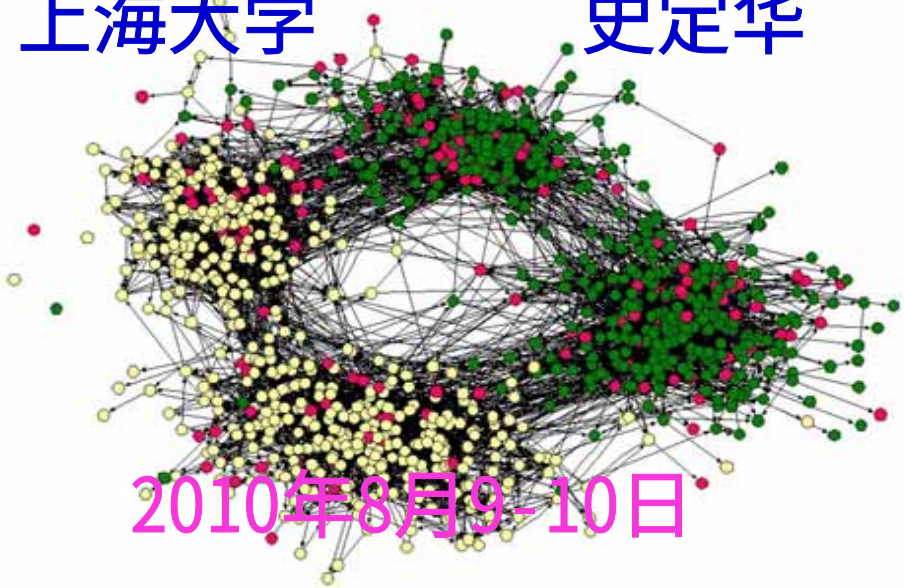
系统科学与复杂网络专题研讨会

无标度网络：

基础理论和应用研究

上海大学　　史定华



2010年8月9-10日

引起轰动的重要文献

小世界网，无标度网，可导航网

[1] D. J. Watts, S. H. Strogatz, *Nature* 393, 440(1998)

[2] Barabási A.-L. and Albert R., *Science* 286, 509(1999)

[3] J. M. Kleinberg, *Nature* 406, 845(2000)

无标度网的系列文章

[4] Albert R., *et al*., *Nature* 401, 130(1999)万维网直径

[5] Jeong H. *et al*., *Nature* 407, 651(2000)代谢网络统计

[6] Ravasz E. *et al*., *Science* 297, 1551(2002)层次网络

[7] Albert R., *et al*., *Nature* 406, 378(2000)稳健而又脆弱

无标度网：过去与未来

[8] Barabási A.-L., *Science* 325, 412(2009)

无标度网络：过去与未来[8]

- 网络并非固定结点随机连线，而显示出增长和择优连线，使得网络度分布为幂律。
- 实际网络都收敛于类似结构与年龄规模无关。
- 网络的结构和演化不可分割，即动态非静态。
- 无标度网络涌现网络核心，故稳健而又脆弱。
- 除非探讨网络拓扑，否则无法理解复杂系统。
- 还没有发现能够解释网络动力学的普适框架。
- 复杂系统单元之间连接(相互作用)如此重要，这正是现在我们关注网络的原因。

1. 无标度网和B-A模型

Barabási和Albert在《科学》上发表的“随机网络中标度涌现”论文[2]尤其引人关注。根据对万维网结构调查得到的情况，论文的主要思想有：从许多实际复杂网络度分布的统计结果发现具有幂律尾部是其普遍的特征。他们提出一个择优增长的动态模型去解释产生幂律的机制；并且认为正确的模型其网络度分布应该独立于时间；从此具有幂律度分布的网络被称为无标度**(scale-free)**网络。

B-A模型及其不明确处

B-A模型定义

(1)初始：开始给定 m_0 个结点；

(2)增长：在每个时间步增加一个新结点和m条新连线；

(3)择优：新结点按照度择优概率 Π 选择旧结点i与之连线，不允许重复连线。

Bollobás等人[9]指出B-A模型不明确处

(1)初始网络没有设定，孤立结点无法择优连线；

(2) $m > 1$时如何进行择优连接？ 依次还是同时；

(3)旧结点i获得连线的概率为什么等于$m\Pi$？

Bollobás等人的LCD模型[9]

允许结点自连线和结点之间重复连线

LCD模型定义

(1)开始一个顶点和一条自连边的图 G_1^1 **；**

(2)从 G_1^{t-1} **到** G_1^t **择优增长概率** $\Pi(i=s)=\begin{cases} k_s/(2t-1), 1\le s\le t-1 \\ 1/(2t-1), \quad s=t \end{cases}$ **；**

(3)将图 G_1^{mt} **顶点合并得图** G_m^t **。**

LCD模型讨论

(1)定义明确，图过程是动态的，却有静态的描述；

(2)*m*=1时，与B-A模型也不相同；

(3)旧结点*i*获得连线的概率不容易确定。

Dorogovtsev等人的吸引模型[10]

只允许结点之间重复连线

吸引模型定义

(1)初始：两个结点一条连线；

(2)增长：在每个时间步增加一个新结点和m条新连线；

(3)择优：新结点按照度择优概率Π选择旧结点i与之连线，允许重复连线。

吸引模型讨论

(1)定义明确，$m=1$时与B-A模型相同；

(2)旧结点i获得连线的概率好确定，但不等于mΠ。

Holme和Kim的模型[11]

不允许自连线和重复连线

H-K模型(特殊情况)定义

(1)初始m+1个结点全连通；

(2)每步增加一个新结点和m条连线；

(3)第1条按度择优概率Π选择旧结点i与之连线，其余m–1条在选中i的邻域随机连线。

H-K模型讨论

(1)定义明确， m=1时与B-A模型相同；

(2)旧结点i获得连线的概率精确等于mΠ；

(3)群集系数大，不是所有连线都择优(有点模块择优)。

时间独立还是时间相依[12]

正确的模型其网络度分布应不应该独立于时间?

B-A模型通过固定加线数和度择优连线来反映与时间无关。即使如此，也只能近似地保证$P(k,t)$与t无关。

只有复制模型[13]：(1)随机选一旧结点复制为新结点；(2)新结点指向旧结点连一条线；(3)新结点再复制旧结点指向更老结点的连线，才能保证$P(k,t)$与t无关。

实际网络度分布必须考虑网络有限规模的影响!

