

一种基于边数随机增长的 BA 网络模型*

王金龙^{a,b}, 刘方爱^{a,b}

(山东师范大学 a. 信息科学与工程学院; b. 山东省分布式计算机软件新技术重点实验室, 济南 250014)

摘要: 针对基本 BA 网络模型模拟现实网络的局限性, 采用新节点的度数增加服从泊松分布的随机函数来代替传统 BA 模型中的常数, 修改了节点度数的增长方式, 提出了一种基于边数随机增长的改进网络模型。从理论分析和实验验证两方面对该模型进行了分析和研究。理论证明过程中, 利用率方程法计算度分布并得到了解析解, 而计算机仿真则验证了度分布解析解的正确性。仿真结果表明改进后的网络模型符合现实网络的一定特性。

关键词: 复杂网络; BA 网络模型; 幂律分布; 泊松分布

中图分类号: TP393 **文献标志码:** A **文章编号:** 1001-3695(2013)05-1475-02

doi:10.3969/j.issn.1001-3695.2013.05.049

Improved BA network model with randomly growing edges

WANG Jin-long^{a,b}, LIU Fang-ai^{a,b}

(a. School of Information Science & Engineering, b. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Shandong Normal University, Jinan 250014, China)

Abstract: Due to the limitation of BA network model in the research on the real word networks, this paper used a random number which followed Poisson distribution as the number of edges of the new node to replace the original constant in BA mode. It modified the degree distribution of the node in a way, and proposed an improved model. It calculated the analytical solution of the degree distribution. Through the computer simulation results, it proved the correctness of the distribution of analytical solutions. Simulation results show that the improved network model conforms to certain characters of the real word networks.

Key words: complex network; BA network model; power law distribution; Poisson distribution

现实生活中, 万维网、因特网、交通网、电力网、科学合作网、生物网、社会网等各种网络遍布人们生活的各个方面, 而这些网络都可以称之为复杂网络。复杂网络可以用来描述从技术到生物直至社会各类开放复杂系统的骨架, 而且是研究其拓扑结构和动力学性质的有力工具。近年来, 复杂网络已成为国内外一个引人注目的新兴研究领域, 特别是小世界网络和无标度概念的提出, 更是掀起了一轮新的研究热潮。为了研究网络的各种特性, 人们提出了很多网络模型来描述现实网络, 其中以 Erdős 等人^[1]提出的随机网络模型、Watts 等人^[2]提出的小世界网络模型以及 Barabási 等人^[3]提出的无标度网络模型为代表。

网络可以看成是由节点和连线组成, 其中节点表示系统的单个元素, 两节点的连线表示元素之间的相互关系。节点的度可定义为该节点与其他节点的连线数目。 $P(k)$ 表示网络中一个随机节点的度恰好为 k 的概率, 则网络中节点的度的分布情况可用分布函数 $P(k)$ 来描述。如果节点的度分布函数具有幂律形式, 度没有明显的特征长度, 则网络被称为无标度网络。近年来大量实证研究表明^[4,5], 包括信息网络、社会网络以及生物网络等在内的许多复杂网络都具有无标度特性。国内外学者利用无标度网络理论对很多现实网络作出了各个方面的大量分析和研究^[6-10], 探索无标度网络背后的形成机理和组织原则, 并建立恰当的模型去拟合现实网络仍然是一个研究热点。

1 传统 BA 网络模型及存在的问题

Barabási 等人^[3]在研究万维网时发现了无标度特性的产生主要取决于增长和择优两条网络演化机理, 并指出两条演化机理缺少任何一个都不能产生无标度特性, 从而揭示了幂律分布的形成机理。

1.1 传统 BA 模型构造算法

Barabási 等人同时给出了一个无标度网络模型 (BA 模型), 具体构造算法如下:

a) 增长。给定 m_0 个节点的网络, 每次引入一个新的节点, 并且连到 m 个已存在的节点上, $m \leq m_0$ 。

b) 择优。一个新节点与一个已经存在的节点 i 相连接, 其概率为 $\Pi(k_i) = k_i / \sum_j k_j$, 其中 k_i 表示旧节点 i 的度数。

