, *EPL* 96, 48007 (2011).
- 吕琳媛, 复杂网络链路预测, *电子科技大学学报* 39, 651-661 (2010).
- 刘宏鲲, 吕琳媛, 周涛, 利用链路预测推断网络演化机制, *中国科学G* 41, 816-823 (2011)
- 朱郁筱, 吕琳媛*, 推荐系统评价指标综述, *电子科技大学学报(自然科学版)*, 第41卷第2, 163-175 (2012)
- Y.-X. Zhu, **L. Lü**, Q.-M. Zhang, T. Zhou, Uncovering missing links with cold ends, *Physica A* 391, 5769-5778 (2012)
- Q.-M. Zhang, **L. Lü***, W.-Q. Wang, Y.-X. Zhu, T. Zhou, Potential Theory for Directed Networks, *PLoS ONE* 8(2), e55437 (2013).
- 吕琳媛, 周涛, 链路预测, 高等教育出版社, 2013年8月。
- **L. Lü***, L. Pan, T. Zhou, Y.-C. Zhang, H. E. Stanley, Toward link predictability of complex networks. *PNAS* 112(8), 2325-2330 (2015)

高等教育出版社 隆重推出

网络科学与工程丛书

网络度分布理论

1

Introduction to Complex Networks:
Models, Structures and Dynamics

2

网络科学导论

3

链路预测

4

复杂网络协调性理论

5

陈关荣



史定华



汪小帆



李翔



吕琳媛



周涛



陈立生



尹万红

名誉主编: 郭雷 院士 金芳蓉 院士 李德毅 院士

主 编: 陈关荣

副主 编: 史定华 汪小帆

委 员: 曹进德 陈增强 狄增如 段志生 方锦清 傅新楚

胡晓峰 来润斌 李翔 刘宗华 陆君安 吕金虎

汪秉宏 王青云 谢智刚 张翼成 周昌松 周涛

NSF

4

网络科学与工程丛书

链路预测

Link Prediction

■ 吕琳媛 周涛 著

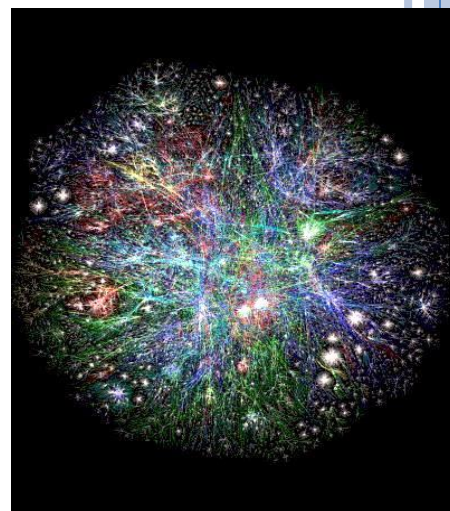
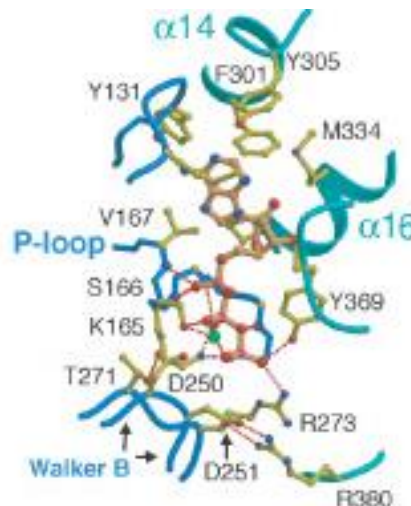
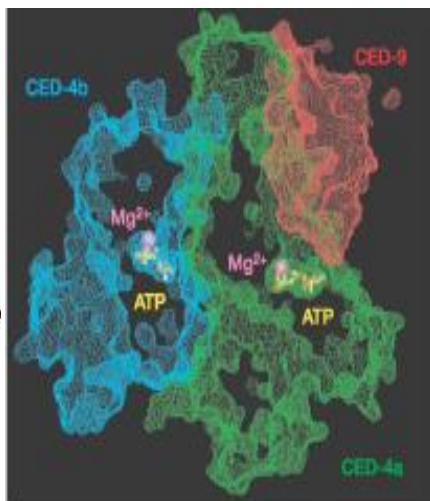
高等教育出版社
HIGHER EDUCATION PRESS

重要节点挖掘

研究背景和意义



- 食物链中哪些物种对整个生态的影响最大?
- “种子信息”发给哪些手机用户可以获得更多的转发?
- 全球经济体系中哪些国家或地区对于体系的健康发展至关重要?
- 高价雇佣微博大号做新产品的推广和营销真的有用吗?如果有用,又如何找到合适的达人?

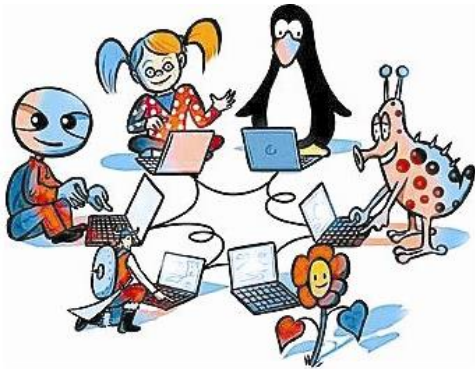


何为“重要”节点？

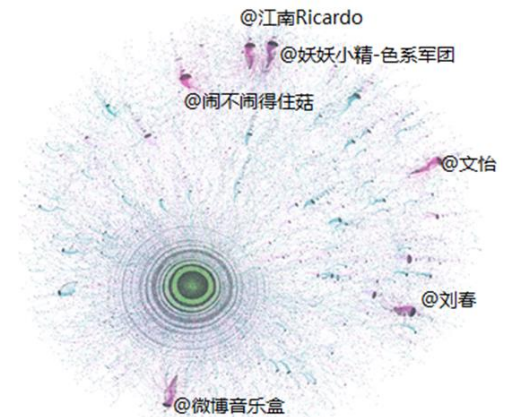
- 所谓的重要节点是指相比网络其他节点而言能够在更大程度上影响网络的结构与功能的一些特殊节点。重要节点一般数量非常少，但其影响却可以快速地波及到网络中大部分节点。
- **结构重要性：** 包括度分布、平均距离、连通性、聚类系数、度相关性等
- **功能重要性：** 涉及网络的抗毁性、传播、同步、控制等
 - 传播
 - 高影响力的节点能够带来更大范围的传播
 - 重要节点与临界值之关系[Phys Rev E 63 (2001) 066117];
重要节点与免疫策略的制定[New J Phys 12(2010) 023015]
 - 交通
 - 移除介数大的节点的连边可以改善交通运输能力
 - [Phys Rev E 76 (2007) 017101]
 - 同步
 - 通过牵制和控制少量重要节点提高同步能力
 - [Phys Rev E 74 (2006) 047102, New J Phys 14 (2012) 083006]

研究问题：如何挖掘社交网络中高传播影响力用户？

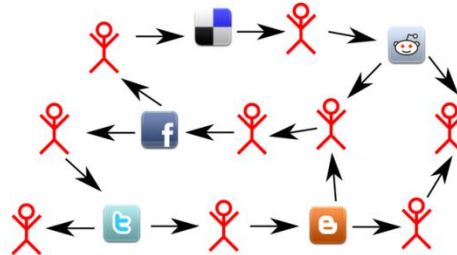
Making friends



Information spreading



Information acquisition



Social marketing



Social Networks

前人的方法

1. 无向网络中节点重要性的刻画

○ 度中心性

- 最简单的指标
- 刻画节点的受欢迎程度

○ 介数中心性

- 网络中节点对的最短路径中经过i节点的数目占有所有最短路径的比例
- 刻画了节点i对网络中节点对之间沿最短路径传输信息的控制力

○ 接近中心性

- 节点i与网络中其他节点的距离的平均值，越小越重要
- 刻画节点对信息流动的观察视野

● Degree Centrality

$$D_i = \frac{k_i}{N-1}$$

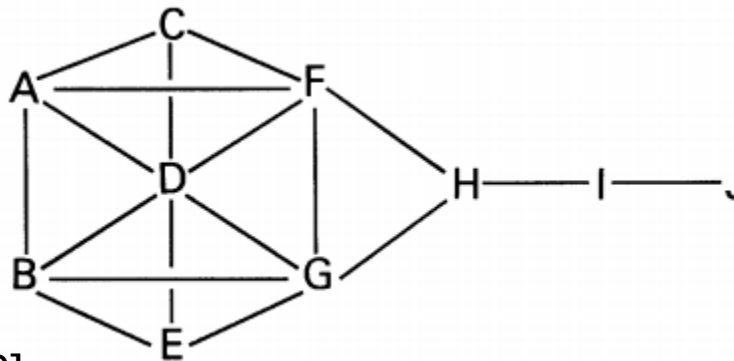
● Betweenness Centrality

$$B_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

● Closeness Centrality

$$C_i = \frac{N}{\sum_{j \in V \setminus i} d_{ij}}$$

风筝网络

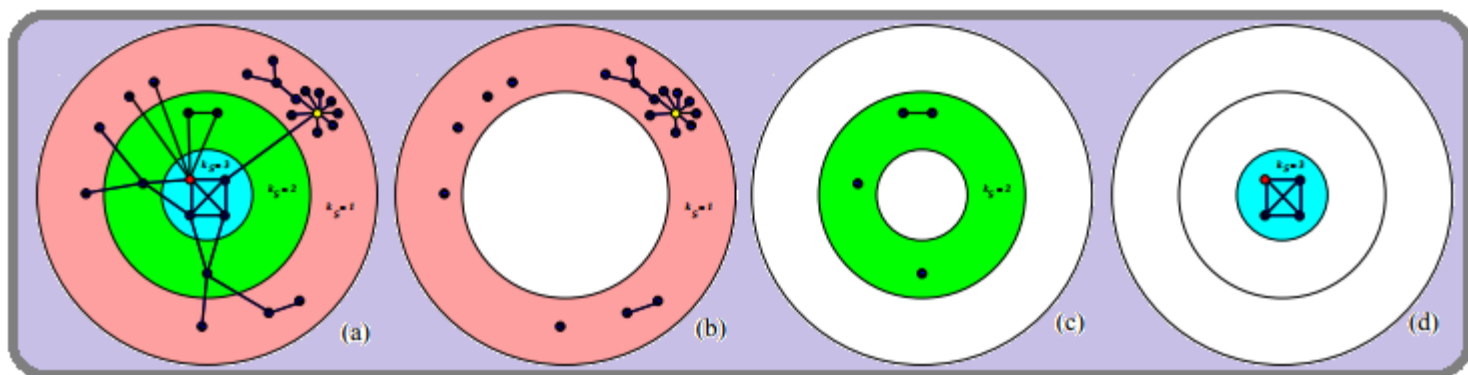


[Admin. Sci. Quart. 35 (1990) 342-369]

度中心性		介数中心性		接近中心性	
D	0.667	H	0.389	F	0.643
F	0.556	F	0.231	G	0.643
G	0.556	G	0.231	D	0.600
A	0.444	I	0.222	H	0.600
B	0.444	D	0.102	A	0.529
C	0.333	A	0.023	B	0.529
E	0.333	B	0.023	C	0.500
H	0.333	C	0.000	E	0.500
I	0.222	E	0.000	I	0.429
J	0.111	J	0.000	J	0.310

节点位置的重要性

○ K-壳分解确定节点的位置

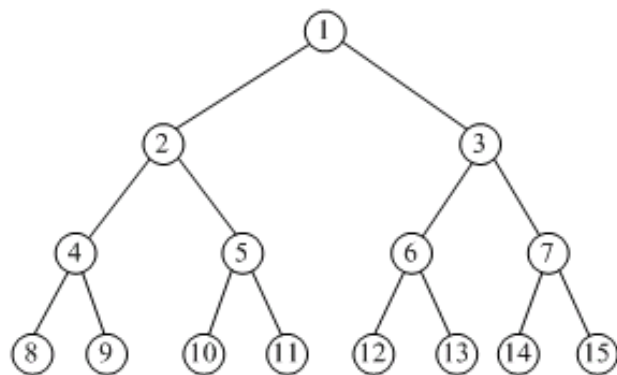


○ 主要结论:

- **单个传染源的情况**: 处于网络核心位置的节点的传播影响力更大, 即核值大的节点传播影响力更大。
- **多感染源的情况**: 依赖于传染源之间的距离, 应选择处在不同层且不直接相连的大度节点。

发现问题：K-壳分解的局限性

树形图



BA网络

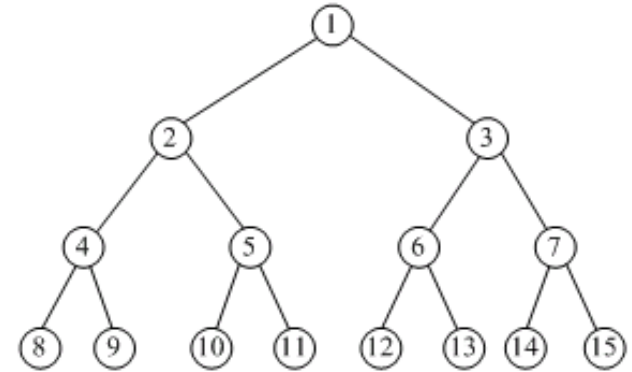


LOCALRANK

[Physica A 391, 1777-1787 (2012)]

$$Q(u) = \sum_{w \in \Gamma(u)} N(w) \quad C_L(v) = \sum_{u \in \Gamma(v)} Q(u)$$

$N(w)$ is the number of the nearest and the next nearest neighbors of node w .



