网络基本的拓扑性质

王守道

2020 年 10 月 29 日

Outline

引言

上章所研究的图往往都是节点较少或者具有某种规则的图，可以通过图示的方法得出某些性质。但是现在实际涉及的网络结构往往涉及上百万个节点。对于这种情况，不能简单地通过图示的方法来研究网络的性质。
同时对于大规模的网络结构，所关心的问题与少数节点的网络变得不同, 而关心地是一些宏观层面的东西。
这一章将从宏观角度研究网络的三个拓扑性质：

1. 平均路径长度
2. 聚类系数
3. 度分布

引言

从研究来看，研究个体数量较小的系统和较大的系统往往采取不同的方法。如 Anderson 在 1972 年的文章 *More is different*:

The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other.

无向网络中的巨片

经验表明很多大规模网络都是不连通的，但会存在特别大的连通子集，这个子集占网络中节点的大部分。这一连通子集称为**巨片** (Giant component)。

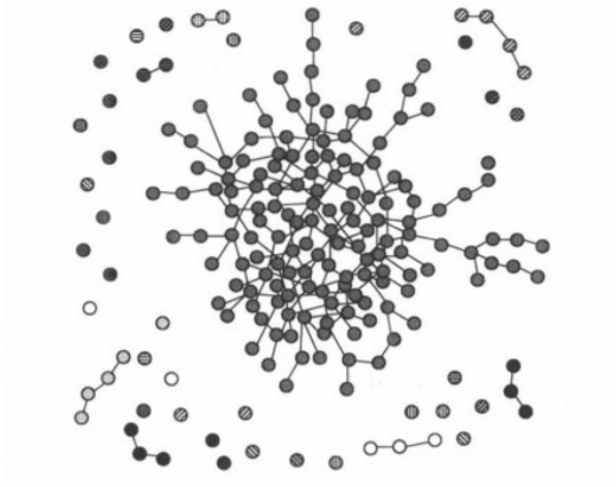

图: 具有单个连通巨片的网络示意

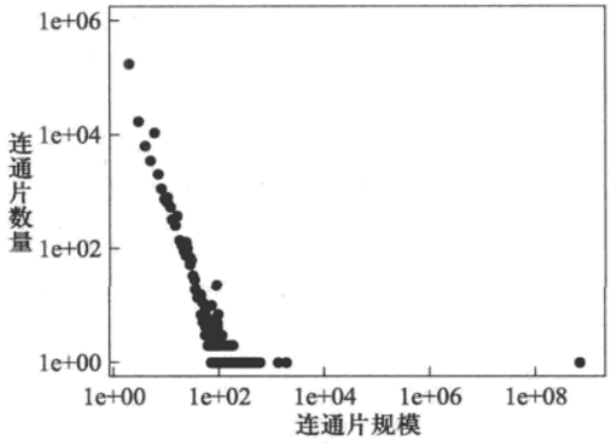

图: Facebook 用户关系网络的连通片规模的分布

有向网络中的蝴蝶结结构

大规模网络往往不是强连通也不是弱连通的，很多网络有一个弱连通巨片，这一巨片具有蝴蝶结构。包含强连通核、入部、出部和卷须。

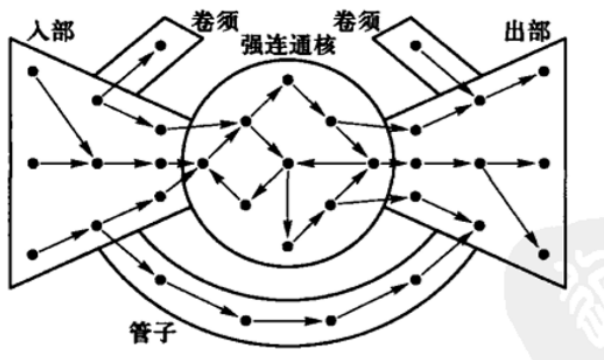

图: 有向网络弱连通巨片的蝴蝶结构

度与平均度

无向网络中节点 i 的度 k_i 定义为与节点直接相连的边的数目。
网络中所有节点度的平均值称为平均度，记为 <k>。
给定网络 G 的邻接矩阵 $\boldsymbol{A} = (a_{ij})_{N\times N}$，有：

$$k_i = \sum_{j=1} a_{ij} = \sum_{j=1} a_{ji}$$

$$\langle k\rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{1}{N}\sum_{i,j=1}^{N} a_{ij}.$$

网络中节点度与网络边数 M 之间有如下关系：

$$2M = N\langle k\rangle = \sum_{i=1}^{N} k_i = \sum_{i,j=1}^{N} a_{ij}$$

$$M = \frac{1}{2}N\langle k\rangle = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{i,j=1}^{N} a_{ij},$$

$$\langle k\rangle = \frac{2M}{N}$$

出度与入度

出度和入度是针对有向网络定义的，节点的出度指的是从节点指向其他节点的边的数目，节点的入度是其他节点指向该节点边的数目。用邻接矩阵表示如下：

$$k_i^{out} = \sum_{j=1}^{N} a_{ij}, \quad k_i^{in} = \sum_{j=1}^{N} a_{jt}$$

在有向网络中，尽管单个节点的出度和入度不相同，但是所有节点平均出度入度相同：

$$\left\langle k^{out} \right\rangle = \left\langle k^{in} \right\rangle = \frac{1}{N} \sum_{i,j=1}^{N} a_{ij} = \frac{M}{N}$$