代替假定 $P(k) \sim k^{-\gamma}$，提出研究时间相依的幂律度分布

$$P(k,t) = C(t)k^{-\gamma}$$

其中度指数相对网络规模稳健，系数需要重点探讨。

是无标度还是标度丰富[14]

度分布不是网络拓扑的唯一指标

加洲理工学院Li等人[14]认为仅以度分布是幂律作为无标度网定义不尽合理，因为幂律是方差跨度极大的高可变分布，由于存在极大的波动性，所以幂律相同的度分布网络完全可能具有极其不同的拓扑结构和特性。因特网和代谢网就是两个典型的案例，它们面对蓄意攻击仍然是“稳健而非脆弱”。

提出网络无标度程度的测量问题

因特网和代谢网的度分布都是幂律，根据他们的测度计算，两者都是标度丰富的网络。

2. 代谢网络和层次网络模型

Barabási等人在《科学》上发表的“代谢网络模块化层次组织”论文[6]是另一篇重要论文。**Barabási**等人[5,6]统计了大量的代谢网络，发现它们是无标度的，度指数接近**2.2**以及群集系数与度***k***存在反比关系。

为此，他们提出可以用**4**结点全连通模块生成的层次网络来建模。他们得到该层次网络是无标度的，其度指数接近**2.26**，以及群集系数与度***k***存在反比关系。$C(k) \sim k^{-1}$**(**存在标度指数问题？**)**

不知何故？总结没有提及这两篇文章。

几何增长网络模型

文献上对其度指数一直颇有争议

- 第一个确定性层次网络， Barabási得到度指数

$$\ln 3/\ln 2 \qquad (1)$$

- 后来觉得不妥，他们又将度指数修正为

$$\gamma = 1 + (\ln 3/\ln 2) \qquad (2)$$

- 这也是Dorogovtsev[15]对伪分形图得到的度指数。
- 在PRL上引入阿波罗网[16]得到的度指数也是如此;
- 然而四年后，在PRL上又出现反复，重新回到(1)。

这会不会是总结不提及的原因？

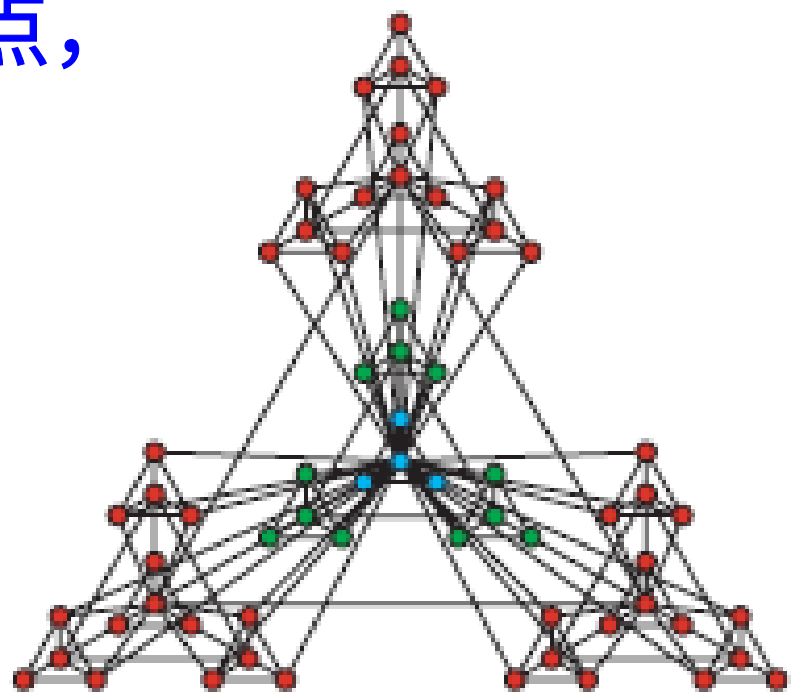
网络度分布统计方法

- 代谢网络以大肠杆菌(E. Coil)为例

总结点数为851，入度 715，出度 702。

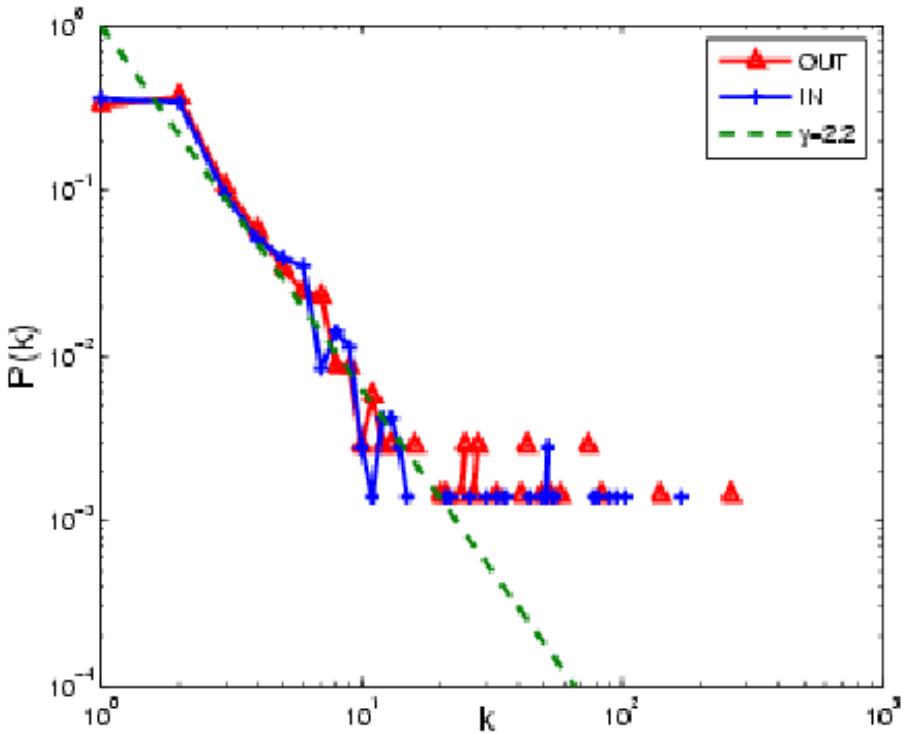
- 层次网络以4结点连通图模块为例

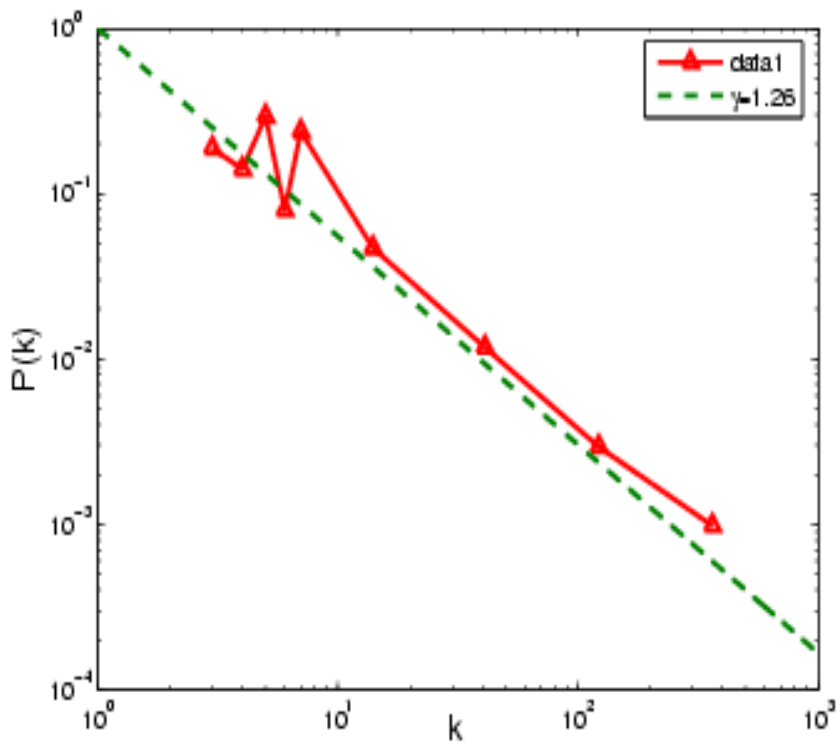
当n=5时有1024个结点，规模就将超过。网络只有度3，4，5，6，7，14，41，122，结点数192，144，300，81， 243，48，12，3，1。