1.2 BA 模型扩展及存在的问题

为了更好地模拟现实世界的网络, 人们在 BA 模型基础上给出了很多扩展模型, 其中一类重要的模型为增长网络模型, 主要分成两种:

a) 边数 m 为固定常数, 而择优概率 $\Pi(k_i)$ 发生改变, 这种模型由于每个时间步长边数增长速度恒定, 所以被称为平稳增长网络模型, 主要有随机和择优混合模型^[11]和适应度模型^[12]。

b) 择优概率 $\Pi(k_i)$ 不变, 仍然采用 BA 模型的择优概率, 边数 $m = m(t)$ 为关于时间 t 的函数。这种模型由于边数增长

收稿日期: 2012-08-29; 修回日期: 2012-10-01 基金项目: 国家自然科学基金资助项目(90612003); 山东省科技发展计划资助项目(2011GGH20123)

作者简介: 王金龙(1979-), 男, 山东临沂人, 博士研究生, 主要研究方向为计算机网络、复杂网络(wjl_1979lx@163.com); 刘方爱(1962-), 男, 教授, 博导, 主要研究方向为网络与网络资源管理、分布式计算。

速度不是恒定的,所以被称为非平稳增长模型,主要有幂律增长模型^[13] ($m(t) = mt^\theta$) 和对数增长模型^[14] ($m(t) = m \ln t$)。

以上可以看出:对于平稳增长模型,每个新节点加入的边数都相同;而非平稳增长模型的边数增长则是关于时间 t 的增长函数。无论哪种模型在具体描述真实网络时虽然具有一定的现实意义,但是现实网络中边数的增长既不是常数也无法用精确函数来描述。实际网络中,当一个节点加入后,和已有的节点建立联系时,往往增加的边数可能是随机的。例如,每一个新成员加入到一个社区后,建立朋友关系的个数总是不一定的,带有很大的随机性。为了更加符合现实网络特性,本文考虑用边数 m 为服从泊松分布的随机函数来代替平稳增长模型中的常数和而非平稳增长中的确定函数,建立一个改进的 BA 模型,并考察模型的度分布的情况。本文首先给出改进模型,然后在理论上分析了度分布的解析式,进一步给出仿真结果,并与标准的 BA 模型作了比较分析,并验证了理论的正确性。

2 基于边数随机增长的 BA 模型

根据现实网络特点,在标准的 BA 网络模型基础上加以改进,假定边数增长函数为一个随机数 $m(t)$ 服从参数为 λ 的泊松分布,其算法构造如下:

a) 增长。给定 m_0 节点的网络,每次引入一个新的节点,并且连到 $m(t)$ 个已存在的节点上,其中 $m(t) \leq m_0$ 为一个随机数,且 $m(t)$ 服从参数为 λ 的泊松分布 $P(\lambda)$ 。

b) 择优。一个新节点与一个已经存在的节点 i 相连接,其概率为 $\Pi(k_i) = k_i / \sum_j k_j$,其中 k_i 表示旧节点 i 的度数。

定理 1 基于边数随机增长的 BA 网络模型,网络度分布为

$$P(k) = \frac{2e^{-\lambda}}{(k+2)(k+1)k} \sum_{i=0}^{k-1} \frac{\lambda^{i+1}(i+2)}{i!} \quad k=1,2,\dots$$

证明 令 $N_k(t)$ 表示在 t 时间步时,网络中有 k 条连线的节点数, $m(t) \sim P(\lambda)$ 。设 $A = A(k, m(t))$ 服从两点分布,当 $m(t) = k$ 时 $A=1$,其概率为 $P(m(t)=k)$;当 $m(t) \neq k$ 时 $A=0$,其概率为 $1 - P(m(t)=k)$ 。由率方程法可知, $N_k(t)$ 的增长速率为

$$\frac{d(N_k(t))}{dt} = E(m(t)) \left[(k-1) \frac{N_{k-1}(t)}{\sum_k k N_k(t)} - k \frac{N_k(t)}{\sum_k k N_k(t)} \right] + E(A) \quad (1)$$

其中: $E(m(t))$ 、 $E(A)$ 分别为随机变量 $m(t)$ 和 A 的数学期望。

由大数定律可知: $N_k(t) \approx tP(k)$, $\sum_k k N_k(t) = 2\lambda t$ 。又 $E(m(t)) = \lambda$, $E(A) = P(m=k) = \frac{\lambda^k}{k!} e^{-\lambda}$,代入式(1)得

$$P(k) = \frac{(k-1)}{k+2} P(k-1) + \frac{\lambda^k}{(k+2)k!} e^{-\lambda} \quad (2)$$