对于加权网络的度的定义，只需将上述定义中的边换位对应边的权重即可。

网络稀疏性和稠密化

网络的**密度** (density) 定义为网络中实际存在的边数和最大可能的边数之比。对于无向网络，有：

$$\rho = \frac{M}{\frac{1}{2}N(N-1)}$$

有向网络将分母中的 $\frac{1}{2}$ 去掉即可。
对于一个网络，当 $N \to \infty$ 时，网络密度趋于非零常数，说明该网络是稠密的。当网络密度趋于 0 时，说明该网络是稀疏的。
对比平均度和网络密度的定义，有以下关系：

$$\langle k \rangle = \frac{2M}{N} = (N-1)\rho \approx N\rho$$

因此可将平均度作为判别网络稀疏或稠密的标准。

网络稀疏性和稠密化

将 t 时刻网络中节点数和边数分别记为 N(t) 和 M(t)。

1. 线性关系：$M(t) \sim N(t)$，平均度为一个常数。
2. 平方关系：$M(t) \sim N^2(t)$，每个节点都会与一定比例的其他节点相连，网络稠密。
3. 一般情况：$M(t) \sim N^{\alpha}(t), \quad 1 < \alpha < 2$，稠密化幂律。网络会变得越来越稠密，但是相比平方关仍然是稀疏的。

对 3 的关系取对数，可以得到：

$$\ln M(t) \approx \alpha(\ln N(t)) + C, \quad 1 < \alpha < 2$$

我们可以通过线性拟合来判断其关系是否有 power law。

网络稀疏性和稠密化

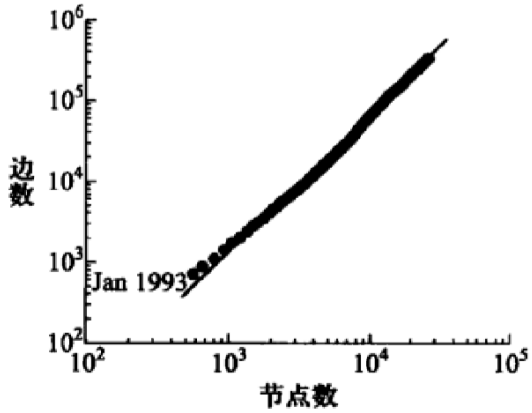

图: 科研合作网络

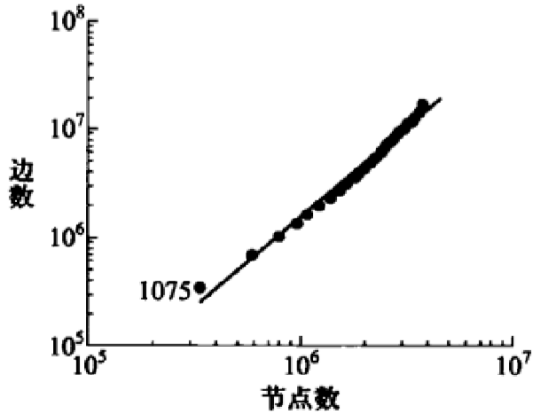

图: 专利网络

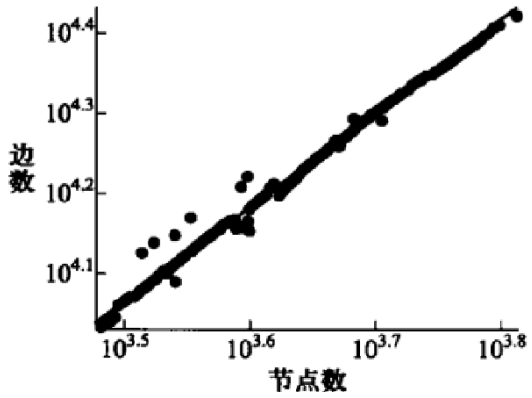

图: 自制层 Internet

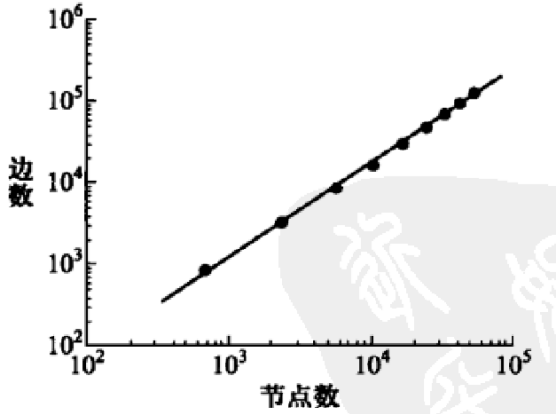

图: 从属网络

无权无向网络情况

- 最短路径：网络中连接两点边数最少的路径
- 距离：i 和 j 经过最短路径的距离，d_{ij}

网络的平均距离 L 指的是任意两个节点距离的平均值，即：

$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i>j} d_{ij}$$

对于无权网络，平均距离可以通过 BFS 在 $O(MN)$ 时间复杂度得到。
在多数情况下，整个网络都不是连通的，按照上面的方法计算会出现 $d_{ij} = \infty$ 的情况，这时可用简谐平均来代替：

$$GE = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i=j} \frac{1}{d_{ij}}$$