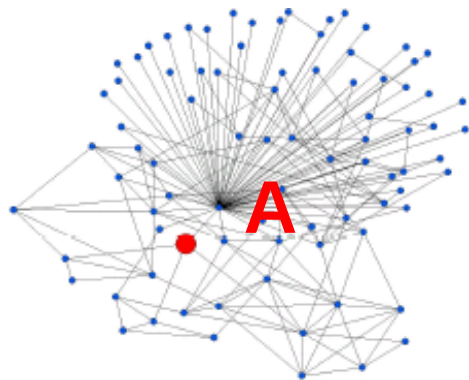
$$N(1)=6; N(2)=N(3)=8; N(4)=N(5)=N(6)=N(7)=5; N(8)=N(9)=\dots=N(15)=3.$$

$$\begin{aligned} Q(1) &= N(2) + N(3) = 16; \\ Q(2) &= N(1) + N(4) + N(5) = 6 + 5 + 5 = 16 = Q(3); \\ Q(4) &= N(2) + N(8) + N(9) = 8 + 3 + 3 = 14 = Q(5) = Q(6) = Q(7); \\ Q(8) &= N(4) = 5 = Q(9) = Q(10) = \dots = Q(15). \end{aligned}$$

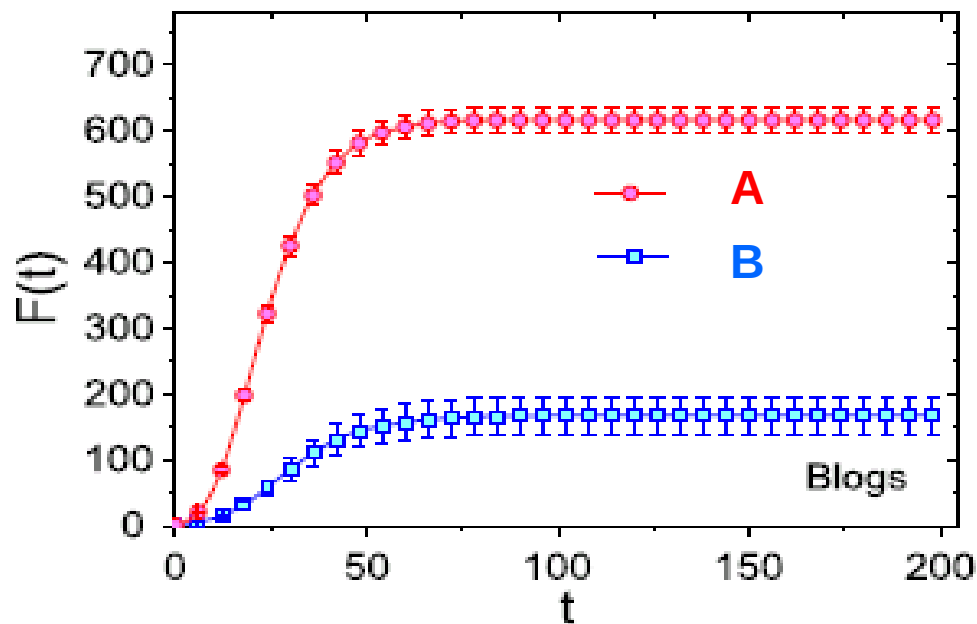
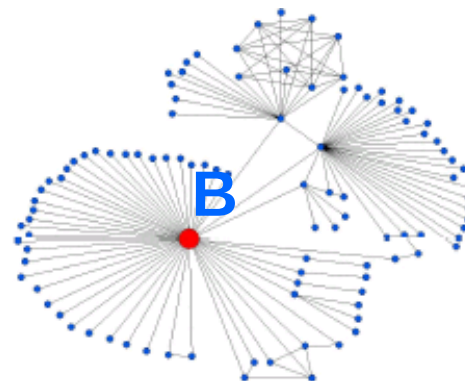
$$\begin{aligned} C_L(1) &= Q(2) + Q(3) = 32; \\ C_L(2) &= Q(1) + Q(4) + Q(5) = 16 + 14 + 14 = 44 = C_L(3); \\ C_L(4) &= Q(2) + Q(8) + Q(9) = 16 + 5 + 5 = 26 = C_L(5) = C_L(6) = C_L(7); \\ C_L(8) &= Q(4) = 14 = C_L(9) = C_L(10) = \dots = C_L(15). \end{aligned}$$

MSN交友网络传播影响力实验结果

Small degree (4) Large CL (14008)



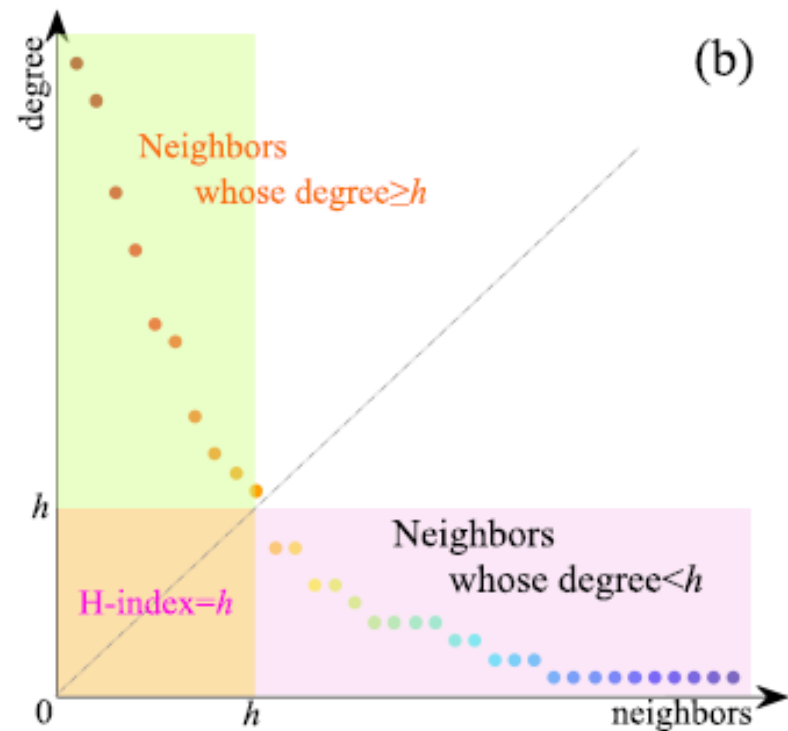
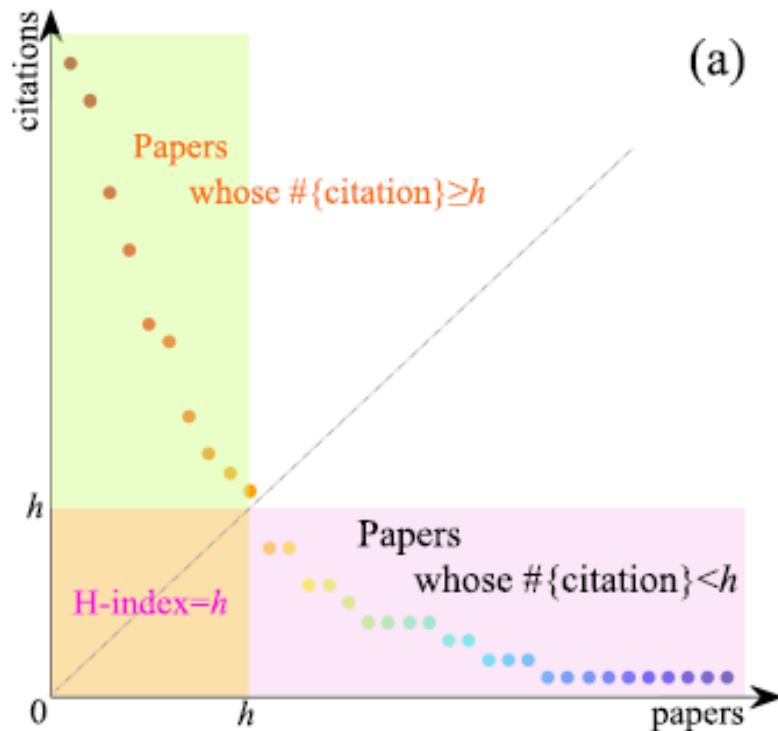
Large degree (50) Small CL (9388)



RELATION BETWEEN DEGREE, H-INDEX AND CORENESS

For a scholar

For a node on network



- The **H-index** of a node i , h_i , is the largest value such that node i has at least h_i neighbors of degree no less than h_i .

THE MAGIC H FUNCTION

- We construct an operator H , which acts on a finite number of reals $(x_1, x_2, \dots, x_n)$, and returns an integer

$$y = H(x_1, x_2, \dots, x_n) > 0,$$

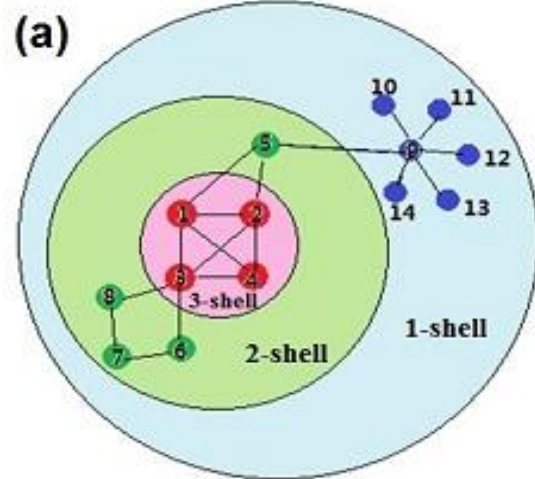
where y is the maximum integer such that there exist at least y elements in $(x_1, x_2, \dots, x_n)$, each of which is no less than y .

- If we assign the sequence as the degree of node i 's neighbors, then we have $y = h_i^{(1)}$.

$$\text{Degree} \xrightarrow{H} h^{(1)} \xrightarrow{H} h^{(2)} \xrightarrow{H} h^{(3)} \xrightarrow{H} \dots \xrightarrow{H} h^{(n)} \xrightarrow{H} \text{Coreness}$$

- Converge at time ∞

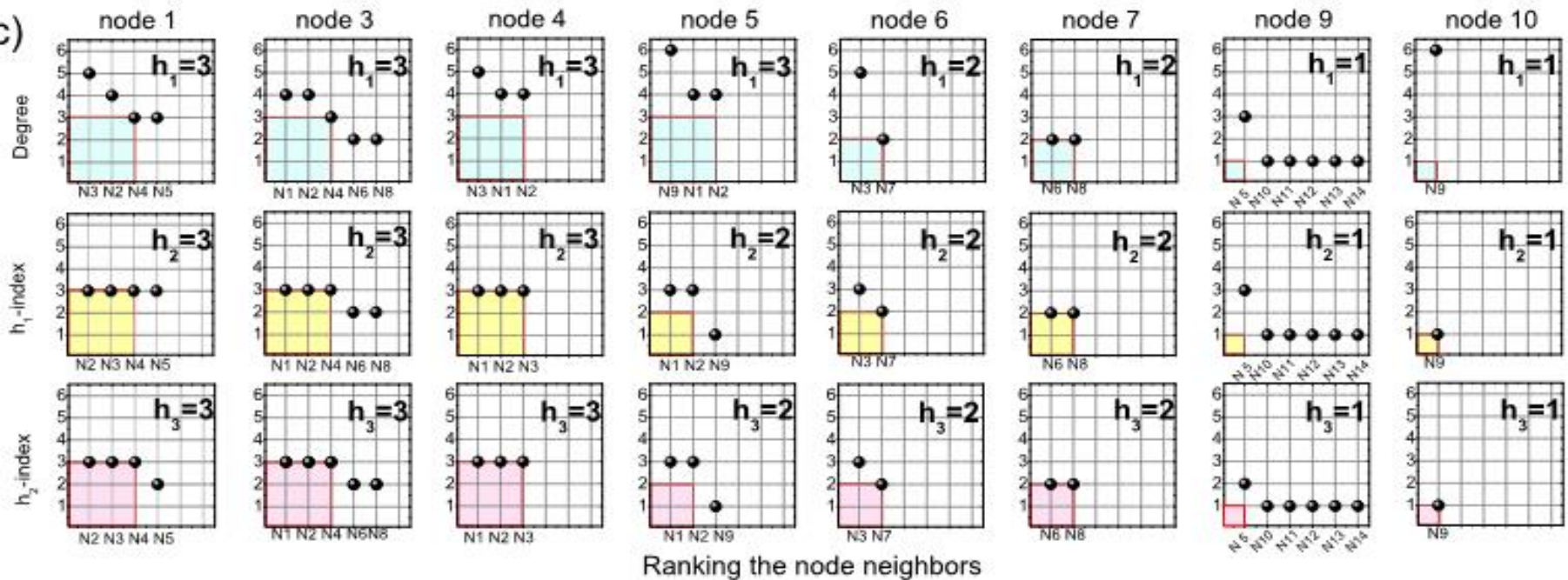
Theorem: Coreness $= \lim h$



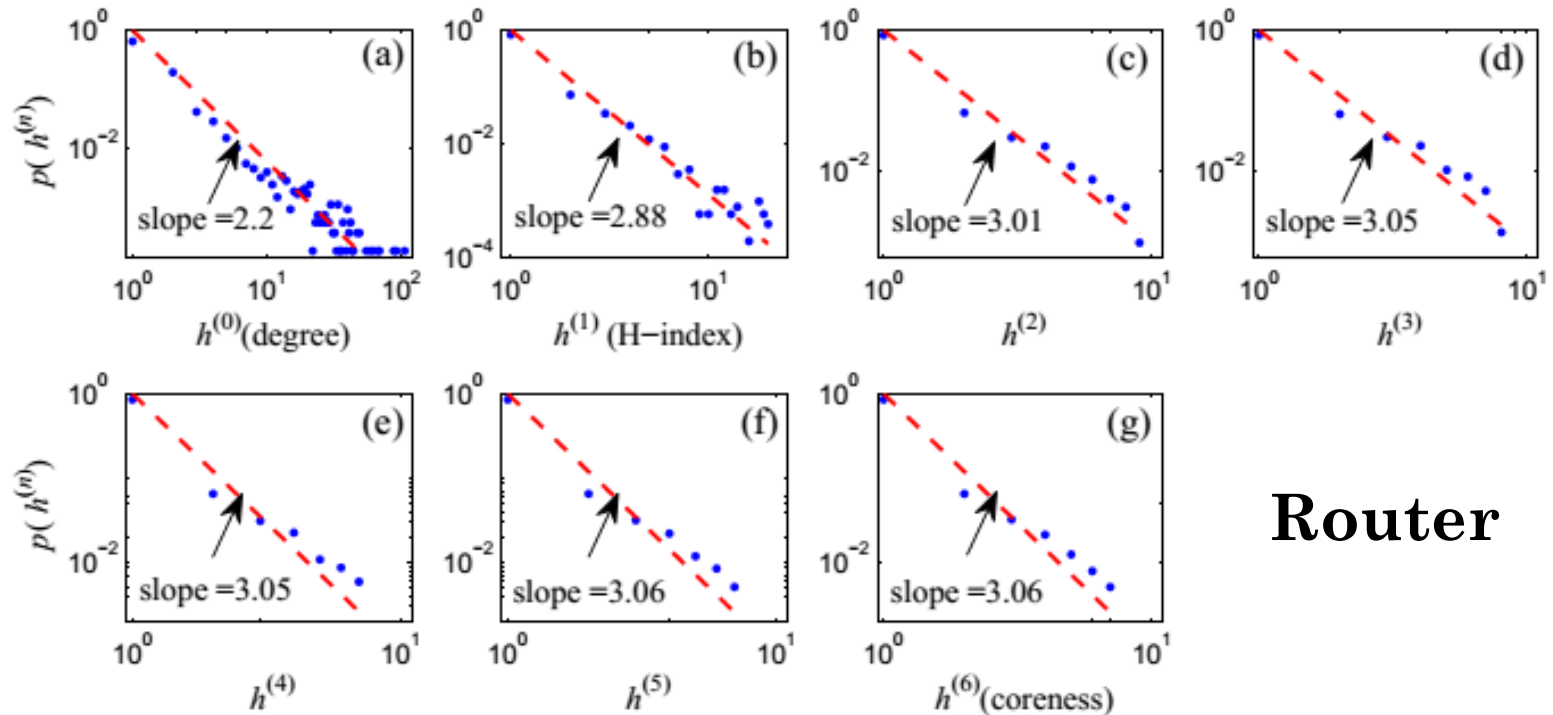
(b)