两边同时乘以 $(k+2)(k+1)k$,并令 $Q(k) = (k+2)(k+1)kP(k)$,得差分方程

$$Q(k) = Q(k-1) + f(k), f(k) = \frac{\lambda^k(k+1)}{(k-1)!} e^{-\lambda}, Q(0) = 0 \quad (3)$$

取母函数 $u(x) = \sum_{k=1}^{\infty} Q(k)x^k$,代入差分方程(3),整理可得 $u(x) = \sum_{k=1}^{\infty} \left(\sum_{i=1}^{k-1} f(k-i) \right) x^k$,比较两式可得 $Q(k) = \sum_{i=1}^k f(k-i)$,因此有 $P(k) = \frac{2e^{-\lambda}}{(k+2)(k+1)k} \sum_{i=0}^{k-1} \frac{\lambda^{i+1}(i+2)}{i!}$ 。

定理 2 当 k 充分大时, $P(k) \approx \frac{2(\lambda^2 + 2\lambda)}{(k+2)(k+1)k} \propto 2(\lambda^2 + 2\lambda)k^{-3}$ 。

证明 由泰勒展开式定理可知 $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$,则当 k 比较大时, $e^{\lambda} \approx \sum_{n=0}^k \frac{\lambda^n}{n!}$,所以

$$\sum_{i=0}^{k-1} \frac{\lambda^{i+1}(i+2)}{i!} = \sum_{i=0}^{k-1} \left(\frac{\lambda^{i+1}}{i!} + \frac{2\lambda^{i+1}}{i!} \right) = \lambda^2 \sum_{i=1}^{k-1} \frac{\lambda^{i-1}}{(i-1)!} + 2\lambda \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \approx (\lambda^2 + 2\lambda)e^{\lambda} \quad (4)$$

将式(4)代入定理1得

$$P(k) \approx \frac{2(\lambda^2 + 2\lambda)}{(k+2)(k+1)k} \propto 2(\lambda^2 + 2\lambda)k^{-3} \quad (5)$$

可以看出,改进后的 BA 网络模型,当每个新节点增加的边数变为一个随机数,定理1说明当 k 比较小时,此时节点度分布要受到参数 λ 的影响, λ 取值越大,度分布受到的影响也越大,从而偏离幂律分布;定理2则表明,当 k 较大时,此时节点的度分布与参数 λ 的取值无关,依然服从幂律分布。文献[10]中增加的网络边数服从二项分布时也得到了类似结果,因此可以发现随机性的加入会影响到 BA 网络模型的度分布。

3 模型仿真分析

本文先对模型中主要参数加以说明: m_0 为初始网络的规模, N 为演化网络的最大规模, m 为原始 BA 模型的每个时间步长增加的边数, $m(t)$ 为改进模型的每个时间步长增加的边数, $P(\lambda)$ 为参数为 λ 的泊松分布。利用 MATLAB 分别对 BA 网络模型和改进后的无标度模型的度分布根据不同的参数进行仿真实验,坐标均采用了双对数坐标。

图1为网络规模 $m_0 = 5$, $N = 5000$, 边数 $m = 1$ 的 BA 模型和 $\lambda = 1$ 时的改进模型的度分布。由图1可以发现,当参数 $\lambda = 1$ 取值较小时,BA 模型和改进后的模型差别不大,依然服从幂律分布。图2为网络规模 $m_0 = 5$, $N = 5000$, 边数 $m = 3$ 的 BA 模型和 $\lambda = 3$ 时的改进模型的度分布。可以发现当 $\lambda = 3$ 时,度分布上表现为当 k 比较大时,改进的模型仍服从幂律分布,而当 k 较小时,则表现出了弯曲现象,不符合幂律分布。这与定理1和2的理论分析是相吻合的。