无权无向网络情况

网络直径指的是存在有限距离的节点距离最大值：$D = max_{i,j} d_{ij}$

更多情况下，我们关注的是网络中，用户距离的分布，大多数用户之间的距离, 记：

- f(d): 网络中距离为 d 的节点对数量占网络中连通节点对的比例（概率密度）
- g(d): 网络中距离不超过 d 的节点对的数量占网络中连通节点对的比例（概率分布）

如果整数 D 满足，

$$g(D-1) < 0.9, \quad g(D) \geqslant 0.9$$

称 D 为网络的有效直径，这种定义有效直径可以通过插值的方法推广到实数范围。

加权有向网络情况

上述对于无权无向网络的定义可以推广到有权有向网络中，只需要考虑边的权重和方向即可。但此时对于距离的求解不能通过 BFS 求解，需要用到 Dijkstra 算法。

1. 维护两个节点集 S 和 Q。初始时，S 只包含起点 s；Q 包含除 s 外的其他顶点，且 Q 中顶点的距离为起点 s 到相邻点的距离 [Q 中顶点 v 的距离为 (s,v) 的长度，若 s 和 v 不相邻，则 v 的距离为 ∞]
2. 从 Q 中选出与 S 中节点相邻，且距离 s 最近的节点 k，将 k 放入 S 中，从 Q 中移除 k
3. 更新 s 到 U 中各个节点的距离
4. 重复 2、3 步直到覆盖所有节点