Node	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Degree	4	4	5	3	3	2	2	2	6	1	1	1	1	1
h_1 -index	3	3	3	3	3	2	2	2	1	1	1	1	1	1
h_2 -index (Coreness)	3	3	3	3	2	2	2	2	1	1	1	1	1	1
h_3 -index	3	3	3	3	2	2	2	2	1	1	1	1	1	1

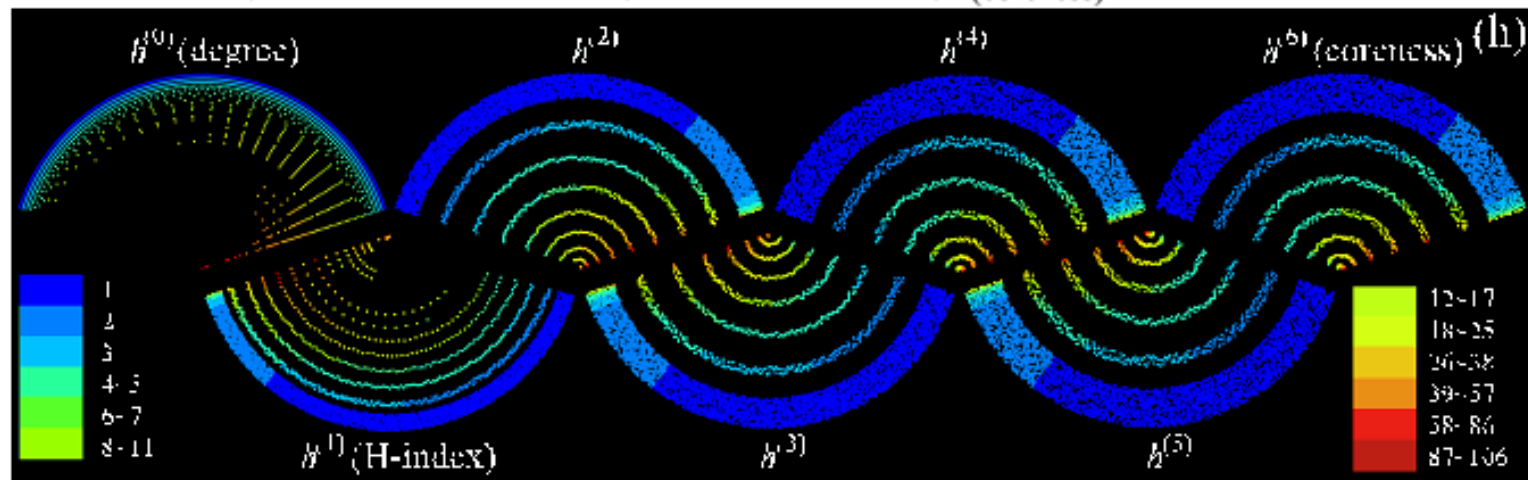
(c)



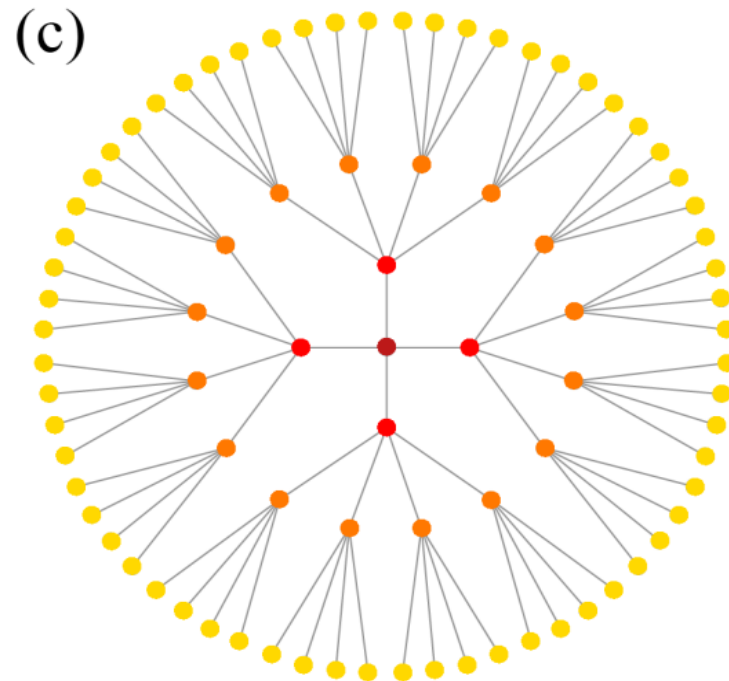
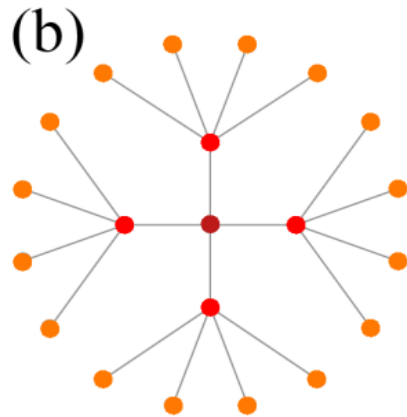
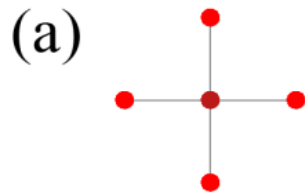
HETEROGENEOUS DISTRIBUTIONS



Router

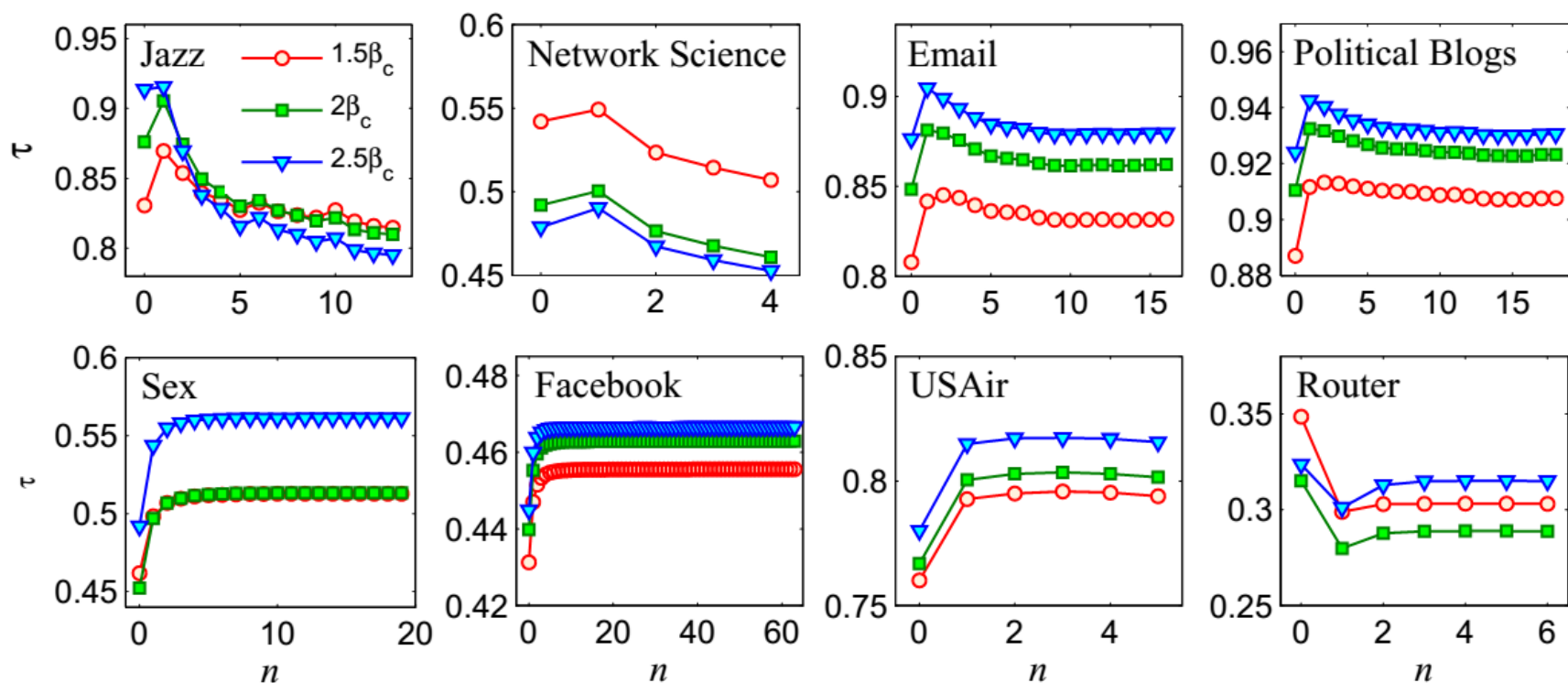


For a hierarchical tree of D levels we need $D-1$ indices $h^{(0)}, h^{(1)}, \dots, h^{(D-2)}$ to quantify node influence.

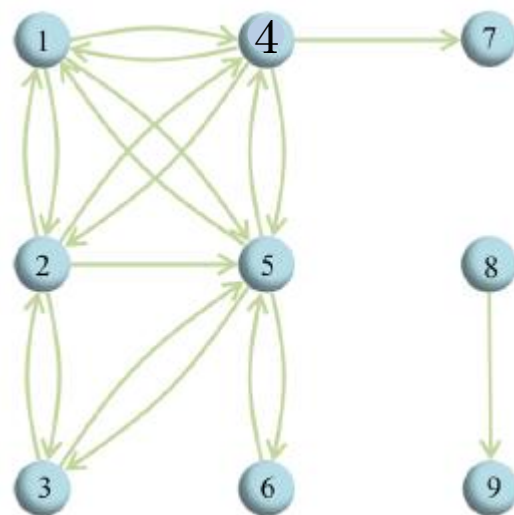


	Degree	h	h	Coreness
Level-1	4	4	4	1
Level-2	5	4	1	1
Level-3	5	1	1	1
Level-4	1	1	1	1

传播影响力

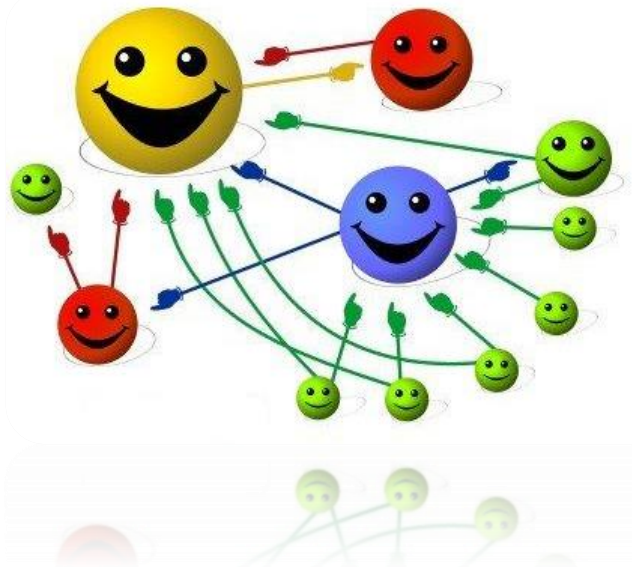


有向网络的H-INDICES



Node	$h_{in}^{(0)}$	$h_{in}^{(1)}$	$h_{in}^{(2)}$	$h_{in}^{(\infty)}$	$h_{out}^{(0)}$	$h_{out}^{(1)}$	$h_{out}^{(2)}$	$h_{out}^{(\infty)}$
1	3	3	2	2	3	3	2	2
2	3	2	2	2	4	3	2	2
3	2	2	2	2	2	2	2	2
4	3	3	2	2	4	3	2	2
5	5	3	2	2	4	2	2	2
6	1	1	1	1	1	1	1	1
7	1	1	1	1	0	0	0	0
8	0	0	0	0	1	0	0	0
9	1	0	0	0	0	0	0	0

2. 有向网络重要节点挖掘



从A页面到B页面的链接解释为A页面给B页面投票，
根据投票来源和投票目标的等级来决定新的等级

Algorithm

$$s_i(t+1) = c + (1-c) \sum_{j=1}^N \left[\frac{A_{ji}}{k_j^{out}} (1 - \delta_{k_j^{out},0}) + \frac{1}{N} \delta_{k_j^{out},0} \right] s_j(t)$$

$A_{ij} = 1$ if i points to j , k_j^{out} is the number of leaders of i ;

c is the probability for a web surfer to jump to a random website, called return probability.