方法1——画度频率图

- 度频率方法





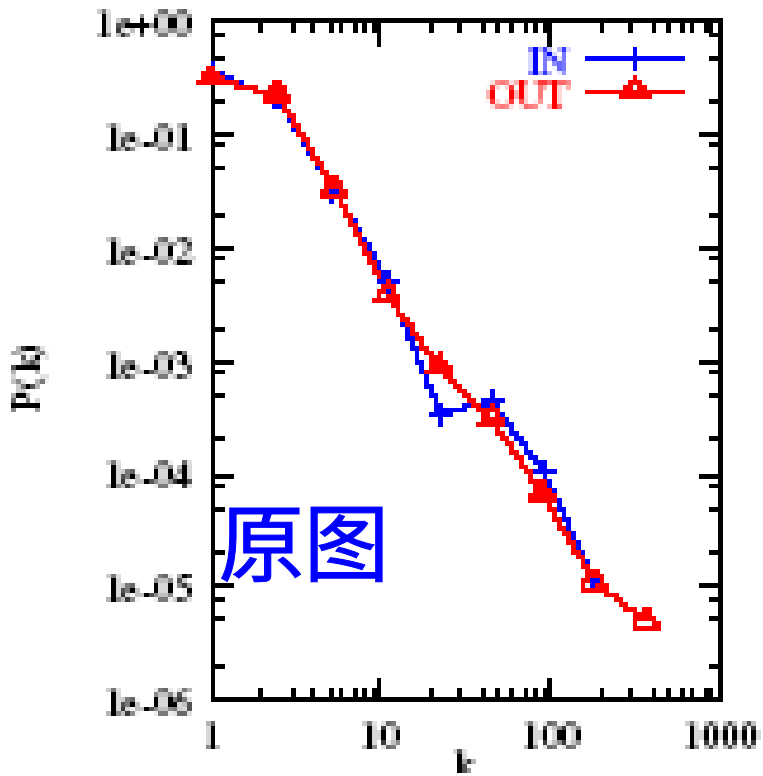
- 代谢网络度指数2.2；层次网络度指数1.26
- 常用方法，尾部摆动，容易误导

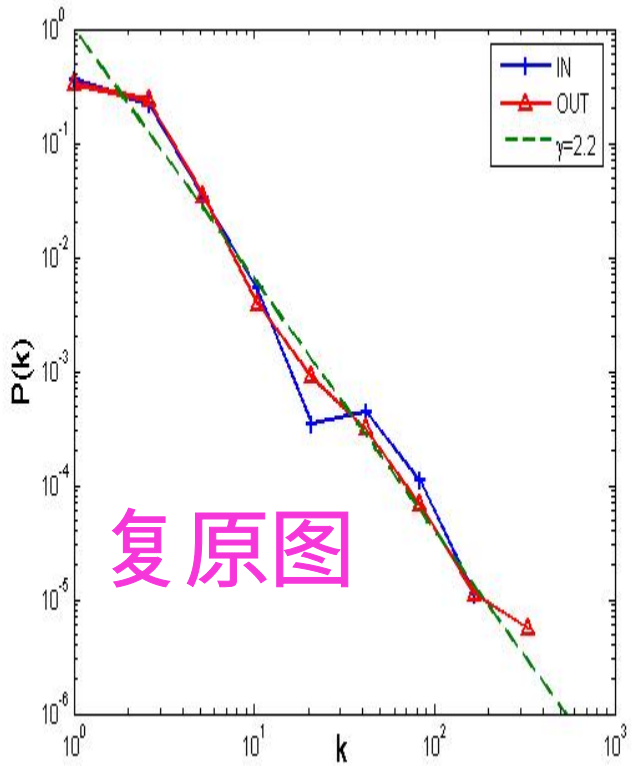
方法2——logarithmic binning

☞logarithmic binning对数盒子方法

中心：1.3, 2.6, 5.2, 10.4, 20.8, 41.6, 83.2, 166.4, 332.8

盒子：[1,2)，[2,4)，[4,8)，[8,16)，…，[512,1024)