基于堆的数据结构的 Dijkstra 算法的时间复杂度为 $O(m \log n)$。

加权有向网络

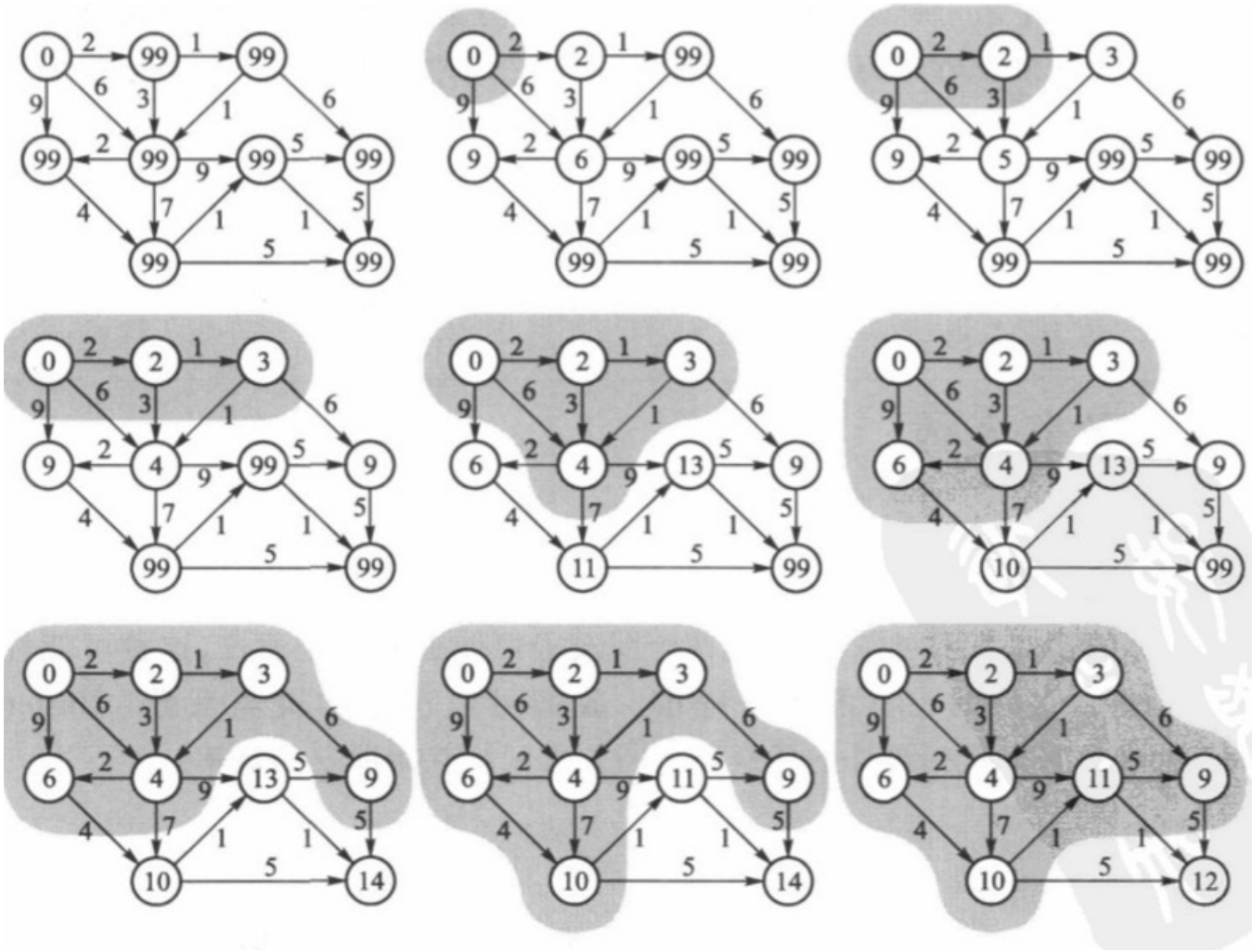

图: Dijkstra 算法示例

无权无向网络情况

聚类系数用来刻画与当前节点相连的两个节点也相连的概率，假设节点 i 的度为 k_i, 其聚类系数 C_i 定义为：

$$C_i = \frac{E_i}{(k_i(k_i-1))/2} = \frac{2E_i}{k_i(k_i-1)}$$

其中 E_i 表示节点 i 的 k_i 个邻接点之间实际存在的边数。
从几何角度来看，聚类系数可以定义如下：

$$C_i = \frac{\text{包含节点}i\text{的三角形的数目}}{\text{以节点}i\text{为中心的连通三元组的数目}}$$

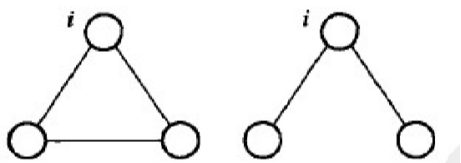

图: 以节点 i 为中心的连通双元组的两种可能

无权无向网络情况

从邻接矩阵角度，给定网络的邻接矩阵 $\mathbf{A} = (a_{ij})_{N\times N}$, 此时包含节点 i 的三角形数目为：

$$E_i = \frac{1}{2}\sum_{j,k} a_{ij}a_{jk}a_{ki} = \sum_{k>j} a_{ij}a_{jk}a_{ki}$$

同样的，以 i 为中心的连通三元组个数可用邻接矩阵表示为 $\sum_{j\neq i, i\neq j, k\neq i} a_{ij}a_{ik}$.
因此，节点 i 的聚类系数可用邻接矩阵表示如下：

$$C_i = \frac{2E_i}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)}\sum_{j,k=1}^{N} a_{ij}a_{jk}a_{ki} = \frac{\sum_{j\neq i,k\neq j,k\neq i} a_{ij}a_{ik}a_{jk}}{\sum_{j\neq i,t\neq j,k\neq i} a_{ij}a_{ik}}$$

无权无向网络情况

对于大规模网络，我们喜欢考虑其整体行为或平均行为，在求得单个聚类系数的基础上，可以得到网络聚类系数的平均值：
$C = \frac{1}{N}\sum_{i=1}^{N} C_i$.
同样的，我们也可以研究具有某一性质的节点的聚类性质的分布，可以得到度为 k 的节点的聚类系数的平均值：

$$C(k) = \frac{\sum_i C_i \delta_{k_i,k}}{\sum_i \delta_{k_i,k}}$$

许多实际的网络表明，$C(k)$ 具有幂律形式 $C(k) \sim k^{-\alpha}(\alpha > 0)$

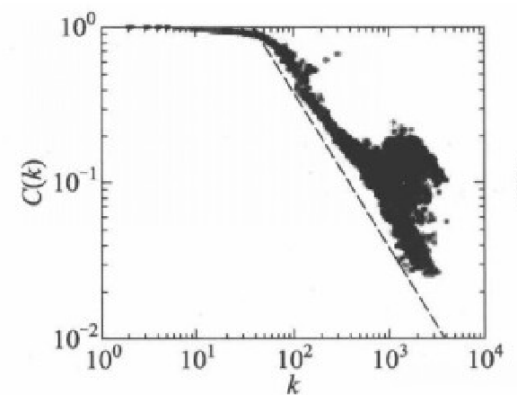

图: 电影演员网

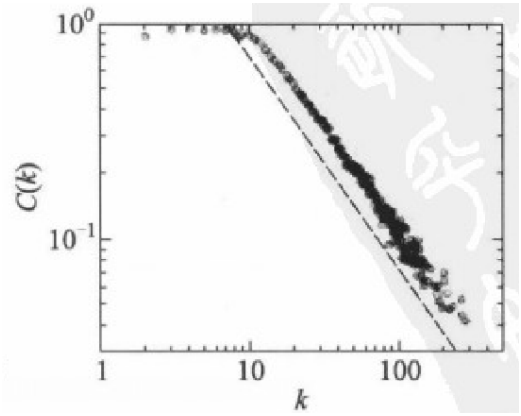

图: 语义网

加权网络情况

对于加权网络，其聚类系数不能简单的从无权网络移植。给定一个加权网络 G 的邻接矩阵 $\mathbf{A} = (a_{ij})$ 和非负权值矩阵 $\mathbf{W} = (w_{ij})$，我们想得到如下形式的节点聚类系数：

$$\tilde{C}_i = \frac{1}{k_i (k_i - 1)} \sum_{j,k} \omega_{ijk} a_{ij} a_{ik} a_{jk}$$

关键是如何通过 $\mathbf{W}$ 确定 w_{ijk}。w_{ijk} 一般有以下限制条件：

1. 当 i，j，k 不构成三角形时，w_{ijk} 可以任意取值
2. 在无权情况下，当 i、j 和 k 构成三角形时有 $w_{ijk} = 1$
3. 为保证 $\tilde{C}_i \in [0, 1]$，有 $w_{ijk} \in [0, 1]$

加权网络情况

取法 1：将 w_{ijk} 取为节点 i 与它的两个邻节点 j 和 k 之间的两条边的权值的归一化平均值，即

$$\omega_{ijk} = \frac{1}{\langle w_i \rangle} \frac{w_{ij} + w_{ik}}{2}$$

其中 $< w_i >$ 是以节点 i 为一个端点的所有边权值的平均值，即

$$\langle w_i \rangle = \frac{1}{k_i} \sum_j w_i$$

代入可得此种取法的聚类系数定义如下：

$$\tilde{C}_i^{(1)} = \frac{1}{k_i (k_i - 1)} \sum_{j,k} \frac{w_{ij} + w_{ik}}{2 \langle w_i \rangle} a_{ij} a_{ik} a_{jk}$$