局限性:

- (i) 随机跳转概率 c 对每个节点都相同，这意味着上网者从冷热门网站跳转到其他网站的概率相同！！
- (ii) 最优参数 c 是一个经验值，对于不同系统不具有普适性。

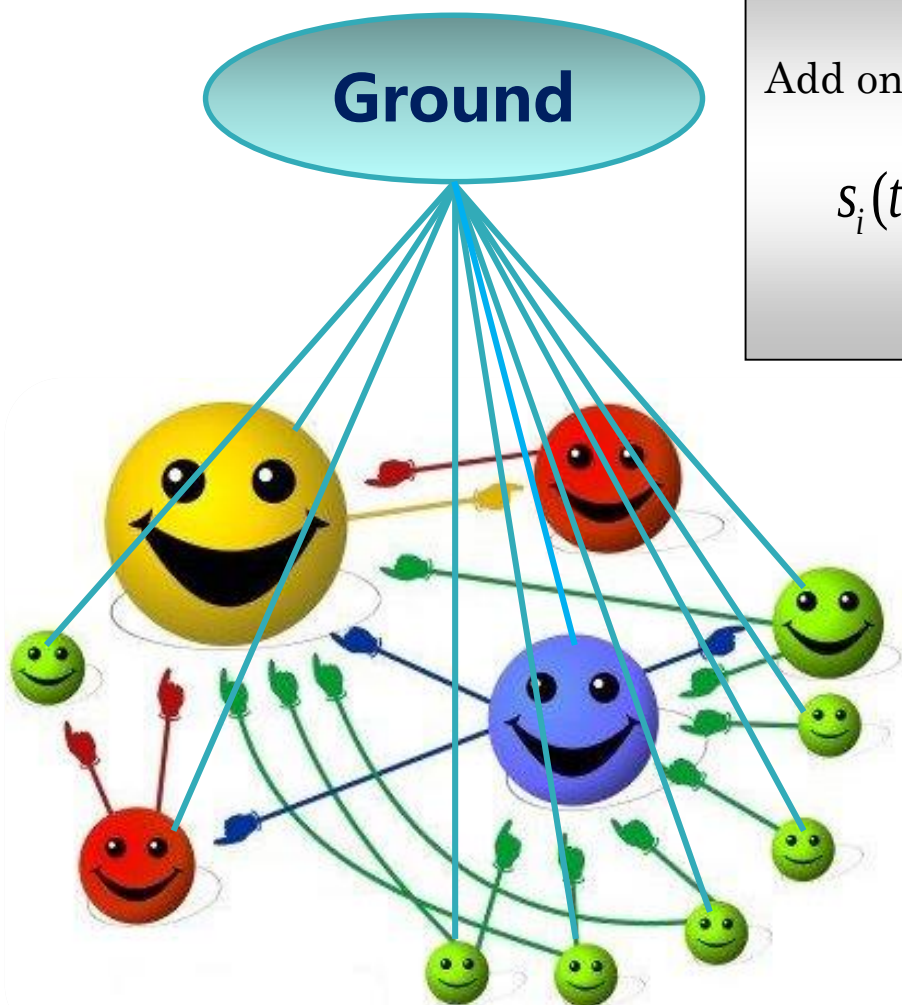
LEADER RANK

[PLoS One 6 (2011) e21202]

Algorithm

Add one ground node and N bi-directed links

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{A_{ji}}{k_j^{out}} s_j(t) \quad S_i = s_i(t_c) + \frac{s_g(t_c)}{N}$$



强连通 解唯一

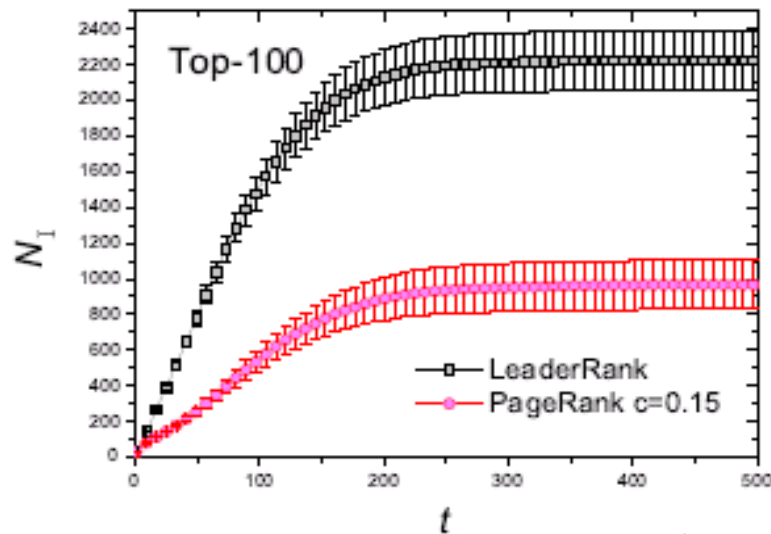
快速收敛

自适应性调节

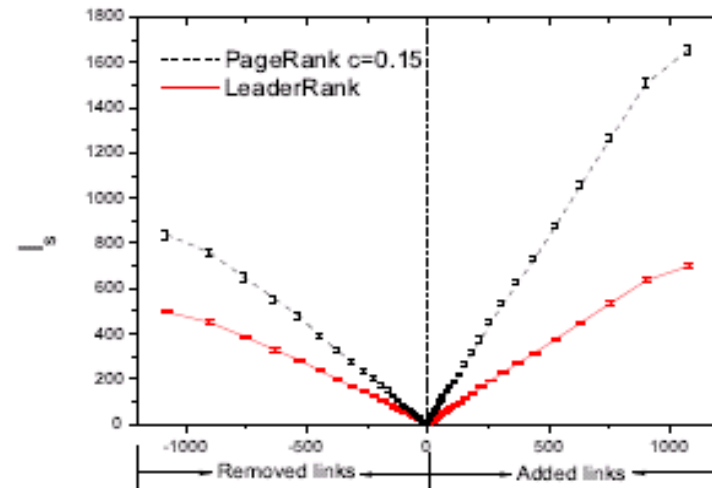
无参数

Experiments on Delicious network

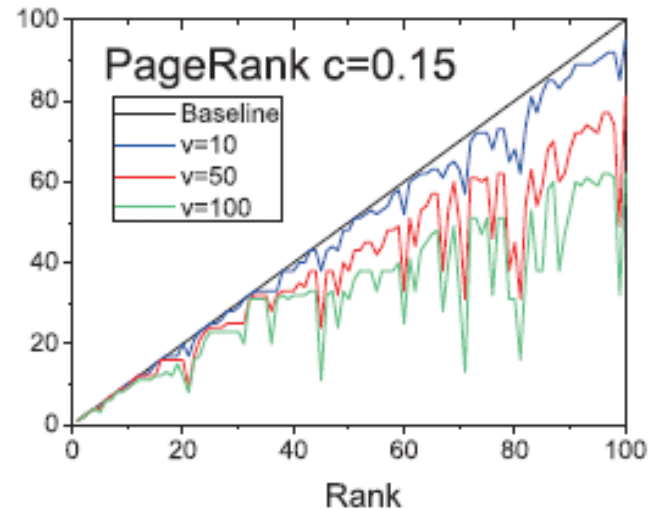
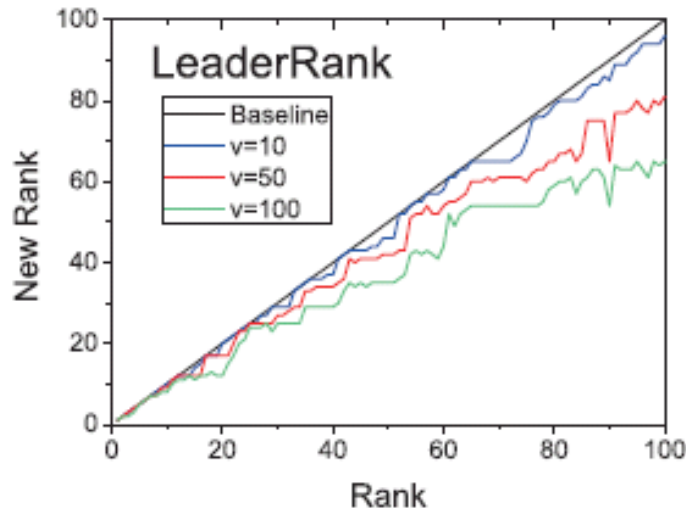
Effectiveness



Robustness



Resilience to spammers



3. 基于渗流理论的节点排序方法

如何识别网络中有影响力的一群人

度中心性

- 刻画节点

介数中心性

- 网络中节点i与j之间有最短路径的条数

接近中心性

- 节点i与网络中所有节点的平均距离越短，节点i越重要，节点i的接近中心性越高

K-壳分解

- 有效识别网络中的核心节点

PageRank

.....



Centrality

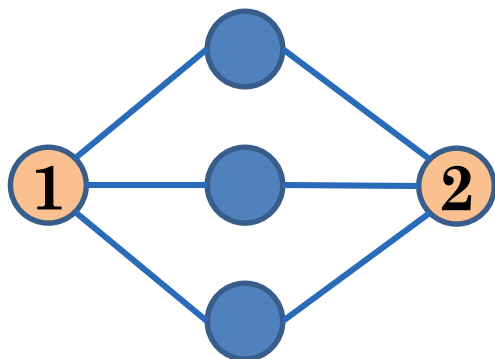
$\frac{1}{N} \sum_{j=1}^N \tau_{st}(i)$
s Centrality

$\frac{\tau_{st}(i)}{\sigma_{st}}$
Centrality

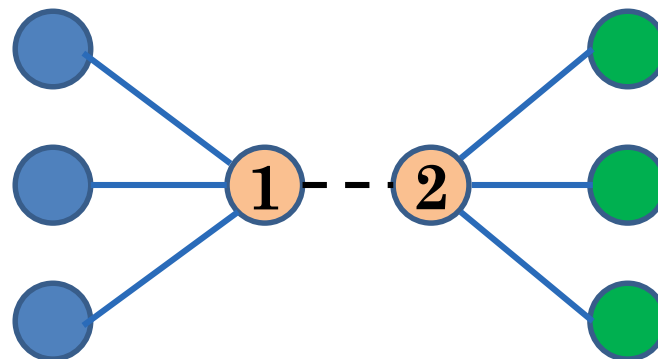
$\frac{1}{N} \sum_{j=1}^N \frac{1}{d_{ij}}$

$\sum_{j=1}^N \frac{A_{ji}}{k_j^{\text{out}}} s_j(t)$

关键问题：降低信息冗余度



$N1=3, N2=3, N1+N2=3$



$N1=3, N2=3, N1+N2=6$

基于网络渗流理论的节点排序方法

PROCEDURES FOR OUR MATHOD

To find L coordinate influential spreaders:

- Step1: each link is reserved (occupied) with p , remove with $1-p$.
- Step2: calculate the size of each cluster, the number of node in it.
- Step3: choose the top- L largest clusters and assign one score to the largest degree node in each cluster.
- Step4: repeat 1-3 for several independent runs, we obtain the final score of a node. All nodes are ranked according to their scores in a descending order and **those L nodes with the highest scores** are suggested to be the set of initial spreaders.