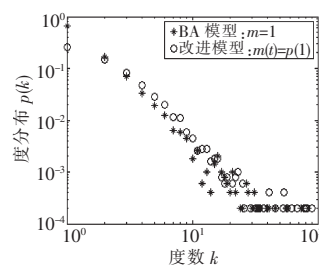


图1 BA模型与网络模型度分布比较

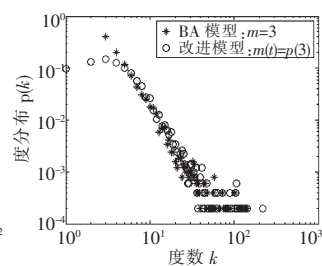


图2 BA模型与网络模型度分布比较

图3为参数 $\lambda = 3$ 不变,网络规模 N 分别为 1000、3000、5000 时的度分布。显然当 N 发生改变时,度分布基本上没有发生改变,依然保持着 k 较小时发生弯曲,而当 k 较大时服从幂律分布的特性,可见网络的规模不影响度的分布。图4为网络规模 $N = 5000$ 不变,参数 λ 分别为 1、2、3 时的度分布,通过图形发现参数 λ 取值越大, k 较小时弯曲越大,但是当 k 较大时,则与 λ 取值无关,仍然服从幂律分布。

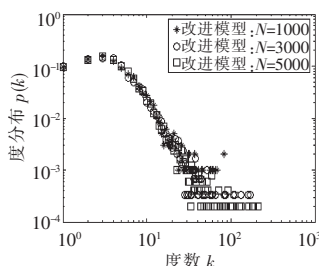


图3 不同规模的改进网络模型 ($m_0=5, m(t)=p(3)$)

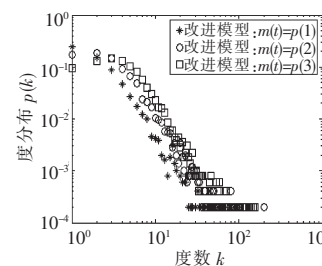


图4 不同参数的改进网络模型 ($m_0=10, N=5000$)

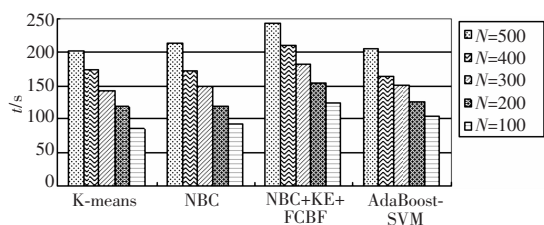


图7 四种算法计算复杂度仿真

6 结束语

基于机器学习的流量分类方法利用业务流的统计特征进行识别,无须关心数据包的内容,使得基于机器学习的流量分类成为该领域一个新兴的研究方向。立足于算法的计算复杂度和准确性,本文提出一种基于 AdaBoost-SVM 流量分类器。首先,该方法利用 K-L 变换从大量冗余特征中遴选出本征特征;然后基于 AdaBoost 将一次分类等分成若干层弱分类过程,其中每层构造了非线性 SVM 弱分类器,并通过分层组合和迭代权重的方法聚焦在分离域内容易被错误分类的样本上;最后基于 Moore 数据集对算法进行了仿真实验,验证了 AdaBoost-SVM 既保持了 K-means 算法计算简单的优势,又提高了分类器的性能。另外,AdaBoost-SVM 算法依赖全部已标记的样本流,这在实际操作中需要大量的人工操作。本文下一步将结合半监督学习机制,研究如何利用少量已标记样本提高流量分类的性能。

参考文献:

- [1] CNNIC. 中国互联网络发展状况统计报告 [EB/OL]. [2012-09-01]. <http://www.cnnic.cn>.
- [2] Cisco. Networking solutions white paper [EB/OL]. [2012-09-01].

(上接第 1476 页)

4 结束语

本文在 BA 无标度网络的基础上,考虑了网络增长时节点度数增加的随机性,通过引入了度数增长服从泊松分布的改进模型,修改了新节点度数增长的方式,并利用率方程法分析了改进后模型的度分布的解析解。利用 MATLAB 仿真验证了理论的正确性,并发现只有在度分布 k 较大时才服从幂律分布,而在度分布 k 较小时总是有规律的弯曲,这和 BA 模型明显不同。文献[4,8]表明度分布弯曲现象在一些现实网络中也存在,因此改进后的网络模型既保留了度分布服从幂律分布这一现实网络特性,又比 BA 模型更好地拟合现实网络,利用它来刻画和分析现实中的复杂网络具有一定的研究价值。同时该模型也存在不足:a)模型中仅考虑了边数增加的随机性,而没有考虑节点数增加的随机性,考虑影响度分布的因素过于单一;b)模型中虽然体现了度分布弯曲现象,但是没有揭示现象背后的本质原因。因此,在下一步的工作中,将对网络模型作进一步修改,增加节点数的随机性使之更加符合现实网络,同时进一步分析度分布的弯曲现象,并试图揭示背后的原因。

参考文献:

- [1] ERDÖS P, RÉNYI A. On the evolution of random graphs[J]. *Publications of the Mathematical Institute of the Hungarian Academy of the Sciences*, 1960(5): 17-61.
- [2] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks [J]. *Nature*, 1998, 393: 440-442.

http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.