加权网络情况

取法 2：将 w_{ijk} 取为节点 i 与它的两个邻节点 j 和 k 组成三角形的三条边归一化的几何平均:

$$\omega_{ijk} = \left(\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}\right)^{1/3}$$

其中 $\hat{w}_{ij} \in [0,1]$ 为归一化权值，$\hat{w}_{ij} = \frac{w_{ij}}{\max_{k,l} w_{kl}}$ 代入可得聚类系数又一种定义方式：

$$\tilde{C}_i^{(2)} = \frac{1}{k_i\left(k_i-1\right)} \sum_{j,k} \left(\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}\right)^{1/3} a_{ij}a_{ik}a_{jk}$$

加权网络情况

取法 3：在无权网络中，聚类系数定义为包含节点 i 的三角形数目除以可能包含 i 的三角形数目的上界，仿照这个定义带权重的网络的聚类系数可定义如下：

$$\tilde{C}_i = \frac{\frac{1}{2}\sum_{j,k}\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}}{\frac{1}{2}\left(\left(\sum_k \hat{w}_{ik}\right)^2 - \sum_k \hat{w}_{ik}^2\right)} = \frac{\sum_{j,k}\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}}{\left(\sum_k \hat{w}_{ik}\right)^2 - \sum_k \hat{w}_{ik}^2} = \frac{\sum_{j,k}\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}}{\sum_{j\neq k}\hat{w}_{ij}\hat{w}_{ik}}$$

上式的分子为包含节点 i 的三角形数目 E_i 的加权形式，对应的分母为分子可能的上界。

度分布的概念

一个网络的度分布 p_k 定义为，对于一个给定的度 k，网络中度为 k 节点的比例占总结点数的比例。
对于有向网络，只需将上述度分别换位出度和入度即可。

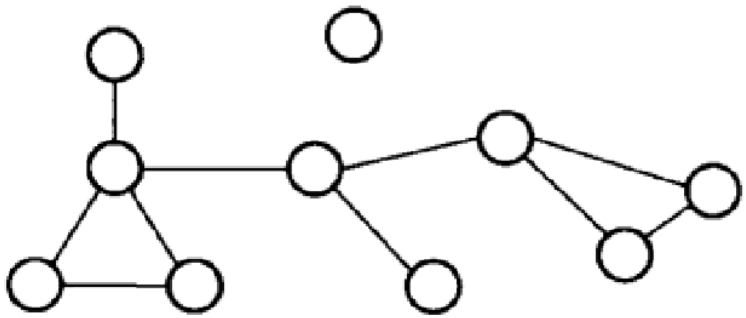

图: 一个无向网络的例子

$$p_0 = \frac{1}{10}, p_1 = \frac{2}{10}, p_2 = \frac{4}{10}, p_3 = \frac{2}{10}, p_4 = \frac{1}{10}, p_k = 0, k > 4$$

重尾分布

重尾分布是尾部比指数分布还要厚的分布，重尾分布又包含长尾分布和肥尾分布：

- 重尾分布：如果一个随机变量的累积分布函数满足，$\lim_{x\to\infty} e^{\lambda x}\Pr[X>x]=\infty \quad \text{for all } \lambda>0$ 则称其为重尾分布。
- 长尾分布：如果对于一个随机变量，对于所有 t，满足：$\lim_{x\to\infty}\Pr[X>x+t \mid X>x]=1$ 则称其为长尾分布。
- 肥尾分布：如果对于一个随机变量，满足 $\lim_{x\to\infty}\Pr[X>x]\sim x^{-\alpha}, \alpha>0$ 则称其为肥尾分布。

幂律分布及其检验

Barabasi 发表在 *Nature* 和 *Science* 文章表明，现实中的网络分布大多不服从正态分布或泊松分布，而是服从有以下规律的幂律分布：

$$P(k) \sim k^{-\gamma}$$

对于大规模网络来说，服从正态分布或幂律分布会有很大区别，指数下降要比幂函数下降快很多。
一个 WWW 例子:

- 幂律分布：$P\left(k^{in}=1000\right) \sim 1000^{-2}=10^{-6}$,N=$10^9$
- 正态分布：$P\left(k^{in}=1000\right) \sim e^{-1000} \approx 2^{-1000}$,N=$10^9$

幂律分布及其检验

对于网络中度分布 $P(k)$ 可以通过双对数来验证其是否为幂律分布。假设要验证是否有常数 C 和幂指数 γ 使得满足：

$$P(k) = Ck^{-\gamma}$$

对上式两端取对数，可得

$$\ln P(k) = \ln C - \gamma \ln k$$

说明 $\ln P(k)$ 与 $\ln k$ 有线性关系，可通过最小二乘的方法判断。可能存在两个问题:

1. 并非所有度值都服从幂律分布，只是当度值较大时存在幂律分布
2. 在度值很高时，由于度值是有限的，可能会出现截断现象，不服从幂律分布

幂律分布及其检验

除了上述问题，判断幂律分布还会受到服从幂律分布区间内噪声的干扰。一种常见的光滑化处理方法是转换为累积度分布。累积度 P_k 的定义如下:

$$P_k = \sum_{k'=k} P\left(k'\right)$$

如果一个网络的度分布为幂律分布，那么累积度分布近似也服从幂律分布:

$$P_k = C\sum_{k=1}^{\infty} k'^{-\gamma} \simeq C\int_k^{\infty} k'^{-\gamma}dk' = \frac{C}{\gamma-1}k^{-(\gamma-1)}$$