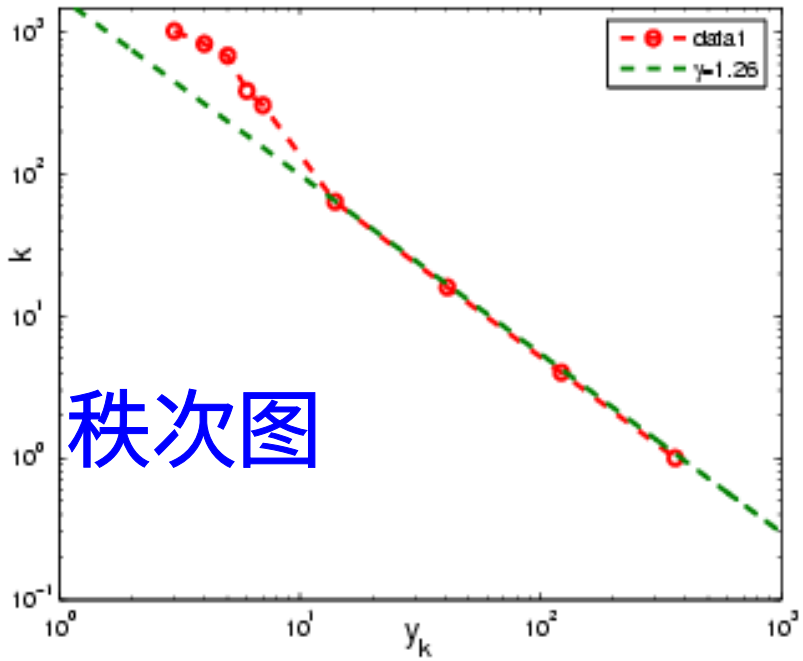




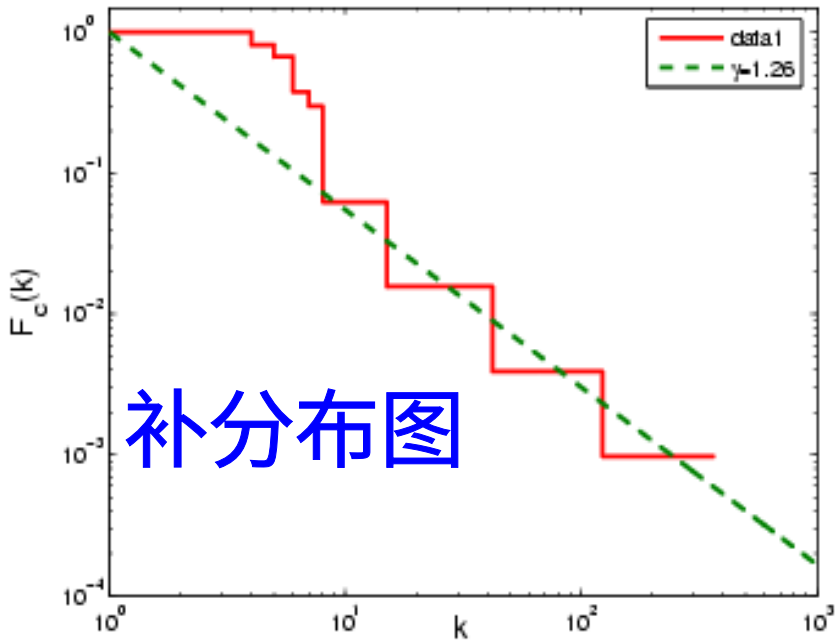
方法3——画秩次图

上述logarithmic binning比较陌生(代谢网)

☞度秩次补分布方法



秩次图



补分布图

☞推荐方法，指数加1，比较准确

实际网络的恰当模型

- B-A模型不是因特网络恰当模型，因为因特网标度丰富[17]，稳健而非脆弱！
- 层次网络模型也不是代谢网络恰当模型，因为代谢网络标度丰富[18]，稳健而非脆弱！
- 启发式最优模型可能更适合作为因特网和代谢网络模型[17,18]。
- 部分复制则可能更适合作为万维网的模型[13]。等价于确定的自然数整除网络，见最后面。

3. 无标度网络的动力学特性1

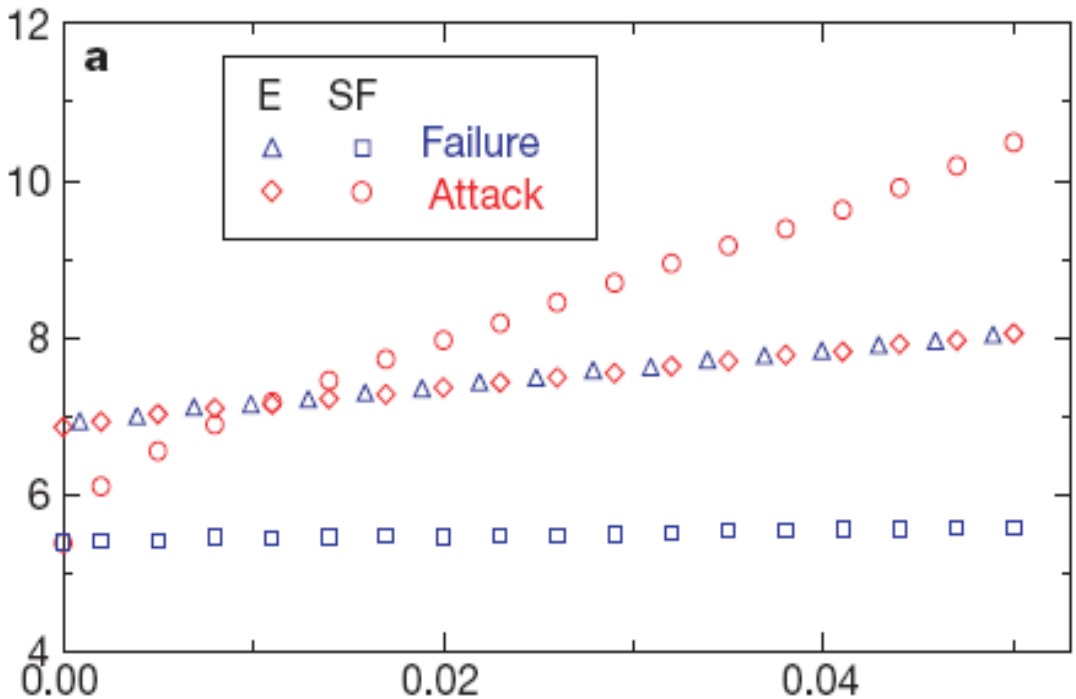
Barabási等人在《自然》上[7]发表的无标度网“稳健而又脆弱”论文是第一篇动力学文献。

