

有向网络的复制机制及其相关网络分析

毛小燕, 周晖杰

(宁波大学 科学技术学院, 浙江 宁波 315212)

摘要: 鉴于有向网络比无向网络包含了更多的内在信息, 而复杂网络中的基本统计量往往仅适用于无向网络, 使得有向网络的研究相对缺少, 由此提出了一个有向网络的统计量, 并分析该统计量在相关有向网络研究中的有效性. 考虑到复制是有向网络增长的一个主要动力, 定义了有向网络结点复制率和有向网络复制率的概念, 并利用结点入度分布和复制率研究了有向规则网络、复制模型网络及自然数网络. 结果显示, 完全复制模型和自然数网络的入度具有无标度特性, 其入度分布的幂律指数 γ 都为 2, 2 个有向网络的复制率 $c=1$, 而部分复制模型的复制率 $c=p$. 因此, 有向网络的入度分布、复制率都能很好地解释完全复制模型与自然数网络的相关性, 可作为重要统计量广泛应用于有向网络研究中.

关键词: 有向网络; 复制模型; 自然数网络; 入度分布; 复制率

中图分类号: TP393; N94

文献标志码: A

文章编号: 1001-5132 (2015) 02-0028-04

Barabási 和 Albert 从动态的、增长的角度研究了复杂网络幂律度分布的形成机理, 提出了 BA 模型^[1], 此后极大地激发了科学界的研究热情. 无向网络被广泛研究, 引入度分布、集聚系数、最短距离、介数、度相关性、连通集团的规模分布、结点重要性^[2]、网络同配性^[3-4]、结点同配性^[5]等统计量. 然而, 现实网络往往具有有向特性, 若将有向网络处理成无向网络, 并运用无向网络的统计量来研究, 这会导致很多内在信息的缺失. 另外, 现实有向网络的形成往往依赖于复制增长, 即新增结点的连接更倾向于复制旧结点的连接机制, 比如基因调控网络的关键在于复制增长, 也是生物进化的主要动力. 鉴于复制增长在有向网络中的重要性, 笔者提出了一个适合有向网络的统计量——复制率, 包括结点复制率和网络复制率; 其次研究了通过复制增长的有向网络模型的入度分布和网络复制率, 分析复制增长模型的无标度特性; 最后从有向网络入度分布和复制率角度分析了复制模型与自然数网络模型的内在联系, 并运用复杂网络方法研究自然数网络及素数提出了一些想法.

1 有向网络复制率定义

引文网中结点代表论文, 从论文 B 到论文 A 的有向边代表 B 引用 A . 人们在引用 1 篇论文时, 往往把它引用过的部分文献也一起复制引用, 因此复制增长是引文网形成的主要动力^[6]. 另外, 基因调控网络、代谢网络、蛋白质相互作用网络等都具有复制增长特征. 基于复制是一些重要有向网络增长的主要动力, 提出了有向网络复制率的概念.

定义 1 设 e_{ij} 表示有向网络 G 中结点 v_i 指向结点 v_j 的边, 若结点 v_j 的出度为 k_j , 结点 v_i 和结点 v_j 指向同个结点的个数为 k_{ij} , 则结点 v_i 对结点 v_j 的复制率定义为:

$$c(e_{ij}) = k_{ij} / k_j, \quad (1)$$

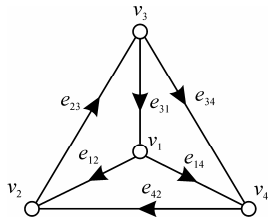
简记为 c_{ij} , 规定: 若 $k_j = 0$, 则 $c_{ij} = 1$. 显然, 网络结点的复制率 $c_{ij} \in [0, 1]$.

定义 2 若 $N(e)$ 表示有向网络 G 中有向边的总数, 那么有向网络 G 的复制率定义为:

$$c = \left(\sum_{e_{ij} \in N(e)} c_{ij} \right) / N(e), \quad (2)$$

显然, 有向网络的复制率 $c \in [0, 1]$.

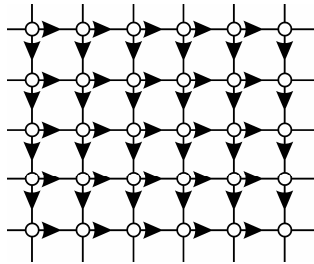
1 个四阶比赛网络如图 1 所示, 对有向边 e_{12} , 结点 v_2 只有 1 条指向 v_3 的边 e_{23} , 而结点 v_1 没有指向 v_3 的边, 因此结点 v_1 对结点 v_2 的复制率 $c_{12} = 0$. 对有向边 e_{31} , 结点 v_1 有 2 条边分别指向 v_2 和 v_4 , 结点 v_3 有 1 条边指向 v_4 , 但没有指向 v_2 的边, 因此结点 v_3 对结点 v_1 的复制率 $c_{31} = 0.5$. 对有向边 e_{14} , 结点 v_4 有 1 条边分别指向 v_2 , 结点 v_1 也有 1 条边指向 v_2 , 因此结点 v_1 对结点 v_4 的复制率 $c_{14} = 1$. 那么整个网络复制率 $c = (c_{12} + c_{14} + c_{23} + c_{31} + c_{34} + c_{42}) / 6 = (0 + 1 + 0 + 0.5 + 0 + 0) / 6 = 0.25$.