这样处理的缺点:

1. 间接反映节点度的分布，没有度分布那么容易解释
2. 累积度分布相邻点之间的关系并不是独立的，不符合最小二乘拟合的假设
3. 即使符合最小二乘，这样估计 γ 也是有偏的

幂律分布及其检验

为了解决上述问题，常用的方法是极大似然直接估计幂指数:

$$\gamma = 1 + \tilde{N}\left[\sum_i \ln \frac{k_i}{k_{\min} - 0.5}\right]^{-1}$$

这里 $k_{\min}$ 是使幂律成立的最小度值，$\tilde{N}$ 是度不小于 $k_{\min}$ 的节点数。上述估计的统计误差为：

$$\sigma = \sqrt{\tilde{N}}\left[\sum_i \ln \frac{k_i}{k_{\min} - 0.5}\right]^{-1} = \frac{\gamma - 1}{\sqrt{\tilde{N}}}$$

幂律分布的性质

无标度性质：
考虑到一个概率分布函数 $f(x)$，假设 $f(1)f'(1) \neq 0$。如果对任意给定常数 a，存在常数 b 使得函数 $f(x)$ 满足以下" 无标度条件":

$$f(ax) = bf(x)$$

那么必有:

$$f(x) = f(1)x^{-\gamma}, \quad \gamma = -f'(1)/f(1)$$

即幂律分布是唯一满足无标度条件的概率分布函数。
具有无标度性质的分布函数具有长尾性质。

无标度网络的性质和应用

无标度网络同时具有鲁棒性和脆弱性，下图是对无标度网络和随机网络进行不同类型的攻击后的结果:

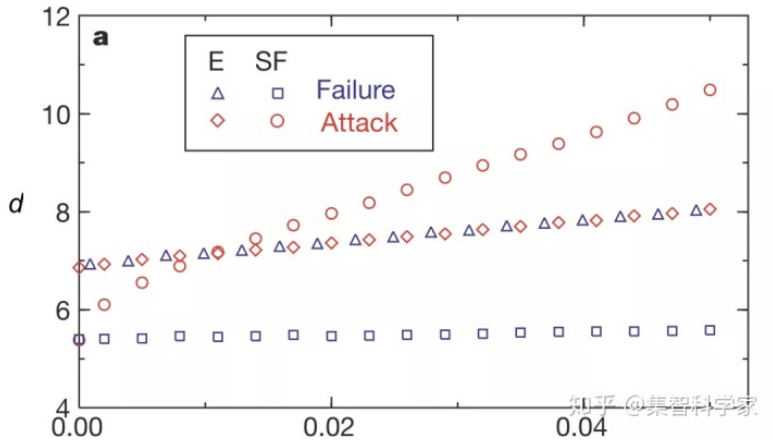

图: 移除的节点比例与特征路径长度示意图

启发：

- 社交网络中移除节点度高的”谣言传播者”可终结谣言
- 金融网络中去除度高的节点的金融风险可阻止风险的传播

无标度网络历史和生成原因

无标度网络受到重视是由 Barabási 于 1999 年在 *Science* 上一篇文章提出，和另一篇 Duncan Watts 发表在 *Nature* 的小世界网络共同宣告网络科学领域的形成。
事实上 Barabási 对幂律分布的发现是重新发现。事实上有许多科学家曾发现了此规律。

- 维弗雷多·帕累托（Vilfredo Pareto ，1848—1923），福利经济学先驱；
- 乔治·金斯利·齐普夫（George Kingsley Zipf，1902-1950），计量语言学先驱；
- 德里克·普莱斯（Derek John de Solla Price，1922-1983），科学计量学之父；
- 赫伯特·亚历山大·西蒙（Herbert Alexander Simon，1916-2001），人工智能先驱；
- 波努瓦·曼德布罗特（Benoit B. Mandelbrot，1924-2010），分形之父；
- 艾伯特-拉斯洛·巴拉巴西（Albert-László Barabási，1967- ），当今网络科学研究代表人物、无标度网络概念提出者。