他们模拟比较了ER随机图和BA无标度网络对结点删除的影响。

两种方式：一是某些结点随机失效；二是蓄意攻击中枢点集。

10000个结点
20000条连线

横坐标
删除结点比例
纵坐标
网络直径变化



稳健而又脆弱特性的原因

- **Barabási等人解释[7]**

他们认为这是ER随机图结点同质(泊松分布)，而BA无标度网结点异质(幂律分布)有hub nodes所致。

- **Bollobás等人证明[19]**

他们对类似于BA模型的LCD模型做了严格证明，BA网比ER图更稳健而又更脆弱的原因是由于择优连线。

- **Li等人提出的质疑[14]**

认为与是否存在网络核心有关。网络核心是指高度数结点相互连接组成的子网络，即所谓网络结点同配。Newman[20]则认为技术、生物网络往往结点异配。

网络核心和度–度相关性

☞Li等人[14]引入无标度程度来测量。

令$D=\{d_1,\ldots d_N\}$表示网络度序列，Li等人[7]提议用

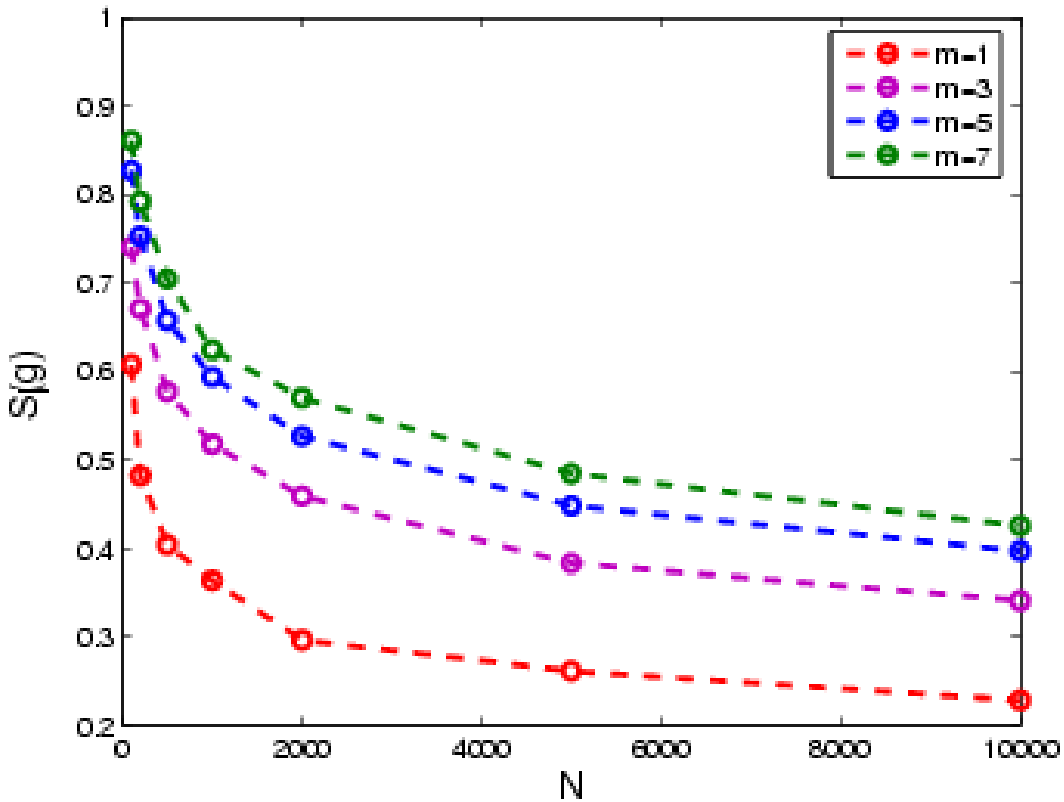
$$S(g)=s(g)/s_{\max}$$

测量无标度程度

其中 $s_{\max}$ 是

$$s(g)=\sum_{(i,j)\in E} d_i d_j$$

在度分布相同时的最大值。



BA模型

网络核心和度–度相关性

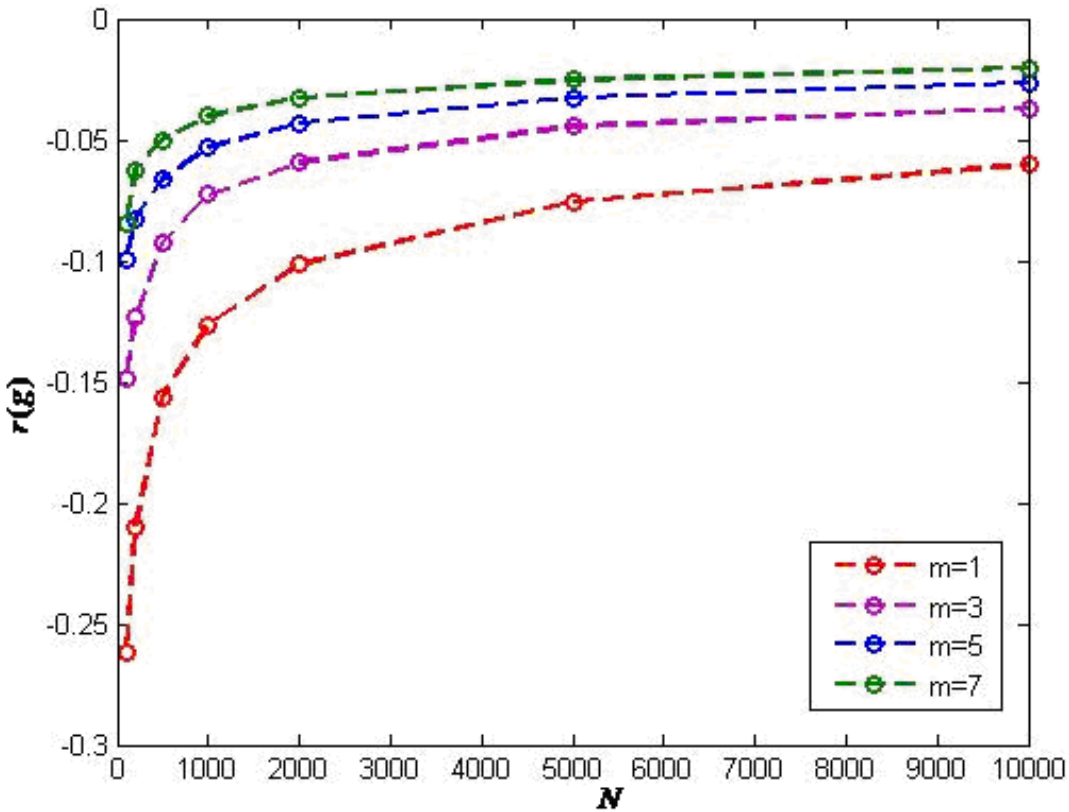
☞ **Newman[20]则引入相关系数来测量。**

$$r(g)=\frac{\sum_{(i,j)\in E} d_i d_j \Big/ |E| - \left[\sum_{(i,j)\in E} \frac{1}{2}(d_i+d_j)\Big/|E|\right]^2}{\sum_{(i,j)\in E} \frac{1}{2}(d_i^2+d_j^2)\Big/|E| - \left[\sum_{(i,j)\in E} \frac{1}{2}(d_i+d_j)\Big/|E|\right]^2}$$

$r(g)>0$ 同配

$r(g)<0$ 异配

他发现社会网络大多同配，而技术网络和生物网络倾向异配。BA网络 $r(g)\to 0$。



度–度相关性的更佳测度?