在具有四个社团结构的网络上识别4个节点

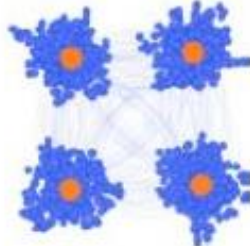
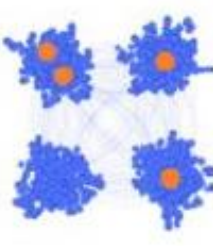
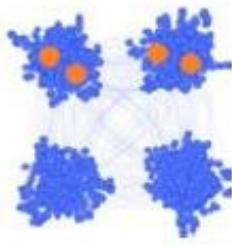
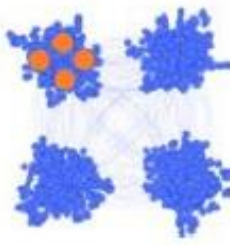
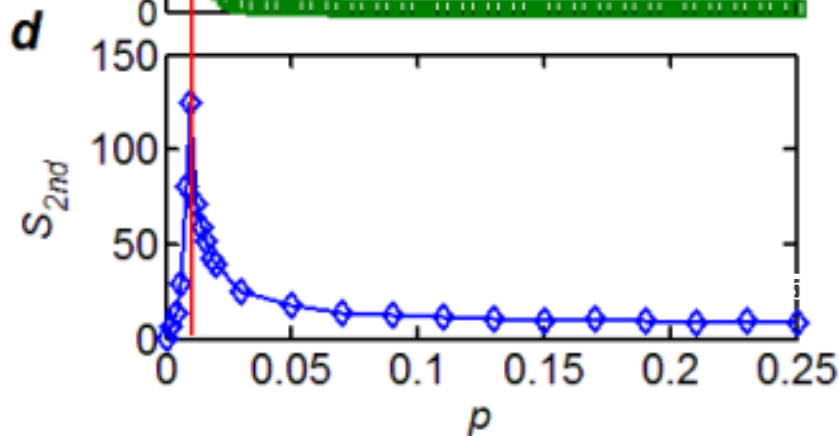
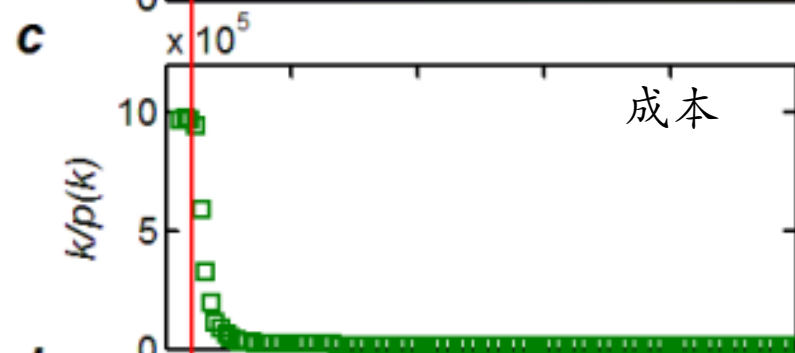
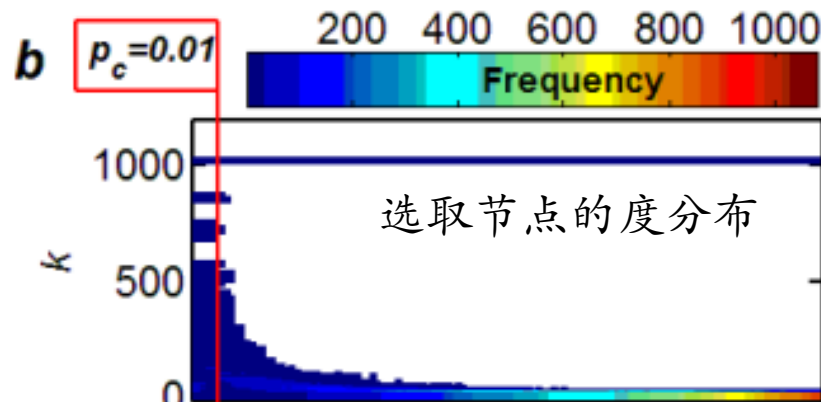
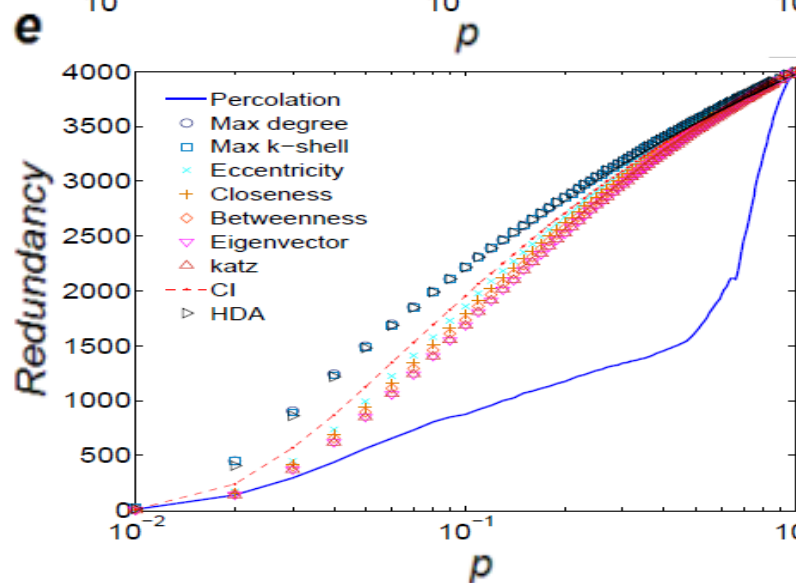
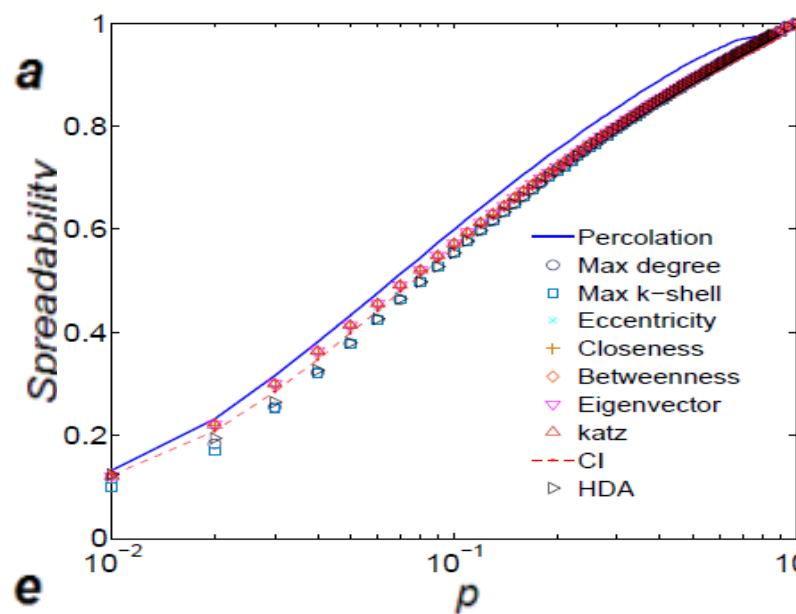
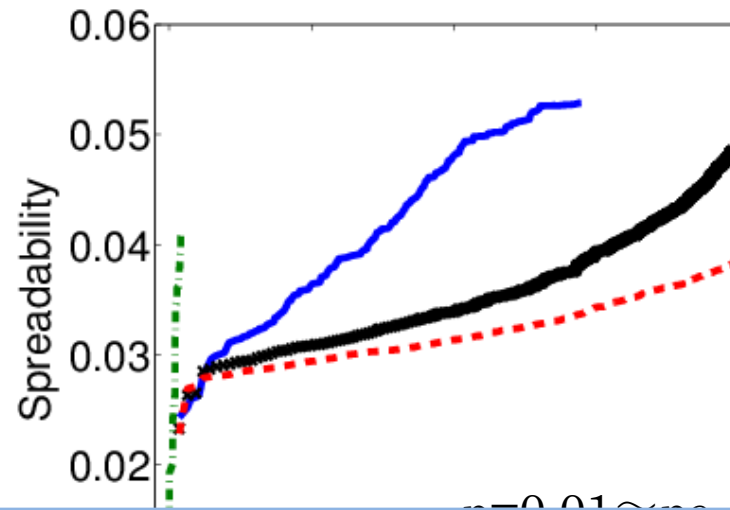
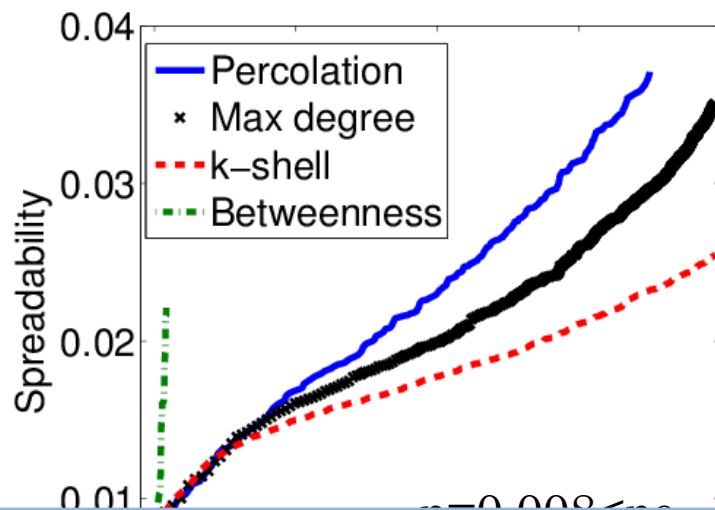
Type					
Percolation	93.9	6.1	0	0	0
Max degree	11	54.4	14	19.5	1.1
K-shell	9.4	43.6	11.9	16.2	18.9
Betweenness	13.1	64.5	11.7	10.6	0.1

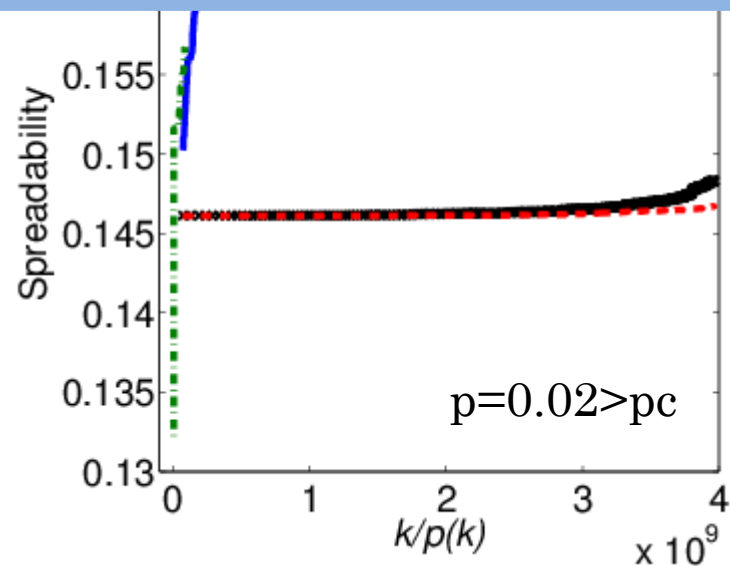
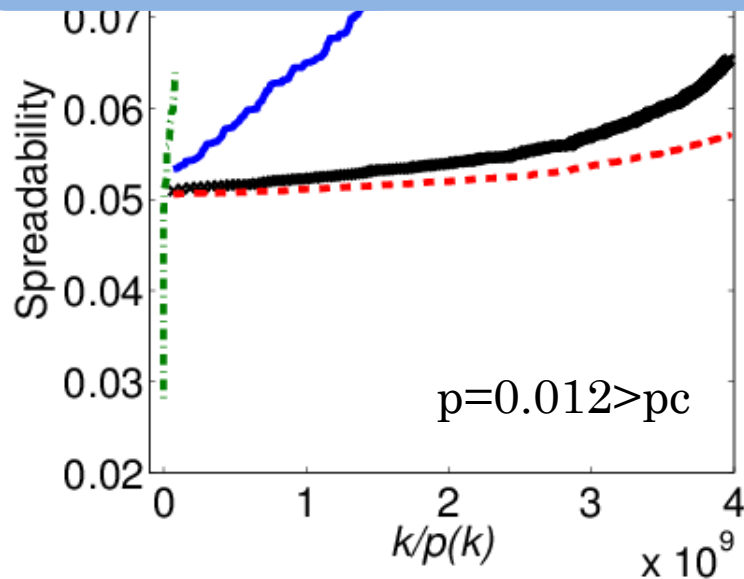
Figure : The results of identifying four influencers on modeled networks with four communities. The procedure to construct the networks is shown in Method and Materials. Each number indicates the probability of the group of selected nodes being in different types. The results are obtained base on 1000 implements. The number 93.9 means that among 1000 implements, there are 939 times that the four selected nodes belong to four different communities. The critical point of the network is about $p_c = 0.27 \pm 0.00$. For our percolation method, we set $p = 0.28$.

在含有6万节点的Facebook中找4000个节点



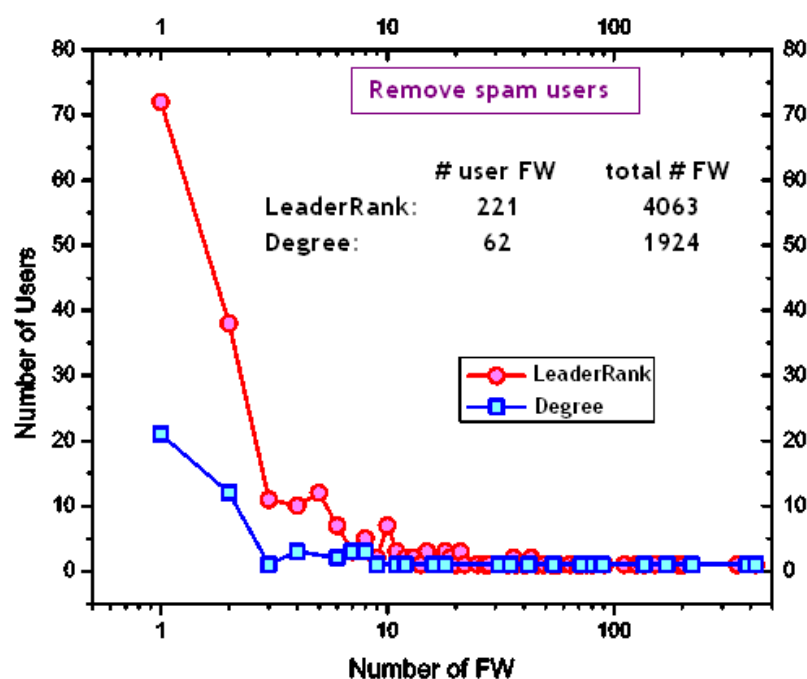
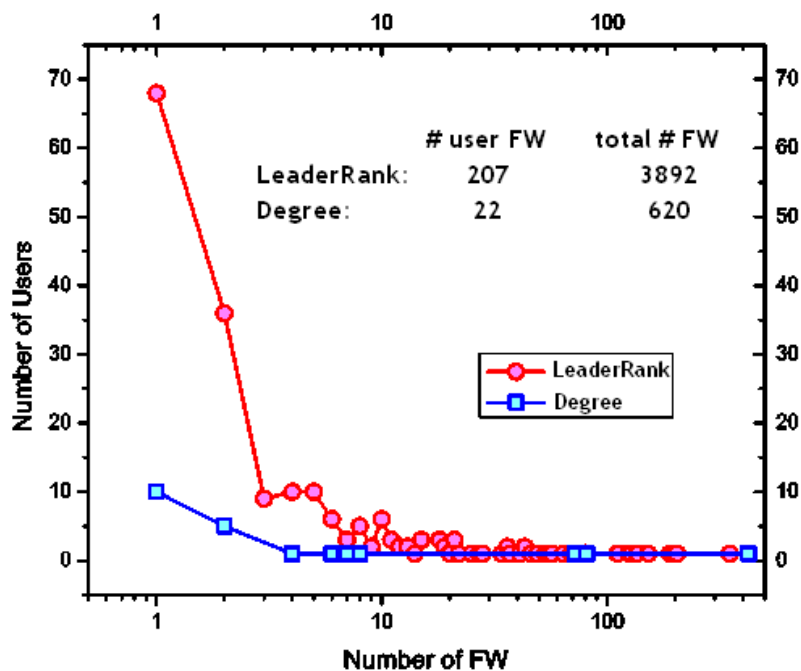


Percolation方法具有高“性价比”！



应用1：短信高转发率用户识别系统

合作公司：中国移动珠海分公司



- 分别用LeaderRank算法和通信用户数（degree）选取1000个种子用户。
- 结果显示LeaderRank算法将效果提高10倍，在去除垃圾用户之后还可以提高3倍多。