无标度网络历史和生成原因

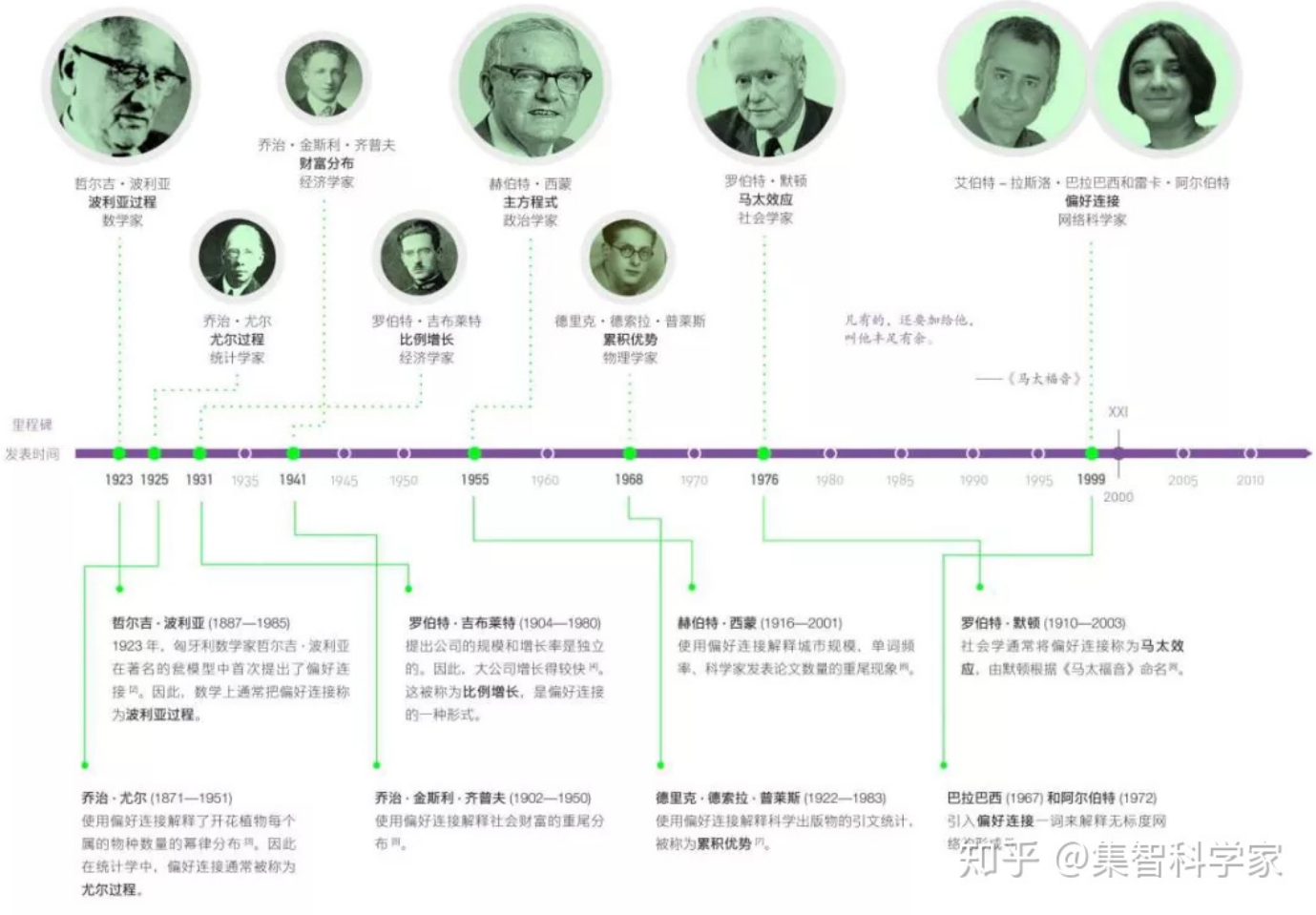

图: 偏好链接研究史

无标度网络历史和生成原因

产生无标度网络原因主要有两个理论方向：随机框架和优化框架

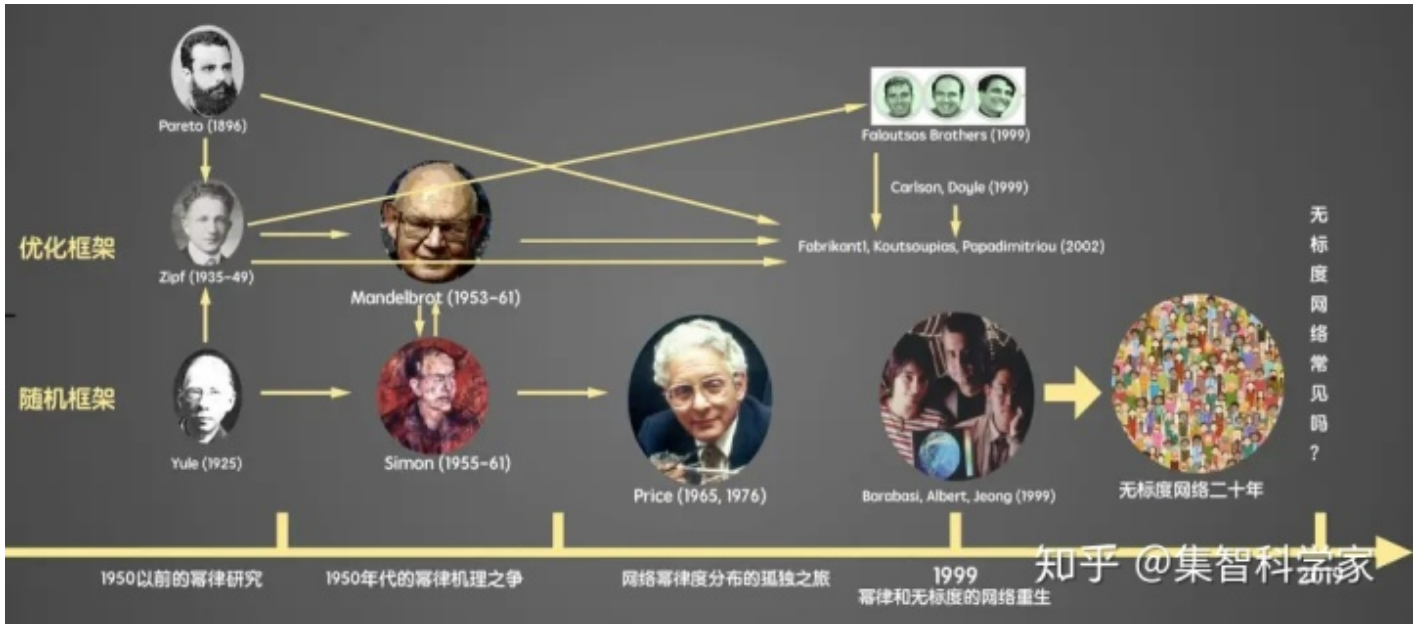

图: 幂律分布和无标度网络研究中一些历史性节点之间的引用关系网络

无标度网络历史和生成原因

随机框架：以普莱斯的文献网络为例

1. **增长机制**：文章的数量是不断增长的；新发表的文章会引用早前发表的一些文章作为参考文献。
2. **累计优势机制**：早先发表的一篇文章被一篇新发表的文章引用的概率与它已经被引用的次数成正比。