☞网络拓扑测度

Barabási说从度分布到度相关性，不同拓扑特征的广泛存在性被作为研究不同现象以及做出预测的跳板。

☞测度的合理性

网络度分布

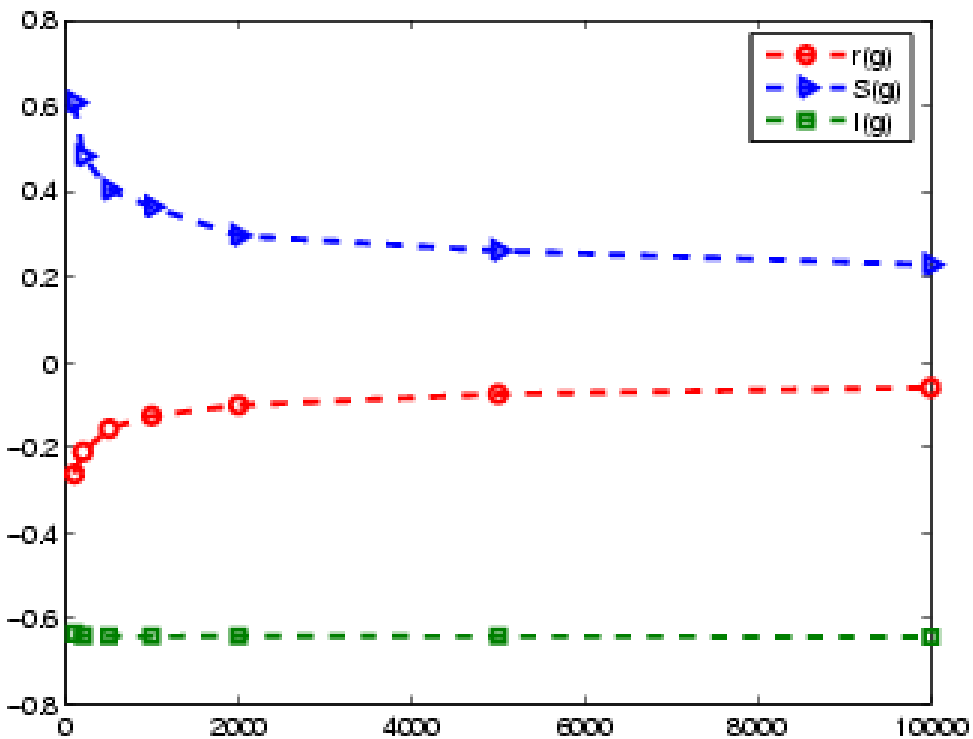
度指数独立规模

度度相关性

依赖于网络规模

联合度分布太困难

是否有更佳测度?



无标度网络的动力学特性2

- 无标度网络另一重要动力学特性是疾病的传播阈值收敛于0。最近发现在无标度网络上的动力学模型有两类不同的有限网络影响。一类只依赖网络规模；另一类同时依赖网络的规模和上割[21]，定义为 $\int_{k_c}^{\infty} P(k)dk = 1/N$ 。
- 上割 k_c 用到无限网络的度分布，且为常数。直接利用网络最大度$k_m(t)$比上割更合理。最大度是个随机变量，其发散和扰动的规律如何？对BA模型，开方发散和对数正态扰动.

网络拓扑学和动力学

☞网络拓扑学

Barabási说除非探讨其网络拓扑，否则没有办法去理解复杂系统。

网络度分布研究已有较好的基础，特别是动力学指数抓住了要害。

但是度－度相关性等其它测度还需要理清。

☞网络动力学

Barabási说共性是存在的，我们只是还没有发现能够解释他们普遍性的框架。

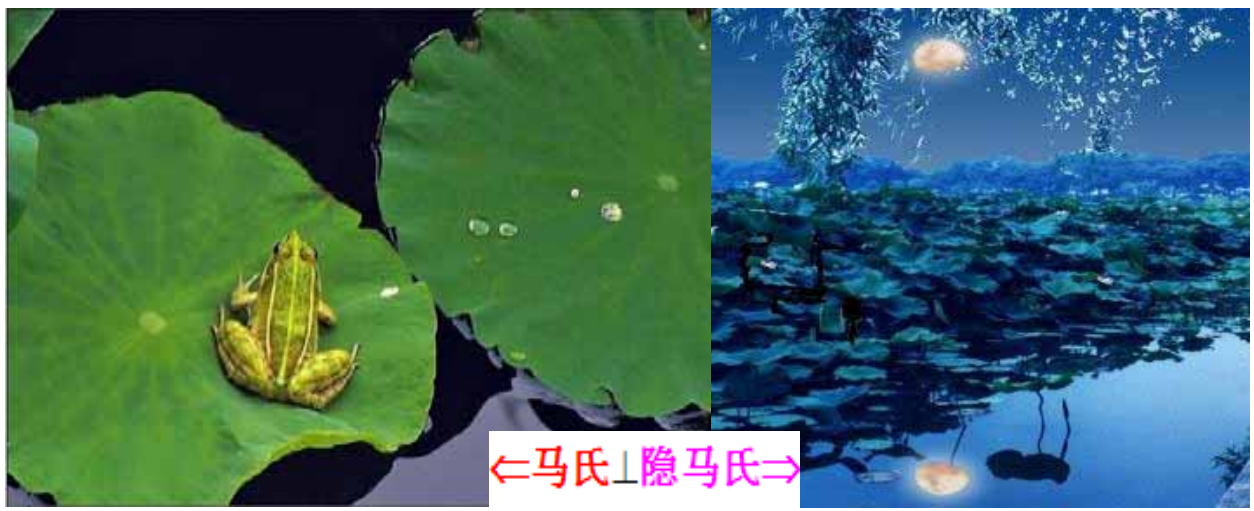
这是我们需要攻克的下一个前沿问题。

4. 复杂网络与马氏过程

☞应用研究——搜索引擎是成功范例

☞基础理论——相对论与几何学，量子力学与泛函分形，复杂网络(统计力学)与马氏过程？

上网冲浪就像青蛙在荷页上跳动



“荷塘月色”只闻蛙声，不见蛙跳

荷塘春色蛙欢跳(是否涌现的桥梁？)

网络搜索成功范例

☞可导航网络——2006年应用数学大奖

Kleinberg[3]提出可导航网络模型和分散搜索算法，极大地刺激和推动了互联网信息检索的研究。网页评级

☞网络搜索——复杂网络成功应用范例

PageRank[22]

佩奇和布林考虑网页重要性, 一是看有多少超级链接指向它(入度)；二是要看那个页面重要不重要(出度)。

BrowseRank[23]

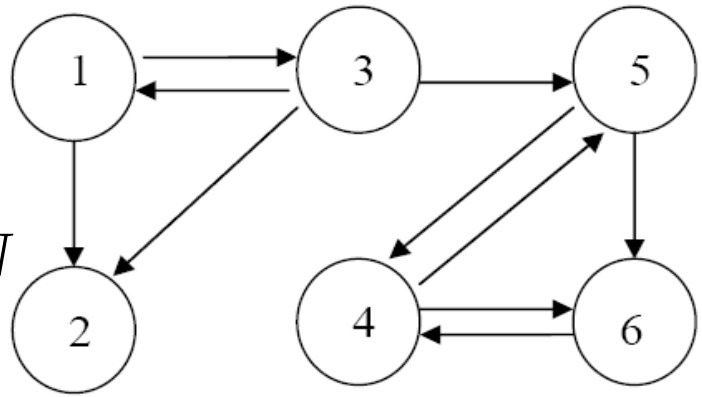
Google搜索引擎虽然取得巨大成功，但排序只依赖页面的连线(度)，因此会面临垃圾网站的严重干扰。马院士团队提出考虑用户浏览过程的网页排序。

搜索引擎的马氏过程

PageRank的马氏链

$$P=\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\bar{P}=\alpha P+(1-\alpha)ee'/N$$