应用2：腾讯应用宝好友识别

腾讯社交网络运营部

比随机提高7倍！



比原有算法提高 **?** 倍！

应用3：达推网

——基于社交网络的达人营销系统

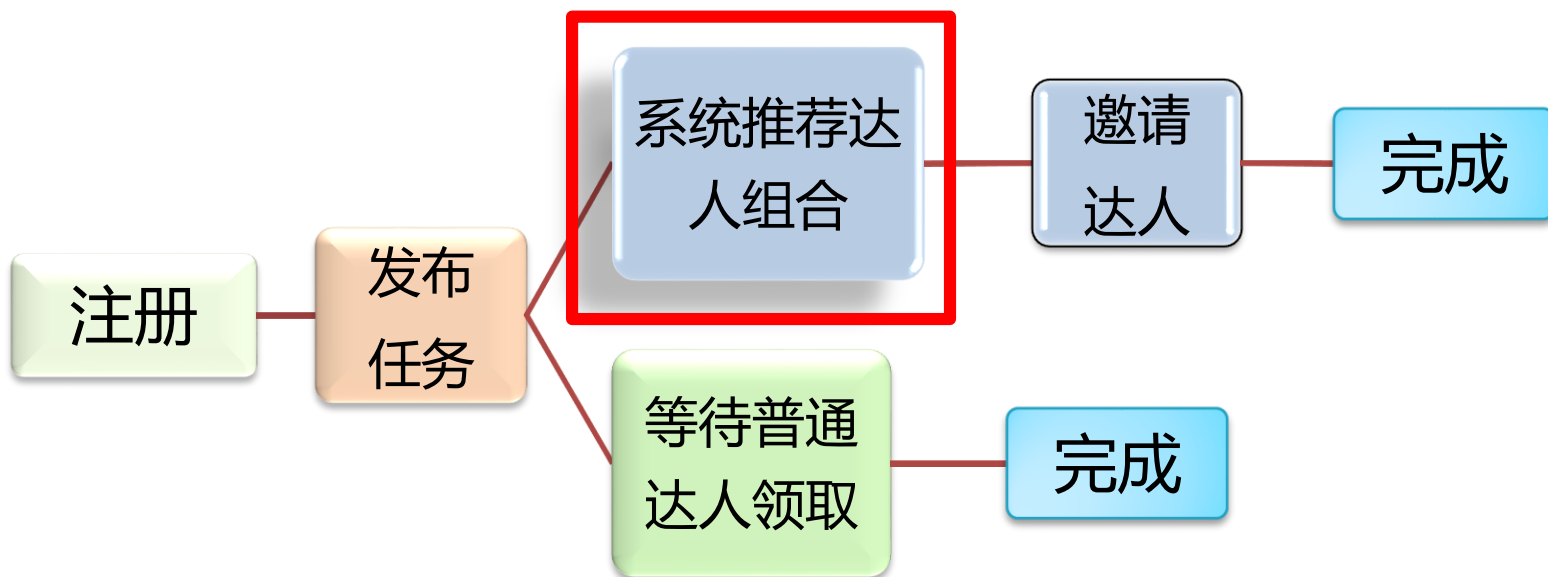
www.idatui.cn



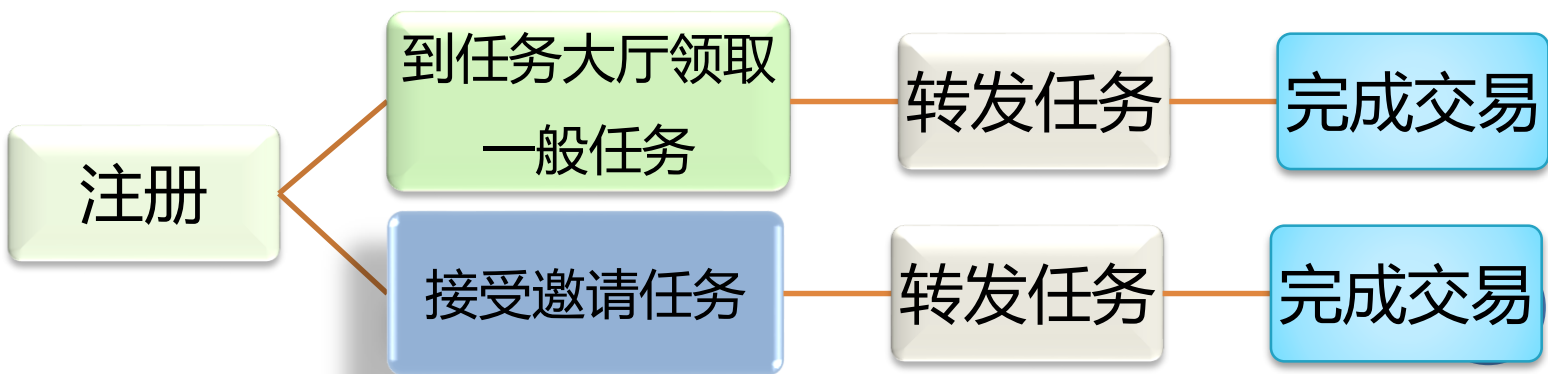
达推网服务流程



中小卖家



达人



总结与展望

○ 主要存在的问题包括

- 一是理论研究跟不上应用需求
- 二是结构重要性和功能重要性脱轨
- 三是理论和真实实验脱轨

○ 八个未来研究方向

- 针对大规模动态变化的网络设计快速有效的重要节点挖掘方法.
- 重要节点对网络特定结构和功能的影响.
- 在相互关联的网络中如何确定一个节点的重要性.
- 利用功能表现挖掘重要节点的反问题研究.
- 含时网络中用户的影响力评价方法及预测.
- 重要节点在网络演化过程中的作用以及对网络控制的意义.
- 用真实网络验证各种算法得出的重要节点的真实影响力.
- 基于已有成果开发实际产品并在真实平台上验证实际应用效果.

参考文献

- **L. Lü**, T. Zhou, Q.-M. Zhang, H. Stanley. The H-index of a network node and its relation to degree and coreness, *Nature Communications*, to be published.
- **L. Lü**, Y.-C. Zhang, C. H. Yeung, T. Zhou, Leaders in Social Networks, the delicious case, *PLoS ONE* 6(6): e21202 (2011).
- D. Chen, **L. Lü***, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, *Physica A* 391, 1777-1787 (2012).
- Q. Li, T. Zhou, **L. Lü**, Chen, Identifying influential spreaders by weighted LeaderRank, *Physica A* 404: 47-55 (2014).
- D.-B. Chen, H Gao, **L. Lü***, T Zhou, Identifying influential nodes in large-scale directed networks: The role of clustering, *PLoS ONE* 8(10): e77455 (2013).
- Y.-B. Zhou, **L. Lü***, M. Li, Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity, *New J. Phys.* 14, 033033 (2012).
- 任晓龙、吕琳媛*, 科学通报 59 (13):1175-1197 (2014).

科学通报 2014年 第59卷 第13期: 1175 ~ 1197

特邀评述

www.scichina.com csb.scichina.com



《中国科学》杂志社
SCIENCE CHINA PRESS

网络重要节点排序方法综述

任晓龙, 吕琳媛*

杭州师范大学阿里巴巴复杂科学研究中心, 杭州 310036

* 联系人, E-mail: linyuan.lv@gmail.com

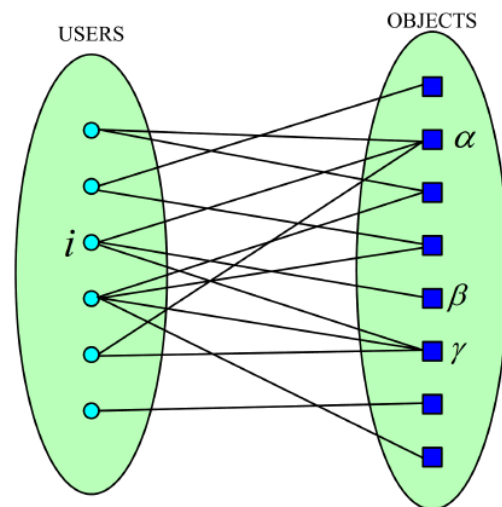
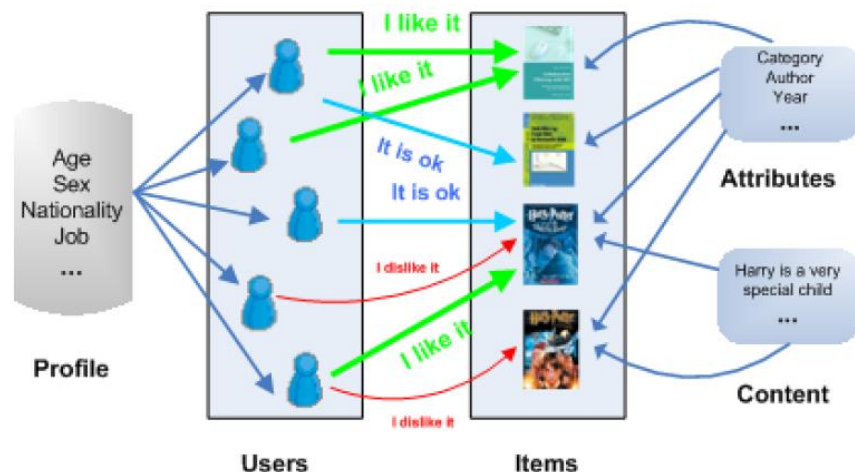
2013-11-21 收稿, 2014-02-25 接受

国家自然科学基金(11205042)、杭州师范大学科研启动基金和 CCF-腾讯科研基金资助

个性化推荐

什么是个性化推荐?

- 推荐系统就是应用已知用户和产品的信息，挖掘用户的喜好，从而帮助用户找到喜欢的产品。
- 可以获得的信息包括：用户曾经购买商品的历史记录，用户的属性信息（性别，年龄，职位...），产品的属性信息，内容，时间信息，用户关系信息等。
- 给定目标用户推荐算法会为其产生一个推荐列表：即将他没有买过的商品按照算法预期的用户喜好程度从大到小排序获得。
- 目标是：尽可能把用户真正喜欢的商品排在前面推荐给用户。



常用评价指标

Accuracy (精确度)

- Overall Ranking: AUC, Ranking Score, MAE, RMSE, PCC, NDPM
- Top Recommended Objects: Precision, Recall, F1-Measure

Diversity (多样性)

- Intra-diversity: Recommendations to a user are diverse
- Inter-diversity: Recommendations to different users are diverse

Novelty (新颖性)

- Popularity
- Self-information

Coverage (覆盖率)

- Measures the percentage of objects that an algorithm is able to recommend to users in the system

$$R_{i\alpha} = \frac{r_{\alpha}}{|U - E_i^T|}$$

$$I_i(L) = \frac{1}{L(L-1)} \sum_{\alpha \neq \beta} s(o_{\alpha}, o_{\beta})$$

$$H_{ij} = 1 - \frac{Q_{ij}(L)}{L}$$

$$N(L) = \frac{1}{ML} \sum_{i=1}^M \sum_{\alpha \in O_r^i} k_{\alpha}$$

$$U_{\alpha} = \log_2(M / k_{\alpha})$$

$$COV(L) = N_d / N$$

M is the number of users

N is the number of objects

J. L. Herlocker *et al.*, ACM Trans. Inf. Syst. 22 (2004) 5

L. Lü *et al.*, Physics Reports, 519: 1-49 (2012)

推荐系统评价指标

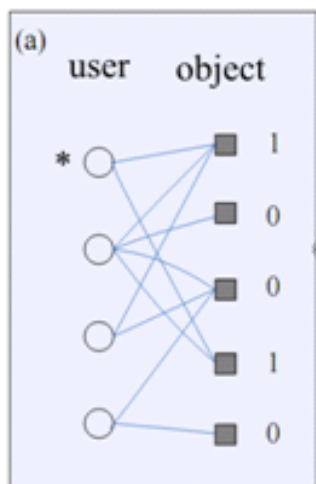
Name	Symbol	Preference	Scope	Rank	<i>L</i>
MAE	<i>MAE</i>	Small	Rating accuracy	No	No
RMSE	<i>RMSE</i>	Small	Rating accuracy	No	No
Pearson	<i>PCC</i>	Large	Rating correlation	No	No
Spearman	ρ	Large	Rating correlation	Yes	No
Kendall's Tau	τ	Large	Rating correlation	Yes	No
NDPM	<i>NDPM</i>	Small	Ranking correlation	Yes	No
Precision	<i>P(L)</i>	Large	Classification accuracy	No	Yes
Recall	<i>R(L)</i>	Large	Classification accuracy	No	Yes
F_1 -score	$F_1(L)$	Large	Classification accuracy	No	Yes
AUC	<i>AUC</i>	Large	Classification accuracy	No	No
Ranking score	<i>RS</i>	Small	Ranking accuracy	Yes	No
Half-life utility	<i>HL(L)</i>	Large	Satisfaction	Yes	Yes
Discounted Cumulative Gain	<i>DCG(b, L)</i>	Large	Satisfaction and precision	Yes	Yes
Rank-biased Precision	<i>RBP(p, L)</i>	Large	Satisfaction and precision	Yes	Yes
Hamming distance	<i>H(L)</i>	Large	Inter-diversity	No	Yes
Intra-similarity	<i>I(L)</i>	Small	Intra-diversity	No	Yes
Popularity	<i>N(L)</i>	Small	Surprisal and novelty	No	Yes
Self-information	<i>U(L)</i>	Large	Unexpectedness	No	Yes
Coverage	<i>COV(L)</i>	Large	Coverage and diversity	No	Yes

朱郁筱，吕琳媛，推荐系统评价指标综述，
电子科技大学学报，41(2):163-175 (2012)

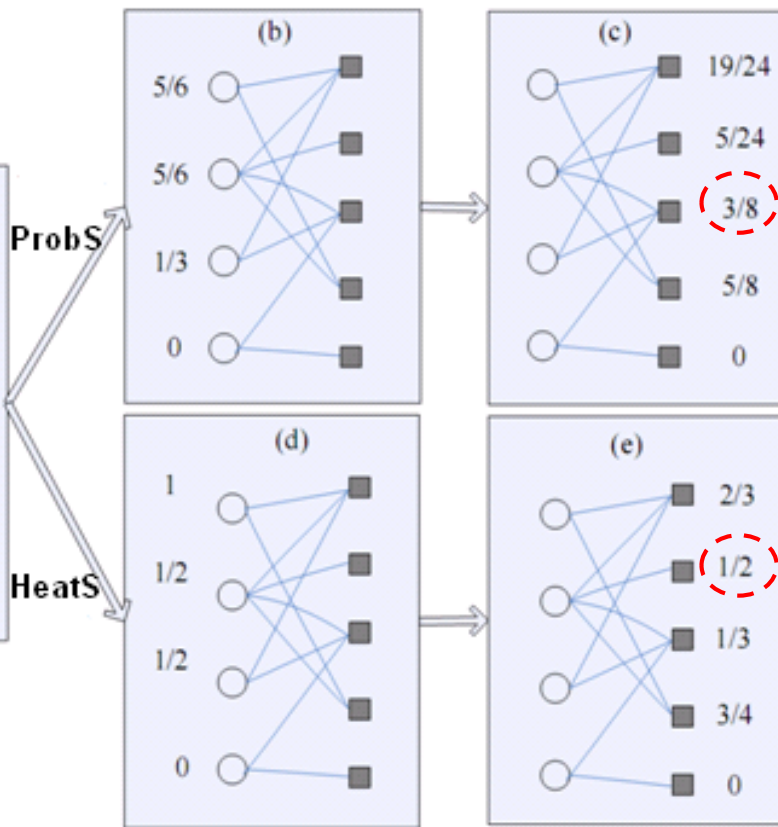
热传导+物质扩散

$f' = Wf$ f is the initial resource vector, W is the transfer matrix

$$w_{\alpha\beta} = \frac{1}{k_{\beta}} \sum_{l=1}^M \frac{a_{l\alpha} a_{l\beta}}{k_l}$$