图 1 四阶比赛图($c = 0.25$)

2 有向模型网络及其复制率

2.1 有向规则网络

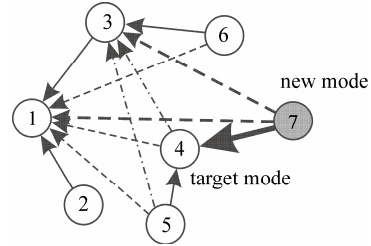
图 2 是 1 个二维无限规则有向网络, 任何 1 个第 i 行、第 j 列的结点 v_{ij} , 只有 2 条分别指向结点 $v_{i(j+1)}$ 和结点 $v_{(i+1)j}$ 的边, 而其入度边的起始结点都不指向其出度边的终点, 因此网络结点的复制率为 0, 整个有向网络的复制率 $c = 0$.

图 2 二维无限规则有向网络($c = 0$)

2.2 复制增长模型网络

鉴于复制增长是有向网络的一个内在机制, 由此提出了一个部分复制增长模型, 具体算法: (1) 随机选择 1 个旧结点复制为新结点; (2) 新结点指向旧结点连 1 条线; (3) 新结点以概率 p 复制旧结点指向更老结点的连线。当 $p = 1$ 时, 该模型就成了 Krapivsky 等^[7]提出的完全复制增长模型, 其演化过程如图 3 所示。新结点按加入顺序编号, 开始有 1 个根结点, 每步执行下述操作。例如第 7 步, 随

机选择到靶结点 4 进行复制, 复制得到的黑色新结点 7 首先与靶结点 4 连一条线, 再与靶结点 4 指向的更老结点 3 和 1 分别连线。但不与结点 5 连线, 因为结点 5 比靶结点 4 年轻。换句话说, 灰色新结点每步只复制靶结点的出度连线。

图 3 完全复制有向网络($c = 1$)

对于以概率 p 连线的部分复制模型, 任意 1 个结点复制率 $c_{ij} = p$, 网络复制率 $c = p$, 当 $p = 1$ 时, 有 $c_{ij} = 1$, $c = 1$.

对部分全复制模型, 1 个具有入度 $k_i(t)$ 的结点和其 k_i 个后面有连线指向它的结点被选中的概率为 $(pk_i + 1) / t$. 因此结点 i 的入度 $k_i(t)$ 应该近似地满足下述动力学方程:

$$\partial k_i / \partial t = (pk_i + 1) / t, \quad k_i(i) = 0,$$

上述动力学方程的解是:

$$k_i(t) = ((t/i)^p - 1) / p.$$

现在让结点 i 在现有结点中随机地选择, 可得网络瞬时度分布如下:

$$\begin{aligned} P\{k_i(t) < k\} &= P\{i > (t/(pk+1))^{1/p}\} = \\ &= 1 - (t/(pk+1))^{1/p} / t, \\ P_{in}(k, t) &= \frac{\partial P\{k_i(t) < k\}}{\partial k} = \frac{1}{(pk+1)^{1/p+1}} \sim k^{-(1+1/p)}. \end{aligned} \quad (3)$$

当 $p = 1$ 时, 有:

$$P_{in}(k, t) = \frac{\partial P\{k_i(t) < k\}}{\partial k} = \frac{1}{(k+1)^2} \sim k^{-2}. \quad (4)$$

3 自然数网络及其复制机制

3.1 自然数整除网络的定义

给定自然数集合的 1 个子集: $I_N = \{1, 2, 3, \dots, N\}$, 子集中每个元素表示 1 个结点, $\forall i, j \in I_N$, 且 $i \neq j$, 若 i 能被 j 整除, 即 $j \% i = 0$, 则连接 1 条从结点 j 指向结点 i 的有向线段。例如取 $N = 6$, 则有 $I_6 = \{1, 2, 3, 4, 5, 6\}$, 对于自然数 2, 其能被 4 和 6 整除, 则连接 2 条有向线段, 分别是结点 4 指向结点 2、结点 6 指向结点 2, 其他结点的连线以此类

推,可形成1个规模为6的有向网络,称之为如图4所示的自然数整除网络^[8].