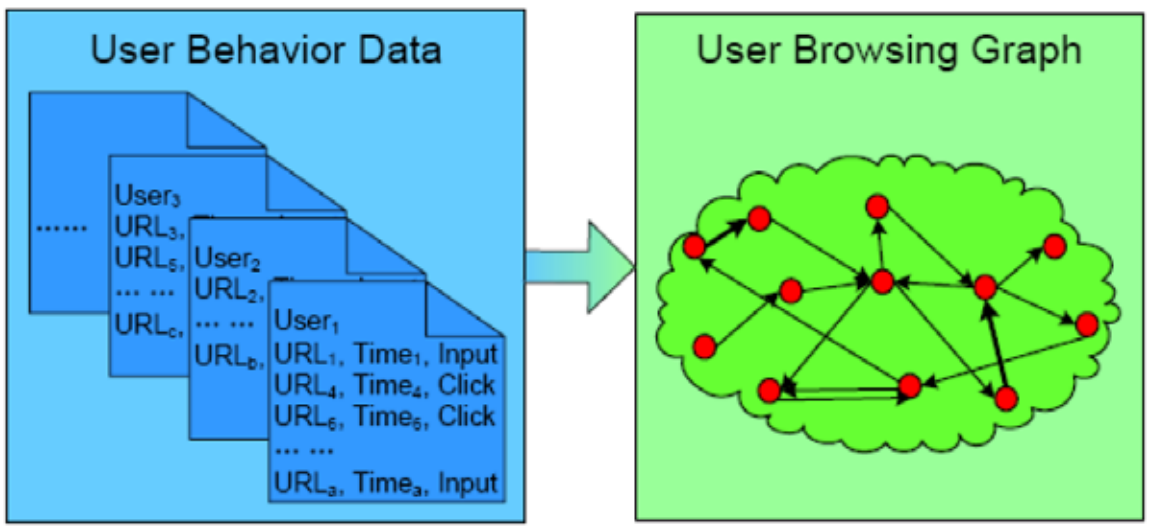


BrowseRank的Q-过程

用户浏览网页的停留时间$1/q_i$

用户浏览网页的跳转时间$1/q_{ij}$

$P=[q_{ij}/q_i]$矩阵



两个网络马氏链

☞结点度非齐次马氏链[24]

对许多网络模型，$K_i(t)$是一个马氏链。例如，网络最大度 $K_m(t)=\max\{K_i(t)\}$ 是一个鞅。

☞结点数向量马氏链[25]

对许多网络模型，$N(t)=\{N_k(t),k=1,2,\cdots\}$是一个马氏链

其中 $N_k(t)$表示在*t*时刻网络中具有度数为*k*的结点数。

☞差分方程极限定理[25]

方程 $\Delta(t+1)-\Delta(t)+\dfrac{b(t)}{c(t)}\Delta(t)=d(t)$有极限 $\lim\limits_{t\to\infty}\dfrac{\Delta(t)}{t}=\dfrac{l}{1+b}$

如果 $\lim\limits_{t\to\infty}d(t)=l$，$c(t+1)-c(t)=1$ ，$\lim\limits_{t\to\infty}b(t)=b\geq 0$。

一类增长网络的理论结果

☞网络度分布稳定性条件

初始分布：带入连线数分布稳定 $\lim_{i\to\infty}\alpha_i(h)=\alpha(h)$

转移概率：至少与t^{-1}同阶 $f(k,t)=g(k)O(t^{-r}),\ r\geq 1$

☞网络(度分布)无标度条件

无标度：存在整数J当$h\geq J$时有$\alpha(h)=0$，即有限分布

以及 $\lim_{t\to\infty}tf(k,t)=\beta k+\omega,\ 0<\beta\leq 1$；随机：$\beta=0,\omega>0$

☞度分布的迭代计算公式

收敛速率：R-线性收敛

误差上界：c未定，m最小 $\left|P(k,t)-P(k)\right|\leq\begin{cases}ct^{-\beta(m+\omega)}, & m>1\\ ct^{-\beta(m+1+\omega)}, & m\leq 1\end{cases}$

科学完整性争论与务实

“科学完整性”(Integrity of science)——

强调对目的、责任、准则、规范、价值的道德坚持。

- ☞科学结论从基本定律推导而来，受到实验室实验、自然界的观察以及数学和计算机建模的支持。
- ☞科学本身就被设计成寻找并改正错误的过程。

复杂网络在发展过程中存在欠妥并不可怕，贵在勇于修正，但许多杂志对此做得不够好。

科学不是政治宣言，不是封面女郎，不是商业广告，也不是影响因子等夺人眼球的东西，而是踏踏实实探寻和追求真理的过程，这才是亘古不变的正道。

其它参考文献

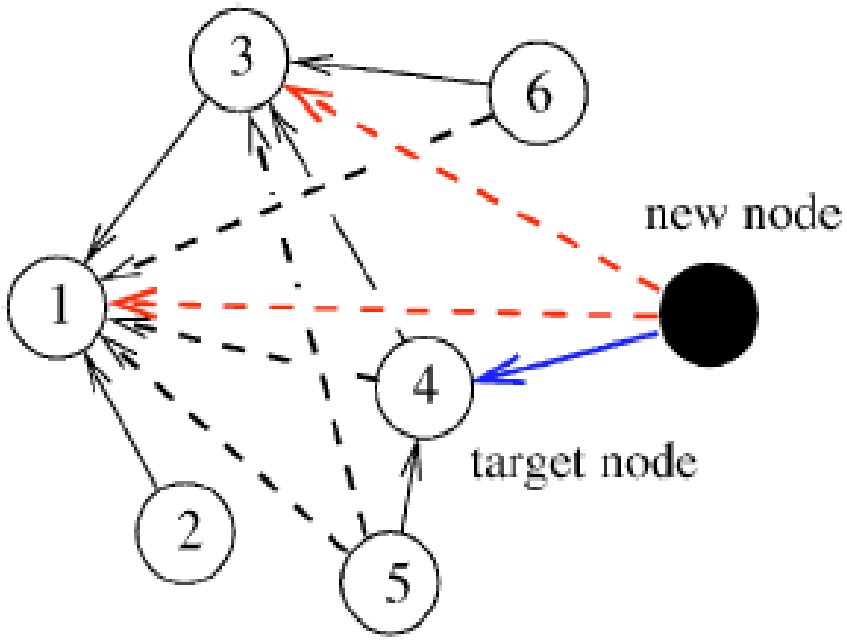
[9] B. Bollobás *et al.*, *Ran. Struc. and Algorith.* 18, 279(2001)
[10] S. N. Dorogovtsev *et al.*, *Phys. Rev. Lett.* 85, 4633(2000)
[11] P. Holme, B. J. Kim, *Phys. Rev. E* 65, 026107(2002)
[12] D. H. Shi *et al.*, *Physics Procedia* 3, 2010
[13] P. L. Krapivsky, S. Redner, *Phys. Rev. E* 71, 036118(2005)
[14] L. Li *et al.*, *Internet Math.* 2, 431(2005)
[15] S. N. Dorogovtsev *et al.*, *Phys. Rev. E* 65, 066122(2002)
[16] J. S. Andrade *et al.*, *Phys. Rev. Lett.* 94, 2005, 018702 (102, 2009, 079901)
[17] C. Doyle *et al.*, *PNAS* 102, 14497(2005)
[18] R. Tanaka, *Phys. Rev. Lett.* 94, 168101(2005)
[19] B. Bollobás, O. Riordan, *Internet Math.* 1, 1(2004)
[20] M. E. J. Newman, *Phys. Rev. Lett.* 89, 208701(2002)
[21] C. Castellano, R. Pastor-Satorras, *Phys. Rev. Lett.* 100, 148701(2008)
[22] S. Brin, L. Page, *Computer Networks and ISDN Systems* 30, 107(1998)
[23] Y. Liu *et al.*, *The 31st ACM AIGIR Conference*, 2008
[24] D. H. Shi *et al.*, *Phys. Rev. E* 71, 036140(2005)
[25] H. Xu, D. H. Shi, *Chin. Phys. Lett.* 26, 038901(2009)