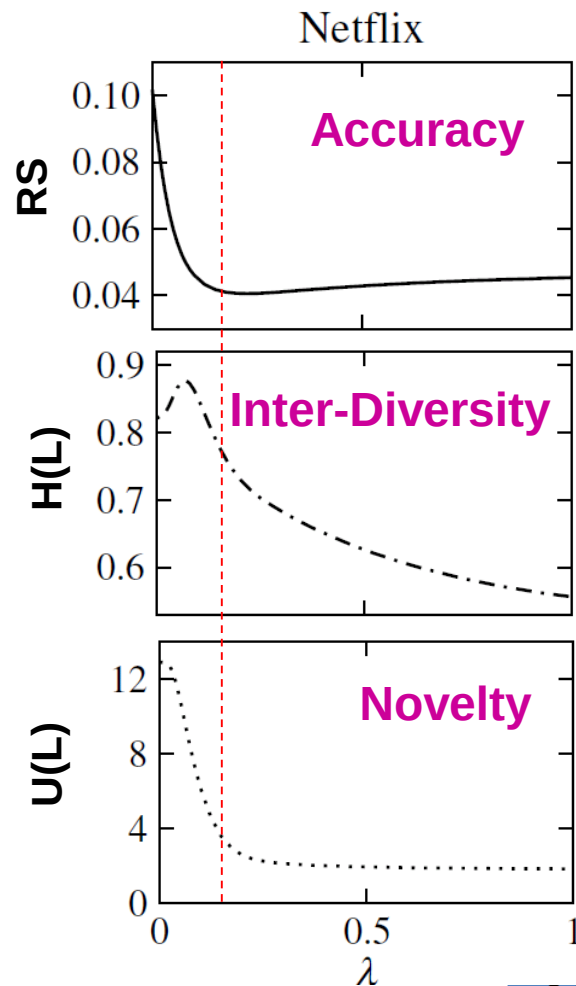
$$w_{\alpha\beta} = \frac{1}{k_{\alpha}} \sum_{l=1}^M \frac{a_{l\alpha} a_{l\beta}}{k_l}$$



Hybrid-PH

$$w_{\alpha\beta} = \frac{1}{k_{\alpha}^{1-\lambda} k_{\beta}^{\lambda}} \sum_{l=1}^M \frac{a_{l\alpha} a_{l\beta}}{k_l}$$

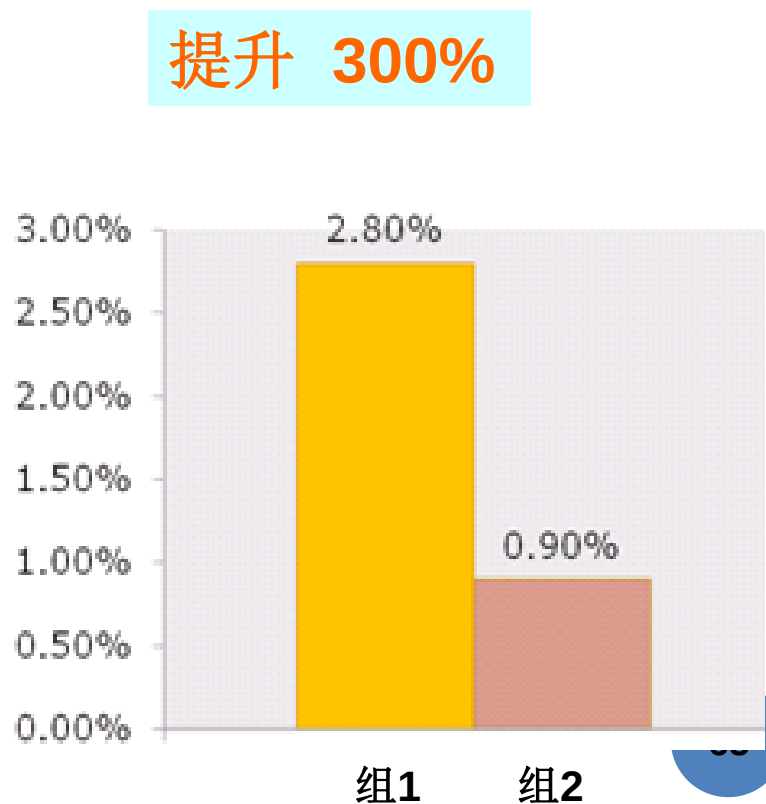
[PNAS 107 (2010) 4511]



应用1：电子书推荐系统（与杭州移动合作）

- 为用户（至少在最近一个月内活跃）推荐他可能感兴趣的电子书.
- 组1: 个性化推荐
- 组2: 按流行度

实验组	推送人数	成功数	成功率
组1	16708	467	2.8%
组2	18415	165	0.9%



应用2：百分点个性化推荐引擎



为300余家电商企业提供个性化推荐服务，每天为超过3000万独立访问用户服务，网页点击超过1.5亿次，日推荐栏展示接近10亿次。

交叉推荐

- 如果我们分析名鞋库的数据时，发现用户2和用户1具有很强的相似性，我们就可以把用户2在果皮网购买的产品推荐给用户1，即使用户1对于果皮网而言是一个全新的用户。

用户进行交叉购物的一个示意图。



解决冷启动问题！

应用3：收宝包收藏夹

好评

销量

商家

信誉
服务
描述
物流

- 可以全网一键收藏的个人购物智能助手。



双十一千万别用收宝包！

扫它抓住收宝包
分分钟成为赏金猎人



收 OR 狗帶

It's a question.....



100000000000

Bounty -

CONFRONTI DI PRESSIONI E TEMPERATURE IN UNO DEI
SISTEMI DI MONITORING: SCELTA DEGLI STRUMENTI E
VALUTAZIONE DELLA PRESSIONE E TEMPERATURA IN UNO DEI

SHOU

阿里巴巴商学院出品

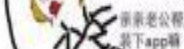
尼美啊！
我回来了
想我没？



我保证不装淘宝京
东折800.....



海宝京东一号店曹燕街新八巴
通通都加神。看你这个沙雕怎
么玩



像這，我有這個



FFFFF
FFFFF
FFFFF
FFFUU
UUUU
UUUU
UUUU
UUUU
UUUU

天！竟然可以全网收藏，
智能评价。天。不错，我也去
try try



拿去



几个月后



這凌亂的頭髮！

这不整的表轻里

你有多久了？

龙舌！



老舍



拉！



老公！把信刷
我你的卡了！
你用爆了！



媒体关注

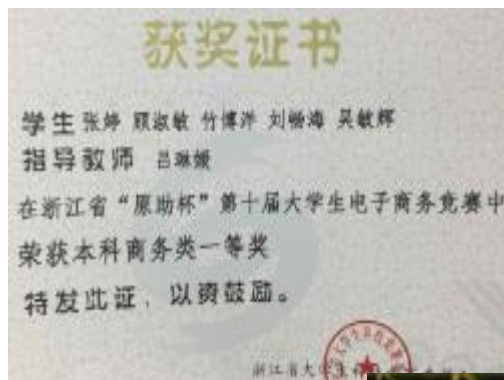


顾淑敏团队那个悬在半空的创意，是一个全网一键收藏宝贝的购物管家，叫做“收宝包收藏夹”。它能根据每个人的购物喜好，给某件网购产品打个分。这个分数，可以成为用户网购时的参考。

这个想法得到了万善的认可，供职于电商企业的他，深知基于人性的个性化分数，才是用户真正需要的购物指数。他对团队的建议只有三个字，那就是把算法“做夯实”。

答辩结束后，评委之一、迈阿顿创业投资有限公司总经理王进就拉着顾淑敏和她的两个小伙伴张婷、刘畅海见了几个投资人。看来，他们动的小心思——现场招兵买马，最后还真的奏效了。

2015年8月 杭州新闻媒体60分对我们团队的采访



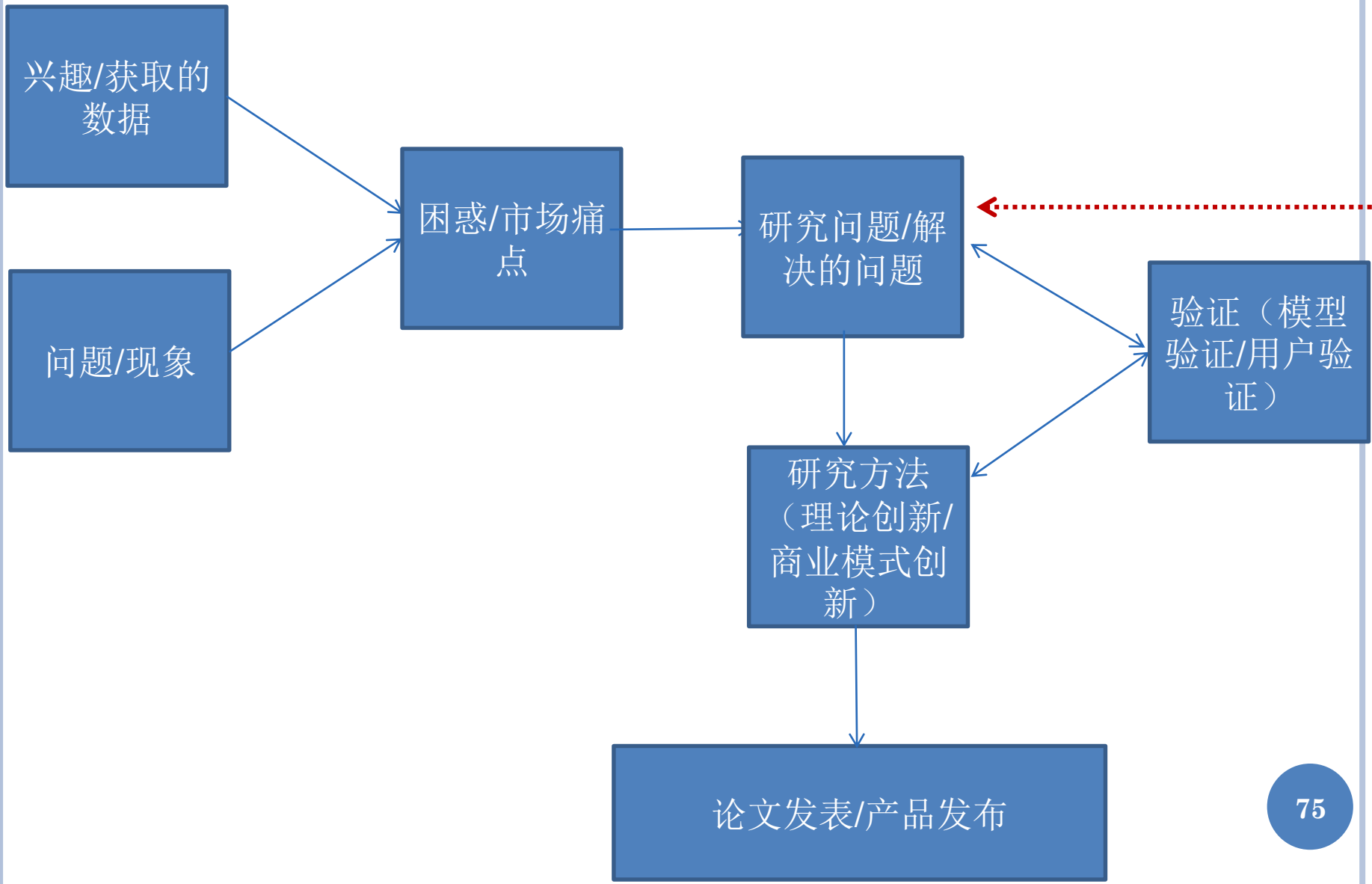
11月20日 省长李强、市长张鸿铭莅临





科研 VS 创业

- 定义“问题”：从兴趣和数据出发找到问题
- 找到“巨人”：相关文献搜索
- 搭好“云梯”：了解前人的工作，看到不足，排除雷区
- 站上“肩膀”：在前人基础上创新理论，解决问题或更好地解决问题，理论验证
- 一览众山：撰写论文，发表论文，分享成果。
- 定义“问题”：从兴趣和数据出发找到问题，市场痛点
- 找到“巨人”：已有哪些竞争对手
- 搭好“云梯”：了解竞争对手的方法，分析不足，找到切入点
- 站上“肩膀”：创新方法和商业模式，解决问题或更好地解决问题，内测验证，天使用户反馈
- 一览众山：产品发布，拜见观众。



「修身、齐家、治国、平天下」

- 四书五经《大学》里有一句话说儒家的理想，做一个君子应该怎么样：修身齐家治国平天下。
- 1、修身：作为科研团队的负责人，或者创业公司的创始人，你都是支柱，如果你不能不断**提升**自己，那么整个团队和公司都不可能提升的，事情也不可能做大。所以要不停的学习和提升，把握科学研究的前沿，了解创业环境趋势。
- 2、齐家：对科研而言，家就是课题组成员。对创业而言，家就是创业团队成员。由于个人精力是有限的，团队成员的质量、投入度都直接影响你是否能最终治国平天下。齐家，就是需要**协调**团队中不同性格、不同分工、不同角色成员的关系。

- 3、**治国**：你不可能自己去干每个事情，要有人查文献、写算法、编程序，你需要**管理**好团队。国就是一个生态系统，包含了多个群落。比如我的团队就分为理论研究组+应用实践组。前者做基础研究，后者做项目和产品开发。两个群落有各自的管理文化、评价标准。比如，基础研究组毕业可能需要几篇SCI文章和国际会议交流，而实践组要求能够完成一个具体的产品开发和设计，有1-2项专利等。这两个组共同构成一个生态。
- 4、**平天下**：就要求你能够占据一个独有的山头。所谓「天下」就是你的目标市场或科研领域，做科研就是要有自己的研究领域，做创业就是要占据细分市场。按照互联网创业思想，互联网创业成功的关键是垄断，而非竞争。这一点其实在今天的科研领域也适用。「平」强调的是你不仅要**彻底垄断**。一种模式：广撒网，多点开花。另一种模式：深耕一个细分领域，做深做透，所谓的“一针捅破天”。

选择一个好天下至关重要！

不管是创业，还是科研，归根到底都是要解决问题的，一个是科学问题，一个是市场和用户问题。如果你解决的问题不够大，你创造的价值就不够大，所以选择重要的有价值的研究问题和市场问题！

大天下，大梦想！



梦想极像了一个魔法棒，他可以让千疮百孔的人满血复活。



然后你会发现，生活中没有一条路是白走的，每个转弯都有它的意义，而至于未来怎么样，你要努力走下去才知道啊！

坚持梦想，活在当下！





Thank you!