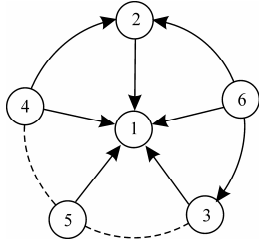


图4 自然数整除网络

自然数整除网络具有以下特点:(1)该网络是1个有向网络;(2)在给定 n (网络规模)的条件下是1个有限网络,但当 $n \rightarrow \infty$ 时,记 $I = \lim_{n \rightarrow \infty} I_n = \{1, 2, 3, \dots, n, \dots\}$,变成了真正意义上的无限有向网络.

3.2 网络入度分布的理论推导

对 $\forall x \in \mathbb{R}$, 都有 $[x] \leq x < [x] + 1$ 或 $x - 1 < [x] \leq x$, 其中 $[x]$ 表示不超过 x 的最大整数. 给定规模为 n 的自然数整除网络 $I_n = \{1, 2, 3, \dots, n\}$, 若 $0 \leq k \leq n-1$ 且 $i \in I_n$, 当 $[n/(k+2)] < i \leq [n/(k+1)]$ 时, 都有 $(k+1)i \leq (k+1)[n/(k+1)] \leq (k+1)n/(k+1) = n$ 且 $(k+2)i \geq (k+2)[n/(k+2)] > (k+2)n/(k+2) = n$, 对 $[n/(k+2)] < i \leq [n/(k+1)]$ 的自然数其入度都为 k , 可得入度 k 的结点所占比例为 $P_n(k) = \{[n/(k+1)] - [n/(k+2)]\} / n$, 因此对无限自然数整除网络 $I = \lim_{n \rightarrow \infty} I_n = \{1, 2, 3, \dots, n, \dots\}$, 入度 k 的结点分布为:

$$P(k) = \lim_{n \rightarrow \infty} P_n(k) = \lim_{n \rightarrow \infty} \frac{[n/(k+1)] - [n/(k+2)]}{n},$$

而 $(n/(k+1) - n/(k+2) - 1)/n < ((n/(k+1)) - (n/(k+2))) / n < (n/(k+1) - n/(k+2) + 1)/n$, 且

$$\lim_{n \rightarrow \infty} \frac{n/(k+1) - n/(k+2) - 1}{n} = \lim_{N \rightarrow \infty} \frac{n/(k+1) - n/(k+2) + 1}{n} = \frac{1}{(k+1)(k+2)},$$

由夹逼准则可得:

$$\lim_{n \rightarrow \infty} \frac{[n/(k+1)] - [n/(k+2)]}{n} = \frac{1}{(k+1)(k+2)},$$

即:

$$P(k) = 1 / ((k+1)(k+2)), \quad (5)$$

则无限有向网络的平均入度为:

$$\bar{k} = \sum_{k=0}^{\infty} kP(k) = \sum_{k=1}^{\infty} \frac{k}{(k+1)(k+2)} = \sum_{k=1}^{\infty} \frac{(k+1)-1}{(k+1)(k+2)} = \sum_{k=1}^{\infty} \frac{1}{(k+2)} -$$

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)(k+2)} = \sum_{k=3}^{\infty} \frac{1}{k} - \frac{1}{2} = \sum_{k=1}^{\infty} \frac{1}{k} - 2 = \lim_{n \rightarrow \infty} \ln(n+1) + r - 2,$$

其中, $r = 0.577\ 215\ 664\ 9$ 为 Euler 常数.

由此说明平均入度以对数方式增长, 自然数 n 值越大, n 是素数的可能性越小, 即素数的分布越往后越稀疏.

通过模拟复制增长模型网络, 并统计网络入度、出度分布, 模拟结果如图5所示.

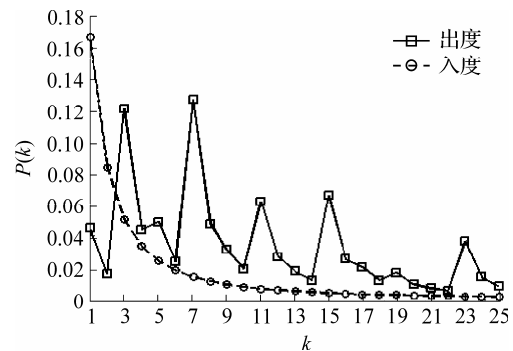


图5 自然数整除网络度分布

图5中虚线表示入度分布, 近似呈现幂律分布, 实线表示出度分布, 出度为1的结点分布反映了素数的分布情况. 而对于出度分布, 不具有幂律特征, 但值得注意的是, 线往往在 $4n-1$, $n=1, 2, \dots$ 处有较大峰值, 如图中的3, 7, 11, 15, 19, 23等, 而这其中大多数是素数, 那么是否与素数定理之间有一定的联系呢? 在文献[8]中, 史定华等人则给出了复杂网络研究素数定理的一些初步结论.