优化框架：以曼德布罗对语言信息研究为例

经典的香农信息论问题研究的是对消息构造最小代价编码，而曼德布罗特指出语言的统计结构问题事实上是这一经典问题的逆问题，即用尽可能少的成本传递尽可能多的信息。这种目标优化会导致第 j 个使用最多的单词出现的频率服从幂律分布。

无标度网络受到的挑战

安然·克洛赛（Aaron Clauset）等人研究发现无标度网络（幂律分布）在现实网络中，并不常见。但是大多数学者认为，研究度分布是否是长尾的，比其是否为幂律的重要的多。甚至有一些学者对于度分布是否在网络科学中为核心概念产生质疑。

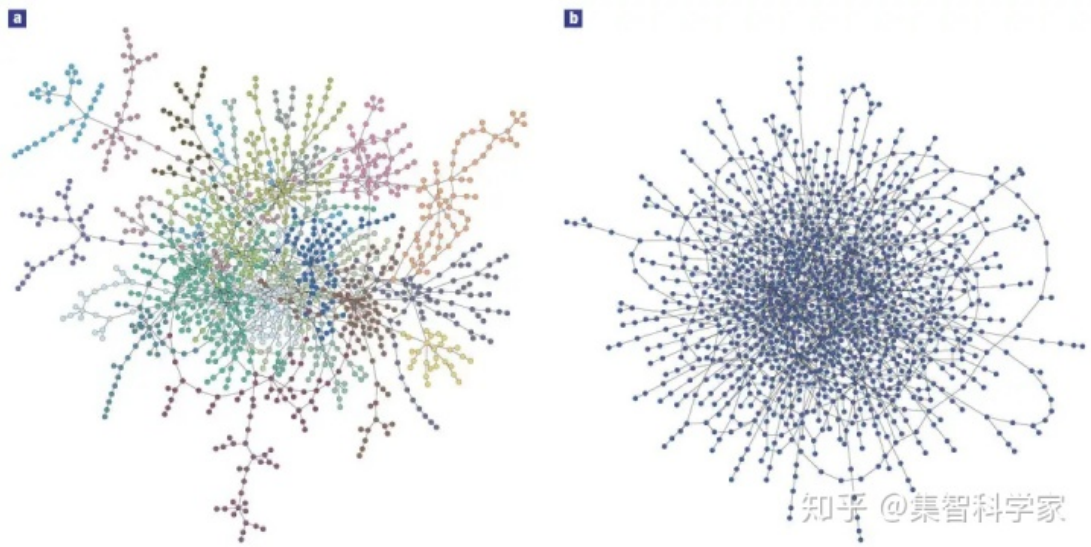

图: 两个具有完全相同度分布的网络

无标度分布的归一化

可以通过概率分布的性质来求得幂律分布前面的系数 C:

$$\sum_{k=k_{\min}}^{\infty} Ck^{-\gamma} = 1 \Longrightarrow C = \frac{1}{\zeta(\gamma, k_{\min})}, \zeta(\gamma, k_{\min}) \equiv \sum_{k=k_{\min}}^{\infty} k^{-\gamma}$$

由此可以通过近似积分的形式求得归一化常数：

$$C = \frac{1}{\sum_{k=k_{\min}}^{\infty} k^{-\gamma}} \sim \frac{1}{\int_{k_{\min}}^{\infty} k^{-\gamma} \mathrm{d}k} = (\gamma - 1)(k_{\min})^{\gamma - 1}$$

从而幂律度分布 p_k 和累计度分布 P_k 可以写成:

$$p_k \sim \frac{\gamma - 1}{k_{\min}} \left(\frac{k}{k_{\min}}\right)^{-\gamma}, P_k \sim \left(\frac{k}{k_{\min}}\right)^{-(\gamma - 1)}$$

幂律分布矩的性质

考虑到度分布在 $k \geq k_{\min}$ 时服从幂律分布，度分布的 m 阶矩为：

$$\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k = \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^{m-\gamma}$$

利用近似积分，有：

$$\begin{aligned} \langle k^{\min} \rangle &\simeq \sum_{k=0}^{k_{\min}-1} k^m p_k + C \int_{k_{\min}}^{\infty} k^{m-\gamma} \, \mathrm{d}k \\ &= \sum_{k=0}^{k_{\min}-1} k^m p_k + \frac{C}{m-\gamma+1} \left[k^{m-\gamma+1} \right]_{k_{\min}}^{\infty} \end{aligned}$$

从上可以看出幂律分布存在有限 m 阶矩的条件是 $\gamma > m+1$。大多数网络的幂律分布的 $2 < \gamma \leq 3$。

Thank you!