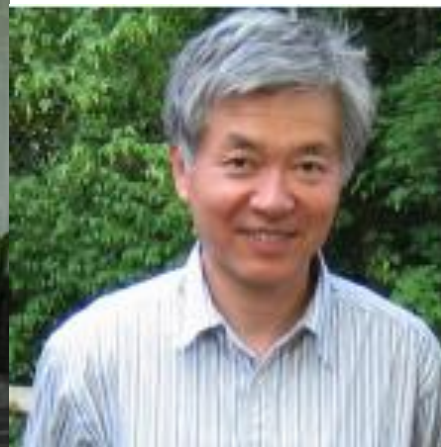
感谢曾经帮助过我的人，感谢生命中的一切挫折，
也感谢自己，感谢生活对我拳打脚踢之时，我仍甘之如饴！

阿里巴巴复杂科学研究中心

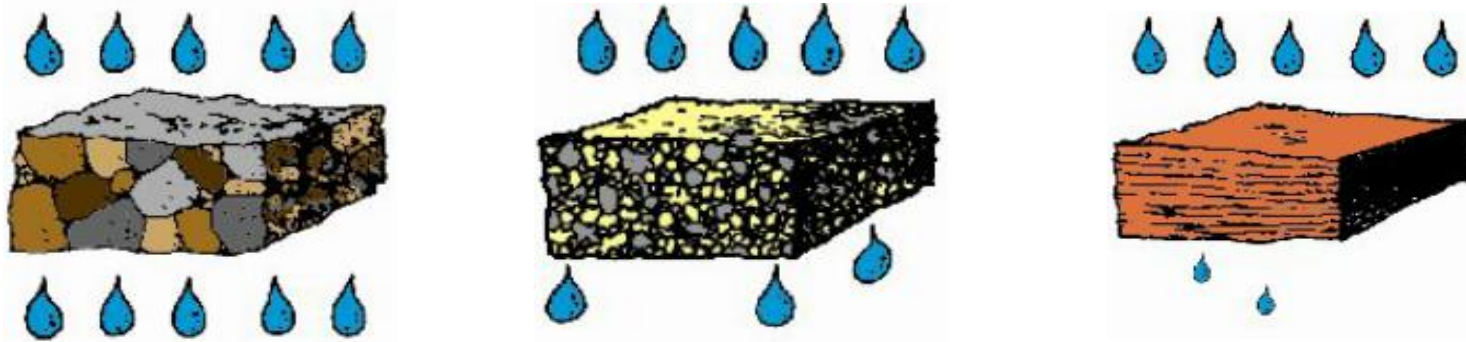
- 成立时间：2013年9月14日
- 地址：杭州师范大学怒园1号楼（近阿里巴巴淘宝城）
- 中心主页：<http://www.eugene.cn>
- 宗旨：以基础科学研究为主，立足产业实践前沿，结合产业发展需求
- 研究领域：人类行为动力学，社会经济复杂性，复杂系统信息挖掘
- 开放基金：支持博士生，博士后，青年教师从事科学研究工作
- 中欧联合实验室：联合申请国际合作项目&联合培养研究生



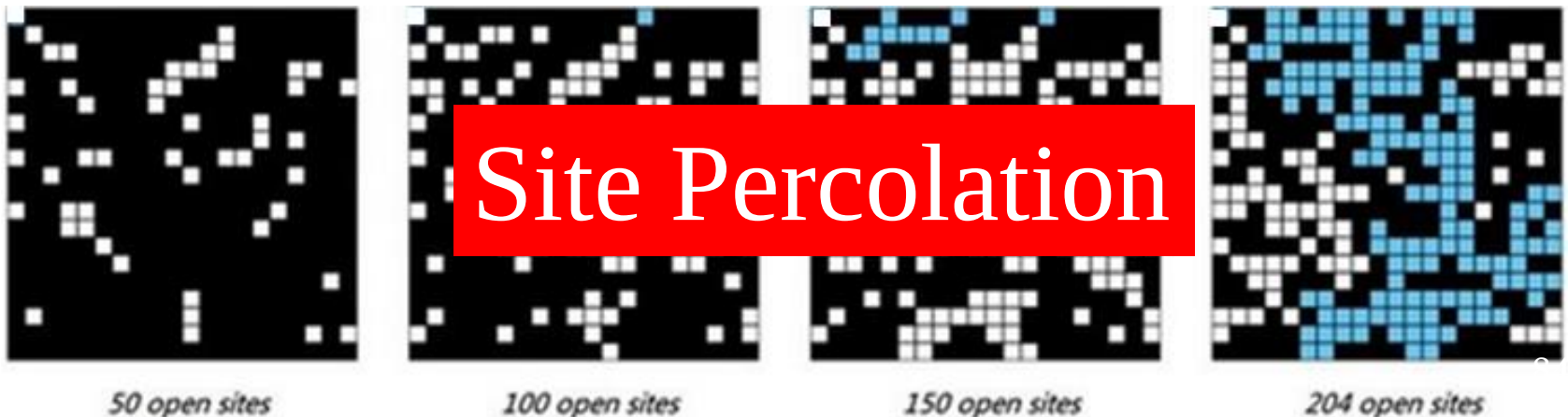
欢迎大家来杭州访问交流



什么是复杂网络上的渗流？

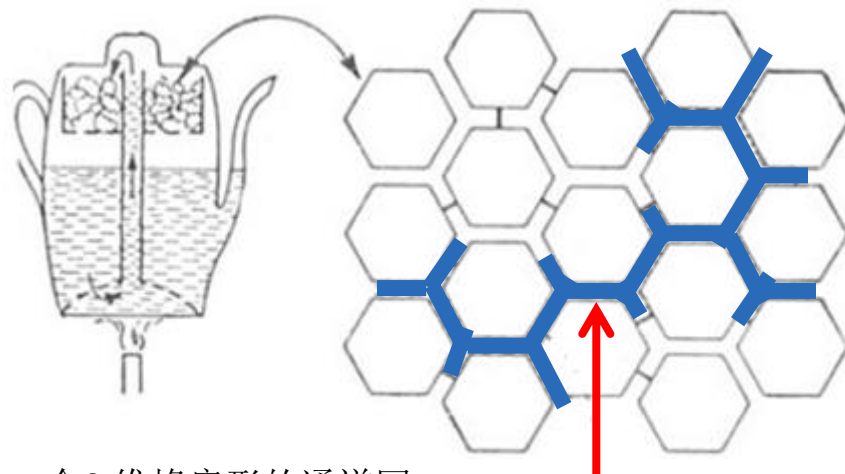
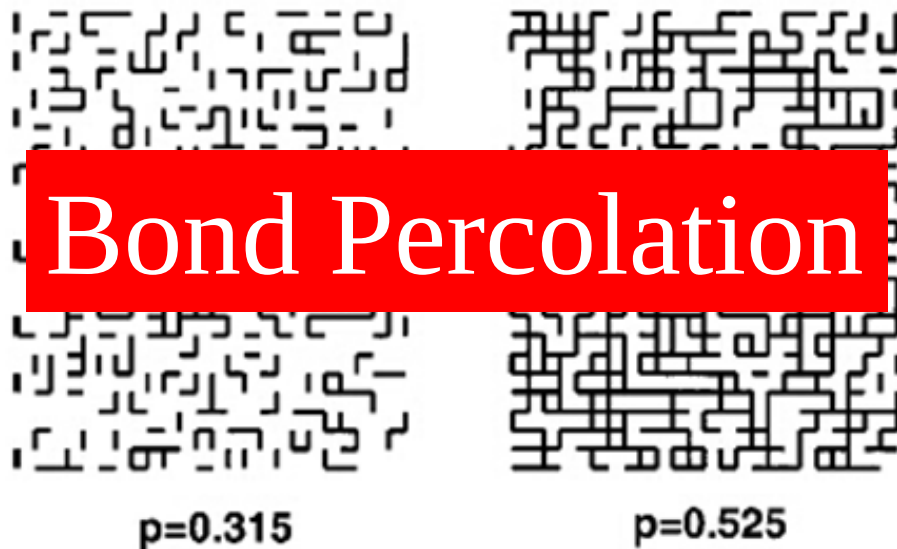


- 渗流（percolation），也称逾渗，这个词汇是由数学家J.M. Hammersley 在1957年创造的，最初目的是为了描述流体在无序介质中作随机的扩展和流动。



正方形点阵上座逾渗的临界浓度 $p_c=0.59$

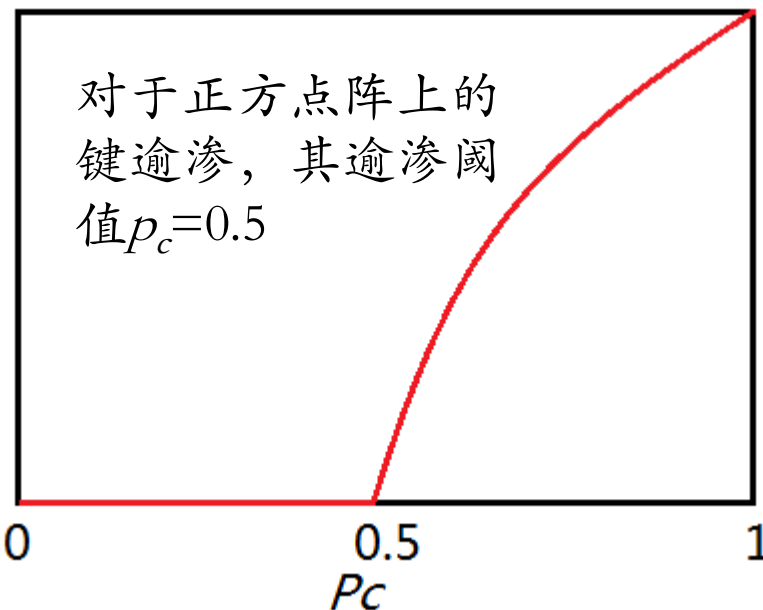
Bond Percolation



一个2维蜂房形的通道网络中的键逾渗连通集团

渗流通路

极大簇



连通集团:

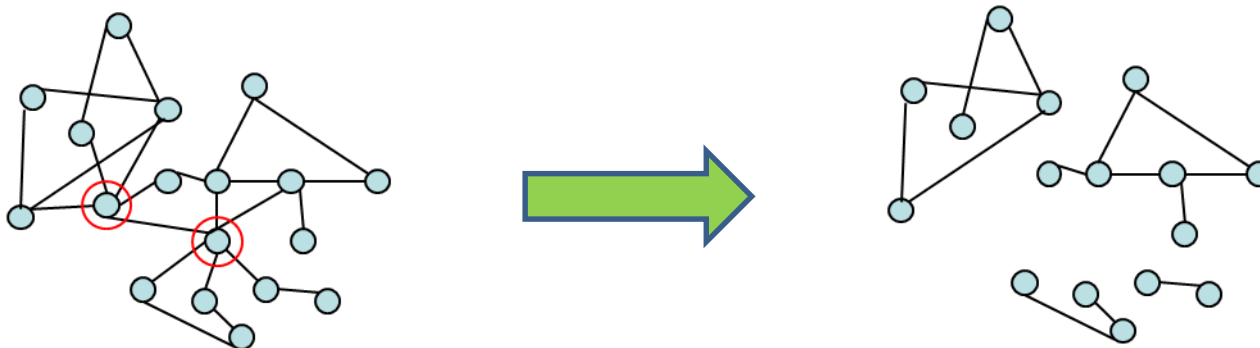
- 座逾渗, 若两个已占座(点)可以通过由一系列最近邻的已占座连成的路径连接起来, 则这两个座属于同一集团, 所有属于同一个集团的座构成一个连通集团。
- 键逾渗, 若两条键(边)可以通过至少一条由联键连成的路径连接起来, 则这两条键属于同一集团。

渗流通路, 也称渗流集群, 无限集群, 其规模随着网络规模增大而扩展。在 p_c 以上, 系统存在逾渗通路, 在此以下就不存在逾渗通路。

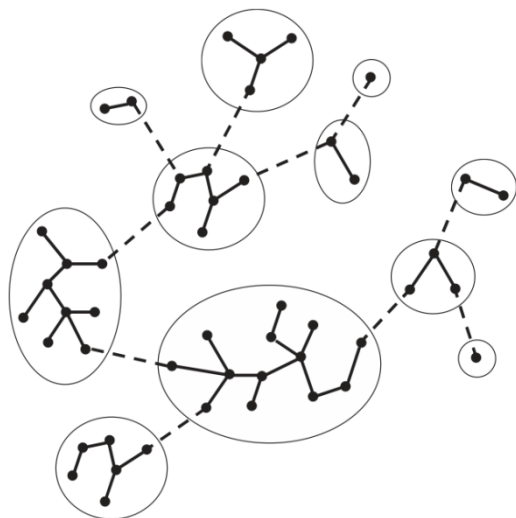
网络渗流与传播过程

Site Percolation

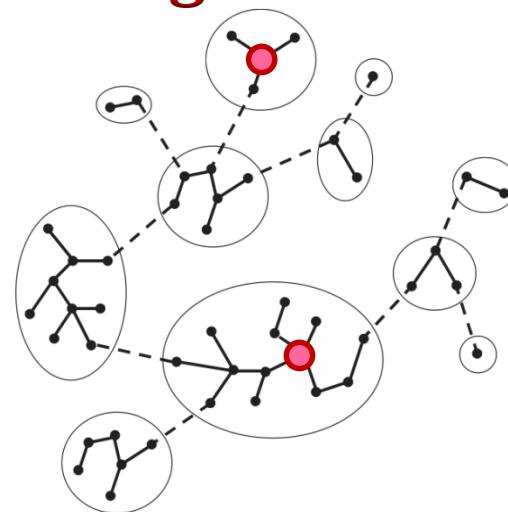
每个点以概率 p 保留，以概率 $1-p$ 移除



Bond Percolation



Spreading (SIR)



每条边以概率 p 保留，以概率 $1-p$ 移除

已感染节点以概率 p 感染其易感邻居节点

推荐的书

- 《网络科学导论》
- 《链路预测》
- 《智慧社会》
- 《复杂》
- 《涌现》
- 《链接》
- 《爆发》
- 《大自然如何工作》
- 《预知社会》
- 《黑天鹅》
- 《反脆弱》
- 《大数据时代》
- 《人类简史：从动物到上地》
- 《科技想要什么》
- 《商业的常识》
- 《精益创业》
- 《人生十论》