- [3] 杨家海,吴建平,安常青. 互网络测量理论与应用 [M]. 北京:人民邮电出版社,2009.
- [4] KARAGIANNIS T, BROID A. Transport layer identification of P2P traffic [C]//Proc of the 4th ACM SIGCOMM Internet Measurement Conference. 2004: 121-134.
- [5] SUBHABRATA S, SPATSCHECK O. Accurate, scalable in-network identification of P2P traffic using application signatures [C]//Proc of the 13th International World Wide Web Conference. 2004: 512-521.
- [6] CASCARANO N, CIMINIERA L, RISSO F. Improving cost and accuracy of DPI traffic classifiers [C]//Proc of ACM Symposium on Applied Computing. 2010: 641-646.
- [7] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning [C]//Proc of the 30th IEEE Conference on Local Computer Networks. 2005: 15-17.
- [8] ROUGHAN M, SEN S, SPATSCHECK O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification [C]//Proc of the 4th ACM SIGCOMM Internet Measurement Conference. 2004: 135-148.
- [9] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques [C]//Proc of SIGMETRICS. 2005: 50-60.
- [10] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119-139.
- [11] MOORE A W, ZUEV D. Discriminators for use in flow based classification, RR-05-13 [R]. London: Queen Mary University, 2005.
- [12] HAN Jian-wei, KAMBER M. 数据挖掘概念与技术 [M]. 北京:机械工业出版社,2007.
- [13] THEODORIDIS S, KOUTROUMBAS K. Pattern recognition [M]. 3rd ed. Beijing: Publishing House of Electronics Industry, 2010.

- [3] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks [J]. *Science*, 1999, 286(5439): 509-512.
- [4] ALBERT R, BARABASI A L. Statistical mechanics of complex networks [J]. *Review of Modern Physics*, 2002, 74(1): 47-97.
- [5] CLAUSET A, SHALIZI C R, NEWMAN M E J. Power-law distributions in empirical data [J]. *SIAM Review*, 2009, 51(4): 661-703.
- [6] BARABASI A L. Scale-free networks: a decade and beyond [J]. *Science*, 2009, 325(5939): 412-413.
- [7] 林兵,郭文忠,陈国龙,等. 标度网络中基于最短路径免疫策略的病毒传播研究 [J]. *计算机科学*, 2012, 39(6): 136-138.
- [8] 窦炳琳,李澍淦,张世永. 基于结构的社会网络分析 [J]. *计算机学报*, 2012, 35(4): 741-752.
- [9] 何敏华,张端明,王海艳,等. 基于无标度网络拓扑结构变化的舆论演化模型 [J]. *物理学报*, 2010, 59(8): 5175-5180.
- [10] 王建伟,荣莉莉,于凯. 基于节点批量生长机制的无标度网络演化模型 [J]. *系统工程学报*, 2010, 25(5): 581-585.
- [11] 陈琴琴,陈丹青. 基于二项分布随机增长的无标度网络模型 [J]. *数学研究*, 2010, 43(2): 185-191.
- [12] LIU Zong-hua, LAI Ying-cheng, YE Nong, et al. Connective distribution and attack tolerance of general networks with both preferential and random attachments [J]. *Physics Letters A*, 2002, 303(5-6): 337-344.
- [13] BIANCONI G, BARABASI A L. Competition and multiscaling in evolving networks [J]. *Europhysics Letters*, 2001, 54(4): 436-442.
- [14] DOROGOVTSY S N, MENDES J F F. Effect of the accelerating growth of communication networks on their structure [J]. *Physical Review*, 2001, 63(2): 025101.
- [15] SHI Ding-hua, CHEN Qing-hua, LIU Li-ming. Markov chain-based numerical method for degree distribution of growing networks [J]. *Physical Review*, 2005, 71(3): 036140.