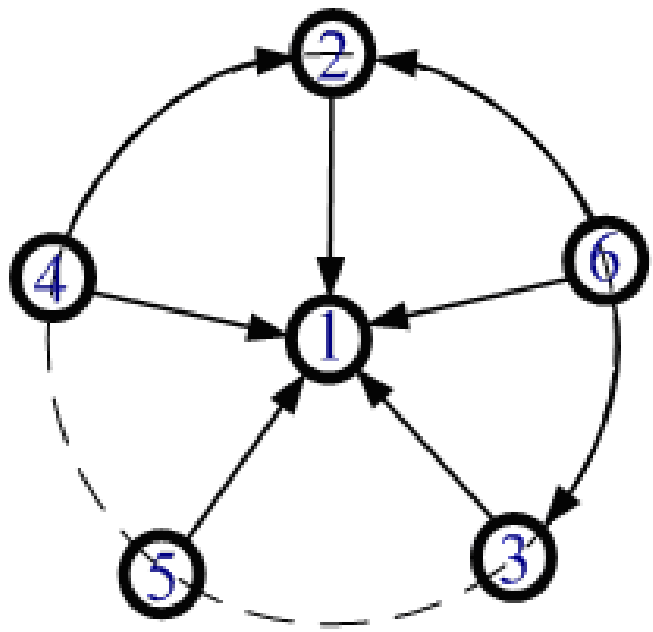
复制模型和自然数整除

左图复制模型见 **P. L. Krapivsky, S. Redner,**

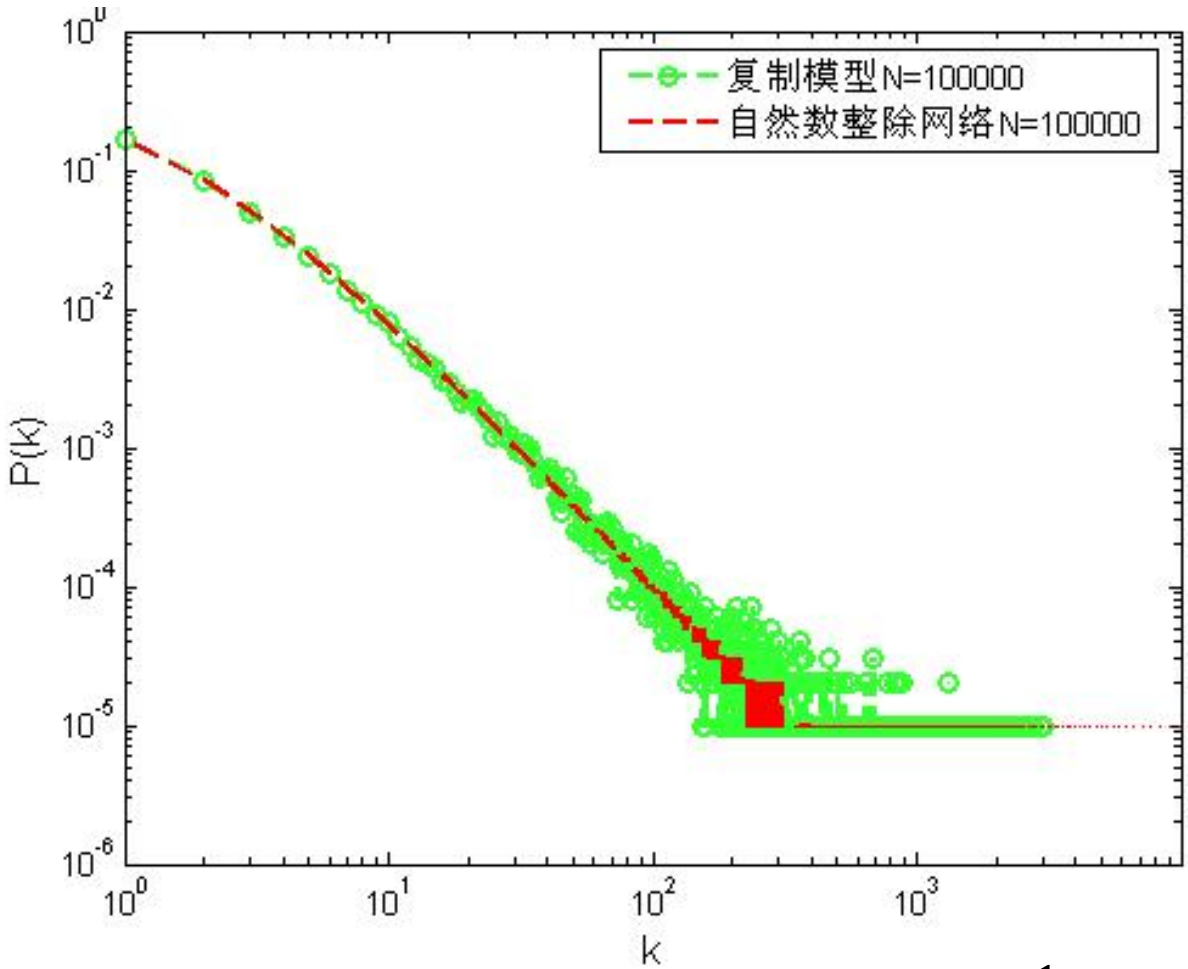
***Phys. Rev. E* 71, 036118(2005)** 随机

右图自然数整除，$N = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ 确定





两个模型度分布比较



可证明度分布完全一致 $P_{in}(k)=\dfrac{1}{(k+1)(k+2)}\approx k^{-2}$

复杂网络与素数定理猜想

由算术基本定理，可以认为素数是组成自然数系统的基本单元。对有限自然数系统，基本单元多少？

素数定理猜想

在N个自然数中素数的个数，记为 $\pi(N)$

素数定理猜想 $\pi(N) \sim N/\ln N$

Legendre1798年的观察结果是 $\pi(N) \approx N/(\ln N - 1.08366)$

Gauss1793年猜测素数分布密度是 $\rho(x) = \ln^{-1}(x)$

Euler1737年得到极为重要的乘积公式，但错过猜想

根据网络平均度，我们可以得到猜想 $\rho_{net}(x) = \ln^{-1}(x+c)$

黎曼猜想

证明素数定理需要知道 $\pi(N)$精确公式。黎曼给出了，但与黎曼猜想零点有关。**Erdös1949年**给出初等证明。