3.3 自然数整除网络与复制模型

前面(4)式得到 $P_{in}(k) = 1/(k+1)^2 \sim k^{-2}$, 这说明复制加速增长模型的网络入度分布是一个具有衰减指数 $\gamma=2$ 的幂律分布, 它的平均度近似地按照对数 $\ln t$ 增长, 如图6中黑色直线.

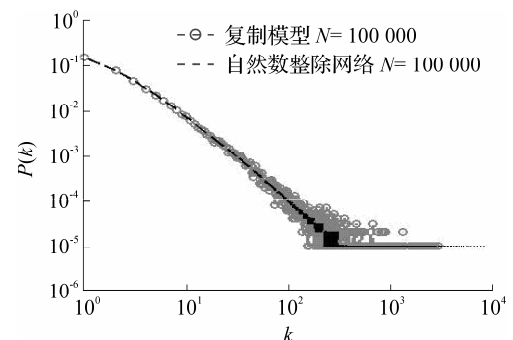


图6 自然数整除网络与复制模型网络度分布

现将自然数整除网络和复制增长模型网络的度分布图放在一起, 取 $N = 10^5$, 模拟结果如图 6 所示(双对数坐标). 从图 6 中不难发现, 自然数整除网络与复制模型网络的入度分布具有高度的相似性, 其头部都呈现弯曲特征, 而尾部都呈现 1 条直线. 这主要是因为复制增长模型的增长机制和自然数整除网络的连接机制具有一定的相似性, 比如在自然数整除网络, 结点 8 与结点 4 连接, 则结点 4 的孩子结点 1 和 2, 也能被结点 8 整除, 故与结点 8 也连接, 相当于对结点 4 的孩子结点的一个完全复制过程, 尤其是网络的复制率都为 1.

4 总结

复杂网络是当前研究的一个重要关注点, 择优增长作为无标度网络形成的内在机制, 在无向网络拓扑分析中, 大量的统计量被广泛的研究. 复制增长往往是有向网络演化的重要机制, 度分布不能完全代表网络拓扑, 但又缺少统计量来对有向网络内在规律的研究, 若将有向网络简单处理成无向网络, 必然会导致相关信息的丢失, 甚至会产生错误的结果. 为此, 基于有向网络的复制增长特征, 笔者提出了有向网络复制率的概念, 结合有

向网络的入度分布, 能很好地解释复制模型网络和自然数网络之间的相关性. 然而, 该统计量没有考虑到带有权值的有向网络, 这值得进一步深入研究.

参考文献:

- [1] Barabási A L, Albert R and Jeong H. Mean-field theory for scale-free random networks[J]. Physica A, 1999, 272:173-187.
- [2] Du Yuxian, Gao Cai, Hu Yong, et al. A new method of identifying influential nodes in complex networks based on TOPSIS[J]. Physica A, 2014, 399:57-69.
- [3] Newman M E J. Assortative mixing in networks[J]. Phys Rev Lett, 2002, 89(20):1-5.
- [4] 毛小燕. 增长网络结点度相关性的混合系数[J]. 宁波大学学报: 理工版, 2011, 24(4):92-96.
- [5] Gnana T, Mahendra P, Dharshana K, et al. Node assortativity in complex networks: An alternative approach [J]. Procedia Computer Science, 2014, 29:2449-2461.
- [6] Broder A, Kumar S R, Maghoul F, et al. Graph structure in the web[J]. Computer Networks, 2000, 33:309-320.
- [7] Krapivsky P L, Redner S. Network growth by copying[J]. Phys Rev E, 2005, 36118(71):1-7.
- [8] 史定华, 周晖杰. 自然数网络和素数定理[J]. 复杂系统与复杂性科学, 2010, 7(4):52-54.

Copying Mechanism of Directed Networks and Relevant Networks Analysis

MAO Xiao-yan, ZHOU Hui-Jie

(College of Science and Technology, Ningbo University, Ningbo 315212, China)

Abstract: Directed networks contain much more internal information than undirected ones. Many of the basic statistics for complex networks can only be applied to undirected networks, causing relative lack in the research on directed ones. To remedy this shortcoming, the paper proposes the concept of copying rate and analyzes the effectiveness of the statistics for directed networks. Based on statistics of in-degree distribution and copying rate for directed networks, the directed regular networks are selected for study purposes, including both the copying model networks and natural number network. On the rigorous derivation in in-degree distribution and copying rate of directed networks, the result reveals that completely copying model and natural number network have a power-law in-degree distribution with $\gamma = 2$, and copying rate for completely copying model and natural number network is 1, which explains the correlation between completely copying model and natural number network.

Key words: directed networks; copying model; natural number network; in-degree distribution; copying rate

(责任编辑 章践立)